HIGH QUALITY SPEECH SYNTHESIS USING A SMALL SPEECH DATASET

Pavel Chistikov[1], *Andrey Talanov*[2]

[1]St. Petersburg National Research University Of Information Technologies, Mechanics and Optics,
49, Kronverkskiy pr., St. Petersburg, Russia, 197101
http://en.ifmo.ru, chistikov@speechpro.com
[2]Speech Technology Center Ltd., 4 Krasutskogo st., St. Petersburg, Russia, 196084
http://www.speechpro.com, andre@speechpro.com

ABSTRACT

We propose an approach to synthesizing high-quality speech under the conditions of a small dataset. A robust method for solving this problem is vital for voice restoration (recreation of lost fragments of records based on available speech material of a well-known person, e.g. an actor). The proposed TTS system is a hybrid system which includes the advantages of both HMM- and Unit Selection-based TTS systems. The approach described in the paper is based on statistical models of intonation parameters and special algorithms of speech element concatenation and modification. Listening tests show that it is possible to synthesize high-quality speech even with a small speech database (approximately one hour of speech).

Index Terms— speech synthesis, voice restoration, hidden Markov models, Unit Selection, speech modification.

1. INTRODUCTION

In recent years speech synthesis technology has been strongly improved, a lot of research has been carried out. As a result synthesized speech sounds very natural now and we can hear it in more and more places. The most popular approach for obtaining high-quality speech is Unit Selection [1,2]. HMM-based TTS systems sound worse due to buzzy effects [3]. But the price of such good quality is the necessity to have a large speech database (up to 10 hours) [2,4]. It is worth noting that each sound file contained in the database must be labeled with high accuracy, which increases costs and time expenses [2]. Such TTS systems can be used for call centers, audio book reading, voice assistance systems, etc. That is an area where it is not important whose voice is used; the quality and pleasantness of the voice are fundamental.

However, there are some applications where it is crucial to synthesize a specific voice. These include voice cloning, as well as voice restoration, i.e. the recreation of lost fragments of records based on available speech material of a well-known person (e.g. an actor, public speaker, etc.). The complexity of this task lies in the small amount of available speech data. The main reason for this is that usually the existing recordings of the speaker are acoustically various: they were made with different microphones in dissimilar conditions over a long time period. This means that we only have a small speech dataset of sufficiently good quality to create a voice suitable for a text-to-speech system.

There is some research on synthesizing speech from under-resourced speech data [5-8]. All of them are based on speech corpora of a non-target speaker. Then the target voice is produced by adaptation techniques applied directly to speech elements or acoustic models (which are usually HMMs). Both of these approaches do not produce natural speech due to a deteriorating effect of applying adaptation in the acoustic domain. To avoid this problem, we propose a hybrid TTS system [9] where the intonation of the target speaker is modeled on another speaker's database, and adaptation techniques are applied to emulate voice characteristics with maximum precision. Speech synthesis is performed by using the target speaker speech database and a Unit Selection algorithm with special methods for modification and concatenation of speech elements. Using such complex techniques makes it possible to synthesize high-quality speech even with a small speech database (approximately one hour of speech), which is confirmed by expert listening tests.

This paper is organized as follows: the description of the proposed system is presented in Section 2, which comprises intonation modeling, unit selection and the modification and concatenation algorithms; experimental results illustrating the system's performance are included in Section 3; conclusions and future developments are presented in Section 4.

2. THE PROPOSED SYSTEM

Structurally, the text-to-speech system consists of two parts: the training part (Figure 1) and the synthesis part (Figure 2).

