Statistical Study of Speaker's Peculiarities of Utterances into Phrases Segmentation

Boris Lobanov & Liliya Tsirulnik

United Institute of Informatics Problems
National Academy of Science, Belarus
lobanov@newman.bas-net.by, liliya_tsirulnik@ssrlab.com

Abstract

The report is concerned with the experimental study of the idiosyncrasy of utterance-into-phrase segmentation observed in the speech of a popular Russian TV-anchorman and two TV-news readers. Comparative statistical estimation of relative frequencies of occurrence of pauses of various duration, frequencies of occurrence of phrases and pairs of phrases with a different number of accent units were computed, as well as frequencies of occurrence of phrase break between different consecutive parts of speech. On the basis of the results obtained a stochastic algorithm of personalized utterance-into-phrase segmentation is developed. The algorithm is intended to be implemented to the system of individual voice cloning using a text-to-speech synthesis.

1. Introduction

It is well known that the location of prosodic phrase breaks in natural speech depends on the grammatical categories of the text, the morphological and syntactical construction of utterance, the semantic characteristics of the text, speech style, certain language and extra-linguistic factors.

There are a number of models for predicting the location of prosodic phrase breaks for an utterance to be spoken by a TTS system [1-4]. There are several studies in Russian natural speech, dealing with acoustical and perceptual peculiarities of utterance-into-phrase segmentation [5-7]. Most of the studies performed were oriented to revealing the language-specific peculiarities of utterance-into-phrase segmentation. Similarly, most of the models developed assign the phrase breaks, which a speaker ***might*** assign in natural speech. Our goal in this work is to predict the location of breaks and duration of pauses caused by a ***particular speaker***'s speech idiosyncrasy.

In [8, 9] TTS synthesis was suggested as a computer means for personal voice and speaking manner "cloning", granting the closest possible simulation of acoustic, phonetic and prosodic features of the synthetic speech. The task of cloning is that of preserving, as fully as possible, the personal acoustic peculiarities of the voice, the phonetic peculiarities of the pronunciation of segmental sounds and the individual prosodic features, i.e. the individual peculiarities of the tonal, rhythmical and dynamic ordering of speech.

The primary objective of the research is the verification of differences in relative frequencies of occurrence of pauses of various duration, relative frequencies of occurrence of phrases and pairs of phrases with a different number of accent units (AU), and possibility of occurrence of prosodic phrase breaks between different consecutive parts of speech for different speakers. The next goal is to create stochastic algorithm useful for cloning the individual manner of utterance into phrase segmentation based on the experimental results thus obtained.

2. Experimental procedure

2.1. Speech corpora

The experimental study of individual peculiarities in utterances into phrases segmentation effected by a popular Russian TV-anchorman and two TV-news readers is based on the speech corpora obtained from audio recordings of their TV programs. The corpus for each speaker contains about 15-20 minutes of recordings that corresponds to about 100-130 sentences for different speakers or more than 1,000 orthographic words in the tapescript.

2.2. Speech corpora and tapescript processing

The procedure of speech corpora and a tapescript processing is shown on fig. 1.

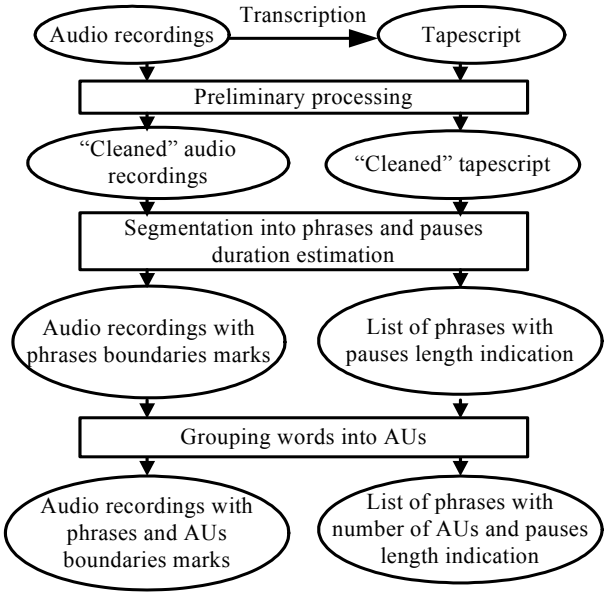


Figure 1: Procedure of speech corpora and a tapescript processing

The audio recordings were initially transcribed verbatim by auditory analysis. On the basis of the verbatim thus obtained the final transcript was composed by removing the mispronounced words and sounds. The recordings were also

