Style Transplantation in Neural Network-based Speech Synthesis

Siniša B. Suzić[1,2], Tijana V. Delić[1], Darko J. Pekar[2], Vlado D. Delić[1], Milan S. Sečujski[✉1,2]

[1]University of Novi Sad, Faculty of Technical Sciences,
Trg Dositeja Obradovića 6, Novi Sad, Serbia (e-mail: sinisa.suzic@uns.ac.rs, tijanadelic@uns.ac.rs, vlado.delic@uns.ac.rs, secujski@uns.ac.rs)

[2]AlfaNum – Speech Technologies, Bulevar Vojvode Stepe 40/7, Novi Sad, Serbia (e-mail: darko.pekar@alfanum.co.rs)

Abstract: The paper proposes a novel deep neural network (DNN) architecture aimed at improving the expressiveness of text-to-speech synthesis (TTS) by learning the properties of a particular speech style from a multi-speaker, multi-style speech corpus, and transplanting it into the speech of a new speaker, whose actual speech in the target style is missing from the training corpus. In most research on this topic speech styles are identified with corresponding emotional expressions, which was the approach accepted in this research as well, and the entire process is conventionally referred to as "emotion transplantation". The proposed architecture builds on the concept of shared hidden layer DNN architecture, which was originally used for multi-speaker modelling, principally by introducing the style code as an auxiliary input. In this way, the mapping between linguistic and acoustic features performed by the DNN was made style dependent. The results of both subjective or objective evaluation of the quality of synthesized speech as well as the quality of style reproduction show that in case the emotional speech data available for training is limited, the performance of the proposed system represents a small but clear improvement to the state of the art. The system used as a baseline reference is based on the standard approach which uses both speaker code and style code as auxiliary inputs.

Keywords: deep neural networks; human-computer interaction; affective computing; text-to-speech; emotion transplantation

1 Introduction and Related Work

Until recently, the speech technology research community focused its attention on elementary machine capabilities necessary to sustain speech interaction with a human, such as reasonably natural speech synthesis and sufficiently accurate automatic speech recognition. However, particularly in the last decade, this research focus has shifted to more sophisticated capabilities of cognitive systems,

including emotional capabilities represented by affective computing, which are often related to the global state of the cognitive system, rather than speech as just one of the modalities of human-machine interaction [1]. As a concept, affective computing was first introduced in [2], motivated by the realization that human perception, reasoning and decision making are intricately linked with emotion. As there is strong evidence that humans evaluate their interaction with machines along criteria analogous to those used in conventional social interaction with other humans [3], there is a need for a cognitive system to perceive, understand and emulate human emotions. The need for a system to have *emotional appearance*, i.e. that its behavior gives the appearance that it has emotions, was formally established in [4] as one of the four key factors of cognitive systems related to the emulation of human emotions. From the point of view of speech production, this means that a cognitive system should not only be able to produce speech that sounds natural, but that it should create an impression that it actually has emotions, that it empathizes with its collocutor and that it is able to establish social bonds such as trust. One of the greatest technological steps in the pursuit of a cognitive system that would be able to emulate a human collocutor to that extent is the advent of deep neural networks (DNN).

In the area of parametric text-to-speech speech synthesis (TTS), deep neural networks have been initially employed for acoustic modelling, owing to their ability to learn complex mappings between input linguistic representation of text and corresponding acoustic features of speech [5]. It has been shown that DNNs clearly outperform conventional hidden Markov models (HMM), which use decision tree-based state tying, in terms of naturalness and overall quality of synthesized speech. Modelling output probabilities of HMMs using restricted Boltzmann machines and deep belief networks has also shown good results [6]. The use of DNNs has soon been extended to other speech synthesis tasks, such as prosody modelling [7] as well as modelling of acoustic trajectories [8]. Deep neural networks have also been used for signal processing tasks, such as the extraction of low dimensional excitation parameters by auto-encoders [9] as well as DNN-based post-filtering, aimed at the recovery of fine spectral structure of speech which was lost during acoustic modelling [10]. Even the most widely cited deficiency of parametric TTS, which is a somewhat muffled character of synthesized speech due to the use of a vocoder, has been recently addressed by methods aimed at synthesis of raw speech waveforms [11]. Owing to their superior performance and the ability to solve a huge variety of different tasks given a sufficient amount of training data, DNNs now represent the state of the art in text-to-speech synthesis.

Meanwhile, the emergence of applications such as smart environments, virtual assistants and intelligent robots [12], has increased the demand for high-quality speech synthesis systems which would be able to use different voices, speak in different styles and convey different emotional states of the artificial speaker [13]. For instance, for a conversational robot intended to support medical therapy of children with developmental disorders it is desirable that it should be able to

address the patient in a variety of styles, fitting a specific situation [14]. A high degree of naturalness of human-computer interaction, exemplified in a wide range of available speech styles, is also beneficial in case the human collocutor is a person with non-standard cognitive characteristics [15]. All these requirements have shifted the focus of research towards developing speech synthesis methods oriented on obtaining more economical use of speech data. Namely, it would quickly become unfeasible to record and process a new speech corpus for each particular speaker/style combination, having in mind that the development of speech corpora is an extremely time-consuming process which requires a significant amount of human effort. A number of different approaches can be used instead, and most of them have initially been employed for speaker-dependent DNN-based TTS. A multitask learning framework based on a DNN with shared hidden layers and multiple speaker-dependent output layers has been proposed in [16], while a range of speaker-adaptation methods for DNNs has been investigated in [17]. The introduction of additional speaker-dependent inputs to the DNN was proposed in [18], and further extended by explicit handling of speaker gender and age [19]. The problem of developing a style-dependent expressive TTS has also been given a lot of attention in the research community. Early solutions, based on hidden Markov models, included HMM style modeling by either using a separate acoustic model for each style or using a single model which considers the style to be one of the linguistic features used [20]. Various approaches to style interpolation have also been proposed, including direct interpolation between models [21] or a single multiple-regression hidden semi-Markov model (MRHSMM) based on style vectors [22].

