Evaluation of a Segmental Durations Model for TTS

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: In this paper we present a condensed description of a European Portuguese segmental duration's model for TTS purposes and concentrate on its evaluation. This model is based on artificial neural networks. The evaluation of the model quality was made by comparison with read speech. The standard deviation reached in test set is 19.5 ms and the linear correlation coefficient is 0.84. The model is perceptually evaluated with 4.12 against 4.30 for natural human read speech in a scale of 5.

1 Introduction

The presented segmental duration's model is part of a global prosodic model for a European Portuguese TTS system, which is under development in the authors' Institutions. It is based on artificial neural networks that process the input of linguistic information relative to the context of each phoneme, and outputs the predicted duration for each of its segments.

A series of durations' models can be found in the literature for other languages, mostly for American and British English. The most prominent ones will be mentioned in the following.

Campbell introduced the concept of Z-score [1] to distribute the duration estimated, by a neural network, for a syllable, considering that it would be the more stable unit for prediction of duration. He measured a linear correlation coefficient (r) between speakers taking the syllable as unit of **r**=0.92 and only **r**=0.76 for segments. He reported an **r**=0.93 for the syllables in his model. This concept isn't however generally accepted. Others authors, like Van Santen [2] use the phoneme as the segmental unit in order to predict durations in a Sum-of-Products model. The author reported **r**=0.93 in his database. Barbosa and Bailly [3] employed the concept of Inter-Perceptual Center Groups (IPCG) as the stable unit, and applied a neural network to predict its' duration and subsequently the Z-score to determine the duration of each phoneme inside the IPCG. This model can deal with speech rate. They reported standard deviation for French σ=43ms, and later, Barbosa reported a σ=36ms for Brazilian Portuguese [4]. Other relevant models are the Klatt model [5] based on a Sum-of-Products; the rule-based algorithm for French presented by Zellner [6] for two different speech rates, obtaining an **r**=0.85 and arguing that this value corresponds to the typical inter-speaker correlation; the look-up table based model for Galician [7] with a **rmse** (root-mean squared error) value of 19.6 ms in the training data; the neural net-

N.J. Mamede et al. (Eds.): PROPOR 2003, LNAI 2721, pp. 40–48, 2003.

work based models for Spanish [8] and for Arabic [9], achieving **r**=0.87; the CART-based model for Korean [10] with **r**=0.78. The final model we consider in this introduction was developed by Mixdorff as an integrated model for durations and F0 for German [11], having achieved **r**=0.80 for the durations.

Existing durations' models can be classified as statistical, mathematical or rule-based models. Besides the present one, examples of other statistical models can be [1,3,4,8–11], although [1] and [3] use the Z-score concept. These types of models became interesting with the availability of large databases. Examples of mathematical models can be [2] and [5]. Rule based models are [6] and [7].

The basic idea behind our approach comes from the fact that the duration of a segment depends, in a complex manner not only on a set of contextual features derived from both the speech signal and the underlying linguistic structure, but also on random causes. We therefore try to take into consideration most of the known relevant features of different kinds that are candidates to be influential on duration value and try to determine the complex dependency function in a robust efficient statistical manner that fits the selected database. This is known in advance not to contain all possible different combinations of features. Additionally the considered set of features is not exhaustive.

Inter-speaker and intra-speaker variability is well known and should be considered in the results analysis. In that way, what can be expected from such a model is an acceptable timing for the sequence of phonemes, and not exactly the same timing imposed by the speaker. This can only be evaluated perceptually.

The data that was used for the training and testing of the model was extracted from the database described in [12]. This database consists of tagged speech tracks of a set of texts extracted from newspapers that were read by a professional male radio broadcast speaker at the average speech rate of 12.2 phonemes/second. The dimension of the part of the data that was used in the present work, covers a total of 101 paragraphs containing a few hundreds phrases, essentially of declarative and interrogative types, with dimensions from one word to more than one hundred, consisting in a total of 18,700 segments of 21 minutes of speech. Phonemes were selected as the basic segment allowing the smallest granularity of the modeling.

Section 2 describes the model and Sect. 3 describes the evaluation.

