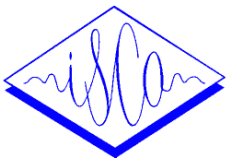

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

DURATION MODELLING FOR THE GREEK LANGUAGE

G. Epitropakis, D. Tambakas, N. Fakotakis and G. Kokkinakis

Wire Communications Laboratory
University of Patras
Patras, Greece

ABSTRACT

A novel two-level model of timing in speech is described. In this model, phoneme durations are first calculated at the lower-level of segmental characteristics and contextual effects, and at a second stage, the already predicted durations are modified to reflect the higher-level of rhythmic and structural organisation of the utterance. The model is based on results obtained from analysis of isolated words for the first level and results from continuous speech analysis for the second one. The complete model can be used in both isolated word and continuous speech recognition systems, and Text-To-Speech systems. The model has been implemented in the Greek language.

Keywords: *Duration models, Text-To-Speech systems, Isolated Word Recognition systems, Greek language.*

10.21437/Eurospeech.1993-451

1. INTRODUCTION

The prediction of the phoneme duration in a specific phonemic and prosodic context is a crucial factor for the performance of both Text-To-Speech (TTS) and Speech-To-Text (STT) systems. At least ten factors [1,2,3,4] have to be taken into account in order to accurately predict the phoneme duration: segmental identity, preceding phonemes, following phonemes, syllabic stress, length of the word that the phoneme belongs to, within-word position, word emphasis, nature of the phrase that the phoneme belongs to, phrase length and within-phrase position. In an effort to model the contribution of all these factors for predicting speech timing, various durational systems have been proposed in the past:

- *Sequential rule systems* [5], that modify a default value of the left-hand term according to the right-hand parts of the rules. The main disadvantage of such rule-based systems is that someone has to exactly determine the order in which the rules have to be fired and this is not very simple in most of cases.
- *Neural network and tree-based systems* [6] with good performance but time consuming and complicated training too. In addition, in these systems it is very difficult for the researcher to manually refine the weights used, since the latter are lacking in physical substance.
- *Duration models* [1,7,8] where equations employing factor scales are used. Although these models seem to overcome the deficiencies of the previously described systems, they have proved to fail when are rigorously tested [6]. This is mainly due to the way that the interaction of the factors taken into account, is perceived. Namely, the models have come out from data sets in which only two or three factors were varied and the results obtained could not be generalized.

The model described in this article assumes timing to be a result of interaction between "processes" operating in two discrete levels in speech hierarchy [9,10]. By the term "process", we mean the contribution of the ten factors, mentioned above and, influencing the prediction of the phoneme durations. The model first calculates phoneme durations taking into account the first six factors related to the phoneme's segmental characteristics, the manner, the place and the context of coarticulation. The way that these factors affect the phoneme durations is established upon the results obtained from analysis of a speech database consisting of isolated words. Then the model supplements the already predicted durations with the appropriate elements of the higher-level prosodic influences, i.e. the phrasal and rhythmic processes operating at the level of the syllable and above. The assignment of such characteristics is performed by using the last four factors in a way that has come out from continuous speech analysis. The first-level model can be used in isolated word speech recognition systems, while the finally resulting model can be used in continuous speech (recognition and synthesis) systems. The parameters of the complete model have been established for each phoneme of the Greek language through extensive analysis of the speech data.

2. SPEECH DATA-BASE

A speech database has been created for analysis purposes. It consists of three data-sets:

Set-A: 500 words with good coverage of phonemes, spoken by 8 speakers (4 male, 4 female).

Set-B: 180 sentences that cover the greatest possible syntactic structures of the Greek language, spoken by 2 speakers (1 male and 1 female speaker already recorded for the Set-A construction).

Set-C: Utterances of 10 minutes total duration collected from radio broadcasters.

The above speech database has been labelled manually. A dynamic database has been constructed containing a total of about 35,000 entries. Each entry consists of the phoneme label, its

duration, its phonemic context and the length of the word it belongs to. For the Set-B and Set-C additional information has been included referring to the grammatical class and the syntactic category of the word that the phoneme belongs to. In addition, appropriate prosodic labels have been used indicating the location of the phoneme at the end of an intonational phrase (pre-boundary segment), the type and the length of the intonational phrase. At last, a user-friendly interface has been constructed in order to retrieve the appropriate information.