One of the most recent lines of research in the domain of multi-speaker, multi-style TTS is based on learning a transformation that maps the neutral speech style of a particular speaker into the desired speech style, even in cases when the target speaker/style combination is missing from the training corpus [23]. This approach is referred to as style transplantation or emotion transplantation, having in mind that styles are frequently identified according to emotional expressions that they carry. The term has since been extended to refer to any method aimed at obtaining synthetic speech with a particular speaker/style combination where the model was trained without access to any speech in that speaker/style combination. Most common HMM-based approaches to style transplantation include the use of constrained structural maximum a posteriori linear regression (CSMAPLR) [24] and emotion additive models [25]. The introduction of DNNs into this field has given rise to new approaches, including modifications of DNN architecture so as to exhibit structures that explicitly separate speaker and speech style contributions [26], as well as adaptation of an expressive single speaker DNN to a new speaker's voice [27].

The rest of the paper is organized as follows. Section 2 briefly describes the speech corpus used in the experiments. Section 3 proposes a novel deep neural network (DNN) architecture for style transplantation, which builds on the concept

of shared hidden layer DNN used in multi-speaker modelling [16]. Section 3 also briefly presents the auxiliary input model described in [26], which is used as a baseline reference in this research. Section 4 presents the experiment setup, while Section 5 discusses the results of subjective and objective evaluation of the quality of synthetic speech as well as the quality of style reproduction. The concluding section of the paper summarizes the main findings and outlines the plans for future research.

2 Speech Corpus

The speech corpus used in this research contains multi-style speech data in American English collected from two speakers. Both speakers are professional voice talents, one male and the other female. The training section of the corpus for each speaker contains 2 hours of neutral speech style (excluding silent segments) as well as 10 minutes of speech acted in three different styles – happy, apologetic and stern. The styles were described to the voice talents, as well as to the listeners who subsequently performed subjective evaluation, as follows:

- **happy** – the style of a call centre agent who delivers some very good news to the caller, such as: 'You have just won ten thousand dollars!';
- **apologetic** – the style of a call centre agent informing the caller that the caller's account has been blocked due to a company error;
- **stern** – the style of a technical support agent dealing with a difficult customer who keeps misunderstanding simple instructions, which is why the agent has to be strict and may even sound a little annoyed.

The corpus was originally designed with the aim of commercial application in a call centre environment, and was obtained for the purpose of this research in its original form (see Acknowledgement). The semantic content of all sentences in the style-dependent section of the corpus was mostly neutral with respect to any of the styles. The entire corpus was phonetically and prosodically annotated. The prosodic annotation followed the extended Tone and Break Indices (ToBI) set of conventions, described in [28].

Some general statistics of the speech corpus are shown in Table 1. It can be noticed that the male speaker generally spoke faster than the female one, and that both speakers spoke the fastest in the neutral style and the slowest in the stern style. Although the average fundamental frequencies are significantly different in the male and the female section of the corpus, both speakers had the highest average f_0 in the happy style and the lowest in the neutral style. It is also interesting to note that the standard deviation of f_0 is almost equal between speakers for the neutral and the stern style, while for the remaining two styles

there are significant differences. In the happy style, the standard deviation of f_0 in the male speaker is significantly higher than in the female speaker, while in the apologetic style the opposite is the case. This illustrates the well-known fact that emotional expressions can be extremely speaker-dependent [26], which makes the task of style transplantation even more difficult.

Table 1
General statistics of the speech corpus used in the experiments

	Male			Female		
	speech rate [phone/s]	average f_0 [Hz]	std f_0 [Hz]	speech rate [phone/s]	average f_0 [Hz]	std f_0 [Hz]
Neutral	12.7	98.7	34.1	11.5	188.3	34.1
Happy	11.4	170.2	71.4	11.0	239.7	53.3
Apologetic	10.8	101.9	25.1	9.7	215.7	38.4
Stern	9.5	131.0	50.4	9.5	216.3	50.6

3 Deep Neural Network Architecture

This research investigates the possibility of using a style-dependent shared hidden layer DNN architecture to generate speech in any speaker/style combination, even in those that may not be present in the training set. The proposed architecture, shown in Fig. 1, represents an upgrade of the shared hidden layer architecture introduced in [16], originally used for multi-speaker TTS. Similarly to [16], the proposed architecture includes a section containing hidden layers which are shared between all speakers and which implement a speaker-independent global linguistic feature transformation. Furthermore, the proposed architecture includes a separate output section for each speaker, which is expected to model his or her acoustic space. However, there are two significant differences with respect to [16]. Firstly, an additional input to the network is used to provide the style information, by analogy with the way speaker codes were used in [18]. This input represents a one-hot style code $\mathbf{s} = [s_1\ s_2\ \ldots\ s_K]^T$, which for the style m has a fixed 1-of-K form:

$$s_k = \begin{cases} 1, & k = m \\ 0, & k \neq m \end{cases}, \quad k = 1, 2, \ldots, K, \tag{1}$$

where K is the total number of styles, including the neutral style. The second difference with respect to [16] is the existence of an additional speaker-dependent hidden layer, which allows more sophisticated modelling of speaker acoustic spaces. The architecture obtained in this way is flexible and controllable, allowing synthesis in any speaker/style combination at runtime. It also explicitly separates speaker and style factors, i.e. it does not represent a black box. During the training

process, speaker and style dependent acoustic features that correspond to a given set of linguistic features are presented to the network through the speaker's output section. In this way, the shared layers are trained using the data from all speakers, while the output sections are trained using the data only from the speaker to which they correspond. In addition, if the number of neurons in one or more shared hidden layers is reduced, this effectively creates a bottleneck, which enforces a more compact representation of the speaker-independent global linguistic feature transformation implemented by shared hidden layers. Preliminary experiments have confirmed the assumption that bottlenecking can improve the performance of the model, as will be explained in more detail in the following section. The proposed architecture will be referred to as style-dependent shared hidden layer model (SDSM).

