EVALUATION OF A NEURAL NETWORK SEGMENTAL DURATION MODEL FOR PORTUGUESE

*João Paulo Teixeira and Diamantino Fre*itas

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 duration model, that, as far as the authors know, is the first published for European Portuguese, with objective and subjective evaluations. The model is aimed at TTS applications and is based on an ANN, trained with a resilient back-propagation algorithm. Using a substantial amount of training data and a carefully selected set of input factors, the standard deviation of the error of segmental duration estimations reaches 19 ms and the correlation coefficient goes above 0.9. Several models have been published for other languages with objective and subjective good performances. The methodology of construction of the model, the importance of the used factors and the neural network will be presented, together with the evaluation of the model, allowing a comparison with other models for other languages.

1. INTRODUCTION

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

Several types of duration models were studied in the preparation of this work, concerning the selection of text parameters and the model architecture itself. Relevant examples are: - models considering the Inter-Perceptual Center Groups (IPCG) produced for French [1] and Brazilian Portuguese [2], capable of generation of pauses and respective durations; - models considering several different Sum-of-Products for each type of segment [3] achieving a very good performance and requiring large amounts of input data; - the traditional Klatt model [4]; - a syllable-based model introducing the Z-score concept [5]; - a rule-based algorithm for the French language for two cadencies of speech [6]; a look-up table based model for Galician [7], a neural network based model for Spanish [8] and a Bayesian belief network model [9][10].

The common idea behind this work is to take all the relevant factors influencing the duration of a segment in a read utterance and use them as input to a tool that can automatically learn how each factor influences the duration and how the factors combine and are related to each other. Based on a sufficiently large amount of examples, this model will be able to efficiently predict durations of segments.

The data used for training, validation and testing the model were extracted from the FEUP-IPB database [11]. This database consists of several texts extracted from newspapers that were read by a professional radio broadcast speaker (average 12.2 phonemes/second). The dimension of the part of the database that was used in the present work is 7 texts, in a total of 200 phases of practically all types and sizes, consisting in a total of 18.700 segments of 21 minutes of speech.

2. DESCRIPTION OF THE MODEL

A large number of factors were considered as candidates in the beginning of the work. One by one, they were studied and taken in and out in order to evaluate their relative importance for the results. Some times a group of more than one factor was considered and taken out jointly to check for consistency. The conclusion was that the result is many times different from considering the factors isolated. This is because these factors interact significantly. After several experiments, considering different combinations of factors, the set of factors was finally established. Some factors were coded in varying ways, in order to find the best solution.

The final set of factors of the model of duration of the present segment and their codification is as follow:

a. Position relative to the tonic syllable in the duration group – coded in one of 5 positions;
b. Type of syllable – one of nine types, according to the sequences of consonants and vowels;
c. Type of previous syllable – same as in b;
d. Type of vowel in the syllable according to length – one of five;
e. Type of vowel in last syllable – same as in d;
f. Type of vowel in next syllable – same as in d;
g. Length of the duration group – number of syllables and phonemes;

h. Relative position of the duration group in the sentence – first; middle; last;
i. Suppression or not of last vowel;
j. Phonetic identity of segment – one of the 44 different segments considered in the inventory of the database (excluding pause and aspiration);
k. Context segments identities– previous (-1) and next three (+1, +2, +3) segments – each coded as in j.

During the process of selection of the factor to be used, a qualitative measurement of its relative importance comes out. Factor j is clearly the most important. Then come factors a, d, g, h, i and k as next important, and factors b, c, e and f are less important.

In the last list of factors any of these does not alone improve significantly the performance of the model. Anyhow, when considered jointly, these less important factors improve the model performance.

The ANN is a feed-forward fully connected network, with one 10-neurons hidden layer activated by log-sigmoid transfer functions (figure 1). The output is one neuron activated by a tan-sigmoid transfer function. This neuron codes the predicted duration in values between 0 and 1. This codification is linear in the range 0 and 250 ms. The input neurons code the set of factors.

