ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
UDC/UDK 004.934.5:[811.512.161:801.61]

A RULE BASED PROSODY MODEL FOR TURKISH TEXT-TO-SPEECH SYNTHESIS

Ibrahim Baran Uslu, Hakki Gokhan Ilk, Asim Egemen Yilmaz

Original scientific paper

This paper presents our novel prosody model in a Turkish text-to-speech synthesis (TTS) system. After developing a TTS system driven by parametric features consisting of duration, pitch and energy modifications, we try to figure out some prosody rules in order to increase the naturalness of our synthesizer. Since the inflected verbs in Turkish can be stand-alone sentences with the suffixes they take, we build a perceptual prosody model by defining rules on the stress patterns of verb inflections. Affirmative, negative and interrogative (both positive and negative) forms of many verbs were examined in a systematic way. Not only verbs, but in the same way, some phrases were examined for obtaining a proper prosody. According to the results of listening tests, the defined rules based on duration, pitch and energy modification weights, result in perceptually better speech synthesis, namely about 1,78/5,0 improvement in average in the CMOS (Comparative Mean Opinion Score) test. This improvement shows the success of our novel prosody model.

Keywords: *CMOS test, diphone, natural speech, prosody, PSOLA, text-to-speech synthesis (TTS), verb inflection*

Prozodijski model za sintezu turskog teksta u govor na temelju pravila

Izvorni znanstveni članak

Ovaj članak predstavlja naš novi prozodijski model u sustavu za sintezu turskog teksta u govor (TTS). Nakon razvijanja TTS sustava vođenog parametrijskim osobinama koje se sastoje od promjena trajanja, visine i jačine glasa, pokušavamo postaviti neka prozodijska pravila kako bi se povećala prirodnost našeg sintetizatora. Budući da u turskom jeziku glagoli koji se sprežu mogu biti samostalne rečenice uz sufikse koji im se dodaju, sastavljamo perceptualni prozodijski model definiranjem pravila o obrascima naglasaka kod sprezanja glagola. Sistematski su se proučavali potvrdni, negativni i upitni (i potvrdni i negativni) oblici mnogih glagola. Nisu se proučavali samo glagoli već, na isti način, i neke fraze kako bi se postigla ispravna prozodija. Prema rezultatima testova slušanja, definirana pravila zasnovana na promjenama trajanja, visine i jačine glasa, dovode do perceptualno bolje govorne sinteze, naime u prosjeku do 1,78/5,0 poboljšanja u CMSO testu (Comparative Mean Opinion Score). To poboljšanje predstavlja uspjeh našeg novog prozodijskog modela.

Ključne riječi: *CMOS test, difon, prirodni govor, prozodija, PSOLA, sinteza teksta u govor (TTS), sprezanje glagola*

1 Introduction

Text-to-speech synthesis is a common research area for many languages. It has many important application areas such as reading and communication aid for the blind and vocally handicapped people, voice response and warning systems, announcement systems, man-machine communication, educational tools in language learning, etc. The goal of the research studies in TTS is to synthesize intelligible and natural speech as close as possible to human voice. As the intelligibility and naturalness increase, TTS systems will find more application areas in daily life.

TTS systems of different languages have common blocks, but especially the prosody block is specific to the language. One of the most commonly used techniques for speech synthesis is concatenative synthesis [1], which combines pre-recorded speech segments after proper duration, pitch and energy modifications. In our study, we concatenate the speech units with the well-known PSOLA (Pitch Synchronous OverLap and Add) method [2] and we use diphones as speech units. We also perform duration and pitch modifications with PSOLA. There is no unit selection in our work, so we have only one diphone database [3] as source for proper intonation.

Some of the hurdles against the goal of natural concatenative speech synthesis are:

- recording of a huge database of speech units (probably with different pitch frequencies),
- duration modelling for the decided speech units,
- finding proper prosody rules for the synthesis,
- spectral differences between the concatenated units,
- differences between the energy levels of the segments and etc.

