COMBINING HMM AND UNIT SELECTION TECHNOLOGIES TO INCREASE NATURALNESS OF SYNTHESIZED SPEECH

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

Speech Technology Center ltd, St. Petersburg, Russia

We propose a text-to-speech system based on the two most popular approaches: statistical speech synthesis (based on hidden Markov models) and concatenative speech synthesis (based on Unit Selection). TTS systems based on Unit Selection generate speech that is quite natural but highly variable in quality. On the other hand, statistical parametric systems produce speech with much more consistent quality but reduced naturalness due to their vocoding nature. Combining both approaches improves the overall naturalness of synthesized speech. To reduce variability of Unit Selection results, we calculate a statistical generalization of the speaker's intonation. We created a methodology of voice model building in order to solve the task of speech parameterization. The model is a set of HMM models whose state parameters are clustered to provide good quality of synthesized speech even under conditions of insufficient training data. MFCC coefficients, pitch, energy and duration values are used as fundamental features. Objective and subjective experiments show that our method increases the naturalness of synthesized speech.

Key words: speech processing, speech synthesis, text-to-speech system, hidden Markov model, unit selection, voice model

1. Introduction

Speech synthesis (text-to-speech, TTS) is a process of transforming the character sequence of any text to a sequence of speech samples [1–3]. There are several approaches to doing this. The basic approaches are the following: rule-based speech synthesis (formant synthesis), articulatory speech synthesis, concatenative speech synthesis, and speech synthesis based on statistical models [4–8].

Currently the most popular approaches are the following: the Unit Selection algorithm (speech element selection) and statistical models (HMM TTS). The first one makes it possible to synthesize speech with maximum naturalness, given an accurately segmented voice database of a large size (10 hours and more). On the other hand, the second approach, which produces synthesized speech that is less natural, has the advantages presented below.

1. The HMM-based method provides an easy way to modify voice characteristics by using speaker adaptation/interpolation techniques. The Unit Selection algorithm generates speech with a constant style that is the same as the style of the speech in the database.
2. Speech generated by the HMM method is less natural for listeners. However, it is smoother, without detectable phone boundaries (pitch or energy leaps) which are usual for concatenative synthesis. In addition, the quality of Unit Selection TTS can be strongly reduced when some of the necessary speech elements are absent in the database. When voice models are used, absent speech elements are synthesized based on mean values which are closest to the required ones. It is possible due to tree-based context clustering, and the method provides good intelligibility when the amount of contexts is insufficient.
3. Applying the HMM-based speech synthesis method makes it possible to create a new TTS voice in much less time and to reduce the memory size required for storing the voice data.

We propose a hybrid TTS system that combines both approaches: looking for a matching sequence of speech elements in the speaker's speech database by means of the classic Unit selection algorithm, and employing a statistical intonation model which was trained on the same database. Experiments show that the naturalness of synthesized speech is increased compared to systems based only on Unit Selection or hidden Markov models.

2. System description

Structurally, the system is divided in two parts (Figure 1): the training part (the preparation stage) and the synthesis part. A speech database is created based on the speech corpus containing 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) [9–12]. Then the speech database is indexed to provide fast search for target elements by the following features: phone name, names of phones before and after the current phone, mel-frequency cepstral coefficients (MFCC) at phone boundaries, energy, pitch, and phone duration.

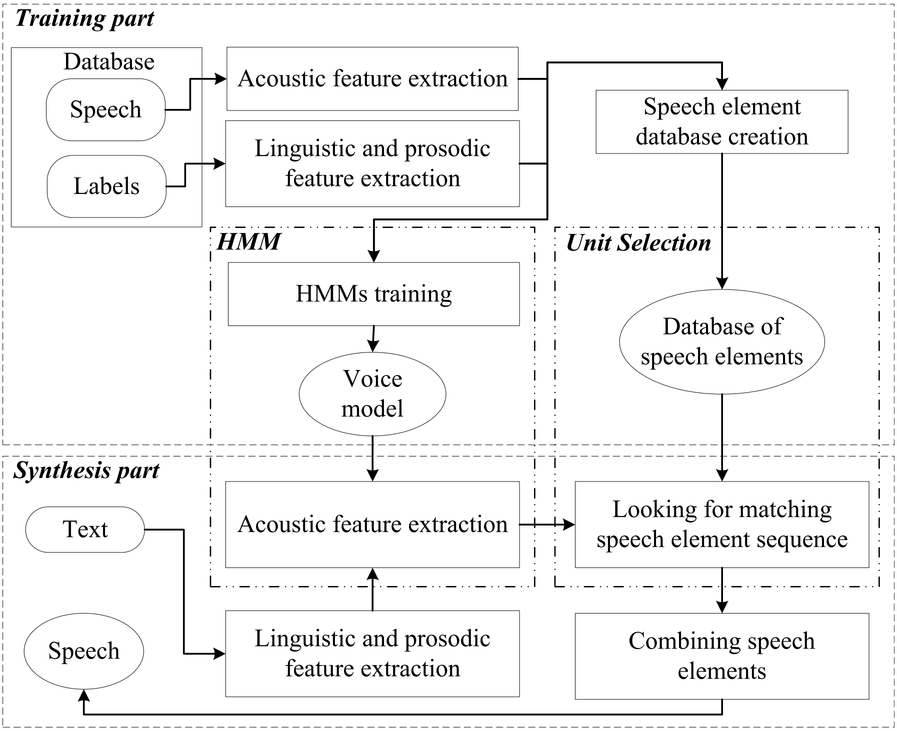
Training part
Database
Speech
Labels
Acoustic feature extraction
Linguistic and prosodic feature extraction
Speech element database creation
HMM
HMMs training
Voice model
Unit Selection
Database of speech elements
Synthesis part
Text
Speech
Acoustic feature extraction
Looking for matching speech element sequence
Linguistic and prosodic feature extraction
Combining speech elements

