

An Infrastructure for Turkish Prosody Generation in Text-to-Speech Synthesis

M. Oğuzhan Külekci¹ and Kemal Oflazer²

¹ TÜBİTAK-UEKAE

Gebze, Kocaeli, Turkey 41470

² Faculty of Engineering and Natural Sciences

Sabancı University

Tuzla, Istanbul, Turkey 34956

kulekci@uekae.tubitak.gov.tr

oflazer@sabanciuniv.edu

Abstract. Text-to-speech engines benefit from natural language processing while generating the appropriate prosody. In this study, we investigate the natural language processing infrastructure for Turkish prosody generation in three steps as pronunciation disambiguation, phonological phrase detection and intonation level assignment. We focus on phrase boundary detection and intonation assignment. We propose a phonological phrase detection scheme based on syntactic analysis for Turkish and assign one of three intonation levels to words in detected phrases. Empirical observations on 100 sentences show that the proposed scheme works with approximately 85% accuracy.

1 Introduction

TTS systems are now able to generate highly intelligible synthetic speech from unedited text input [1], but they have some deficiencies in naturalness [2]. As the researchers aim to build synthesizers that produce speech close to human speech as much as possible, more attention has to be paid for prosody generation.

In practice, the prosody generation process of a sentence begins with the words in it. For each word, the position of the primary stress along with the correct set of phonemes must be specified at the beginning. There may be different pronunciations with different phonemes or primary stress positions corresponding to a word, and such pronunciation ambiguities must be resolved properly according to the context.

Phonetic transcriptions of words selected by the end of the pronunciation disambiguation process include the position of the primary stress. Although this is enough for inner-word prosody, detection of the words that are to be accented or deaccented in a sentence with deeper syntactic and semantic analyses are to be performed for further inter-word prosodic events. Phrase boundary detection is an important issue in synthesis of natural sounding speech both to adjust the durations between the tokens and to find out which ones to be accented or deaccented.

Most text-to-speech systems perform this boundary detection based on content word/function word distinction. This approach divides the words of a given utterance into two as content words and function words named as *chunks* and *chinks* respectively. The phrases are assumed to begin with a chunk and continue by any number of chinks [3]. For example, in sentence "[**She read**] [**the important pages**] [**in the park**]", the words **the** and **in** are function words, and according to chinks and chunks algorithm the phrases are marked between the brackets.

Another approach to detect phonological phrases is to use the syntactic analysis of the sentence. Although syntactic structure provides a good basis for prosodic structure, the effect of the semantic and discourse also has great impact [4–6]. Lindström *et al.* [6] proposed to use dependency graphs which are of the form head and modifiers. The idea is deployed on a Swedish text-to-speech system, where the output of a morphosyntactic component is used to build a dependency graph of utterances. The feasibility of the system is demonstrated by comparing the results with the human read sentences and the authors reported that it seems appropriate to use also dependency graphs in prosody generation.

Although content/function word heuristics works fine on some right headed languages such as English,³ it is not suitable for some languages such as Turkish. This is because of the free word order structure of the language, and also the difficulty in content/function word distinction, which is not very clear in Turkish. Hence, alternative solutions must be investigated for phonological phrase detection in languages similar to Turkish.

2 Pronunciation Disambiguation

Words typically have different pronunciations depending on their syntactic, and semantic properties in context. In Turkish, differences in pronunciation stem from differences in the phonemes used, the length of the vowel and the location of the primary stress [8]. The selection of the correct pronunciation requires a disambiguation process that needs to look at local morphosyntactic and semantic information to determine the correct pronunciation among alternatives. Disambiguating morphology serves a good starting basis for disambiguation of pronunciations, although it by itself, does not disambiguate all ambiguous cases of pronunciation. For example, determining the correct morphological analysis of the word **okuma** in Turkish, distinguishes between the possible pronunciations of this word in the sentences '**Okuma kitabı belirlendi.**' (*Reading book has been determined.*) and '**Saçma sapan şeyleri okuma.**' (*Don't read those silly things.*) In the former, **okuma** is an infinitive form derived from verb **okumak** (*to read*) and corresponds to phonetic representation /o-ku-"ma/ in SAMPA representation.⁴ Note that " indicates the stressed syllable, and - indicates a

³ Hirschberg [7] discussed that especially in synthesis of longer texts, this approach is problematic also for English.

⁴ SAMPA (Speech Assessment Methods Pronunciation Alphabet) is an international machine-readable pronunciation alphabet. For further information, please refer to

syllable boundary. In the latter case the same word functions as an imperative form of the same verb, and pronunciation is represented with /o-"ku-ma/ where the primary stress is on the second syllable. A text-to-speech system would have to take this into account for the generation of proper prosody.

