EVALUATION OF NATURALNESS OF SYNTHESIZED SPEECH WITH DIFFERENT PROSODIC MODELS

Solomennik A. I. (solomennik-a@speechpro.com)
Speech Technology Ltd., Minsk, Belarus

Chistikov P. G. (chistikov@speechpro.com)
Speech Technology Center Ltd, St. Petersburg, Russia

Obtaining natural synthesized speech is the main goal of modern research in the field of speech synthesis. It strongly depends on the prosody model used in the text-to-speech (TTS) system. This paper deals with speech synthesis evaluation with respect to the prosodic model used. Our Russian VitalVoice TTS is a unit selection concatenative system. We describe two approaches to prosody prediction used in VitalVoice Russian TTS. These are a rule-based approach and a hidden Markov model (HMM) based hybrid approach. We conduct an experiment for evaluating the naturalness of synthesized speech. Four variants of synthesized speech depending on the applied approach and the speech corpus size were tested. We also included natural speech samples into the test. Subjects had to rate the samples from 0 to 5 depending on their naturalness. The experiment shows that speech synthesized using the hybrid HMM-based approach sounds more natural than other synthetic variants. We discuss the results and the ways for further investigation and improvements in the last section.

Key words: speech synthesis, unit selection, naturalness evaluation, prosodic modeling

1. Introduction

The task of speech synthesis or text-to-speech (TTS) is to convert a written text to sounds. The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood, i. e. the quality of synthetic speech depends primarily on two main factors: its intelligibility and naturalness. It is possible to say that the problem of intelligibility for speech synthesis is already solved [Taylor 2009: 474]. Extensive research in the field of speech synthesis during the last few decades allowed synthetic speech to sound quite natural, and its characteristics come close to those of human speech.

At present the two main and most popular methods of natural-sounding speech synthesis are unit selection concatenative synthesis and so-called hidden Markov model (HMM) synthesis based on statistic models.

Unit selection synthesis [Black, Hunt 1996] is based on determining the best sequence of candidate units from a speech corpus. Then these candidates are concatenated

to form the resulting words and sentences. This process may be followed by modification of prosodic features of units (duration, energy and pitch) to match prescribed values.

HMM-based TTS is also called statistical parametric synthesis. A TTS system of this type models frequency spectrum, fundamental frequency (pitch) and duration of speech by HMM and then generates speech waveforms directly from HMM based on the maximum likelihood criterion [Masuko 2002; Zen et al. 2004]. Although HMM TTS provides an easy way to modify voice characteristics, speech generated without natural units usually sounds less natural than unit selection synthesis. This is the reason why we use unit selection in our TTS system.

However, naturalness of speech depends not only on segmental quality. Prosodic features including pitch, duration and energy and the way of achieving their required values are by no means less important. There are several approaches to the task [Krivnova 2000]. In the next sections we consider two ways to obtain them which are used in our VitalVoice Russian TTS system [Oparin, Talanov 2007].

2. Rule-based approach

The first approach is rule-based. It consists of two steps. During the first step we define the intonation type of the phrase (i.e. syntagma) and the word bearing the nuclear pitch accent depending on punctuation, parts of speech of words in the phrase and presence of special trigger words (question words, conjunctions, etc.). This is performed by manually constructed rules. It is worth mentioning that phrase boundaries are already defined at this stage [Khomitsevich, Solomennik 2010]. At present we have six intonation types that are reliably derived from the text: completeness, incompleteness, general and special questions and two types of exclamations. This is a reduced set of types from [Volskaya, Skrelin 2009].

At the second stage (after phonetic transcription) allophones receive tone, duration and energy values [Volskaya, Skrelin 1998]. These parameters depend on the voice used and the intonation type. For long and short phrases we use different parameters. For pitch they set declination (based on average pitch) and deviation from it depending on stress and its type. Duration and energy are also specified depending on the position in the phrase and stress as deviations from average.

The parameters are manually adjusted with respect to statistics. So, for a new voice we can immediately apply only a model from a different voice combined with the average characteristics of the new voice. But for accurate tuning we need some additional time to obtain appropriate quality.

3. Hybrid approach

Our hybrid HMM plus unit selection approach is described in detail in [Chistikov, Korolkov 2012]. It combines all the advantages of both methods. Features used for model training and then for generating the necessary physical characteristics of allophones are listed in Table 1:

Table 1. Features used in the statistic intonation model

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 type 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	

The speech parameters are obtained from HMMs whose observation vectors consist of mel-frequency cepstral coefficients (MFCC), pitch and duration features; the speech signal is generated by a unit selection algorithm using the obtained speech parameters. The phonetic and linguistic information for the training parameters derives from the speech corpus markup [Prodan et al. 2009].