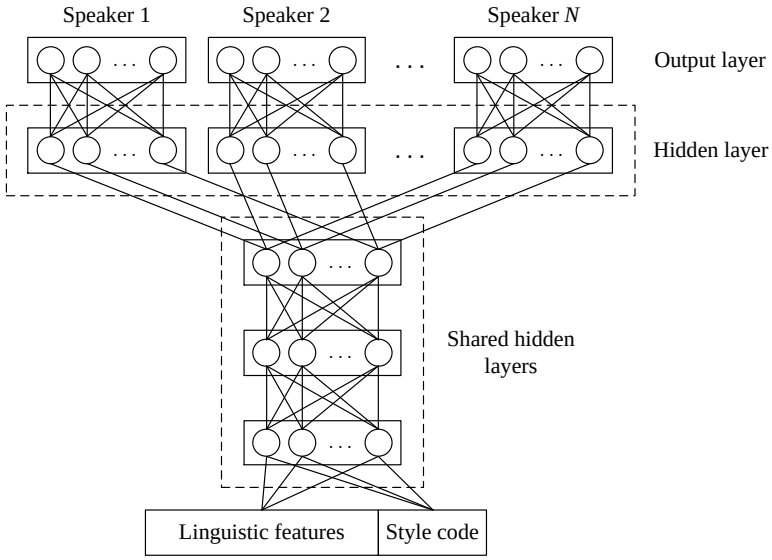

Figure 1
Style-dependent shared hidden layer model (SDSM)

In the subjective and objective evaluation, the proposed structure is compared to the reference, which is the auxiliary input model (AIM) of [20], shown in Fig. 2. This architecture is motivated by the work described in [18], and provides both speaker and style related information at the input. This structure does not explicitly separate the speaker and style factors, but distributes their contributions across the entire DNN. It should also be noted that [18] proposes another architecture, referred to as parallel model, whose performance was shown to be slightly above AIM. However, the choice of AIM as the reference for this research is justified by the fact that the advantages of the parallel model are lost when the quantity of training data is small, as is the case in this research [16, 29]. For the sake of comparison, the multi-speaker corpus used in [18] contains speech data

from 16 speakers, and there was approximately one hour of speech data for each speaker/style combination that existed in the corpus.

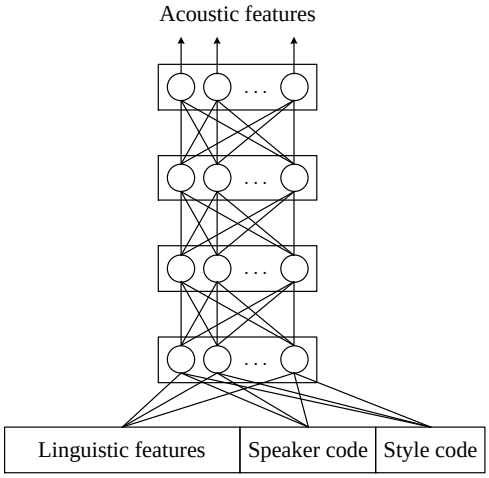

Figure 2
Auxiliary input model (AIM) of [20]

4 Experiment Setup

To produce samples of synthesized speech to be used for evaluation, SDSM and AIM were used for modelling acoustic features, while phonetic segment durations were taken from the original recordings. Acoustic features included 40 mel-generalized cepstral coefficients (MGC), band aperiodicity (BAP) and log f_0 values, first and second derivatives of these features, as well as an additional binary feature indicating whether the current frame was voiced or unvoiced (VUV). Consequently, the size of the acoustic feature vector was 127. The acoustic features were extracted from the speech recordings using the WORLD vocoder [30] for the purpose of training, and the same vocoder was used to convert the predicted acoustic features into speech at synthesis time. In each experiment, the inputs to both SDSM and AIM included 540 linguistic features, 9 features specifying within-phone positional information as well as the style code (one hot, 1-of-4). Furthermore, AIM included the speaker code (one hot, 1-of-2) as an additional input.

The depth of both SDSM and AIM was 4 hidden layers. In the case of SDSM, the three shared hidden layers contained regular neurons using the tangent hyperbolic activation function, while the hidden layer in each of the speaker-dependent output sections used LSTM neurons. In both cases the output layers used linear

activation and the cost function used was mean square error. The number of neurons in each hidden layer of both architectures was initially set to 1024. However, preliminary experiments on SDSM showed that the introduction of a bottleneck in one or more shared hidden layers can slightly improve the performance of the model. More specifically, among all the candidates which were investigated, the most promising SDSM architecture was found to be 1024-512-64-512, i.e. an architecture with a gradual bottleneck in the shared hidden layers and the size of the speaker-dependent LSTM hidden layer reduced as well. As for the AIM model, preliminary experiments showed that its performance did not improve with bottlenecking, which is why a simple model with 1024 neurons in all 4 hidden layers was used in that case. Each of the models was trained for 40 epochs with a gradually decreasing learning rate, starting from 0.01. Stochastic gradient descent with momentum and L2 regularization was used for optimization.

Besides the general evaluation of the quality of synthesized speech, both SDSM and AIM were evaluated for their ability to reproduce trained style as well as transplanted style. The term "trained style" refers to the production of synthetic speech by a model which had access to speech in the target speaker/style combination during training, while "transplanted style" refers to the case when speech in the target speaker/style combination was withheld during training. Speech samples with trained style were produced by models trained on all available speech data, i.e. 2 hours of neutral speech and 10 minutes of each of the 3 other speech styles (happy, apologetic and stern) for each of the two speakers. On the other hand, speech samples with transplanted style were obtained by models trained on all speech data excluding the 10 minutes of speech data in the target speaker/style combination.

5 Results and Discussion

In order to compare the proposed approach (SDSM) with the reference (AIM), they were both evaluated through objective measures as well as listening tests. In all cases the evaluation was carried out on utterances that did not appear in the training set in any of the speaker/style combinations[1].

