EXPERIMENTAL RESULTS ON PROSODY FOR ROMANIAN TEXT TO SPEECH SYNTHESIS

MIRCEA GIURGIU

Abstract. This paper presents several experimental results on the study of the factors that affect the prosody in Romanian in order to create the basis of a computational model for prosody description in a system for Text To Speech (TTS) synthesis. Different manifestation levels of prosodic phenomena together with their importance for TTS are presented. The experiments reported here, mainly take into consideration the linguistic aspect of prosody in a qualitative and quantitative manner: the influence of the accent at word and sentence level, the intonation, the rhythm and the speech rate. The evaluated parameters are: fundamental frequency, the intensity and the duration. Such parameters are going to be predicted as well as possible by a linguistic computational model which will be developed in the future.

1. INTRODUCTION

Most of the Text To Speech (TTS) systems have two main components: a text-to-segment conversion components and a segment-to-speech synthesiser. The first component converts the input text into specific encoded strings of synthesis segments. Such synthesis segments are different from system to system and they depend on the synthesis approach chosen. The second component converts the segments into actual speech output.

The synthesis approach could be: parametric synthesis or concatenative synthesis. In a parametric synthesis system (Allen *et al.* 1987) the text is first converted into phonological units, which are then divided into several acoustic units. Each acoustic unit is generated on the basis of a sequence of control signals according to the context and the transitional effects. In a concatenative synthesis system (Sato 1992), the acoustic units are previously recorded, stored and recalled in the concatenation process. Our research tasks on TTS are focusing on concatenative speech synthesis.

The first problem in concatenative speech synthesis is the selection of the acoustic units. Some of the existing telecommunications applications are using phrases or words, but this is applicable only if the system requires a high synthesis quality and the set of synthesised messages is very small. For an unlimited TTS of a language it is practically impossible to store all the words, that's why the actual TTS systems are using smaller acoustic units such as: phonemes, diphones, triphones or subphonemic segments.

RRL, **LI**, *3–4*, p. 467–475, Bucureşti, 2006

The phoneme appears to be an attractive linguistic unit for speech synthesis because of the limited number of phonemes in any language. Still, one major reason for not being practically used is that the boundaries between the phonemes usually corresponds to areas that are acoustically volatile. Also, in the synthesis process it appears an unsatisfactory coarticulatory effect (Allen *et al.* 1987).

In the case of diphone concatenation, the acoustic segment captures all the transitional information that is usually present between the phonemes. A diphone is composed of the final portion of one unit and the initial portion of the succeeding unit (Isard, Miller 1986).

Triphones are alternate units that are used in speech synthesis. To generate a CVC syllable, one would require a triphone containing the CV transition, the V portion as well as the VC transition. The extremely large number of triphones is a discouraging factor (Sagisaka, Sato 1985).

The subphonemic segments capture the transitional as well as transient information that is available between consonants and vowels, as well as between vowels and vowels and between other continuant sounds. While capturing such information, one can simultaneously attempt to segment the stationary and non-varying portions of the signals and to economise by suppressing repetitive elements (Bhaskararao, Venkata 1992).

The goal of the reported experiments is to find out what factors have to be considered in the prosody analysis in Romanian as well as to question how a prosodic model would be able to generate the acoustic parameters for a TTS.

2. PROSODIC ASPECTS OF SPEECH

The classic definition of prosody refers to the speech features whose domain is not a single phonetic segment, but larger units of more than one segment, possibly whole sentence. Consequently, prosodic phenomena are often called supra-segmental speech features. They appear to be used to structure the speech flow and are perceived as stress or accentuation, or as other modifications of intonation, rhythm and loudness. There are four principal manifestation levels of prosodic phenomena: a) linguistic level, b) articulatory level, c) acoustic level, d) perceptual level.

a) The linguistic intention level: the speaker can be assumed to employ prosodic coding with a certain intention. This intention can influence both linguistic and paralinguistic expression. By linguistic expression is meant any oral expression using language signs. Paralinguistic phenomena include non-verbal vocalisations that make an utterance to sound angry, urgent or ironic. Examples of linguistic distinctions that tend to be communicated by prosodic means are the question-statement distinction or the semantic emphasis of an element. Systematic knowledge of how these phenomena are used in human speech can be expected to play a significant role in improving the naturalness of the synthetic speech. From linguistic point of view, prosody is generally thought of as relating different

linguistic elements to each other, above all accentuating certain elements of a text, by marking boundaries and by defining transition between words or phrases. Linguistically, the prosodic elements relate either to tone, intonation or accent.

