ИСПОЛЬЗОВАНИЕ СТАТИСТИЧЕСКИХ МЕТОДОВ ДЛЯ ОПРЕДЕЛЕНИЯ МЕСТ И ДЛИТЕЛЬНОСТЕЙ ПАУЗ ПРИ АВТОМАТИЧЕСКОМ СИНТЕЗЕ РУССКОЙ РЕЧИ

Хомицевич О. Г. (khomitsevich@speechpro.com),
Чистиков П. Г. (chistikov@speechpro.com)

ООО «ЦРТ», Санкт-Петербург

Ключевые слова: синтез речи, расстановка пауз, паузирование, интонационное членение, просодический анализ, статистические методы

USING STATISTICAL METHODS FOR PROSODIC BOUNDARY DETECTION AND BREAK DURATION PREDICTION IN A RUSSIAN TTS SYSTEM

Khomitsevich O. G. (khomitsevich@speechpro.com),
Chistikov P. G. (chistikov@speechpro.com)

Speech Technology Center Ltd, St. Petersburg, Russia

The paper deals with statistical methods for predicting positions and durations of prosodic breaks in a Russian TTS system. We use CART and Random Forest classifiers to calculate probabilities for break placement and break durations, using grammatical feature tags, punctuation, word and syllable counts and other features to train the classifier. The classifiers are trained using a large high-quality speech database consisting of read speech. The experimental results for prosodic break prediction shown an improvement compared to the rule-based algorithm currently integrated in the VitalVoice TTS system; the Random Forest classifier shows the best results, although the large size of the model makes it more difficult to use in a commercial TTS system. To make the system more flexible and deal with the remaining break placement errors, we propose combining probabilities and rules in a working TTS system, which is the direction of our future research. We observe good results in experiments with predicting pause durations. A statistical model of break duration prediction has been implemented in the TTS system in order to make synthesized speech more natural.

Keywords: speech synthesis, TTS, text-to-speech, prosodic breaks, prosodic boundaries, pauses, statistical models

1. Introduction

Natural-sounding prosody is a key component for a successful Text-to-Speech (TTS) system, and correct prosodic segmentation of speech is necessary for achieving this goal. In natural speech, if an utterance is sufficiently long, it is normally divided into prosodic phrases, which are marked by intonational unity and are usually separated by pauses. Large chunks of speech pronounced without any breaks sound monotonous and are uncomfortable for the listener. In addition, accurate break placement enhances the intelligibility of speech, while pauses in the wrong positions can distort the meaning of a sentence or make it incomprehensible.

The way our natural speech is segmented prosodically depends on various factors. A major factor is syntactic structure; prosodic breaks often fall between syntactic constituents, so that syntactic structure can be seen as "mapped" onto prosodic phrases [1, 2]. However, the length of the sentence, semantics of certain words, and other features also play a role [3]. In a TTS system, these factors can be captured either by explicit rules defining which words in the synthesized sentence should be followed by a pause [4, 5], or by statistical models trained on large speech corpora and predicting probabilities of prosodic breaks. The latter method has become prevalent in the recent years (see, for example, [6–9]), and in this paper we will explore this approach as applied to a Russian TTS system.

2. Break detection using statistical methods

The principle behind automatic prosodic segmentation of speech is training a classifier on a large speech database which is labeled with word boundaries, Part-of-Speech (POS) and other grammatical tags, punctuation marks that were present in the original text (in case of read speech), and phrase breaks in the speech signal. Features like grammatical form, the place of the word in the sentence, the length of the sentence, the presence or absence of a punctuation mark, etc, are used by the classifier to predict the location and length of phrasal breaks in the synthesized speech.

This method has yielded good results for English and a number of other languages (as reported in [8] for Spanish, [9] for Arabic, etc), although some problems are bound to arise if the method is applied to Russian. Unlike English, Russian has relatively free word order, which means there is a lot of variation in possible POS sequences, and data sparseness can be an issue for model training. Russian also has rich morphology, which greatly increases the number of grammatical tags required for labeling text (and also increases variation in word form combinations). A large number of word forms in Russian are homonymous, so correct homonym resolution is essential for phrasal break detection, and errors in grammatical labeling of homonyms often lead to errors in break placement.

