Segmental Durations Predicted With a Neural Network

João Paulo Teixeira and Diamantino Freitas***

*Polytechnic Institute of Bragança and
**Faculty of Engineering of University of Porto, Portugal
joaopt@ipb.pt, dfreitas@fe.up.pt

Abstract

This paper presents a segmental durations' model applied to the European Portuguese language for TTS purposes. The model is based on a feed-forward neural network, trained with a back-propagation algorithm, and has as input a set of phonological and contextual features, automatically extracted from the text. The relative importance of each feature, concerning the correlation with segmental durations and improvements in the performance of the model, is presented. Finally the model is evaluated objectively and subjectively by a perceptual test.

1. Introduction

The present model is part of a global prosody model, which is presently under development in the authors' Institutions, and the basic motivation is to use it in a Portuguese TTS system.

Other durations' models are published in literature for other languages, and can be grouped in rule-based models, mathematical models and statistical models.

Rule-based models should allow a straightforward knowledge of the effects of each feature in the duration of the segments. Examples of this type of models are the Klatt rule-based model [1], the rule-based algorithm for French [2], presented by Zellner for different speech rates, and the look-up-table for Galician [3].

Mathematical models usually appear as a Sum-of-Products, where the features are statistically weighted and summed to produce the segmental duration [4].

Statistical duration models become more and more used with the availability of large phonetically labelled data-bases. Neural networks and regression trees are the more often used tools, applied in different ways for different languages and using different type of segments. Campbell [5] introduced the concept of Z-score to distribute the duration estimated by a neural network, for a syllable, among its segments. He argued in favour that the syllable is the more stable unit. Barbosa and Bailly also presented a two steps model for French [6] and Brazilian Portuguese [7]. In the first step, using a neural network, they estimate the duration of the Inter-Perceptual Centre Groups (IPGC), arguing that is the more stable unit. In the second step they distribute the duration of the IPCG among its segments, using the Z-score concept. This model can deal with different speech rates, and pauses. Other neural network-based models were also presented for Spanish [8] and Arabic [9]. Example of a CART-based model applied for Korean can be found in [10].

In the next sections a neural network based model is presented to predicted segmental durations using a set of features carefully analysed for European Portuguese. In section 5 the model is evaluated.

2. Corpus

The data used for training and test was extracted from the FEUP-IPB database [11]. This database consists of several texts extracted from newspapers that were read by a male professional radio broadcast speaker at the average speech rate 12.2 phonemes/second. This database was then manually labelled in three levels: 1- phonetic level, considering 46 different segments classes and also marking the tonic syllable; 2- word level marking beginning and end of words; and 3- phrase level marking the beginning and end of phrases as well all orthographic marks. Seven texts from the data-base were used, in a total of 101 paragraphs, mainly with declaratives and interrogatives of very different dimensions, from one to one hundred words, consisting in a total of 18.700 segments in 21 minutes of speech. The training set consists of 6 texts containing ≈15000 segments and the test set consists in one text containing ≈3000 segments. The relative frequencies of the phonemes are identical in both sets.

3. Features

The considered features were extracted by processing data from the corpus' labels with algorithms for syllabification [12] and grouping words in the so-called accent groups. These groups act like prosodic words aggregating neighbour particles, and were created according to following rules:

- Groups have more than 2 syllables in total.
- Groups never end with words of less than 3 phonemes.
- Phrase marks are always group boundaries.
- If more than one tonic syllable exists in the assembled group of words, then only the last is considered as tonic.

An example of application of the concept of accent groups is presented in the following sentence ('a strong reserve with justice situation'): "*uma forte / reserva / em relação / à situação / da justiça"*.

A large number of features were considered as candidates in the first phase of the work. The final set was established by the study of each feature considering its linear correlation with the output and observing its influence in the performance of the model by taking it out of the initial set of features. Some times more than one feature were considered as a group and taken out together to check for consistency. The known interaction between some features [4], changes the measured influences on the duration when they were considered isolated or as a group. The correlation coefficient appears to be a good indicator for the influence of the features, but for the ones that

are coded in more than one node, it is difficult to connect its correlations with the feature's importance.