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

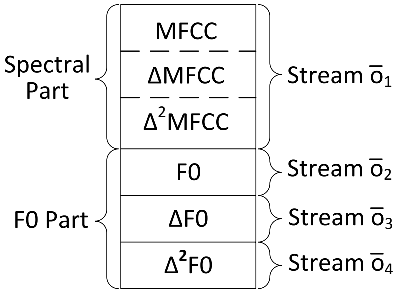

Fig. 2. Observation vector

The procedure of voice parameter modeling begins with the extraction of the feature set for all sound files [13, 14]. Each member of the set represents a short part of the signal (frame) with the length of 25 ms. The features contain the following parameters:

- Sequence $\{C_1, \ldots, C_K\}$ 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.

After that, linguistic and prosodic features for each allophone of all the sentences of the training database are calculated. The description of the linguistic and prosodic features is presented in Table 1.

In the next step, the 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 4 streams, $\overline{o}^{i} = [\overline{o}_1^{iT}, \overline{o}_2^{iT}, \overline{o}_3^{iT}, \overline{o}_4^{iT}]^T$ as illustrated in Figure 2, 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; and stream 4 is a vector composed by F0 delta-delta components.

For each k-th HMM the durations of the N states are considered as a vector $\overline{d}^{k} = [\overline{d}_1^{k}, \ldots, \overline{d}_N^{k}]^T$, 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 [16].

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	End punctuation type (comma, full stop, etc.)
Number of words in the current sentence	

During the voice model building, a tree-based clustering technique is applied to the HMM-states of MFCC and their delta and delta-delta components, F0 values and their delta and delta-delta components, as well as to the state duration models. In the end of the process, 4N + 1 different acoustic decision trees are generated: N trees for MFCC and their delta and delta-delta components, 3N trees for F0 features, and one tree for state duration (Figure 3). 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.

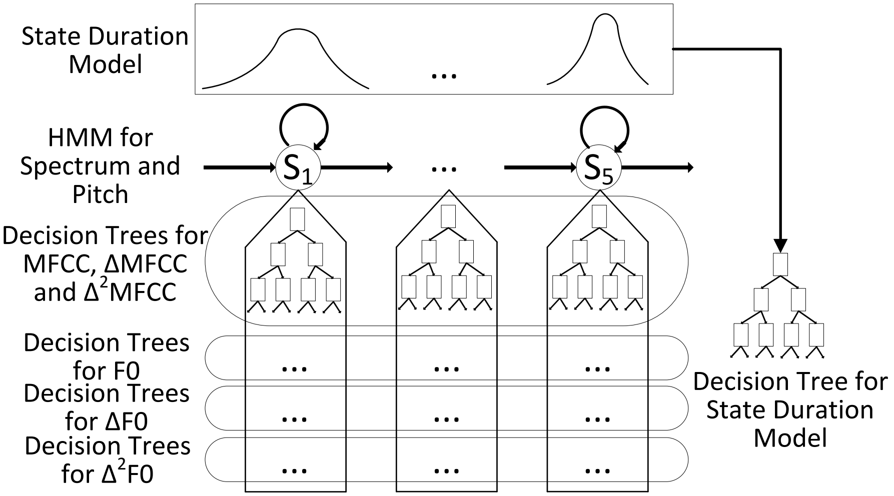
State Duration Model
...
HMM for Spectrum and Pitch
S1
...
S5
Decision Trees for MFCC, ΔMFCC and Δ2MFCC
Decision Trees for F0
Decision Trees for ΔF0
Decision Trees for Δ2F0
...
Decision Tree for State Duration Model

Fig. 3. Voice model

Text-to-speech system input is a 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. Special metrics (target cost and concatenation cost) are used to estimate the suitability of each selected allophone [17].