4. Experiment

In our experiment we follow the recommendations of the state standard specification GOST R 50840-95 “Speech transmission through communication channels. Methods for quality, intelligibility and recognizability evaluation” [State standard specification 50840-95 1995]. This standard specification is also applied to speech synthesizers evaluation.

The new female TTS voice Julia was tested. The evaluated synthetic speech variants were the following:

1. Rule-based prosody on a small speech corpus of 20 minutes (with manually corrected labels).
2. Rule-based prosody on a speech corpus of about 2.5 hours of speech (with manually corrected labels).
3. HMM-based prosody on the same speech corpus (2.5 hours, manual correction).
4. Rule-based prosody on a large (6 hours) automatically labeled speech corpus (without manual correction).

17 listeners, 8 female and 9 male aged from 20 to 55 were subjects for the listening test. Among them 11 were trained (i. e. in one way or another closely familiar with synthetic speech) while the other 6 had little or no contact with synthetic speech before.

They were given the task to rate the naturalness of 4 synthetic and one natural speech variants of seven test utterances:

(1) *Если хочешь быть здоров, советует Татьяна Илье, чисть зубы пастой «Жемчуг»!*
(2) *Вчера на московском заводе малолитражных автомобилей состоялось собрание молодежи и комсомольцев.*
(3) *В клумбах сочинской здравницы «Пуща», сообщает нам автоинспектор, обожгли шихту.*
(4) *Тропический какаду — это крупный попугай? Ты не злословишь?*
(5) *Актеры и актрисы драматического театра часто покупают в этой аптеке антибиотики.*
(6) *Нам с вами сидеть и обсуждать эти слухи некогда!*
(7) *Так ты считаешь, что техникой мы обеспечены на весь сезон?*

Ratings could vary from 0 to 5 with a step of 0.1 with clear description of rates (from [State standard specification 50840-95 1995]):

Table 2. Rates and their meaning

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

Five variants of each utterance were given in a random order with possibility of listening for each utterance several times if needed. The obtained ratings are as follows:

Table 3. Evaluation results

TTS type	Mean	Standard deviation
20 min. database	3.6	0.9
Rule-based prosody (2.5 hours)	4.1	0.7
HMM-based prosody (2.5 hours)	**4.3**	0.6
Auto-labeled database (6 hours)	3.7	0.8
Natural speech	4.9	0.1

If we exclude results for two subjects that show more than 20 % deviation from mean ratings and normalize the score to the rating of natural speech (as recommended by the standard specification) we will have 4.4 and 4.5 for rule-based and hybrid approaches respectively. All the synthetic types appeared to be in the same I class (rates from 3.6 to 4.5) of quality (according to [State standard specification 50840-95 1995]).

It should be mentioned that there was a clear connection between the rates and the subject's familiarity with synthetic speech. This may be seen in the diagram below where "a" means "naive" listener and "b" — a listener familiar with TTS (rates were averaged for all of four TTS types):

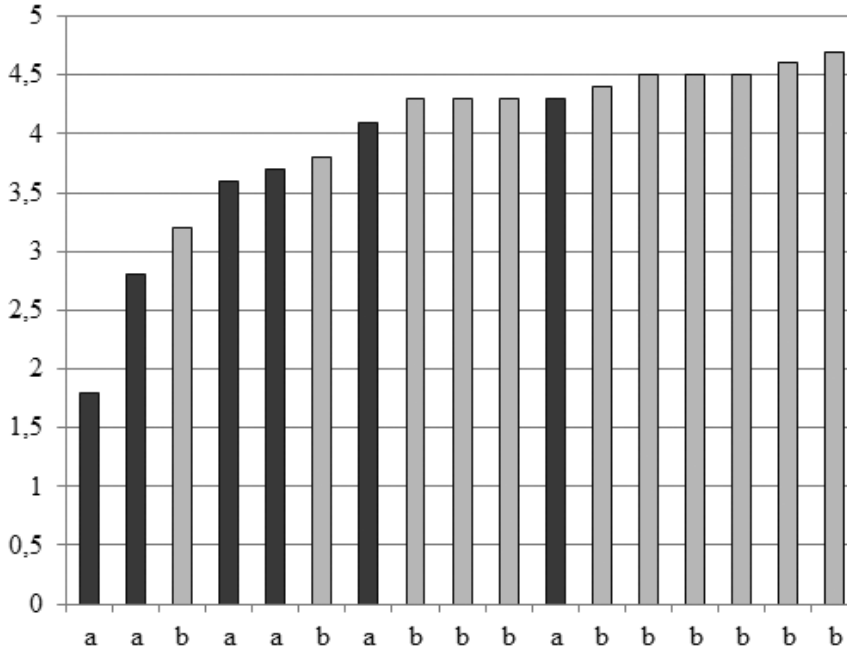

Fig. 1. Mean rates for different types of synthetic speech with respect to familiarity to TTS ("a" — "naive" listener, "b" — familiar to TTS)

We can observe that subjects unaccustomed to synthetic speech tend to give lower rates than others.

