DOI 10.15622/sp.60.8

S.B. SUZIĆ, T.V. DELIĆ, S.J. OSTROGONAC, S.V. ĐURIĆ, D.J. PEKAR

STYLE-CODE METHOD FOR MULTI-STYLE PARAMETRIC TEXT-TO-SPEECH SYNTESIS

Suzić S.B., Delić T.V., Ostrogonac S.J., Đurić S.V., Pekar D.J. **Style-Code Method for Multi-Style Parametric Text-to-Speech Synthesis.**

Abstract. Modern text-to-speech systems generally achieve good intelligibility. The one of the main drawbacks of these systems is the lack of expressiveness in comparison to natural human speech. It is very unpleasant when automated system conveys positive and negative message in completely the same way. The introduction of parametric methods in speech synthesis gave possibility to easily change speaker characteristics and speaking styles. In this paper a simple method for incorporating styles into synthesized speech by using style codes is presented.

The proposed method requires just a couple of minutes of target style and moderate amount of neutral speech. It is successfully applied to both hidden Markov models and deep neural networks-based synthesis, giving style code as additional input to the model. Listening tests confirmed that better style expressiveness is achieved by deep neural networks synthesis compared to hidden Markov model synthesis. It is also proved that quality of speech synthesized by deep neural networks in a certain style is comparable with the speech synthesized in neutral style, although the neutral-speech-database is about 10 times bigger. DNN based TTS with style codes are further investigated by comparing the quality of speech produced by single-style modeling and multi-style modeling systems. Objective and subjective measures confirmed that there is no significant difference between these two approaches.

Keywords: text-to-speech synthesis, expressive speech synthesis, deep neural networks, speech style, style code, one-hot vector.

1. Introduction. Text-to-speech (TTS) synthesis, a set of techniques that enable computers to convert text to human voice, has been a popular research area during the last few decades. This technology has a wide range of possible usage scenarios. Initially, it was used as a reading aid for blind people. It was also successfully applied in call centers for reading different types of information to the customers. Nowadays, audiobooks are generated by TTS systems and personal assistant applications use this technology to deliver information to its users. There are few different approaches to converting text to speech. Concatenative synthesis [1], a method based on concatenation of authentic speech segments from some prerecorded database, produces high-quality speech, the naturalness of which is still considered to be state-of-the-art. However, in some cases, audible glitches appear, usually in contexts that are not covered by the speech database. Furthermore, creating a new synthetic voice can only be done by obtaining a whole new speech database and spending a significant amount of resources on

the database preparation. A first technique in which some of these drawbacks where overcome is parametric synthesis based on hidden Markov models (HMM) [2]. It produces speech of constant quality, and even smaller speech databases can be used for getting a voice of decent quality. However, because of some drawbacks in modeling approach the speech synthesized by HMM system sounds muffled [3].

In recent years, prevalent research methods for text-to-speech synthesis have been based on deep neural networks (DNN). Reason for this prevalence is considered to be related to immanent characteristics of DNNs that are so-called *deep structures*, in contrast with the already mentioned HMMs that are so-called *shallow structures* [4]. As deep structures are proven to be more appropriate for modeling complex relations between input and output data [5], it was expected that DNNs would be suitable for modeling relations between linguistic features and acoustic parameters. Different papers examined this approach [6-7] and concluded that DNNs are appropriate for usage in TTS, since they provide synthesized speech of high quality. Furthermore, it has been proven that DNNs are better than HMMs in this context, since synthesized voice has even higher quality, comparable to that of concatenative synthesis [4].

The two most important requirements that synthesized speech should fulfill are intelligibility and naturalness. The research community mostly agrees that modern TTS systems achieve good performance regarding these two requirements [8]. The major critique of TTS systems is the uniformity of synthesized speech. Namely, most of the speech is generated in same speaking style and yet modern applications require not only high-quality naturally sounding speech, but also the possibility of changing speaking style, thus allowing users to exchange subtext information. For example, the style in which some news is generated should be different from the style in which commercials or warnings are synthesized, as explained in [9]. In [10] it is stated that some aircraft accidents investigators think that neutral voice, in which warnings in critical situation are generated, are the reason why the passengers do not perceive these situations as potentially dangerous.

Although different speaking styles can be associated with some emotional states, the term speaking style is more general. Emotional state is usually associated with speaker’s inner state which affects speech characteristics. The term speaking style in this paper is used to mark any deviation in speech characteristics compared to neutral speech and does not consider the cause or intention for this change.

This paper presents an expansion of [11], the goal of which was to make a multi-style model using style codes in order to enable speech synthesis in different speaking styles, both for HMM and DNN synthesis. Although it is already proven that DNN synthesis achieves better results than HMM [12-14], it is yet to be tested if proposed approach can be used in both synthesis technics and if there is difference in their performance. Since it is assumed that only a small amount of speech material with new speaking style is available, the aim is to test if the speech synthesized in target style has similar quality as the speech synthesized in neutral style, for which much more training material is used. The main contribution of this paper is extensive analysis of the performance of the style code approach in DNN synthesis framework.

The rest of the paper is organized as follows. In section II, parametric approaches to speech synthesis are explained. The review of methods used in style modeling is given in section III, which is followed by an explanation of technic used for creation of multi-style DNN model in section IV. In the section V, the results are presented and discussed. Finally, conclusions are drawn and directions of further research are mentioned.

2. Parametric text-to-speech synthesis. Parametric speech synthesis consists of two phases. First, in training phase, acoustic features are extracted with a vocoder and models are trained on the extracted features. In synthesis phase, models are used to generate acoustic parameters which are converted to the speech samples by a vocoder.

HMM based TTS.

In order to overcome some of the shortcomings of concatenative synthesis (complicated implementation of a new voice, occasional glitches in speech, memory space requirements), statistical parametric methods emerged as the best solution. [15]. These systems model the parameters extracted from the speech by using some generative modeling approaches. The most widely used systems are based on hidden Markov models (HMM). Actually, the terms HMM synthesis and statistical speech synthesis are being used as synonyms in literature. However, one of the parameters that need to be modeled is fundamental frequency, which is not defined in unvoiced regions and therefore usage of standard HMM modeling is not fully applicable. Bearing that in mind, an extended HMM model called multi-space distribution hidden Markov model (MSD-HMM) is proposed and successfully applied in speech synthesis [16].

In the field of automatic speech recognition (ASR) based on HMM, the speech units that are modeled are the triphones, phonemes with known preceding and succeeding phoneme identities. For the purposes of speech synthesis, modeling unit takes a much wider context. Besides the preceding and succeeding phoneme identity, the context factors that are taken into account are different phonetic, linguistic and prosodic information. Since there is not enough training data to adequately model all the contexts, tree-based context clustering is used, meaning that similar contexts share distribution parameters [17]. This, naturally, introduces some unwanted smoothing of the parameters and influences naturalness of synthesized speech. Various techniques have been developed in order to address this problem [18].