Features were coded in different ways taking in consideration the minimization of number of nodes and the maximization of performance.

The final list of features, the correlations (r) with segment's durations and the number of nodes used to code, are presented in ***Table 1***. Detailed specifications of each feature are presented in the following (phonemes are represented in SAMPA code):

Table 1: Segment Features.

Phonologic level	*Feature*	*# nodes*	*Correlation |r|*
Phoneme	Segment identity	44	0.01 to 0.26
	Consonant in the end of word	1	0.08
Phoneme context	Previous segment (-1)	20	0.05 to 0.23
	Next segment (+1)	12	0.05 to 0.28
	Next segment (+2)	4	0.08 to 0.14
	Next segment (+3)	2	0.05 to 0.11
Syllable	Type	1	0.18
	Vowel	1	0.21
Syllable Context	Type of previous syllable	1	0.06
	Vowel in previous syllable	1	0.08
	Vowel of next syllable	1	0.15
	Distance to tonic syllable	1	0.15
Foot	Position in group	2	0.03 to 0.15
	Position in Phrase	2	0.04 to 0.24
	Distance to next pause	1	0.20
Accent group	Length	2	0.03 to 0.05
Phrase	Position of accent group	3	0.02 to 0.11

- Segment Identity: is coded activating the node correspondent to the segment. The segments are: 9 vowels, 4 semi-vowels, 5 nasal vowels, 6 plosive consonants (closure part), 6 plosive consonants (burst part), 3 nasal consonants, 5 liquid consonants and 6 fricative consonants. This is the major feature.
- Consonant at the end of word: codes in one node if the actual segment is [r, l*, S] (l* is a velar l, in the end of syllable) in end of word position. This fact should slightly increase the length of the segment. It is a minor feature.
- Previous segment (-1): the duration of the segment is statistically correlated with the type of the previous segment. The closure part of plosive consonants and fricative [S] are correlated with shorter segments in the next position. In case of closure segments this correlation is high because next segment is the burst part of plosive consonants that are very short. On the other way the burst part of plosive consonants [t, k, b, d, g], consonants [n, J, l, r, R, v, z] and pause, are correlated with longer segments in the next position. These 20 segments are coded by activation of the correspondent node. It is a major feature.
- Next segment (+1): the duration is statistically correlated with the type of the next segment. Segments [a, 6, u, 6~, o~, t, d] are correlated with shorter segments in previous position. Segments [l*, v] and closure part of plosive consonants [t, d] are correlated with longer segments in the previous position, and pause is even highly correlated with longer segments in the previous position. These 12 segments are coded by activation of correspondent node. It is a major feature.
- Next segment (+2): the duration is statistically correlated with the type of the second next segment. Segment [r] is correlated with shorter last segments but one. Burst part of stop consonants [t, d] and pause are correlated with longer last segments but one. These 12 segments are coded by activation of correspondent node.
- Next segment (+3): segment [u] and pause are correlated with longer antepenultimate segments. This feature is coded by activation of correspondent node.
- Type: the syllables considered are of the types: V, C, VC, CV, CC, VCC, CVC, CCV, and CCVC. Types C and CC result from elision of vowel. Syllables beginning with vowel are correlated with longer segments. Syllables with consonant clusters are correlated with shorter segments. This feature was coded in one node with values between 0 and 1 according to the correlation of the respective type of syllable with segments length.
- Vowel: codes the type of vowel in the syllable according to its average length. The considered 5 types are: long [a, E, e, O, o], medium [6, i], short [@, u], diphthong and nasal vowel. Long and nasal vowels are correlated with longer segments in the syllable. The others are slightly correlated with shorter segments in syllable. The feature was coded in one node with values between 0 and 1 according to the correlation of the respective type of vowel in the syllable with segments length.
- Type of previous syllable: Some types of syllables are slightly correlated with segments in next syllable. Syllables of types VC, CC and CVC, are slightly correlated with shorter segments in next syllable. The feature was coded in one node with values between 0 and 1 according to the correlation of the respective type of syllable with segments length (different of feature type). It is a minor feature.
- Vowel in previous syllable: long and nasal vowels as well as diphthongs are negatively correlated with length of segments in next syllable. Medium and short vowels are positively correlated. The feature was coded in one node with values between 0 and 1 according to the respective correlation with segments length (different of feature vowel). It is a minor feature.
- Vowel of next syllable: length of segments is positively correlated with short vowels in the next syllable. The other types of vowels are negatively correlated. The feature was coded in one node with values between 0 and 1 according to the respective correlation with segments

