Study of Idiosyncrasy of Syntagmatic Segmentation for Personal Speaking Manner Cloning by TTS

Boris Lobanov, Liliya Tsirulnik, Anatoly Fiodorov

United Institute of Informatics Problems,
National Academy of Sciences, Belarus
6 Surganov Str., Minsk 220012, Belarus
lobanov@newman.bas-net.by

ABSTRACT

The report is concerned with the experimental study of the idiosyncrasy of syntagmatic segmentation observed in the speech of the one popular TV-anchorman and two TV-news speakers. The study was carried out within the framework of the on-going research into the "cloning" of individual voice quality and speaking manner. The audio recordings were initially transcribed by a skilled phonetician who marked out primary and secondary stresses, identified syntagmatic boundaries and its intonation types. Comparative statistical estimation of the number of accentual units in syntagms, duration of inter-syntagm pauses, probabilities of syntagm combinations, etc, were computed. The results of the investigation have been applied to the system of individual voice cloning using a text-to-speech synthesis system.

1. Introduction

In [1, 2] text-to-speech synthesis was suggested as a computer mean for personal voice and speaking manner "cloning", granting the closest possible simulation of acoustic, phonetic and prosodic features of the synthetic speech. A task of the cloning is set 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 organization of speech. The results of the experimental study of individual peculiarities of speech segmentation into syntagms effected by the one popular TV-anchorman and two TV-news readers is based on the audio recordings of their TV programs. The paper presents some statistical results by comparative analysis of the individual peculiarities of speech segmentation into syntagms. The total space of sound files for each speaker approximated 25 MB that amounted to about 1,000 orthographic words in the typescript. Our primary objective of the investigation is to figure out fundamental algorithms for cloning individual manner of syntagmatic segmentation. The next goal will be the verification of differences in the others speech prosodic characteristics for several speakers.

2. Experimental Procedure

2.1. Preliminary processing of audio recordings and transcripts

The audio recordings were initially transcribed verbatim by auditory analysis. The doctoring of the original script and its audio recordings was based on the transcripts thus obtained. Mispronounced words and sounds were removed from the recordings as well fragments with all sorts of interference (breathing, noises, music, low-volume words, etc). Adjustment of acoustic parameters on the audio recordings was carried out in terms of sounds level and sound frequency-response equalization when needed.

2.2. Segmentation of audio recordings and transcripts into syntagms

Syntagm is taken to mean an intonationally separate piece of utterance or an entire utterance. Syntagmatic boundaries were marked out in the transcripts and audio recordings by consecutive auditory analysis. The final decision about the syntagmatic boundary was taken on the basis of several features, such as, breath-pauses, complete production of an available intonation type of syntagm, specified dynamic contour (sound amplitude envelope) and rhythmic structure (speech sound durations). When analyzing audio recordings into syntagms, punctuation marks in the typescripts along with other formal clues in the script were also taken into account.

2.3. Prosodic marking of syntagms

After segmentation of audio recordings into syntagms, each syntagm was consecutively analyzed aurally and marked out in the following way. Each word within a syntagm was stress-marked, i.e. stress placement and its type: strong (+), weak (-) or no stress. The words with no stress were grouped in the typescripts with one of the adjoining words into one phonetic word. Then the words with weak stresses were united into a single accentual unit (AU) containing a word with a strong stress, thus AU-boundaries were marked out in the script.

After that, in the course of auditory analysis, every syntagm was assigned a specific intonation subtype along with the number of AUs in a syntagm (for example, 2-C means "two-AUs syntagm with 'incomplete' intonation contour") as well as the duration of an inter-syntagm pause. The following labels were used for marking intonation types: C (comma) means "incomplete intonation", P (point) means "complete intonation", Q–question, E–exclamation.

Given in Table 1 is an illustration of transcript segmentation into syntagms and word-stress placement for a fragment of file 1 (speaker – TV-anchorman) following its auditory analysis. Forward slash (/) marks the boundary of a AUs.