3. ANALYSIS METHODOLOGY

The aim of a duration model developed in the context of a speech system (TTS/STT), is to separate the wide interval obtained by natural speech analysis, in which the duration of a specific phoneme varies, into shorter intervals which would describe more accurately the phoneme duration in different cases. This is necessary because using the mean duration of phonemes, neither a TTS-system could output a naturally-sounded speech, nor a STT-system could recognize the input speech with high accuracy. The additional task of the duration model is to describe in terms of well-weighted factors the sequence of the textual features that cause the phoneme duration to be in the appropriate time interval.

By their very nature as interactively functioning features, durational factors are difficult to be observed, analyzed and transcribed if one tries to investigate all of them at the same time. The analysis methodology we propose here, tries to isolate subsets of factors from speech data and to treat these subsets separately. The whole analysis is carried out on two basic successive levels:

Analysis Level-1

Step-1: The Set-A of the speech database consisting of well-varying in length words is analyzed assuming that from these data only the contribution of the first three factors (see Table 1) can be modeled.

Step-2: The time intervals obtained from step-1 are refined by exploiting the contribution of the word length and within-word position that the phoneme belongs to.

Analysis Level-2

Since the contribution of the supra-segmental prosodic factors is up till now assumed to be negligible (the analysis level-1 is carried out on isolated word recordings), we have to exploit their functional attributes by analysing continuous speech recordings. Extensive measurements have proved that the already extracted duration rules could generally be used also for continuous speech, but the predicted durations have to be reduced appropriately, since the speaking rate is higher in continuous speech. Furthermore, using the labelled Set-B and Set-C of the speech database, the last four factors of Table 1 are modeled, in such a way that the predicted (from the results of the level-1 analysis) durations are modified to reflect the prosodic effects of continuous speech.

Concerning the accentuation factor, the problem is simply faced by treating the accented phonemes as different phonemes in the Greek phonetic alphabet.

Table 1 summarizes the factors to be taken into account for predicting the appearance of the phoneme duration in a specific short time interval and the corresponding analysis level that our method proposes for each factor's study and further modelling.

Factor	Analysis Level
Segmental identity	Isolated Word speech data (Level-1)
Preceding phonemes	Isolated Word speech data (Level-1)
Following phonemes	Isolated Word speech data (Level-1)
Accentuation	Different phonemes
Within-word position	Isolated Word speech data (Level-1)
Word length	Isolated Word speech data (Level-1)
Word emphasis	Continuous Speech analysis (Level-2)
Phrase nature	Continuous Speech analysis (Level-2)
Phrase length	Continuous Speech analysis (Level-2)
Within-phrase position	Continuous Speech analysis (Level-2)

Table 1: Factors influencing the phoneme duration and their treatment in successive analysis levels.

3.1 Analysis Level-1

In this level of the analysis procedure only data from isolated word recordings are treated.

All the phonemes of the Greek language (37, including the stressed and unstressed vowels as different phonemes) have been classified according to the manner and the place of the coarticulation during the time of their production. Table 2 shows the groups in which the phonemes (CPA) of the Greek language have been classified.

Consonants	*Labial*	p, b, f, v
	Dental	t, d, T, D
	Sibilants	C/, Z/, s, z
	Palatal/Velar	k, g, G/, X, K, g/, C
	Alveolar	l, r, L
	Nasal	m, n, N, n/, m/
Vowels	*Open*	a, e, o, 'a, 'e, 'o
	Closed	i, u, 'i, 'u

Table 2: Classification of the Greek phonemes according to the manner and the place of coarticulation.