Target cost estimation is given in equation (1):

$$C^t\left(u_i,t_i\right)=\sum_{k=1}^{p} w_k^t C_k^t\left(u_i,t_i\right) \quad (1),$$

where u_i is an element from the database; t_i is the target element; C_k^t is a distance between k-th features of elements; w_k^t is the weight of the k-th feature.

In other words, target cost is the weighted sum of differences in features between the target element and an element from the database. Any suitable linguistic and prosodic characteristics can be used as features. Usually the following information is used: pitch, duration, context, position in the syllable, position in the word, number of stressed syllables in the utterance, etc.

Selected elements should be not only close to the targets, they should also concatenate well with each other. Concatenation cost is defined as the weighted sum of differences in features between two successive selected elements:

$$C^c\left(u_{i-1},u_i\right)=\sum_{k=2}^{q} w_k^c C_k^c\left(u_{i-1},u_i\right) \quad (2),$$

where u_{i-1} is the previous element; u_i is the current element; C_k^c is the distance between k-th features of elements; w_k^c is the weight of the k-th feature.

The final cost for the whole sequence of n elements is the sum of the target cost and the concatenation cost:

$$C\left(u,t\right)=\sum_{i=1}^{n} C^t\left(u_i,t_i\right)+\sum_{i=2}^{n} C^c\left(u_{i-1},u_i\right) \quad (3).$$

The purpose of the Unit Selection algorithm is to select a sequence of elements that minimizes the final cost equation (3).

In the final step, the selected sequence of elements is concatenated to form the speech signal which is the result of TTS system work.

3. Experiments

Figures 4–6 present the results of the system's work. They are oscillograms, spectrograms, and pitch envelopes for the utterance "это очень важно!" ("eto očen' važno", Russian for "it is very important!"). A natural phrase is at the top of each figure, and its synthesized equivalent is at the bottom. It should be mentioned that this phrase had been excluded from the training data set.

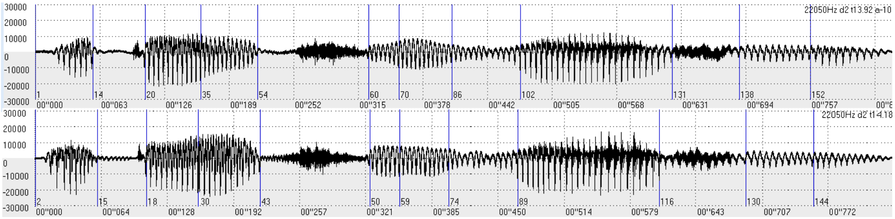

Fig. 4. Oscillograms for the natural sentence *"это очень важно!"* ("it is very important") (top) and its synthesized version (bottom)

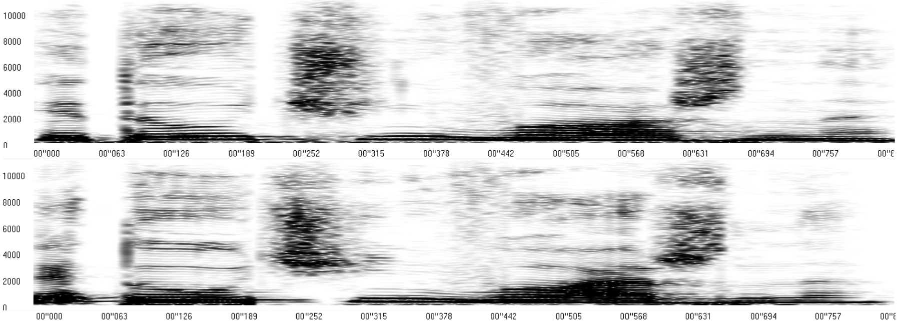

Fig. 5. Spectrograms for the natural sentence *"это очень важно!"* ("it is very important") (top) and its synthesized version (bottom)

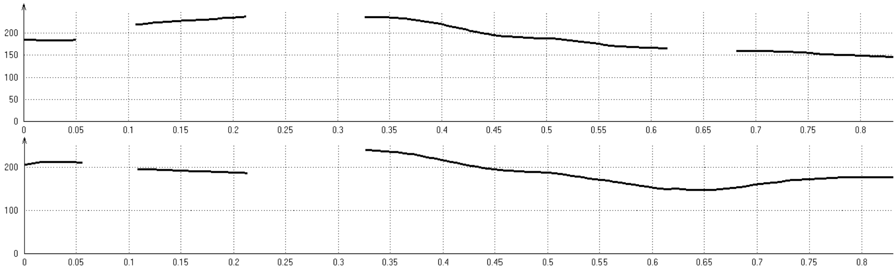

Fig. 6. Pitch envelopes for the natural sentence *"это очень важно!"* ("it is very important") (top) and its synthesized version (bottom)

From the figures above you can notice that the synthesized utterance has almost the same tempo and spectrum characteristics as the natural equivalent uttered by a real speaker. It is due to the modeling of parameters based on hidden Markov models.