cleaned from fragments with all sorts of interference (breathing, noises, music, etc).
The next steps of processing – segmentation into phrases and grouping words into accentual units (AUs) – was performed by the experienced phoneticians.
A phrase is taken to mean a prosodically separate piece of utterance or an entire utterance. Phrases boundaries were marked out in the transcripts and audio recordings by consecutive auditory analysis. The final decision about the phrase boundary was taken on the basis of several features, such as breath-pauses, complete realization of a selected specific intonation type, specific dynamic contour (sound amplitude envelope) and rhythmic structure (speech sound duration). When analyzing audio recordings into phrases, punctuation marks and other formal clues in the tapescripts were also taken into account.
After the assigning phrase boundaries, each phrase was consecutively analyzed aurally and marked out in the following way. Each stressed word within a phrase was marked with primary (+) or secondary (-) stress, respectively. The words with no stress were joined with the neighbouring stressed words thus forming phonetic words. Then the adjoining words with weak stresses and strong stresses were united into AUs in such a way, that each AU has one strong stress.Then AU-boundaries were marked out in the script.
Given in Table 1 is an illustration of a tapescript broken down into phrases with word-stress placement for a fragment of TV-anchorman natural speech following its auditory analysis. Forward slash (/) marks the AUs boundary.

Table 1: *List of phrases with number of AUs and pauses duration indicaors*

N°	Accents and AUs marked text	Number of AUs	Pauses duration (ms)
1	/Здра+вствуйте/,	1	200
2	/дороги+е/ /люби+тели/ /путеше+ствий/.	3	1100
3	/Сего+дня/	1	75
4	/мы+/ /отпра+вимся/ /сва+ми/	3	50
5	/вФинля+ндию/,	1	750
6	/и+/,	1	150
7	/мне+/ /ка+жется/,	2	0
8	/что- э+то/ /путеше+ствие/ /бу+дет/ /длявас+/	4	400
9	/интере+сным/,	1	0
10	/поско+льку/	1	400
11	/путеше+ствие/ /э+то/ /нето+лько/ /впростра+нстве/,	4	50
12	/но+/	1	120
13	/и- вовре+мени/.	1	650
14	/Мы+/ /расска+жем/ /в+ам/	3	120
15	/о+/ /стари+нных /	2	900
16	/фи+нских/ /крепостя+х/.	2	400

The result of the processing is a list of phrases with specifying a number of AUs and pause duration.

3. Experimental results and discussion

The quantitative characteristics of idiosyncrasy of prosodic phrasing observed in the speech of three speakers included: relative frequency of occurrence of pauses of various duration, relative frequencies of occurrence of phrases and pairs of phrases with a different number of AUs, and possibility of phrase break appearance between different consecutive parts of speech. The result of the first three characteristics estimation are shown in figures 2-4, namely:
in figure 2 - relative frequencies of occurrence of pauses of various duration for different speakers;
in figure 3 - relative frequency of occurrence of phrases with a various number of AUs;
in figures 4 - relative frequency of occurrence of pairs of phrases with a various number of AUs.

