# Use of Phoneme Dedicated Artificial Neural Networks to Predict Segmental Durations

*João Paulo Teixeira and Diamantino Freitas**

ESTiG – Instituto Politécnico de Bragança, Portugal joaopt@ipb.pt
* Faculdade de Engenharia da Universidade do Porto, Portugal dfreitas@fe.up.pt

## Abstract

The results of two alternative models to predict segmental durations in speech synthesis, both based on Artificial Neural Networks (ANNs) are discussed. The ANN model consists in just one ANN trained to predict the segmental durations for all phonemes. The phoneme dedicated ANN model consists in a set of ANNs, each one dedicated to predict the segmental duration of a specific phoneme. Both models are compared with the same input information extracted from one European Portuguese database. Objective and subjective measurements of performance of both approaches are compared. A slight preference was denoted for the phoneme dedicated ANN model.

## 1. Introduction

The issue of predicting durations of speech segments, whatever the target segment is, was already object of several important publications. Different authors proposed different models, for different languages, and even the type of segment unit is not consensual. The models can be grouped in rule-based models, mathematical models and statistical models. Next lines summarize very shortly the most relevant ones of each type.

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] and the rule-based algorithm for French [2], presented by Zellner for different speech rates. The Klatt model is possibly the best-known model developed already in 1976 based on eq. (1). $D_p$ is the predicted duration for segment p, $D_{min,\ p}$ is the minimum duration for segment p, $D_{in}$ is the output from preceding rules. For the first segment of the sequence, $D_{in}$ equals the inherent duration of segment p. Finally, k is a parameter reflecting the contribution to duration of a set of features expressed by the eq.(2), where $k_{fi}$ is the value of feature i. k has a value between 0 and 1 for shortening rules and superior to one for lengthening rules.

$$D_p = D_{\min,p} + k \times \left( D_{in} - D_{\min,p} \right) \tag{1}$$

$$k = \prod_{i=1}^{N} k_{fi} \tag{2}$$

The rule-based algorithm for French [2], proceeds in two phases. In the first phase predicts the syllable duration based on the type of word the syllable belongs to (lexical VS grammatical), the position of the syllable in the word, group, sentence, etc. In the second phase the distribution of that duration to the component segments of each syllable is made. The logic of that distribution varies with different types of syllabic structure. For the two stages jointly, the results were never inferior to 0.7 for slow and fast speech rates.

Mathematical models usually appear as a Sum-of-Products, where the features are statistically weighted and summed to produce the segmental duration. The previous models already used some type of sum of weighted features. One example is the Jan van Santen model [3] proposed in 1994 that is composed by a tree that can handle the linguistic heterogeneity of the segments, allowing a separate treatment for each category and its own sum-of-products model at the end of the tree, generically given by eq. (3). Each model differs from the remaining because the features affecting each category also differ. The reported results refer to the correlation coefficient considering all types of segments of 0.93 for the parameter determination database and 0.88 for other databases, which is excellent.

$$Dur(p) = \sum_{i \in T} \prod_{j \in I_i} S_{i,j}\left( p_j \right) \tag{3}$$

Statistical duration models become more and more used with the availability of large phonetically labelled databases. 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 [4] introduced the concept of Z-score to distribute the duration predicted by a neural network, for a syllable, among its segments in the logarithmic domain. He argued in favour that the syllable is the more stable unit in the logarithmic domain. The Z-score distribution develops the elasticity concept, according to which the duration of syllable segments is obtained through the application of a single z score, in the logarithmic domain, in the eq. (4), so that the sum of segmental durations equals the syllable duration, eq. (5).

$$Dur_i = \exp\left( \mu_i + z\sigma_i \right) \tag{4}$$

$$\sum_i Dur_i = syllable\ duration \tag{5}$$

Where $\mu_i$ , and $\sigma_i$ are respectively the mean and standard deviation of the transformed durations or logarithmic duration for segment i. The author presented a correlation coefficient of 0.93 for syllable duration prediction.

