Prominence detected by listeners

for future speech synthesis application

Maria Eskevich
Saint Petersburg State University
Saint Petersburg, Russia
maria.eskevich@gmail.com

Abstract

The point of interest in the present investigation is to find out and to make a pilot statistical presentation of the prominence distinguished by native speakers in read aloud texts taken from the Russian corpus for text-to-speech unit-selection synthesis.

The TTS system uses the linguistic information encoded in the input text. Therefore the parameters which are easily extracted from the text (part of speech classes, number of syllables) are admitted as the basis for the classification of the words detected as prominent by listeners.

On further steps the TTS system has to assign prosodic structure and its suprasegmental acoustic parameters. The professionally made phonetic segmentation and analysis of syntagmatic structures of the material are compared with the judgments of native speakers in order to find some of these acoustic correlates.

1 Introduction

Prediction of word prominence might help us to build more natural synthesized speech and pay more attention to some parts of speech in process of speech recognizing because it brings more valuable information. The person who is making it prominent (speaker or writer) is doing it consciously, however it is probably that this person is not aware of the physical mechanism (changing of pitch, intensity or syllable duration).

According to Taylor (2008) there are different levels of prominence:

(1) conceived by the author of some written text which normally is not intended to be read and therefore in this text special constructions are used to emphasize important things (e.g. he said it *angrily*, he said *aloud*)

(2) conceived by people who are reading the text aloud

(3) conceived by people who are talking and emphasizing something in their utterances

In all these cases, native speakers understand where the authors put the emphasis and the speech seems natural and normal for them.

While doing speech synthesis we have to predict prominence by using information available in the text. In the sentence we have word-accent and sentence-level stress. The former is put according to the rules of the language (for example on the last syllable as in French) or according to the dictionary (as in English or in Russian). The latter is put according to the meaning of the sentence and communicative intention.

We have also to distinguish three levels of word accents: accented, unaccented, and clitisized. Clitisized words are unaccented, but additionally lack word stress (Cole, 1997).

On the other hand, we may distinguish the words according to main lexical classes – function words and content words (Holmes at al., 2001). Function words (or grammatical words) are words that have little lexical meaning or have ambiguous meaning, but instead serve to express grammatical relationships with the other words within sentence, or specify the attitude or mood of the speaker (Skrelin at al., 1997). Content words always have meaning (Noun names the object; Verb, Adjective and Adverb name its features). Function words and Pronouns belong to

Kristiina Jokinen and Eckhard Bick (Eds.)
NODALIDA 2009 Conference Proceedings, pp. 251–254

the closed-class words and content words to the open-class words.

It seems that content words should have accent and can have prominence. The clitisized function words (as particles in Russian) may add some extra meaning, and in this way they may increase prominence of the content word.

The structural data of the input text are processed by the algorithm which predicts the prosodic structure and the prominence of some words. Later on the TTS system changes the acoustic characteristics of concatenated units in a way the researchers suppose it has to be realized in the signal.

Streefkerk at al. (2001) tried to predict prominence using rules based mainly on the word-class classification and achieved the score of 92.6 % right prediction. The prominence was considered as gradual parameter and the value was counted as the sum of marks assigned while applying the rules. Content words received one mark, then additional mark for special parts of speech within content words) and also on the polysyllabic structure of content words (polysyllabic words from the classes Pronoun, Verb, Adverb). The information about the word-class of the previous word was partially used, only for the case of the Noun preceded by an Adjective. This limitation seems reasonable since the other research confirmed that the word class and the clause position are more relevant for prominence prediction than word class of context. For Russian the experiments of the perception of the combination Noun + Adjective and Adjective + Noun also did not reveal strong difference in prominence perception (Altuhova, 2007).

In the prosodic organization of Russian text-to-speech synthesis there were several stages of accentuation assignment: content words were unified with cliticized words; on the level of phrase the content words received stress; then the last content word in phrase received additional syntagmatic stress; and in the end special logic stress derived from the special syntactic factors (Skrelin at al., 1997). These words marked with phrasal stress and special logic stress are supposed to be perceived by listeners as prominent.

2 Experiment

2.1 Material/Method

Corpus for Russian text-to-speech unit-selection synthesis is created at the Saint Petersburg State University. For this pilot experiment 100 sentences read by 2 speakers (male (MS) and female (FS)) were taken. Both speakers are professional announcers that is why it is possible to assume that the quality of their voice is not going to change and become more monotonous due to tiredness caused by reading.

The speech material was presented via headphones and judgments were made on printout of the text. There was no response time limitation. The subjects could decide for themselves how many times to replay the utterance.

8 Russian native speakers (3 male and 5 female), aged from 25 to 31, passed the listening tests and gave their responses regarding the prominence in each sentence and its position in each phrase. It was not explained in details what kind of prominence they should find. The request was to indicate the prominence where they hear it and to evaluate it from 1 to 10 points. The amount of points assigned to the word is the level of prominence intensity felt by the native speaker.

As a result the statistics was built on the basis of the judgments made by listeners. The word was considered prominent if at least four listeners marked it as prominent. Since the data might be affected by the restricted number of speakers (just two voices) the results are given separately.

3 Results

The average length of the sentence is 8.79 words. The total amount of words is 863.

All words in text are divided into two general classes – content words (Verbs, Nouns, Pronouns, Adverbs, Adjectives, Numerals) and function words (Conjunctions, Propositions, Particles and Interjections).