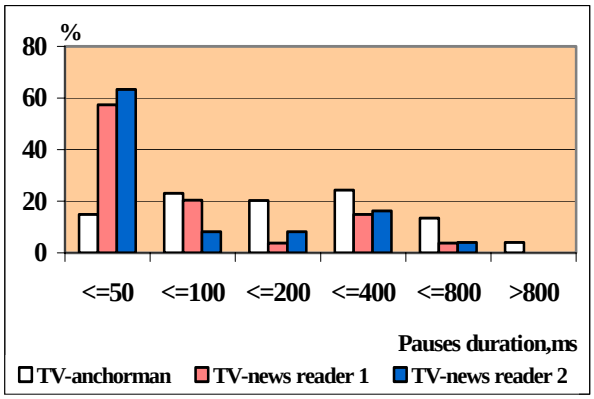


Figure 2: *Relative frequency of occurrence of pauses of various duration.*

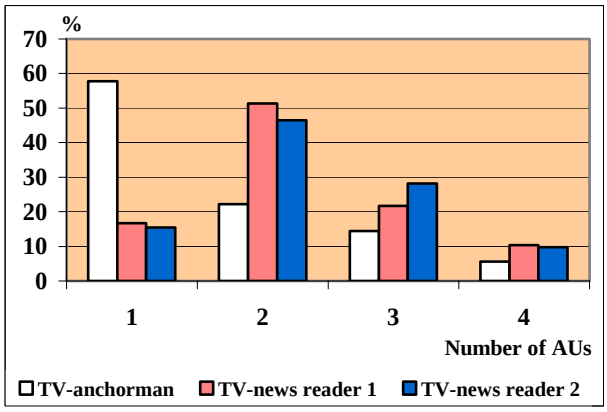
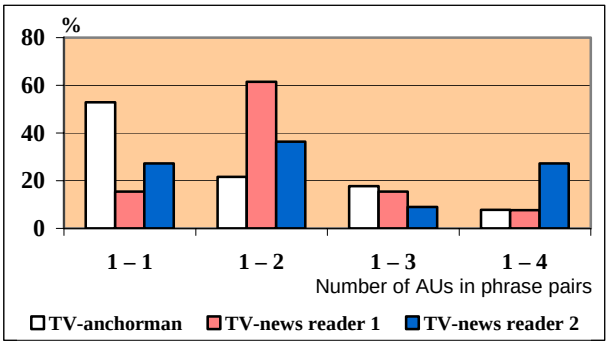
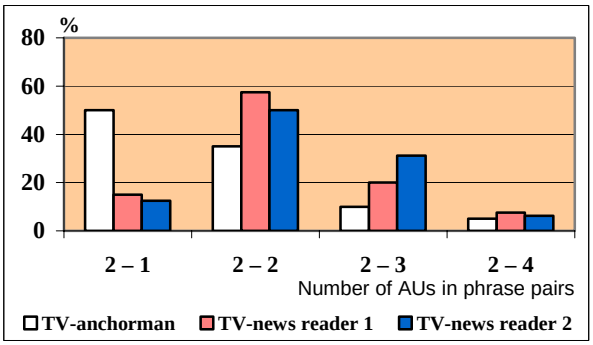


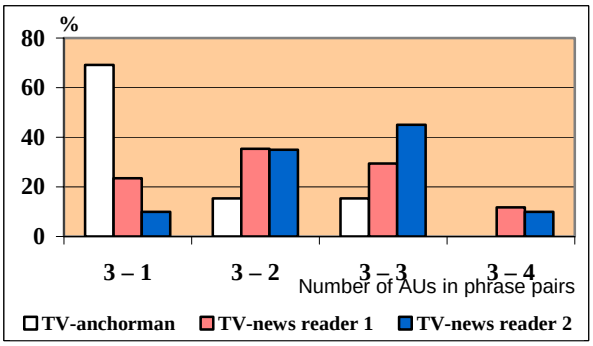
Figure 3: *Relative frequency of occurrence of phrases with a various number of AUs.*



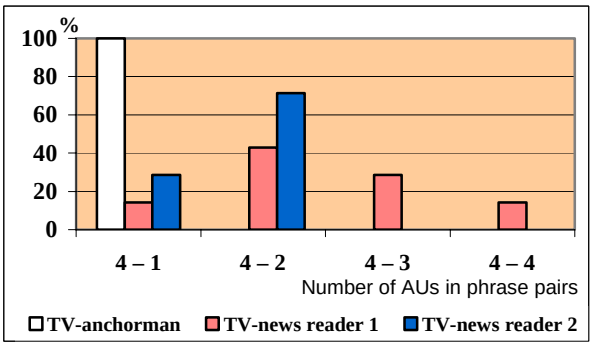
a)



b)



c)



d)

Figure 4: *Relative frequency of occurrence of pairs of phrases with a various number of AUs with each pair beginning with a) a one AU phrase, b)* a *two AU phrase, c) a three AU phrase, d) a four AU phrase.*

Referring to Fig. 2, it can be seen that the distribution of inter-phrase pause duration is rather even only in TV-anchorman, the other two speakers – TV-news readers – clearly showing the maximum frequency for minimum-duration phrase pause. Alternatively, with reference to Fig. 3, it will be observed that only the TV-anchorman's speech demonstrates a pronounced predominance of one-AU phrases and a comparatively even distribution of two- and three-AUs phrases. The analysis of distributions in Fig. 4 suggests similar conclusions.