Sometimes morphology may not be enough to differentiate between the possible pronunciations. Besides morphological disambiguation, word sense disambiguation and named entity recognition are the subsidiary tools for pronunciation disambiguation. The word **adet** is such an example that word sense disambiguation should be used to determine the reading. It has two pronunciations as /a-"det/ and /a:-"det/ corresponding to meanings *piece* and *tradition* respectively. Part-of-speech tags of both are noun and their all inflectional forms have exactly the same morphological analysis. Thus, it is not possible to disambiguate them using syntactic properties and word sense disambiguation should be applied to catch the correct sense, which also determines the correct reading, in a given context. As another example, the word **Aydın** may correspond to a city in Turkey, a man's name or an ordinary adjective meaning bright or intellectual. It has the pronunciation renderings /"aj-d1n/, /aj-"d1n/ and /aj-"d1n/ respectively. If the morphological disambiguation results that the word is used as an ordinary adjective in a given context, then the ambiguity is resolved. Otherwise, named entity recognition must be performed to find out the correct meaning (a city or a person's name) which determines the correct pronunciation.

In this study, we used the disambiguator developed by Külekci and Oflazer [9], and more detailed explanations of this disambiguator may be found in [10]. Given an input sentence, the disambiguator returns both the morphological parses and corresponding pronunciations of each word in it. Note that, the rules described below for phonological phrase detection use these disambiguated morphological parses to establish syntactic analyses.

3 Phonological Phrase Detection

Dependency parsing was proposed as an alternative for phrase boundary detection [6]. Although it requires much deeper analysis than simple chunks and chunks algorithm, this approach fits better for Turkish. After the morphological disambiguation, prosodic phrases may be detected by applying simple dependency rules between the consecutive words. Here, it is not required to extract the whole dependency graph of a given utterance, but instead a light parsing is enough. Relations between the distant words are not important for prosodic structure since the aim is to find the phonologic relationships of neighboring words. Dependency parsing of Turkish has been studied with an extended finite state approach by Oflazer [11]. Figure 1 demonstrates the relations between the words of a sample sentence from that work.

On the example sentence, *subject*, *object*, and *determiner* relations are not between the consecutive words. Hence, they don't carry valuable information

www.phon.ucl.ac.uk/home/sampa. We use the SAMPA notation to represent pronunciations in the text, where necessary.

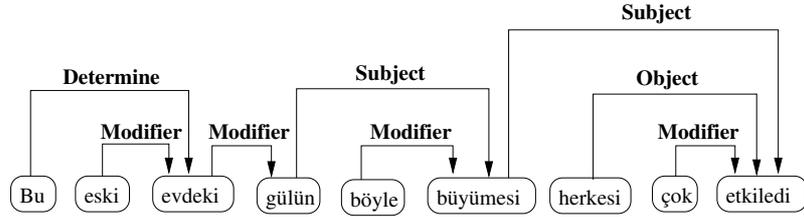


Fig. 1. The dependency structure of a sample Turkish sentence.

for phonological phrase detection. On the other side, although there is no link between the words **bu** and **eski**, they must be in the same phrase. Thus, dependency parsing alone seems not enough for phonological phrase boundary detection problem, and some extra rules must be compiled to take care of the prosodic interferences that are not handled in syntactic structure.

The following rules to search the relations between consecutive words in a sentence were empirically constructed based on the noun phrase structure and dependency parsing of Turkish [11]:

1. An adjective, determiner, or number followed by a noun defines a syntactic relationship that the preceding token modifies the succeeding one. Some explanatory examples of such situations may be given as: **güzel ev** (*sweet home*), **birçok araba** (*many cars*), **100 dolar** (*hundred dollar*).
2. Any number of consecutive adjectives, determiners, numbers, or adverbs forms a group of modifiers as in phrase **bu eski evdeki** (*in this old house*), where **bu** is a determiner and **eski** is an adjective. Note that although dependency parsing do not link these words, from a phonological point of view they are to be processed in the same phrase.
3. The word is a noun, pronoun, or postposition, followed by an adjective, adverb or noun which is derived from a verb root. This is another type of modify relation observed frequently in Turkish. For example, on the sample sentence given in figure 1, the words **böyle büyümesi** (*such grow*) demonstrate the structure of this relation.
4. Postpositional phrases constitute phonological phrases. An example is : **başlangıcından beri** (*since its beginning*).
5. A noun in genitive or nominative case followed by another noun in any case constitute a phonological phrase if the possessive agreement of the second one matches with the number/person agreement of the first noun, e.g., **üyelerinin yerine** (*in place of its members*).
6. Similar to rule 5, if a noun in genitive or nominative case is followed by a derived adjective with **Rel** tag, a phonological link is to be established between them. An example is: **adanın kuzeyindeki** (*in northern of the island*). Note that, most probably, the adjective further modifies the succeeding noun which will then constitute another phonological phrase.