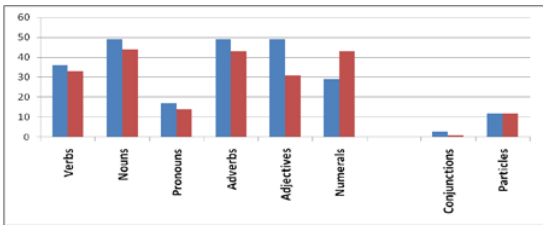

Figure 1. Distribution of the words according to the parts of speech classification (percentage). First column – Male speaker, second column – Female speaker

Figure 1 shows the distribution of words used in experiment among the parts of speech and the number of words (percentage) within corresponding part of speech detected as prominent by listeners. The data for the male and the female speakers are given separately. 29 % of the words in the male speaker's sentences and 25 % of the words in the female speaker's sentences were marked as prominent. It means that each sentence had one or more prominent words. The number of prominent content words (239 for MS and 206 for FS) is much greater than the number of prominent function words (13 for MS and 11 for FS) as it was expected and found by Widera at al. (1997) for German.

Even though different speakers have their own style of pronunciation, they are giving a comparable level of prominence to the words since they read the same text. The slight difference in percentage may consequently show individual characteristics and tendency to emphasize more or less, but the order of numbers stays the same. This implies that the text contains some linguistic information.

The average length of prominent words is presented in Figures 2. It shows the percentage of the words with 1-6 syllables in each part of the speech (only prominent words). It turns out that the content words, such as Verbs, Nouns, Adverbs and Adjectives, which contain 2-3 syllables are more prominent (2 syllables - 40-50% and 3 syllables – 30-40 % of all prominent words). Among the content parts of speech only pronouns are mainly presented by 1-syllable words. It can be explained by the average length of this class of words in Russian (1-2 syllables) and the fact that this is a closed-class.

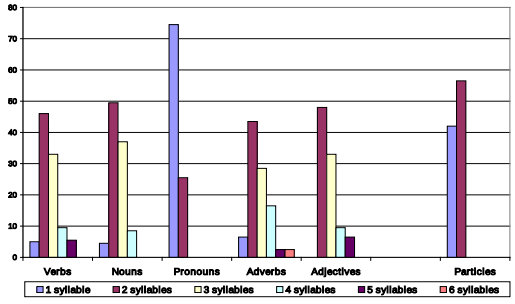

Figure 2. Average number of prominent words of each part of speech (%) with corresponding number of syllables (for 2 speakers).

3.1 Discussion

The distribution of the prominence assignment among content words shows that the part of speech tagging of the input text might help to predict prominence. For Dutch the additional marks were given to Noun, Adjective, Numeral and Negation (Streefkerk at al., 2001), but the results of experiment shows that for Russian these extra marks can be added to Verb as well. The polysyllabic structure in Russian also differs and the results show that for Russian the polysyllabic Verb, Noun, Adverbs, Adjectives and monosyllabic Pronouns can receive such additional marks.

All the data were listened by a phonetician to assign logical stress which is supposed to be consistently perceived by native speakers as prominence.

When the logical stress coincides with phrasal stress, it is perceived by listeners in 92 % for MS and 82 % for FS. There are some cases when less than three listeners perceive prominence, but there are no cases when it is not perceived at all. It means that this type of pattern can be used for sentences with predicted prominence. The other question is how to derive the information about logical stress from written text when it is emphasized by font and has to be done by means of syntactic analysis.

On the other hand, there are 33 occurrences (22 for MS and 10 for FS) when logical stress does not coincide with phrasal stress. And these cases are perceived by speakers as prominent in 64 % for MS and 60 % for FS. It is interesting that in 25 % cases for both speakers only two listeners marked prominence. These listeners differ from other ones as they received musical education and seem to have ear for music that might be the reason for detecting pitch changing as good as phoneticians. However there are some cases that are not perceived as prominence at all.

3.2 Conclusion and Future Work

As the experiment has shown, listeners quite easily distinguish prominent words and are mainly uniform in assigning it. Further interrelated directions of research are prediction of prominence on the basis of the text (part of speech classification and assigning of prominence) and further acoustical analysis of words marked as prominent and the pitch contours they make part in, thus investigating how to set prominent parame-

ters for which words have to be emphasized according to the previous written text analysis. It was found that the distribution of the prominence within part of speech classification can be added for Russian speech synthesis as well as it is added for other languages. The coincidence of marks of professional phonetician segmentation and of native speakers means that the correlates of prominence are presented in the signal and can be found in further acoustic analysis.

References

Cole R., Mariani J., Uszkoreit H., Zaenen A. and Zue V., 1997 *Survey of the state of the art in human language technology*, Cambridge University Press.

Holmes, J. and Holmes, W. 2001. *Speech synthesis and recognition. London and New York.*

Taylor P., 2008. *Text-to-Speech Synthesis* Cambridge University Press.

Widera C. and Portele T., 1997. *Prediction of word prominence.* In Proc. European Conf. on Speech Communication and Technology.

Streefkerk, B. M., Pols, L. C. W., and Bosch, L. F. M., 2001. *Up to what level can acoustical and textual features predict prominence*", Proc. Eurospeech'01, Vol. 2, Aalborg, Denmark: 811-814,

Raux A. and A. Black, 2003. *A unit selection approach to F0 modeling and its application to emphasis*. ASRU 2003, St Thomas, US Virgin Islands.

Altuhova E., 2007. *Changing of the pitch on an adjective followed by a noun and its' interpretation for intonation division*. Materials of the XXXVI International Philological conference [in Russian]

Skrelin P.A., Svetozarova N.D. and Volskaja N.B., 1997. *Modelling of prosodic organization of Russian speech.* In Bulletin of Phonetic Fund of Russian language. N. 7. Saint Petersburg-Bohum. [in Russian]

ISSN 1736-6305 Vol. 4
http://hdl.handle.net/10062/9206