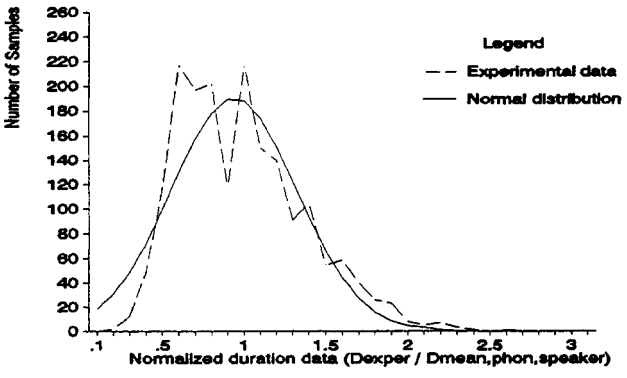

Fig. 1: Normalized experimental data and, normal distribution of the Greek phoneme /i/.

For each phoneme of the Greek language, the mean, maximum and minimum duration for every different speaker has been estimated. In order to avoid speaker variability effects, all the experimental data (Set-A) have been normalized according to the phoneme's mean duration for each different speaker. Figure 1 shows as an example, the normalized distribution of the

experimental data (dashed line), and the normal distribution (solid line) that has been proved to fit better to the data concerning the phoneme /i/. For presentation reasons, we will refer from now on to the normal distribution, although other mathematical distributions (beta, gama, etc.) proved to be the appropriate ones for other phonemes of the language.

3.1.1 Analysis Level-1: *Step-1*

Assuming that the phoneme's mean duration (D_m) is the default value, we examine the effect of the preceding and following phonemes on its variations. As expected, the mean duration changes are subject to contextual constraints. The contextual effect on a predicted duration (D_{pre}) is modeled by the multiplication factor $p_{context}$ as follows:

$$D_{pre} = D_m \ (\ 1 + p_{context} \ ) \qquad (1)$$

For each phoneme of the language, a two dimensional matrix (Table 3) is constructed with rows including all possible preceding contextual groups and columns including all possible following contextual groups. The cells of this matrix contain two numbers: The first one, is the $p_{context}$ coefficient for the specific environment of the phoneme and the second one is the time interval in which the duration of the phoneme in the specific environment could vary, as a percentage of the distribution width ($D_{maximum}$ - $D_{minimum}$). E.g. the phoneme /i/ in the context dental (left) and palatal (right) has a $p_{context}$=-.26 and its predicted duration varies in an interval equal to 44% of the whole interval of /i/ variation. Additional information needed when the phoneme is located in the start or the end of a word is also included in the matrixes.

Left \ Right	Labial	Dental	Sibilant	Palatal	Alveolar	Nasal	Open	Closed	End
Labial	-.04 / 22	-.15 / 19	-.48 / 9	-.32 / 29	-.49 / 17		.05 / 20	-.36 / 6	.37 / 33
Dental		-.15 / 25	-.2 / 29	-.26 / 44	-.23 / 29	-.4 / 30	-.04 / 40	-.09 / 33	.5 / 44
Sibilant		-.09 / 36	-.26 / 18	-.23 / 20	-.1 / 27	-.32 / 20			.33 / 47
Palatal	-.18 / 16	-.34 / 23	-.45 / 35	-.16 / 24	.33 / 29	-.47 / 23	-.26 / 21	-.26 / 21	.48 / 42
Alveolar	-.2 / 20	.12 / 34	-.32 / 29	-.28 / 23	-.29 / 23	-.36 / 28	-.09 / 47	-.09 / 47	.28 / 40
Nasal		-.26 / 13	-.23 / 27	-.21 / 28	-.16 / 20		.01 / 22	.01 / 22	.4 / 46
Open									
Closed									
Start	-.46 / 23	.04 / 42	-.65 / 15	-.45 / 24					

Table 3: $P_{context}$ coefficients and time intervals produced for the phoneme /i/ under all the possible contexts.

Empty cells in the matrix indicate that the corresponding environments have not been detected in the database used.

3.1.2 Analysis Level-1: *Step-2*

Although the above extracted coefficient $p_{context}$ predicts the mean change of duration from the default value (D_m) due to the specific context, the time intervals in which D_{pre} could vary are too wide for some cases. Figure 2 shows the time intervals for three of the rules produced.

