Journal of Universal Computer Science, vol. 26, no. 4 (2020), 434-453
submitted: 4/11/19, accepted: 7/3/20, appeared: 28/4/20 CC BY-ND 4.0

Speaker/Style-Dependent Neural Network Speech Synthesis Based on Speaker/Style Embedding

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

Darko Pekar
(AlfaNum Speech Technologies Ltd., Novi Sad, Serbia
darko.pekar@alfanum.co.rs)

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

Anton Smirnov
(AlfaNum Speech Technologies Ltd., Novi Sad, Serbia
anton.smirnov@alfanum.co.rs)

Tijana Nosek
(University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia
tijanadelic@uns.ac.rs)

Abstract: The paper presents a novel architecture and method for training neural networks to produce synthesized speech in a particular voice and speaking style, based on a small quantity of target speaker/style training data. The method is based on neural network embedding, i.e. mapping of discrete variables into continuous vectors in a low-dimensional space, which has been shown to be a very successful universal deep learning technique. In this particular case, different speaker/style combinations are mapped into different points in a low-dimensional space, which enables the network to capture the similarities and differences between speakers and speaking styles more efficiently. The initial model from which speaker/style adaptation was carried out was a multi-speaker/multi-style model based on 8.5 hours of American English speech data which corresponds to 16 different speaker/style combinations. The results of the experiments show that both versions of the obtained system, one using 10 minutes and the other as little as 30 seconds of target data, outperform the state of the art in parametric speaker/style-dependent speech synthesis. This opens a wide range of application of speaker/style dependent speech synthesis based on small quantities of training data, in domains ranging from customer interaction in call centers to robot-assisted medical therapy.

Keywords: Deep neural networks, Embedding, Speaker adaptation, Text-to-speech synthesis
Categories: H.1.2, H.5.2, I.2.4, I.2.6

1 Introduction

Innovative approaches in the research and development in the area of speech synthesis in the last two decades have led to a breakthrough in both the quality of

synthesized speech and the flexibility of the system. Namely, classical text-to-speech systems relied on concatenation of speech segments and were thus able only to reproduce speech in the voice and speech style of the speaker who provided the speech data used by the system [Hunt & Black, 96]. The first major step forward was the introduction of statistical parametric speech synthesis based on Hidden Markov models (HMM), which was able to learn from speech data instead of merely reproducing it [Yoshimura, 99]. This approach overcame the necessity for the existence of very large speech corpora [Morioka, 04; Gutkin, 10], and offered a range of new possibilities for synthesis in different voices or speaking styles [Tamura, 98; Yamagishi, 04; King, 08; Yamagishi, 09; Kanagawa, 13; Trueba, 13; Ohtani, 15]. However, largely due to the limited capability of HMMs to generalize from the training data, which led to inferior accuracy of acoustic models and a tendency to over-smooth acoustic parameters [Zen, 09], the status of HMMs as the state of the art in speech synthesis began to fade with the advent of deep neural networks.

The application of deep neural networks (DNN) to speech synthesis was inspired by the assumption that a human speech production system transforms information from the linguistic level to its acoustical representation through a layered hierarchical structure [Yu & Deng, 11]. Deep neural networks were initially used to improve the performance of HMM-based TTS, e.g. by more sophisticated modelling of speech excitation using autoencoders [Vishnubhotla, 10]. However, they were soon found to directly outperform HMMs in acoustic modelling, owing to their superior ability to learn complex mappings from the linguistic representation of information to the corresponding acoustic features [Zen, 13]. A systematic review on the use of DNNs in acoustic modelling for speech synthesis can be found in [Ling, 15]. Approaches aimed at avoiding the parameterization process and working with speech waveforms directly have also been proposed recently [van den Oord, 16; Wang, 17], improving the quality of synthesized speech still further. A significant part of the research community focused their attention on exploiting the potential of DNNs to modify speaker characteristics [Nakashika, 13; Wu, 13], or constructing models and architectures that will support synthesis in multiple voices or speaking styles. This line of research was encouraged by market needs for applications such as virtual assistants, smart homes and intelligent robots [Ondáš, 13], which are often required to use different voices or to accommodate their speaking style to the context of the communication, in order to create an impression that the system has emotions and that it empathizes with the user [Picard, 03, Eide, 04]. A multitask learning framework based on a DNN with shared hidden layers and multiple speaker-dependent output layers has been proposed in [Fan, 15], while different speaker-adaptation methods for DNNs have been investigated in [Wu, 15]. A DNN architecture with additional speaker-dependent inputs was proposed in [Hojo, 16], and this approach was further extended by supplementing the input information by speaker gender and age [Luong, 17]. To enable the network to reproduce the voice of a particular speaker in a style that is absent from the training corpus (which is referred to as emotion or style transplantation), the research presented in [Inoue, 17] proposed a network architecture which explicitly separates speaker and speech style contributions, while the one presented in [Suzić, 19] built on the multi-speaker DNN with shared hidden layers proposed in [Fan, 15], by extending it with a single style-dependent input and introducing an additional bottleneck layer. Other lines of research, such as the one

presented in [Parker, 18], focused on the development of methods for adaptation of a multi-style single-speaker DNN to a new speaker's voice. In one way or another, all these approaches address the practical impossibility of recording and processing a new training speech corpus for each new speaker/style combination for which the need may arise.

This research proposes a DNN architecture and a two-step training procedure aimed at obtaining speaker/style-dependent speech synthesis based on very small quantities of training data by the target speaker and in the target speech style, which produces synthesized speech of very good quality even in case both the style and the speaker are withheld from the network during the training of the multi-speaker multi-style model. This has been made possible owing to embedding, which is a powerful deep learning technique based on mapping discrete (often binary) vectors from a high-dimensional space to continuous vectors in a low-dimensional space, and which has been used for a variety of machine learning tasks ranging from text tagging [Wang, 15] to automatic image captioning [Mao, 14]. In the context of speech synthesis, both speaker and speech style are traditionally represented as one-hot vectors, which can be considered an ignorant representation, since the similarity of two voices is not related in any way to the distance between corresponding points in the high-dimensional space [Lorenzo-Trueba, 18]. The architecture and training procedure presented in this paper overcome this deficiency by performing joint embedding of the speaker and style, representing them in a low-dimensional space in a more logical way, which helps the network to efficiently generalize on unseen speech data. To use the available speech data even more economically, the embedding is jointly performed not only on speaker and style ID's, but on cluster ID's as well, where the term "cluster" refers to the portion of a speaker/style dependent speech corpus which is consistent in terms of acoustic and prosodic quality. Namely, one of the practical problems in obtaining a high quality speech corpus for training, which is rarely mentioned in the literature, is maintaining the consistency of the acoustic and prosodic quality of the voice and speaking style, especially when the recording is performed in multiple sessions or the speaker takes a break within a session. This often results in parts of the corpus being slightly different in volume, timbre or even the particular way the speaker has chosen to render a speech style (e.g. "happy"). Rather than discarding the parts of a speech corpus that deviate from the corpus segment that can be termed as "default", we have opted for dividing each speaker/style-specific speech corpus into consistent clusters. Consequently, instead of supplying two non-linguistic inputs to the network (speaker ID and speech style ID), now a third input (cluster ID) is added, and these three inputs are jointly represented as a single one-hot vector, which is converted into an appropriate joint embedding through the training procedure. The effects of the division of speech data into clusters have been analyzed in [Sečujski, 19], and it has been shown to slightly improve the quality of speech synthesis. The architecture and training procedure proposed in this research result in a multi-speaker multi-style text-to-speech synthesis able to reproduce speech of very high quality in any speaker/style/cluster combination present in the general training corpus, but is also easily adaptable to a new speaker/style, with a relatively small quantity of adaptation data needed. We have shown that 10 minutes of adaptation data are sufficient to achieve speech quality and voice similarity slightly above the state of the art in parametric speech synthesis, and

