DATA-DRIVEN SPEECH PARAMETER GENERATION FOR RUSSIAN TEXT-TO-SPEECH SYSTEM

Chistikov P. G. (chistikov@speechpro.com),
Korolkov E. A. (korolkov@speechpro.com)

Speech Technology Center ltd, St. Petersburg, Russia

We propose a speech parameter generation approach for Russian based on hidden Markov models. The speech parameter sequence is generated from HMMs whose observation vectors contain speech characteristics. As a baseline we use the spectrum represented by mel-frequency cepstral coefficients (MFCC), pitch and duration parameters. All of them can be easily complemented by any other parameters, improving the quality. For the creation of the voice model we use linguistic and prosodic features which are the observations of every allophone in the utterance. This paper also presents the results of research into selecting the most effective features to characterize an allophone. Experimental results show that Russian speech can be successfully parameterized and an arbitrary utterance can be synthesized from the generated parameters.

Key words: speech processing, speech synthesis, text-to-speech system, tree based clustering, speech parameterization

1. Introduction

Voice analysis and synthesis have been vastly studied in recent years. Many applications and methods have been developed in these areas. The increasing availability of large speech databases makes it possible to construct speech synthesis systems by means of a data-driven or corpus-based approach, by applying statistical learning algorithms. These systems can be automatically trained and are able not only to produce natural and high-quality synthetic speech but also, importantly, to imitate voice characteristics of the original speaker.

For constructing such a system the use of hidden Markov models (HMMs) is widely employed [1–3]. HMMs have been successfully applied to model the sequence of speech spectra in speech recognition systems. The quality of HMM-based speech recognition systems has been improved by techniques which utilize the flexibility of HMMs. These are mixtures of Gaussian densities, tying mechanism, context-dependent modeling dynamic feature parameters and speaker and environment adaptation methods. By applying these techniques to a TTS system we can also support the contemporary tendency towards the synthesis of voices with different styles and emotions [4–5].

In this research we apply the HMM-based approach to speech parameterization to Russian. In our work we use spectrum, pitch and duration as baseline parameters which can be easily complemented by dynamic features and any other features to improve the quality of synthesized speech. It is worth noting that the voice model can be created with no need for large databases [6–8].

The paper is organized as follows: in Section 2 all the procedures that are carried out by the TTS system are described, from the source database to the synthesized speech. Section 3 describes the voice model building methods to synthesize a given utterance. In Section 4, linguistic and prosodic features are presented. Section 5 deals with the system implementation and experimental results. The conclusions are given in Section 6.

2. TTS Engine description

The speech generation mechanism is a subpart of the fully-functional text-to-speech system which has been constructed on the basis of the Vital Voice [9] system and HTK [10]. The procedures of voice model building and synthesis of an arbitrary utterance by the speech synthesis engine are shown in Figure 1.

