Phoneme Dedicated ANN Improves Segmental Duration Model

*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 Phoneme Dedicated Artificial Neural Network (PDANN) segmental duration model consists of a set of ANNs trained specifically for each phoneme segment in order to avoid miscellaneous influence of different types of phoneme segments. Therefore, each ANN is dedicated to predict the duration of a specific phoneme segment. Objective and subjective measurements of the performance of the PDANN model were compared with those of a typical ANN model using the same input features and database. The results indicate a slight, but clear, perceptually perceived preference towards the PDANN.

1. Introduction

The issue of predicting the duration of the speech segments, whatever the target segment is, has been the object of several publications. Various authors proposed different models, for different languages [1, 2, 3, 4, 5]. The segmental unit is not consensual among the authors. The elementary segment unit is the phoneme, but some models predict the duration for larger units like syllables or Inter-Perceptual Centre Groups. The latter consider these units more stable than the phoneme for the specific idiom, and in a second step divide the duration between the phonemes using the z-score concept (mentioned below).

The models can be grouped in rule-based models, mathematical models and statistical models.

The rule-based models require a straightforward knowledge of the effects of each feature in the duration of the segments.

Mathematical models usually appear as a Sum-of-Products, where the features are statistically weighted and summed to produce the segmental duration. Different Sum-of-Products for different type of segments can improve the final level, though they require more sophisticated statistical analyses to find the correct features for each type of segment and the respective weight. Well-constructed models can achieve excellent performance.

The statistical models require large labeled databases but they can easily provide very good results, using the right set of features and the appropriate methodology. Usually, statistical models use the Classification and Regression Trees (CART) and Artificial Neural Networks (ANN).

The remaining of this section provides a brief summary of the most relevant types of models.

1.1. Rule- based models

The Klatt model [1] is probably the best-known rule-based model**,** developed in 1976 and is 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 of the preceding segment. 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 the duration of a set of features expressed by 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)greater than 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. 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 of the component segments of each syllable is made according to its structure.

1.2. Mathematical models

The model [3] proposed in 1994 is composed of 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 sum-of-product differs from the others because the features affecting each category also differ. The reported results refer to the linear correlation coefficient, *r,* equal to 0.88.

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

1.3. Statistical models

Statistical duration models have become widely used with the availability of large phonetically labelled databases. ANN and CART are the most often used tools, and are applied in different ways for different languages and using different types of segments. Recent works have shown better results using ANN compared to those that use CART.

The Z-score concept was introduced [4] in order to distribute the duration predicted with an ANN for a syllable, among its segments in the logarithmic domain. The author argued that the syllable was the more stable unit in the logarithmic domain for the English language. The Z-score distribution develops the elasticity concept, where the duration of syllable segments is obtained through the application of a single z score applied to all segments of the syllable, in the logarithmic domain, Eq. (4), so that the sum of segmental durations equals the syllable duration, Eq. (5). Therefore, the duration of all segments of the syllable will be affected by the same portion of standard deviation lengthening (or shortening) in the logarithmic domain.

$$Dur_i = \exp(\mu_i + z\sigma_i) \quad (4)$$

$$\sum_i Dur_i = syllable\ duration \quad (5)$$

Where μ_i , and σ_i are respectively the mean and standard deviation of the transformed durations for segment i.

A two steps model for French [5] and later for Brazilian Portuguese [6] was developed by P. Barbosa. In the first step, the author estimates the duration of the Inter-Perceptual Centre Groups (IPCG), arguing that it is the more stable unit, using an ANN. In the second step**,** the duration of the IPCG is distributed among its segments, using the Z-score concept.

Next section describes the database and the hierarchic organization of the phonologic structures used in this study. Section 3 describes the PDANN model, their input features and their relevance. Section 4 presents a comparison between the results of the proposed model with those of a typical ANN model. The discussions and conclusions are made in sections 5 and 6, respectively.

2. Database

The data used for training and testing were extracted from the European Portuguese FEUP-IPB database [7]. This database consists of several labeled speech tracks of read speech. The speech waveforms were manually labelled in three levels: the phonetic, word and phrase levels. Seven tracks of the database were used, in a total of 101 utterances relative to paragraphs of high variety of lengths, from 1 to 100 words. Mainly declarative and interrogative types of sentences were selected, in a total of 21 minutes of speech and about 18000 phoneme segments.

The corpus was divided into three sets: a training set consisting of 12000 phonemes; a validation set with 3000 phonemes; and the test set with the remaining 3000 phonemes. The relative frequency of the phonemes occurrence is identical in the three sets.

The occlusive consonants were divided in the occlusion and the burst parts.

A total number of 44 phonemes or phoneme segments were used and consist of: 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.

The hierarchical structure was organized according to the increasing order of levels: phoneme, syllable, word, accent group, phrase and paragraph.