that with as little as 30 seconds of adaptation data it is possible to maintain virtually the same degree of voice similarity, although not without losing some of the general quality of synthesis.

The remainder of the paper is organized as follows. Section 2 briefly describes the speech corpora used in the experiments. Section 3 gives a detailed description of the proposed DNN architecture and the training procedure for constructing a multi-speaker multi-style speech synthesis model, and describes the two-step procedure for adapting this model to a new speaker/style. Section 4 presents the results of objective and subjective evaluation of the ability of the proposed model to adapt to a new speaker/style, comparing it to two baseline models from the literature in terms of the quality of synthesis as well as speaker similarity. The concluding section of the paper summarizes the main findings and briefly presents the plans for future research.

2 Data

The data used to construct the multi-speaker multi-style speech synthesis model, as well as other models used in some of the experiments, consists of a total of 8 hours and 38 minutes of speech from 6 speakers, whose contributions varied in sizes, numbers of speech styles as well as acoustic quality, as shown in Table 1. All speech data was sampled at a rate of 22.05 kHz and 16 bits per sample were used.

Speaker	**Gender**	**Quality**	**Style**	**Time [hh:mm:ss]**	**Total time per speaker [hh:mm:ss]**
F1	female	studio	Neutral	01:30:03	02:32:59
			Apologetic	00:17:42	
			Happy	00:21:24	
			Promotional	00:23:50	
M1	male	studio	Neutral	01:38:07	03:34:11
			Angry	00:16:55	
			Apologetic	00:15:58	
			Happy	00:26:13	
			Promotional	00:28:04	
			Stern	00:28:54	
F2	female	studio	Friendly	00:31:42	01:00:25
			Promotional	00:28:43	
M2	male	studio	Friendly	00:18:26	00:39:46
			Promotional	00:21:20	
F3	female	source: YouTube	Neutral	00:26:46	00:26:46
M3	male	source: YouTube	Neutral	00:24:17	00:24:17
				Total time [hh:mm:ss]:	08:38:24

Table 1: Speech corpora used for construction of multi-speaker multi-style model ("time" refers to the time left when leading and trailing silences are trimmed and silent phonetic segments, such as mid-phrase silences, excluded)

As can be seen, there are two principal speakers, M1 and F1, whose contributions include the largest number of speech styles, and whose contribution to the neutral style is the greatest. In the neutral segments of each of these two corpora 4 clusters were manually identified. This task was actually quite simple since all clusters contain contiguous utterances and the boundaries between clusters correspond to mid-session breaks as well as breaks between sessions. In order to avoid the bias towards M1 and F1, as well as to increase the basis for the multi-speaker multi-style model, the available speech data was artificially augmented by introducing changes in pitch and/or spectral envelope into the utterances of all 6 original speakers. Although the data obtained in this way is correlated with the original data, augmentation implicitly regularizes the model and improves its ability to generalize [Vapnik & Chervonenkis, 71]. As such, data augmentation as an approach to overcome data scarcity has been used from the earliest days of machine learning [Simard, 92]. Using different portions of the original speech corpora, as well as additional utterances from some of the original speakers, 10 new artificial speakers were created, bringing the total number of speakers to 16 (with 67 unique speaker/style/cluster combinations), and the total duration of the speech corpus to 21 hours and 50 minutes. Modification of an utterance consisted of extracting its spectral envelope using the WORLD vocoder [Morise, 16], extracting its pitch contour by an autocorrelation-based pitch-extraction algorithm based on [Pekar and Obradović, 01], and rescaling the spectral envelope, pitch contour and speech rate to the extent chosen so that speech resynthesized by the WORLD vocoder sounds like a different, yet natural speaker. In 7 of the 10 artificial speakers this procedure effected an apparent gender switch, but a rough balance between male and female speakers in the resulting speech corpus was maintained. In all cases the speech style ID was copied from the original speech data, and different clusters of the neutral speech style were created by performing slightly different modifications of the original data. The artificial speaker/style combinations created by augmenting speakers with less speech data available (F2, M2, F3 and M3) were obtained by modifying utterances that already appear in the original speech corpus, while speaker/style combinations created from F1 and M1 were obtained by modifying not only utterances from the original F1 and M1 corpora, but some previously unseen utterances as well, owing to greater availability of speech data for these speakers. The entire speech corpus was phonetically and prosodically annotated, with prosodic annotation following the extended Tone and Break Indices (ToBI) set of conventions, proposed in [Sečujski, 18].

In order to evaluate the ability of the system to adapt to a new speaker and style, two relatively small corpora were used, one containing speech from a female speaker (F4) and the other from a male speaker (M4). Both these corpora were withheld during the training of the multi-speaker multi-style model. The speech style in these two corpora can be termed as roughly neutral, although it should be noted that this information it actually irrelevant, since, just as the model is able to adapt to an unknown speaker, it is also able to adapt to an unknown style.

3 Models

This section will describe a novel DNN architecture and two-step training procedure aimed at obtaining speaker/style-dependent speech synthesis using very small training

datasets, and compare it with two other methods for speaker/style-dependent speech synthesis already described in the literature. All three methods are based on a cascade of two independent neural networks – one predicting phonetic segment durations, and the other predicting acoustic feature vectors for each frame. The principal input to both networks is the vector of 577 linguistic features extracted from text, related to the current phone. In the synthesis stage, the output of the duration model is used as supplementary input of the acoustic model, augmented with the information on the duration of particular hidden Markov model (HMM) states of each phone, which is obtained in the training phase from HMM models through the alignment procedure described in [Suzić, 17]. In all experiments each of the two networks has 4 hidden layers of size 1024, where the first three use tangent hyperbolic activation function, while the fourth one is composed of LSTM units. Stochastic gradient descent was used as optimizer in backpropagation algorithm, using one utterance as a batch. In other words, backpropagation occurs after the networks have seen one utterance, regardless of the number of phones (in the case of the first network) or frames (in the case of the second one).