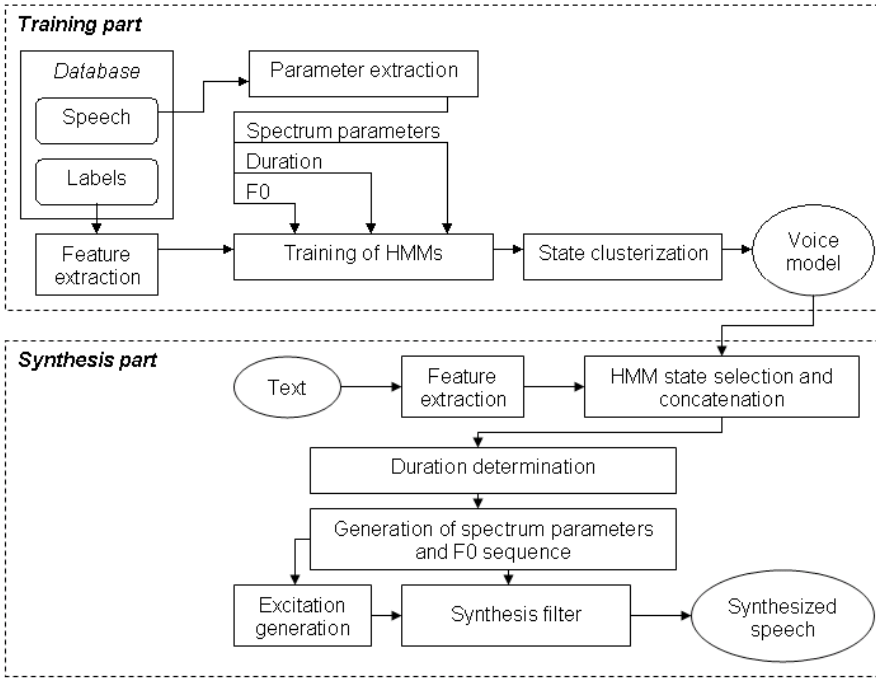

Fig. 1. Diagram illustrating the basic steps conducted by the speech synthesis engine

2.1. Training

The training starts with parameter extraction. For each allophone from the database, linguistic and prosodic features and voice parameters are calculated. Then context-dependent HMMs are trained. In the last step, HMM states are clustered based on linguistic and prosodic features. Eventually we have the voice model. Section 3 gives these steps in detail.

2.2. Synthesis

The synthesis procedure starts with transcribing the utterance to an allophone sequence and generating the linguistic and prosodic features which are eventually used to select corresponding leaves from each of the 2N+1 (where N is the number of states in HMM) decision trees generated by the context-clustering procedure in the training step. At the end of this stage, three logical HMM sequences, whose states correspond to the selected leaves, are obtained.

The three above-mentioned HMM sequences are used to derive spectrum parameters, fundamental frequencies and state durations. Then excitation is generated based on F0 and energy, and synthesized speech is extracted by the synthesis filter based on excitation and spectra [11–13].

3. Voice model building

Initially a sequence of fundamental frequencies $\{F0^1, \ldots, F0^K\}$, including voicing decision information (if F0 is 0 then the frame is considered unvoiced), where K is the total number of frames of all sentences from the training database, is extracted on a short-term basis. Simultaneously, a sequence of spectrum parameter (mel-cepstral coefficients [14]) vectors which represent speech envelope spectra, $\{c^1, \ldots, c^K\}$, is obtained. Each MFCC vector $\overline{c}^i=[c_0^i \ldots c_M^i]$, where the i indicates the frame number and $[\cdot]^T$ means transposition, is derived through an M-th order mel-cepstral analysis.

After that, linguistic and prosodic features for each allophone of all the sentences of the training database are estimated. The description of linguistic and prosodic features for Russian is given in Section 4.

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

The observation vector $\overline{o}^i$ is the output of an HMM state n according to a probability given by

$$\beta_n(\overline{o}^i) = \prod_{j=1}^{2}[\sum_{l=1}^{R_j} \omega_{njl} \mathrm{N}(\overline{o}_j^i; \mu_{njl}, \Sigma_{njl})], \tag{1}$$

where $\overline{o}_j^i$ means a Gaussian distribution with mean vector μ and covariance matrix Σ, ω_{njl} is the weight for the l-th mixture component of the j-th stream of vector $\overline{o}_j^i$ (output of the state n) with R_j being the corresponding number of mixture components. The stream vectors are modelled by single-mixture continuous Gaussian distributions where the dimensionality is 3(M+1) for $\overline{o}_1^i$ and 3 for $\overline{o}_2^i$.

For each k-th HMM the durations of the N states are considered as vectors $\overline{d}^k=[\overline{d}_1^k,...,\overline{d}_N^k]^T$, where $\overline{d}_n^k$ represents the duration of the n-th state. Furthermore, each duration vector is modelled 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].

During the voice model building, a tree-based clustering technique is applied to the HMM-states of MFCC and F0 values, as well as to the state duration models. In the end of the process, 2N+1 different acoustic decision trees are generated: N trees for mel-cepstral coefficients, N trees for F0 features and finally one tree for state duration.

