Segmental Duration Modeling in Turkish

Ozlem OZTURK, Tolga CILOGLU[1]

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

Abstract

Naturalness of synthetic speech highly 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 using machine-learning algorithms, especially *Classification and Regression Trees* (CART). 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 (POS) information, word length in number of syllables and position of word in utterance extracted from a speech corpus of approximately 700 sentences. Obtained models predict segment durations better than mean duration approximations (~0.77 *Correlation Coefficient*, CC, and 20.4 ms *Root-Mean Squared Error*, RMSE). Attributes phone identity, neighboring phone identities, lexical stress, syllable type, POS, phrase break information, and location of word in the phrase constitute best predictor set for phoneme duration modeling.

Index Terms: segmental duration modeling

1. Introduction

Prosody refers to characteristics of speech such as intonation, timing, stress, loudness, and other acoustical properties imposed by articulatory, emotional, mental, and intentional states of the speaker. One of the most prominent components of prosody is considered as timing or duration. Duration plays as much important role as intonation in the encoding/decoding of speech by the speaker/listener. Duration can be defined as the time taken to utter an acoustic unit such as phoneme, syllable, etc. In this study, it is aimed to predict the phoneme durations of a Turkish sentence given its written form so that resultant phoneme durations resemble those of natural speech.

Various methods exist for building duration models [1]-[13]. Those that combine linguistic expert knowledge with manual analysis of quite limited amount of data are generally known as *rule-based* approaches. Rule-based heuristic systems such as Klatt's duration modeling system [8] which assigns a percent increase or decrease to the inherent duration of the segment which is specified as one of its distinctive properties are case-dependent and hence, exhibit less flexibility.

State-of-the-art is dominated by *corpus-based* approaches [1]-[7] [9]-[13]. They have appeared due to the increasing computational power and availability of large corpora. Corpus-based (data-driven) approaches utilize large text and speech corpora to map linguistic features such as phonetic context, number of words in sentence, number of syllables in word to timing of synthetic speech. Corpus-based modeling involves machine learning techniques such as Artificial Neural Networks (ANN) [3][5][6][12], and Classification and Regression Trees (CART) [1] [4] [7] [9] [10] to reveal the relation between timing of speech and linguistic features.

In this study, a CART based method is used to map linguistic features to phoneme durations. CART is a predictive model that can be viewed as a tree. CART provide *interpretability* so that underlying dynamics between input space and outputs can be clearly identified. They can also be applied to any data and requires less parameter tuning [14].

For phoneme duration modeling, a collection of attributes is defined such as phoneme identity, left/right context, lexical stress, Part-of-Speech (POS), and etc. Relevancies of attributes affecting phoneme duration in Turkish are determined by means of statistical analyses. Using classification and regression trees durational attributes are mapped to phoneme durations. The performance of the mapping is evaluated by objective measures such as correlation coefficient (CC), mean absolute error (MAE), and root mean squared error (RMSE).

Focusing on the most influencing research, an overview of different approaches to duration modeling is given in the Introduction. Section 2 introduces the text and speech databases used in feature extraction and model development. Section 3 introduces attributes used for phoneme duration modeling. Phoneme duration modeling studies are presented and corresponding results are discussed in Section 4. Last section comprises final conclusions and future directions.

2. Speech Database

Corpora design is a fundamental issue for building appropriate prosody, in particular, duration models. A speech database can be built randomly or by means of optimizing the units acoustically or with respect to their textual properties. The database used in this study was optimized to provide phonetic variability. The speech files are annotated with respect to SAMPA [15] units first by forced alignment then by manual corrections. No allophonic variations are used for the vowels and the consonant 'r' but allophones of 'g', 'k', 'n', and 'l' are used. Long vowels are distinguished from their short counterparts. Resulting speech corpus contains 36855 phonemes. The lists of phonemes and their frequency in the speech corpus are given in Table 1. Phoneme duration distribution of the corpus approximates gamma distribution.

3. Feature Set

Various durational attributes have been used in the literature for duration modeling. Some of them are listed in Table 2.