Table 1. *Syntagms intonation type and duration of pauses*

No of the sintagm	Accents marked text	Intonation type	Pauses duration (ms)
1.	/Здра+вствуйте/,	1-C	200
2.	/дороги+е/ /люби+тели/ /путеше+ствий/.	3-P	1100
3.	/Сего+дня/	1-C	75
4.	/мы+/ /отпра+вимся/ /сва+ми/	3-C	50
5.	/вФинля+ндию/,	1-C	750
6.	/и+/,	1-C	150
7.	/мне+/ /ка+жется/,	2-C	0
8.	/что- э+то/ /путеше+ствие/ /бу+дет/ /длявас+/	4-C	400
9.	/интере+сным/,	1-C	0
10.	/поско+льку/	1-C	400
11.	/путеше+ствие/ /э+то/ /нето+лько/ /впростра+нстве/,	4-C	50
12.	/но+/	1-C	120
13.	/и- вовре+мени/.	1-P	650
14.	/Мы+/ /расска+жем/ /в+ам/	1-C	120
15.	/о+/ /стари+нных /	2-C	900
16.	/фи+нских/ /крепостя+х/.	2-P	400
17.	/Крепостя+х/ /за+мках/.	2-P	2400

3. Statistical characteristics of syntagmatic segmentation

Statistical analysis of the experimental study of audio recordings was aimed at obtaining quantitative characteristics useful in terms of personalization of synthesized speech. These characteristics included: relative frequencies of occurrence for syntagms with different number of AUs, frequency of pauses of various duration, frequency of occurrence for pairs of syntagms with various number of AUs. In figures 1-3 are listed the essential results of the statistical analysis of test material in the course of investigation.

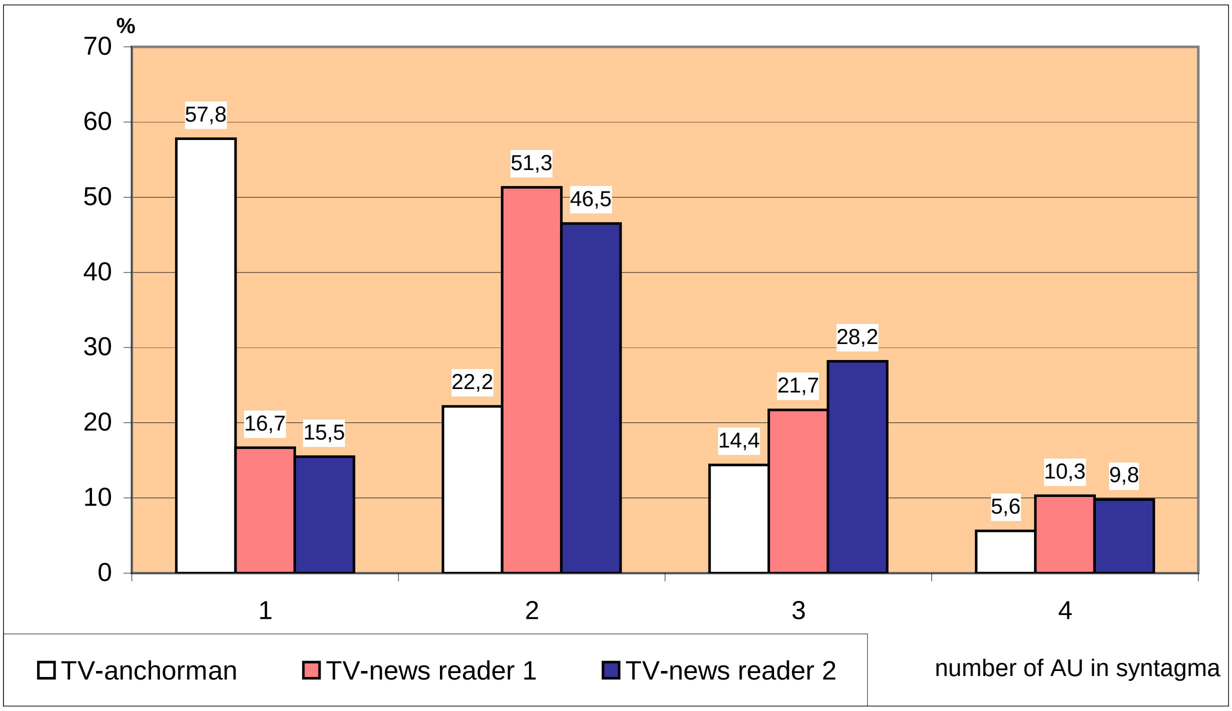


Fig.1. *Relative frequency of occurrence of syntagms with various number of AU. Speakers: TV-anchorman , TV-news reader 1(male), TV-news reader 2(female)*