4. Contextual features

When speech synthesis systems are constructed, some parameters are necessary to provide a natural rendering of the prosody. These parameters might include context-dependent items, for instance preceding/following phone, syllable, word, sentence etc [17].

The determination of contextual features for a particular language is based on linguistic and prosodic parameters. Apart from this theoretical approach, empirical analysis can also be implemented in order to tune the features by extending the factors that are important and eliminating the ones which are not [18].

The contextual features listed in Table 1, which were selected as the most informative ones to build the voice model, were first obtained from those used in HMM-based Brazilian Portuguese speech synthesis [18] and eventually adjusted through theoretical and empirical methods to Russian.

5. Implementation

5.1. The corpus

In this work we used the speech corpora (one male and one female voices) initially prepared for the Vital Voice TTS system [9]. They correspond to six hours of speech excluding silence regions. The databases were recorded at a sample rate of 22 050 Hz with 16 bits per sample.

The phonetic labeling of the database was carried out using the phone set described in [16]. Time label boundaries were obtained manually. Further, syllable and word labeling were also manually conducted for each sentence.

Table 1. Contextual features

Allophone features	
Phone before previous	Phone after next
Previous phone	Phone position from the beginning of the syllable
Current phone	Phone position from the end of the syllable
Next phone	
Syllable features	
Previous syllable	Syllable position from the end of the word
Current syllable	Syllable position from the beginning of the sentence
Next syllable	Syllable position from the end of the sentence
Number of phones in the previous syllable	Number of stressed syllables before current syllable in the sentence
Number of phones in the current syllable	Number of stressed syllables after current syllable in the sentence
Number of phones in the next syllable	Vowel name in the current syllable
Syllable position from the beginning of the word	
Word features	
Part of speech of the previous word	Number of syllables in the current word
Part of speech of the current word	Number of syllables in the next word
Part of speech of the next word	Word position from the beginning of the sentence
Number of syllables in the previous word	Word position from the end of the sentence
Sentence features	
Number of syllables in the current sentence	Sentence type (declarative, exclamation etc.)
Number of words in the current sentence	

5.2. Parameter extraction

Fundamental frequencies and mel-cepstral coefficients were extracted from the speech corpus in every 10ms frame. MFCCs were obtained through a 25-th order analysis (M=25) by means of a 25ms Henning window.

5.3. Generated decision trees

Figures 2 and 3 show the top part of the decision tree for mel-cepstral coefficients of [a1] phoneme, and state durations, respectively. By observing Figure 2 and Figure 3, and assuming that top nodes are more important when selecting the parameter which is being clustered, we can notice that the information about the syllable, word and sentence is more crucial for state duration. On the other hand, questions related to phonemes are more significant to the tree of mel-cepstral coefficients.