In the synthesis stage, the most probable parameter sequence should be generated based on input text and known models' parameters, $\boldsymbol{\lambda}$. If the HMM state sequence, $\boldsymbol{q}$, is known, the solution to this problem can be found by solving the following optimization equation:

$$\hat{\boldsymbol{O}} = \arg\max_{\boldsymbol{o}} P(\boldsymbol{O}|\boldsymbol{q},\boldsymbol{\lambda}), \quad (1)$$

The solution to the optimization problem from Equation 1, as well as some other algorithms for generating speech parameters, can be found in [19]. It is proven that inclusion of dynamic features (first and second derivatives) improve overall quality of synthesized speech.

DNN based TTS.

In order to improve modeling of a layered, hierarchical structure of a human speech production system, deep neural networks (DNNs) are becoming dominant in speech synthesis. DNNs manage to achieve better generalization and thus synthesized speech is less smoothed compared to the one synthesized using HMM based approach.

When neural networks are applied to speech synthesis linguistic features are used as inputs. Extraction of linguistic features is carried out on the phoneme level. Therefore, linguistic features most often contain information on phoneme identity, as well as the identity of phoneme's contexts, phoneme's accent, etc. Those features are extracted by a separate front-end module and delivered into two neural networks (Figure 1). The first one is trained to predict phonemes' states durations. In order to get durations on a state level, as targets during the training, initial alignment has to be performed. This procedure is usually done by using monophone models and a couple of rounds of Baum-Welch

algorithm [20]. Inputs and targets of the duration network are phoneme-aligned. The second network is trained independently from the first one, and its task is to predict the acoustic features. Beside the features used as inputs of the duration network, it requires additional inputs specifying within-phone positional information (including state-level durations). Target features for this network are extracted from the training dataset by the vocoder. Inputs and targets are frame-aligned. Networks are trained using back-propagation algorithm.

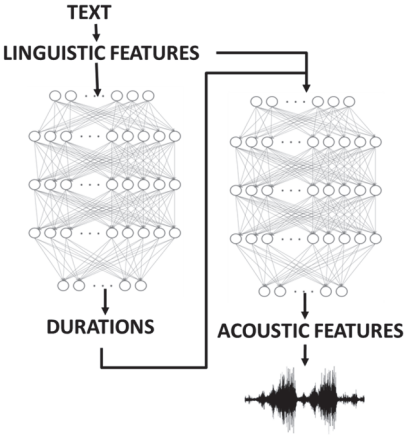

Fig. 1. DNN based TTS model

In the stage of synthesis, state-level durations predicted from the first network are used to extract additional features for the second network, which predicts acoustic features required by vocoder to produce waveforms. It was found that better results are achieved when dynamic acoustic features are used along with the static ones. For this reason, first and second derivatives of acoustic features are also used as targets during the training of the second network. In the synthesis stage, after those are predicted, they are only used by MLPG algorithm [19] in order to slightly correct static acoustic features trajectories. Those static features are smoothed this way and after that propagated to the vocoder.

3. Style modeling in TTS. In [21] approaches to an expressive speech synthesis are categorized in three different groups: implicit control, explicit control and playback approach. As it was mentioned before, expressive speech can be considered as an example of using different styles in speech.

Explicit control approaches are based on using some transformation rules to the prosody of a sentence synthesized in neutral style in order to get speech in some other style. These transformation rules are learned on a database of expressive speech. The examples of these transformations can be found in [22].

Playback approaches to an expressive speech synthesis create separate synthesizers for desired styles based only on expressive speech data. In [23] concatenative speech synthesis system capable of producing speech in three different styles is presented. Each style is synthesized using approximately 1 hour of corresponding speech data. HMM based TTS system that uses different acoustic models for each desired style is presented in [24]. In the same paper, another technique for expressive speech synthesis, which embeds style information in input linguistic features, is proposed. The authors have compared these approaches and concluded that they produce the same results regarding expressive speech quality.

Implicit approaches can be applied in statistical speech synthesis and are based on interpolation between different models. In [25] method which adapts a model of neutral speech to some desired style is described.

The comparison of concatenative and HMM based synthesis regarding expressive speech is given in [26]. It is concluded that concatenative synthesis performs better regarding overall emotion intensity, whereas the HMM approach is better for emotion intensity manipulation.

4. Style-code modeling in DNN based TTS. When we initially started experiments with proposed techniques there were almost no attempts at the expressive speech synthesis using DNN. Meanwhile, several papers were published where similar techniques are proposed [27-29].

Some research related to manipulation of voice intended for speaker modeling includes voice conversion [30], speaker adaptation [31] and multi-speaker synthesis [32]. Voice conversion approaches are based on parallel corpus of source and target speaker. The aim is to make conversion function, which when applied on speech of the source speaker should make it sound as the target speaker. The speaker adaptation approaches try to adapt already trained models towards some target speaker and do not require parallel corpus. Starting models are usually trained on a large training corpus and a small amount of target speaker speech material is used for model adaptation. Multi-speaker synthesis requires a collection of smaller datasets for training, and then, in the synthesis phase, decides on the voice that is to be synthesized.

In [31], [32] and [33] multi-speaker model is made by training a DNN on a speech material from a few to more than 100 speakers. One of the methods involves standard DNN-based single-speaker system, but extended with the feature that represents *speaker code* (Figure 2) [31]. The speaker code can be represented as a one-hot vector, or extended with additional information about a certain speaker, like gender, age, etc. [33]. This extension provides even better results, resembling on the idea of using *i-vectors* as speaker code [34]. In synthesis stage, the network will synthesize the speech in some speaker's voice by setting appropriate speaker code.

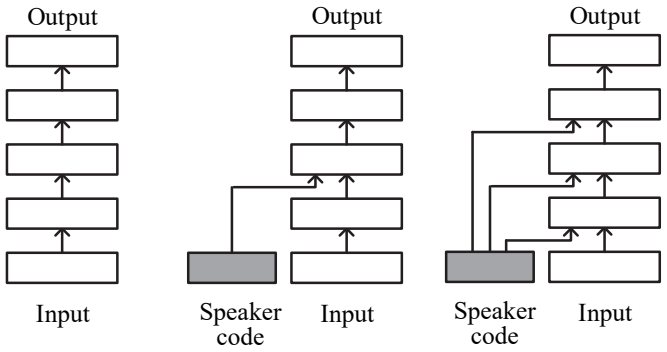

Fig. 2. The idea presented in [30] for multi-speaker DNN TTS model

The idea presented in this paper is to use a single one-hot vector to represent speaking style. It is solved by adding lexical questions of type *"is_style_x"*, where *x* can be neutral, angry, happy, etc. In case of using DNN model, since the input of the neural network represents binary label with value 1 on places representing questions on which answers are positive, part of the input will be one-hot vector indicating speaking style (Figure 3). In case of using HMM modeling, similar idea is used. Input label is extended with information that indicates if a phone belongs to speech of a certain style. In HMM modeling this actually corresponds to the idea presented in [24].

Usually, just a part of the sentence will be said with high expression of emotions and that is the reason why the style code is given per word, although the whole sentences were labeled with a marker of just one style in used database. This way, it should be possible to produce just a part of a sentence with an emotion, and the rest in neutral style.