The main purpose of the training part is creating the target voice model (to emulate speaker parameters:

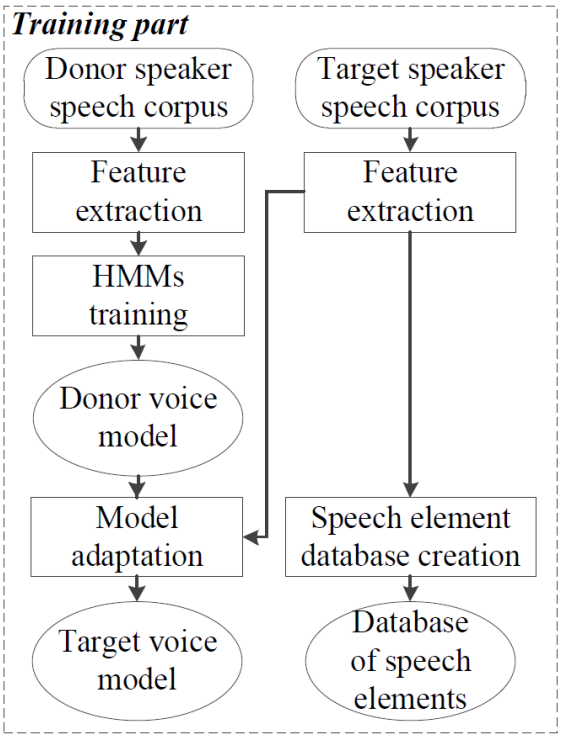

Fig. 1. *Diagram illustrating the basic steps conducted by the training part*

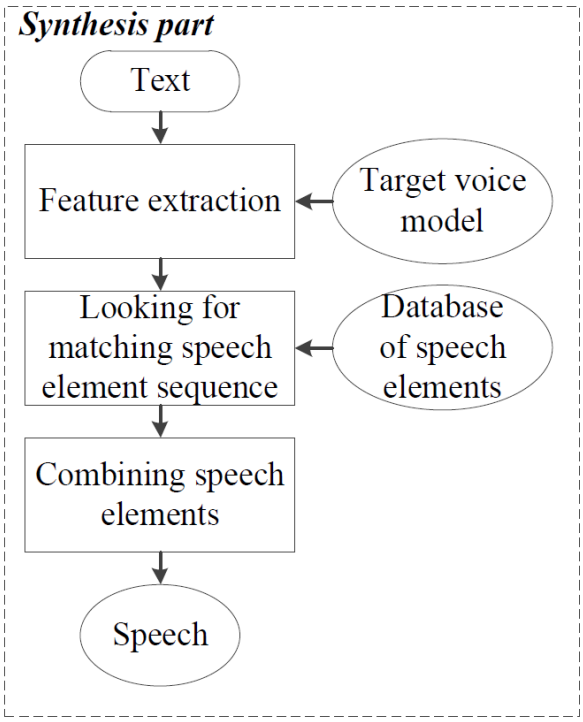

Fig. 2. *Diagram illustrating the basic steps conducted by the synthesis part*

mel-frequency cepstral coefficients (MFCC), pitch (F0), energy and duration) and the database of speech elements. To perform this we need two speech datasets: a donor speaker speech dataset (contains about 8-10 hours of speech of a speaker in the same language as the target speaker) and a target speaker speech dataset (a small speech dataset of the voice to be synthesized). Each speech dataset contains a set of sound files (each file contains a single recorded sentence) and a set of corresponding label files (these contain information about the speech elements in each sound file) [10-12]. First of all, linguistic and acoustic features [9, 14] are calculated for both of the speech datasets. Then, on the one hand, a speech database is created based on the target speaker speech material. The speech database contains an indexed element set to provide fast search by the following features: phone name, names of phones before and after the current phone, MFCC at phone boundaries, energy, pitch, and phone duration. On the other hand, the target speaker voice model is trained. This model is a set of HMMs which generalize sound element parameters (MFCC, pitch, energy and duration) in different contexts. A detailed description of the target speaker voice model creation is presented in section 2.1.

Speech synthesis is performed based on the target speaker voice model and the database of speech elements prepared at the previous step. The TTS system input is raw text without any manual preprocessing. Based on the input text, the target allophone sequence is formed, and linguistic and prosodic features are calculated for each allophone. The type and structure of features are the same as those used at the stage of the speech database building. Using this information and the voice model, acoustic features are calculated for each allophone: MFCC, pitch, energy and duration. Then the most appropriate speech elements are selected from the database based on the calculated acoustic features using the Unit Selection algorithm [4]. Then the selected sound elements must be smoothed and concatenated to each other in order to produce synthesized speech. Those final steps are the most important, particularly taking into account the small size of the target speaker database. In these conditions it is likely to be impossible to find appropriate speech elements corresponding well enough to the model, which is necessary to realize appropriate intonation, and to each other, which is extremely important for smoothness of synthesized speech. So it is necessary to have special techniques that ensure the fulfilment of both these requirements. They are described in sections 2.2 and 2.3 correspondingly.

