Proceedings of the International Multiconference on
Computer Science and Information Technology pp. 435–438
ISBN 978-83-60810-27-9
ISSN 1896-7094

Automatic Detection of Prominent Words in Russian Speech

Daniil Kocharov
Saint-Petersburg State University
Universitetskaya emb., 11, 199034,
Saint-Petersburg, Russia
Email: kocharov@phonetics.pu.ru

***Abstract*—An experimental research with a goal to automatically detect prominent words in Russian speech is presented in this paper. The proposed automatic prominent word detection system could be further used as a module of an automatic speech recognition system or as a tool to highlight prominent words within a speech corpus for unit selection text-to-speech synthesis. The detection procedure is based on the use of prosodic features such as speech signal intensity, fundamental frequency and speech segment duration. A large corpus of Russian speech of over 200 000 running words was used to evaluate the proposed prosodic features and statistical method of speech data processing. The proposed system is speaker-independent and achieves an efficiency of 84.2 %.**

I. Introduction

THE SOLUTION of the prominent word detection task is to be used within the field of speech technologies while developing automatic speech recognition, unit-selection text-to-speech synthesis, spoken term detection, video and audio data indexing. For example, natural speech understanding systems need to know not only "what" has been said, but also "how" it has been pronounced. Intonation prominence is very important linguistic information for speech understanding, i.e. [1] showed that the use of word prominence degree helped to disambiguate the meaning of utterances.

The procedures of speech signal processing, prosodic features extraction and statistical speech data processing were developed during a series of investigations. The experiments used the data of CORPRES speech corpus created at the Department of Phonetics of Saint Petersburg State University [6]. The paper contains the experimental results and efficiency evaluation of the developed automatic prominent word detection system.

II. Prosodic Features of Word Prominence

A speaker prosodically emphasizes a word in an utterance to make it stand out of the surrounding words. The most common cure for doing that is a pitch accent that is acoustically expressed by the increase of local pitch maxima and minima. Intonational prominence reflects different aspects of pragmatics. It can express attitudes such as doubt, uncertainty and surprise, demonstrate anaphora references, the location of rheme and theme in the utterance, as well as show whether the utterance is a question or statement. Recently many research efforts have been dedicated to prominence detection due to its importance in such fields as natural speech understanding and emotion recognition in spontaneous speech. Almost all researchers assume that relative syllable and sound length, melody and loudness are highly connected with prominence [2]. The first research goal was to find the most efficient acoustic features for automatic prominent word detection.

A. Relative syllables and sounds length

The relative length of syllables and sounds is an obvious prominence feature. The speaker usually stretches a word if s/he wants to emphasize it. As there is no reliable algorithm of syllable detection for Russian and there is no syllable transcription in CORPRES, the speech corpus of Russian speech used in the present experiment, relative syllable length was not used. Two temporal features were used. The first feature is total word length in milliseconds. The second one is relative sound length within a current word that is expressed by the ratio of sound length within a current word and the length mean for the sound. The means were calculated for sound samples within the speech corpus. These features are possible to calculate as there is phonetic transcription with precise speech sound boundaries and orthographic transcription with precise word boundaries in CORPRES. Thus, there is no need to detect speech segment boundaries automatically, but it should be done by means of the automatic speech recognition in real-life applications.

B. Melodic features

The majority of researchers support the idea that melodic features are the most crucial for speech prominence, i.e. see [8]. In present research, the melodic contour of every word was examined separately. An original recently developed method of melody processing [6] was used. This method showed its efficiency in the system of automatic interpretation of tone unit prosody.

The melodic features were extracted from a preprocessed and smoothed melodic contour. The melodic contour is achieved as a result of automatic pitch detection system that has been developed earlier at the Department of Phonetics of Saint-Petersburg State University. The goal of preprocessing

This work was supported in part by the "Research Potential Development of Higher Education Institutions" Federal Program ("Automatic Interpretation of Prosody" Project)

is to eliminate microprosodic events and to get a smoothed melodic contour. This allows to get rid of calculation errors occurring within microprosodic events. Automatic melody preprocessing consists of the following four steps:

1) detection of voiced parts of the speech signal;
2) pitch detection within voiced parts;
3) microprosody and laryngalization rule-based processing;
4) melodic contour smoothing based on the algorithm of moving average.

