Computer Speech and Language (2002) **16**, 183–203
doi: 10.1006/csla.2002.0190
Available online at http://www.idealibrary.com on IDEAL®

Selection of the most significant parameters for duration modelling in a Spanish text-to-speech system using neural networks

Ricardo Córdoba,†§ Juan M. Montero,† Juana M. Gutiérrez,† José A. Vallejo,† Emilia Enriquez‡ and José M. Pardo†

†Grupo de Tecnología del Habla—Departamento de Ingeniería Electrónica, E.T.S.I. Telecomunicación—Universidad Politécnica de Madrid, Ciudad Universitaria s/n, 28040 Madrid, Spain, ‡Facultad de Filología, Universidad Nacional de Educación a Distancia, Senda del Rey s/n, 28040 Madrid, Spain

Abstract

Accurate prediction of segmental duration from text in a text-to-speech system is difficult for several reasons. One which is especially relevant is the great quantity of contextual factors that affect timing and it is difficult to find the right way to model them. There are many parameters that affect duration, but not all of them are always relevant and some can even be counterproductive because of the possibility of overtraining.

The main motivation of this paper has been to reduce the error in the duration estimation. To this end, it is of the utmost importance to find the factors that most influence duration in a given language. The approach we have taken is to use a neural network, which is completely configurable, and experiment with the different combinations of parameters that yield the minimum error in the estimation.

We have oriented our work mainly towards the following aspects: the most significant parameters that can be used as input to the automatic model, and the best way to code these parameters. We have studied first the effect of each parameter alone and, after that, we have included all parameters together to have our final system.

Another important aspect of this study is the generation of a suite of software tools and design protocols that will be used in future tasks with different speakers and databases. The applications for automatic modelling are obvious: adapt the prosody to a new speaker, to a new environment, to "restricted-domain" sentences, etc., in a fast, semi-automatic and inexpensive way. After the database labelling, it is a matter of minutes to prepare the inputs to the network for the new situation, and the network is trained in 1 h.

The result has been a system that predicts duration with very good results (19 ms in RMS) and that clearly improves our previous rule-based system.

§ Author for correspondence: E-mail: cordoba@die.upm.es

0885–2308/02/$35.00

1. Introduction

The primary goal of this study was to develop an automatic system to model duration for a Spanish text-to-speech system and achieve better results than those obtained with our previous system, which used a rule-based approach (Santos, 1984; Pardo, Martínez, Quilis & Muñoz, 1987). We will compare them in this paper. The rule-based approach follows a classic multiplicative model, also known as the Klatt model (Allen, Hunnicut & Klatt, 1987). Beginning from an initial inherent duration based on the phoneme identity, this value is multiplied by a series of factors based on the characteristics of the phoneme, its position, stress, etc.

Now, our objective is to develop a self-learning model: using a general-prosody segmented database for training and considering the characteristics (or parameters) of every phoneme, the model has to predict the estimated duration of the phoneme, without inferring any rule. At the same time, we wanted the system to be very flexible, so we could easily adapt it to new contexts, especially to restricted-domain prosody, an area of research that we are also very interested in. For example, in some banking applications, the text-to-speech system has to generate sentences with a very similar structure. Applying the environment presented in this paper to this application has produced outstanding results.

Many studies have been successfully carried out lately in the field of automatic estimation of a duration model, using different techniques and input parameters to obtain the model. These automatic techniques are mainly of two types: decision trees and neural networks (the objective of this paper). Another line of investigation with very good results is the statistical sum-of-products method. We will now highlight recent work in all these areas.

In all the systems, regardless of the technique used for the modelling, it is crucial to find the parameters that are most significant for duration modelling. So, we can take advantage of previous studies dedicated to duration modelling, but using other techniques. For example, the technique using a sum-of-products model from van Santen (1994) used the following parameters: phoneme identity, surrounding phonemes, pitch accent, syllabic stress, within-syllable position, within-word position (initial/not initial, final/not final), within-utterance position (last syllable/penultimate syllable/other). In van Son and van Santen (1997) they used the following parameters: phoneme identity, stress, position in the word (initial, medium, and final), word length in syllables (one, two, three, or more than three) and the frontedness of the syllabic vowel. In Möbius and van Santen (1996) the sum-of-products model is applied to German. A very large set of possible parameters is considered, and the most relevant parameters (statistically determined) are the syllable stress, the word class, and the presence of phrase and word boundaries. In van Santen, Shih, Möbius, Tzoukermann and Tanenblatt (1997) where multi-lingual duration modelling is described, they distinguish three groups of core factors: phone identity, stress-related and locational. We have studied all of them in this paper and adopted a similar solution for the locational factors: location of the segment in the syllable, syllable in the word, word in the phrase and phrase in the utterance.

Eloquens (Quazza & Heuvel, 1997) is a system for Italian using decision trees. The parameters considered in this system are: phoneme identity, stress, window of phonemes, and characteristics of higher order units (syllable, word, and phrase). In Deans, Breen and Jackson (1999), they propose a CART-based approach using prosodic features obtained from the TTS system, but the results were not as good as expected.