[1] Examples of speech samples used for both objective and subjective evaluation are available at the URL: www.alfanum.ftn.uns.ac.rs/style_transplant.

5.1 Objective Evaluation

The samples containing either trained or transplanted style were evaluated by calculating the distance between a number of acoustic features in 20 synthesized utterances with a particular speaker/style combination and the same features in the original utterances. The acoustic features under consideration were: the root mean square error (RMSE) of f_0, correlation of f_0, mean square error of mel-generalized cepstral coefficients (MCD – mel cepstral distance), mean square error of band aperiodicities and the percentage of correctly predicted frame voicing. However, since all objective measures have shown similar behaviour, only the results related to MCD and f_0 are discussed in detail.

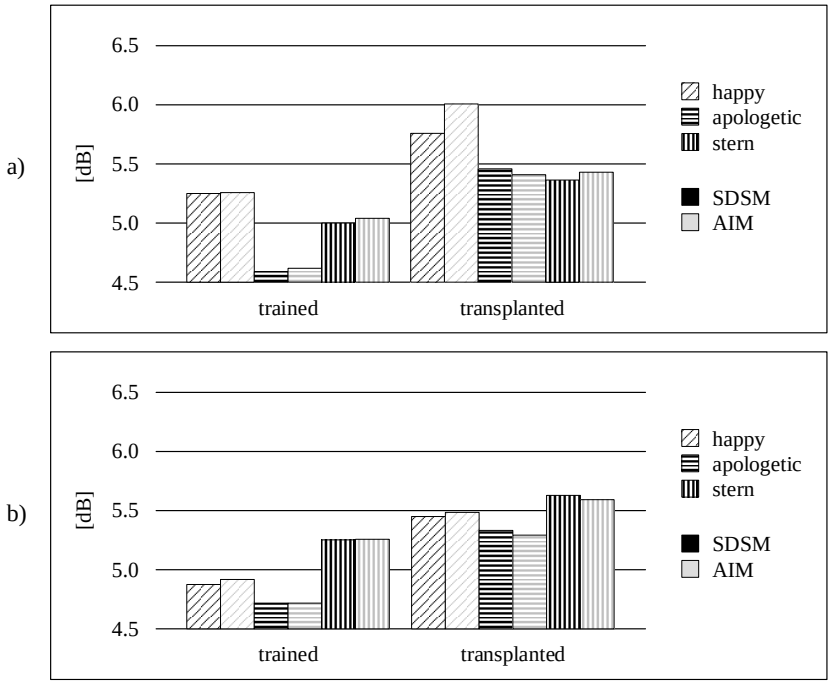

Figure 3
Comparison of trained and transplanted happy, apologetic and stern speech style: MCD for SDSM and AIM: (a) male speaker; (b) female speaker

Figure 3 shows the results related to MCD. It can be seen that transplanted styles do not perform as well as trained styles. This is not surprising, but the difference is less than 0.62 in the case of the male speaker and less than 0.88 in the case of the female speaker. For the male speaker, it can be seen that for all three styles both models, AIM and SDSM, behave almost identically in case of trained style, with differences below 0.04. Similar holds for transplanted apologetic and stern styles, with differences below 0.07, while for the happy style SDSM outperforms AIM

by 0.24. It should be noted that this is the style which shows the greatest difference in the variability of f_0 between the two speakers (Table 1). This indicates a difference in the level of emotional expressions of happiness between the speakers, which also suggests that this style may have been the hardest one to transplant. This conclusion is also supported by the fact that the average MCD for the transplanted happy style was the highest (5.61, as opposed to 5.40 for apologetic and 5.50 for stern). In the case of the female speaker, objective distance measures are generally smaller and differences between the two models are practically negligible (less than 0.05 in all cases). It can thus be concluded that, as regards objective evaluation with respect to MCD, SDSM slightly surpasses AIM, in that it shows better results in the case which is arguably the most challenging one, while in all other cases there is practically no difference between them. The case where SDSM clearly surpassed AIM is the transplantation of female happy style into a male happy style, which exhibited a significantly higher standard deviation of f_0 (Table 1). A higher level of emotional expression of happiness in the male speaker may also explain why better results were obtained for the transplantation of this style from the male to the female speaker than vice versa.

The results of the objective evaluation with respect to RMSE of f_0, shown in Fig. 4, exhibit similar behaviour. The most difficult case was again the transplantation of the happy style from the female to the male speaker, and again SDSM showed a clear advantage, surpassing AIM by 13.6 Hz. However, as regards RMSE of f_0 there was another case where SDSM surpassed AIM, which is the transplantation of the apologetic style from the male to the female speaker. This conclusion is also consistent with Table 1, which shows that the variability of f_0 in the apologetic style was significantly higher for the female speaker. Therefore, a general conclusion may be that the system can be expected to transplant a style from the speaker with lower variability of f_0 to the speaker with higher variability of f_0 (and, arguably, more intense emotional expression) less accurately than vice versa, at least in terms of objective parameters. In all cases except the two mentioned above, the performances of both models exhibited relatively small differences. Most notably, the AIM model was better at the transplantation of the apologetic style from the female to the male speaker (3.6 Hz) as well as at reproducing the trained happy style of the male speaker (3.5 Hz). In all remaining cases the results of the objective evaluation of both models with respect to RMSE of f_0 were almost identical (all differences were below 1 Hz). As was the case with the objective evaluation regarding MCD, transplanted styles again do not perform as well as trained styles. As expected, the difference is the greatest in the two cases which include the transplantation from a speaker with lower f_0 variability to the speaker with greater f_0 variability. In these two cases (transplantation of happy from female to male and transplantation of apologetic from male to female), SDSM clearly outperformed AIM, by 13.6 Hz and 19.0 Hz respectively, while in all other cases the difference was below 3.7 Hz. It should also be noted that the difference between the transplanted and trained styles is, for both models, the lowest for the stern style, which was found to exhibit almost identical standard deviation of f_0 in

both speakers. Similarity of the levels of emotional expression between two speakers of a particular style is clearly one of the key factors for the success of style transplantation.