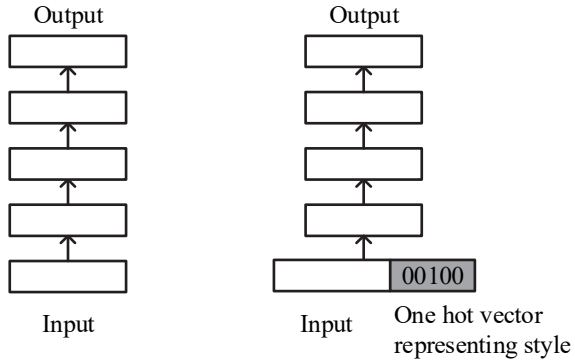

Fig. 3. The idea presented in [34] but modified for multi-style DNN TTS

In multi-style modeling, it is expected that the database consists of a lot more material of neutral speech than any other style. In case of multi-speaker modeling, such big differences between databases of speakers can badly influence final model (e.g. if some features of different speakers are averaged). On the other hand, this effect should have smaller impact in multi-style approach since all sentences are uttered by the same speaker. This could be actually beneficial for this model since the states are not covered with small database of certain speaking style can be compensated by neutral style and it will not badly influence the final model, since the voice is the same in the two databases.

5. Experiments and results.

DNN and HMM based systems.

DNN system is built using Merlin toolkit [35]. It consists of two neural networks, one for duration and one for acoustic modeling. Both networks contain 4 hidden layers with 1024 units per layer. Neurons in first 3 layers use tangent hyperbolic activation function, while the 4^{th} layer is based on long short-term memory (LSTM) architecture [36]. The output layer is linear. Input for duration model consists of answers to 554 closed lexical questions. These include questions about phoneme identity, number of phonemes and syllables in a word, ToBI tags [37] attached to a phoneme, etc. One-hot vector representing style is not included in the overall question number. Outputs are durations per HMM state. Input for acoustic model, beside the answers to the lexical questions, contains 9 additional features regarding state and phoneme durations, which were also introduced in Merlin toolkit. The acoustic

features are extracted by WORLD vocoder [38]. Since it produces smoothed envelopes, they are converted to mel-generalized cepstral coefficients (MGCs). At the end, the output feature vector for acoustic network contains 40 MGCs, 1 band aperiodicity (BAP) feature, logarithm of fundamental frequency, first and second derivatives of previously mentioned features and one feature representing information if a given frame is voiced or unvoiced (VUV).

HMM system is built by using HTS toolkit [39]. HMM models that are used in the system are 5-state left-to-right models with no skip, where each state is represented by a single Gaussian with the diagonal covariance matrix. Same acoustical features as for DNN system are used, except for the VUV feature. Input label consists of lexical features extracted in similar way as for DNN synthesis, but the number of available features was higher. Namely, HMM contained a combination of some features and initial experiments in DNN synthesis showed that usage of these complex questions does not improve the quality of synthesis. This can be explained by the DNN's capability to model some complex relations that HMMs cannot.

Databases.

In order to compare performance of the proposed technic in case of using HMM and DNN, 4h and 20 min of neutral speech and 20 min of speech in angry style were used. The same database is used in case of investigating quality of the synthesized speech, while in case of further investigation of the proposed technic for the DNN based synthesizer, the database is expanded with two more speaking styles – happy and apologetic. More precisely, 2h of neutral speech and 10 min per style are used for these experiments. The whole database is pronounced by a native American English male voice talent. It is recorded in a professional studio. Some statistics can be seen in Table 1.

Table 1. Statistics of the database

Style	Speech rate [phoneme/s]	Average f_0 [Hz]	std f_0 [Hz]
Neutral	12.7	98.7	34.1
Apologetic	10.8	101.9	25.1
Happy	11.4	170.2	71.4
Angry	10.9	103.9	30.3

It can be noticed that the happy style has significantly higher average fundamental frequency (f_0) than other styles, as well as more than two times higher standard deviation of f_0 in comparison to any other style. The neutral style is the fastest and has the lowest average f_0. Apologetic and angry styles have very similar characteristics with around 15% smaller standard deviation of f_0 in case of apologetic style.

Performance of the proposed technic for different synthesis method.

The two systems were first compared objectively. All objective measures were calculated as differences between acoustic features predicted by neural network and acoustic features extracted directly from original recordings. The objective measures that were used include:

– mel-cepstral distance (MCD) — mean square error of mel generalized cepstral coefficients calculated in decibels,

– mean square error of band aperiodicities (BAP) calculated in decibels,

– root mean square error of fundamental frequency (RMSE f_0) calculated in hertz,

– correlation of fundamental frequency (CORR f_0),

– percentage of correctly predicted voiced frames (V/UV).

The objective measures are calculated for test files only, for both systems and they are shown in Table 2. During the feature generation the original durations created during the DNN alignments procedure were used. Based on the results from Table 2 it can be concluded that DNN system outperforms HMM regarding all objective measures.

Table 2. Objective comparison between HMM and DNN systems

	MCD [dB]	BAP [dB]	RMSE f_0 [Hz]	CORR f_0	V/UV [%]
HMM	6.73	0.18	23.60	0.5	8.64
DNN	4.29	0.15	20.84	0.63	5.52

One example of fundamental frequency trajectory created by the tested systems is shown in Figure 4. Both trajectories are compared with trajectory extracted from the original recording. It can be seen that DNN-predicted trajectory resembles much better the original one than the HMM-predicted trajectory.

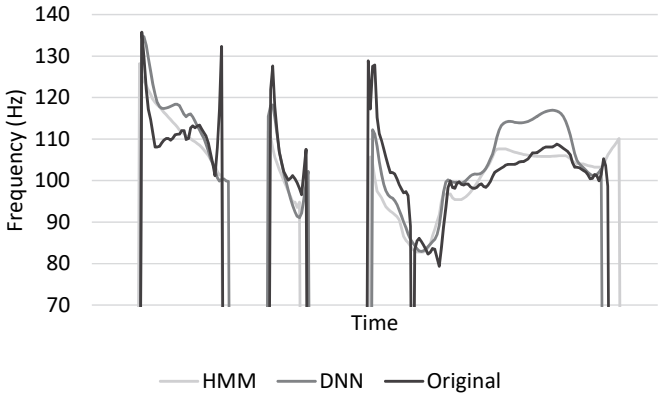

Fig. 4. Comparison of fundamental frequency trajectory created by the HMM and DNN systems with original trajectory

But it is known that objective measures are not always fully correlated with subjective assessment of overall speech quality. Subjective listening tests are considered as the most reliable. Due to this reason, for the comparison of the obtained results from HMM and DNN synthesizers by the proposed style-code method, preference tests were conducted among 24 amateur non-native listeners. They were asked to recognize which sentence, in each of 20 pairs, was pronounced in angry style. One of the possible options was also *No preference* (i.e. none of the sentences stands out). The first 10 pairs consisted of sentences pronounced in neutral and angry style, synthesized by DNN synthesizer, while the remaining 10 pairs (also pronounced in both styles per pair) were synthesized by HMM synthesizer. These results are given in Figure 5. It clearly shows much better results in case of DNN synthesis. As many as 39% of answers were *No preference* in case of HMM synthesis, and even in 1% listeners chose neutral sentence as the one where the angry emotion was better expressed. That leads to the conclusion that HMM model is not able to produce entirely clear difference between styles in case of the proposed technic. On the other hand, only 12% answers were *No preference*, in case of DNN synthesis, while remaining 88% of answers accurately recognized angry style. Since even in spontaneous human communication it is not always easily recognizable if the speaker intended to express some emotion, 88% can be considered as high accuracy and the proposed technic can be considered as very powerful although pretty simple.