The following melodic features were selected for prominence detection based on the analysis of other research and solutions as well as a series of experiments: maximum, minimum, mean and standard deviation of the fundamental frequency within a word. The rate of fundamental frequency change is also taken into account to model not only the fundamental frequency itself but the extent of its change as well. Thus, maximum, minimum, mean and standard deviation of the fundamental frequency change within a word were applied for this purpose. These features are applied based on the idea that the change of fundamental frequency is higher within a prominent word.

C. Intensity of speech signal and its spectrum

Speech loudness also correlates with prominence. Speech loudness corresponds with speech signal intensity or, more precisely, with its spectrum intensity. Meanwhile there are two main ways of modeling speech loudness. The first one is a calculation of spectrum intensity within certain, most significant frequency bands. The second is a calculation of speech signal intensity. The latter does not require FFT calculation that leads to much faster feature extraction. It is worth saying that the efficiency of signal based features is equal to the efficiency of spectrum based ones. The following features were used to express word loudness: maximum, minimum, mean and standard deviation of speech signal intensity within a current word.

The discrete extraction of signal intensity features is used. The speech signal within a word is divided into processing windows. Window length is 10 ms and window step is also 10 ms, that is windows do not overlap. The signal intensity within a processing window is a mean of signal amplitude values within the window. Thus, we calculate a single signal intensity value every 10 ms. This value array is used to obtain the features listed above.

D. D. General overview of proposed acoustic features

The acoustic features described above express three different aspects of speech prosody: melodic, dynamic and temporal description of prosody. They are independent at the first glance, that is changing one of them does not influence the others. But that is not really true. Changing one of them will influence the perception of the others. The perceptual characteristics of speech are more important than its real acoustic characteristics when one considers speech prominence. That is why all the features should be taken into account. For example, an increase in signal intensity leads to an increase in perceived pitch and an increase in fundamental frequency leads to an increase in perceived loudness [3]. It is quite obvious how to extract temporal and dynamic information. The task of key melodic feature extraction still requires a solution because it is not obvious what melodic features are the most important for automatic prosody interpretation. However, to make the feature extraction process consistent, it was decided to use the following list of features to model word prominence:

1) total word length, relative sounds length;
2) maximum, minimum, mean and standard deviation of the fundamental frequency within a word;
3) maximum, minimum, mean and standard deviation of the fundamental frequency change within a word;
4) maximum, minimum, mean and standard deviation of the speech signal intensity within a word.

The use of these acoustic features is based on the following reasons. First of all, it is well known that short words are rarely prominent, but the melodic contour within them usually changes greatly and rate of F0 change is a correlate of prominence. Thus, the use of word length as an acoustic feature helps to detect such words as non-prominent. The melodic and dynamic features are designed following the same principles: maximum, minimum, mean and standard deviation are calculated. This allows to estimate the range and variance of prosodic features. On the other hand, it allows to examine how words differ from each other, especially by statistical measures such as mean and standard deviation. These prosodic features are essential and almost all other features are based on them.

III. Statistical Processing of Prosodic Data

The choice of statistical data processing and acoustic modeling method that allows to achieve the best efficiency of automatic prominent word detection is no less crucial than the choice of acoustic features. The main statistical framework applied in speech technology at the moment is hidden Markov models (HMM) that would be perfect for the solution of this task within an automatic speech recognition system. HMM is the best choice when one needs to reveal a context dependency of objects or a dependency of certain objects appearing next to preceding objects. The solution of current task does not require this; it is possible to make an assumption about context independence of word prominence from the prominence of preceding words. Thus it was decided to use another method. It seems reasonable to use classification and regression trees (CART) to detect prominent words as it was done in [5]. CART is an effective classification method when classified objects are independent from each other. Besides that, CART allows to define a relative significance of features for a classification task. It is especially valuable from scientific point of view and allows us to develop, test and apply acoustic features that are more and more effective and reliable for a task of modeling and detecting prominent words. Usually the entropy is used as a splitting criterion in a CART framework. However, it has been decided to use probability of prominent words as a splitting criterion in the current system. There are two