2 Description of the Model

2.1 Duration Features

A large number of features were considered as candidates in the beginning of the work. One by one, they were studied and taken out in order to evaluate their relative importances. In selected cases, a set of a few features was considered and taken out jointly to check for consistency. The conclusion is, in general, that the result is different from considering the isolated features. This is because these features interact non-linearly in a significant manner. After several experiments, considering different sets of features and the correlation with segment's duration, one was finally established as giving the best optimization of the performance of the neural network approximation. The coding of features' values is also an important issue, so some features were coded

in varying ways, in order to find the best trend and solution. The final set of features of the model and their codifications is listed in order of decreasing importance:

a. Identity of segment – one of the 44 phoneme segments considered in the inventory of the data-base (Table 3);

b. Position relative to the tonic syllable in the so-called accent group – coded in 5 levels according to its correlation with durations;

c. Contextual segments identities – previous (-1) and next three (+1, +2, +3) segments – signaling some significant specific phones in referred position and silences (20 phones in position -1; 12 phones in position +1; 4 phones in position +2; 2 phones in position +3);

d. Type of vowel length in the syllable – coded in 5 levels according to its correlation with durations;

e. Length of the accent group – number of syllables and phonemes;

f. Relative position of the accent group in the sentence – first; other; last;

g. Suppression or non-suppression of last vowel;

h. Type of syllable – coded in 9 levels according to the correlation with durations;

i. Type of previous syllable – same as previous;

j. Type of vowel in previous syllable – same as d;

k. Type of vowel in next syllable – same as d;

Features *b*, *e* and *f* are linked with the so-called accent groups that we consider as random groups of words with more than two syllables, aggregating neighbor particles. These groups work like prosodic words having only one tonic syllable. Any how they aren't exactly prosodic words if one considers the multiple definitions in the literature.

In feature *d* we consider 5 types of vowels according to the average duration. These 5 types are: long {a, E, e, O, o}; medium {6, i}; short {u, @}; diphthong and nasal.

Feature *g* codes the eventual suppression of the last vowel in the word as can be found in [12], because this event usually lengthens the remaining consonant, like in the word 'sete' (read {sEt} – SAMPA code).

The type of syllables mentioned in features *h*, *i* and *j*, are: V, C, CC, (both resulting from suppressed vowel) VC, CV, VCC, CVC, CCV, CCVC.

During the above described process of selecting the features to be used, a qualitative measurement of the relative importance comes out. Three groups of features can be distinguishing according to relevance. The first is feature *a,* clearly the most important one. The second group in relevance is composed of features *b*, *c*, *d*, *e*, *f* and *g*. The third group, with features that alone are not very important, but together assume some relevance, is formed by features *h*, *i*, *j* and *k*.

2.2 Neural Network

The model consists in a feed-forward neural network, fully connected. The output is one neuron that codes the desired duration in values between 0 and 1. This codification is linear in correspondence to the range 0 and 250 ms. The input neurons receive the set of coded features. Similar levels of performance (r=0.833 to 0.839) are achieved with different network architectures (2-4-1, 4-2-1, 6-1, 10-1), activating functions (hyperbolic logarithmic, hyperbolic tangent and linear) and training algorithms (Levenberg-Marquardt [13] and Resilient Back-propagation [14]).

If the number of weights of the net is not fewer than the number of training situations, and the training is excessive, an over-fitting may occur. In order to avoid this problem, two sets of data were used. One set for training with 14,900 segments and another set for test with 3,000 segments. The test vectors were used to stop training early if further training on the training set will hurt generalization to the test set. The cost function used for training was the mean squared error between output and target values.