3.1 Baseline method 1

The first method used as a baseline, presented in [Delić, 18], represents one of the simplest methods for creating a voice of new speaker with a very small amount of speech training data. Its main idea is to create a speaker-dependent text-to-speech (SD TTS) model, initially trained on a large speech corpus from one speaker, and then adapt it to another speaker with a very small quantity of training data. The adaptation process differs from the standard training of SD TTS [Zen 13; Delić, 17] only in the starting point, i.e. it starts from an already trained model instead of a randomly initialized one, and it proceeds in an identical way. It was shown that such an approach, using only 10 minutes of training data from the target speaker, produces results that are comparable to the results obtained from a regular SD TTS trained on a 3-hour speech corpus. Due to the limited availability of training data, the research presented in [Delić, 18] analyzed 2 SD TTS models: one based on a speech corpus from the male speaker M1 and the other based on a speech corpus from the female speaker F1, both in American English, which were identical to the ones used in this research. Since both corpora included multiple speech styles, the inputs to SD TTS models were extended with the information related to the style and cluster, both in the form of a one-hot vector, as was previously done in [Suzić, 18]. For the purpose of this research, speech data from the same two speakers, M1 and F1, was used to obtain two speaker-dependent TTS models that served as a basis for adaptation to the speakers M4 and F4, respectively.

3.2 Baseline method 2

The second method used as a baseline represents a slight modification of the approach described in detail in [Suzić, 18], where it is referred to as “separate output layer”. This approach builds upon the idea presented in [Fan, 15], which proposes an architecture based on shared hidden layers and multiple speaker-dependent output layers. In the second baseline approach the shared part of the network is assumed to represent a global linguistic feature transformation, while separate output layers are

used for different speaker/style combinations. In the adaptation phase only a specific speaker/style-dependent output layer is adapted using the limited speaker/style-specific data, following the adaptation procedure proposed in [Fan, 15]. The modification with respect to [Suzić, 18] lies in the introduction of an additional speaker/style-dependent hidden layer into the network structure. Similarly to the case of baseline model 1, the inputs are extended with the style and cluster codes in the form of one-hot vectors, but in this case all of the speech data listed in the Table 1. was used for training the multi-speaker/multi-style model that was subsequently adapted to M4 and F4.

3.3 The proposed method

With the idea of improving the multi-speaker model as a starting point for speaker/style adaptation, we supplement the inputs of both neural networks (one that predicts durations and the other, which predicts acoustic features) with the information about the speaker, speaking style and cluster in an embedded form, as shown in Fig. 1. As previously explained, both networks are presented with 577 binary linguistic features at their inputs, with the output of the duration network serving as an additional input for the network predicting acoustic features. However, in the proposed model the input layer of each network is extended with an *N*-dimensional vector containing the joint embedding of the speaker ID, speaking style ID and cluster ID, all of them originally represented in the form of a single one-hot vector of length

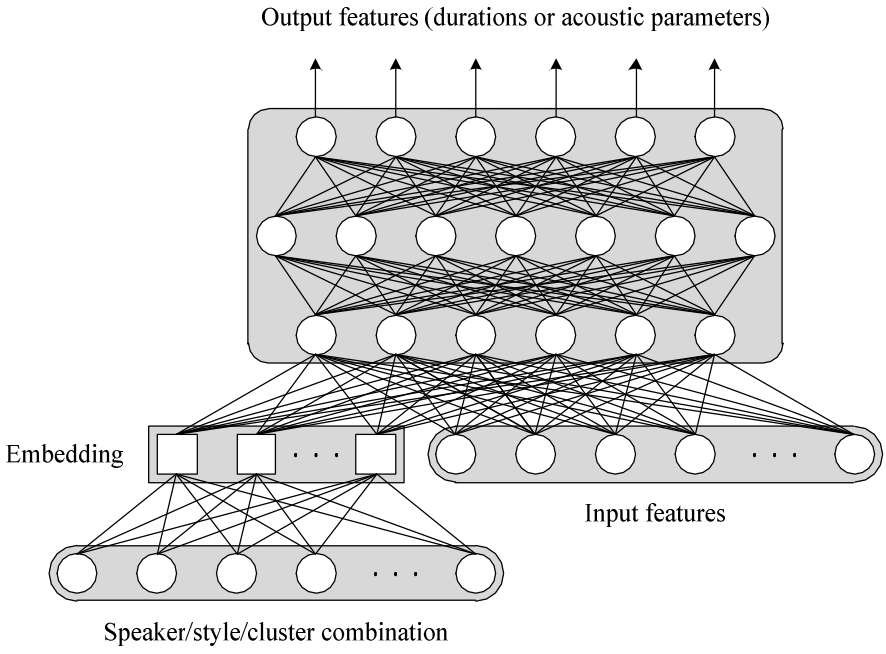

Figure 1: The architecture of the proposed model for the two neural networks that predict either phonetic segment durations or acoustic features.

67, which is the number of unique SSCs existing in the training corpus. In this way it is left to the network to represent a particular SSC in a space of lower dimensionality (in our research it was set to $N = 15$). The idea of representing the speaker, the style and the cluster using 3 separate one-hot vectors was discarded since it would imply the questionable assumption that every speaker renders a speaking style in a similar way. The main advantage of the approach based on embedding is that the network itself has the opportunity to establish similarities and differences between particular speakers, styles or clusters, and based on this information, it is expected to position particular speaker/style/cluster combinations (SSC) closer or further from each other in the embedding space. This, in turn, will help the main network to generalize more easily, since the distance between two SSCs in the embedding space will correspond to the general difference between them. Once trained, the network will be able to synthesize speech that corresponds to a particular SSC given the corresponding point in the embedding space. Furthermore, given a random point in the embedding space, the network will be able to produce a new, previously "unseen" voice.

In the initial training of the multi-speaker model, each of the networks is presented with the linguistic features as well as the one-hot vector representing the speaker, style and cluster combination (SSC) at the input, and with the corresponding values (durations or acoustic features) at the output. In this way the networks themselves perform the embeddings for each SSC, and as a result, each SSC will be mapped into two points in the corresponding embedding spaces – one related to phonetic segment durations and the other to acoustic features. In both cases it is expected that the proximity of two SSCs in either of the two embedding spaces will reflect their perceptual similarity. The result of the initial training is a multi-speaker model, able to provide speech that can sound like any SSC that was present in the training, given the corresponding embeddings in both networks.