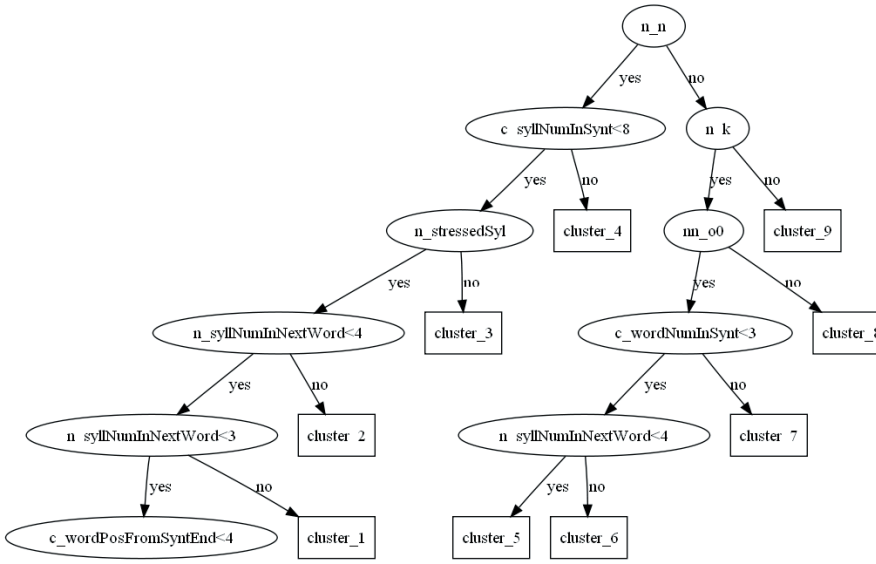

Fig. 2. Top of the decision-tree constructed to cluster the third HMM state for mel-cepstral coefficients of [a1] phoneme

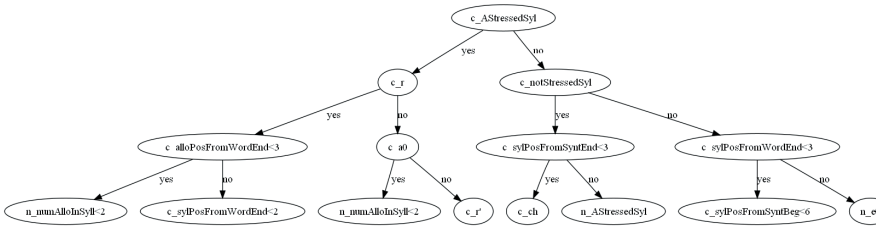

Fig. 3. Top of the decision-tree constructed to cluster the distribution of state durations

5.4. Example of speech synthesis

Figure 4 presents the spectrograms for the natural sentence “Это моя сестра” (this is my sister) and its synthesized version. It should be noted that the utterance is not part of the training database.

Aside from the reproduction of phones it can be also observed from Figure 4 that the synthesized version presents a speaking rate similar to that of the natural speech. This shows an important characteristic of the HMM-based speech synthesis approach: the ability to imitate the prosody of the speech corpus which was used to build the voice model.

Although it has been reported that even with a small database it is possible to synthesize speech, lack of data strongly affects the quality of synthesized speech. Once the HMMs do not properly reflect the characteristics of some linguistic and prosodic feature sets, inconsistent parameters may be generated during synthesizing, and as a result we could have poor quality speech. Thus the process of preparing the database is very important for building appropriate voice models.

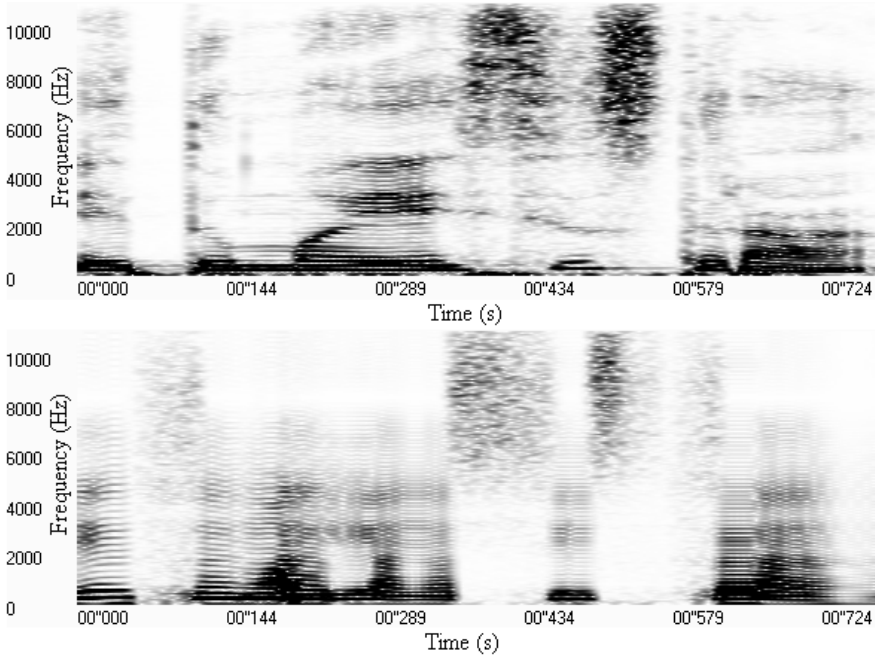

Fig. 4. Spectrograms for the natural sentence “Это моя сестра” (this is my sister) (top) and its synthesized version (bottom)