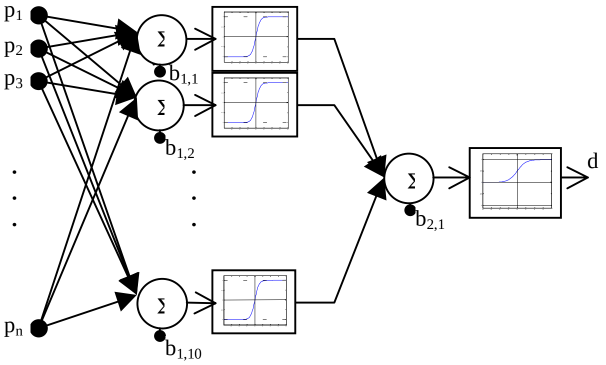

Figure 1: Architecture of the neural network.

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 problem may occur. The error obtained over the training set is driven to a very small value, but when new data is presented to the network the error becomes larger. The network has “memorized” the training examples, but it has not “learned” to generalize to new situations. In order to avoid this problem, an additional set of data besides the training (13700 segments) and the test (2000 segments) sets was used for validation (3000 segments). The validation vectors are used to stop the training phase as early as needed when further training will hurt generalization to the validation set. Test set performance can be used to measure how well the network generalizes beyond training and validation sets.

The performance function used for training was the mean squared error taken between the output and the target values.

Multi-layer networks typically use sigmoid functions in the hidden layers. These functions compress an infinite input range into a finite output range. Sigmoid functions are characterized by the fact that their slope must approach zero as the input gets larger. This can cause a problem when using steepest descending algorithms to train, since the gradient can have a very small magnitude and cause small changes in the weights and bias. The purpose of the used resilient back-propagation algorithm is to eliminate these harmful effects of the magnitude of the derivatives. Only the sign of the derivative and not tme magnitude is used to determine the direction of the weights update [12]. With this algorithm the training session takes about 30 seconds in one 850 MHz-clocked PC.

3. MODEL EVALUATION

The evaluation was done with the validation and test sets, not used for training.

Three indicators were used to evaluate the performance of the model. The standard deviation of the error (**σ**), and the mean of the absolute error (**δ**) were used, according to the following expressions:

$$\sigma = \sqrt{\frac{\sum_i x_i^2}{N}}\,,\ x_i = e_i - \bar{e}\,,\ e_i = d_{i\_original} - d_{i\_predicted} \qquad (1)$$

where x_i is the difference between the values of the error value of each segment and the mean error. The error being given by the difference between predicted and original durations, for each segment.

When the mean error value is null, as in this case, **σ** is equal to the **rmse** (root-mean-square error); **rmse** and **δ** are given by:

$$rmse = \sqrt{\frac{\sum_i e_i^2}{N}}\,, \qquad \delta = \frac{\sum_i |e_i|}{N} \qquad (2)$$

The linear correlation coefficient (**r**) was the third indicator selected. It is given, together with $V_{A,B}$ for vectors A and B by:

$$r_{A,B} = \frac{V_{A,B}}{\sigma_A . \sigma_B}\,,\ \text{with}\ V_{A,B} = \frac{\sum_i (a_i - \bar{a}).(b_i - \bar{b})}{N} \qquad (3)$$

where $V_{A,B}$ is the variance between vectors *A=[a1 a2 … aN]* and *B=[b1 b2 … bN].*

The general performance, considering all types of phonemes, is **σ** =20 ms., **δ** =15 ms. and **r**=0.82.

Each model has its own characteristics and peculiarities. Some of these models mentioned in the introduction also include estimation of pauses, others are applicable for several speech rates, etc.. Any type of comparison of the prediction models’ performances can not be seen as a definitive comparison due to differences in the material under analysis, like the numbers of segments considered, the differences in languages, and possibly the most relevant are the databases used, that usually are different.