3 Model Evaluation

Two indicators were used to evaluate the performance of the model: the standard deviation (**σ**) (Eq. (1)) of the error (e) and the linear correlation coefficient (**r**) (Eq. (3)). Considering the error as the difference between target and predicted values of duration of segments, the standard deviation of the error (**σ**) was used, according to the following expressions:

$$\sigma = \sqrt{\frac{\sum_i x_i^2}{N}}, \quad x_i = e_i - \overline{e}, \quad \mathrm{e_i} = d_{i\_original} - d_{i\_predicted} \tag{1}$$

where, x_i is the difference between the error of each segment and the mean error. The error is given by the difference between predicted and original duration, for each segment. When the mean error is null, as happens in this case, **σ** is equal to the **rmse,** given by Eq. (2):

$$rmse = \sqrt{\frac{\sum_i e_i^2}{N}} \tag{2}$$

$$r_{A,B} = \frac{V_{A,B}}{\sigma_A . \sigma_B}, \quad V_{A,B} = \frac{\sum_i \left(a_i - \overline{a}\right).\left(b_i - \overline{b}\right)}{N} \tag{3}$$

The linear correlation coefficient (**r**) was the second indicator selected, and is given by Eq. (3), where $V_{A,B}$ is the variance between vectors A=[a_1 a_2 ... a_N] and B=[b_1 b_2 ... b_N]. A and B are predicted and target duration vectors.

The performance in the test and training sets, considering all types of phonemes, is given in Table 1.

Table 1. General best performance

set	*r*	*σ (ms)*
Test	**0.839**	**19.46**
Training	**0.834**	**19.85**

Table 2. Performance of the model (**r** and **σ**), and average duration for each type of segment

Vowel	r	σ (ms)	Aver. (ms)	Cons.	r	σ (ms)	Aver. (ms)
a	0.63	26.8	110	**p**	0.25	9.2	20
6	0.65	21.1	68	**!p**	0.39	17.8	64
E	0.62	23.1	97	**t**	0.76	12.8	29
e	0.71	28.2	95	**!t**	0.59	16.5	48
@	0.63	29.5	53	**k**	0.41	14.4	37
i	0.58	23	69	**!k**	0.27	16.6	59
O	0.61	25.8	106	**b**	0.79	11.2	17
o	0.63	26.4	97	**!b**	0.23	15.1	43
u	0.56	24.1	57	**d**	0.76	10.9	20
j	0.59	21.2	49	**!d**	0.2	17.2	41
w	0.68	20	44	**g**	0.73	8.9	20
j~	0.28	18.6	64	**!g**	0.26	12.8	44
w~	0.53	25.2	53	**m**	0.31	18.6	62
6~	0.69	24.9	75	**n**	0.33	17.9	54
e~	0.65	23.6	107	**J**	0.3	16.7	68
i~	0.74	27.9	109	**l**	0.23	19.4	52
o~	0.69	25.9	98	**l***	0.73	20.9	68
u~	0.79	26.9	86	**L**	0.68	15.3	56
Aver.	**0.63**	**23.8**		**r**	0.63	12.4	32
				R	0.38	19	73
				v	0.45	19.7	65
				f	0.56	22.7	93
				z	0.37	16.6	70
				s	0.59	24.7	103
				S	0.68	24.5	89
				Z	0.54	21.4	78
				Aver.	**0.50**	**16.3**	

Phonemes are presented in SAMPA code. l* is a velar l.

! Represents the occlusive part of stop consonants.

In the left part of Table 2, the vowels have, in a weighted average, **r**=0.63 and **σ**=24 ms. In the right part of the table, **r**=0.50 and **σ**=16 ms are presented as weighted average values for consonants. The average value for each phone is very well fitted by the neural network.

Figure 1 plots the original versus the predicted durations in the test set for one simulation with r=0.839. There are no major errors. These errors are quite low in short segments and naturally higher in longer ones.