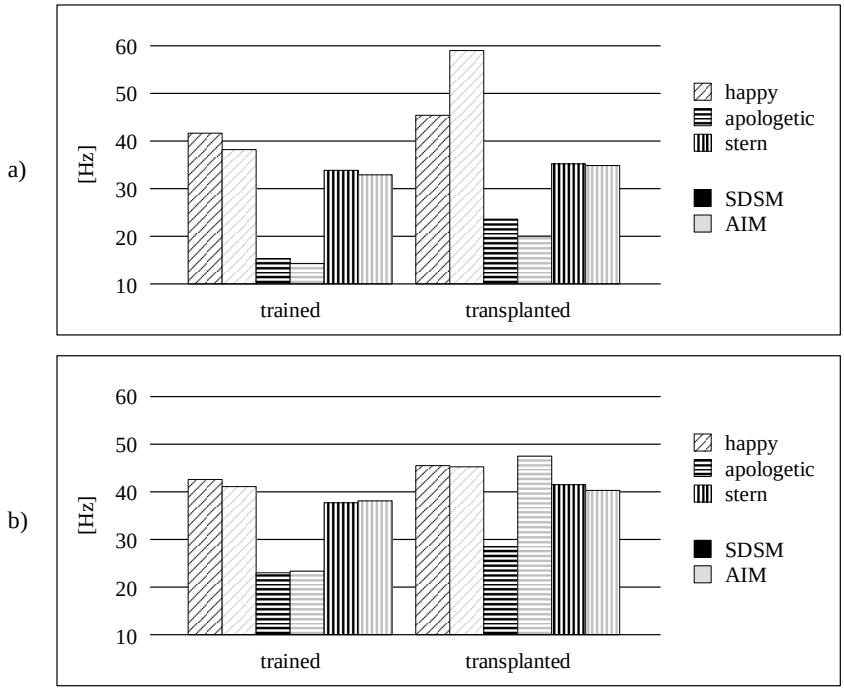

Figure 4
Comparison of trained and transplanted happy, apologetic and stern speech style: RMSE of f_0 for SDSM and AIM: (a) male speaker; (b) female speaker

5.2 Subjective Evaluation

Speech samples synthesized by both models were also evaluated through two independent listening tests. The first test required the listeners to classify samples of synthesized speech into one of the three styles, while the second one required the listeners to grade them according to the correspondence to the intended style as well as general quality of synthesis.

5.2.1 Style Classification

Besides speech samples corresponding to the trained and transplanted style generated by SDSM and AIM models, the first listening test also included speech samples obtained by resynthesis of speech from acoustic features extracted from original recordings (referred to as copy synthesis). Each of the 30 listeners was

presented with 60 utterances, 20 for each of the 3 styles, in random order. The 20 utterances corresponding to a particular style included 4 utterances obtained by each of the 5 approaches mentioned above. The listeners were required to identify each utterance as happy, apologetic or stern (no other options were given). The confusion matrices obtained by style classification for copy synthesis as well as trained and transplanted styles are given in Table 2, and the corresponding classification accuracies are shown in Fig. 5 as well.

Table 2

Confusion matrices for style classification of utterances obtained by copy synthesis as well as utterances synthesized by SDSM and AIM models

	[%]	Copy synthesis			Trained						Transplanted					
					SDSM			AIM			SDSM			AIM		
		H	A	S	H	A	S	H	A	S	H	A	S	H	A	S
Male	Happy	**80**	0	20	**83**	0	17	**87**	0	13	**80**	10	10	**50**	17	33
	Apologetic	8	**78**	13	0	**87**	13	7	**73**	20	10	**37**	53	23	**32**	45
	Stern	23	43	**33**	23	53	**23**	33	17	**50**	37	20	**43**	40	17	**43**
	Accuracy	63.9			64.4			70.0			53.3			41.7		
Female	Happy	**98**	2	0	**93**	0	7	**63**	3	33	**52**	10	38	**30**	5	65
	Apologetic	3	**83**	13	2	**92**	7	3	**93**	3	5	**60**	35	5	**62**	33
	Stern	12	0	**88**	17	2	**82**	20	5	**75**	35	17	**48**	35	7	**58**
	Accuracy	90.0			88.9			77.2			53.3			50.0		
Average	Happy	**89**	1	10	**88**	0	12	**75**	2	23	**66**	10	24	**40**	11	49
	Apologetic	6	**81**	13	1	**89**	10	5	**83**	12	8	**48**	44	14	**47**	39
	Stern	18	22	**61**	20	28	**53**	27	11	**63**	36	18	**46**	38	12	**51**
	Accuracy	76.9			76.7			73.6			53.3			45.8		

It should firstly be noted that, while the accuracy of classification of copy synthesis in case of the female speaker is satisfactory (90.0%), in case of the male speaker it is lower (63.9%), which is mostly due to frequent misclassification of the stern style. As for the evaluation of synthesized utterances containing trained or transplanted style, it can be seen that utterances obtained by SDSM are more accurately classified than those obtained by AIM (76.7% vs. 73.6% on trained style and 53.3% vs. 45.8% on transplanted style), although the accuracy rates vary across different methods and styles. It can also be seen that without exception, the accuracy of classifying transplanted styles is lower than in case of trained styles, which is in line with findings reported in [26]. Style classification of synthesized samples of the female voice is generally more accurate, which is in agreement with the results of the classification of copy synthesis. It can also be noted that, for both speakers, trained style samples synthesized by SDSM are classified almost as accurately as copy synthesis, while the same does not hold for AIM. On the other hand, although all these results are in agreement with the results of objective evaluation in that they show a slight advantage of SDSM over AIM in style transplantation, there are differences when individual styles are considered.