Features that are considered to affect phonetic duration in Turkish are determined and extracted from both speech and text corpus. Each phone in the database is assigned a feature vector describing the phone and the values of its attributes. The attributes and their values used in this study can be divided into two groups as categorical and numerical.

Table 1. *SAMPA symbols of the phonemes in speech corpus and their frequencies*

Phone	Freq.	Phone	Freq.	Phone	Freq.	Phone	Freq.
a	5790	gj	546	n	3627	t	1761
a:	268	h	459	N	156	tS	547
b	1292	1	2415	o	1521	u	1980
c	1007	1:	42	o:	31	u:	84
d	2142	i	4378	2	493	v	391
dZ	731	i:	141	2:	1	w	178
e	4451	j	1931	p	436	y	972
e:	94	k	1389	r	3570	y:	14
f	235	l	1656	s	1503	z	757
g	163	5	1705	S	747	Z	133
G	685	m	2228	silence	2000		

3.1. Categorical features

- Identity: The phonetic description of current, preceding and following phonemes. Each phoneme can take one of 42 SAMPA symbols and each preceding/following phoneme can be either one of 42 SAMPA or silence.
- Lexical Stress: There exist two levels for lexical stress: Accented (A) or Not-Accented (NA). A segment is associated with an A if the vowel of the parent syllable is stressed and an NA otherwise.
- Position in Syllable: A three level representation is used to code phoneme position in syllable: Nucleus (N), Onset (O) and Coda (C).
- Syllable Type: Two levels are used to denote parent syllable types: Heavy (H) and Light (L).
- Part-of-Speech: Each phoneme in the database is annotated with the major POS tag of the parent word such as NOUN, PRONoun, VERB, QUEStion, INFinitive, POSTPronoun, CONJunction, ADVerb, ADJective, CompoundNOUN, or EXClamation. These tags are obtained through a morphological analysis procedure.
- Phrase Break Information: Speech corpus has been evaluated perceptually several times and major perceptual breaks in the utterances are marked manually. The marks mainly correspond to the speaker's breathings. The feature is represented by three levels: Segment takes a Phrase Initial (PI) value if it immediately follows a phrase break, a Phrase medial (PM) value if there is no phrase break engagement and a Phrase Final (PF) if a phrase break immediately follows the segment.

3.2. Numerical features

- Syllable Position in Word: Syllables of the same word are counted from the left starting from 1. The database contains words of at most 10 syllables; however, there is no word that contains 9 syllables.
- Word/Syllable Position in Sentence: Words/Syllables are counted from left starting from 1. The longest sentence contains 19 words and 45 syllables. All phonemes of the parent word take the same value.
- Word Length: Each phoneme of the same word is annotated with the total number of syllables in that word. The attribute values are numeric and ranges from 1 to 10.
- Total number of Words (Syllables) in Sentence (Word): Each phoneme of a sentence is represented by the total number of words (syllables) in the sentence (word). The value of the feature changes in between 3 and 19.
- Number of Words from (to) the Preceding (Following) Phrase Break: The attributes identify the number of words between the parent word and the preceding (following) phrase break counting from 0.
- Number of Syllables from (to) the Preceding (Following) Phrase Break: This attribute is almost the same as the number of words from the preceding phrase break attribute counting from 0.

3.3. Statistical Observations

- Phrase-final lengthening is observed.
- Differences between durations of voiced and voiceless consonants are significant; voiceless consonants are longer (~30-40 ms) in duration than their voiced counterparts.
- When followed by a voiced consonant, phoneme durations increase except for vowel + voiced-plosive combination. Voiced-fricative-followers influence voiceless phoneme durations (~30 ms) more than voiced plosive (~12 ms) and affricate (~14 ms) followers.
- Voiced consonants are slightly longer when they occur in coda position. For example, of the two '*l*'s of the word *eylülden* given in Figure 1, the one at onset position has a duration of 56 ms while the other has 78ms. Affricates, nasals, plosives and liquids occurring at onset are significantly longer in duration (around 20-30 ms) than the ones occurring at coda.