2.1. Intonation modeling

The modeling of intonation parameters begins with the extraction of the feature set from all sound files [13]. Each member of the set represents a short part of the signal

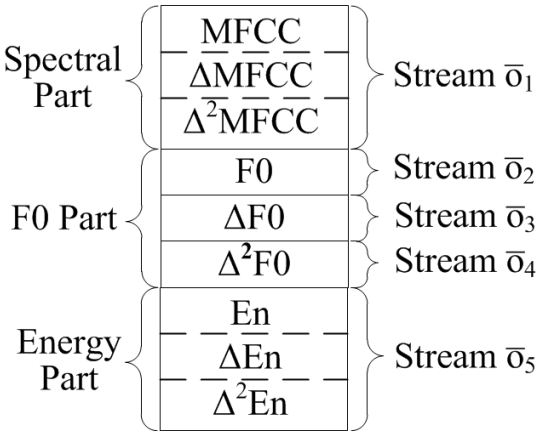

Fig. 3. *Observation vector*

(frame) with the length of 25 ms. The features contain the following parameters:

- Sequence $\{\mathbf{C}_1, \ldots, \mathbf{C}_K\}$ of MFCC vectors [15], where each vector consists of 25 coefficients and characterizes the spectrum envelope of the signal for the frame; K is the total number of frames.
- Sequence $\{F0_1, \ldots, F0_K\}$ of pitch values.
- Sequence $\{E_1, \ldots, E_K\}$ of energy values.

After that, linguistic and prosodic features for each allophone of all the sentences of the training database are calculated [9, 14].

In the next step, the HMM prototypes for each allophone in the donor speech dataset are created. Each HMM corresponds to a no-skip N-state left-to-right model with N = 5. Each output observation vector $\bar{o}^i$ for the i-th frame consists of 5 streams $\bar{o}^i = [\bar{o}_1^{iT}, \bar{o}_2^{iT}, \bar{o}_3^{iT}, \bar{o}_4^{iT}, \bar{o}_5^{iT}]^T$, as illustrated in Figure 3, where stream 1 is a vector composed by MFCCs, their delta and delta-delta components; stream 2 is a vector composed by F0s; stream 3 is a vector composed by F0 delta components; stream 4 is a vector composed by F0 delta-delta components; and stream 5 is a vector composed by energies, their delta and delta-delta components.

For each k-th HMM the durations of the N states are regarded as a vector $\bar{d}^k = [\bar{d}_1^k, \ldots, \bar{d}_N^k]^T$, where $\bar{d}_n^k$ represents the duration of the n-th state. Furthermore, each duration vector is modeled by an N-dimensional single-mixture Gaussian distribution. The output probabilities of the state duration vectors are thus re-estimated by Baum-Welch iterations in the same way as the output probabilities of the speech parameters [15].

At the final step, the donor intonation model is adapted so as to make it as close as possible to the target speaker voice parameters. The adaptation is performed using the procedure proposed in [16]. During the voice model building, a tree-based clustering technique is applied to the HMM-states of MFCC, F0 and energy values and their delta and delta-delta as well as to the state duration models. In the end of the process, 5N + 1 different acoustic decision trees are generated: N trees for MFCC and their delta and delta-delta components, 3N trees for F0 features, N trees for energy features and one tree for state duration. Performing this stage makes it possible to generate speech parameters for elements absent in the database, which provides intelligible output even under conditions of insufficient training data. Eventually we have the target speaker intonation model which is then used to predict target voice parameters for a synthesized utterance.

2.2. Speech element modification

When the most appropriate speech element sequence is selected, the F0 and duration parameters are adjusted according to the predicted ones. This step is needed to insure the proper intonation of the synthesized sentence. In our system we use the LP model [17] to get the prediction of the residual e[n], modify it by the TD-PSOLA [17], and eventually use the obtained modified prediction of the residual e'[n] to recover the source signal with a new pitch.