We conducted a MOS (mean opinion score) evaluation to estimate the naturalness of the synthesized speech. Table 2 presents the results of the comparison for two systems: the proposed hybrid system and the system based on Unit Selection only. The comparison was performed by five experts for two voices (one male and one female); the results in the table have been averaged. The values ranged from 0 (unnatural, "mechanical" speech) to 5 (completely natural speech). The synthesized sentences were also compared to the same sentences pronounced by the speaker (they were not included in the training data set). The results show that the hybrid TTS approach increases the naturalness of synthesized speech.

Table 2. Comparison of the proposed system and the Unit Selection system

Type of TTS		Natural speech
Unit Selection	Hybrid	
4,0	4,3	4,8

4. Conclusions

This paper describes an approach for building a Russian TTS system based on the integration of hidden Markov models and Unit Selection. The TTS engine is based on a method where the speech parameters are obtained from HMMs whose observation vectors consist of MFCC, pitch and duration features; the speech signal is generated by a Unit Selection algorithm using the obtained speech parameters. We developed a voice model creation method for constructing a natural intonation contour. The experimental results confirm the improved quality of synthesized speech. It is also worth noting that the final speech quality can be improved by tuning Unit Selection weights and optimizing the training feature set.

References

1. *Dines J.* (2003), Model based trainable speech synthesis and its applications, Ph. D. Thesis, Queensland University of Technology, Brisbame, Australia.
2. *Dutoit Th.* (2002), Introduction au traitement de la parole, Faculte Polytechnique de Mons.
3. *Stilianou Y.* (1996), Harmonic plus noise models for speech, combined with statisticals methods, for speech and speaker modification, Ph.D. Thesis, Ecole Nationale Superieure des Telecommunications, Paris, France.
4. *Klatt D. H.* (1987), Review of text-to-speech conversion for English, Journal of the Acoustical Society of America, Vol. 82, pp. 737–793.
5. *Tokuda K.* (2011), HMM-based Speech Synthesis System (HTS), available at: http://hts.sp.nitech.ac.jp.
6. *Huang X., Acero A., Adcock J., Goldsmith J., Liu J., Whistler A.* (1996), Trainable Text-to-Speech System, Proc. of the International Conference on Spoken Language Processing, Philadelphia, PA, Vol. 4, pp. 2387–2390.
7. *Donovan R. E., Eide E. M.* (1998), The IBM Trainable Speech Synthsis System, Proc. ICSLP'98, Sydney, Australia.
8. *Donovan R. E., Ittycheriah A., Franz M., Ramabhadran B., Eide E., Viswanathan M., Bakis R., Hamza W.* (2001), Current Status of the IBM Trainable Speech Synthesis System, Proc. 4th ESCA Tutorial and ResearchWorkshop on Speech Synthesis, Atholl place Hotel, Scotland, UK.
9. *Prodan A., Chistikov P., Talanov A.* (2010), Voice building system for Russian TTS system "Vital Voice", Proceedings of the Dialogue-2010 International Conference, № 9 (16), pp. 394-399.
10. *Smirnova N., Chistikov P.* (2011), 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, № 10 (17), pp. 632–643.
11. *Chistikov P., Khomitsevich O.* (2011), On-line automatic sentence boundary detection in a Russian ASR system, Vestnik MGTU. Priborostroenie, Special Issue "Biometric Technologies, pp. 115–123.

12. *Chistikov P., Khomitsevich O.* (2011), On-line automatic sentence boundary detection in a Russian ASR system, SPECOM 2011 International Conference, pp. 112–117.
13. *Chistikov P.* (2012), Speech parameter modeling at Russian Text-to-Speech system, Proceedings of the 1st All-Russian researcher congress, № 2, Editor-in-chief PhD, prof. V. O. Nikiforov, SPb: ITMO, pp. 227–228.
14. *Chistikov P., Korolkov E.* (2012), Data-driven Speech Parameter Generation For Russian Text-to-Speech System, Proceedings of the Dialogue-2012 International Conference, № 11 (18), pp. 103–111.
15. *Fukada T., Tokuda K., Kobayashi T., Imai S.* (1992), An adaptive algorithm for mel-cepstral analysis of speech, Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 137–140.
16. *Zen H., Tokuda K., Masuko T., Kobayashi T., Kitamura T.* (2004), Hidden semi-Markov model based speech synthesis, Proc. of the International Conference on Spoken Language Processing (ICSLP), pp. 1393–1396.
17. *Black A. W., Hunt A. J.* (1996), Unit Selection in a Concatenative Speech Synthesis Using a Large Speech Database, In Proceedings of ICASSP 96, Atlanta, Georgia, Vol. 1, pp. 373–376.