The input features were extracted by processing data from the corpus labels and grouping words in the so-called accent groups. These groups act like prosodic words aggregating neighbouring particles, and were created according to the 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 the justice situation'): "*uma forte / reserva / em relação / à situação / da justiça"*.

3. PDANN model

Since phonological syllables, in European Portuguese, frequently derive from the collapse of weaker vowels, syllables cannot be regarded as rhythmic units, as opposed to the situation in other languages. Therefore, the model uses the phoneme or parts of the phoneme (in the case of occlusive consonants) as the segmental unit.

The Phoneme Dedicated Artificial Neural Network model consists of a set of 44 ANN. Each ANN is dedicated to predict the duration of each phoneme segment. All 44 ANN have the same architecture and input features.

The idea behind the use of dedicated ANN consists basically in avoiding the possibility of one feature that has opposite influence in the duration of different phonemes getting the same influence given by the ANN trained with data for all phonemes. This occurs in the usage of only one ANN. In the case of the PDANN, the training is performed only with the data related to the respective phoneme without the influence of data relative to other phonemes.

Each ANN is specialized for the respective phoneme segment. Therefore a small improvement in the prediction performance is expected.

Each specialized feed-forward ANN has 55 input nodes, 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. The output has one node with a linear activating function, corresponding to the coded phoneme segment duration.

The computational cost of the proposed model becomes lower because the PDANN has a reduced structure dimension than the equivalent ANN model.

3.1. Training

The training was performed with a Levenberg-Marquardt back-propagation algorithm over the training set and using the validation set in order to avoid over-fitting. The cost function used was the mean squared error between output and target values. Some pre-processing is performed in order to normalise the input and output data.

Each ANN was trained only with the data relative to the respective phoneme.

Input featuresTable *1* presents the 55 input nodes corresponding to the input features, described below, and their linear correlation, *r*, with phoneme segment duration. For features with more than one input node only the major *r* is presented.

The phoneme segments in neighbouring positions were selected by their relevant correlation with the actual phoneme segmental duration.

- Consonant at the end of word: codes, in one node, if the actual segment {/r/, /l~/[1] or /ʃ/} is in the end of the 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 /ʃ/ are correlated with shorter segments in the next position. On the other way the burst part of plosive consonants /t/, /k/, /b/, /d/, /g/, consonants /n/, /ɲ/, /l/, /r/,

[1] /l~/ means /l/ in syllable final position.

/ʁ/, /v/, /z/ and pause, are correlated with longer segments in the next position. These 20 segments are coded by the activation of the correspondent node. This is an important feature.

Table 1: Segment Features

Phonologic level	*Feature*	*# nodes*	*|r|*
Phoneme	Cons. at the end of word	1	0.082
Phoneme context	Previous segment (-1)	20	<0.227
	Next segment (+1)	12	<0.282
	Next segment (+2)	4	<0.141
	Next segment (+3)	2	<0.110
Syllable	Type	1	0.175
	Vowel	1	0.208
Syllable Context	Type of previous syllable	1	0.055
	Vowel in previous syllable	1	0.075
	Vowel of next syllable	1	0.151
	Distance to tonic syllable	1	0.145
Foot	Position in group	2	<0.153
	Position in Phrase	2	<0.244
	Distance to next pause	1	0.203
Accent group	Length	2	<0.052
Phrase	Position of accent group	3	<0.114

- Next segment (+1): segments /a/, /ɐ/, /u/, /ɐ̃/, /õ/, /t/, /d/ are correlated with shorter segments in previous position. Segments /l~/, /v/ and closure part of plosive consonants /t/ and /d/ are correlated with longer segments in the previous position. Pause is even highly correlated with longer segments in the previous position. These 12 segments are coded by the activation of the correspondent node. It is a major feature.
- Next segment (+2): segment /r/ is correlated with shorter last segments but one. Burst part of stop consonants /t/ and /d/ and pause are correlated with longer last segments but one. These 4 segments are coded by the activation of the correspondent node.
- Next segment (+3): segment /u/ and pause are correlated with longer antepenultimate segments. This feature is coded by the activation of the correspondent node.
- Type: the syllables are considered of the following types (where V means a vowel and C a consonant): V, C, VC, CV, CC, VCC, CVC, CCV, and CCVC. Types C and CC result from elision of vowel (very frequent in EP [7]). 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. 5 types of vowels were considered: long /a/, /ɛ/, /e/, /ɔ/ and /o/; medium /ɐ/ and /i/; short /i̥/ and /u/; diphthongs and nasal vowels. Long and nasal vowels are correlated with longer segments in 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 the next syllable. Syllables of types VC, CC and CVC, are slightly correlated with shorter segments in the next syllable. The feature was coded in one node according to the correlation of the respective type of syllable with segments length (different correlation of previous 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 the next syllable. Medium and short vowels are positively correlated. The feature was coded according to the respective correlation with segments length (different of feature 'vowel'). It is a minor feature.
- Vowel of next syllable: the segments length is positively correlated with short vowels in next syllable. The other types of vowels are negatively correlated. The feature was coded in one node according to the respective correlation with segments length (different of features 'vowel' and 'vowel in previous syllable').
- Distance to tonic syllable: five positions were considered to characterize distance to tonic syllable in the accent group: tonic syllable, previous syllable, before previous, next syllable and after next. As it is well known, syllable tonicity is highly correlated with the length of segments, and next and after next are also positively correlated with the length of segments. In opposition, the other categories are negatively correlated. The feature was coded in one node according to the respective correlation.
- Position in group: it is the segment position inside the accent group, taken both from the beginning and end of the group. The position relative to the end of the group is highly and negatively correlated with segments length. The position relative to the beginning is positively correlated with segments length, as expected. It is coded in two nodes.
- Position in Phrase: it is the segment position inside the phrase, both from beginning and end of the phrase. Phrase is delimited by orthographic punctuation. The position relative to the end of the phrase is highly and negatively correlated with segments length. The position near to the beginning is slightly correlated with segments length. It is coded in two nodes.
- Distance to next pause: it is the distance in the 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 an important feature.
- Length: it is the number of phonemes and syllables of the accent group. It is slightly correlated with longer segments. It is coded in two nodes. This is a minor feature.
- Position of accent group: it is the position of the group inside the phrase (beginning, middle and end). Beginning position and specially end position are correlated with longer segments. In opposition, middle

