Duration Modeling For Turkish Text-to-Speech Synthesis System

Ö. ÖZTÜRK[1], Ö. SALOR[2], T. ÇİLOĞLU[2], M. DEMİREKLER[2]

Dept. of Electrical and Electronics Eng.,
[1]Dokuz Eylul Univ., Izmir, Turkey, [2]Middle East Tech. Univ., Ankara, Turkey
ozturk@eee.deu.edu.tr, salor@metu.edu.tr, ciltolga@metu.edu.tr, demirekler@metu.edu.tr

Abstract

Naturalness of synthetic speech depends on appropriate modeling of prosodic aspects. Mostly, three prosody components are modeled: segmental duration, pitch contour and intensity. In this study, we present our work on modeling segmental duration in Turkish by using machine-learning algorithms. The models predict phone durations based on attributes such as phone identity, neighboring phone identities, lexical stress, position of syllable in word, part-of-speech information, word length in number of syllables and position of word in utterance. Obtained models predict segment durations better than mean duration approximations.

1. Introduction

A fundamental issue in Text-to-Speech (TTS) conversion is naturalness of the synthesized speech. In order to obtain natural sounding synthetic speech, prosodic attributes of speech such as pitch frequency, duration and intensity should be modeled appropriately (Santen, 1997). In natural speech, segment durations are highly correlated to context. Similar/same phones differ in their energy, duration and fundamental frequency depending on their context (Möbius, 1996; Santen, 1997; Venditti, 1998). This process could be viewed as a mixed classification problem (Lee, 1999).

The aim of this study has been set as the determination of phone durations in Turkish according to a set of qualities including those like phonetic, syllabic, word-internal and sentence-internal contexts.

Over the years, there have been emphases on two approaches for duration modeling: rule-based systems such as Klatt's duration model (Klatt, 1987) and corpus-based statistical systems such as decision trees and Sum-of-Products (SOP) model (Batusek, 2002; Dusterhoff, Black, Taylor, 1999; Febrer, Padrell, Bonafonte, 1998; Lee, 1999; Möbius, Santen, 1996; Santen, 1997; Venditti, Santen, 1998). In this work, two corpus-based methods – Linear regression and J48, a modified version of C4.5 decision tree algorithm – have been employed in modeling segment durations in Turkish (Witten, Frank, 1999). Both subjective and objective comparisons reveal that duration models obtained via decision tree learning provide better results and both models are superior to mean duration modeling in all aspects.

In this study, we present our work on modeling segmental duration in Turkish by using machine-learning algorithms. The rest of the paper is organized as follows: Section 2 presents the prepared speech database. Section 3 summarizes the developed feature set. In Section 4, contextual attributes are described. Section 5 describes the duration analysis and modeling studies. Section 6 reports the performance of the models. Conclusions and future projects are discussed in the last section.

2. Speech Database

As a reference for duration calculations, recordings of 190 sentences form a 27-year old, non-professional native female speaker have been used. Recordings have been aligned in a former study, (Salor, Pellom, Çiloğlu, 2002), using the University of Colorado's speech recognition system, Sonic. This speech corpus consists of a total of 8005 METU-bet[1] phones (419 silence, 3246 vowels and 4340 consonants).

It has to be mentioned that the size of this database is not truely sufficient for the purpose; the formation of a richer corpus is an ongoing work.

3. Feature Set

Contextual attributes of a phone have been represented as a feature vector; the elements of a vector is a subset of the items listed below:

- **Phn:** METU-bet representation for the current phone.
- **Left_Content:** Left content feature. It might have the value plosive (1283), nasal (796), fricative (1112), lateral (531), rolled (618), and high/low vowel (1823/1423).
- **Right_Content:** Right content feature.
- **Left_Id:** Left content identity in orthographic form. The levels might be a, b, c, C, d, e, f, g, h, I, i, j, k, l, m, n, o, O, p, r, s, S, t, u, U, v, y, z, SIL.
- **Right_Id:** Right content identity in orthographic form.
- **Stress:** Lexical stress of the current phone. The syllable containing the phone may either be Accented (2393) or Not-Accented (5193).
- **Posin_Syl:** Location of the phone in the syllable. It may take values such as Onset (2920), Nucleus (3665), or Coda (1420).
- **Syl_Pos:** Location of the syllable in the word. The syllable carrying the syllable may be the Initial (2599), Middle (2542), or Final (2445) syllable of a word.

[1] METU-bet is the phonetic alphabet (containing 39 units including SILence.) developed in the course of the work presented in (Salor, Pellom, Çiloğlu, 2002).