A model which uses the embedded representation of SSCs can be adapted to a new voice and/or a new speaking style using a relatively small quantity of adaptation data, through a procedure that consists of two phases. The first phase is aimed at establishing the embedding for the new speaker/style, and it begins by random initialization of the values in the embedding layers of both networks. In this phase of the adaptation process, only the values in the embedding layers of both networks are adjusted through back-propagation while the rest of the networks is kept unchanged. The model with embedding layers adapted in such a way synthesizes speech that already resembles the target speaker/style to some extent. However, the quality of synthesized speech can be further improved through the second phase of the adaptation process, in which the same training data is used again, but the embedding layer is frozen, while the weights in the networks are modified according to the back-propagated error. In the experiments that will be described in the following section we demonstrate the ability of the initially trained multi-speaker multi-style model to synthesize speech in speaker/style combinations seen during the training, as well as its ability to synthesize speech in a previously unseen speaker/style combination (even for a previously unseen speaker and an unknown style) after the two-phase adaptation process. Through these experiments we also investigate the influence of a range of factors, most notably the relative importance of each phase of the proposed adaptation process as well as the quantity of target speech data available for adaptation.

4 Experiments

In this research the proposed model is trained on the same data as the two baseline models described in Section 3. However, while the multi-speaker models (baseline 2 and the proposed model) were trained on the entire database presented in Table 1, the baseline model 1, not being a multi-speaker model, was trained only on M1 and F1 in order to create two speaker-dependent models. To test the capability of all three models to adapt to an unknown speaker and an unknown style, speech data from speakers M4 and F4 were used for adaptation. Since the particular aim of this research is to investigate the case when the quantity of target speech data is very small, the adaptation experiments were conducted with databases containing 10 minutes and as little as 30 seconds of target speech data. For the adaptation of the baseline model 1 the initial speaker-dependent model of the same gender was used in each case. As the baseline 2 model actually encompasses 16 different speakers (6 of them real and 10 obtained by augmentation), those used as starting points for adaptation in this research were the ones that correspond to M1 or F1 (depending on the gender of the target speaker). In the proposed model, the dimension of the embedding was set to $N = 15$, although it has been shown that it is of surprisingly little importance to the performance of the synthesizer (values ranging from 4 to 40 were tested). The capability of the proposed model to synthesize speech that corresponds to the intended speaker/style was firstly evaluated through the objective distance between appropriate acoustic features of the original and synthetic speech, which was followed by a series of listening tests specifically aimed at establishing the relevance of the position of the SSC points in each of the two embedding spaces, the relevance of each phase in the two-phase adaptation process, as well as the influence of training data. Examples of speech samples used for both objective and subjective evaluation are available at the URL: www.alfanum.ftn.uns.ac.rs/embedding.

4.1 Objective evaluation

In order to objectively evaluate the three models, the values of objective acoustic parameters were compared between speech synthesized by each of the three models and original target speech data in case both phases of the adaptation process were carried out. The target speech data used for evaluation was withheld form all training phases. The acoustic parameters taken into account include the root mean square error (RMSE) and correlation for f_0, RMSE and correlation for the duration of phones as well as mel cepstral distance (MCD). The results are presented in Fig. 2.

Firstly, it can be seen that the correlation between the predicted f_0 contour and the actual one, as well as the correlation between the predicted phone durations and the actual ones, exhibit only minor differences among the three models, but that the proposed model consistently achieves the best performance, regardless of whether 10 minutes or 30 seconds of target speech data were used for adaptation. It can also be seen that the differences are slightly greater in case of the adaptation on less target speech data. The baseline model 1 appears to be the most sensitive to the decrease of the quantity of adaptation data, although the differences are not significant in this case either. The differences between the models are much more obvious in case of RMSE of f_0 and phone durations. In most cases the proposed model outperforms the two

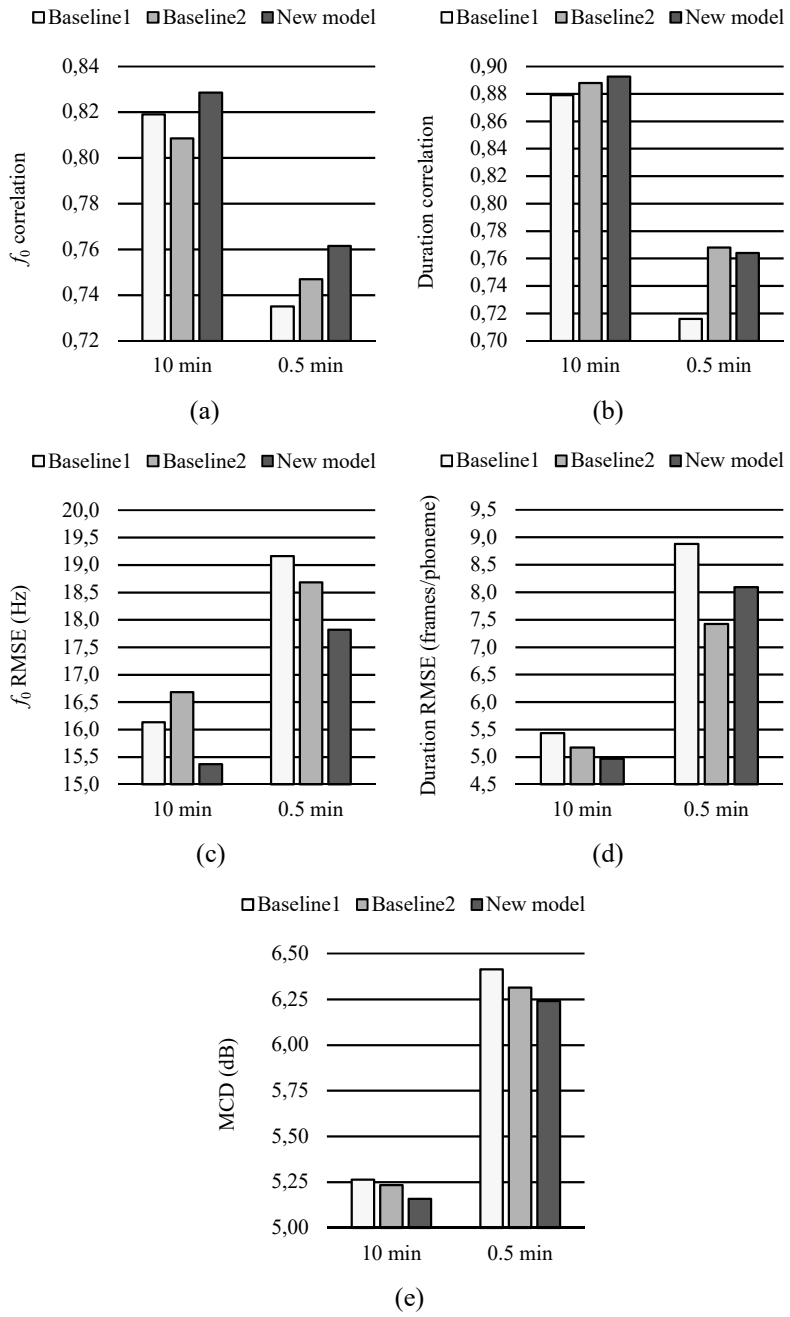

Figure 2: The results of the objective evaluation of the proposed model against the two baseline models: (a) correlation of f_0; (b) correlation of phone durations; (c) RMSE of f_0; (d) RMSE of phone durations; (e) mel cepstral distance (MCD).

baseline models, and the baseline model 1 seems to be least successful. The differences among the models are again more visible in case of the smaller adaptation dataset. As regards MCD, the differences among the models are practically negligible, but the proposed model consistently exhibits the best performance, and the baseline model 1 is the least successful one.