position groups are slightly correlated with shorter segments. It is coded by activation of correspondent node. It is an important feature.

4. Results

4.1. ANN model used as a reference

In order to compare the results of the PDANN model, a reference model was used. The reference model was built using only one ANN with the same input features plus the identity of the segments [8]. Therefore, the input features increased by 44 nodes, in a total of 99 nodes. The training was performed using the training set for all phonemes. Training, validation and test sets were the same.

4.2. Objective results

The standard deviation, σ, and the linear correlation coefficient, r, between original (measured) and predicted segment durations were determined. Table 2 shows a σ and r visibly improved with the PDANN model.

Table 2: Prediction accuracy in test set

	ANN	*PDANN*
σ (ms)	**19.5**	**18.2**
r	**0.839**	**0.861**

4.3. Subjective results

A perceptual test was performed using a Mean Opinion Score (MOS) scale (1- **Unsatisfactory**, 2- **Poor**, 3- **Fair**, 4- **Good**, 5- **Excellent**).

Five paragraphs were used for the tests. A total of 4 stimuli per paragraph were presented in random order to 20 listeners in a blind test. The four stimuli correspond to the original stimuli, re-synthesized with predicted durations by the PDANN model, re-synthesized with predicted duration with the ANN model, and re-synthesized with the average duration of each type of phoneme (named as no model).

Table 3 presents the MOS and standard deviation corresponding to the 100 opinions for each stimulus. Once again the preference went to the PDANN rather than to the ANN model. The overall MOS for the PDANN were "Good".

An analysis of variance (ANOVA) test showed a confidence level of 74% in favor of the PDANN against the ANN model [8].

Table 3: MOS and standard deviation of the perceptual test.

	original	*PDANN*	*ANN*	*No model*
MOS	4.27	3.93	3.78	2.88
σ	0.81	0.96	0.91	1.10

5. Discussion

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

One can argue about the reasonability of the proposed PDANN model for it improves the σ error by only 1.3 ms. But the preference for the PDANN in the perceptual test over 100 stimuli and the confidence level of 74% provides evidence that the improvement is perceptually perceived.

The total number o weights of the whole PDANN model became increased, claiming a larger number of training vectors particularly for those less frequent phonemes. Therefore, it is important to increase the number of training vectors for those phonemes.

6. Conclusions

An ANN-based statistical model has been presented. The innovative approach developed in the study consists in using one dedicated ANN for each phoneme segment with a specific training in order to avoid miscellaneous influence of different types of phoneme segments. The model was compared with an equivalent ANN model using the same set of input features.

The objective and also the subjective measurements show a slight**,** but clear, perceptually perceived preference towards the PDANN.

The relatively slight insignificant improvement with the proposed approach suggests that the segmental durations models reached a ceiling level, and there seems to be no room for further improvement. Thus, it is reasonable to assume that the incorporation of other type of features such as non linguistic and even paralinguistic features are the way forward.

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", Eurospeech 93, vol. 2, pag. 1081-1084.

[5] Barbosa P., "*Caractérisation et génération automatique de la structuration rythmique du français*.", Thèse de Docteur de L'Institut National Polytechnique de Grenoble 1994.

[6] Barbosa, P., A Model of Segment (and Pause) Duration Generation for Brazilian Portuguese Text-to-Speech Synthesis. *Proceedings of Eurospeech'97*, Rodes, pages 2655-2658.

[7] 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.

[8] Teixeira, J. P. A Prosody Model to TTS Systems, Doctoral Thesis, Faculty of Engineering of Univ. of Porto, 2004.