Barbosa and Bailly also presented a two steps model for French [5] and Brazilian Portuguese [6]. In the first step, using a neural network, they estimate the duration of the Inter-Perceptual Centre Groups (IPGC), arguing that it 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 [7] and Arabic [8]. Example of a CART-based

model applied for Korean can be found in [9].

Since phonological syllables, in European Portuguese, frequently derive from the collapse of weaker vowels, syllables cannot be regarded as rhythmic units, as opposed to other languages. Therefore, the phoneme was used as the segmental unit in the present models.

Next section describes the models based on ANNs to be compared. In section 4 the results obtained from measurements and from perceptual test of both models are compared. In Section 5 a discussion and some considerations about the models are presented.

## 2. The compared models

The models to be compared are very similar except that the so-called ANN model consists of just one ANN, and the Phoneme Dedicated ANN (PDANN model) consists of 44 ANNs (44 is the number of different phonemes in the database). All 44 ANNs, basically have the same architecture and input features of the ANN of the first model. Each one was trained only with the data relative to the respective phoneme. The paper attempts to measure precisely the effect of this difference.

### 2.1. ANN Model

Several experiments took place in order to optimize the performance of the model concerning the architecture (number of layers, number of nodes in each layer) activating functions, number of input features, and their codification, as described in [10]. The selected feed-forward ANN has 99 input nodes (detailed below), two hidden layers with 4 nodes with the hyperbolic tangent activating function in the first hidden layer and 2 nodes with the hyperbolic logarithmic activating function in the second one, and one output node with a linear activating function. The output codes the predicted duration.

Table 1 presents the 99 input nodes corresponding to the input features and their correlation (r) with segment's durations. Detailed specifications of each feature can be found in [10] and [11].

*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 |

Training was performed with a Levenberg-Marquardt [12] back-propagation algorithm 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. Some pre-processing is performed in order to normalise the input and output data.

### 2.2. PDANN model

As already mentioned the PDANN consists of 44 ANN with a similar architecture to the previous one. Each ANN is dedicated to predict durations of each type of phoneme. Therefore, the segment identity will be used to select the ANN to predict the segment duration and is not needed in the input nodes saving 44 nodes in the input. Everything but the input layer is identical to previous ANN. The training of each ANN was performed using only the data correspondent to the respective phoneme.

One of the advantages of the PDANN model is the fact that a given phoneme segment duration cannot be "disturbed" in any direction by the influence of the other segments' features. However, that may also become a disadvantage, since the parameter information for a given segment is not applied to others. This becomes more relevant when the number of stimuli for each segment is clearly not enough to train a sizeable network.

## 3. Results

Results presented in this section were measured (objectively and subjectively) under the test set of the FEUP/IPB European Portuguese Database of speech [13]. The database was divided in the training set with ≈15000 segments and the test set with ≈3000 segments. The relative frequencies of the phonemes are identical in both sets.

Objective results were also objectively measured by mean of a mathematic formula in opposition to the subjective ones that were measured by means of a subjective evaluation made by a group of subjects – a perceptual evaluation.

### 3.1. Objective results

The standard deviation and the linear correlation coefficient between original (measured) and predicted segment durations were determined according to the expression in Table 2. The σ and r values were improved with the PDANN model.

*Table 2:* Prediction accuracy in test set

| ***Equation*** | ***ANN*** | ***PDANN*** |
|---|---|---|
| $\sigma = \sqrt{\frac{\sum_i d_i^2}{N}}, \quad d_i = e_i - \bar{e}, \quad e_i = x_i - y_i$ | **19.5** (ms) | **18.2** (ms) |
| $r_{X,Y} = \frac{V_{X,Y}}{\sigma_X . \sigma_Y}, \quad V_{X,Y} = \frac{\sum_i (x_i - \bar{x}).(y_i - \bar{y})}{N}$ | **0.839** | **0.861** |