4.2 Subjective evaluation

A series of listening tests was carried out to corroborate the results of the objective evaluation and to establishing the influence of various factors to the quality of speech synthesized after the initial model is adapted to the target speech data.

4.2.1 Experiment 1

The aim of this experiment was to evaluate how successful the proposed model is in producing synthetic speech that is intended to resemble a particular speaker and speaking style in case only a small quantity of target speech data is available, and it also investigates the influence of the relationship between the position of the SSC points in each of the two embedding spaces and the degree to which the synthesized speech corresponds to the target speaker and style. Furthermore, the experiment also demonstrates the positive effect of the second phase of the adaptation process, which has been shown to increase the similarity of the synthesized speech to the intended speaker/style combination. The experiment investigates only the proposed model and does not include any comparison to the baseline models.

The experiment was set up as a MUSHRA listening test, and conducted among 26 listeners. Each listener was presented with 10 tasks, including 5 sentences in the voices of 2 speakers (M4 or F4). In each task, the listener was presented with the following 5 versions of the same utterance, in a randomized order:

- Hidden reference recording (original recording of the source speaker);
- Synthesis after just the first adaptation phase has been carried out on the initial model;
- Synthesis after the first adaptation phase has been carried out and then the obtained embedding was modified by 10%;
- Synthesis after the first adaptation phase has been carried out and then the obtained embedding was modified by 20%;
- Synthesis after both phases of the adaptation procedure have been carried out on the initial model without modifying the embedding obtained in the first phase.

In this experiment adaptation was performed using 10 minutes of target speech data. In cases the obtained embedding was modified, the modification was carried out for each of the 15 dimensions of the embedding independently, as follows. Firstly the reference range corresponding to a dimension was established as the sample standard deviation of its 67 values (one for each unique SSC) multiplied by 6, and then the existing coordinate was modified by ±10% or ±20% of the reference range. As it is common in MUSHRA tests, the reference recording was explicitly marked as such, but it was also hidden among the 5 utterances needing to be graded. The listeners

were asked to rate speaker similarity between the reference and each of the 5 utterances on a scale of 0 to 100. As the listeners tend to give lower grades to less appealing voices, which would obscure the influence of the factors that were considered as relevant for this experiment, the grade given to the hidden reference was scaled up to the maximum grade, and the remaining grades were scaled accordingly. Furthermore, in order to simplify the comparison of the results across all experiments, all grades are shown as rescaled to the interval 0-5.

The results, shown in Fig. 3, reveal that the first phase of the adaptation alone is sufficient for the model to be able to produce speech that resembles the voice of the target speaker to some extent. It has also been shown that the position of the embedding obtained through initial model training is relevant, since if it is modified, some of the resemblance to the target speaker is lost (a modification of each coordinate by 10% yields a relatively small degradation, but an increase to 20% of the original value reduced the mean score from 2.22 to 1.06). The experiment has also shown the importance of the second phase of adaptation, since the grade obtained after both adaptation phases had been carried out is significantly higher than any grade obtained after the first phase alone. A relatively wide margin still exists between the synthesized and the original speech, and it is quite likely that it is due to the relatively poor coverage of the embedding space by the SSCs existing in the training corpus. With more diverse data available for training the initial multi-speaker/multi-style model, we expect that the synthesis both before and after adaptation to a new speaker/style will suffer from less audible artefacts, and will be perceived as better by the listeners.

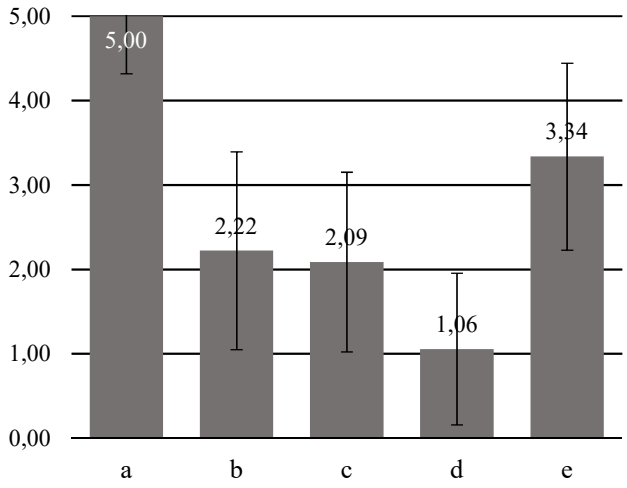

Figure 3: Subjective assessment of speaker similarity to the reference recording, rescaled to 5.00: (a) reference recording; (b) synthesis after the first phase of adaptation of the initial model; (c) synthesis after the first phase of adaptation and thus obtained embedding modified by 10%; (d) synthesis after the first phase of adaptation of the initial model and thus obtained embedding modified by 20%; (e) synthesis after both phases of adaptation.

4.2.2 Experiment 2

The aim of this experiment was to compare the quality of synthesis by the proposed model with respect to the two baseline models after adaptation, regardless of speaker similarity with respect to the reference utterance, through a MUSHRA listening test with 24 participants. In each of the 20 tasks, the listeners were informed that the reference utterance, marked as such, represents a recording of natural speech, and they were asked to grade, in terms of intelligibility and naturalness but without regard to speaker similarity, the following versions of the same utterance, which appeared in a randomized order:

- Hidden reference recording (original recording of the source speaker);
- Synthesis by the baseline model 1 after adaptation;
- Synthesis by the baseline model 2 after adaptation;
- Synthesis by the proposed model after the embedding obtained in the initial training is reset to 0 and only the second phase of adaptation is carried out;
- Synthesis by the proposed model after both phases of adaptation.

Out of the 20 tasks, in 10 of them models were adapted using 10 minutes of target speech data, and in the remaining 10 only 0.5 minutes of target speech data were used for adaptation. In each of the two cases there were 5 utterances by each of the 2 speakers (M4 and F4).