b) The articulatory manifestation level: prosodic phenomena are physically manifested at a series of modifications of articulatory movement. Such phenomena do not result in separate, identifiable articulations. For example, the stressed syllable /ve/ in "*ve*selă" (happy, jubilant) does not involve an articulatory movement distinctive of a more neutral, destressed articulation of the same syllable in "ve*se*lă" (serving dishes). Pertinent physical observations of prosodic manifestations thus typically include variations in the amplitude or air pressure.

c) The acoustic realisation level: it may be observed and quantified using acoustic signal analysis. The main acoustic parameters bearing on prosody are: fundamental frequency, intensity and duration.

d) The perceptual level: it refers to the perceptual reactions to prosodic phenomena and it may be quantified by: pauses, length, pitch/melody and loudness.

In the following experiments priority is given to relating linguistic distinctions to acoustic aspects of prosody.

3. EXPERIMENTAL RESULTS ON PROSODY ANALYSIS FOR ROMANIAN

3.1. The influence of the accent on the prosody parameters at the word level

These experiments considered the evaluation of the relation between the intensity and the fundamental frequency (F0), called also pitch, for stressed syllables in the Romanian words: veselă/veselă (happy/serving dishes), casa/casa (house/to quash), marca/marca (label/to score), factura/factura (invoice/to invoice), lumina/lumina (light/to lighten).

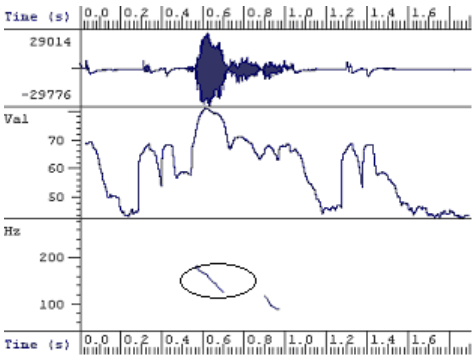

Fig. 1 – Prosody parameters for "casa" (the house)

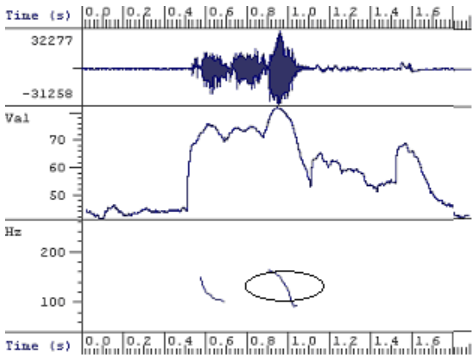

Fig. 2 – Prosody parameters for "casa" (to quash)

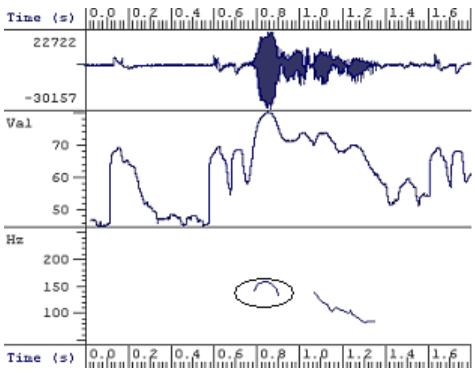

Fig. 3 – Prosody parameters for "veselă" (happy)

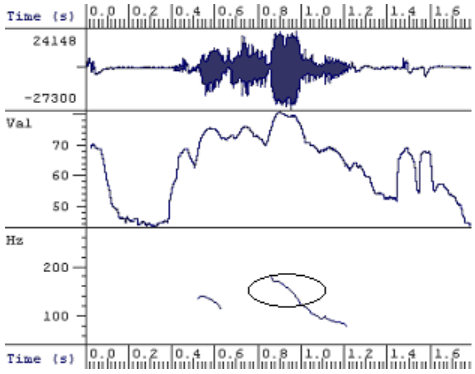

Fig. 4 – Prosody parameters for "veselă" (serving dishes)

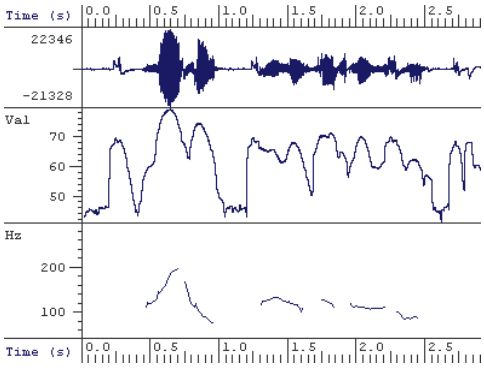

Fig. 5 – The prosody for "Mama vine azi la mine" (My mother comes today to me)