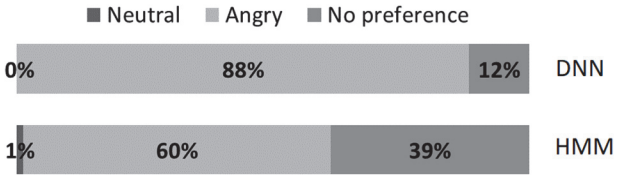

Fig. 5. Results of preference test regarding style expression

In order to directly compare DNN and HMM synthesis in angry style, the second listening test was conducted. It consisted of 10 pair of sentences where each pair of sentences was synthesized by both synthesizers. Listeners were asked to choose sentence in each pair in which angry emotion was more emphasized. As shown in Figure 6, exceptionally high preference for DNN synthesized sentences clearly confirms that the proposed technic has better performances in case of DNN.

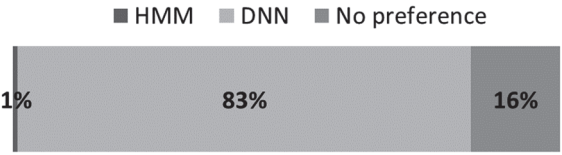

Fig. 6. Direct comparison of expressive speech synthesized by HMM and DNN based synthesizers

Quality of synthesized speech.

As previously mentioned, the key requirements for synthesized speech is intelligibility and naturalness. Intelligibility seized to be an issue a long time ago, but the naturalness is still a big issue in the field. Naturalness is defined as resemblance to human speech. Since for the experiment, almost 10 times bigger amount of neutral speech than angry-style speech was used, it was important to investigate if mismatch in the amount of the used material caused some loss in quality for speech synthesized in angry style, compared to the neutral one. In order to do that, additional listening test was conducted. In this test listeners were asked to grade the quality of 10 sentences synthesized by DNN synthesizer by giving grades between 1 (very bad) and 5 (very good). Among these, 5 are pronounced in angry and 5 in neutral style. Quality of

synthesis is defined to listeners as resemblance to human speech and the absence of the artifacts.

The obtained average grades for the synthesis in neutral style and the synthesis in angry style are almost the same, 3.9 against 3.8 in favor of neutral style. This proves the hypothesis that all states not covered by the small angry-style database are compensated by neutral style.

Objective comparison of results obtained with three different styles.

Since all of the previous experiments were performed by using only a single style (other than neutral), we also tested the performance of the approach in reproduction of other styles. In order to test the proposed method performance, separate synthesizers were constructed, where each one is capable of producing a single style. These synthesizers were created by using 2h of neutral speech and 10 minutes of intended style (angry, apologetic or happy). The results were compared only objectively since subjective comparison of different speaking styles does not represent an adequate approach, due to significant differences in the manners in which the styles are expressed.

For each synthesizer, objective measures were calculated by using 30 test sentences of corresponding speaking style, which were not a part of the training data. The results are given in Table 3. Among the presented styles, the best objective measures are achieved with apologetic style and mostly the worst for happy style. It is the most emphasized for RMSE of f_0 which is 41.48 Hz in case of happy style. This can be explained by its characteristics — mean frequency as well as standard deviation are the highest among all the styles (see Table 1). On the other hand, correlation of f_0 is the best among all three styles.

Table 3 Objective measures 1-style modeling

	MCD [dB]	BAP [dB]	RMSE f_0 [Hz]	CORR f_0	V/UV [%]
Happy	5.50	0.19	41.48	0.79	5.59
Apologetic	4.70	0.13	16.85	0.73	4.88
Angry	4.79	0.17	18.67	0.62	5.68

The obtained results show that some differences exist in modeling different styles, which can be explained by the differences in original part of databases.

It should also be noticed that the objective measures in Tables 3 and 4 for angry style are worse compared to ones from Table 2. This can be

explained by the amount of material of this style being used in training process. Specifically, the amount of angry style material used in experiments whose results are presented in Table 2 was 20 minutes, while in other 2 experiments only 10 minutes of angry speech was used.

Table 4 Objective measures 3-style modeling

	MCD [dB]	BAP [dB]	RMSE f_0 [Hz]	CORR f_0	V/UV [%]
Happy	5.46	0.19	42.85	0.77	5.64
Apologetic	4.67	0.13	17.11	0.72	4.92
Angry	4.75	0.17	18.46	0.63	5.70

Comparison of single-style and multi-style modeling.

As already mentioned in previous sections, the style-code approach can be applied to simultaneous modeling of an arbitrary number of styles. Therefore, we wanted to compare single-style modeling with multi-style approach. For these purposes we constructed a new synthesizer, which was trained on full previously introduced database — 2 hours of neutral speech and 10 minutes of speech of each of the three additional styles. The objective measures calculated in the same manner as for single-style are given in Table 4. As in the case of single-style modeling, the best results are obtained for the apologetic style, while the measures for the happy style again were the worst among all three styles.

The average measures per all three styles for those two approaches are presented in Table 5. These results indicate that measures for both approaches differ only in minor.

Table 5 Average objective measures for 1-style and 3-style modeling

	MCD [dB]	BAP [dB]	RMSE f_0 [Hz]	CORR f_0	V/UV [%]
1-style	5.00	0.16	25.55	0.71	5.38
3-style	4.96	0.16	26.14	0.71	5.42

We also conducted one subjective test, in which 15 non-native listeners were included. The task consisted of 30 pairs of sentences. In each pair there was one sentence synthesized with 1-style approach and another synthesized with 3-style approach. Each of three styles was represented with 10 pairs. Participants had to choose in which of the two sentences the

presented style was better expressed. The intended style of each pair of sentences was clearly presented to the participants. There was also *No preference* as a possible answer.

The results are given in Figure 7. It can be seen that, on average, multi-style model is preferred over the other one. Only for angry style percentage is the same. *No preference* answer is chosen most often in case of apologetic and angry style, while in case of the happy style, multi-style model is chosen significantly more often than the other two possible answers.

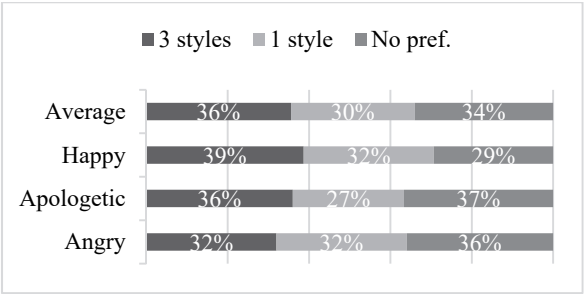

Fig.7. Results of subjective tests for comparing single-style and multi-style modeling

As conclusion it can be stated that there is no significant difference between modeling one style per model or multiple at once, although the slight advantage can be given to multi-style model, probably because of better generalization.

Analysis based on parallel corpus.