Despite these problems, statistical methods of phrasal break placement and break length prediction are a promising approach for Russian TTS systems, first of all because they aim to model the natural behavior of speakers, rather than rely on rigid rules and constants. They are also easier to implement because they do not require much expert linguistic knowledge, though tuning the system for practical use may require additional linguistic constraints. In this paper we describe methods of phrasal break prediction using CART and RF classifiers, which are tested using the VitalVoice Russian TTS system [10].

3. Experimental setup

We conducted experiments using the CART classifier [11] for predicting both break placement and break duration, and the Random Forest classifier [12] for break placement only.

CART is a recursive partitioning method based on minimization of partition goodness criterion (1):

where

$$G(C_1, C_2) = \frac{D(C_1)T(C_1)+D(C_2)T(C_2)}{T(C_1)+T(C_2)} \quad (1),$$

$$D(C) = \frac{2*\left(\sum_{i=1}^{|C|}\sum_{j=1}^{|C|}\left[d\left(U_i,U_j\right)\right]\right)}{|C|^2-|C|} \quad (2),$$

$$T(C) = .5(|C|^2 - |C|) \quad (3),$$

|C| is a size of cluster C, d(U,V) is a distance between U и V vectors, stop criterion is the minimal number of items in the cluster (in our work this number is 3).

A Random Forest classifies data using a given set of features by means of a hierarchy (a “tree”) of queries, based on the predictive value of each feature at each point. The classifier is capable of processing large amounts of training data. The leaves of each tree in the forest store the class distribution of all samples falling into the corresponding region of the feature space, which then serve as predictors for test samples. In our system, we use a forest containing 100 trees, and the probabilistic value is calculated by dividing the number of trees classifying the target class by the total number of trees. Each tree is built on the basis of 60% of randomized training data. This prevents the data from being dependent on noise in the training set.

For break placement prediction, we examined each sentence separately, because in the TTS system text is divided into sentences during the normalization process, and each sentence is processed separately. As for break duration, both intrasentential and intersentential breaks were taken into account.

We used the following features for the classification:

- Punctuation features. All common punctuation marks are included in the feature set. These features are calculated for the current word as well as two preceding and two following words.

- Word and syllable count features: number of words and syllables in the sentence, number of words and syllables from the previous break to the current word, from the current word to the end of the sentence, etc.
- Morphological word features. Morphological information is calculated using the VitalVoice speech synthesis engine which includes a morphogrammatical dictionary. Since using all grammatical features of Russian words would result in an enormous number of tags, which would be too large for the classifier to cope with, we decided to limit the grammatical features to part of speech and case. We also use the information on whether or not the word is a proper noun (name, geographical location, ect). Another word feature is capitalization of the first letter of the word. These features are calculated for the current word as well as two preceding and two following words. In addition, we use specific features intended for capturing grammatical agreement between words: whether or not the grammatical form of the current word matches that of the following word and the second word on the right.

Both in model training and testing, homonym resolution is necessary to minimize the number of errors due to incorrect feature calculation. We use homonym resolution provided by the VitalVoice TTS system, which labels 96% of homonyms correctly [13].

The speech database used in the experiments was originally recorded as the Unit Selection speech database for the VitalVoice TTS system and consists of read speech by nine speakers (four male and five female). The texts read by the speakers are contemporary Russian works of fiction as well as newspaper articles on the topics of politics and technology. The database comprises over 50 hours of speech, which contain over 38,000 phrasal breaks. It was divided into a training set and a test set.

4. Results and discussion

4.1. Break placement

Evaluating automatic phrase break placement is not straightforward, since the accuracy of the classification can be estimated in different ways. One way is to calculate the accuracy of the prediction (break or no break) for each pair of words (a value sometimes called "juncture"). However, significantly more junctures in speech are non-breaks than breaks, so this measure will usually yield high results even for a poorly performing system. If we only consider breaks, then two types of errors have to be taken into account: breaks added by the algorithm that were not there in the data (False Alarms, FA), and breaks incorrectly skipped by the algorithm (False Rejections, FR). Some authors [6] devise their own evaluation system which incorporates both measures; however, we prefer to use the standard precision/recall/F-score evaluation.