7. A verb with a preceding noun in any case form a phonological phrase. This is akin to subject/object relationships of dependency parsing, e.g., **hatırlatmak istiyorum** (*I want to remind*).
8. A verb preceded by an adverb form a phonological phrase together. In this situation the adverb is the modifier of the verb. An example is: **şöyle anlattı** (*he/she explained such that*).

For each word in a sentence, it is investigated whether there is a phonological link to the preceding or succeeding word conforming to one of the rules above. Each rule binds two words. A word can be linked to both preceding and succeeding tokens by the defined rules. That constructs longer chains of words which are actually the phonological phrases we are searching for. Table 1 shows the word length of the detected phonological phrases in a one-million words corpus. Note that in this table, length 1 corresponds to the tokens that are not assigned to a phrase by the rules above. It is observed that the length of a detected phonological phrase is most of the time smaller than 5.

Word length of the phrase	% Frequency of observation
1	57.26
2	21.61
3	11.12
4	5.42
5	2.44
>5	2.15

Table 1. Word length distribution of the detected phonological phrases in a one-million words corpus.

The result of the proposed phonological phrase boundary detection process on the example sentence **Milli Savunma Bakanlığı dövizli askerlik konusunda çözüm arayışına girdi.** (*Ministry of defense had begun searching for solutions on completing the military service with money.*) is depicted below. Note that $\langle PRx \rangle$ and $\langle /PRx \rangle$ mark the beginning and end of the applied number x phrase rule.

$\langle PR5 \rangle \langle PR3 \rangle$ **Milli Savunma** $\langle /PR3 \rangle$ **Bakanlığı** $\langle /PR5 \rangle$
 $\langle PR5 \rangle \langle PR1 \rangle$ **dövizli askerlik** $\langle /PR1 \rangle$ **konusunda** $\langle /PR5 \rangle$
 $\langle PR7 \rangle \langle PR3 \rangle$ **çözüm arayışına** $\langle /PR3 \rangle$ **girdi** $\langle /PR7 \rangle$

4 Intonation Level Assignment in Phonological Phrases

In her book on Turkish phonology, Özsoy [12] argues that the words that modify, determine, or somehow related to the head words are to be accented. She also notes that the speaker or reader specifies the important point of the utterance by

the stressed word. For example, accenting the first word of the phrase **babamın yeni arabası** (*my father's new car*) emphasizes that the owner of the new car is the father, while stressing the second word underlines that the car is the new one rather than the old one. Under normal conditions the second selection is more probable.

In their studies of Turkish stress assignment, Kabak and Vogel [13], and Inkelas and Orgun [14] argue that the leftmost accentable syllable is to be stressed in case of compound noun phrases. The intonation of noun compounds and genitive possessive noun phrases were explicitly explored in the studies of Levi [15], [16]. Although the number of sample structures investigated in her studies are rather limited, Levi discussed that the noun compound phrases have their first component promoted generally while the analysis of accentuation in genitive noun phrases vary. The experiments in her studies showed that the components of a genitive noun phrase may or may not retain their pitch accents. However the reason for that differentiation could not be identified totally.

In our study, we decided to promote the first word of a genitive phrase if the second word begins with a vowel. That is based on the observation that people generally tend to read such phrases as a single lexical item in Turkish promoting the word on the left of the phrase. If the second word is not beginning with a vowel than both words are promoted equally. With the proposed accentuation of genitive phrases, the first word of the phrase **babamın evi** (*my father's house*) is accented, while both of the words retain their pitch accents on **babamın sandalyesi** (*my father's chair*).

Based on these research and observations of Turkish phrasal stress, Table 2 depicts which component is to be promoted by our previously explained rules that detect the phonological link between two consecutive words. The intonations of the phrases detected by the second rule (which connects consecutive modifiers or determiners) and the sixth rule (which is a special case of fifth rule) require their second token to be stressed more. The rest have their first words promoted. Only in some situations of the genitive noun phrases discussed in the previous paragraph, both tokens retain their accents.

Rule #	Promote First Word	Promote Second Word
1	✓	
2		✓
3	✓	
4	✓	
5	✓	
	✓	✓
6		✓
7	✓	
8	✓	

Table 2. The accentuation table of the defined rules.

Initially all of the words in a given utterance are given zero intonation level. While detecting the phonological phrases by the rules, the intonation levels of the promoted words, which are specified in Table 2, are increased accordingly. As each word may be linked to the preceding and succeeding one, the maximum level of intonation defined for a token may be 2 at most. For example, while searching the phrasal connections between the words in **sarı büyük kitap** (*yellow big book*), **büyük** is connected to **sarı** by the second rule and to **kitap** by the first rule. As second token is promoted by the second rule and first token by the first rule, the word **büyük** has an intonation level of 2.