- **Word_POS:** Part-of-Speech information of the word. It could be VERB (1522), NOUN (3944), PROPer noun (7), ADJective (1085), ADVerb (482), INFinitive (130), CONJunction (64), QUEStion (23), CompoundNOUN (44), PRONoun (147), POSTPosition (114), SIL (419).
- **Word_Len:** Length of the word in number of syllables.
- **Duration:** Segment duration in ms.
- **Duration:** Segment duration in discrete levels (due to the limitations resulting from statistical analysis package).

4. Identification of Contextual Attributes

Contextual attributes of the phones have been extracted from the text of the speech corpus. Some of the attributes such as neighboring phone identities, syllable position, position of phone in syllable, and word length could be retrieved directly from text by means of phonetic categorization and syllabification algorithm. However, because of the agglutinative nature of Turkish, the Part-Of-Speech (POS) tags and lexical stress could not be obtained utilizing already developed text analysis tools such as phonetization and syllabification. To overcome these drawbacks, morphological analysis (MA) has been employed.

MA has been done by PC-Kimmo, a freely available software developed by Karttunen (Antworth, 1990). The rules and the lexicon developed by Oflazer (1994) have been integrated with PC-Kimmo and the lexicon has been enlarged to increase its coverage. MA outputs and syllable information are then used to obtain the lexical stress of the words.

Stress, in general, is on the last syllable of the word and migrates to the right with successive suffixation in Turkish. However, there are exceptions to this rule. The rule does not apply to place names and borrowings. Furthermore, there exist stress-blocking morphemes that prevent stress migration to the right (Barker, 1989). The stress assignment algorithm used in this study identifies these exceptions using morphemic information provided by MA.

Figures 1-5 demonstrate the duration pattern for several contextual attributes derived from the dataset. It can be deduced that silence has a wide range of duration values, which might cause a deficiency in the performance of the machine-learning (ML) algorithms. For future studies, a short-pause would be employed for better performance.

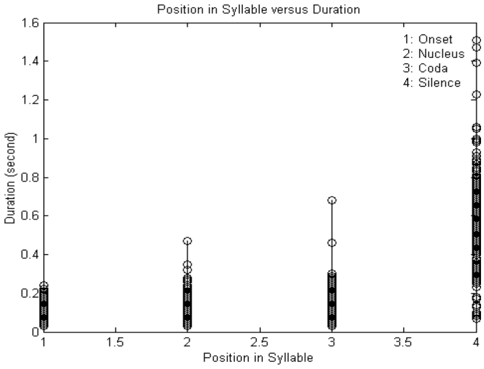

Figure 1 Position in syllable versus duration.

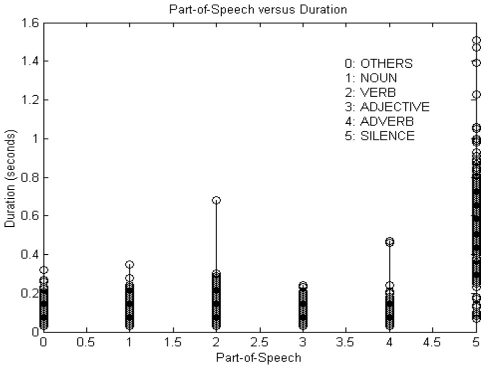

Figure 2 Part-of-speech tags versus duration

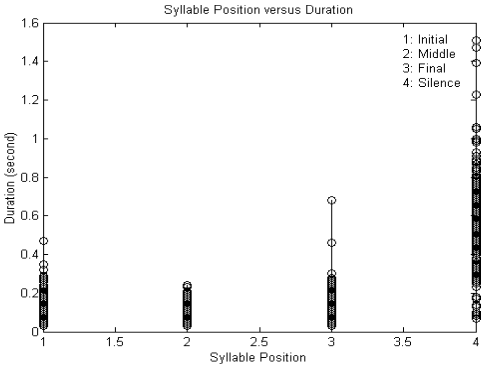

Figure 3 Syllable position versus duration

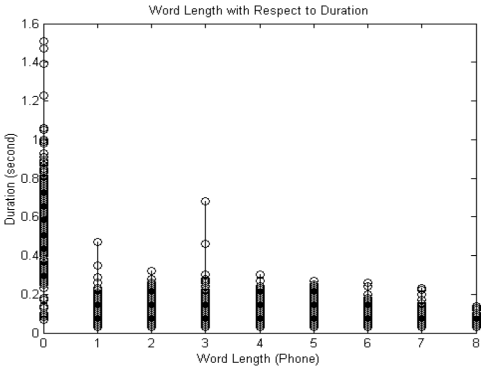

Figure 4 Word length versus duration

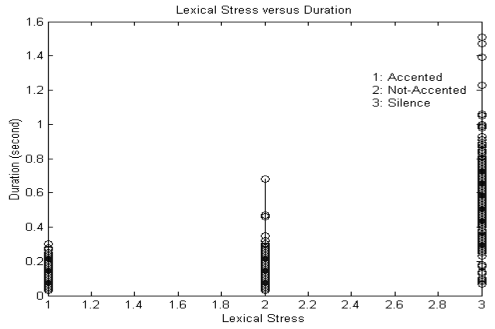

Figure 5 Lexical stress versus duration

5. Duration Analysis and Modeling

The duration analysis and modeling have been performed using a publicly available statistical analysis package, WEKA, developed at the University of Waikato, New Zeland (Witten, Frank, 1999). The package bears a collection of algorithms for solving real-world data mining problems. The software has a uniform interface to a number of standard ML techniques (Ross, 2000).