Differences between the pitch values (fundamental frequencies) and spectral contents of adjacent units introduce a distortion in the synthesized speech. The effect is similar to hearing two different speakers at the same time and therefore a precaution must be taken against this distortion. The first step can be equalizing the pitch frequencies by proper pitch scalings and this may result in a more stable synthesis. Unfortunately, there may still be some spectral envelope mismatches that will degrade the quality [1]. The energy difference between adjacent units will again result in artefacts in the synthesized speech. We try to overcome these problems with the parametric features of our TTS system; however the spectral envelope mismatch problem is beyond the scope of this paper. Due to the agglutinative nature of Turkish, prepositions, auxiliary verbs, indefinite relative clause constructors, etc. neither occur nor appear as isolated words in sentences. Instead, such grammatical structures are used by means of appended suffixes. Eventually, it is quite common to encounter multiple agglutinations in Turkish words in sentences. In case of multiple suffixes, the stress position varies according to the function of each suffix, as well as the stem of the inflected word. As a matter of fact, verb inflection is nothing but a very special case of this phenomenon.

The aim of this work is to find a solution for the prosody modelling problem. Stressed syllables in both the phrases and inflected verbs in selected sentences, are applied some duration, pitch and energy modifications for a proper prosody in the proposed model (the details are given in Section-3). In Section-2, we present a literature

survey for prosody studies for Turkish. In Section-3, we briefly explain the properties of the testbed which is a TTS system developed by us. The details of the proposed prosody model are also illustrated in this section. In Section-4, we give evaluation results of the performed listening tests and finally in Section-5 conclusion and discussions are given.

2 Prosody models for Turkish

Natural speech has its own prosody. Prosody is the harmony of the sentences which is specific to that language. The effect of prosody on speech is important; it can even change the meaning of an utterance. Actually, prosody helps the speaker and the listener to understand/ interpret the spoken material more easily; therefore a proper model for prosody is required to synthesize speech naturally. Stress is an important component of prosody. It is highlighting a syllable of a word by pronouncing it in a stronger and more striking way [4]. The stress of Turkish words is usually on the last syllable. This is the default stress position of words. Every word has exactly one main stress [5], however Turkish is an agglutinative language and the suffixes the words take can change the stress position. According to Sezer stress [6], the stresses in the place names are usually at the beginning or in the middle of the words. Some researchers investigated the generality of the Sezer stress. Both [5] and [7] tried to generalize the cases exceptional to the regular stress in Turkish. Şayli [8], in his work, investigated the duration models for speech segments in Turkish since duration is one of the significant components of prosody. He searched for mean durations of phones and triphones. He also investigated the success of the triphone tree and linear additive models for duration modelling. As we concatenate diphones in our study, we use the mean diphone durations which we calculate by taking the average of the mean phone durations found in [8]. Öztürk [9], in her work, investigated the phone duration models, as well as F_0; fundamental frequency contours for Turkish. Her classification and regression tree based phone duration model takes into account many textual features such as preceding and following phoneme types, number of syllables, syllable position, syllable stress properties, and etc. The most effective three parameters on the mean durations are: phoneme identity, left and right phoneme types and position in the syllable. Fundamental frequency contours were also modelled statistically in terms of syllable pitch frequencies. She concludes in the end, that the models developed would better be evaluated perceptually. In another work, Oskay et al. [10] tried to generalize some intonation rules for sentences, depending on some textual features. They modelled F_0 contours using linear and second order functions. Both male and female voices of affirmative, negative and interrogative sentences were recorded and examined in that study. Külekçi and Oflazer [11], in their work, focused on phrase boundary detection based on syntactic analysis and assigning intonation levels (3 levels) to words in detected phrases. In [12], we applied Fujisaki intonation model to some Turkish sentences for the first time. This approach improved the naturalness of the synthesis with an improvement of 0,15 in the PESQ score.

The original contribution of this study is the modelling of the stresses of verb inflections and phrases for giving a natural prosody to the synthesis. Verb inflection is a characteristic property of Turkish. The inflection rules, as well as the stress positions of the inflected verbs are well known [13, 14]. That is why we chose to build a prosody model for the inflected verbs initially, since they have an important effect in the meaning (affirmative, negative or interrogative) of the sentences in Turkish. In a systematic way, combinations of duration, pitch and energy modifications are applied to the raw synthesis and the listeners' preferences are evaluated. The operations performed in the proposed prosody model are described in detail in the following section.

3 The proposed prosody model

We illustrate the properties of our prosody model in this section. The method, the notation and the details of our work are given here.