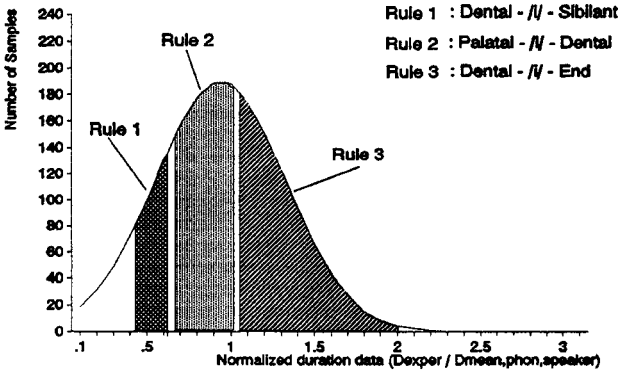

Fig. 2: Resulting time intervals, in which the predicted duration for the phoneme /i/ could vary in a specific context.

For minimizing the time intervals, in order to achieve more fine prediction rules, additional analysis has been carried out, taking into account the effect of the word length factor. From this analysis, an improved version of equation (1) has been produced:

$$D_{pre} = D_m \ (\ 1 + p_{context} \ ) + w_l \ (\ D_{max} - D_{min} \ ) \qquad (2)$$

where,

w_l, is a factor corresponding to the word length,

D_{max}, is the phoneme's statistically obtained maximum duration, and

D_{min} is the phoneme's statistically obtained minimum value.

Figure 3 gives the modified time interval for the third rule shown in figure 2, which now includes two intervals a and b for polysyllabic words and words of with few syllables respectively. Figure 4 shows the minimization of the time intervals (refinement) after the second stage approach.

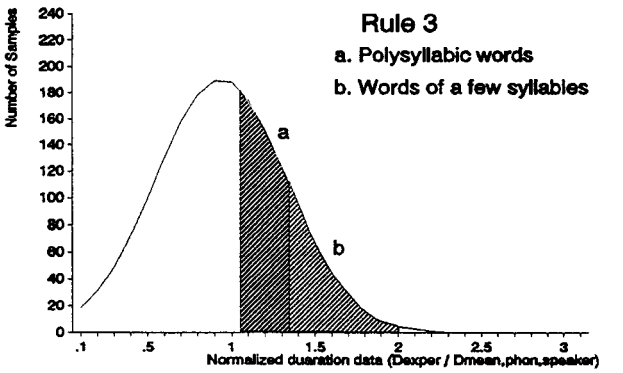

Fig. 3: Time interval of rule-3 (see fig. 2), as modified by the word length factor w_l.

The w_l coefficient is estimated through analysis of the experimental data. Generally, an appropriate definition of w_l for each phoneme and for each specific context has to be achieved, in order to minimize the time intervals. For the example given in figure 3, the rule-3 is split into two subrules by the application of w_l = 0.0873 for words less than 5 phonemes (case b) and w_l = -0.437 for the longer words (case a). This modification results to time intervals of 28% and 16% of the distribution width, against of 44% of the step-1 estimated time interval.

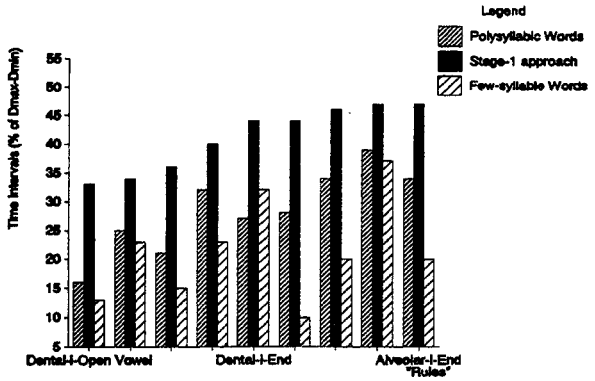

Fig. 4: Finally resulting time intervals for the "rules" predicting the mean duration of phoneme /i/.