The calculation of e[n] is shown in (1):

$$e[n] = s[n] - \bar{a}^T \cdot \bar{s}[n-1], \quad (1)$$

where

$$\bar{s}[n-1] = [s[n-1], s[n-2], \ldots, s[n-P]]^T, \quad (2)$$

$$\bar{a} = [a_1, a_2, \ldots, a_P]^T, \; P = 25. \quad (3)$$

The result is a coefficient vector given by (4):

$$\bar{a}_n = \bar{R}^{-1}[n-1] \cdot \bar{p}[n], \quad (4)$$

where

$$\bar{R}^{-1}[n-1] = \sum_{i=0}^{n} \bar{s}[i-1] \cdot \bar{s}^T[i-1], \quad (5)$$

$$\bar{p}[n] = \sum_{i=0}^{n} \bar{s}[i-1] \cdot s[i], \quad (6)$$

The values of $\bar{p}[n]$ in Equation (4) can be calculated recursively to avoid extra computational load as given by (7):

$$\bar{p}[n] = \bar{s}[n-1] \cdot s[n] + \bar{p}[n-1]. \quad (7)$$

Using the LP coefficients obtained in the analysis cycle, the LP model can be employed in the synthesis cycle with a new excitation signal e'[n] to get the modified signal s'[n] with the desired pitch characteristics, such that:

$$s'[n] = e'[n] + \bar{a}^T \cdot \bar{s}'[n-1]. \quad (8)$$

The LP model is determined for each time sample n, leading to smooth transitions between consecutive models.

Once reliable pitch marks $p_m[n]$ and pitch periods p[n] of the original signal are determined, the pitch contour can be modified as desired. For that purpose, new pitch marks $p'_m[n]$ are determined corresponding to a new pitch period

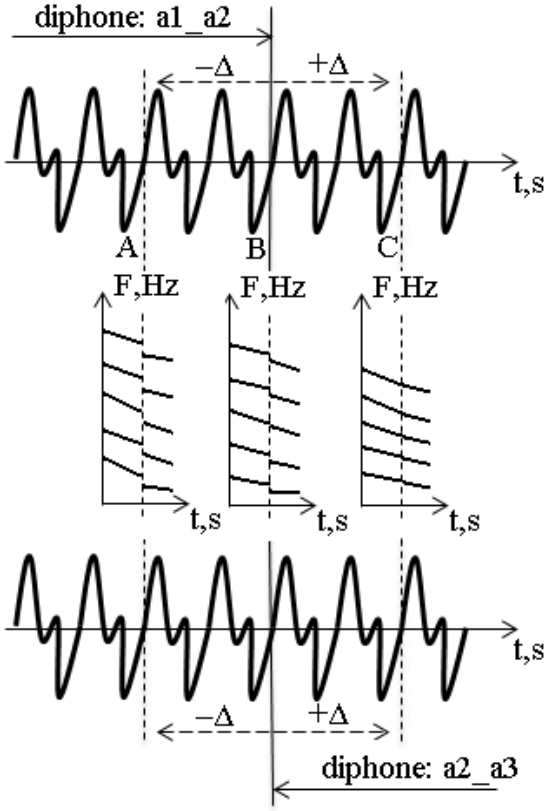

Fig. 4. *Speech element boundary correction*

p'[n], such that $p'[n] = \beta[n] \cdot p[n]$, where β[n] is the pitch period modification factor, which can vary for natural prosody modification, automatic pitch correction and so on. The new pitch marks $p'_m[n]$ are determined by inserting an interval of p'[n] samples between two consecutive marks, so that a pitch mark will be placed at the position n+p'[n] if n has a pitch mark. The next step is to link each new pitch mark $p'_m[n]$ with its corresponding closest peak in the original signal $p_m[n]$. This is done straightforwardly by comparing the time index of $p_m[n]$ and $p'_m[n]$.

In the final step of the new signal generation, each peak in the original signal is segmented by two half-Hanning windows, starting at the preceding pitch mark and ending at the next one. The resulting segments are put together by an overlap and add procedure according to the new pitch period p'[n] obtained previously.