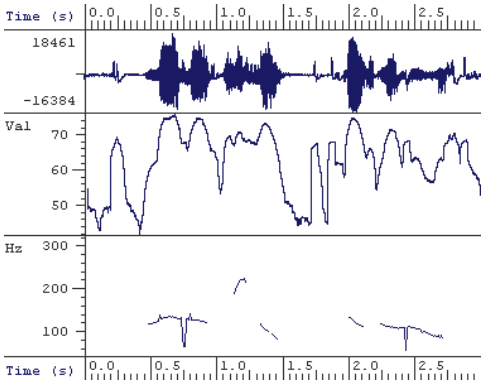

Fig. 6 – The prosody for "Mama vine azi la mine" (My mother comes today to me)

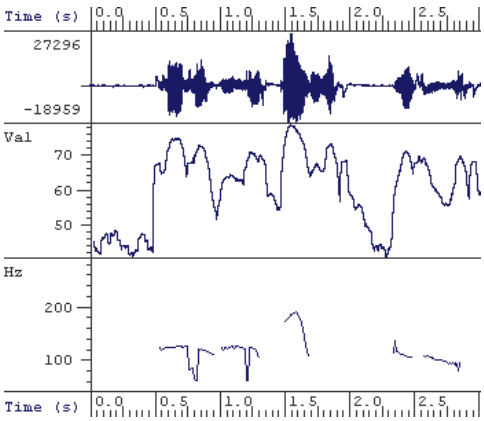

Fig. 7 – The prosody for "Mama vine azi la mine" (My mother comes today to me)

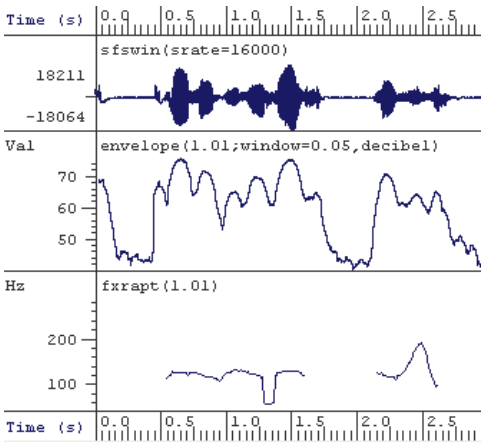

Fig. 8 – The prosody for "Mama vine azi la mine" (My mother comes today to me)

Table I

Acoustic parameters for different prosodic situations (accent) at word level

Words	Syllable	Acoustic parameters					
		Intensity [dB]		Average F0 [Hz]		Duration [ms]	
		stressed	unstressed	stressed	unstressed	stressed	unstressed
casa/casa	*/ca/*	*78*	*73*	*180*	*120*	*230*	*180*
	/sa/	*75*	*68*	*140*	*100*	*220*	*170*
veselă/vesela	*/ve/*	*79*	*72*	*160*	*140*	*190*	*120*
	/se/	*78*	*71*	*130*	*100*	*210*	*140*
marca/marca	*/mar/*	*87*	*79*	*130*	*105*	*290*	*210*
	/ca/	*78*	*71*	*125*	*105*	*210*	*180*

The acoustic realization of stress generally makes use of at least two and often all three acoustic parameters of prosody (fundamental frequency F0, the intensity and duration). The Figures 1, 2, 3, and 4 illustrate the differences with respect to the stressed syllables. It can be seen that the stressed syllables show a higher F0, a greater amplitude and a greater duration than their unstressed variants.

3.2. The influence of the accent at sentence level

The experiments focussed on the analysis of the prosodic pattern at the sentence level for the utterance "Mama vine azi la mine" (My mother comes today to me), when the stress was changed on different words (first on "mama" – my mother, second on "vine" – comes, third on "azi" - today and fourth on "la mine" – to me) in order to induce various semantic intentions (Figures 5, 6, 7, 8). For the same utterance, the supra-segmental features varies a lot, according to the stressed word and the emphasis of the speaker for a particular meaning of the sentence. Despite the reduction of phrasal stress, each word maintains its lexical stress pattern that may be noticed in the figures above.