The results, shown in Fig. 4, reveal that regardless of the amount of target speech data used for adaptation, baseline model 2 was considered least successful by the listeners, while the highest grades were given to the two versions of the proposed model. It is interesting to note that, although the difference between average grades for baseline model 1 and the proposed model is not significant in case 10 minutes of adaptation data were used, the proposed model significantly outperforms the baseline model 1 in case adaptation is performed using only 0.5 minutes of target speech data. In other words, the proposed model appears to be much less sensitive to the scarcity of adaptation data compared to any of the baseline models. Another interesting point related to the proposed model is that, if the embedding obtained in the initial training is reset to 0 and only the second phase of adaptation is performed, this does not significantly degrade the intelligibility and naturalness of synthesis. However, they are still a little higher if the embedding obtained in the initial training and adapted to the new speaker is used.

4.2.3 Experiment 3

The setup of Experiment 3 was exactly the same as in case of Experiment 2 as regards the versions of synthesis that were offered to the listeners in each task, but this time the listeners were asked to evaluate speaker similarity instead of the general quality of the synthesis. The experiment consisted of 10 tasks (5 for each of the two speakers, M4 and F4), and 20 listeners participated in it. Since Experiment 2 has shown that the general quality of synthesis is quite different among the three models in case of adaptation on very little data, to prevent the listeners from being distracted by this difference so that they could focus on speaker similarity instead, adaptation was performed only on 10-minute target speaker datasets. As can be seen from Fig. 5, the

proposed model outperforms both baseline models in terms of producing synthesis in a voice that resembles the original speaker, even in case the embedding is reset to 0 and only the second phase of adaptation is carried out.

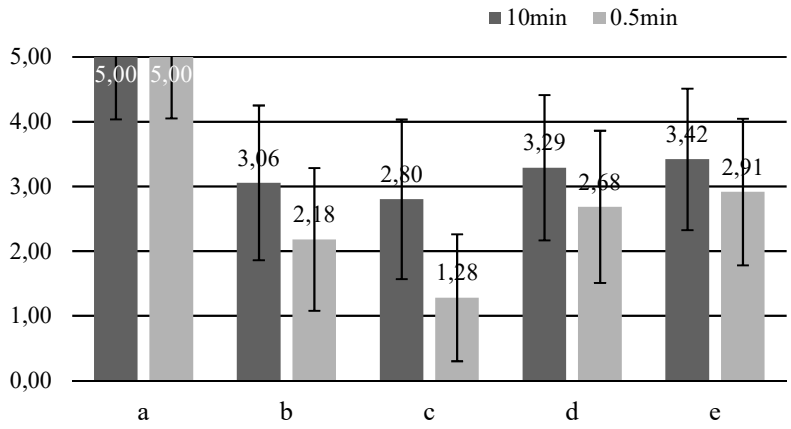

Figure 4: Comparison of the quality of synthesis obtained in different conditions, rescaled to 5.00: (a) reference recording; (b) synthesis by the baseline model 1 after adaptation; (c) synthesis by the baseline model 2 after adaptation; (d) Synthesis by the proposed model after the embedding is reset to 0 and only the second phase of adaptation is carried out; (e) synthesis by the proposed model after both phases of adaptation.

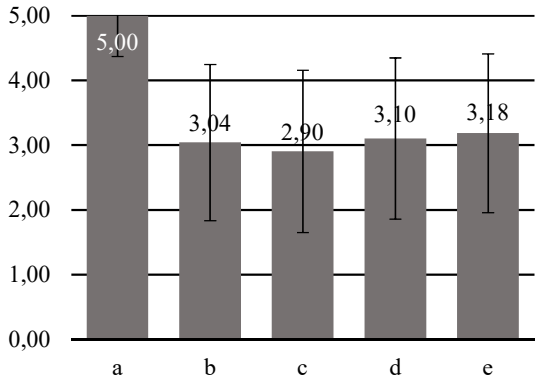

Figure 5: Comparison of the speaker similarity obtained in different conditions, rescaled to 5.00: (a) reference recording; (b) synthesis by the baseline model 1 after adaptation; (c) synthesis by the baseline model 2 after adaptation; (d) Synthesis by t proposed model after the embedding is reset to 0 and only the second phase of adaptation is carried out; (e) synthesis by the proposed model after both phases of adaptation.

4.2.4 Experiment 4

A more comprehensive evaluation of the performance of the proposed model would include its comparison with another speaker-dependent baseline model using not only small but large quantities of target speaker data for training. However, the authors were unable to carry out such an evaluation directly due to the availability of only a small quantity of speaker data for the speakers M4 and F4, having in mind that all the remaining available speakers were already used for the training of the initial model. This experiment represents an attempt of circumventing this limitation by including two types of MUSHRA tasks (10 tasks of each type). In both types of tasks the 32 participants in the listening test were informed that the reference utterance contains a natural recording of speech, and were required to evaluate the general quality (in terms of intelligibility and naturalness) of 3 utterances given in a randomized order. In the tasks of type 1 the following 3 utterances were offered:

- Hidden reference recording (original recording of M1 or F1);
- Synthesis by the baseline model 1 trained on all available data for M1 or F1 (see Table 1), without further adaptation;
- Synthesis by the proposed model using embeddings corresponding to M1 or F1, without further adaptation;

while the tasks of type 2 included the following 3 utterances:

- Hidden reference recording (original recording of M4 or F4);
- Synthesis by the proposed model after both phases of adaptation to M4 or F4, using 10 minutes of target speaker data;
- Synthesis by the proposed model after both phases of adaptation to M4 or F4, using 0.5 minutes of target speaker data.

In each task the 3 given utterances corresponded to the same speaker in order to eliminate the preference that a listener may have for the voice of one speaker over another. This is also the reason why the authors decided against merging the two types of tasks into one. All speakers were equally represented throughout the experiment, i.e. each of them appeared in 5 tasks.

The results of the experiment, with scores rescaled to the interval 0-5, are shown in Fig. 6. Before any general conclusions are drawn, it should be noted that although M1 and F1 did not appear in the same tasks as M4 and F4, it is still possible to compare the perceived quality of synthesis between models and/or versions that did not appear in the same tasks. Most notably, the synthesis by the baseline model 1 trained on all available data for M1 or F1 without further adaptation and synthesis by the proposed model after two-phase adaptation to M4 or F4, using 10 minutes of data (items (a) and (c) in Fig. 6) were perceived to be of similar quality. This shows that the proposed model, given a reasonable multi-speaker/multi-style model as a starting point, and using as little as 10 minutes of adaptation data, is able to achieve a quality of synthesis comparable to that of a standard speaker-dependent model trained on much more target speech data (~3.5 hours in case of M1 and ~2.5 hours in case of F1). Furthermore, synthesis obtained by the baseline model 1 trained on all available data for M1 or F1 (~3.5 and ~2.5 hours respectively) was considered to be of the same quality as the synthesis by the proposed model using embeddings corresponding to

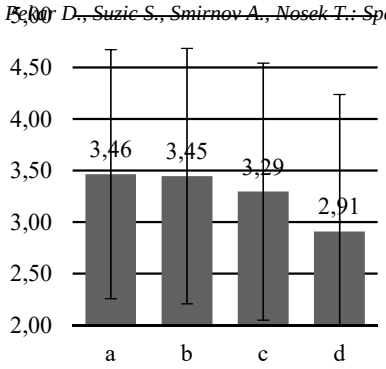

Figure 6: Comparison of the quality of synthesis obtained in different conditions, rescaled to 5.00: (a) Synthesis by the baseline model 1 trained on all available data for M1 or F1 without further adaptation; (b) Synthesis by the proposed model using embeddings corresponding to M1 or F1 without further adaptation; (c) Synthesis by the proposed model after both phases of adaptation to 10 minutes of speech data from M4 or F4; (d) Synthesis by the proposed model after both phases of adaptation to 0.5 minutes of speech data from M4 or F4.