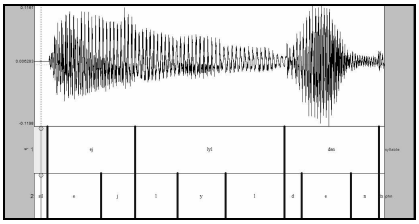

Figure 1. *Speech waveform, phonetic and syllabic segmentation of the word 'ej-lyl-den'.*

- Phonemes have shorter durations in open syllables than in closed syllables. For example, the durations of '*e*'s in the word *türkiyeye* are around 37 ms, while its counterpart in closed syllable of the word *tamamen* has a duration of 103 ms.
- Durations of word-initial and word-final syllables' phonemes are longer
- Single-syllable-words' phonemes attain maximum average duration.
- Phonemes occurring at sentence-final-words are longer than the ones occurring at sentence-initial or sentence-medial words. Percentage lengthening is approximately 20%. For example, '*r*' of the word *diyor* located at sentence-final position has a duration 149 ms, while it's the one in word *itibaren* at sentence medial

position has a duration of 26 ms. However, this lengthening is mostly due to the phonemes occurring at sentence boundary and should be related to phrase final lengthening.

- Average phoneme duration is shortened as the number-of-syllables-in-word increases. Phoneme duration is 41% longer in single-syllable-words than in words having ten syllables.
- Average phoneme duration increases as the number-of-words-from-preceding-phrase-break increases.
- Words immediately followed by a phrase break attain maximum average phoneme durations.

4. Segmental Duration Modeling

Experiments for developing duration models are performed with the REPTree algorithm of WEKA [14]. Vowel and consonant durations are predicted at the same time. Prediction performance of each experiment is evaluated using mean absolute error (MAE), root mean squared error (RMSE), and correlation coefficient (CC).

The speech database is split into two subsets: training dataset is used to develop duration models and test dataset is used to evaluate the performance of the model on unseen data. The test set consists approximately 20% of the database and the remaining phonemes constitute the training set (80%). The total numbers of instances in the training and test sets are 29527 and 7328, respectively.

Each attribute described in Section 3 is evaluated by the tree building method to observe the individual affects on phoneme duration. Phoneme Identity is considered as the discriminating attribute; hence corresponding results are used as a reference (baseline) for the rest of the experiments. Individual performances of the attributes in terms of CC, MAE and RMSE are given in Table 3. Phoneme Identity (1) is the best and Preceding/Following Phoneme Identities (2-3) is the second best predictor. However, such kind of an evaluation does not give an idea about the relative relevance of the attributes when all combinations are considered.

Best attribute set is constructed by means of a greedy procedure which starts with an empty or pre-selected set of attributes and adds more attributes as the resulting learning algorithm's performance is improved. Model performances obtained at each step are given in Table 4.

The optimum predictor set is composed of Phoneme Identity, Preceding/Following Phonemes Identities, Lexical Stress, Syllable Type, POS, Phrase Break Information, and Number of Words to the Following Phrase Break. It should also be noted that when all attribute set is considered, resultant performances are worse than those obtained by using optimum attribute set.

4.1. Improvements

According to our observations, neighboring phonemes turn out to be the most influential attributes on phoneme duration. Regarding this fact, current database is used to predict phoneme durations with increasing number of phonetic context, i.e. for example a window of 5 phonemes. Former results using only improved contextual information out performs the performance measures obtained using optimal attribute set such that the resulting tree predicts phoneme durations with a CC of 0.7815 and a RMSE of 20.0115ms (cf. Table 4).

5. Remarks

Within the scope of this study, phoneme duration modeling in Turkish is performed. To this aim, attributes that may affect phoneme duration in Turkish are defined and extracted from the developed speech and text corpora. Model development is performed using tree building by means of a self-learning algorithm. Learning algorithm is run on the training set and its performance is evaluated on test dataset, hence resulting performances can be regarded as the models' real performance on unseen data. A greedy approach is applied to find an optimal set of attributes to build a regression tree. Resulting model performance is comparable to those reported in literature.

One of the main causes of prediction error is the inevitable inconsistency in the segmentation of phonemes. Consequently, although it is known to be a difficult task, research on devising objective consistency measures in phoneme (or some other suitable unit) segmentation is considered to be an essential future work.