2.3. Speech element concatenation

The last step, when the most appropriate speech elements have been selected and have been adjusted to reliable intonation parameters, is speech element concatenation. The problem is the mismatch of spectrum and pitch components at speech element boundaries. To solve this task we propose the approach detailed in sections 2.3.1 and 2.3.2 respectively.

2.3.1. Speech element boundary correction

Speech element boundary correction is performed to minimize spectrum distortions in the positions of concatenation. The process is illustrated in Figure 4. For example, at the previous step diphones a1_a2 and a2_a3 were selected as the most appropriate. The position B is the original boundary of diphones in the corresponding source files. This position is compared with another two, A and C, which are obtained by shifting B by the offset Δ that is usually two or three F0 periods.

The optimal speech element boundary P_{opt} position is calculated by the Equation (9):

$$P_{opt} = \arg\min_{P=\{A,B,C\}} L_2(P), \quad (9)$$

where

$$L_2(P) = \sqrt{\sum_{i=1}^{M} (c_{Li}(P) - c_{Ri}(P))^2}, \quad (10)$$

$c_{Li}(P)$ is the i-th MFCC coefficient of the left diphone boundary P, $c_{Ri}(P)$ is the i-th MFCC coefficient of the right diphone boundary P, the number of MFCC coefficients M is set by 12.

2.3.2. Pitch smoothing at element boundaries

The main idea of pitch smoothing at speech element boundaries is, on the one hand, to avoid F0 envelope discontinuities and, on the other hand, to keep local F0 fluctuations to make synthesized speech less static and as a result more natural. Let us assume that $\overline{p}_L = \{p_{L1}, p_{L2}, ..., p_{LN}\}$ is the N boundary pitch points of the left speech element and $\overline{p}_R = \{p_{R1}, p_{R2}, ..., p_{RM}\}$ is the M boundary pitch points of the right speech element. They form the mutual pitch envelope $\overline{p} = \{p_{L1}, p_{L2}, ..., p_{LN}, p_{R1}, p_{R2}, ..., p_{RM}\}$ which must be smoothed. The resulting pitch envelope $\overline{p}' = \{p'_1, p'_2, ..., p'_{N+M}\}$ can be calculated as detailed below.

First of all the pitch envelope $\overline{p}$ is represented as the superposition of its filtered part $\overline{p}_m$ and fluctuation part $\overline{p}_f$, where $\overline{p}_m[i] = \alpha \cdot \overline{p}[i] + (1-\alpha) \cdot \overline{p}_m[i-1]$, $\overline{p}_f = \overline{p} - \overline{p}_m$, $0 < \alpha < 1$. Then $\overline{p}_m$ is smoothed based on the Bezier curve as calculated in the Equation (11):

$$\overline{p}'_m[i] = \sum_{j=1}^{N+M} \overline{p}_m[j] \cdot b_{j-1,N+M-1}\left(\frac{i-1}{N+M-1}\right), \quad (11)$$

where $b_{i,n}(t) = C_i^n \cdot t^i \cdot (1-t)^{n-i}$, $C_i^n = \frac{n!}{i! \cdot (n-i)!}$.

The result $\overline{p}'$ is calculated by the Equation (12):

$$\overline{p}' = \overline{p}'_m + \overline{p}_f \,. \quad (12)$$

3. EXPERIMENTAL RESULTS

This section describes some experiments performed with the speech element boundary correction and algorithms of pitch smoothing at element boundaries.

Figure 5 illustrates the results of applying the technique of speech element boundary correction to concatenated diphones; the boundary is marked by the vertical line. The top image shows the spectrum of the concatenation position of diphones. The bottom image shows the spectrum after the application of the correction technique to diphone boundary positions. As we can note from the results, the new spectrum is smoother and looks more natural: there is derivative discontinuity of the spectrum peaks in the top diagram which is absent in the bottom one.

The performance of pitch smoothing at element boundaries is demonstrated in Figure 6. The top diagram shows the original pitch envelope, and the bottom image shows the modified one where the pitch discontinuities (marked by vertical lines) were smoothed while keeping its local fluctuations.