M1 or F1 and no further adaptation. It thus appears to be more reasonable to use a given quantity of training data for a certain speaker as a basis for a multispeaker model based on embeddings than to train a single speaker-dependent model. Finally, it should be noted that the adaptation of the proposed model using 0.5 minutes of data yielded synthetic speech that was, as expected, rated as being of inferior quality than in case the adaptation was done on 10 minutes of data. However, the difference in scores is only 0.38, which can be considered quite small having in mind the difference in the quantity of adaptation data.

5 Conclusions and Future Work

In this research we propose a new deep neural network based speech synthesis model capable of adaptation to a particular speaker and speaking style, and we show that it outperforms two other recently proposed parametric speaker/style-dependent speech synthesis models, particularly in case the quantity of available adaptation data is extremely small. This is achieved owing to the joint representation of speaker, speaking style and cluster by their low-dimensional embedding, whereby the model is able to establish the similarities or differences among speakers and styles, and consequently generalize more accurately. The embedding approach opens up a range of interesting possible applications of the proposed model in any domain where the possibility of quick and efficient adaptation of speech synthesis to a new speaker and/or style is required.

A limitation of this research that cannot be disregarded is the relatively small quantity of speech data on which it was based. Namely, only 8 hours and 38 minutes of actual speech from 6 speakers was available for training, and a total of 20 minutes was available for adaptation, which is why we had to resort to data augmentation.

Although this is a valid technique aimed at overcoming data scarcity, the question remains to what extent a stronger multi-speaker/multi-style basis, including a greater number of speakers/styles, would improve the ability of the proposed system to produce synthetic speech of high intelligibility, naturalness and similarity to the target speaker/style. For that reason, the model will certainly be re-investigated as soon as a significantly greater amount of training data becomes available, and this will also be an opportunity to study the influence of data augmentation to the performance of the model. Another issue that will be further investigated is the influence of the difference in the quantities of available training data related to particular speakers/styles. This research in particular may have suffered from two speakers (M1 and F1) being over-represented in the training data. In the future versions of the proposed model we intend to equalize the influence of all speakers/styles on the training process by introducing weight coefficients corresponding to their relative contributions.

Acknowledgements

The study was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia (grants TR32035 and OI178027), as well as the Provincial Secretariat for Higher Education and Scientific Research of the Autonomous Province of Vojvodina (project CABUNS). Speech resources were provided by Speech Morphing Systems Inc., Campbell, CA, United States.

References

[Delić et al., 17] Delić, T., Sečujski, M., Suzić, S.: A Review of Serbian Parametric Speech Synthesis Based on Deep Neural Networks, TELFOR Journal, 2017, Vol. 9, No. 1, 32-37.

[Delić et al., 18] Delić, T., Suzić, S., Sečujski, M., Pekar, D.: Rapid development of new TTS voices by neural network adaptation, In Proc. 17th INFOTEH, Jahorina, Bosnia and Herzegovina, March 2018, 1-6.

[Eide et al., 04] Eide, E., Aaron, A., Bakis, R., Hamza, W., Picheny, M., Pitrelli, J.: A corpus based approach to <ahem/> expressive speech synthesis, In Proc. 5th ISCA Speech Synthesis Workshop, Pittsburgh, PA, United States, June 2004, 79-84.

[Fan et al., 15] Fan, Y., Qian, Y., Soong, F. K., He, L.: Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis, In Proc. ICASSP 2015, Brisbane, Australia, April 2015, 4475-4479.

[Gutkin et al., 10] Gutkin, A., Gonzalvo, X., Breuer, S., Taylor, P.: Quantized HMMs for low footprint text-to-speech synthesis, In Proc. Interspeech 2010, Chiba, Japan, September 2010, 837-840.

[Hojo et al., 16] Hojo, N., Ijima, Y., Mizuno, H.: An investigation of DNN-based speech synthesis using speaker codes, In Proc. Interspeech 2016, San Francisco, CA, United States, September 2016, 2278–2282.

[Hunt and Black, 96] Hunt, A., Black, A.: Unit selection in a concatenative speech synthesis system using a large speech database, In Proc. ICASSP 1996, Atlanta, Georgia, May 1996, 373–376.

[Inoue et al., 17] Inoue, K., Hara, S., Abe, M., Hojo, N., Ijima, Y.: An investigation to transplant emotional expressions in DNN-based TTS synthesis, In Proc. APSIPA Annual Summit and Conference, Kuala Lumpur, Malaysia, December 2017, 1253-1258.

[Kanagawa et al., 13] Kanagawa, H., Nose, T., Kobayashi, T.: Speaker-independent style conversion for HMM-based expressive speech synthesis, In Proc. ICASSP 2013, Vancouver, Canada, May 2013, 7864-7868.

[King et al., 08] King, S., Tokuda, K., Zen, H., Yamagishi, J.: Unsupervised adaptation for HMM-based speech synthesis, In Proc. Interspeech 2008, Brisbane, Australia, September 2008, 1869-1872.

[Ling et al., 15] Ling, Z.-H., Kang, S.-Y., Zen, H., Senior, A., Schuster, M., Qian, X.-J., Meng, H. M., Deng, L.: Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends, IEEE Signal Processing Magazine, May 2015, Vol. 32, No. 3, 35-52.

[Lorenzo-Trueba et al., 18] Lorenzo-Trueba, J., Henter, G.E., Takaki, S., Yamagishi, J., Morino, Y., Ochiai, Y.: Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis, Speech Communication, May 2018, Vol. 99, 135-143.

[Luong et al., 17] Luong H., Takaki S., Henter G., Yamagishi J.: Adapting and controlling DNN-based speech synthesis using input codes, In Proc. ICASSP 2017, New Orleans, LA, United States, March 2017, 4905-4909.

[Mao et al., 14] Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A. L.: Explain images with multimodal recurrent neural networks, October 2014, arXiv:1410.1090.

[Morioka et al., 04] Morioka, Y., Kataoka, S., Zen, H., Nankaku, Y., Tokuda, K., Kitamura, T.: Miniaturization of HMM-based speech synthesis, In Proc. Autumn Meeting of ASJ, 2004, 325-326, (in Japanese).