There are a few sentences in our database originally recorded in different styles but with the same content spoken (i.e. parallel corpus). We have chosen one of these sentences and performed further analysis in order to check the impact of style code to parameter generation, as well as the impact of some linguistic differences to generated parameters. The chosen sentence is originally recorded in happy and apologetic style and annotated in accordance to the actual prosodic events. In this particular sentence, there are 2 more phrase breakes in apologetic sentence in comparison to the happy one. All other prosodic annotations are the same.

In this experiment, the focus was on predicted fundamental frequency trajectories. Although the same content was spoken in both sentences the length of corresponding phonemes is not the same. Namely, according to Table 1 the phoneme rate is slightly higher for apologetic style.

This makes the direct comparison not possible. In order to make direct comparison feasable original durations were used.

Figure 8 shows generated trajectories when lingustic annotations and durations of original apologetic sentence are used. It can be seen that fundamental frequency of the original recording (black line) does not vary much (mostly it is between 60 Hz and 120 Hz) and its average is around 95 Hz, which is in accordance to the style characteristic (Table 1). It can also be noticed that the synthesis in apologetic style follow the original trajectory very well. On the other hand, synthesis of the same sentence with the same durations and annonations and switching only the style code to happy produced different curve. The average frequency is increased for about 65 Hz, and it varies a lot more than the other two curves in Figure 8, reaching even 300 Hz. This behaviour is also in accordance to happy style characteristics presented in Table 1.

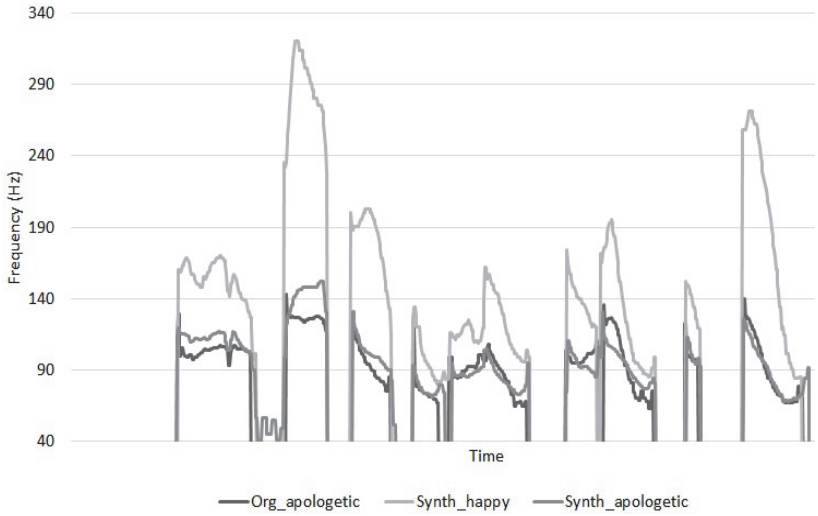

Fig. 8. Fundamental frequency trajectories of the sentences originally anotated as apologetic, synthesized in happy and apologetic style

The case when annotations and durations for original happy sentence are used are shown in Figure 9. Again, it can be seen that the sentence synthesized in apologetic style, has a lot lower fundamental frequency compared to both, original and sentence synthesized in happy style. Synthesized happy sentence follows the original trajectory of fundamental

frequency very well, but sometimes fails to track some rush changes from original recording. This can be explained by certain smoothing that is introduced by the model.

The case when annotations and durations for original happy sentence are used are shown in Figure 9. Again, it can be seen that the sentence synthesized in apologetic style, has a lot lower fundamental frequency compared to both, original and sentence synthesized in happy style. Synthesized happy sentence follows the original trajectory of fundamental frequency very well, but sometimes fails to track some rush changes from original recording. This can be explained by certain smoothing that is introduced by the model.

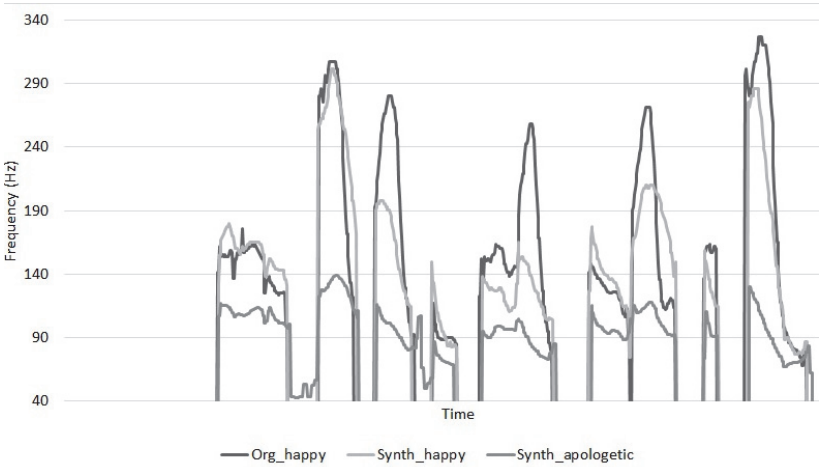

Fig. 9. Fundamental frequency trajectories of the sentences originally anotated as happy, synthesized in happy and apologetic style

Both examples prove that the style code itself succesfully predicts important style characteristics and that these characteristics are not highly dependent on linguistic input.

The impact of linguistic differences on the generated trajectories can be analyzed in Figure 10. These trajectories are generated using linguistic features from both original recordings, which have the same spoken content but differ in their annotations. The biggest differences in the generated trajectories can be observed in the middle of the sentence where the phrase brakes are annotated differently.

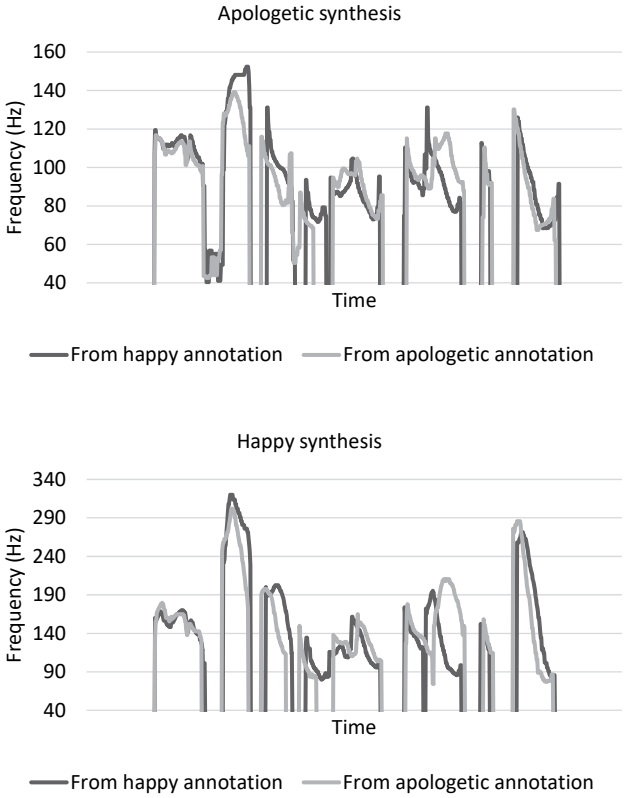

Fig. 10. Comparison of linguistic features impact on fundamental frequency trajectory

One more characteristic that the model learned, but cannot be seen from figures, is loudness of certain styles. In case of apologetic style, the speech is noticeable quieter, e.g. in this particular sentence, the difference between synthesized happy and apologetic sentence is around 10 dB.