A complex evaluation procedure of the Mean Opinion Score was performed to evaluate the results of this new system under the conditions of different sizes of the available speech database. The assessment procedure consists of two different parts: speech naturalness and speech intelligibility evaluation. The criteria are based on the standard [18] and presented in Table 1 and Table 2 respectively. The assessment results are presented on Figure 7 and Figure 8 respectively.

Table 1. *Criteria for speech naturalness evaluation*

Speech characteristics	Rates
Natural-sounding speech, some subtle distortion present. Wheeze, rattle missing. High recognizability	> 4.5
Some violation of naturalness and recognizability, a weak presence of one type of distortion (burr, twang, wheeze, rattle, etc.)	3.6 – 4.5
Audible violation of naturalness and recognizability, presence of several types of distortion (burr, twang, wheeze, rattle, etc.)	2.6 – 3.5
Constant presence of distortions (burr, twang, wheeze, rattle, etc.). A significant violation of naturalness and recognizability	1.7 – 2.5
Strong mechanical distortion: burr, twang, wheeze, rattle, etc., mechanical voice. A significant loss of naturalness and recognizability is observed	< 1.7

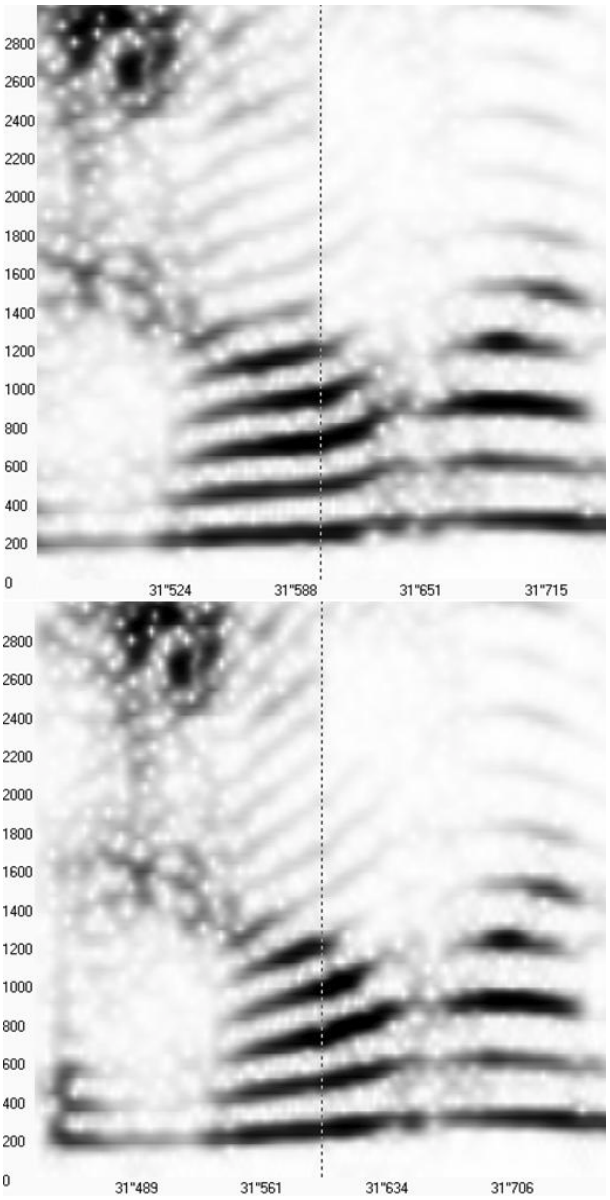

Fig. 5. *Fragments of spectrum in the position of speech element concatenation: top - original speech elements, bottom - speech elements with corrected boundaries*

Table 2. *Criteria for speech intelligibility evaluation*

Speech characteristics	Rates
Absolutely intelligible speech	5
Intelligible speech, understanding without difficulties	4.6 – 4.9
Intelligible speech, understanding with small difficulties	3.6 – 4.5
Almost intelligible speech, understanding with difficulties	2.6 – 3.5
Partly intelligible speech, understanding with huge difficulties	< 2.5