3.3. Experiments on intonation

The intonation is what is perceived as speech melody. The sentence mode, such as declarative or interrogative mode are communicated by means of variation of the melody. For example, the two sentences: a) "Ai reuşit!" (You have succeeded !) and b) "Ai reuşit?" (Have you succeeded?) are distinguished by a difference in intonation patterns. The first sentence carries a falling and the second carries a rising intonation pattern at the end of the sentence (Figures 9, 10). In its acoustic manifestation, intonation is primarily related to fundamental frequency (F0). So for example, the utterance final rises and falls of perceived intonation in the examples above is well captured by the F0 inflections shown in the associated experiments. In this example the interrogative intonation is not only characterised

by a rising intonation, but also by a greater dip in the fundamental frequency prior to the word-final rise in intonation. This phenomenon is specific for multisyllabic words with non-initial stress. So, we may conclude at this stage there are variations in F0 induced by stress and variations induced by intonation.

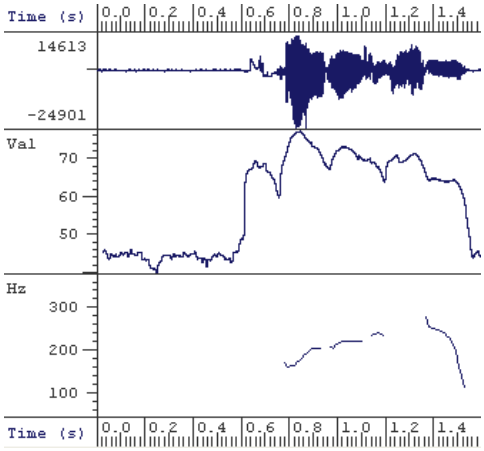

Fig. 9 – Falling intonation pattern for "Ai reuşit!"

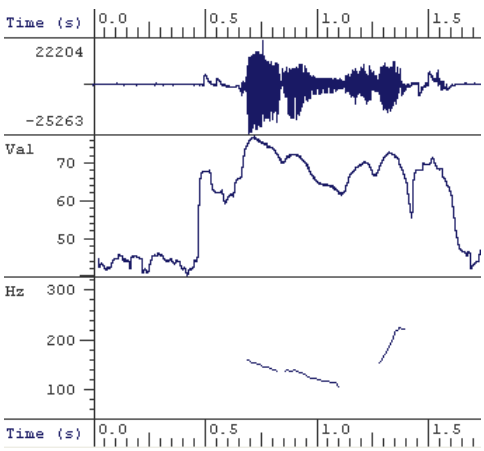

Fig. 10 – Rising intonation for "Ai reuşit?"

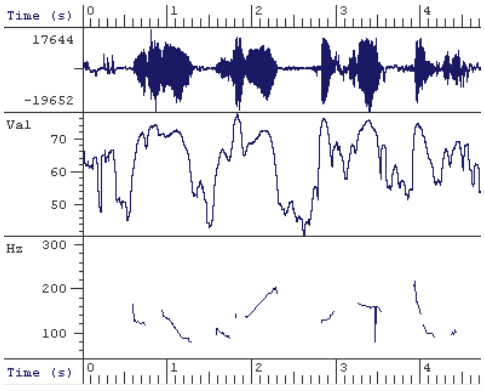

Fig. 11 – Intonation pattern for the sentence: "Moşia, moşie, foncția, foncție"

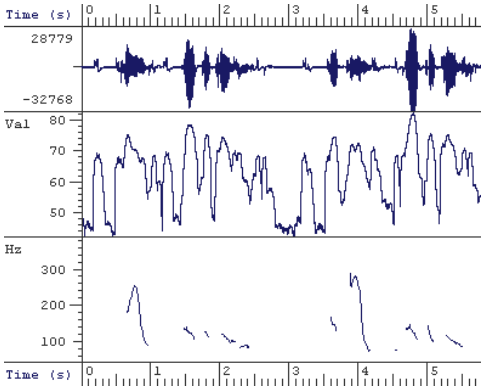

Fig. 12 – Intonation pattern for the sentence: "Când pleacă trenul ?"

An exception for the intonational rising pattern rule at the end of interrogative sentences is in the case of partial interrogation sentences, relatively short in duration, such as: "Când pleacă trenul?" (When is departing the train?, Fig. 12). In such cases, the rising pattern is present at the beginning of the sentence because of the accent on the word *când* ('when') and it decreases to the end. Also, there is a specific intonation with rising and falling patterns for sentences like "Moşia, moşie, foncția, foncție" (Fig. 11).

A further question is if there are interactions between stress and intonations which are reflected in the variation of F0. Our experiments show that F0 interactions are manifold. First, the intonation is very hard to predict. The research of the last decades has shown that: F0 structure is very complex, it has many inter- and intra-speaker variability, it is subject to variations due to the number of syllables, placement of main and secondary stress and interactions with intonational variables. Second, intonation used in most speech synthesis is quite rudimentary. Few systems implement intonational modulations that go much beyond the obligatory question-final F0 raises and fewer implements word-level modulations. More details are presented in Giurgiu and Peev (2005).

