Speech Technology Center

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

Olga Khomitsevich, Pavel Chistikov
{khomitsevich,chistikov}@speechpro.com
01.06.2013

Prosodic boundaries in TTS

- Natural-sounding prosody is key for Text-to-Speech (TTS).
- Prosodic boundaries help to:
  - Make speech more comfortable for the listener;
  - Disambiguate sentences and make speech more intelligible.
- It is difficult to predict correct break placement and duration automatically because many factors are at play:
  - Syntactic structure of the sentence;
  - Sentence length;
  - Semantics, emphasis;
  - Etc...

Methods for break prediction in TTS

- **Rule-based** methods (*used in baseline Vital Voice TTS*):
  - Rely on expert knowledge;
  - Take a long time to develop;
  - Are difficult to develop due to the complexity of the data.
- **Statistical** methods (*present work*):
  - Easy and fast to train given large annotated corpora;
  - But: subject to data sparceness problem;
  - May be difficult for languages with free word order and rich morphology due to large numbers of feature combinations.

Experimental setup: classifiers

- **CART**: predicting break placement and break duration.
  - CART is a recursive partitioning method based on minimization of partition goodness criterion.
- **Random Forest**: predicting break placement.
  - 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;
  - We use a forest containing 100 trees; each tree is built on the basis of 60% of randomized training data.
- The classifiers are used to predict the probability of a break after a word and/or the duration of the break.

Experimental setup

- Word **features** used for classification:
  - Punctuation;
  - Sentence length and position of the word in the sentence;
  - Morphological features, capitalization;
  - Features are computed for the current word and two previous/following words.
- Speech **database**:
  - Read speech (TTS Unit Selection database);
  - Over 50 hrs of speech (over 38000 phrasal breaks);
  - Divided into training and testing datasets.

Experimental results: break placement

	Baseline TTS	CART	Random Forest
Correct junctures	43254 (90.45%)	44358 (92.76%)	44865 (93.82%)
Correct breaks	5042 (81.51%)	5176 (83.67%)	4695 (75.90%)
FA	3421 (55.30%)	2451 (39.62%)	1463 (23.65%)
FR	1144 (18.49%)	1010 (16.33%)	1491 (24.10%)
Recall	0.82	0.84	0.76
Precision	0.60	0.68	0.76
F-score	0.69	0.75	**0.76**

Experimental results: break placement

- Both classifiers show an improvement on the baseline;
- **Random Forest** yields the best results;
- F-score values are comparable with those reported in the literature for English.
- **However**, automatic testing does not reflect possible variations in break placement.
- It is important to avoid "serious" errors:
  - Breaks in impossible locations;
  - Omission of necessary breaks.
- Combination of rules and statistical models may be needed.

Experimental results: break duration

	Sentence-external breaks	Sentence-internal breaks
General model	0.25	0.23
Specialized models	0.19	0.16

Break durations were predicted for break positions in the database;

The table shows:

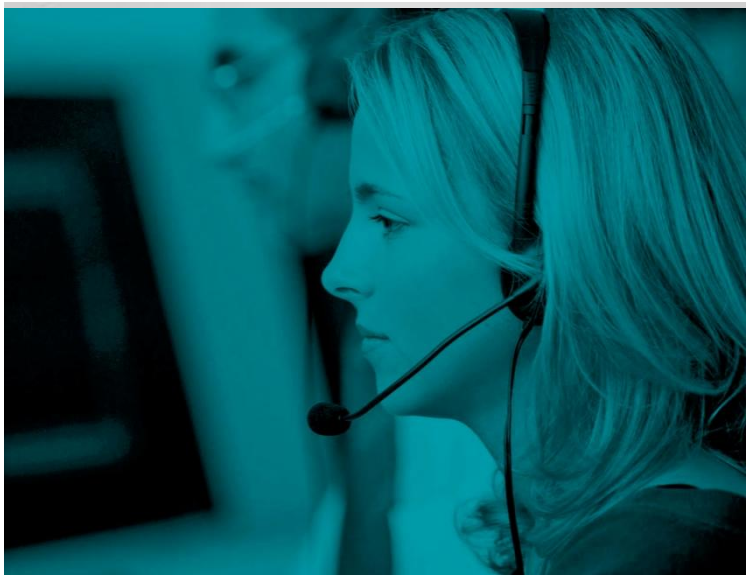
- A model for predicting break durations both between and inside sentences (general model);
- A combination of two separate models for intra-sentential and inter-sentential breaks (specialized models);
- NRMSD (Normalized Root-Mean-Square Deviation) measure was used.

The specialized models give a better approximation both for sentence-internal and sentence-external breaks.

Conclusions

- Break placement models based on CART and RF classifiers give more accurate test results than the baseline rule-based algorithm.
- The CART model displays more errors than the Random Forest model.
- Break duration prediction works better when sentence-internal and sentence-external breaks are modeled separately.
- A hybrid algorithm combining statistical models and rules may be efficient for applied TTS systems.

Thank you for your attention!



ABOUT THE COMPANY

ABOUT THE COMPANY

Speech Technology Center (STC) is an international leader in speech technology and multimodal biometrics. It has over 20 years of research, development and implementation experience in Russia and internationally.

STC is a leading global provider of innovative systems in high-quality recording, audio and video processing and analysis, speech synthesis and recognition, and real-time, high-accuracy voice and facial biometrics solutions. STC innovations are used in both public and commercial sectors, from small expert laboratories, to large, distributed contact centers, to nation-wide security systems.

STC is ISO-9001: 2008 certified.

CONTACTS

Russia
4 Krasutskogo street, St. Petersburg, 196084
Tel.: +7 812 331 0665
Fax: +7 812 327 9297
Email: info@speechpro.com

USA
Suite 316, 369 Lexington ave
New York, NY, 10017
Tel.: +1 646 237 7895
Email: sales-usa@speechpro.com