3.1 The testbed developed for Turkish TTS system

A testbed is developed for Turkish text-to-speech synthesis, as shown in Fig. 1, using Matlab GUI. The text to be converted into speech is typed in the *text edit box*. Given in Fig. 1 is the text: *"soruları kavramalısın" (you should comprehend the questions)*. This text is automatically converted to lower-case and parsed into words and diphones. Words are detected according to the space character, and diphones are determined according to Turkish pronunciation rules. The system can concatenate diphones from two databases (from one male and one female voice), using the PSOLA method. In this work we synthesized speech with female voice using diphones obtained from [3]. Duration, pitch and energy modifications over individual diphones can be controlled in percentages as shown in Fig. 2, Fig. 3 and Fig. 4, respectively. Another adjustable parameter is the number of overlap pitch cycles as shown in Fig. 4. Here, "a1-a2, l1-l2, k1-k2 and r1-r2" are used for different phonemes. There are 29 letters and 43 phonemes in Turkish [4]. The operations performed in the testbed can be best explained via mathematical expressions and formulations. In the following subsection, the nomenclature will be presented.

3.2 Nomenclature and notation

In order to follow the notation easily, it is better to list the definitions in a table. The symbols used for phone, diphone and word durations, as well as diphone pitch frequency and energy values are given in the nomenclature in Tab. 1.

Table 1 Nomenclature

$t_i^{(w)}$	Duration of the i^{th} word
$t_{i,j}^{(d)}$	Duration of the j^{th} diphone in the i^{th} word
$t_{i,j,k}^{(p)}$	Duration of the k^{th} phone in the j^{th} diphone in the i^{th} word
$p_{i,j}^{(d)}$	Original pitch frequency of the j^{th} diphone in the i^{th} word
$(p_{i,j}^{(d)})'$	Modified pitch frequency of the j^{th} diphone in the i^{th} word
$e_{i,j}^{(d)}$	Original energy of the j^{th} diphone in the i^{th} word
$(e_{i,j}^{(d)})'$	Modified energy of the j^{th} diphone in the i^{th} word

The first operation performed in the testbed is to bring the loaded diphones' durations to mean duration values, calculated by averaging the mean phone durations obtained from [8].

This operation is given in Eq. (1).

$$t_{i,j}^{(d)} = \frac{t_{i,j,1}^{(p)} + t_{i,j,2}^{(p)}}{2}. \quad (1)$$

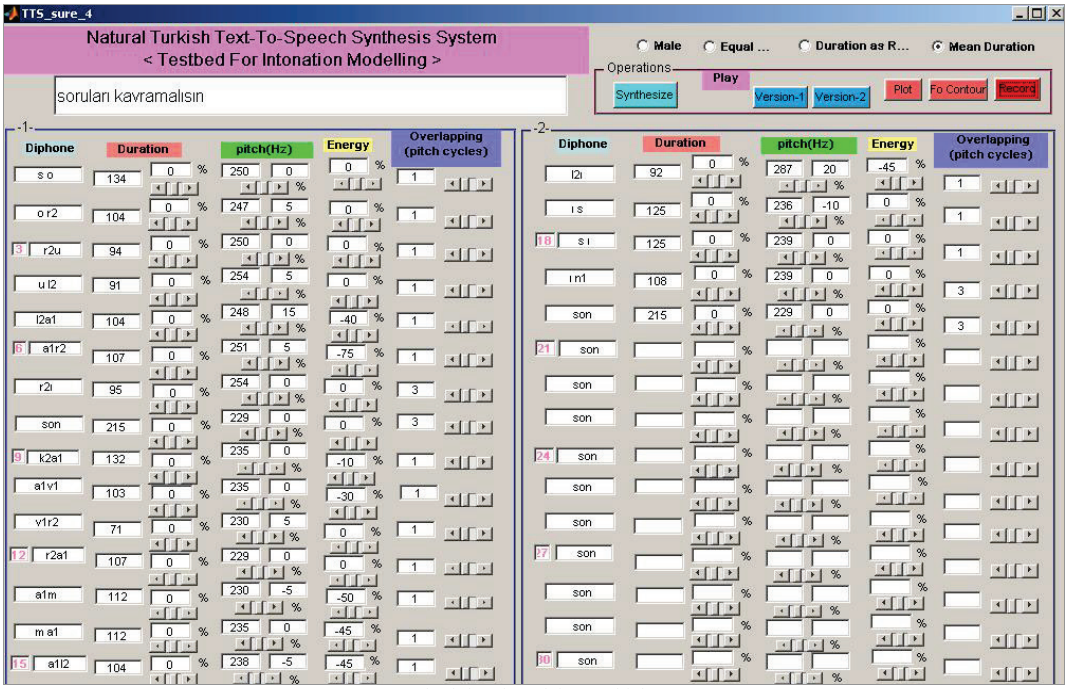

Figure 1 Testbed developed for Turkish TTS System

Here, $t_{i,j}^{(d)}$ is the duration of the j^{th} diphone (j = 1,..., K) in the i^{th} word. $t_{i,j,k=1}^{(p)}$ and $t_{i,j,k=2}^{(p)}$ are the mean phone durations (k = 1: former, k = 2: latter phoneme). According to the number of diphones in the words, diphone duration adjustment parameter: α_d is applied as given in Eq. (2).