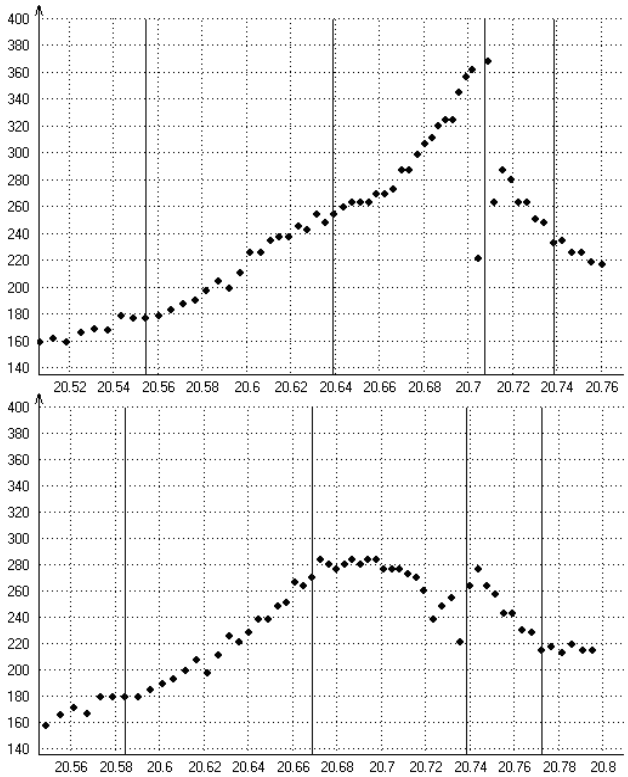

Fig. 6. *Fragments of pitch envelope: top - original envelope, bottom - smoothed envelope*

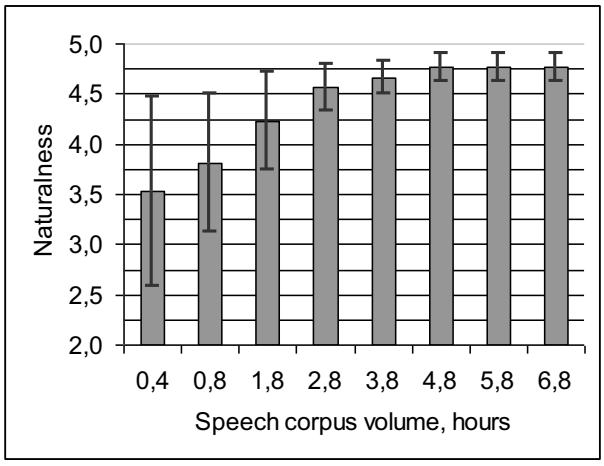

Fig. 7. *Speech naturalness evaluation results*

4. CONCLUSIONS

In this paper, we presented a description of a system for high-quality speech synthesis under the conditions of a small speech dataset. Our main aim was to solve the problem of voice restoration as well as voice creation when available speech data is strongly limited. It can be performed by using a hybrid approach (HMM-based plus Unit Selection techniques) where the intonation of the target speaker is modeled based on another speaker's database, adaptation techniques are applied to emulate voice characteristics with maximum precision, and the selected optimal elements are adjusted to predicted parameters by special modification and concatenation methods described in the paper. Experiments and subjective expert evaluation results demonstrate that high-quality speech synthesis can be achieved even with a small speech database. Moreover, the proposed approach reduces requirements for accuracy of database labeling thanks to spectrum adjustment, and improves synthesized speech quality in general.

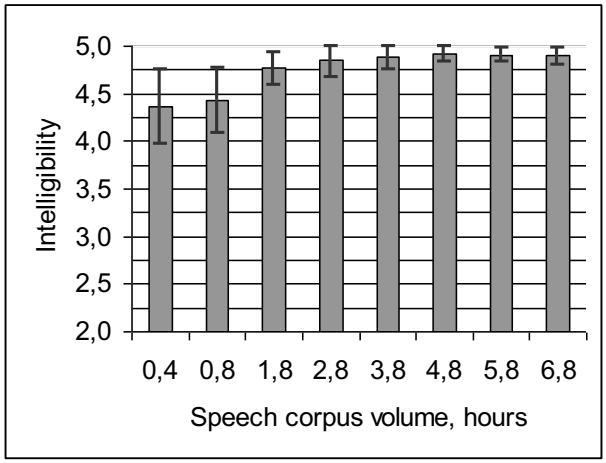

Fig. 8. *Speech intelligibility evaluation results*

5. ACKNOWLEDGEMENT

This work was partially financially supported by Government of Russian Federation, Grant 074-U01.