length (different of features vowel and vowel in previous syllable).

- Distance to tonic syllable: five categories were considered to characterize distance to tonic syllable in the accent group: tonic syllable, previous syllable, before previous, next syllable and after next. As is well known syllable tonicity is highly correlated with length of segments, but also next and after next are positively correlated with length of segments. In opposition, the other categories are negatively correlated. The feature was coded in one node with values between 0 and 1 according to the respective correlation.
- Position in group: is the segment count value inside the accent group, taken both from the beginning and end of group. The position relative to the end of group is highly and negatively correlated with segment's length. The position relative to the beginning is positively correlated with segment's length, as expected, by opposition. It is coded in two nodes.
- Position in Phrase: is the segment position count inside the phrase, both from beginning and end of phrase. Phrase is delimited by orthographic punctuation. The position relative to the end of phrase is highly and negatively correlated with segment's length. The position near to the beginning is slightly correlated with segment's length. It is coded in two nodes.
- Distance to next pause: is the distance in number of segments to the next pause. Is highly and negatively correlated with longer segments. As the segments are closer to a pause the longer are their durations. It is coded in one node. This is a major feature.
- Length: the number of phonemes and syllables, of the accent group. Is slightly correlated with longer segments. Coded in two nodes. This is a minor feature.
- Position of accent group: is the position of group inside the phrase (beginning, middle and end). Beginning position and specially end position are correlated with longer segments. In opposite middle position groups are slightly correlated with shorter segments. It is coded by activation of correspondent node. This is a major feature.

Other types of features were also considered but they didn't improve performance. These features were: type of next syllable; length of phrase; type of sentence; previous segments (-2 and -3).

4. Neural network

The model consists of a fully connected feed-forward neural network. The output is one neuron that codes the segment duration in values between 0 and 1. This codification is linear in correspondence to the range 0 and 250 msec. The 99 input nodes receive the set of coded features.

Training was done over the training set and using the test set for cross validation in order to avoid over-fitting. The test vector was used to stop training early if further training on the training set will hurt generalization capacity to the test set. The cost function used for training was the mean squared error between output and target values.

Similar levels of performance are achieved with different network architectures, varying in the number of intermediate layers, activating functions and training algorithms. ***Table 2*** reports the best performances for the best configurations. The activating functions are hyperbolic logarithmic (Log), hyperbolic tangent (Tan) and linear (Lin). The back-propagation training algorithms are Levenberg-Marquardt [13] and Resilient Back-propagation [14].

Table 2: Neural Network configurations and best performances.

Nodes in layers	Activating Functions	Training Algorithm	Value of r in test set
2-4-1	Log-Log-Lin	Lev.-Marq.	0.836
2-4-1	Tan-Log-Lin	Lev.-Marq.	0.838
4-2-1	Tan-Log-Lin	Lev.-Marq.	0.839
10-1	Tan-Lin	Resilient–Backpr.	0.835
10-1	Tan-Log	Resilient –Backpr.	0.837
2-4-1	Tan-Log-Lin	Resilient –Backpr.	0.836
6-1	Tan-Log	Resilient –Backpr.	0.836

5. Model evaluation

In order to objectively evaluate the prediction accuracy, between predicted values and actual duration values, standard deviation of the difference (σ) and linear correlation coefficient (r) were computed. ***Table 3*** presents the indicator's equations and respective scores, in test set. ***Fig. 1*** presents the target duration values (T) versos predicted durations (A), in test set.