$$t_i^{(w)} = \alpha_d \sum_{j=1}^{K} t_{i,j}^{(d)}. \quad (2)$$

Here, $t_i^{(w)}$ is the duration of the i^{th} word and K is the number of diphones in the i^{th} word.

The values for the α_d parameter are determined intuitively, and given in Tab. 2. The reason for these values can be explained as follows. For short words, we increase the diphone mean durations by 20 % in order to make them longer and thus increase their intelligibility. For longer words, we decrease the diphone mean durations by 5 % in order to avoid a slower reading.

Table 2 Values selected for diphone duration adjustment parameter

α_d = 1,20	number of diphones in the word ≤ 3
α_d = 1,00	4 ≤ number of diphones in the word ≤ 6
α_d = 0,95	number of diphones in the word ≥ 7

Similar to duration modifications (see Fig. 2), the following pitch modifications (Eq. (3)) are performed.

$$(p_{i,j}^{(d)})' = \beta_j (p_{i,j}^{(d)}), \quad (3)$$

where $p_{i,j}^{(d)}$ is the original pitch value of the j^{th} diphone and $(p_{i,j}^{(d)})'$ is the modified pitch value of the j^{th} diphone as described in Tab. 1.

$\beta_j = 1 + \sigma_j$, where β_j is the pitch frequency scale factor and σ_j is the corresponding percent increase/ decrease parameter for the j^{th} diphone, adjusted through the testbed (see Fig. 3).

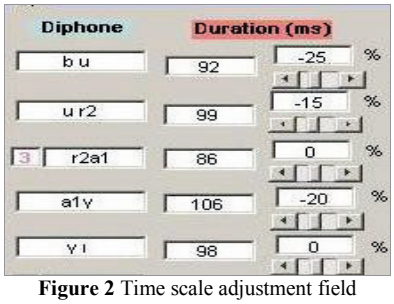

Figure 2 Time scale adjustment field

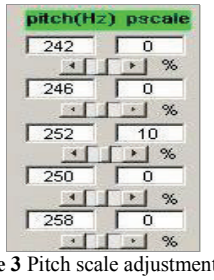

Figure 3 Pitch scale adjustment field

As an example; for σ_j = −0,15 then β_j = 0,85 and the pitch frequency of the j^{th} diphone will be decreased by 15 % from its original value, which is obtained from a pitch

estimation algorithm. Similar to pitch modifications, the following energy modifications (Eq. (4)) are performed.

$$(e_{i,j}^{(d)})' = \gamma_j (e_{i,j}^{(d)}), \quad (4)$$

where; $e_{i,j}^{(d)}$ is the original energy of the j^{th} diphone and $(e_{i,j}^{(d)})'$ is the modified energy of the j^{th} diphone as described in Tab. 1.

$\gamma_j = 1 + \lambda_j$, where γ_j is the energy scale factor and λ_j is the corresponding percent energy increase/decrease parameter for the jth diphone, adjusted through the testbed (see Fig. 4).

As an example; for λ_j = +0,20 then γ_j = 1,20 and the energy of the j^{th} diphone will be increased by 20 %.

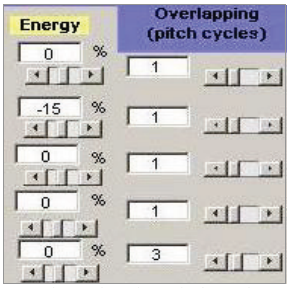

Figure 4 Energy and overlapping pitch cycles adjustment field

3.3 Signal processing infrastructure

The developed testbed performs several speech processing tasks. Initially voiced/ unvoiced classification is performed according to four criteria, namely: energy, zero crossing level, correlation coefficient at unit sample delay and spectral flatness [15]. The next operation is to equalize the fundamental frequencies of two neighbour diphones using a moving average filter. If there are differences between the energy levels of the diphones, they are also equalized as given in Eq. (5) and Eq. (6).