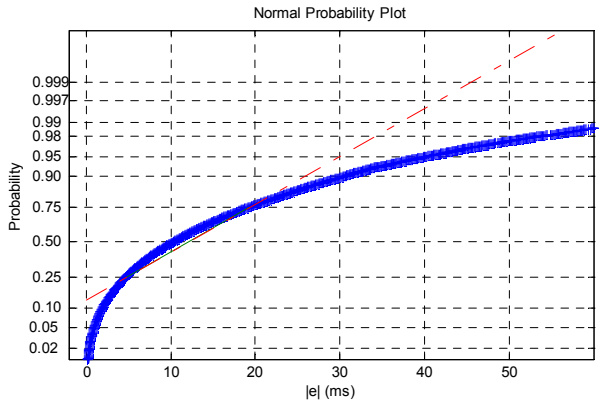


*Figure 1*: Normal probability distribution and absolute error curve for every segment in both sets with ANN model.

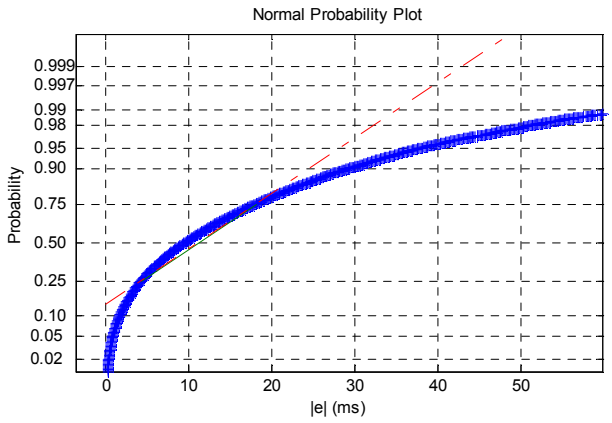


*Figure 2*: Normal probability distribution and absolute error curve for every segment in both sets with PDANN model.

In order to analyze the error distribution |e| in both models and considering that the measured standard deviation and linear correlation coefficient were similar in the training and test sets, both sets (training and test) were analyzed and used to produce representations of the probability density functions of the absolute error, depicted in Fig. 1 and Fig. 2.

It can be seen, once again, the improvement obtained with the PDANN model. This model predicts 75% of the segments with an error inferior to 18 ms, against the 20 ms for the ANN model, 90 % of segments with an error inferior to 29 ms and 95% of segments with an error inferior to 37 ms, against the 40 ms for the ANN model.

### 3.2. Subjective results

A change of the standard category-judgment test [14] was introduced, consisting of not giving reference of excellence or unsatisfactory category. Instead, two original stimuli (without modifications) and one stimulus with segments with the average duration by segment (henceforth "No model") were used. The two original stimuli were used to evaluate the consistency of answers by the listeners, since the stimuli were exactly the same. The "No model" stimuli were produced by changing the original duration of each segment in the stimulus to the average duration in the database for each identity of segment. These stimuli are called "No model" because durations can be easily taken from a very simple table with the 44 different types of segments and respective average durations. The "No model" stimuli are not comparable with an unsatisfactory reference, because, in fact, they produce a fair timing for several sentences with no emphatic prosody.

The test material was divided into paragraphs. A total of 5 stimuli per paragraph were presented in random order to 20 listeners in a blind test, without knowing whether they were listening the original or the manipulated version. Listeners were informed about the type of modifications introduced in the original sound and asked to concentrate in timing acceptability. They could hear the stimuli as many times as they wanted and were asked to classify each stimulus in a scale from 1 to 5 (1- **Unsatisfactory**, 2- **Poor**, 3- **Fair**, 4- **Good**, 5- **Excellent**).