Table1: Performance of the present model (**r** and σ)

Vowel	r	σ (ms)
a	0.59	27.9
6	0.66	20.9
E	0.60	24.3
e	0.72	27.8
@	0.49	33.0
i	0.56	23.4
O	0.67	24.3
o	0.59	27.6
u	0.56	24.3
j	0.61	20.8
w	0.70	19.3
j~	0.39	17.3
w~	0.72	20.6
6~	0.72	24.2
e~	0.48	27.7
i~	0.74	27.6
o~	0.62	28.2
u~	0.70	30.6
Aver.	**0.62**	**25.0**

Cons.	r	σ (ms)
p	0.38	8.3
!p	0.44	17.5
t	0.73	13.5
!t	0.58	16.6
k	0.52	13.6
!k	0.36	16.4
b	0.86	9.3
!b	0.31	15.1
d	0.79	10.4
!d	0.37	16.3
g	0.72	9.0
!g	0.40	12.3
m	0.33	19.0
n	0.40	17.9
J	0.40	16.3
l	0.33	19.2
l*	0.61	24.8
L	0.46	18.6
r	0.60	12.8
R	0.27	20.4
v	0.49	19.5
f	0.60	22.3
z	0.34	17.8
s	0.58	25.3
S	0.66	25.0
Z	0.55	21.2
Aver.	**0.50**	**16.9**

Phonemes are presented in SAMPA code.

l* is a velar l.

! Represents the occlusive part of stop consonants.

Another important issue is that significantly different values for each indicator can be obtained using different (in content or length) sets with the same prediction model. This means that the results are strongly dependent of the used database. Also a better **r** doesn't mean always a better σ, as can be observed in the segments l* and L of table 1, not enabling a consistent ranking of models.

The Sum of Products model in [3], reports a value of **r**=0.88 when tested with a different database, with all types of segments. Brigitte Zellner in [6] reports an **r** value not less than 0.7 for all cases, for two different speech rates in her proposed algorithm for French. Barbosa and Bailly, in [1], report a value of σ=43 ms, for normal speech rate, from their IPGC model for French. Later, Barbosa reported a value of σ=36 ms for the adaptation of the IPGC model for Brazilian Portuguese. In [7], Salgado and Banga reported a value of σ=19.6 ms in the training set for a Galician Language Model. At the same time Cordoba *et al*, in [8], reported a value of δ=14.3 ms. as the best score for the neural network model for Spanish. For our model, in the left part of Table1, the vowels present an average **r**=0.62 and σ=25 ms. The right part of the table, presents **r**=0.5 and σ=16.9 ms. as the average values for consonants.

Very interesting scores were presented by Goubanova et al in [9] and [10]. In [9] the scores (**rmse**=5 ms and average **r** of 0.94) for the proposed Bayesian belief network (BN) are compared with other models, for vowels. The scores in [10] for BN were for a different database (**rmse**=3 ms and **r**=0.56 for vowels and **rmse**=2 ms and **r**=0.38 for consonants). The **rmse** is remarkably low being the outcome from a comparison with labeled segments in the database. What is the meaning for **rmse** values with higher precision than the precision in labeling of segments? Van Santen, for example, mentions an average error in the database used in his work [3] of 3 ms, but also refers a significant variability (21.4 ms. of standard deviation) when the same speaker repeats the same word in the same context. Why doesn't the value of **r** doesn't increase, following the high precision of **rmse** in [10]? Which is the most important one?

Another important issue is the impact on the scores obtained for **rmse,** caused by the relative dimension of the database used for training. A significant decrease in **rmse** should be expected from an increase in this dimension. Rules for **r** are much more difficult to find.

Table 2: Objective scores of the material used.

Paragraph	N. of segm.	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 subjective evaluation of the model presented in this paper was done through a perceptual test. Five paragraphs were used. Three realizations of each paragraph were randomly presented to 17 listeners, individually, for evaluation in a 0-20 scale: the original; another with the segmental durations modified according to the prediction model; and a third with the average durations. The modifications were done with a TD-PSOLA algorithm. The listener is not which case corresponds to each realization and subjects could hear as many times as they want. Table 2 presents, for each paragraph, the number of contained segments and σ and **r** values, for the predicted durations.

Fig. 2 shows the average evaluation by listeners. For most of them the model is very close to the original, and in three cases the model is even preferred.