In Table 1 we present the results for automatic break placement (CART and Random Forest) compared to the results of the "baseline" rule-based algorithm that is currently implemented in the standard version of the VitalVoice TTS system [14]. The test set contained 47,819 junctures (word pairs inside sentences) and 6,186 phrase breaks.

Table 1. Results of automatic break detection

	Baseline TTS	CART	Random Forest
Correct junctures	43,254 (90.45%)	44,358 (92.76%)	44,723 (93.53%)
Correct breaks	5,042 (81.51%)	5,176 (83.67%)	4,624 (74.75%)
FA	3,421 (55.30%)	2,451 (39.62%)	1,534 (24.80%)
FR	1,144 (18.49%)	1,010 (16.33%)	1,562 (25.25%)
Recall	0.82	0.84	0.75
Precision	0.60	0.68	0.75
F-score	0.69	0.75	0.75

The results of both classifiers show an improvement on the baseline system: they yield a higher F-score, and the rates of FA to FR errors are more balanced. Both the CART and the RF classifiers show a maximum F-score of 75%; however, RF can be tuned so that the Precision and Recall counts are equal. CART shows a higher percentage of correct breaks due to a lower level of False Rejection errors; however, the RF classifier gives the highest percentage of correct junctures. Overall, RF can be considered the best-performing model.

The results of the classifier also compare well with those reported in the literature. For instance, for English [6] reports up to 91.1% correct junctures and the F-score of up to 71.9; [7] improves their result and attains the F-score of 74.4, which is basically equal to our result.

However, some comments on the model's performance are in order. First of all, an automatic performance test evaluates the breaks locally, without estimating the overall naturalness of the whole utterance (a discussion of this issue is given in [15]). So the fact that an utterance can usually be segmented in several correct ways remains unaccounted for. This problem is especially obvious if we consider different speakers' performance when reading the same text. Our speech database contained the same text read by several speakers, so we had a chance to find out whether they placed prosodic breaks at the same word junctures. We took a text read by three speakers (about 500 sentences) and, taking the breaks placed by one of the speakers as the model to be "tested", checked how it would fare if the other two speakers would be taken as target performances. Then we repeated the experiment with a second speaker. The results (F-scores) are given in Table 2.

Table 2. Comparison of speakers' break placement (F-score)

	Speaker 1	Speaker 2	Speaker 3
Speaker 1 as model	*1.00*	0.69	0.71
Speaker 2 as model	0.71	*1.00*	0.76

As was expected, not all breaks made by two different speakers when pronouncing the same text actually overlap; if we compare three or more speakers, the discrepancy would probably be even greater. On the other hand, treating only those breaks that coincide for several speakers as necessary and ignoring all others is clearly wrong, because that would yield too few breaks. Interestingly, a statistical model predicts a speaker's breaks better than another human speaker; this can be explained by the fact that the model is able to generalize over multiple speakers' behavior.

Another aspect of this problem is that an automatic test that compares phrase breaks placed by the algorithm to those present in actual speech does not reflect the relative "gravity" of possible mistakes. Intuitively, some prosodic breaks in a sentence seem necessary, while others can be omitted; on the other hand, some word combination can in principle be separated by a pause, while others should be pronounced without a break. These distinctions are hard to formalize, so an automatic error detection system treats all errors as "equal". Of course, an automatic classifier should learn to avoid serious errors if the training database is sufficiently large, but in practice data sparceness is often a problem, especially for the CART classifier. In the course of subjective tests, we have identified several types of "egregious" errors that significantly worsen the impression of a model's performance, even if the overall error count is low:

- False alarm errors (inserted breaks): pauses after prepositions, conjunctions and other function words; pauses between agreeing words.
- False rejection errors (deleted breaks): lack of pauses on commas and other punctuation marks.