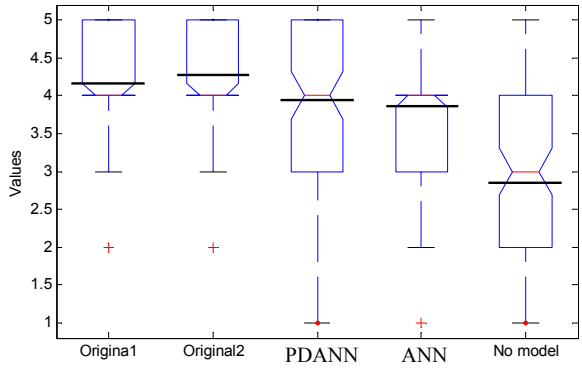


*Figure 3*: Analyses of opinion scores.

Fig. 3 illustrates the subjects' opinion analysis for each type of stimulus over 100 opinions. Mean Opinion Scores (MOS) are represented by a black thick line. The (blue) boxes

represent the lower and upper quartile. Thin (red) lines represent the median score. Minimum and maximum values are presented with the (black) thin lines. Plus signals (red) represent the outliers. Picture evidences the equality in original1 and original2. Although "No model" presents a quite good score, it is still far from the ANN model and even more far from the PDANN model. The PDANN model is close to the original and a little bit better than the ANN model.

## 4. Discussion

The reduction of 1.3 ms of the σ value of the distance between original and predicted durations using PDANN and the increase in the correlation from 0.839 to 0.861 denotes the evident improvement of the results with the dedicated ANNs.

The perceptual test confirmed a slight preference of the PDANN model over the ANN model. For some subjects the PDANN model was even preferred against the original stimuli. In some tested paragraphs the PDANN model was also preferred instead of original stimuli. In general the PDANN model is very close to original, with an average (original1 and original2) MOS distance of 0.27. This result evidence the improved results achieved by the usage of phoneme dedicated ANN.

The architecture of an ANN should be carefully designed in order to guarantee that the available number of training vectors is several times larger (at least more than 5) than the number of weights of the ANN. Otherwise the training set will not be sufficient to optimize all ANN weights. In the ANN model the number of weights is 413, and the number of training vectors is about 15.000, about 36 times larger. But in the PDANN model the number of weights of each ANN is 237, and there is not enough training vectors to guarantee a good optimization for all ANN weights.

Therefore, it is needed to increase the number of training vectors for some phonemes, and reduce the number of input features. The reduction of input features already proved to make this solution more useful when a reduction of the input nodes occurred during the development of the model, making the PDANN model more competitive.

A secondary effect of the low number of training vectors is the lower elasticity in predicted durations with PDANN for several phonemes.

## 5. Conclusions

Two ANN based models for prediction of segmental durations was compared. The first model is a classic ANN model with one ANN optimized concerning its features and architecture for European Portuguese. The second model is a phoneme dedicated ANN model, based on the features and architecture of the first one but consisting of one ANN for each type of phoneme.

Both models achieved a GOOD result in a perceptual test, and their correlation coefficients between original and predicted durations are at the very best of the state-of-the-art level.

The objective and also the subjective measurements denote a slight but clear preference towards the PDANN, in spite of the requirements to increase the number of input training vectors for some phonemes and to reduce the number of input nodes.

## 6. Future developments

The problem identified in the discussion section and mentioned as a requirement in the conclusions is one of the planned future developments. A second possibility to improve the PDANN that will be implemented is the optimization of the architecture and set of features, individually for each ANN.

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

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

[5] 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.

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

[7] 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.

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

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

[10] Teixeira, J. P. "A Prosody Model to TTS Systems", Doctoral thesis to fulfil the degree of Doctor in Electrotechnical and Computer Engineering in Faculdade de Engenharia da Universidade do Porto, 2004. http://www.ipb.pt/~joaopt/publicacoes/publicacoes.html

[11] Teixeira, J. P. and Freitas, D. Segmental Durations Predicted With a Neural Network", Proceedings of Eurospeech'03, pag.169-172.

[12] 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.

[13] 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.

[14] Standard Publication No. 297, IEEE, (1969). IEEE Recommended Pratice for Speech Quality Measurements. *IEEE Transations on Audio and Electroacoustics*. Vol. AU-17, no.3. 1969.