[Morise et al., 16] Morise, M., Yokomori, F., Ozawa, K.: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, IEICE Transactions on Information Systems, July 2016, Vol. 99, 1877-1884.

[Nakashika et al., 13] Nakashika, T., Takashima, R., Takiguchi, T., Ariki, Y.: Voice conversion in high order eigen space using deep belief nets, In Proc. Interspeech 2013, Lyon, France, August 2013, 369-372.

[Ohtani et al., 15] Ohtani, Y., Nasu, Y., Morita, M., Akamine, M.: Emotional transplant in statistical speech synthesis based on emotion additive model, In Proc. Interspeech 2015, Dresden, Germany, September 2015, 274-278.

[Ondáš et al., 13] Ondáš, S., Juhár, J., Pleva, M., Lojka, M., Kiktová, E., Sulír, M., Čižmár, A., Holcer, R.: Speech Technologies for Advanced Applications in Service Robotics, Acta Polytechnica Hungarica, May 2013, Vol. 10, No. 5, 45-61.

[Parker et al., 18] Parker, J., Stylianou, Y., Cipolla, R.: Adaptation of an expressive single speaker deep neural network speech synthesis system, In Proc. ICASSP 2018, Calgary, Canada, April 2018, 5309-5313.

[Pekar and Obradović, 01] Pekar, D., Obradović, R.: C++ Library for Digital Signal Processing - slib, In Proc. TELFOR, Belgrade, Serbia, November 2001, 7.7:1-4.

[Picard, 03] Picard, R. W.: What does it mean for a computer to “have” emotions?, Chapter in: Emotions in Humans and Artifacts, Trappl, R., Petta P., Payr, S. (Eds.), MIT Press, Cambridge, MA, 2003, 213-235.

[Sečujski et al., 18] Sečujski, M., Ostrogonac, S., Suzić, S., Pekar, D.: Learning prosodic stress from data in neural network based text-to-speech synthesis, SPIIRAS Proceedings Journal, Saint Petersburg, Russia, August 2018, Vol. 4, No. 59, 192-215.

[Sečujski et al., 19] Sečujski, M., Nosek, T., Suzić, S., Pekar, D.: Improvement of the quality of neural network based speech synthesis through intra-speaker clustering, In Proc. TAKTONS, Novi Sad, Serbia, November 2019, 9-10.

[Simard et al., 92] Simard, P., Victorri, B., LeCun, Y., Denker, J.: Tangent prop – a formalism for specifying selected invariances in an adaptive network, In Proc. 4th NIPS, Denver, Colorado, December 1991, 895-903.

[Suzić et al., 17] Suzić, S., Delić, T., Pekar, D., Ostojić, V.: Novel alignment method for DNN TTS training using HMM synthesis models, In Proc. SISY, Subotica, Serbia, September 2017, 271-276.

[Suzić et al., 18] Suzić, S., Delić, T., Jovanović, V., Sečujski, M., Pekar, D., Delić, V.: A comparison of multi-style DNN-based TTS approaches using small datasets, 13th International Scientific-Technical Conference on Electromechanics and Robotics "Zavalishin's Readings", ER(ZR), St. Petersburg, Russia, April 2018, 1-6.

[Suzić et al., 19] Suzić S., Delić T., Pekar D., Sečujski M.: Style Transplatation in Neural Network-based Speech Synthesis, Acta Polytechnica Hungarica, Jun 2019, Vol. 16, no. 6, 171-189.

[Tamura et al., 98] Tamura M., Masuko T., Tokuda K., Kobayashi T.: Speaker adaptation for HMM-based speech synthesis system using MLLR, In Proc. ESCA/COCOSDA Workshop on Speech Synthesis, Blue Mountains, Australia, November 1998, 273-276.

[Trueba et al., 13] Trueba L., Chicote R., Yamagishi J., Watts O., Montero J.: Towards speaking style transplantation in speech synthesis, In Proc. 8th ISCA Speech Synthesis Workshop, Barcelona, Spain, August 2013, 159-163.

[van den Oord et al., 16] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K. W.: WaveNet: A generative model for raw audio, Computing Research Repository, September 2016, arXiv:1609.03499v2.

[Vapnik and Chervonenkis, 71] Vapnik, V. N., Chervonenkis, A. Y.: On the uniform convergence of relative frequencies of events to their probabilities, Theory of Probability and its Applications, 1971, Vol. 16, No. 2, 264-280.

[Vishnubhotla et al., 10] Vishnubhotla, R., Fernandez, S., Ramabhadran, B.: An autoencoder neural network based low-dimensionality approach to excitation modelling for HMM-based text-to-speech, In Proc. ICASSP 2010, Dallas, TX, United States, March 2010, 4614-4617.

[Wang et al., 15] P. Wang, Y. Qian, F. K. Soong, L. He, H. Zhao: A Unified Tagging Solution: Bidirectional LSTM Recurrent Neural Network with Word Embedding, November 2015, arXiv:1511.00215v1.

[Wang et al., 17] Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Zh., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., Saurous, R. A.: Tacotron: towards end-to-end speech synthesis, April 2017, arXiv:1703.10135v2.

[Wu et al., 13] Wu Z.-Z., Chang, E.S., Li, H.-Z.: Conditional restricted Boltzmann machine for voice conversion, In Proc. ChinaSIP, Beijing, China, July 2013, 104-108.

[Wang et al., 15] Wang, P., Qian, Y., Soong, F. K., He, L., Zhao, H.: A unified tagging solution: bidirectional LSTM recurrent neural network with word embedding, November 2015, arXiv:1511.00215v1.

[Yamagishi et al., 04] Yamagishi, J., Tachibana, M., Masuko, T., Kobayashi, T.: Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis, In Proc. ICASSP 2004, Montreal, Canada, May 2004, Vol. 1, 5-8.

[Yamagishi et al., 09] Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata K., Isogai J.: Analysis of Speaker Adaptation Algorithms for HMM-based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm, IEEE Audio, Speech, and Language Processing, January 2009, Vol. 17, No. 1, 66-83.

[Yoshimura et al., 99] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T.: Simultaneous modeling of spectrum, pitch and duration in HMMbased speech synthesis, In Proc. EUROSPEECH 1999, Budapest, Hungary, September 1999, 2374-2350.

[Yu and Deng, 11] Yu, D., Deng, L.: Deep learning and its applications to signal and information processing, IEEE Signal Processing Magazine, February 2011, Vol. 28, No. 1, 145-154.

[Zen et al., 09] Zen, H., Tokuda, K., Black, A.: Statistical parametric speech synthesis, Speech Communication, November 2009, Vol. 51, No. 11, 1039-1064.

[Zen et al., 13] Zen, H., Senior, A., Schuster, M.: Statistical parametric speech synthesis using deep neural networks, In Proc. ICASSP 2013, Vancouver, Canada, May 2013, 7962-7966.