4. Conclusion. In this paper we present a simple but very effective method for incorporating styles into statistical speech synthesis. It is based on style codes, similar to earlier introduced speaker codes, and consists of adding one-hot vector representing style to the auxiliary inputs used in NN-based speech synthesis.

The proposed method is also applicable to HMM based speech synthesis. Objective and subjective results show that the proposed method achieves higher performances in case of DNN systems. Objective and

subjective results also suggest that, although the amount of speech material of certain style is much smaller compared to material in neutral style, quality of speech synthesized in certain style is preserved. It is proved that the method is applicable to any speaking style and that there is no significant difference if the model learns multiple styles at once or one by one.

Although the duration is important feature in expressing some style, the focus of this paper was on analysis of the acoustic features. Some analysis on the duration prediction performance should be performed.

Bearing in mind that humans are able to control the level of expressed emotion in speech and rarely one emotion is completely distinguished from another, some future work should investigate possibilities of controlling level of expressed emotion as well as mixing them.

References

1. Hunt A.J., Black A.W. Unit selection in a concatenative speech synthesis system using a large speech database. Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP-96). 1996. vol. 1. pp. 373–376.
2. Tokuda K. et al. Speech synthesis based on hidden Markov models. Proceedings of the IEEE. 2013. vol. 101. no. 5. pp. 1234–1252.
3. Watts O. et al. From HMMs to DNNs: where do the improvements come from? International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. pp. 5505–5509.
4. Ling Z.H. et al. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Processing Magazine*. 2015. vol. 32. no. 3. pp. 35–52.
5. Yu D., Deng L, Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*. 2014. vol. 7. no. 3-4. pp. 198–387.
6. Qian Y., Fan Y., Hu W., Soong F.K. On the Training Aspects of Deep Neural Network (DNN) for Parametric TTS Synthesis. International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014. pp. 3829–3833.
7. Delić T., Sečujski M., Sinteza govora na srpskom jeziku zasnovana na veštačkim neuralnim mrežama. Telecommunication forum (TELFOR 2016). 2016. pp. 403–406.
8. Solomennik A.I., Chistikov P.G., Evaluation of naturalness of synthesized speech with different prosodic models. Proceedeings International conference on Computational Linguistics and Intellectual Technologies “Dialogue 2013”. 2013. 7 p.
9. Abe M. Speaking Styles: Statistical Analysis and Synthesis by a Text-to-Speech System. Progress in speech synthesis. 1997. pp. 495–510.
10. Rusko M. et al. Expressive Speech Synthesis for Critical Situations. *Computing and Informatics*. 2015. vol. 33. no. 6. pp. 1312–1332.
11. Delić T. et al. Multi- style Statistical Parametric TTS. Proceedings Digital speech and image processing (DOGS 2017). 2017. pp. 5–8.
12. Wu Z., Valentini-Botinhao C., Watts O., King S. Deep Neural Networks employing multi-task learning and stacked bottleneck features for speech synthesis. International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. pp. 4460–4464.
13. Watts O. et al. From HMMs to DNNs: Where do the improvements come from? International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. pp. 5505–5509.

14. 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. pp. 32–37.
15. Zen H., Tokuda K., Black A.W. Statistical parametric speech synthesis. *Speech Communication*. 2009. vol. 51. no. 11. pp. 1039–1064.
16. Zen H. et al. A hidden semi-Markov model-based speech synthesis system. *IEICE transactions on information and systems*. 2007. vol. 90. no. 5. pp. 825–834.
17. Yoshimura T. et al. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. Sixth European Conference on Speech Communication and Technology. 1999. 4 p.
18. Toda T., Tokuda K. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis, *IEICE transactions on information and systems*. 2007. vol. E90-D. no. 5. pp. 816–824.
19. Tokuda K. et al. Speech Parameter Generation Algorithms for HMM-based Speech Synthesis. International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2000. pp. 1315–1318.
20. Rabiner L.R. A tutorial on hidden Markov models and selectedapplications in speech recognition. Proceedings of the IEEE. 1989. vol. 77. no. 2. pp. 257–286.
21. Schröder M. Expressive speech synthesis: Past, present, and possible futures. Affective information processing. 2009. pp. 111–126.
22. Tao J., Kang Y., Li A. Prosody conversion from neutral speech to emotional speech. *IEEE Transactions on Audio, Speech, and Language Processing*. 2006. vol. 14. no. 4. pp. 1145–1153.
23. Iida A., Campbell N., Higuchi F., Yasumura M. A corpus-based speech synthesis system with emotion. *Speech Communication*. 2003. vol. 40. no. 1–2. pp. 161–187.
24. Yamagishi J., Onishi K., Masuko T., Kobayashi T. Modeling of various speaking styles and emotions for HMM-based speech synthesis. Eighth European Conference on Speech Communication and Technology. 2003. pp. 2461–2464.
25. Yamagishi J. et al. Model adaptation approach to speech synthesis with diverse voices and styles. International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007). 2007. vol. 4. p. IV-1233–IV-1236.
26. Barra-Chicote R. et al. Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. *Speech Communication.* 2010. vol. 52. no. 5. pp. 394–404.
27. Inoue K. et al. An investigation to transplant emotional expressions in DNN-based TTS synthesis. Proc. APSIPA Annual Summit and Conference. 2017. pp. 1253–1258.
28. An S., Ling Z., Dai L. Emotional statistical parametric speech synthesis using LSTM-RNNs. Proc. APSIPA Annual Summit and Conference. 2017. pp. 1613–1616.
29. Lorenzo-Trueba J. et al. Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis. *Speech Communication*. 2018. vol. 99. pp. 135–143.
30. Stylianou Y., Cappe O., Moulines E. Statistical Methods for Voice Quality Transformation. Fourth European Conference on Speech Communication and Technology. 1995. pp. 447–450.
31. Hojo N., Ijima Y., Mizuno H. An Investigation of DNN-Based Speech Synthesis Using Speaker Codes. INTERSPEECH. 2016. pp. 2278–2282.
32. Fan Y., Qian Y., Soong F.K., He L. Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis. International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. pp. 4475–4479.
33. Luong H.T., Takaki S., Henter G.E., Yamagishi J. Adapting and controlling DNN-based speech synthesis using input codes. International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017. pp. 4905–4909.

34. Yang S., Wu Z., Xie L. On the Training of DNN-based Average Voice Model for Speech Synthesis. Proc. Signal and Information Processing Association Annual Summit and Conference (APSIPA). 2016. pp. 1–6.
35. Wu Z., Watts O., King S. Merlin: An Open Source Neural Network Speech Synthesis System. Proc. 9th ISCA Speech Synthesis Workshop (SSW9). 2016. pp. 218–223.
36. Fan Y., Qian Y., Xie F.L., Soong F.K. TTS synthesis with bidirectional LSTM based recurrent neural networks. Fifteenth Annual Conference of the International Speech Communication Association INTERSPEECH. 2014. pp. 1964–1968.
37. Silverman K. et al. ToBI: A standard for labeling English prosody. Proc. International Conference on Spoken Language Processing (ICSLP). 1992. pp. 867–870.
38. Morise M., Yokomori F., Ozawa K. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE transactions on information and systems*. 2016. vol. E99-D. no. 7. pp. 1877–1884.
39. Zen H. et al. The HMM-based speech synthesis system (HTS) version 2.0. Proc. Sixth ISCA Workshop on Speech Synthesis. 2007. pp. 294–299.