One other factor that may affect the performance duration modeling is the speaking rate. Although the speech database used in this study is developed under control, the speaker can not sustain a constant speaking rate since the recordings last a few days. In most of the studies, this phenomenon is underestimated. However, speaking rate does not affect phonemes' durations linearly. Therefore, the effects of speaking rate on segmental duration will be considered.

In concatenative synthesis, using variable unit sizes highly improves speech quality and eliminates the need for duration prediction. However, it still remains as a hard problem to accommodate the duration of larger units during concatenation since two consecutive segments taken from different contexts may not generate a natural sounding synthetic speech.

Performances of developed models are evaluated quantitatively throughout this study. However, prosody is meaningful perceptually. Hence, perceptual evaluations can be carried out to evaluate model performances as a future work.

6. Acknowledgements

We are grateful to E. Akdemir, T. Koç, and Y. Özbek for their support in segmenting the speech corpus. The study is supported by METU Research foundation (BAP-2005-03-01-06).

7. References

[1] Batusek, R., (2002), "A Duration Model for Czech Text-to-Speech Synthesis", in Proc. of Speech Prosody 2002, Aix-en-Provence, France, pp. 167-170.

[2] Campbell, N., (2000), "Timing in Speech: A Multi-Level Process", in M. Horne (ed), Prosody: Theory and Experiment, Kluwer Academic , Dordrecht, pp. 281-335.

[3] Chen, S. H., Hwang, S. H., Wang, Y. R., (1996), "A Mandarin Text-to-Speech System", in Computational Linguistics and Chinese Lang. Processing, Computational Linguistic Soc. of R.O.C., vol.1, no.1, pp. 87-100.

[4] Chung, H., (2002), "Duration models and the perceptual evaluation of spoken Korean", in Proceedings of Speech Prosody, Aix-en-Provence, France, pp. 219-222.

[5] Cordoba, R., Vallejo, J. A., Montero, J. M., Gutierrez-Arriola, J., Lopez, M. A., Pardo, J. M., (1999), "Automatic Modeling of Duration in Spanish Text-to-Speech System Using Neural Networks", in Proceedings of EUROSPEECH, Budapest, Hungary, pp. 1619-1622.

[6] Cordoba, R., Montero, J. M., Gutierrez-Arriola, J, Vallejo, J. A., Enriquez, E., Pardo, J. M., (2002), "Selection of the Most Significant Parameters for Duration Modeling in a Spanish TTS System Using Neural Networks", Computer Speech and Language, Elsevier, Vol. 16, pp 183-203.

[7] Febrer, A., Padrell, J., and Bonafonte, A., (1998), "Modeling Phone Duration: Application to Catalan TTS", Proc. of 3rd ESCA/COCOSDA Workshop on Speech Synt., NSW, Australia, pp. 43-46.

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

[9] Krishna, N. S., Murthy, H. A., (2004), "Duration Modeling of Indian Languages Hindi and Telugu", Proc. of 5th ISCA ITRW on Speech Synt., Pittsburgh, USA, pp.197-202.

[10] Lee, S.and Oh, Y. W., (1999a), "Tree-Based Modeling of Prosodic Phrasing and Segmental Duration for Korean TTS Systems", Speech Comm., Elsevier, Vol. 28, pp 283-300.

[11] Möbius, B. and van Santen, J. P. H., (1996), "Modeling Segmental Duration in German TTS", Proc. of Int. Conf. on Spoken Language Processing, Philadelphia, USA, Vol. 4, pp 2395-2398.

[12] Sreenivasa, K. R., Yegnanarayana, B., (2004), "Modeling Syllable Duration in Indian Languages Using Neural Networks" Proc. of Int. Conf. on Acoustics, Speech and Signal Processing, Quebec, Canada, pp. 313-316.

[13] Venditti, J. J., van Santen, J. P. H., (1998), "Modeling Vowel Duration for Japanese TTS", Proc. of Int. Conf. on Spoken Lang. Processing, Sydney, Australia, paper 0786.

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

[15] Wells, J.C., "SAMPA for Turkish", <http://www.phon.ucl.ac.uk/home/sampa/ turkish.htm>.