Table 3: Prediction accuracy.

Equation	*score*
$\sigma = \sqrt{\frac{\sum_i d_i^2}{N}}, \quad d_i = e_i - \bar{e}, \quad e_i = x_i - y_i$	19.5 (msec)
$r_{X,Y} = \frac{V_{X,Y}}{\sigma_X . \sigma_Y}, \quad V_{X,Y} = \frac{\sum_i (x_i - \bar{x}).(y_i - \bar{y})}{N}$	0.839

5.1. Perceptual evaluation

Five paragraphs from the test set, with 30 words and 160 segments in average, were used for perceptual evaluation. Three stimuli of each paragraph were presented to 19 subjects for evaluation in a scale from 0 to 5, in a blind test. One stimulus was natural speech (natural); another was a time-warped natural speech with durations predicted by the model (model); and the last stimulus, also time-warped speech, with the average duration value for each type of segment (no-model). Time-warped modifications were done with the TD-PSOLA algorithm.

The average score, by subjects, classifies the model as very close to the natural one, and in four cases the model is even preferred by subjects. Also, the average score, by paragraphs, classify the model as very close to the natural and even preferred in one paragraph. ***Fig. 2*** presents the scores of perceptual evaluation for natural, model and non-model utterances. The natural utterances achieved a score of **4.30**, the

model utterances **4.12** and non-model utterances **3.53**. Analysis of variance of the natural, model and non-model scores, gives a significance higher than 99.9% (p<1e-12 for F=31.4). Therefore, there is sufficient evidence to reject the hypothesis that the levels are all the same. The model is 0.18 point (in 5) far from natural read speech, while non-model is 0.77 far from natural read speech and 0.59 far from the model.

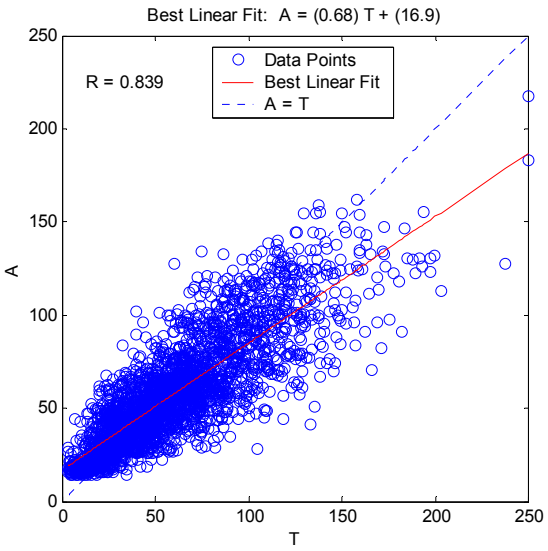

Fig. 1: Target (T) versos predicted (A) durations in test set.

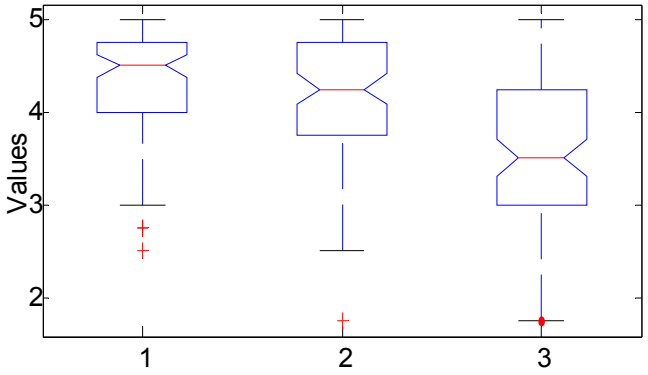

Fig. 2: Mean Opinion Score for: 1-natural; 2-model; 3-non-model.