Below is a sample sentence demonstrating the output of the phrasing and intonation level assignment of the whole system. The number written in bold between the braces at the end of each word indicates the level of intonation assigned for that word by the proposed system.

<PR5> <PR1> <PR2> <PR3> Kars'ta(**1**) yakalanan(**0**) </PR3> 500
 (**2**) </PR2> tüp(**1**) </PR1> zehirin(**0**) </PR5> <PR7> <PR3> <PR5>
 <PR1> <PR2> iki(**0**) milyar(**2**) </PR2> lira(**1**) </PR1> değerinde(**1**)
 </PR5> olduğu(**1**) </PR3> açıklandı(**0**) </PR7> (*It is stated that the 500
 tubes of poison captured in Kars cost 2 billion Turkish liras.*)

5 Results and Conclusion

The first step in generating the correct prosody is the detection of proper pronunciations of words according to the given context. In their study, Külekci and Oflazer [9] stated that they achieved Turkish pronunciation disambiguation with 99.54% recall and 97.95% precision by using the distinguishing tag based morphological disambiguator. In this study, we proposed a heuristic approach for phonological phrase detection and intonational level assignment in Turkish by using the outputs of that disambiguator.

Eight rules, which are based on dependency parsing, have been constructed to explore phonological connections between consecutive words. If there is such a relationship between any consecutive words in a sentence, they are linked. Chains of these links constitute the phonological phrases.

For intonational level assignment, each rule is associated with an accentuation that defines which word of a couple is to be stressed more. The words in a phonological phrase are assigned an intonation level based on these accentuations defined for each rule.

Empirical observation performed on 100 sentences showed that approximately 85% of the time correct intonations are assigned to words. However, the decision of correctness is subjective here, and the real performance can only be understood if the system is connected to a Turkish TTS synthesizer, which we plan to achieve as a next step.

It must be noted that there are not so many studies in the area of phrasal prosodic events of Turkish, and actually even the existing ones do not cover all the aspects to build a working system. Thus, while designing the heuristic and evaluating the results, empirical observations are taken into account. It

is believed that deeper phonological analysis of the phrasal structures will lead to better systems in practice. This attempt of phonological phrase boundary detection in Turkish may be applied to other languages which are not suitable for using function/content word distinction in phrase detection.

References

1. Nooteboom, S.: Text and prosody. In Santen, J., Olive, J., Sproat, R., Hirschberg, J., eds.: *Progress in Speech Synthesis*. Springer-Verlag (1997) 431–434
2. Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., Syrdal, A.: The AT&T Next-Gen TTS system. Joint Meeting of ASA, EAA and DAGA, Berlin, Germany (1999)
3. Liberman, M., Church, K.: Text analysis and word pronunciation in text-to-speech synthesis. In Furui, S., Sondhi, M., eds.: *Advances in Speech Signal Processing*. Dekker (1992) 791–831
4. Bachenko, J., Fitzpatrick, E.: A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics* **16** (1990) 155–170
5. Wang, M., Hirschberg, J.: Predicting intonational phrasing from text. In: *Proceedings of ACL'91*, University of California, Berkeley, California (1991) 285–292
6. Lindström, A., Bretan, I., Ljungqvist, M.: Prosody generation in text-to-speech conversion using dependency graphs. In: *Proceedings of ICSLP'96*. Volume 3., Philadelphia, PA, USA (1996) 1341–1345
7. Hirschberg, J.: Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence* **63** (1993) 305–340
8. Oflazer, K., Inkelas, S.: The architecture and the implementation of a finite state pronunciation lexicon for Turkish. *Computer Speech and Language* (2006)
9. Külekci, M., Oflazer, K.: Pronunciation disambiguation in Turkish. *Lecture Notes in Computer Science, ISCSIS 2005 Proceedings* **3733** (2005) 636–645
10. Külekci, M.: Statistical Morphological Disambiguation with Application to Disambiguation of Pronunciations in Turkish. PhD thesis, Sabancı University (2006)
11. Oflazer, K.: Dependency parsing with an extended finite-state approach. *Computational Linguistics* (2002)
12. Özsoy, A.S.: *Türkçe'nin Yapısı-I Sesbilim*. Boğaziçi University (2004)
13. Kabak, B., Vogel, I.: The phonological word and stress assignment in Turkish. *Phonology* (2001) 315–360
14. Inkelas, S., Orgun, C.: Turkish stress: A review. *Phonology* (2003)
15. Levi, S.: Limitations on tonal crowding in turkish intonation. In: *Proceedings of 9th International Phonology Conference*. (2002)
16. Levi, S.: The realization of noun compounds and genitive possessive noun phrases. Technical report, University of Washington (2002)