For example, while both objective and subjective evaluation suggest that SDSM is better at transplanting the happy style from female to male, there is less agreement in the other case where objective evaluation gives preference to SDSM, which is the transplantation of the apologetic style from male to female. Namely, although in that case the objective distance is lower in case of SDSM, style classification accuracy is approximately the same in both models.

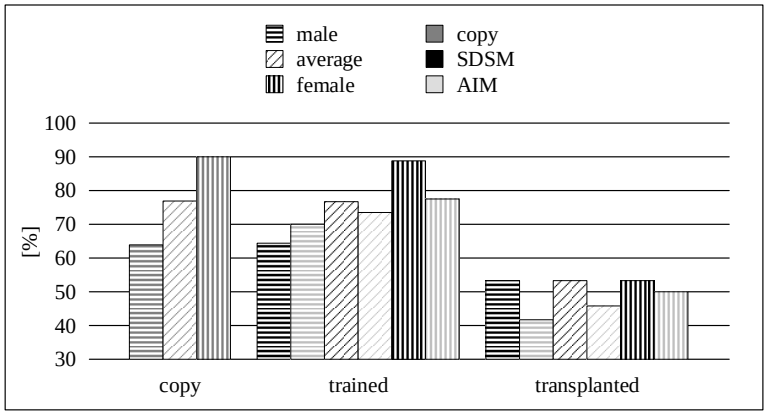

Figure 5
Accuracy of style classification of utterances synthesized by SDSM and AIM, with copy synthesis as reference

5.2.2 Evaluation of Style Reproduction and Quality of Synthesis

To additionally investigate the correspondence of different synthesis versions to the intended style, as well as to evaluate their general quality, a MUSHRA-style evaluation was carried out [31]. The synthesis versions presented to 18 listeners included speech samples corresponding to the trained and transplanted style generated by SDSM and AIM models, samples obtained by copy synthesis, as well as samples synthesized by AIM in the neutral style, which were introduced in order to facilitate the evaluation of style reproduction.

The listening test included two cycles. In the first cycle the listeners were required to evaluate different versions of given 18 utterances for their correspondence to the intended style. For each of the 18 utterances the listeners were presented with a MUSHRA screen containing the copy-synthesis reference (labelled as such), as well as 6 different versions of the utterance (including a hidden reference) in random order, and they were required to grade them on a scale from 0 to 100. In the first cycle the listeners were explicitly instructed to disregard the semantic content of the utterance as well as issues related to general synthesis quality (presence of artifacts or buzziness, unnaturalness of intonation), and only to judge how successfully the intended speech style was reproduced. In the second cycle the same testing framework was used again (6 versions of 18 utterances), but this time the listeners were required to evaluate the speech samples for their general

quality. They were explicitly instructed to disregard the intended style, and only to judge how successfully the synthesis imitates the speech of the original speaker.

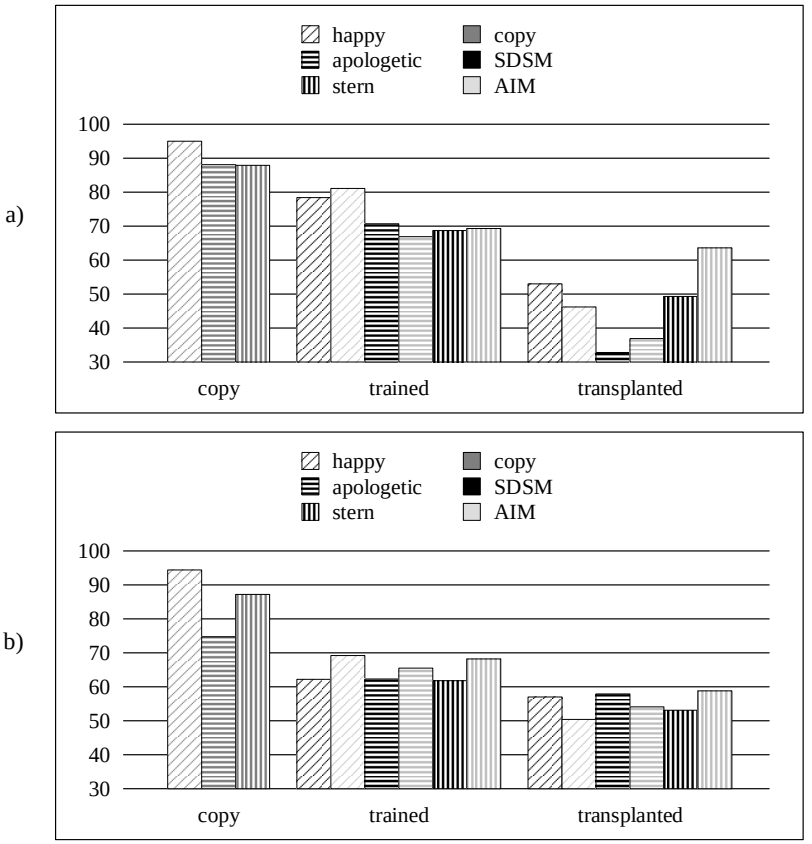

Figure 6
Evaluation of style reproduction in speech synthesized by SDSM and AIM, with copy synthesis as reference: (a) male speaker; (b) female speaker

The results of the evaluation of style reproduction are shown in Fig. 6. Although there is a certain variability of scores with respect to style and speaker, some general conclusions can be drawn. Firstly, it can be seen that, as expected, copy synthesis has consistently obtained the highest scores (mean 87.8, male speaker 90.3 and female speaker 85.4). This result also indicates that speech synthesized in the voice of the male speaker, although more difficult for style classification, nevertheless conforms well to the expectations of listeners when they are aware what the target style is. The grades obtained for trained styles are generally higher than for transplanted styles for both models (SDSM: 67.3 vs. 50.5; AIM: 70.0 vs. 51.7). Although the performance of SDSM was rated as slightly inferior to AIM in case of trained styles, the average perceived difference was practically negligible in case of transplanted styles.