Two of the ML algorithms offered by WEKA have been utilized: Linear Regression and J48, a modified version of c4.5. J48 performs a categorical decision thus the duration of each segment has to be quantized prior to the processing. To this aim, the discretizer embedded in the WEKA package has been used. Experiments with J48 were conducted over equally spaced 100-levels for duration values.

Experimentation has been carried out for investigating the performance of different feature subsets. The feature combinations used for training are as follows:

- **Factor Set1**: Phn, LeftContent, RightContent, Stress, Posinsyl, SylPos, Duration (ms)
- **Factor Set2:** Phn, LeftContent, RightContent, Stress, Posinsyl, SylPos, Duration (100 linearly spaced duration levels)
- **Factor Set3:** Phn, LeftId, RightId, Stress, Posinsyl, SylPos, Duration (ms)
- **Factor Set4:** Phn, LeftId, RightId, Stress, Posinsyl, SylPos, Duration (100 linearly spaced duration levels)
- **Factor Set5:** Phn, LeftContent, RightContent, Stress, Posinsyl, SylPos, WorPOS, Duration (ms)
- **Factor Set6:** Phn, LeftContent, RightContent, Stress, Posinsyl, SylPos, WorPOS, Duration (100 linearly spaced duration levels)
- **Factor Set7:** Phn, LeftId, RightId, Stress, Posinsyl, SylPos, WorPOS, Duration (ms)
- **Factor Set8:** Phn, LeftId, RightId, Stress, Posinsyl, SylPos, WorPOS, Duration (100 linearly spaced duration levels)
- **Factor Set9:** Phn, LeftContent, RightContent, Stress, Posinsyl, SylPos, WorPOS, WordLen, Duration (ms)
- **Factor Set10:** Phn, LeftContent, RightContent, Stress, Posinsyl, SylPos, WorPOS, WordLen, Duration (100 linearly spaced duration levels)
- **Factor Set11:** Phn, LeftId, RightId, Stress, Posinsyl, SylPos, WorPOS, WordLen, Duration (ms)
- **Factor Set12:** Phn, LeftId, RightId, Stress, Posinsyl, SylPos, WorPOS, WordLen, Duration (100 linearly spaced duration levels)

The results were then directed to Festival speech synthesis system. Modules required by the synthesizer were written in Scheme, (Dybvig, 1996), a Festival specific language.

In our baseline system, mean phoneme durations are provided in the duration module. The mean durations are obtained from the aligned speech database of METU (Salor, Pellom, Çiloğlu, 2002), using 120 speakers with 40 sentences each. The mean durations are given in Table 1.

Phone	Duration (sec)	Phone	Duration (sec)
AA	0.054	G	0.068
AAG	0.108	GG	0.061
A	0.046	H	0.058
AG	0.092	J	0.094
E	0.050	K	0.091
EG	0.098	KK	0.090
EE	0.045	L	0.047
EEG	0.093	LL	0.050
I	0.034	M	0.068
IG	0.075	N	0.076
IY	0.035	NN	0.059
IYG	0.070	P	0.096
O	0.061	R	0.055
OG	0.123	RR	0.042
OE	0.062	RH	0.076
OEG	0.124	S	0.118
U	0.039	SH	0.114
UG	0.079	T	0.084
UE	0.035	V	0.063
UEG	0.071	VV	0.051
B	0.066	Y	0.058
C	0.078	Z	0.088
CH	0.105	ZH	0.129
D	0.056	GH	NA
F	0.091	SIL	0.607

Table 1. Turkish phones and their mean durations. These have been used as baseline durations to compare intelligibility of the proposed duration modeling.

6. Evaluations

The performance of the system is evaluated both objectively and subjectively. The overall data set has been used both for training and testing for quantitative evaluation. Table 2 shows numerical results obtained from the duration modeling procedure.

It can be deduced that each model's performance increases with the addition of new attributes to the training set. For example, adding WordPOS and WordLen to the training feature set improves the performance of J48 considerably: the correlation coefficient increases from 0.889 to 0.9107 while RMSE decreases from 0.0555 to 0.0513. However, we still think that the desired level has not been reached yet. Current research is towards expanding the data set to improve the modeling quality.