Table 2. *Some of the attributes used in literature.*

Languages	**Attributes**	**Segments**
Czech [1]	current/previous/next-phone-identities, syllable/word/phrase-lengths-in-phones, phone-position-in-syl.-from-beginning/end, phone-position-in-word-from-beginning/ end and word-position-in-phrase	Phoneme
English [2]	#-of-phones-in-the-syl., nature-of-syllabic-peak, position-in-tone-group, type-of-foot, stress and word-class	Syllable
Spanish [5][6]	phone-identity, contextual-phones, stress, stress-in-the-syllable, syllable-beginning-with-vocal, diphthong, phone-in-a-function-word, phrase-type, position in phrase and number-of-units-in-the phrase	Phoneme
Catalan [7]	vowel-identity, stress, sentence-position, post-vocalic-phone-class and manner-of-articulation	Phoneme
Hindu and Telugu [9]	seg.-identity, seg.-features, previous/next-seg.-features, parent-syl.-structure, position-in-parent-syl., parent-syl.-initial/final, parent-syl.-position-type, #-of-syl.-in-parent-word, position-of-parent-syl., parent-syl.-break-information, phrase-length-in-#-of-words, position-of-phrase-in-utterance, and #-of-phrases-in-utterance	Phoneme
German [11]	Segment-identity, segment-type, word-class, position-of-phrase-in-utterance, phrase-length-in-number-of-words, position-of-word-in-phrase, word-length-in-number-of-syllables, position-of-syllable-in-word, stress, segment-position-in-syllable, segmental-context, segmental-context-type	Phoneme
Japanese [13]	current/preceding/following-phone-identities, left/right-prosodic-context, accent-status, syl.-structure and special-morpheme-status	Vowel

Table 3. *Single attribute performances given in increasing RMSE order.*

Index	**Attribute**	**CC**	**MAE (ms)**	**RMSE (ms)**	**Index**	**Attribute**	**CC**	**MAE (ms)**	**RMSE (ms)**
1	Phoneme Identity	0.5958	18.2003	25.7872	7	Syllable Position in Word	0.1218	24.4285	31.8344
2-3	Preceding/Following Phoneme Identities	0.53	20.8325	27.1914	9	POS	0.0873	24.7954	31.9577
5	Position in Syllable	0.3106	23.3704	30.5724	16	Number of Syllables from the Prev. Phrase Break	0.0713	24.6631	31.9872
12	Phrase Break Information	0.2641	24.3414	30.9329	8	Word Position in Sentence	0.0539	24.7744	32.0196
17	Number of Syllables to Fol. Phrase Break	0.2443	24.5178	31.0977	15	Syllable Position in Sentence	0.0386	24.7759	32.0445
6	Syllable Type	0.1473	24.4769	31.7265	13	Number of Words from the Prev. Phrase Break	0.0234	24.7784	32.0597
14	Number of Words to the Fol. Phrase Break	0.1381	24.8184	31.7601	4	Lexical Stress	0.0193	24.7751	32.0604
10	Total Number of Syllables in Word	0.1212	24.5606	31.8327	11	Total Number of Words in Sentence	0	24.7806	32.0658

Table 4. *Prediction error performances obtained at each level of greedy algorithm.*

Level	**Attributes**	**CC**	**MAE (ms)**	**RMSE (ms)**
1	1	0.5958	18.2003	25.7872
3	1, 2-3	0.7576	15.1605	20.9321
4	1, 2-3, 6	0.7706	14.7089	20.44
5	1, 2-3, 6, 12	0.7744	14.6039	20.2937
6	1, 2-3, 6, 9, 12	0.7772	14.5887	20.184
7	1, 2-3, 4, 6, 9, 12	0.7798	14.5613	20.0792
8	1, 2-3, 4, 6, 9, 12, 14	0.7806	14.5574	20.0456
9	1, 2-3, 4, 6, 9, 12, 14, 7	0.7807	14.5607	20.0478
17	All	0.7718	14.6678	20.4236
Using a window of 5 phonemes only		0.7815	14.5819	20.011