Suzić Siniša — researcher of Laboratory of Acoustics and Speech Technology of Faculty of Technical Sciences, University of Novi Sad. Research interests: expressive speech synthesis, digital signal processing, dialogue systems, machine learning, deep neural networks. The number of publications — 22. sinisa.suzic@uns.ac.rs; 6, Trg Dositeja Obradovića, 21000, Novi Sad, Serbia; office phone: +381-21-485-2521.

Delić Tijana Vlado — researcher of Laboratory of Acoustics and Speech Technology of Faculty of Technical Sciences, University of Novi Sad. Research interests: expressive speech synthesis, digital signal processing, dialogue systems, machine learning, deep neural networks. The number of publications — 18. tijanadelic@uns.ac.rs; ; office phone: +381(21)485-2521.

Ostrogonac Stevan — senior researcher, AlfaNum – Speech Technologies, software developer, AlfaNum – Speech Technologies. Research interests: text-to-speech synthesis, automatic speech recognition, natural language processing, dialogue systems, development of speech and language resources, machine learning, neural networks. The number of publications — 39. ostrogonac.stevan@alfanum.co.rs; 40, Bulevar Vojvode Stepe, 21000, Novi Sad, Serbia; office phone: +381-64-845-5302.

Đurić Simona — researcher of Laboratory of Acoustics and Speech Technology of Faculty of Technical Sciences, University of Novi Sad. Research interests: expressive speech synthesis, digital signal processing, dialogue systems, machine learning, deep neural networks. The number of publications — 6. simona.djuric@uns.ac.rs; 6, Trg Dositeja Obradovića, 21000, Novi Sad, Serbia; office phone: +381(21)485-2521.

Pekar Darko Jovan — research assistant of the Department for Power, Electronic and Telecommunications Engineering of the Faculty of Technical Sciences, University of Novi Sad, CEO (Chief Executive Officer), AlfaNum Speech Technologies. Research interests: human-computer interaction, speech recognition and synthesis, speaker identification, emotion recognition, speech morphing, numerical simulations, artificial intelligence. The number of publications — 100. darko.pekar@alfanum.co.rs; 40, Bulevar Vojvode Stepe, 21000, Novi Sad, Serbia; office phone: +381-21-485-2521.

Acknowledgements. The research is supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia (grant TR32035).

УДК 006.72 **DOI** 10.15622/sp.60.8

С. Сузич, Т.В. Делич, С. Острогонац, С. Джурич, Д.Й. Пекар

МЕТОД СТИЛЕВЫХ КОДОВ ДЛЯ МНОГОСТИЛЕВОГО ПАРАМЕТРИЧЕСКОГО СИНТЕЗА РЕЧИ ПО ТЕКСТУ

Сузич С., Делич Т.В., Острогонац С., Джурич С., Пекар Д.Й. **Метод стилевых кодов для многостилевого параметрического синтеза речи по тексту.**

Аннотация. Современные системы преобразования текста в речь обычно обеспечивают хорошую разборчивость. Один из главных недостатков этих систем — отсутствие выразительности по сравнению с естественной человеческой речью. Особенно неприятно воспринимается на слух, когда автоматическая система передает утвердительные и отрицательные предложения совершенно одинаково. Введение параметрических методов в синтезе речи дало возможность легко изменять характеристики говорящего и стили речи. В статье представлен простой способ включения стилей в синтезированную речь, используя стилевые коды.

Предлагаемый метод требует не более пары минут заданного стиля и некоторый объем данных нейтральной речи. Он успешно применяется в синтезе речи на глубоких нейронных сетях и в скрытых марковских моделях, предоставляя стилевой код как дополнительный вклад в модель. Аудирование подтвердило, что наибольшая выразительность достигается за счет синтеза глубоких нейронных сетей по сравнению с синтезом скрытых марковских моделей. Также доказано, что качество речи, синтезированное глубокими нейронными сетями в определенном стиле, сопоставимо с речью, синтезированной в нейтральном стиле, хотя база данных нейтральной речи примерно в 10 раз больше. Глубокие нейронные сети на основе синтеза речи по тексту со стилевыми кодами изучаются путем сравнения качества речи, создаваемой системами одностилевого моделирования и многостилевого моделирования. Объективные и субъективные измерения подтвердили, что между этими двумя подходами нет существенной разницы.

Ключевые слова: синтез речи по тексту, экспрессивный синтез речи, глубокие нейронные сети, стиль речи, стилевой код, прямой унитарный вектор.

Сузич Синиша — научный сотрудник лаборатории акустики и речи факультета технических наук, Нови-Садский университет. Область научных интересов: синтез выразительной речи, обработка цифровых сигналов, диалоговая система, машинное обучение, глубокие нейронные сети. Число научных публикаций — 22. sinisa.suzic@uns.ac.rs; Трг Доситея Обрадовича, 6, 21000, Нови Сад, Сербия; р.т.: +381(21)485-2521

Делич Тийана Владо — научный сотрудник лаборатории акустики и речи факультета технических наук, Нови-Садский университет. Область научных интересов: синтез выразительной речи, обработка цифровых сигналов, диалоговая система, машинное обучение, глубокие нейронные сети. Число научных публикаций — 18. tijanadelic@uns.ac.rs; Трг Доситея Обрадовича, 6, 21000, Нови Сад, Сербия; р.т.: +381(21)485-2521.

Острогонац Стеван — старший научный сотрудник, AlfaNum – Speech Technologies Ltd, разработчик программного обеспечения, AlfaNum – Speech Technologies Ltd. Область научных интересов: синтез речи, автоматическое распознавание речи, обработка естественного языка, диалоговая система, разработка речевых и языковых ресурсов, машинное обучение, нейронные сети. Число научных публикаций — 39. ostrogonac.stevan@alfanum.co.rs; бул. Войводе Степе, 40, 21000, Нови Сад, Сербия; р.т.: +381-64-845-5302.

Джурич Симона — научный сотрудник лаборатории акустики и речи факультета технических наук, Нови-Садский университет. Область научных интересов: синтез выразительной речи, обработка цифровых сигналов, диалоговая система, машинное обучение, глубокие нейронные сети. Число научных публикаций — 6. simona.djuric@uns.ac.rs; Трг Доситейа Обрадовича, 6, 21000, Нови Сад, Сербия; р.т.: +381(21)485-2521.

Пекар Дарко Йован — младший научный сотрудник департамента энергетики, электроники и телекоммуникационного инжиниринга факультета технических наук, Нови-Садский университет, главный исполнительный директор, AlfaNum Speech Technologies. Область научных интересов: человеко-машинное взаимодействие, распознавание и синтез речи, идентификация диктора, морфинг речи, статистический анализ, искусственный интеллект. Число научных публикаций — 100. darko.pekar@alfanum.co.rs; бул. Войводе Степе, 40, 21000, Нови Сад, Сербия; р.т.: +381-21-485-2521.