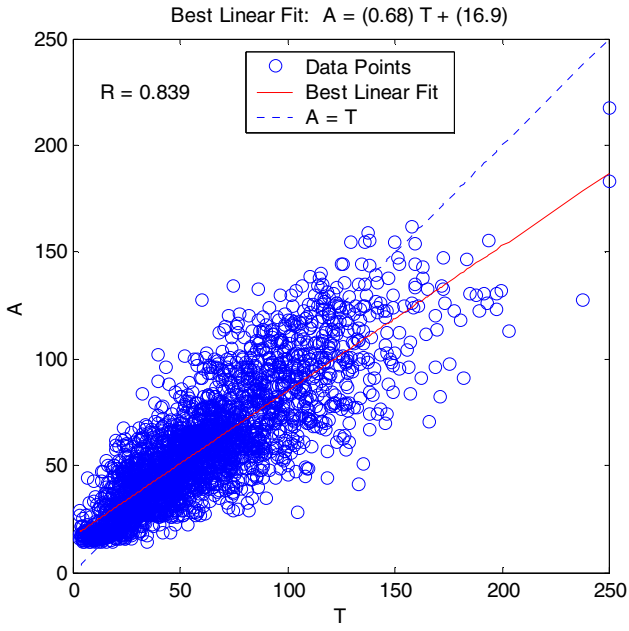

Fig. 1. Best linear fit for original and predicted duration for one simulation in the test set with r=0.839

3.1 Perceptual Evaluation

One last evaluation of the model presented in this paper is the perceptual test. Five paragraphs from the test set were used for this purpose. Three realizations of each paragraph were presented to 19 subjects (8 experts and 11 non-experts) for evaluation in a scale from 0 to 5. One realization was natural speech (original); another was a time-warped natural speech with durations predicted by the model (model); and the last realization, also time-warped speech with the average duration value for each phone (average). Time-warped modifications were done with a TD-PSOLA algorithm.

Table 3. Scores of model for the paragraphs presented to the listeners

Paragraph	N. of seg.	σ (ms)	r
1	36	19.0	0.97
2	164	18.9	0.89
3	177	22.6	0.94
4	209	19.0	0.91
5	204	19.8	0.94

The subjects didn't know which stimulus corresponds to each realization and they could hear as many times as they want. Table 3 presents, for each paragraph, its number of segments, plus **σ** and **r** values, for the predicted durations. In all the cases the scores between experts and non-experts subjects were very similar, so they were merged.

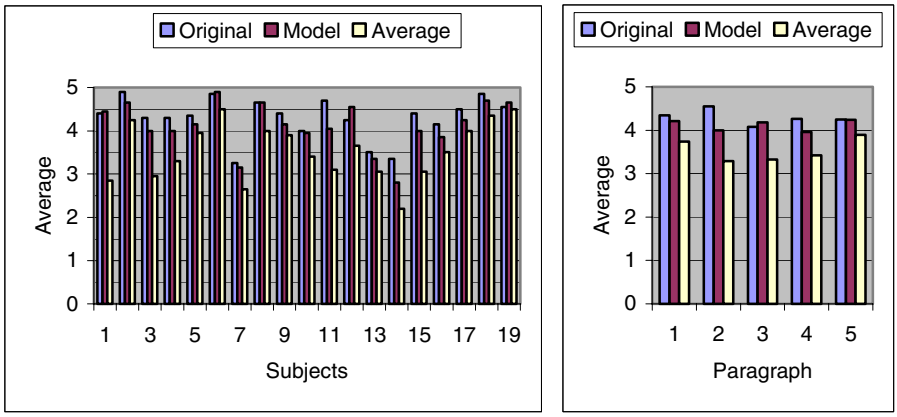

Fig. 2. Average score of perceptual test by subject (*left side*) and by paragraph (*right side*)

Figure 2 (left side) shows the average evaluation by subject for original, modified by the model and fixed average duration. For most of the listeners the model is very close to the original, and in four cases the model is even preferred. Figure 2 (right side) presents the average evaluation by paragraph. Again, the model is very close to the original, and in paragraph 3 is even preferred. Finally, Fig. 3 characterizes the subjects' opinions representing for each of the three sets of realizations the minimum, the ensemble of the lower quartile, median and upper quartile in the notched box, the maximum, the mean with thick lines and outliers. The original utterances achieved a mean score of **4.30**, the ones with durations imposed by the model achieved **4.12** and the ones with durations imposed with average value for each phoneme achieved **3.53**.

In one way, ANOVA gives $p<1e-12$, for an $F=31.4$, meaning a significance higher than 99.9%. The 0.18 points of distance to the original utterances mean that the sentences produced with predicted durations are quite close to natural.