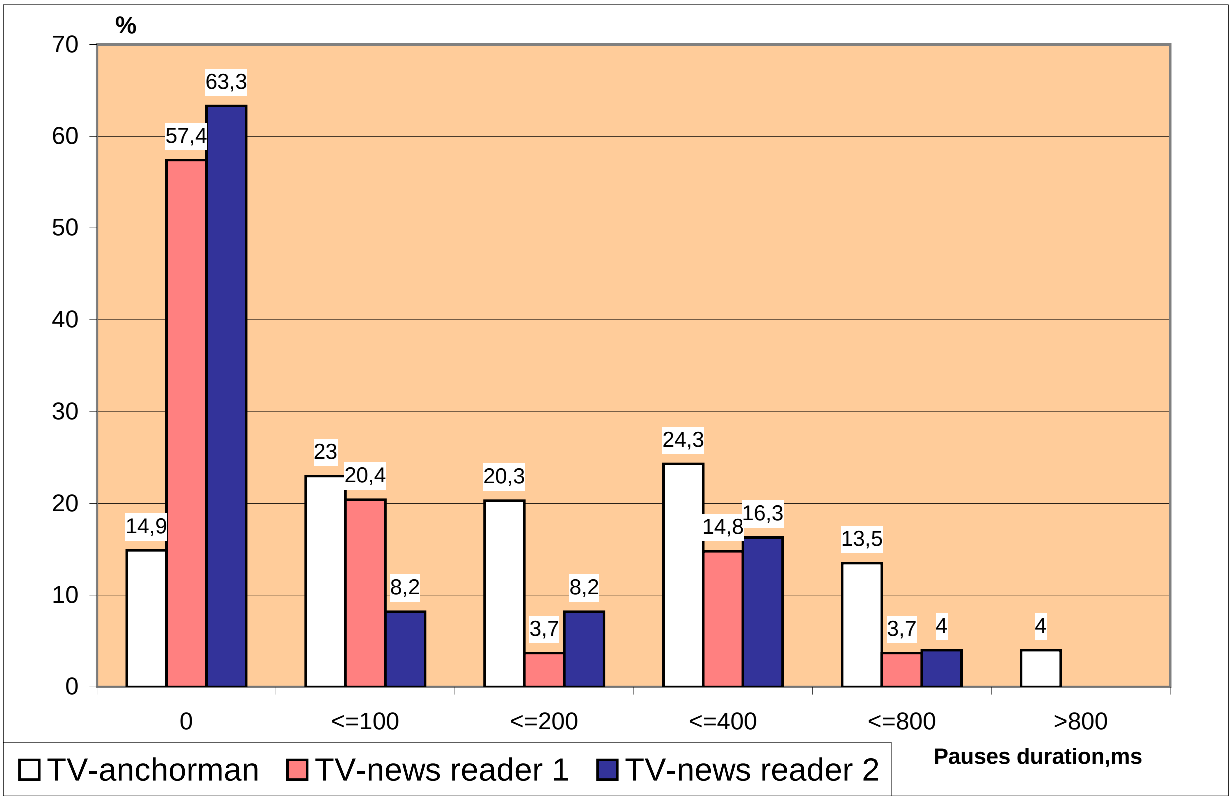
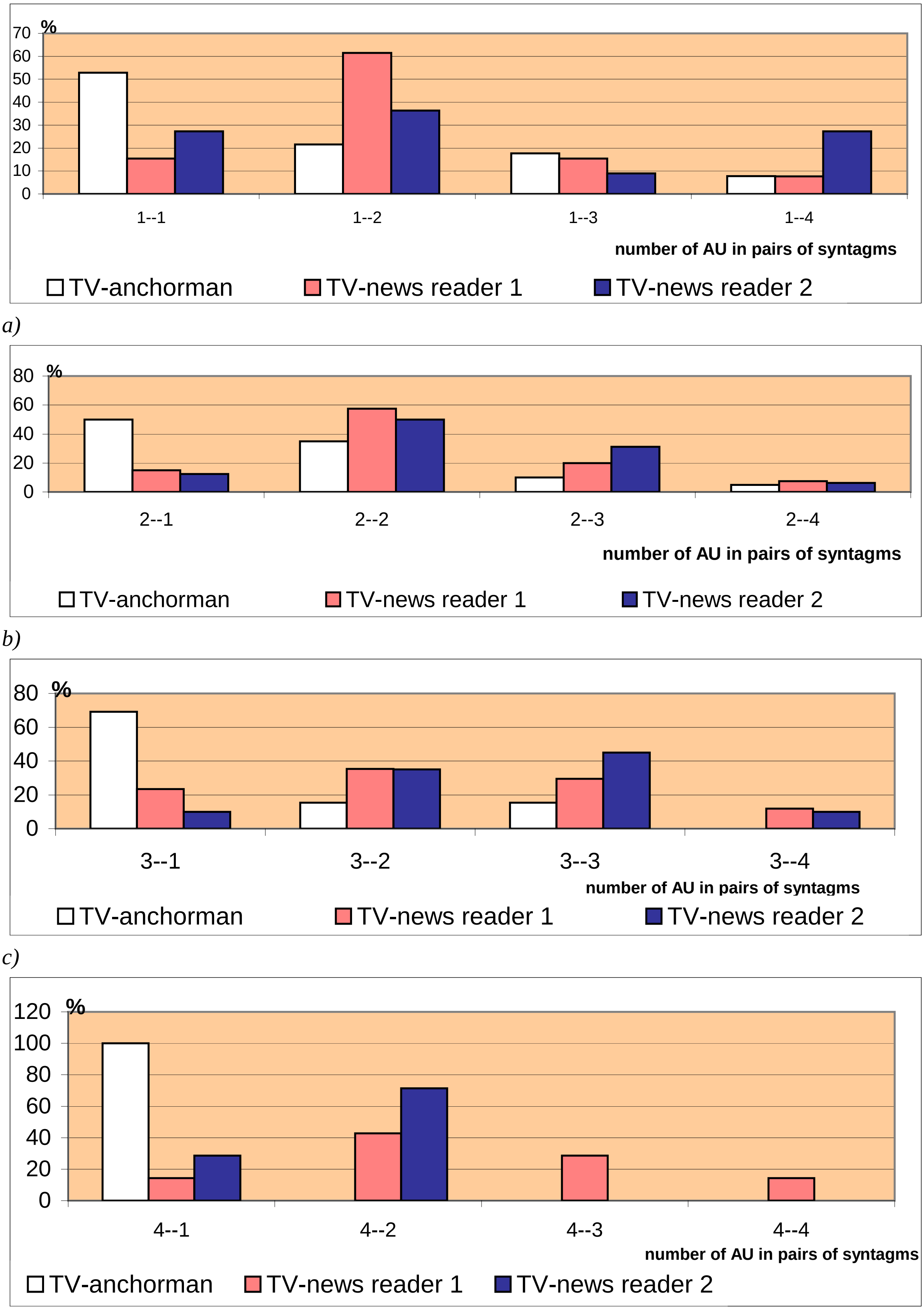