The results of the evaluation of the general quality of synthesis are shown in Fig. 7. Once again, copy synthesis has consistently obtained the highest scores (mean 86.9, male speaker 88.6 and female speaker 85.2). The grades obtained for trained styles are, again, higher than for transplanted styles for both models (SDSM: 66.1 vs. 58.5; AIM: 71.2 vs. 54.8). It can be seen that, although AIM outperforms SDSM in case of trained styles, SDSM produces speech of better quality in case of transplanted styles. Thus, a general conclusion of the MUSHRA evaluation may be that, although AIM outperforms SDSM in both style reproduction and general quality in case of trained styles, this advantage is lost in the style transplantation scenario, in which SDSM produces synthesized speech which corresponds to the intended style almost equally well (−1.2), but has a slight advantage in terms of general quality (+3.7).

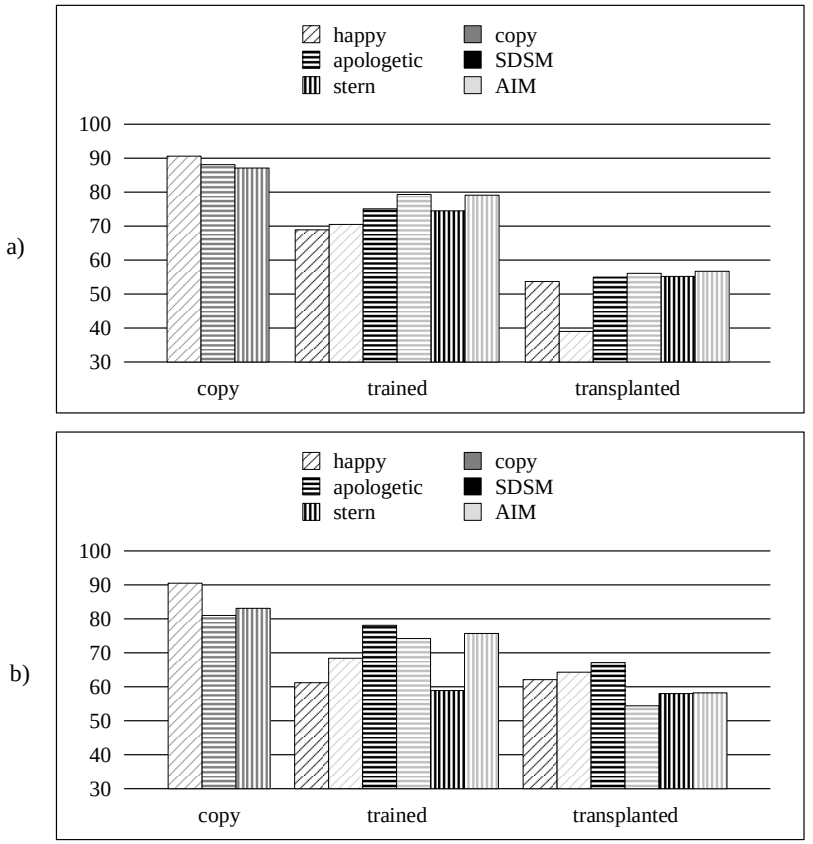

Figure 7
Evaluation of general quality of speech synthesized by SDSM and AIM, with copy synthesis as reference: (a) male speaker; (b) female speaker

Conclusions

In terms of cognitive infocommunications (CogInfoCom), a speech synthesizer capable of delivering speech in a wide range of styles is, at this stage of technical development of speech technology, a vital component of any cognitive technical agent intended to support inter-cognitive, representation-bridging communication with a human collocutor. As it is more comfortable to perceive an artificial agent as a real person than to think about the complexity and all the implications of a communicating machine [32], this innate human tendency to behave naturally in the interaction with computers should only be encouraged by improving the expressiveness of the synthesized voice. According to the definition of cognitive infocommunication as stated in [33], this is an example of merging and extension of cognitive capabilities of both communicating parties, resulting in an engineering application in which an artificial and a natural cognitive system are enabled to work together more effectively.

In this research, the support for text-to-speech synthesis in different speech styles is enabled by using a novel deep neural network architecture for style transplantation (SDSM), which was compared to the state of the art (AIM) through objective and subjective evaluation. Although the results of the comparison are not conclusive in case both models have been trained on speech data containing speech in the desired speaker/style combination, SDSM shows a small, but clear advantage over AIM in case the speech style is transplanted from another speaker.

Namely, utterances synthesized by SDSM exhibit slightly lower average values of MCD and RMSE of f_0 in the objective tests, the intended speech style appears, on average, more recognizable in them and, although no significant difference between SDSM and AIM has been found in subjective style reproduction evaluation, SDSM was found to produce speech of slightly higher general quality. The proposed architecture has thus shown to be a promising choice for speech style transplantation based on limited target speaker data, and as such it has a wide range of practical application. It should be noted that in this research not only the quantity of style-dependent data was relatively small, but the multi-speaker model, which serves as the basis for style transplantation, was based on the voices of just two speakers, one quite different from the other. It can be expected that the ability of SDSM to synthesize high-quality speech in a transplanted style will only increase with the availability of training data.

In the future we intend to investigate the influence of a number of factors, most notably the number of speakers in the multi-speaker model and the quantity of style-dependent data, on the reproduction of transplanted speech styles. A number of different modifications of SDSM, such as the introduction of multiple speaker-dependent layers or the use of different types of units per layer will be investigated as well. The extension of the proposed approach to phonetic segment duration modelling will also be one of the directions of our future research.