6. Conclusion

A model based on a back-propagation trained neural network to predict segmental durations was presented. Phonological and contextual features were analysed considering the correlation with segmental durations as well as the influence on the improvement of the model performance.

Almost equally good performance (r between 0.835 and 0.839) was achieved with the different architectures / number of layers / activating function and training algorithms, presented in ***Table 2***.

The present model was objectively (***Table 3***) and subjectively (***Fig. 2***) evaluated. Objective evaluation gives a standard deviation of the difference between target and predicted duration of **19.5 msec**, and a linear correlation coefficient **r=0.839**. Subjective evaluation, with time-warped speech, puts the model at 4.12 points, where the natural speech achieves 4.30 and absence of model reaches 3.53 (in 5).

7. Acknowledgements

This work was developed under a PhD scholarship financed by the PRODEP program. We would like to thank FEUP and IPB for supporting this research.

8. References

[1] Klatt, D. H., "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence", Journal of the Acoustic Society of America, 59, 1209-1221, 1976.

[2] Zellner, B., "Caractérisation et prédiction du débit de parole en français – Une étude de cas", thèse présentée pour obtenir le grade de Docteur en Lettres, Université de Lausanne, 1998.

[3] Salgado, Xavier F., e Banga E. R., "Segmental Duration Modelling in a Text-to-Speech System for the Galician Language", in Eurospeech'99, Budapest.

[4] Van Santen, J. P. H., "Assignment of segmental duration in text-to-speech synthesis", in Computer Speech and Language, 8, 95-128, 1994.

[5] Campbell, W. N., "Predicting Segmental Durations for Accommodation within a Syllable-Level Timing Framework", Proceedings of Eurospeech 93, volume 2, pag. 1081-1084.

[6] Barbosa P., Bailly G., "Generation of pauses within the z-score model", in "Progress in Speech Synthesis", by Van Santen J. P. H., Sproat R. W., Olive J. P. and Hirschberg J. editors. Springer Verlag, New York 1997, pag. 365-381.

[7] Barbosa P., "A Model of Segment (and Pause) Duration Generation for Brazilian Portuguese Text-to-Speech Synthesis", in Eurospeech'97, Rhodes.

[8] Córdoba R., Vallejo J. A., Montero J. M., Gutierrez-Arriola J., López M. A., Pardo J. M., "Automatic Modelling of Duration in a Spanish Text-to-Speech System Using Neural Networks. Eurospeech'99.

[9] Hifny, Y., Rashwan, M., "Duration Modelling for Arabic Text to Speech Synthesis", Proceedings of ICSLP'2002, Denver.

[10] Chung, H., "Segment Duration in Spoken Korean", Proceedings of ICSLP'2002, Denver.

[11] Teixeira, J. P., Freitas, D., Braga, D., Barros, M. J., Latsch, V., "Phonetic Events from the Labelling the European Portuguese Database for Speech Synthesis, FEUP/IPB-DB", in Eurospeech'01, Aalborg.

[12] Gouveia, P. D., Teixeira, J. P. and Freitas, D., (2000). "Divisão Silábica Automática do Texto Escrito e Falado", in proceedings of V PROPOR, Processamento Computacional da Língua Portuguesa Escrita e Falada, p 65-74, Atibaia – S. Paulo, 2000.

[13] Hagan, M. T., Menhaj, M., "Training feed-forward networks with the Marquardt algorithm", IEEE Transactions on Neural Networks, vol. 5, n 6, pp 989-993, 1994.

[14] Riedmiller, M., and H. Braun, "A direct adaptive method for faster back-propagation learning: The RPROP algorithm", Proceedings of the IEEE International Conference on Neural Networks, 1993.