These errors are a particular problem for the CART classifier, though they are rare for the RF classifier. An advantage of a rule-based system is that it can easily exclude such errors by explicitly prohibiting pauses in certain word combinations and forcing them in others.

Finally, with a probability-based prosodic break model it is difficult to control rhythmic qualities of speech. The local character of decisions that the break placement algorithm takes can result in a sentence having too many pauses, while another sentence of approximately the same length and structure may have no breaks altogether. In a text that is sufficiently long, the frequency of breaks averages out and is judged by an automatic test as correct; however, specific sentences can be uncomfortable for the listener.

To sum up, even though a statistical break placement system imitates the performance of a human speaker fairly successfully, it can also make errors that should be avoided in a working TTS system. A possible solution is to "tune" the probability-based system by introducing a number of rules, which is the direction of our ongoing research.

One way to simplify the task for the break placement algorithm is to put obligatory pauses in places of punctuation marks and use the probability-based algorithm only for the text chunks without punctuation. However, in Russian punctuation is sometimes misleading in the sense that it is purely conventional and does not mark a prosodic break. So the rules need to be more elaborate than just placing a break at each punctuation mark.

In addition, breaks in certain word combinations can be prohibited. However, if we just delete breaks, the sentence may end up with too few of them. This issue is connected with the more general problem of rhythm: controlling the length of prosodic phrases and keeping the frequency of breaks constant seems to be necessary.

4.2. Break durations

The CART classifier predicts not only break positions but also the duration of each break it generates. Two versions of the model were trained. The first one predicts both break placement and break duration. The second one predicts break durations separately; that is, given a predetermined position for a prosodic break, the model predicts the break length for this position. This model can be used in combination with a rule-based break placement model or any other classifier.

In our experiments we first trained the classifier to predict the lengths of all prosodic breaks in the dataset: both those inside sentences and between sentences. After that, we decided to separate the two tasks: predicting sentence-internal vs. sentence-external breaks. It should be noted that in spontaneous speech, the notion of a sentence is controversial, and such an approach would probably fail; in that case it would be more productive to distinguish between types of breaks such as long and short breaks. However, since we were dealing with read speech, we felt that speakers were aware of sentences in the text and marked them prosodically, and we wished to imitate this effect in synthesized speech.

Break duration accuracy is much more difficult to evaluate than break placement accuracy because break lengths are not discrete and there can be no yes/no judgements. For our first model (predicting both break placement and duration), the problem is that if we evaluate the lengths of the breaks correctly predicted by the classifier, there still remain the inserted breaks (FA-type errors) whose lengths will be unaccounted for. For this reason, we decided to test the second model and to evaluate the correctness of the break length prediction that the classifier makes for each break position found in the test dataset. We considered a break as correct if its length did not deviate from the predicted length by more than a certain percentage, which we set as either 30% or 50%. The results are given in Table 3.

Table 3. Results for break length prediction

	Correct sentence — external breaks, %	**Correct sentence — internal breaks, %**
General model, 30% window	63.07%	42.74%
Specialized models, 30% window	64.12%	60.36%
General model, 50% window	81.88%	63.68%
Specialized models, 50% window	80.99%	80.16%

This table presents results for the general model (modeling all breaks in the dataset) and the specialized models (two separate models for sentence-external and sentence-internal breaks). We can see that the specialized models give a better approximation both for sentence-internal and sentence-external breaks (except for sentence-external breaks with a 50% evaluation window, where the results for the general model are slightly better). The difference is especially large for sentence-internal breaks, which are apparently not predicted accurately enough by the general model.

The baseline algorithm for break durations in the VitalVoice TTS system uses constants, so all sentence-external breaks have the same length, and there are only three types of sentence-internal breaks differing by their length. Implementing probability-based pause length prediction is promising because it makes synthesized speech sound less monotonous and more varied, which contributes to overall naturalness of speech. Subjective listening experiments with a new TTS system where constants were replaced by predicted values showed positive results.