The models with the highest correlation results are used to perform perceptual listening tests with Festival speech synthesis system.

	Mean Error (sec)	RMSE (sec)	Correlation Coefficient
Set1	0,0318	0,0555	0,889
Set2	0,0274	0,0526	0,9056
Set3	0,0315	0,055	0,8909
Set4	0,0276	0,0636	0,8543
Set5	0,0317	0,0554	0,8894
Set6	0,0262	0,0521	0,9074
Set7	0,0314	0,055	0,8911
Set8	0,0262	0,0624	0,8601
Set9	0,0316	0,0554	0,8895
Set10	0,0249	0,0513	0,9107
Set11	0,0314	0,0549	0,8913
Set12	0,0253	0,0619	0,8623

Table 2 RMSE and Correlation Coefficients obtained from experiments

To test the intelligibility of our system, we have designed a diagnostic rhyme test for Turkish. 50 pairs of words with different confusability groups have been determined. The pairs are determined from several groups of words. These groups include the words that could be confused by human ear due to voicing, nasality, sustention, sibilation, graveness and compactness affects. 10 pairs are selected for each listener from this list. One word from each pair is synthesized by our TTS system. The listener listens to the synthetic waveform and decides which one from the word pair is synthesized. The overall intelligibility of the system from 20 listeners has been found as 86.5%. Then this test has been repeated with the same words but the duration model is replaced by the proposed duration model. The intelligibility of the words increased to 87.3%. It should be mentioned that mean durations used previously were obtained from a larger speech corpus (Salor, Pellom, Çiloğlu, 2002).

7. Conclusions and Future Work

In this study, we presented our work on modeling segmental duration in Turkish by using machine-learning algorithms. The duration analysis and modeling have been performed using a publicly available statistical analysis package WEKA (Witten, Frank, 1999). Two of the ML algorithms offered by WEKA have been utilized: Linear Regression and J48.

A speech database of 190 sentences were utilized during analysis and modeling. It should be mentioned that data were not sufficient to develop duration models of desired level. Though, quite satisfactory results have been obtained. It is important to note that the highest correlation coefficient achieved is 0,91.

Future studies are planned to concern larger audio database representing each attribute combination sufficiently.

References

Antworth E. L. (1990). PC-KIMMO: a two-level processor for morphological analysis. Occasional Publications in Academic Computing No. 16. Dallas, TX: Summer Institute of Linguistics. ISBN 0-88312-639-7, 273 pages.

Barker C. (1989). Extrametricality, the cycle, and Turkish Word Stress. UCSC qualifying paper

Batusek R. (2002). A Duration Model for Czech Text-to-Speech Synthesis. Proc. Of Speech Prosody 2002, France

Dusterhoff E. K., Black W. A., Taylor P. (1999). Using Decision Trees within the Tilt Intonation Model to Predict F0 Contours. Proc. Of EUROSPEECH

Dybvig, R. K. (1996). The Scheme Programming Language, Ansi Scheme. Prentice Hall PTR, New Jersey

Febrer A., Padrell J., Bonafonte A. (1998). Modeling Phone Duration: Application to Catalan TTS. Proc. of 3rd ESCA Workshop on Speech Synthesis, Australia

Klatt H. D. (1987). Review of Text-to-Speech Conversion for English. Journal of the Acoustical Society of America, vol. 82, pp. 737--793

Lee S. (1999). Tree-Based Modeling of Prosody for Korean TTS Systems. PhD. Thesis

Möbius B., Santen P. H. J. (1996). Modeling Segmental duration in German Text-to-Speech Synthesis. Proc. of ICSLP, Vol. 4, pp 2395-2398

Oflazer K. (1994). Two-level description of Turkish morphology. Literary and Linguistic Computing, 9, 175-198.

Ross P. (2000). Data Mining.
http://www.dcs.napier.ac.uk/~peter/vldb/dm/dm.html

Salor Ö., Pellom B., Ciloglu T., et.al. (2002). On Developing New Text and Audio Corpora and Speech Recognition Tools for the Turkish Language. Proc. of the ICSLP, Denver, USA, Sep

Santen P. H. J. (1997). Prosodic Modeling in Text-to-Speech Synthesis. Proc. of EUROSPEECH'97, Rhodes

Venditti J. J., Santen P.H. J. (1998). Modeling Vowel Duration for Japanese Text-to-Speech Synthesis. Proc. of ICSLP'98, Australia

Venditti J. J., Santen P.H. J. (1998). Modeling Segmental Duration for Japanese Text-to-Speech Synthesis. Proc. of 3rd ESCA Workshop on Speech Synthesis, Australia

Witten H. I., Frank E. (1999). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kauffman Publishing