$$E_1 = \frac{1}{K}\sum_{n=1}^{K} d_1^2(n), \; E_2 = \frac{1}{L}\sum_{n=1}^{L} d_2^2(n). \quad (5)$$

Here, E_1 and E_2 are the energy levels of the first and the second diphones of lengths K and L, respectively and $d(n)$ is the time domain waveform for the corresponding diphone.

$$\gamma = \sqrt{\frac{E_1}{E_2}}, \; d_2(n)' = \gamma \cdot d_2(n). \quad (6)$$

Another feature of the testbed is the ability to estimate the pitch contour of the synthesized speech for only the voiced frames. The pitch frequencies are estimated using the autocorrelation method [15]. Fig. 5 illustrates the pitch contour of the synthesized speech using the developed testbed after pitch modification.

The proposed prosody model in this work tries to find rules for giving proper prosody to the inflected verbs and phrases. First of all, taking into account the stress position of the syllables of inflected verbs [13], duration, pitch and energy scaling ratios for the diphones were determined. Using all the features of the designed testbed, many experiments on verb inflections with different pitch frequency, energy and duration modification rates were performed and evaluated by listening tests. The prosody of affirmative, negative and interrogative (both positive and negative) forms of verbs were investigated according to the proposed model. Many verbs were examined, but since the inflection rules of verbs in Turkish can be generalized, only two of the examined verbs *"gezmek" (to travel)* and *"kavramak" (to comprehend)* are illustrated in this section. Then these rates were generalized and applied to noun and adjective phrases in the selected sentences. Eight combinations of duration, pitch frequency and energy modifications, given in Tab. 3 were applied to the stressed diphones of these verbs. As seen in Tab. 3, both individual and combined effects of these acoustical properties were tested.

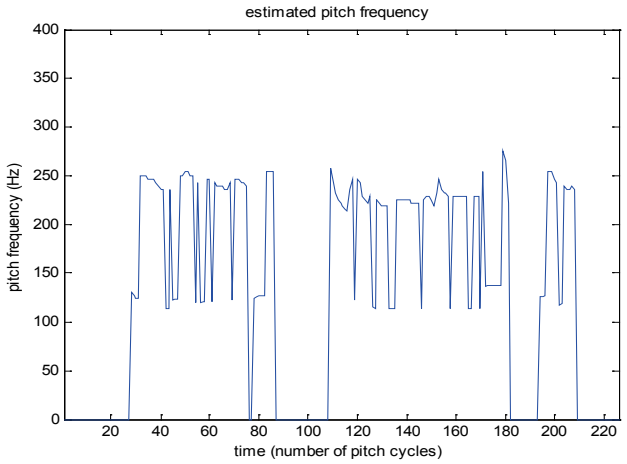

Figure 5 Estimated pitch frequency contour of the sentence: "soruları kavramalısın" (you should comprehend the questions)

Table 3 Combinations of duration, pitch and energy modifications

Combination code	Modification(s) performed
1	none (raw synthesis)
2	duration only
3	pitch frequency only
4	energy only
5	duration + pitch frequency
6	duration + energy
7	pitch frequency + energy
8	duration + pitch frequency + energy

After a set of extensive listening tests done with different rates, it was decided to apply a 40 % increase in the duration for duration modification; to apply a 20 % increase in the pitch frequency for pitch modification and to apply a 40 % increase in the energy of the diphone(s) of the stressed syllable for energy modification. These rates constitute the framework of the proposed prosody model.

In the following section, listening test results will be examined.

4 Listening test results

For the evaluation of the proposed prosody model, CMOS (Comparative Mean Opinion Score) tests which can be examined from the following URL's:

http://demo.reformo.net/baran/index.php
and
http://demo.reformo.net/baran2/index.php

were performed online. 40 naive listeners of ages between 24−44 participated in the CMOS tests. 2 sets of 32 files were loaded randomly and the listeners were asked to compare the modified synthesis with the raw one, without knowing which one is the modified (and how it is modified) and which one is the raw synthesis.

The listeners graded their preferences between –5 and +5, where the grades are given in Tab. 4. The preference results for the investigated verbs: *"gez-mek" (to travel)* and *"kavra-mak" (to comprehend)* are given in Tab. 5 and Tab. 6, respectively in the end of the paper, since they occupy large spaces.