5. Conclusions and future research

In this paper we have presented a probability-based approach to prosodic analysis of speech. The aim of our research was to evaluate different models of break placement and break length prediction for use in a Russian TTS system. The following conclusions can be drawn at the present stage of the research:

1. A break placement algorithm based on a probabilistic model gives better test results than the baseline rule-based algorithm. However, subjective evaluation shows that the presence of errors, even if they are rare, produces a bad impression on listeners, so some additional tweaking is needed in order to include the algorithm in a working TTS system. The CART model displays more errors than the RF model, and these errors are typically more "serious"; however, the RF model slows down the system due to its large size. Adapting statistical break placement methods for practical use will be the subject of future work.
2. CART-based prediction of pause lengths works well, especially if sentence-internal and sentence-external breaks are modeled separately. This model has been included in a new version of the VitalVoice TTS system to replace the old constant-based system, and has received good reviews from expert listeners.

References

1. *Bachenko, J., Fitzpatrick, E.* (1990), A computational grammar of discourse-neutral prosodic phrasing in English, Computational linguistics, Vol. 16 (3), pp. 155–170.
2. *Tepperman, J., Nava, E.* Where should pitch accents and phrase breaks go? A syntax tree transducer solution. Proceedings of the Twelfth Annual Conference of the International Speech Communication Association, Florence, Italy, 2011, pp. 1353–1356.
3. *Zellner, B.* (1994), Pauses and the temporal structure of speech, in Fundamentals of speech synthesis and speech recognition, Chichester, John Wiley, pp. 41–62.
4. *Abney, S.* (1991). Parsing by chunks. Principle-based parsing, Vol. 44, pp. 257–278.
5. *Atterer M.* Assigning Prosodic Structure for Speech Synthesis: A Rule-based Approach. Proceedings of Prosody, Aix-en-Provence, 2002, pp. 147–150.

6. *Black, A. W., Taylor, P.* (1998). Assigning phrase breaks from part-of-speech sequences. Computer Speech & Language, Vol. 12(2), pp. 99–117.
7. *Busser, B., Daelemans, W., Bosch, A. V. D.* Predicting phrase breaks with memory-based learning. 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis, 2001, pp. 29–34.
8. *Torres, H. M., Gurlekian, J. A.* Automatic determination of phrase breaks for Argentine Spanish. Proceedings of Speech Prosody, 2004, pp. 553–556.
9. *Sawalha, M., Brierley, C., Atwell, E.* Predicting Phrase Breaks in Classical and Modern Standard Arabic Text. Proceedings of LREC: Language Resources and Evaluation Conference, 2012.
10. *VitalVoice™* Russian TTS system, demo available at: http://cards.voicefabric.ru/.
11. *Loh, W.-Y.* Classification and Regression Tree Methods, in Encyclopedia of Statistics in Quality and Reliability, Wiley, 2008, pp. 315–323.
12. *Breiman, L., Cutler, A.* Random Forests, available at:
http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.
13. *Khomitsevich O. G., Rybin S. V., Anichkin I. M.* Linguistic analysis for text normalization and homonymy resolution in a Russian TTS system [Ispol'zovanie lingvisticheskogo analiza dlja normalizatsii teksta i snjatija omonimii v sisteme sinteza russkoj rechi.] Izvestija vuzov. Priborostroenie. Tematicheskij vypusk "Rechevye informatsionnye sistemy". [Instrument making. Thematic issue "Speech information systems"] №2, 2013 (in press).
14. *Khomitsevich O. G., Solomennik M. V.* Automatic pause placement in a Russian TTS system [Avtomaticheskaja rasstanovka pauz v sisteme sinteza russkoj rechi po tekstu]. Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2010" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010"]. Bekasovo, 2010. pp. 531–537.
15. *Jian, L., Bolei, H., Hairon, X., Linfang, W., Braga, D., Sheng, Z.* Expand CRF to Model Long Distance Dependencies in Prosodic Break Prediction. Proceedings of the Twelfth Annual Conference of the International Speech Communication Association, Portland, Oregon, 2012.