3.4. Analysis of rhythm, speech rate and duration

Variation in speed of speech production give rise to different perceptual impressions. If the entire utterance is spoken at fast/slow speed, this corresponds to a modification of rhythm. If time variations are of local nature, the durational effects are likely related to stress. Local slowing, resulting in an increase of local duration, is representative of stressing (see durations in Table I for stressed/unstressed syllables). On the other hand local acceleration signifies lessened semantic importance.

Till now we were not able to formalise such durational modifications, mainly because they are not linear. This means that it is not possible to produce natural speech from text, only by accelerating or decelerating normal speech by a fixed rate. The problem is much more complex. Concerning the duration of the pauses between words, a common finding is that they tend to get longer as the sentence proceeds. The duration of such pauses was studied both for spontaneous speaking as for speaking by reading the text (Giurgiu, Peev 2005).

3.5. Other conclusions of the experiments

During the experiments on the acoustic manifestation of the prosody, other observations have been noted and they are briefly reported: a) word grouping – in spontaneous speech the words are grouped according to prosodic principles (stress, syntax, etc); b) syllabification – individual syllables are often clearly discernible from the acoustic signal by a higher intensity; c) content words vs. function words – content words such as lexemes with independent semantic meaning (nouns, verbs, adjectives) are more likely to contain stressed syllables than function words (articles, conjunctions, suffixes); d) there are language-specific prosodic differences such as exact manifestation and placement of stress, the pauses, the communicative function of intonation patterns, etc (Keller 1994).

4. GENERATION OF PROSODIC PARAMETERS IN A TTS SYSTEM

Prosody plays an important role is TTS synthesis, both in terms of intelligibility and naturalness. Fundamental frequency, intensity and duration have to be modelled in accordance with the information from segment structure, lexicon, syntax and semantics. The synthesis system has to predict the acoustic descriptors of the prosody by analysing the text in its complexity.

The factors that need to be taken into account and predicted in the synthesis step for Romania TTS synthesis could be summarised as follows: word stress, phrasal stress, sentence stress and intonation, preceding acoustic unit's prosodic/segmental structure, succeeding acoustic unit's prosodic/segmental structure, the actual prosody of the current acoustic unit. At this stage we are studying the prediction of such parameters, mainly on the basis of statistical processing as well as the use of Artificial Neural Networks (ANN).

5. CONCLUSIONS

The present research is a preliminary study on the evaluation of qualitative, but mostly quantitative aspects of acoustic parameters of the prosody in Romanian. This research is helpful in the application of the resulted knowledge in a Romanian TTS synthesis system based on a linguistic expert system. The system will use the Romanian diphones as acoustic units. In order to predict the appropriate values of the acoustic parameters used in prosody control in the synthesis stage of the speech signal, the system need to be trained with the prosody knowledge extracted from known situation. The principles of developing such a system have been established through the current research.

The experiments demonstrated that the prosody prediction unit needs to take into consideration several factors that have been identified and presented above. The future work will concentrate on the development of a prosodic model able to generate the quantitative values for fundamental frequency, sound intensity and for acoustic unit duration on the basis of text and context analysis.

Acknowledgement. This work was possible under the research grant "Romanian text to speech synthesis using a linguistic expert system for multimodal interfaces", 128/16.08.2004, funded by the Romanian Ministry of Education and Research in the frame of INFOSOC (Information Society) Programme.

REFERENCES

Allen, J. B., M. S. Hunnicut, D. Klatt, 1987, *From text to speech: The MITalk system,* Cambridge, Cambridge University Press.

Bhaskararao, P., N. P. Venkata, 1992, *A report on an unlimited text-to-speech system,* manuscript.

Giurgiu, M., Peev., L., 2005, *Scientific reports on the research "Romanian text to speech synthesis using a linguistic expert system for multimodal interfaces"*, Technical University of Cluj-Napoca.

Isard S. D., D. A. Miller, 1986, *Diphone synthesis techniques,* IEE International Conference on speech input/output techniques., 258, 77–82.

Keller E., 1994, *Fundamentals of speech synthesis and speech recognition*, New-York, John Wiley & Sons Press.

Sagisaka Y., H. Sato, 1985, *VCV compilation speech synthesis using prosodic elements extracted from original speech,* Trans of Institute of Electr and Commun, A64, 551.

Sato, H., 1992, "Speech synthesis using CVC concatenation units and excitation waveform elements*",* Trans. on Acoustical Society of Japan, 83–69.