Fig. 3 presents the average evaluation by paragraph. The model keeps very close to the original, and in paragraph 3 the model duration sequence is even preferred.

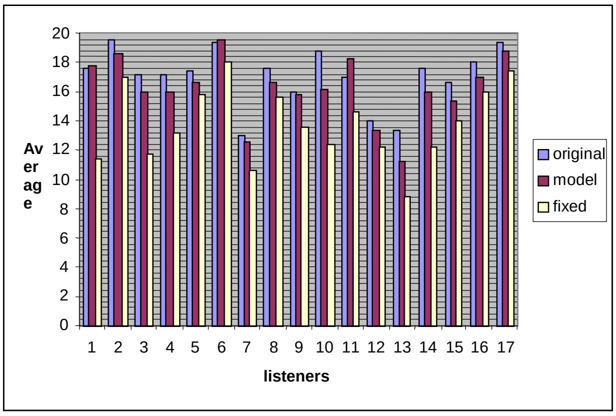

Figure 2: Average evaluation by listeners.

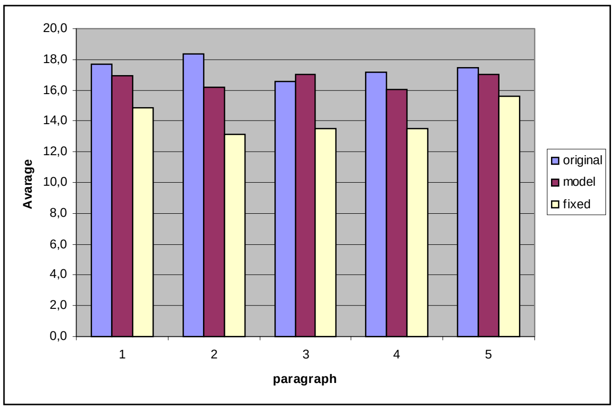

Figure 3: Average evaluation by paragraph.

Finally, the average evaluation, considering all data, is 17.5, 16.6 and 14.1 for original, model and fixed realizations, respectively.

4. CONCLUSION

The evaluation of the model was done comparing output durations with the real durations of speaker's segments. Many times slightly different durations do not mean wrong durations, because the specific speaker production is not the only one admissible and is allowed to vary substantially. What is important is that synthetic speech will sound natural for the model estimated durations. This can only be validated with perceptual tests.

The presented scores considering all type of segments, are at similar quality level as the ones presented for other models and languages. Table 2 presents better scores for **r**, when measured in paragraphs, compared to the one presented for all test data, confirming that the scores are very dependent of database.

Perceptual tests shows that the model is 0.9 in 20 close to the original and 2.4 in 20 far from the fixed realizations.

The model also predicts very consistently the durations of final segments which tend to be quite lengthened in European Portuguese.

A more consistent evaluation framework is needed for the performance assessment of duration and prosody models in general.

5. REFERENCES

[1] 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.

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

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

[4] Klatt, D. H., “Linguistic uses of segmental duration in English: Acoustic and perceptual evidence”, J.A.S.A., 59, 1209-1221, 1976.

[5] Campbell, W. N. and Isard, S. D. “Segment durations in a syllable frame”, Journal of Phonetics, 19 :37-47, 1991.

[6] Zellner, B., “Caractérisation et prédiction du débit de parole en français – Une étude de cas”, Phd Thesis, Université de Lausanne, 1998.

[7] Salgado, Xavier F., Banga E. R., “Segmental Duration Modelling in a Text-to-Speech System for the Galician Language”, in Eurospeech’99, Budapest.

[8] Córdoba, Vallejo, Montero, Gutierrez, López., Pardo, “Automatic Modelling of Duration in a Spanish Text-to-Speech System Using Neural Networks. Eurospeech’99.

[9] Goubanova, O., Taylor, P. “Using Bayesian Belief Networks for model duration in text-to-speech systems”, ICSLP2000, Beijing.

[10] Goubanova, O., “Predicting segmental duration using Bayesian belief networks”, Proc. of 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Scotland, 2001.

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

[12] 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.