The obtained statistical characteristics were united to a double-dimension array, where column and row indices indicate the number of AUs in the first and second phrase of the pair, respectively. The values for relative frequencies of occurrence of phrases with a various number of AUs are placed to the first row (with the column index of 0). The resulting array with corresponding values for the TV-anchorman and generalized values for two TV-news readers, separated by slash, is shown on table 2.

Table 2: *Frequencies of occurrence of pairs of phrases with different number of AUs for TV-anchorman and two TV-news readers*

	1	2	3	4
0	58/16	22/49	14/25	6/10
1	53/21	22/49	17/12	8/18
2	50/14	35/54	10/26	5/7
3	70/17	15/35	15/37	0/11
4	100/21	0/57	0/15	0/7

The last quantitative characteristic – possibility of phrase break appearance between different consecutive parts of speech (POS) does not separated by punctuation mark – was represented by the values of either 0 or 1. The value of 1 was assigned to the pair (p_n, p_m) if the number of corresponding POS pairs in the typescript, separated by phrase break, was appeared in more then 10% cases.

The results for different POS pairs thus obtained were united to a double-dimension array, were column and row indices indicate the first and second POS of the pair correspondingly. The resulting array (of the most frequent POS pairs) with corresponding values for TV-anchorman and generalized values for two TV-news readers, separated by slash, is shown on table 3.

Table 3: *Possibilities of phrase breaks appearance between different consecutive parts of speech for TV-anchorman and two TV-news readers*

	Verb	Adverb	Adjective	Noun	Pronoun	Conjunction	Preposition
Verb	1/1	**1/0**	**1/0**	**1/0**	1/1	1/1	**1/0**
Adverb	**1/0**	0/0	1/1	0/0	0/0	1/1	1/1
Adjective	0/0	0/0	1/1	**1/0**	0/0	1/1	0/0
Noun	1/1	1/1	1/1	**1/0**	**1/0**	1/1	1/1
Pronoun	**1/0**	0/0	**1/0**	**1/0**	0/0	**1/0**	1/1
Conjunction	**1/0**	0/0	**1/0**	**1/0**	**1/0**	0/0	**1/0**
Preposition	0/0	0/0	**1/0**	**1/0**	**1/0**	0/0	0/0

As is evident from table 3, the possibility of phrase break appearance between different consecutive POS diverge considerably for both TV-anchorman and TV-news readers. Significant differences are observed on the possibility of phrase break after pronouns, conjunction and preposition for TV-anchorman that is not typical for TV-news readers.
The obtained statistical arrays (tables 2, 3) are used in a stochastic algorithm of utterance into phrases segmentation.

3. The stochastic algorithm of utterance into phrases segmentation

The algorithm uses the following resources:

- Grammatical vocabulary for Russian. The vocabulary includes a list of word-forms with a specified POS for each word-form.
- The array F(r, s) with frequencies of occurrence of pairs of phrases with different number of AUs (see table 2), where $0 \le r \le 4$, $1 \le s \le 4$. The array elements are natural values in range [0 - 100].
- The array $G(p_n, p_m)$ of possibilities of phrase breaks appearance between p_n and p_m parts of speech, where p_n and p_m are particular parts of speech (see table 3). The array elements are binary values: 0 or 1.

The algorithm input is an orthographic text. The text must preliminarily be transformed to a canonical form, i.e. numbers, formulas, abbreviations and contractions must be converted to a text form.