3.2 Analysis Level-2

The important assumption of our methodology, that the duration model estimated from isolated word speech data could be used as the basis for modelling continuous speech durations , should be verified on the experimental data of Set-B and Set-C. To this end, since the speaking rate is higher in continuous speech, some modifications on the extracted rules had to be made. Namely:

a) The absolute values (in msec) of the mean, maximum and minimum duration of each phoneme had to be reduced by a factor of 15%.

b) The predicted coefficients $p_{context}$ had to be reduced by a factor of 10%.

c) The predicted coefficients w_l were also valid, but for continuous speech, the concept of word had to be broadened in order to include the speech surrounded by two successive speech pauses.

After these transformations, an extensive analysis of Set-B has been carried out, in order to model the prosodic factors influencing the phoneme durations.

The results of this analysis led us to modify the equation (2) as follows:

$$D_{pre} = [\, D_m \,(\, 1 + p_{context} \,) + w_l \,(\, D_{max} - D_{min} \,) \,]\, pr_{context}$$

where, $pr_{context}$ is a multiplicating factor reflecting the increment of the phoneme's duration when it is located at the end of an intonational phrase.

4. MODEL PERFORMANCE

The described model has been applied both to an isolated word speech recognizer and to a text-to-speech synthesizer for the Greek language. In both cases the overall performance of the systems was significantly increased in comparison to results achieved with mean duration values. Measurements are continued with enlarged testing material in order to establish quantitative performance measures of the model for the Greek phonemes and to highlight possible improvements.

5. CONCLUSION

A novel method for modelling phoneme durations has been presented. This method models the various factors affecting phoneme durations in two different levels of speech hierarchy, assuming independence between the factors. The criterion used for treating successive levels is the minimization of the predicted time intervals in which the duration of a phoneme could vary. In this way, in order to predict the phoneme durations three coefficients have been introduced, which depend on the phoneme's context, the length of the word that the phoneme belongs to and the position of the phoneme in the intonational phrase.

The method proved to be time consuming, but the model established has some strong advantages: a) The factor scales (values) that feed the model have a physical meaning. For example, positive values of $p_{context}$ correspond to durations greater than the mean values, while positive values of w_l correspond to polysyllabic words. b) The model is suitable both for isolated and continuous speech. c) The performance of the system evaluated in TTS and STT systems is significantly better than that achieved by using mean duration values.

REFERENCES

[1]: *Klatt, D.H.:* Review of text-to-speech conversion for english. Journal of Acoustical Society of America, 82(3), pp.737-793, 1987

[2]: *Emerard, F.; Mortamet, L.; and Cozannet, A.:* Prosodic processing in a text-to-speech synthesis system using a database and learning procedures. Talking Machines:Theories, Models, and Designs, Baily, Benoit, and Sawallis, pp.225-247, 1992 Elsevier Science Publishers B.V.

[3]: *Umeda, N.:* Vowel duration in American English. Journal of Acoustical Society of America, 58(2), pp.434-445, 1975

[4]: *Kaiki, N.; Takeda, K.; and Sagisaka, Y.:* Linguistic properties in the control of segmental duration for speech synthesis. Talking Machines:Theories, Models, and Designs, Baily, Benoit, and Sawallis, pp.255-263, 1992 Elsevier Science Publishers B.V.

[5]: Esprit project 2104: POLYGLOT-I

[6]: *Riley, M.D.:* Tree-based modelling of segmental durations. Talking Machines:Theories, Models, and Designs, Baily, Benoit, and Sawallis, pp.265-273, 1992 Elsevier Science Publishers B.V.

[7]: *van Santen, J.P.H.; & Olive, J.:* The analysis of Contextual Effects on segmental Duration. Computer Speech & Language, 4, pp.359-390, 1990

[8]: *van Santen, J.P.H.:* Deriving text-to-speech durations from natural speech. Talking Machines:Theories, Models, and Designs, Baily, Benoit, and Sawallis, pp.275-285, 1992 Elsevier Science Publishers B.V.

[9]: *Beckman, M.J.:* Metrical structure versus autosegmental content in phonetic interpretation. Proceedings of the XIIth International Congress of Phonetic Sciences, Aix-en-Provence, France, 1991

[10]: *Campell, W.N.:* Syllable-based segmental duration. Talking Machines:Theories, Models, and Designs, Baily, Benoit, and Sawallis, pp.211-224, 1992 Elsevier Science Publishers B.V.