6. REFERENCES

[1] S. Breuer, S. Bergmann, R. Dragon, S. Möller "Set-up of a Unit-Selection Synthesis with a Prominent Voice", in Proceedings of the 5-th International conference on Language Resources and Evaluation, 2006.

[2] J. Matoušek, D. Tihelka, L. Šmídl "On the Impact of Annotation Errors on Unit-Selection Speech Synthesis", Text, Speech and Dialogue; Lecture Notes in Computer Science, Vol. 7499, pp. 456-463, 2012.

[3] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker independent HMM-based speech synthesis system - hts-2007 system for the blizzard challenge 2007", in Proceedings of the Blizzard Challenge 2007, 2007.

[4] A.W. Black, A.J. Hunt "Unit Selection in a Concatenative Speech Synthesis Using a Large Speech Database", in Proceedings of ICASSP 96, Atlanta, Georgia, Vol. 1, pp. 373-376, 1996.

[5] M. Chi Luong, M. Akagi “A concatenative speech synthesis for monosyllabic languages with limited data”, Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), Asia-Pacific, pp. 1-10, 2012.

[6] M. Fanbo, W. Zhiyong, M. Helen, J. Jia, C. Lianhong “Hierarchical English Emphatic Speech Synthesis Based on HMM with Limited Training Data”, in Proceedings of InterSpeech, ISCA, 2012.

[7] T. Ryosuke, Z. Heiga, T. Keiichi, K. Tadashi, B. Murtaza, N. Shrikanth “Constructing emotional speech synthesizers with limited speech database”, in Proceedings of InterSpeech, Korea, pp. 1185-1188, 2004.

[8] P. Trung-Nghia, L. Chi Mai, A. Masato “A Hybrid TTS between Unit Selection and HMM-based TTS under limited data conditions”, 8th ISCA Speech Synthesis Workshop, Spain, pp. 279-284, 2013.

[9] P. Chistikov, E. Korolkov, A. Talanov “Combining HMM and Unit Selection technologies to increase naturalness of synthesized speech”, Computational Linguistics and Intellectual Technologies, № 12 (19), vol. 2, pp. 607-615, 2013.

[10] A. Prodan, P. Chistikov, A. Talanov “Voice building system for Russian TTS system “Vital Voice”, in Proceedings of the Dialogue-2010 International Conference, № 9 (16), pp. 394-399, 2010.

[11] N. Smirnova, P. Chistikov “Software for Automated Statistical Analysis of Phonetic Units Frequency in Russian Texts and its Application for Speech Technology Tasks”, in Proceedings of the Dialogue-2011 International Conference, № 10 (17), pp. 632-643, 2011.

[12] P. Chistikov, O. Khomitsevich “On-line automatic sentence boundary detection in a Russian ASR system”, SPECOM 2011 International Conference, pp. 112-117, 2011.

[13] P. Chistikov, E. Korolkov “Data-driven Speech Parameter Generation For Russian Text-to-Speech System”, in Proceedings of the Dialogue-2012 International Conference, № 11 (18), pp. 103-111, 2012.

[14] P. Chistikov, O. Khomitsevich “Improving prosodic break detection in a Russian TTS system”, in Proceedings of the 15th International Conference SPECOM 2013, pp. 181–188, 2013.

[15] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura “Hidden semi-Markov model based speech synthesis”, in Proceedings of the International Conference on Spoken Language Processing (ICSLP), pp. 1393-1396, 2004.

[16] J. Yamagishi, T. Kobayashi “Adaptive training for hidden semi-Markov model”, in Proceedings of the ICASSP 2005, pp. 365–368, 2005.

[17] P. Taylor “Text-to-Speech Synthesis”, Cambridge University Press, United Kingdom, 2008.

[18] State standard specification 50840-95 “Speech transmission through communication channels. Methods for quality, intelligibility and recognizability evaluation”. Moscow, 1995.