Table 4 Grades for the CMOS test

Very much better	+5
Much better	+4
Better	+3
Slightly better	+2
A little bit better	+1
About the same	0
A little bit worse	–1
Slightly worse	–2
Worse	–3
Much worse	–4
Very much worse	–5

Table 5 Preference over raw synthesis for the modifications performed for the positive, negative and interrogative forms of the verb: "gez-mek" (to travel) in the sentence: "Burayı gezdim" (I travelled around). Here, "ˈ" denotes the position of the stress.

<affirmative, negative and interrogative (both positive and negative) forms of the sentence: "I travelled around" vs. prosody combinations>	Prosody combination code as illustrated in Tab. 3							
	1	2	3	4	5	6	7	8
*"Burayı gez−**ˈdim**" (I travelled around)*	0,03	−0,3	−1,7	0,1	−2,23	−0,57	−1,43	−2
*"Burayı **ˈgez**−me−dim" (I didn't travel around)*	−0,4	−0,87	−0,93	0,33	−0,63	−0,6	−0,53	−0,7
*"Burayı gez−**ˈdim** mi?" (Did I travel around?)*	0,07	0,07	**1,3**	0,13	0,5	−0,13	**1,37**	0,73
*"Burayı **ˈgez**−me−dim mi?" (Didn't I travel around?)*	−0,03	−0,8	**1,03**	0,06	−0,53	−0,6	**0,83**	0,33

Table 6 Preference over raw synthesis for the modifications performed for the positive, negative and interrogative forms of the verb: "kavra-mak" (to comprehend) in the sentence: "Soruları kavramalısın" (You should comprehend the questions). Here, "ˈ" again denotes the position of the stress.

<affirmative, negative and interrogative (both positive and negative) forms of the sentence: *"You should comprehend the questions" vs. prosody combinations>*	Prosody combination code as illustrated in Tab. 3							
	1	2	3	4	5	6	7	8
*Soruları kav−ra−ma−**ˈlı**−sın (You should comprehend the questions)*	0,07	−0,21	−0,71	0,29	−1,5	**0,86**	−0,07	−1,21
*Soruları kav−**ˈra**−ma−ma−lı−sın (You shouldn't comprehend the questions)*	0,07	0,29	0,36	0,07	0,43	0,07	0,43	0,07
*Soruları kav−ra−ma−**ˈlı** mı−sın? (Should you comprehend the questions?)*	0,14	0,36	**0,86**	0,43	**1,43**	−0,43	**1,57**	0,5
*Soruları kav−**ˈra**−ma−ma−lı mı−sın? (Shouldn't you comprehend the questions?)*	0,36	−0,36	−0,43	−0,36	0,5	0	−0,07	0,43

According to the average grades obtained in Tab. 5 and Tab. 6, the results show us:

- For the affirmative and negative forms, almost all of the modifications make an adverse effect in the quality of the synthesis and therefore should not be performed.
- For the positive interrogative form, for both verbs, the best improvement is achieved with the combined effect of pitch frequency and energy modifications, namely 1,37 point increase in the first and 1,57 point increase in the second verb.
- Furthermore for the positive interrogative form, only pitch frequency modification improves the CMOS score by 1,3 for the first verb, and combined effect of duration and pitch frequency modifications improve by 1,43 for the second verb. We can say that the most important contribution comes from the pitch frequency modification for this form of verb inflections.
- For the negative interrogative form, there is not a definite improvement in the synthesis. Although there is a 1,03 increase with just pitch modification and 0,83 increase with combined pitch frequency and energy modifications in the first verb, the same improvement was not obtained in the second verb. Therefore, it is better to use the raw synthesis, without doing any modifications, for this form as well.

For generalization and completeness, the sentences given in Tab. 7 were also synthesized. Taken into account the numbers of diphones in the words, the duration modification rule was applied according to Tab. 2. The same prosody rules nominated previously for inflected

verbs were applied to the phrases and compounds in these sentences. Two affirmative, two negative, two positive interrogative and two negative interrogative sentences (eight in total) plus one control sentence were examined from the following URL:
http://demo.reformo.net/baran3/index.php

The ninth sentence in the Tabs. 7 and 8 was used for control purpose, since the raw and the enhanced file (file with applied prosody) were completely the same for this synthesis. The scores of the listeners who didn't give less than 2 absolutely to the ninth sentence were not used in the evaluation.