TABLE I
THE EVALUTION RESULTS WITH DIFFERENT SETUPS

Experiment	Efficiency	Precision	Recall
SpInd	84,2	83.3	79.1
SpDep	77.1	81.2	73.4
Male Voices	89.7	90.4	80.1
Female Voices	87.3	88.2	78.8

reasons for that. The first one is the fact that when entropy is calculated all classes are supposed to be equally probable, but the number of prominent words is 4 times smaller than the number of non-prominent words. In case of using entropy this could lead to the situation when there are objects of different classes in all CART leafs: many non-prominent words and several prominent words. The other reason is that there are just two classes in this case: prominent and non-prominent words. Thus, uncertainty degree is unambiguously defined by a probability of one class. The experimental results showed that the probability of prominent words is a much more efficient splitting criterion than entropy.

IV. EXPERIMENTS

A. Experimental Data Description

All the experiments were carried out with the CORPRES (Corpus of Russian Professionally Read Speech) corpus. It consists of recordings from 8 speakers, four men and four women. It contains 25 hours of fully annotated speech [7], three hours per each speaker. The corpus contains the following annotation levels:

1) pitch marks – boundaries of fundamental frequency periods;
2) phonetic events labeling – boundaries and labels of phonetic events;
3) phonetic transcription – boundaries and labels of speech sounds;
4) orthographic transcription – boundaries and labels of words;
5) prosodic transcription – boundaries and labels of tone units and pauses.

There are 211 383 running words in the fully annotated part of the speech corpus and 40 547 of them were labeled by experts as prominent.

B. Experimental Results

Cross-validation has been applied for efficiency measurement during experiments. This method is widely used in cases of lack of data and non-uniform data. Prominent words are non-uniformly distributed over the speech corpus and non-prominent words would be considered as a more probable class. The cross-validation allows to avoid that. It has been decided to use the same efficiency metrics as the ones used in search tasks for the task of prominent word detection can be considered a search task. These are error-rate, precision and recall. A series of experiments was held to estimate the efficiency of automatic detection of prominent words using the above mentioned prosodic features and statistical classifier.

A series of experiments was held:

1. Speaker Independent (SpInd): All data were uniformly devided into 10 parts, i.e. the recordings of every speaker were divided into 10 parts. Nine parts were used as training data and one part was used as test data, thus there were about 22.5 hours of training data and 2.5 hours of test data.

2. Speaker Dependent (SpDep): The purpose of the experiment was to evaluate the efficiency of the system when the recordings of one speaker are used as training data and the recordings of other speakers are used as test data. The data were divided in the following way: the data from 7 speakers (about 22 hours) were used for training and the data from 1 speaker (about 3 hours) were used for evaluation.

3. The last experiment was intended to evaluate the system with gender dependent data. First of all, data from male speakers were separated from data from female speakers. Each part of the data was divided into training data (9/10 of data, about 11.25 hours) and test data (1/10 of data, about 1.25 hours). Thus, four different experiments were held and the results are presented in Table I. The results in the table show several interesting tendencies. The results are much better for the gender dependent system then for gender independent. It might be caused by the significant differences in pitch between female and male voices. This proves the concept that pitch plays a major role in prominence detection.

Another conclusion is that SpInd yields better results then SpDep. This is probably due to the fact that training and test data within SpInd experiment included data from all speakers, while in SpInd experiment training data exluded the data of the test speaker. This shows that the data from seven speakers was not enough to train a speaker independent system and to predict the excluded speaker efficiently.

An overall efficiency of 84.2 % was achieved for speaker independent task. Table II shows the comparison of this result against the results achieved by other researchers in speaker independent systems.