6. Conclusions

This paper described an approach for corpus-based speech parameter generation for a Russian text-to-speech system. The engine is based on a method where the speech parameters are obtained from HMMs whose observation vectors consist of spectum, F0 and duration features. We performed context-clustering to achieve a greater flexibility of the algorithm and to make it possible to use the voice model even when a small database is used. The clusterization procedure is based on linguistic and prosodic features of Russian which were also presented in this paper. Experimental results show that Russian speech can be successfully parameterized and any utterance can be synthesized from the generated parameters.

References

1. *Black A.* Unit selection and emotional speech. Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH). 2003.
2. *Black A., Taylor P., Caley R.* The Festival Speech Synthesis System, http://www.festvox.org/festival.
3. *Campbell N.* Towards synthesizing expressive speech; designing and collective expressive speech data. Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH). 2003.
4. *Fukada T., Tokuda K., Kobayashi T., Imai S.* An adaptive algorithm for mel-cepstral analysis of speech. Proceedings ICASSP-92. 1992, pp. 137–140.
5. *Fukada T., Tokuda K., Kobayashi T., Imai S.* An adaptive algorithm for mel-cepstral analysis of speech. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 1992.
6. *Imai S., Sumita K., Furuichi C.* (1983) Mel log spectrum approximation (MLSA) filter for speech synthesis. IECE Trans. A, Vol. J66-A, no.2, pp.122–129.
7. *Korolkov E, Glavatskih I, Talanov A.* Vital Russian voice text-to-speech system based on unit selection method. Proceedings of the 36th International Philological Conference on formal methods for the analysis of the Russian language. Russia, 2008.
8. *Maia R., Zen H., Tokuda K.* An HMM-based Brazilian Portuguese Speech Synthesis and Its Characteristics. Revista da Sociedade Brasileira de Telecomunicações. 2006.
9. *Maia R., Zen H., Tokuda K., Kitamura T., Resende F.* Towards the development of a Brazilian Portuguese text-to-speech system based on HMM. Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH). Geneva, 2003, pp. 2465–2468.
10. *Shichiri K., Sawabe A., Yoshimura T., Tokuda K., Masuko T., Kobayashi T., Kitamura T.* Eigenvoices for HMM-based speech synthesis. Proceeding of the International Conference on Spoken Language Processing (ICSLP). 2002.
11. *Smirnova N., Chistikov P.* Software for Automated Statistical Analysis of Phonetic Units Frequency in Russian Texts and its Application for Speech Technology Tasks. Proceedings of the Dialogue-2011 International Conference. 2011.

12. *Tokuda K., Kobayashi T., Fukada T., Saito H., Imai S.* (1991) Spectral estimation of speech based on mel-cepstral representation. IEICE Trans. A, Vol. J74-A, no.8, pp.1240–1248.
13. *Tokuda K., Zen H., Black A.* An HMM-based speech synthesis applied to English. Proceedings of IEEE Workshop in Speech Synthesis. 2002.
14. *Yamagishi J., Tachibana M., Masuko T., Kobayashi T.* Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis. Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 2004.
15. *Yoshimura T., Tokuda K., Masuko T., Kobayashi T., Kitamura T.* Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH). 1999.
16. *Yoshimura T., Tokuda K., Masuko T., Kobayashi T., Kitamura T.* Speaker interpolation in HMM-based speech synthesis. Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH). 1997.
17. *Young S., Evermann G., Gales M.* (2006) The HTK Book (for HTK Version 3.4).
18. *Zen H., Tokuda K., Masuko T., Kobayashi T., Kitamura T.* Hidden semi-Markov model based speech synthesis. Proceedings of the International Conference on Spoken Language Processing (ICSLP). 2004.