Fig.2. Relative frequency of occurrence of pauses of various duration (between syntagms with incomplete intonation contours). Speakers: TV-anchorman , TV-news reader 1(male), TV-news reader 2(female)



a)

b)

c)

d)

Fig.3. Relative frequency of occurrence of pairs of syntagms with various number of AUs within which a pair begins with a) one-AU syntagm; b) two- AU syntagm; c) three- AU syntagm; d) four- AU syntagm. Speakers: TV-anchorman , TV-news reader 1(male), TV-news reader 2(female)

4. Discussion of the results

Given in Fig.1-3 statistical characteristics of the peculiarities of syntagmatic segmentation of oral speech for three speakers are most indicative of the marked idiosyncrasy of TV anchorman 's speech as compared to the other two speakers. Referring to Fig. 1, it will be observed that only TV-anchorman speaker demonstrates a pronounced predominance of one-AU syntagms and comparatively even distribution of two- and three-AU syntagms. Alternatively, with reference to Fig. 2, it can be seen that the distribution of inter-syntagm pause durations are distributed rather evenly only in speaker TV-anchorman speaker, the other two speakers clearly showing the maximum frequency for zero-duration syntagms. The analysis of distributions in Fig. 3a,b,c,d suggests similar conclusions.

The resulting experimental characteristics of distributions are proposed to implement for individual speech cloning within a stochastic model of syntagmatic segmentation in question used in text-to-speech synthesizers. The statistical investigation of syntagmatic segmentation is far from being complete. Segmentation regularities conditioned by some morphological and syntactic markers in the text, such as specific parts of speech and punctuation marks, which allow or rule out syntagmatic boundaries have not been given adequate consideration so far. The experience gained through the auditory analysis of audio recordings will be actively applied to the development of computerized segmentation of speech audio files produced by different speakers.

The research was concerned with merely one and not most vital part of prosodic phenomenon which carry speaker's speech idiosyncrasy, namely, regularities of syntagmatic speech segmentation. 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.

Acknowledgements

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.

References

[1] Lobanov, B. and Karnevskaya, E., 2002. TTS-Synthesizer as a Computer Means for Personal Voice "Cloning"(On the example of Russian) [In:] Braun, Angelika / Masthoff, Herbert R. (eds.) (2002): Phonetics and its Applications. Festschrift for Jens-Peter Koester on the Occasion of his 60th Birthday. Stuttgart: Steiner. (Zeitschrift fur Dialektologie und Linguistik. Beiheft 121), pp. 445-452

[2] Lobanov, B. and Tsirulnik, L. Phonetic-Acoustical Problems of Personal Voice Cloning by TTS [In:] Proceedings of the 7th International Workshop "Speech and Computer" – SPECOM'2004. St.-Petersburg, Russia, pp.565 – 571