Поддержка исследований. Работа выполнена при финансовой поддержке Министерства образования, науки и технологического развития Республики Сербия (грант TR32035).

Литература

1. *Hunt A.J., Black A.W.* Unit selection in a concatenative speech synthesis system using a large speech database // Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP-96). 1996. vol. 1. pp. 373–376.
2. *Tokuda K. et al.* Speech synthesis based on hidden Markov models // Proceedings of the IEEE. 2013. vol. 101. no. 5. pp. 1234–1252.
3. *Watts O. et al.* From HMMs to DNNs: where do the improvements come from? // International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. pp. 5505–5509.
4. *Ling Z.H. et al.* Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends // IEEE Signal Processing Magazine. 2015. vol. 32. no. 3. pp. 35–52.
5. *Yu D., Deng L*. Deep learning: methods and applications // Foundations and Trends® in Signal Processing. 2014. vol. 7. no. 3-4. pp. 198–387.
6. *Qian Y., Fan Y., Hu W., Soong F.K.* On the Training Aspects of Deep Neural Network (DNN) for Parametric TTS Synthesis // International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014. pp. 3829–3833.
7. *Delić T., Sečujski M.* Sinteza govora na srpskom jeziku zasnovana na veštačkim neuralnim mrežama // Telecommunication forum (TELFOR 2016). 2016. pp. 403–406.
8. *Solomennik A.I., Chistikov P.G.* Evaluation of naturalness of synthesized speech with different prosodic models // Proceedeings International conference on Computational Linguistics and Intellectual Technologies “Dialogue 2013”. 2013. 7 p.
9. *Abe M.* Speaking Styles: Statistical Analysis and Synthesis by a Text-to-Speech System // Progress in speech synthesis. 1997. pp. 495–510.
10. *Rusko M. et al.* Expressive Speech Synthesis for Critical Situations // Computing and Informatics. 2015. vol. 33. no. 6. pp. 1312–1332.
11. *Delić T. et al.* Multi- style Statistical Parametric TTS // Proceedings Digital speech and image processing (DOGS 2017). 2017. pp. 5–8.
12. *Wu Z., Valentini-Botinhao C., Watts O., King S.* Deep Neural Networks employing multi-task learning and stacked bottleneck features for speech synthesis // International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. pp. 4460–4464.

13. *Watts O. et al.* From HMMs to DNNs: Where do the improvements come from? // International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. pp. 5505–5509.
14. *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. pp. 32–37.
15. *Zen H., Tokuda K., Black A.W.* Statistical parametric speech synthesis // Speech Communication. 2009. vol. 51. no. 11. pp. 1039–1064.
16. *Zen H. et al.* A hidden semi-Markov model-based speech synthesis system // IEICE transactions on information and systems. 2007. vol. 90. no. 5. pp. 825–834.
17. *Yoshimura T. et al.* Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis // Sixth European Conference on Speech Communication and Technology. 1999. 4 p.
18. *Toda T., Tokuda K.* A speech parameter generation algorithm considering global variance for HMM-based speech synthesis // IEICE transactions on information and systems. 2007. vol. E90-D. no. 5. pp. 816–824.
19. *Tokuda K. et al.* Speech Parameter Generation Algorithms for HMM-based Speech Synthesis // International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2000. pp. 1315–1318.
20. *Rabiner L.R.* A tutorial on hidden Markov models and selectedapplications in speech recognition // Proceedings of the IEEE. 1989. vol. 77. no. 2. pp. 257–286.
21. *Schröder M.* Expressive speech synthesis: Past, present, and possible futures // Affective information processing. 2009. pp. 111–126.
22. *Tao J., Kang Y., Li A.* Prosody conversion from neutral speech to emotional speech // IEEE Transactions on Audio, Speech, and Language Processing. 2006. vol. 14. no. 4. pp. 1145–1153.
23. *Iida A., Campbell N., Higuchi F., Yasumura M.* A corpus-based speech synthesis system with emotion // Speech Communication. 2003. vol. 40. no. 1-2. pp. 161–187.
24. *Yamagishi J., Onishi K., Masuko T., Kobayashi T.* Modeling of various speaking styles and emotions for HMM-based speech synthesis // Eighth European Conference on Speech Communication and Technology. 2003. pp. 2461–2464.
25. *Yamagishi J. et al.* Model adaptation approach to speech synthesis with diverse voices and styles // International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007). 2007. vol. 4. p. IV-1233–IV-1236.
26. *Barra-Chicote R. et al.* Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech // Speech Communication. 2010. vol. 52. no. 5. pp. 394–404.
27. *Inoue K. et al.* An investigation to transplant emotional expressions in DNN-based TTS synthesis // Proc. APSIPA Annual Summit and Conference. 2017. pp. 1253–1258.
28. *An S., Ling Z., Dai L.* Emotional statistical parametric speech synthesis using LSTM-RNNs // Proc. APSIPA Annual Summit and Conference. 2017. pp. 1613–1616.
29. *Lorenzo-Trueba J. et al.* Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis // Speech Communication. 2018. vol. 99. pp. 135–143.
30. *Stylianou Y., Cappe O., Moulines E.* Statistical Methods for Voice Quality Transformation // Fourth European Conference on Speech Communication and Technology. 1995. pp. 447–450.
31. *Hojo N., Ijima Y., Mizuno H.* An Investigation of DNN-Based Speech Synthesis Using Speaker Codes // INTERSPEECH. 2016. pp. 2278–2282.

32. *Fan Y., Qian Y., Soong F.K., He L.* Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis // International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. pp. 4475–4479.
33. *Luong H.T., Takaki S., Henter G.E., Yamagishi J.* Adapting and controlling DNN-based speech synthesis using input codes // International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017. pp. 4905–4909.
34. *Yang S., Wu Z., Xie L.* On the Training of DNN-based Average Voice Model for Speech Synthesis // Proc. Signal and Information Processing Association Annual Summit and Conference (APSIPA). 2016. pp. 1–6.
35. *Wu Z., Watts O., King S.* Merlin: An Open Source Neural Network Speech Synthesis System // Proc. 9th ISCA Speech Synthesis Workshop (SSW9). 2016. pp. 218–223.
36. *Fan Y., Qian Y., Xie F.L., Soong F.K.* TTS synthesis with bidirectional LSTM based recurrent neural networks // Fifteenth Annual Conference of the International Speech Communication Association INTERSPEECH. 2014. pp. 1964–1968.
37. *Silverman K. et al.* ToBI: A standard for labeling English prosody // Proceedings of International Conference on Spoken Language Processing (ICSLP). 1992. pp. 867–870.
38. *Morise M., Yokomori F., Ozawa K.* WORLD: a vocoder-based high-quality speech synthesis system for real-time applications // IEICE transactions on information and systems. 2016. vol. E99-D. no. 7. pp. 1877–1884.
39. *Zen H. et al.* The HMM-based speech synthesis system (HTS) version 2.0 // Proceedings of Sixth ISCA Workshop on Speech Synthesis. 2007. pp. 294–299.