Algorithm description:

1. The phrase boundary is assigned after each punctuation mark in a text.
2. For each of the framed phrases phr_i, where i is an order number of a phrase, satisfying the condition 1

$$\left|\text{phr}_{\,i}\right| \ge 5 \qquad (1),$$

where $|phr|$ is a power of phrase – number of words in a phrase, having at least one vowel, the steps 3 – 6 are performed.
3. Each word w_j of phr_i is marked by POS tag (from grammatical vocabulary). The sequence $\{p_j\}$ of POS tags is composed, where each p_j corresponds to a word of the input text.
4. For each (p_j, p_{j+1}) pair the sequence $\{G(p_j, p_{j+1})\}$ is composed from corresponding values of $G(p_n, p_m)$.
5. The elements of $\{G(p_j, p_{j+1})\}$ are multiplied to the value f of $F(r, s)$ array, where f is determined by the formula:

$$f = \begin{cases} F(0, j), & i = 1 \\ F(t, j), & i > 1 \quad \text{and} \quad \left|\text{phr}_{i-1}\right| = t \end{cases} \qquad (2)$$

Then the index l of the maximum element of obtained sequence is founded, so that

$$l = \text{Arg} \quad (\text{Max}_j\{G(p_j, p_{j+1}) * f\})$$

The phrase break is assigned after the l-th word of the phrase phr_i.
6. From the phrase phr_i the next phrase phr_{i+1} is framed after the assigned phrase boundary. While $|phr_{i+1}| \ge 5$, the steps 3-6 are performed.

The proposed stochastic algorithm is intended to be used as part of a multi-voice and multi-lingual TTS-synthesis system [10] under development.

4. Conclusion

The research was concerned with regularities of utterance-into-phrase segmentation that is only one and, perhaps, not the most vital part of prosodic phenomena, which carry the speaker's speech idiosyncrasy. The issues of computational analysis, employment and quantitative estimation of the basic set of speech prosodic features have remained beyond the scope of the present study. They include pitch (fundamental frequency or tone contour – F0), dynamic contour (sound amplitude or volume – A) and sound duration (rhythmic pattern – T). The authors are planning further research into the inventory of these characteristics in terms of their speaker-specific properties.

5. Acknowledgment

This paper was supported by the European Commission under grant INTAS Ref. No 04-77-7404. The authors wish to express their thanks for the support.

6. References

[1] Hirschberg, J.; Prieto, P., 1995. Training intonational phrasing rules automatically for English and Spanish text-to-speech. *Speech Communication* 18, 281-290.
[2] Pavlova E.; Pavlov Y.; Sproat R.; Shih Ch.; van Santen J., 1997. Bell laboratories Russian text-to-speech system. Proc. *Eurospeech'97*. Rhodes-Greece, 2451-2454.
[3] Black, A.; Taylor, P., 1998. Assigning phrase breaks from part-of-speech sequences. *Computer Speech and Language* 12, 99-117.
[4] Read, J.; Cox, S., 2005. Stochastic and syntactic techniques for predicting phrase breaks. Proc. *Interspeech 2005*. Lisbon, Portugal, 3233-3236.
[5] Ventsov, A.; Kasevitch V.; Slepokurova N., 1993. Perceptual segmentation of sounding text. In *Issues in Phonetics.* Moscow: Prometey, 242-273 (in Russian).
[6] Vol'skaya N.B., 2001. Peculiarities of intonation and phrase segmentation of interrogative utterances of different length. *Inl. Conf. "100 years of experimental phonetics in Saint-Petersburg"*. S.-Petersburg: Philological department of SPSU, 54-57 (in Russian).
[7] Hitina M.V., 2004. Peculiarities of segmentation of specific spoken language discourse. In *Language and Speech: problems and solutions*. Moscow: MAX Press, 270-285 (in Russian).
[8] Lobanov, B.; Karnevskaya, E., 2002. TTS-Synthesizer as a Computer Means for Personal Voice "Cloning". In *Phonetics and its Applications*. A. Braun, Herbert R. Masthoff, (eds.). Stuttgart: Steiner, 445-452.
[9] Lobanov, B.; Tsirulnik, L., 2004. Phonetic-Acoustical Problems of Personal Voice Cloning by TTS. Proc. *Int. Conf. SPECOM'2004*. St.-Petersburg, Russia, 565 – 571.
[10] Hoffmann R., Shpilewsky E., Lobanov B., Ronzhin A., 2004. Development of multi-voice and multi-language Text-to-Speech (TTS) and Speech-to-Text (STT) conversion system (languages: Belorussian, Polish, Russian). Proc. *Int. Conf. SPECOM'2004*. Saint-Petersburg, Russia, 657-661.