Table 7 Sentences whose prosody patterns were examined

sentences in affirmative form	1	*"Her şeye rağmen zamanında geldi." (In spite of every trouble he arrived on time.)*
	2	*"Çok çalıştığı için başarılı oldu." (She succeeded because she worked hard.)*
sentences in negative form	3	*"Otobüsle uzun yola hiç gitmedim." (I never travelled long distances by bus.)*
	4	*"Yıllardır güneş yüzü görmedi." (He didn't see the sun for years.)*
sentences in positive interrogative form	5	*"Son sınava yeterince çalıştın mı?" (Did you study enough for the last exam?)*
	6	*"Biz yokken kendine iyi baktın mı?" (Did you take care of yourself while we were abroad?)*
sentences in negative interrogative form	7	*"Görevini en iyi şekilde yapmadın mı?" (Didn't you do your work in the best way?)*
	8	*"Saçımı sana süpürge etmedim mi?" (Didn't I serve you the most?)*
for control purpose	9	*"Peki, yeterince çalışmıyor musun?" (Aren't you working hard enough?)*

Table 8 Obtained CMOS results

	No	Sentence	Average CMOS grade/5,0
sentences in affirmative form	1	*"Her şeye rağmen zamanında geldi." (In spite of every trouble he arrived on time.)*	1,71
	2	*"Çok çalıştığı için başarılı oldu." (She succeeded because she worked hard.)*	1,14
sentences in negative form	3	*"Otobüsle uzun yola hiç gitmedim." (I never travelled long distances by bus.)*	2,11
	4	*"Yıllardır güneş yüzü görmedi." (He didn't see the sun for years.)*	2,97
sentences in positive interrogative form	5	*"Son sınava yeterince çalıştın mı?" (Did you study enough for the last exam?)*	2,25
	6	*"Biz yokken kendine iyi baktın mı?" (Did you take care of yourself while we were abroad?)*	2,42
sentences in negative interrogative form	7	*"Görevini en iyi şekilde yapmadın mı?" (Didn't you do your work in the best way?)*	0,33
	8	*"Saçımı sana süpürge etmedim mi?" (Didn't I serve you the most?)*	1,29
for control purpose	9	*"Peki, yeterince çalışmıyor musun?" (Aren't you working hard enough?)*	0,28

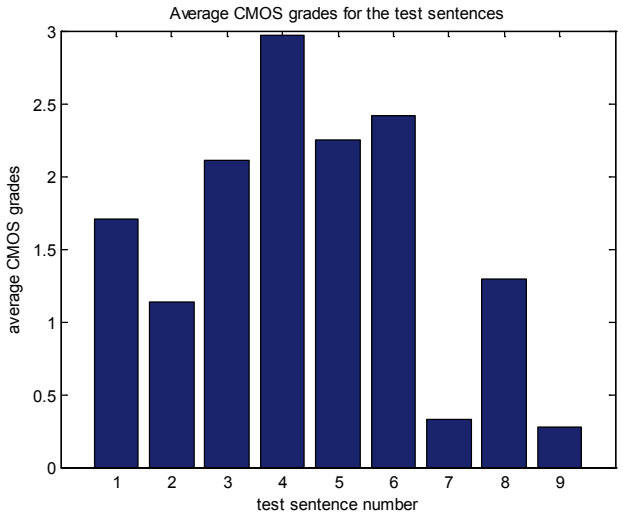

Figure 6 Average CMOS grades given to the sentences

The average grades obtained for these syntheses are given in Tab. 8. All of the results are positive in this case, meaning that our prosody model is preferred in all of the sentences. The histograms of these grades were examined and they were seen to be statistically significant. These grades show that our prosody model improves the naturalness by an increase of 1,78/5,0 points in average in the CMOS score. The distributions of these grades are plotted in Fig. 6.

5 Conclusion

In this study, some rules for the prosody of verbs in Turkish were defined. Verb inflection is a very important aspect of Turkish, since Turkish is an agglutinative language. Perception-based experiments on the prosody of some synthesized verbs and compounds were performed and stress rules based on subjective evaluations were determined.

According to the results obtained, in order to accomplish a proper prosody:

- For verbs in positive interrogative form, 20 % increase in the pitch frequency and 40 % increase in the energy of the stressed diphones are recommended.
- For verbs in affirmative, negative and negative interrogative forms, the rules defined did not make any significant improvement on the raw synthesis. So for these forms they are not recommended to be applied.
- For compounds, 30 % increase in the pitch frequency and 70 % increase in the energy of the stressed diphones are recommended in order to increase the naturalness of the synthesized speech.