4 Conclusion

A statistical model for segmental durations in European Portuguese was presented. This model is based on artificial neural networks that receive linguistically-oriented contextual information for the segment to process and predict its duration. Results are presented and discussed, showing a good objective performance of the model. This evaluation was done comparing model output durations with the target durations.

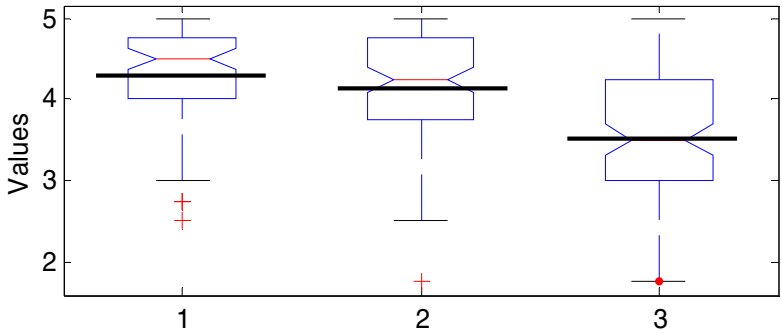

Fig. 3. Opinion score for: (*1-original*); (*2-model*); (*3-average)*

In one way, ANOVA gives p<1e-12, for an F=31.4, meaning a significance higher than 99.9%. The 0.18 points of distance to the original utterances mean that the sentences produced with predicted durations are quite close to natural.

The objective evaluation, considering all type of segments, shows that the model may be considered being at a good quality level, when compared with most of the published work in other languages.

Subjective evaluation was done by a perceptual test and shows that the model is quite close to the original (distance 0.18 in 5) and relatively far from the fixed realizations with averaged durations (0.59 in 5).

Finally, from the observation of several examples, the model predicts quite consistently the durations of final segments of words, where other authors report some troubles.

References

1. Campbell, W.N., "Predicting Segmental Durations for Accommodation within a Syllable-Level Timing Framework", Proceeding Eurospeech 93, volume 2, pag. 1081–1084.
2. Van Santen, J.P.H., "Assignment of segmental duration in text-to-speech synthesis", in Computer Speech and Language, 8, 95–128, 1994.
3. Barbosa P., Bailly G., "Generation of pauses within the z-score model", in "Progress in Speech Synthesis", by Van Santen J.P. et al, editors. Springer-Verlag, 1997.
4. Barbosa P., "A Model of Segment (and Pause) Duration Generation for Brazilian Portuguese Text-to-Speech Synthesis", in Eurospeech'97, Rodes.
5. Klatt, D.H., "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence", JASA, 59, 1209–1221, 1976.
6. Zellner, B., "Caractérisation et prédiction du débit de parole en français – Une étude de cas", PhD, U. de Lausanne, 1998.
7. Salgado, Xavier F., e Banga E.R., "Segmental Duration Modelling in a Text-to-Speech System for the Galician Language", in Eurospeech'99, Budapeste.
8. Córdoba, Vallejo, Montero, Gutierrez, López., Pardo, "Automatic Modelling of Duration in a Spanish Text-to-Speech System Using Neural Networks. Eurospeech'99.
9. Hifny, Y., Rashwan, M., "Duration Modeling for Arabic Text to Speech Synthesis", Proceedings of ICSLP'2002.
10. Chung, H., "Segment Duration in Spoken Korean", Proceedings of ICSLP'2002.

11. Mixdorff, H., “An Integrated Approach to Modeling German Prosody”, Thesis for Dr.-Ing. Habil., Technical University of Dresden, 2002.
12. Teixeira, J.P., Freitas, D., Braga, D., Barros, M.J., Latsch, V., “Phonetic Events from the Labeling the European Portuguese Database for Speech Synthesis, FEUP/IPB-DB”, in Eurospeech’01, Aalborg.
13. Hagan, M.T., Menhaj, M., “Training feedforward networks with the Marquardt algorithm”, IEEE Transactions on Neural Networks, vol. 5, n 6, 1994.
14. Riedmiller, M., and H. Braun, “A direct adaptive method for faster backpropagation learning: The RPROP algorithm”, Proceedings of the IEEE International Conference on Neural Networks, 1993.