The empty cells in the table mean that the authors did not present precision and recall results in their papers. As one can see, the efficiency of the current system is not the best one, but not the worst. However, it is worth highlighting that an amount of experimental data is by two orders of magnitude greater than the amount of data used to test other systems. Thus, the experimental results can be considered as positive and efficient enough to be a baseline for further research in the field of automatic detection of prominent words.

TABLE II
THE EFFICIENCY OF THE AUTOMATIC PROMINENT WORD DETECTION AS COMPARED TO SIMILAR RESULTS ACHIEVED BY OTHER RESEARCHERS

Research	Language	Amount of Data	Efficiency	Precision	Recall
Brenier et. al. [2]	English	2906 words	87,1		
Kroul [4]	Czech	2160 words	91.1		
Tamburini [9]	Italian	4780 syllables	82.5	75.6	77.9
Wang and Narayanan [10]	English	3247 words	76.2	82.1	73.4
Current System	Russian	211383 words	84.2	83.3	79.1

V. CONCLUSION

The paper presented results of the research dedicated to automatic detection of prominent words. Algorithms of prosodic features extraction were developed during this research. Three types of prosodic features were used: melodic, dynamic and temporal. CART with modified splitting criterion was used as a statistical classifier. The efficiency of the developed system was tested in a series of experiments. The efficiency of 84.2 % was achieved, that which is comparable to other research in this field. The undisputable advantage of this system is that it is the first such system of the kind that has been developed for the Russian language and it could undoubtedly be used within automated annotation of speech corpora modules and automatic speech recognition systems.

REFERENCES

[1] M. E. Beckman and J. J. Venditti "Tagging Prosody and Discourse Structure in Elicited Spontaneous Speech," *in Proceedings of Science and Technology Agency Priority Program Symposium on Spontaneous Speech,* Tokyo, Japan, 2000, pp. 87–98.

[2] J. M. Brenier, D. M. Cer, D. Jurafsky "The Detection of Emphatic Words Using Acoustic and Lexical Features," *in Proceedings of International Conference on Speech Communication and Technology 2005,* Lisbon, Portugal, 2005, pp. 3297–3300.

[3] H. Fletcher and W. A. Munson "Loudness, Its Definition, Measurement, and Calculation," *The Journal of the Acoustical Society of America,* vol. 5, 1933, pp. 82–108.

[4] M. Kroul "Automatic Detection of Emphasized Words for Performance Enhancement of a Czech ASR System," *in Proceedings of SPECOM 2009,* St. Petersburg, Russia, 2009, pp. 470–473.

[5] A. Rosenberg and J. Hirschberg "Detecting Pitch Accents at the Word, Syllable and Vowel Level," *in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics Companion Volume: Short Papers,* Colorado, USA, 2009, pp. 81–84.

[6] P. Skrelin and D. Kocharov "Avtomaticheskaja obrabotka prosodicheskogo oformlenija viskazivanija: releventnie prosodicheskie priznaki dla avtomaticheskoj interpretatsii intonatsionnoj modeli," *in Trudi tretiego mezhdistciplinarnogo seminara Analiz russkoj rechi,* St. Petersburg, 2009, pp. 41–46.

[7] P. Skrelin, N. Volskaya, D. Kocharov, K. Evgrafova , O. Glotova, and V. Evdokimova "A Fully Annotated Corpus of Russian Speech," *in Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10),* Valletta, Malta, 2010, pp. 109–112.

[8] B. M. Streefkerk, L. C. W. Pols, L. F. M. Ten Bosch "Acoustical Features as Predictors for Prominence in Read Aloud Dutch Sentences Used in ANNâĂŹs," *in Proceedings of European Conference on Speech Communication and Technology 1999,* Budapest, Hungary, 1999, pp. 551–554.

[9] F. Tamburini "Automatic Prominence Identification and Prosodic Typology," *in Proceedings of Interspeech Conference on Speech Communication and Technology 2005,* Lisbon, Portugal, 2005, pp. 1813–1816.

[10] D. Wang and Sh. Narayanan "An Acoustic Measure for Word Prominence in Spontaneous Speech," *in IEEE Trans. Audio, Speech, and Language Processing,* vol. 15, no. 2, 2007, pp. 690–701.