The acoustical effects that have been defined and the prosody model proposed by this study, improved the CMOS score by 1,78/5,0 points in average on eight test sentences and contributed significantly to the naturalness of the speech synthesized from text.

To our belief, the prosody model proposed in this study can be applied to other Ural−Altaic languages which have similar properties to Turkish for future work.

Acknowledgements

We would like to express our gratitude to Dr. Thomas Drugman and Ph.D. candidate Alexis Moinet from Faculte Polytechnique de Mons for their sharing. Dr. Özgül Salor is appreciated for her diphone database. We also acknowledge everyone who has participated in our listening tests.

6 References

[1] Dutoit, T. An Introduction to Text-to-Speech Synthesis, Kluwer Academic Publishers, 1997.
[2] Moulines, E.; Charpentier, F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. // Speech Communication, 9, (1990), pp. 453–467.
[3] Salor, Ö.; Pellom, B.; Çiloğlu, T.; Hacıoğlu, K.; Demirekler, M. On developing new text and audio corpora and speech recognition tools for the Turkish language, ICSLP-2002: Inter. Conf. On Spoken Language Processing, Denver, Colorado USA, pp. 349–352.
[4] Ergenç, İ. Spoken Language and Pronunciation Dictionary of Turkish, Multilingual, 2002.
[5] Inkelas, S.; Orgun, C. O. Turkish stress: a review. // Phonology, 20, (2003), pp. 139–161.
[6] Sezer, E. On non-final stress in Turkish, Journal of Turkish Studies, 5, (1981), pp. 61–69.
[7] Kabak, B.; Vogel, I. Irregular stress in Turkish, Ms. Thesis University of Konstanz/ University of Delaware, 2005.
[8] Şayli, Ö. Duration analysis and modeling for Turkish text-to-speech synthesis, Master of Science Thesis, Boğaziçi Üniversitesi, 2002.
[9] Öztürk, Ö. Modeling phoneme durations and fundamental frequency contours in Turkish speech, Ph.D. Thesis, ODTÜ Fen Bilimleri Enstitüsü, 2005.
[10] Oskay, B.; Salor, Ö.; Özkan, Ö.; Demirekler, M.; Çiloğlu T. Intonation abstraction from text and its applications for Turkish sentences (in Turkish), 9th IEEE Signal Processing and Communications Applications Symposium, SIU-2001, pp. 238–243.
[11] Külekçi, O.; Oflazer, K. An infrastructure for Turkish prosody generation in text-to-speech synthesis, 15th Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN 2006), June 2006.
[12] Uslu, İ. B.; İlk, H. G. Fujisaki intonation model for Turkish text-to-speech synthesis, 17th IEEE Signal Processing and Communications Applications Symposium, SIU-2009), Antalya, pp. 844–847.
[13] Aydemir, T.; Yılmaz, A. E. A software library for determination of the stress position in Turkish for verb inflection (in Turkish), IEEE Signal Processing and Communications Applications Symp., SIU-2010, Diyarbakır-Turkey, pp. 696–699.
[14] Uslu, İ. B.; Yılmaz, A. E.; İlk, H. G. A new intonation model for inflected verbs in Turkish text-to-speech synthesis (in Turkish), 19th IEEE Signal Processing and Communications Applications Symp. SIU-2011, Antalya-Turkey, pp. 638–641.
[15] Kondoz, A. Digital Speech, 2nd Edition, Wiley Publications, 2004.

Authors' addresses

Ibrahim Baran Uslu, Ph.D.
Atilim University
Faculty of Engineering, Electrical-Electronics Eng. Dept.
Atilim Universitesi Elektrik-Elektronik Muhendisligi Bolumu
Kizilcasar Mahallesi 06836 Incek Ankara-TURKEY
E-mail: baran.uslu@atilim.edu.tr

Hakki Gokhan Ilk, Ph.D.
Ankara University
Faculty of Engineering, Electrical-Electronics Eng. Dept.
Ankara Universitesi Elektrik-Elektronik Muhendisligi Bolumu
06100 Tandogan Ankara-TURKEY
E-mail: h.gokhan.ilk@ankara.edu.tr

Asim Egemen Yilmaz, Ph.D.
Ankara University
Faculty of Engineering, Electrical-Electronics Eng. Dept.
Ankara Universitesi Elektrik-Elektronik Muhendisligi Bolumu
06100 Tandogan Ankara-TURKEY
E-mail: aeyilmaz@ankara.edu.tr