РЕЧЕВЫЕ ТЕХНОЛОГИИ

EVALUATION OF NATURALNESS OF SYNTHESIZED SPEECH WITH DIFFERENT PROSODIC MODELS

Solomennik A. I., Speech Technology Ltd., Minsk, Belarus
Chistikov P. G., Speech Technology Center Ltd,
St. Petersburg, Russia

Dialogue 2013, the 19th International Computational Linguistics Conference, May 29th - June 2nd, Bekasovo

Unit selection and HMM

- At present the two main and most popular methods of natural-sounding speech synthesis
- Quite natural, and nearly 100 % intelligible speech
- US:
  - Determining the best sequence of candidate units from a speech corpus
  - Candidates are concatenated to form the resulting words and sentences
  - Minimum DSP
- HMM:
  - Models frequency spectrum, pitch and duration of speech by HMM
  - Generates speech waveforms directly from HMM based on the maximum likelihood criterion
  - Easy way to modify voice characteristics
  - Usually sounds less natural than US synthesis

Rule-based prosody generation

- (1) Rules define the intonation type (6 types) of the phrase and the word bearing the nuclear pitch accent
  - Punctuation
  - POS
  - Question words, conjunctions, etc.
- (2) Intonation type + length of the phrase + voice parameters -> tone, duration and energy values
  - Declination (based on average pitch) + deviation from it depending on stress and its type
  - Duration and energy depending on the position in the phrase and stress = deviations from average

HMM-based prosody generation

- Speech corpus markup

- HMMs observation vectors:
  - MFCC
  - Pitch
  - Duration

- Speech parameters

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	

Experiment

- 17 listeners: 8 female and 9 male aged from 20 to 55
- 11 were familiar with synthetic speech, 6 had little contact with synthetic speech before
- Evaluated synthetic speech variants:
  1. Rule-based prosody, 20 min, manually corrected labels
  2. Rule-based prosody, 2.5 h, manually corrected labels
  3. HMM-based prosody, 2.5 h, manually corrected labels
  4. Rule-based prosody, 6 h, automatically labeled

Rates

- GOST R 50840-95 “Speech transmission through communication channels. Methods for quality, intelligibility and recognizability evaluation”:

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

Test phrases

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?

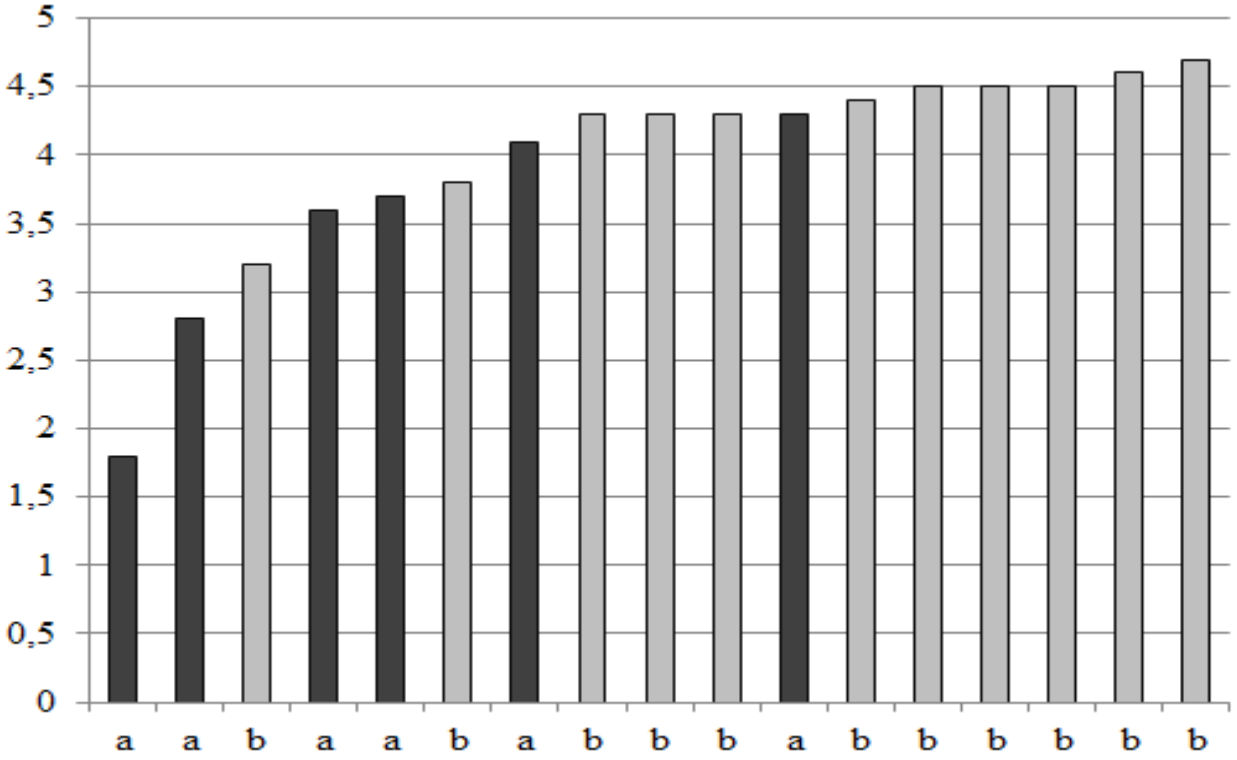




Results

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-labelled database (6 hours)	3.7	0.8
Natural speech	4.9	0.1

Naïve vs. familiar to TTS



“a” – “naive” listener, “b” – familiar to TTS

Conclusions

- Hybrid approach combining HMM-based and unit selection speech synthesis
  - Close to natural sounding Russian synthetic speech
  - Fast adaptation of prosodic prediction for a new voice
- Small phonetically balanced speech corpus can provide us with acceptable quality of synthetic speech
- Ways of improving:
  - Improvement in the algorithm for detecting periods of fundamental frequency
  - Include more verbal features for model training
  - Powerful and generally accepted method of TTS evaluation in Russian

THANK YOU FOR YOUR ATTENTION!

www.speetech.by
www.speechpro.ru