Acknowledgement

The presented study was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia (grants TR32035, OI178027 and E!9944), 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

[1] Baranyi, P., Csapo, A., Sallai, Gy.: Cognitive Infocommunications (CogInfoCom), Springer International, Ch. 1 (2015)

[2] Picard, R. W.: Affective computing. The MIT Press, Cambridge, MA (1995)

[3] Nass, C. I., Yen, C.: The man who lied to his laptop: what machines teach us about human relationships. Current Trade Penguin Group, New York, NY (2010)

[4] Picard, R. W.: What does it mean for a computer to "have" emotions. In: Trappl,. R, Petta, P,. Payr, S. (eds) Emotions in humans and artifacts. MIT Press, Cambridge, MA, pp. 213-235 (2003)

[5] Zen, H., Senior, A., Schuster, M.: Statistical parametric speech synthesis using deep neural networks. Proc. ICASSP 2013, Vancouver, Canada, pp. 7962-7966 (2013)

[6] Zen, H., Senior, A.: Deep mixture density networks for acoustic modelling in statistical parametric speech synthesis. Proc. ICASSP 2014, Florence, Italy, pp. 3844-3848 (2014)

[7] Fernandez, R., Rendel, A., Ramabhadran, B., Hoory, R.: Prosody contour prediction with long short-term memory, bidirectional, deep recurrent neural networks. Proc. Interspeech 2014, Singapore, pp. 2268-2272 (2014)

[8] Fan, Y., Qian, Y., Xie, F., Soong, F. K.: TTS synthesis with bidirectional LSTM based recurrent neural networks. Proc. Interspeech 2014, Singapore, pp. 1964-1968 (2014)

[9] Vishnubhotla, R., Fernandez, S., Ramabhadran, B. An autoencoder neural-network based low-dimensionality approach to excitation modelling for HMM-based text-to-speech. Proc. ICASSP 2010, Dallas, TX, United States, pp. 4614-4617 (2010)

[10] Chen, L.-H., Raitio, T., Valentini-Botinhao, C., Yamagishi, J., Ling, Z.-H.: DNN-based stochastic postfilter for HMM-based speech synthesis. Proc. Interspeech 2014, Singapore, pp. 1954-1958 (2014)

[11] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., Kavukcuoglu, K.: WaveNet:

A generative model for raw audio. Computing Research Repository, arXiv:1609.03499v2 (2016)

[12] 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, Vol. 10, No. 5, pp. 45-61 (2013)

[13] Eide, E., Aaron, A., Bakis, R., Hamza, W., Picheny, M., Pitrelli, J.: A corpus based approach to <ahem/> expressive speech synthesis. Proc. 5th ISCA Speech Synthesis Workshop, Pittsburgh, PA, United States (2005)

[14] Tasevski, J., Gnjatović, M., Borovac, B.: Assessing the children's receptivity to the robot MARKO, Acta Polytechnica Hungarica, Vol. 15, No. 5, pp. 47-66, 2018

[15] Izsó, L.: The Significance of Cognitive Infocommunications in Developing Assistive Technologies for People with Non-Standard Cognitive Characteristics: CogInfoCom for People with Nonstandard Cognitive Characteristics, 6th IEEE Int. Conf. on Cognitive Infocommunications, Győr, Hungary, pp. 77-82, 2015

[16] Fan, Y., Qian, Y., Soong, F. K., He, L.: Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis. Proc. ICASSP 2015, Brisbane, Australia, pp. 4475-4479 (2015)

[17] Wu, Z., Swietojanski, P., Veaux, C., Renals, S., King, S.: A study of speaker adaptation for DNN-based speech synthesis. Proc. Interspeech 2015, Dresden, Germany, pp. 879-883 (2015)

[18] Hojo, N., Ijima, Y., Mizuno, H.: An investigation of DNN-based speech synthesis using speaker codes. Proc. Interspeech 2016, San Francisco, CA, United States, pp. 2278-2282 (2016)

[19] Luong, H., Takaki, S., Henter, G., Yamagishi, J.: Adapting and controlling DNN-based speech synthesis using input codes. Proc. ICASSP 2017, New Orleans, LA, United States, pp. 4905-4909 (2017)

[20] Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T.: Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis. IEICE Transactions on Information and Systems, Vol. E88-D, No. 3, pp. 502-509 (2005)

[21] Tachibana, M., Yamagishi, J., Masuko, T., and Kobayashi, T.: Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. IEICE Transactions on Information and Systems, Vol. E88-D, No. 11, pp. 2484-2491 (2005)

[22] Nose, T., Yamagishi, J., Masuko, T., Kobayashi, T.: A style control technique for HMM-based expressive speech synthesis. IEICE Transactions on Information and Systems, Vol. E90-D, No. 9, pp. 1406-1413 (2007)

[23] Kanagawa, H. Nose, T., Kobayashi, T.: Speaker-independent style conversion for HMM-based expressive speech synthesis. Proc. ICASSP 2013, Vancouver, Canada, pp. 7864-7868 (2013)

[24] Trueba, L., Chicote, R., Yamagishi, J., Watts, O., Montero, J.: Towards speaking style transplantation in speech synthesis. Proc. 8th ISCA Speech Synthesis Workshop, Barcelona, Spain, pp. 159-163 (2013)

[25] 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, pp. 274-278 (2015)

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

[27] Parker, J., Stylianou, Y., Cipolla, R.: Adaptation of an expressive single speaker deep neural network speech synthesis system. Proc. ICASSP 2018, Calgary, Canada (2018)

[28] 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 (accepted for publication) (2018)

[29] Suzić, S., Delić, T., Jovanović, V., Sečujski, M., Pekar, D., Delić, V.: A comparison of multi-style DNN-based TTS approaches using small datasets. Proc. 13th International Conference on Electromechanics and Robotics "Zavalishin's Readings", Saint Petersburg, Russia. DOI: 10.1051/matecconf/201816103005 (2018)

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

[31] Method for the subjective assessment of intermediate quality levels of coding systems. ITU-T Recommendation BS.1534 (2015)

[32] Hamacher, A., Bianchi-Berthouze, N., Pipe, A. G., Eder, K.: Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical human-robot interaction, In 25th IEEE Int. Symposium on Robot and Human Interactive Communication (RO-MAN), (2016)

[33] Baranyi, P., Csapo, A.: Definition and Synergies of Cognitive Info-communications, Acta Polytechnica Hungarica, Vol. 9, No. 1, pp. 67-83, (2012)