5. Conclusion

The obtained results show that by using a hybrid approach combining HMM-based and unit selection speech synthesis we have come close to natural sounding Russian synthetic speech. Also its usage permits fast adaptation of prosodic prediction for a new voice. For these reasons we plan to integrate HMM-based speech parameter generation in our voice-building system [Prodan et al. 2010]. Another important result is that even a small but phonetically balanced [Solomennik, Chistikov 2012] speech corpus can provide us with acceptable quality of synthetic speech.

However, there are still some problems to investigate and several ways of improving our system. Firstly, our evaluation of TTS using the purely automatically labeled speech corpus showed that there is room for improvement in the algorithm for detecting periods of fundamental frequency. Another way to improve prosodic quality is to include more verbal features for model training, primarily special words — potential intonation markers (specific conjunctions, particles etc.). There is also a strong need for a more powerful and at the same time generally accepted method of TTS evaluation in Russian.

References

1. *GOST R 50840-95* (1995), State standard specification 50840-95 “Speech transmission through communication channels. Methods for quality, intelligibility and recognizability evaluation” [GOST R 50840-95. Peredacha rechi po traktam svyazi. Metody ocenki kachestva, razborchivosti i uznavaemosti], Moscow.
2. *Black A. W., Hunt A. J.* (1996), Unit selection in a concatenative speech synthesis using a large speech database, Proceedings of ICASSP 96, Atlanta, Georgia, Vol. 1, pp. 373–376.
3. *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), Bekasovo, pp. 103–111.
4. *Khomitsevich O., Solomennik M.* (2010), Automatic pause placing in Russian text-to-speech system [Avtomaticheskaya rasstanovka pauz v sisteme sinteza russkoy rechi po tekstu], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2010”. [Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii «Dialog 2010»]. Bekasovo, pp. 531–537.
5. *Krivnova O. F.* (2000), Generation of phrase tone contour in speech synthesis systems [Generaciya tonal’nogo kontura frazy v sistemah avtomaticheskogo sinteza rechi], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2000” [Komp’iuternaia

Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii «Dialog 2000»], Protvino, Vol. 2, pp. 211–220.
6. *Masuko T.* (2002), HMM-Based speech synthesis and its applications, Doctoral dissertation, Tokyo Institute of Technology, Tokyo.
7. *Oparin I., Talanov A.* (2007), Outline of a New Hybrid Russian TTS System, Proceedings of the 12th International conference on Speech and Computer, SPECOM 2007, Moscow, pp. 603–608.
8. *Prodan A. I., Korolkov E. A., Oparin I. V., Talanov A. O.* (2009), Multi-tier markup of speech corpus for hybrid Russian TTS system «VitalVoice» [Osobennosti ispol'zovaniya mnogourovnevoy rametki zvukovogo korpusa unit selection v sisteme gibridnogo sinteza «jivoy golos»], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog 2009» [Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii «Dialog 2009»], Bekasovo, pp. 415–419.
9. *Prodan A. I., Talanov A. O., Chistikov P. G.* (2010), Voice building system for hybrid Russian TTS system «VitalVoice» [Sistema podgotovki novogo golosa dlya sistemy sinteza «VitalVoice»], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog 2010» [Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii «Dialog 2010»], Bekasovo, pp. 394–399.
10. *Solomennik A., Chistikov P.* (2012), Automatic generation of text corpora for creating voice databases in a Russian text-to-speech system, Proceedings of the Dialogue-2012 International Conference, № 11 (18), Bekasovo, pp. 607–615.
11. *Taylor P. (2009), Text-to-Speech synthesis, Cambridge University Press, Cambridge*.
12. *Volskaya N. B., Skrelin P. A.* (1998), Intonation modeling for speech synthesis [Modelirovanie intonacii dlya sinteza rechi po tekstu], Ufa.
13. *Volskaya N. B., Skrelin P. A.* (2009), System of intonation models for automatic utterance intonation interpretation: functional and perceptual characteristics [Sistema intonacionnyh modeley dlya avtomaticheskoy interpretacii intonacionnogo oformleniya vyskazyvaniya: funkcional'nye i perceptivnye harakteristiki], Proceedings of the third interdisciplinary workshop "Russian spoken speech analysis" (AR3-2009) [Trudy tret'ego mejdisciplinarnogo seminara «Analiz razgovornoy russkoy rechi» (AR3-2009)], St. Petersburg, pp. 28–40.
14. *Zen H., Tokuda K., Masuko T., Kobayashi T., Kitamura T.* (2004), Hidden semi-Markov model based speech synthesis, Proceedings of the International Conference on Spoken Language Processing, Interspeech 2004, Jeju Island, Korea, pp. 1393–1396.