Neural networks have previously been used with success. In Campbell (1992) a neural network was trained to predict syllable timing. The parameters used are the number of phonemes in the syllable, nature of the syllabic peak, position in tone-group, type of foot, stress and word-class. In Riedi (1995) they apply a neural network to duration modelling in German

with very good results. In Tournemire (1997) a syllable duration modelling is shown, in which the neural network is used to predict the standard deviation of the duration (called the "syllable elasticity factor"). The parameters used are the breaks, the stress, and the information on the composition of the syllable. Corrigan and Massey (1997*a*,*b*) propose a hybrid system using a neural network and a rule-based system. It is the neural network which is responsible for the final estimation of the duration, and the rule-based system is only used to obtain the inputs to the network (which are all binary). In concept, the solution is similar to ours, as we always need a preprocessing step to obtain the inputs to the network, the difference being that we accept different inputs, not only binary. Morlec, Bailly and Aubergé (1997) present the "sequential neural networks", where there is a set of neural networks working in parallel to predict the prosody of the sentence. In Fackrell, Vereecken, Martens and Van Coile (1999) they compare the performance of neural networks and CART techniques for six different languages, including Spanish. The results for both are very similar, which shows that any of them can be used. They also propose an architecture with three levels of cascaded modules for prosody prediction. Another possibility is to use "multivariate adaptive regression splines" (Riedi, 1997). Its objective is to determine automatically the structure and parameters of the model with sparsely distributed data.

Regarding the application of these techniques to Spanish, there are very little references and none is dedicated to neural networks or CART approaches. For example, the TTS from the Spanish dominant telephone company—Telefónica—uses a multiplicative model (Rodríguez-Crespo, Escalada & Torre, 1998; Macarrón, Escalada & Rodríguez, 1991), similar to our previous solution. The factors that were considered as significant in this system were: position of the phoneme within the phrase, phrase length, position of the phoneme within the word with respect to the stressed vowel, left and right context of the phoneme. They suggested that they would use "type of phrase" in future works. As we will see in this paper, we have covered all these parameters in our system and some more. In Villar, López-Gonzalo and Relaño (1999) they use a database of syllables, each one associated with its linguistic properties, and search the closest syllable in the database using a dynamic programming algorithm. This way, they model f0 and duration at the same time. In previous works of this group, they propose very little variance to this approach, as in López-Gonzalo and Rodríguez-García (1996), but they do not provide any result about the accuracy of the estimation (no measure of error in the estimation). So, we cannot compare our results with theirs. In Febrer, Padrell and Bonafonte (1998) they implement a sum-of-products model for Catalan and compare it with the multiplicative model. There is a deep study of factors that affect duration in Catalan using a CART tree. In Fernández-Salgado and Banga (1999), there is a study on duration for Galician, which is very close phonetically to Spanish. They decided that the most significant parameters for duration were: number of allophones in the syllable, position in the syllable, broad phonetic class, stress and final position (they discarded phoneme identity and contextual information). They used a look-up table for the estimation.

In this paper, we will show and analyse new alternatives for the automatic techniques mentioned earlier, specifically for the parameters used and their coding as inputs to the neural network, and for the way to model the duration, as there are several alternatives: the duration itself, a normalized duration, the logarithm, standard deviation, and different combinations of these measures.

TABLE I. Number of units in the database

List	Phonemes	Syllables	Words	Phrases
1	3 224	1355	654	94
2	2 354	1015	541	90
3	3 867	1671	882	138
4	2 440	987	497	164
5	3 256	1314	612	198
All	15 141	6342	3186	684

2. Database used for the modelling

2.1. Contents

Our database consists of five sets of phrases of different lengths and patterns, giving a total of 732 phrases (15 141 phonemes). We have used a single speaker with a neutral voice, looking for homogeneity. In previous studies using many speakers, the variation in our parameters was too high and it was difficult to extract reliable statistical decisions.

We have used this database to model both fundamental frequency (see a thorough study in Vallejo, 1998) and duration. So, some considerations of the design of our database are dedicated to intonation modelling. We split the database into five lists with the following types of sentence in the database:

- List 1: a discourse, with a predominance of declarative patterns, especially all patterns relative to complex phrases, but with very few interrogative and exclamatory sentences. This discourse has the same content as the one used to develop our first rule-based system.
- Lists 2 and 3: colloquial text, read as an interview, with more interrogative and exclamatory sentences to complete list 1.
- Lists 4 and 5: laboratory sentences, designed to complement the previous text. It includes at least one sample of all the possible patterns with one or two stressed syllables in sentences with up to eight syllables. The distribution of this set follows the distribution of a large set of texts in Spanish. The objective is to improve the coverage of all possible vectors of parameters that we will have in our system.

In its design, the goal was to cover not only intonation patterns, but all syntactic and stress patterns in Spanish, and have an appropriate number of each of them to be able to train our model. To this end we have followed the studies developed in Navarro Tomás (1948). Similarly, we wanted to have enough examples of words with a different number of syllables, with one or two stressed syllables and with different positions of the stressed syllable inside the word.

The detailed figures of our database are shown in Table I. Throughout the paper we will also mention the results we have obtained in similar experiments using a female speaker in a restricted-domain environment (Córdoba, Montero, Gutierrez-Arriola & Pardo, 2001). This restricted-domain offers several advantages to the modelling: the variation in the different patterns is reduced, and there are more instances of each vector of parameters in the database. Besides, this database is slightly larger (1735 phrases, 3594 words, 6551 syllables, 20 089 phonemes).

2.2. Separation in training and testing

We broke down the database into three parts: 60% is used for the "training set" (i.e. adjust the parameters in our system), 20% for the "development test set", and 20% for the "validation set". The topology of the network is determined considering the results in the "development test set", and the final system is checked against the "validation set". In this division, we decided to assign one complete phrase of each type alternately to training, testing, and validation.

2.3. Database segmentation

The segmentation of the database has been automatic, using a continuous speech recognition system with HMM models (Ferreiros, Córdoba, Savoji & Pardo, 1995), followed by a manual revision of the marks to correct some of the accuracy problems of the system. In this manual segmentation we were limited to 10 ms steps. The segmentation has been made at the phoneme level, as the phoneme is the base unit in our modelling. To obtain the syllable boundaries we have developed a module to generate those boundaries from text using Spanish rules. In Spanish the syllabication rules are known, fixed and well documented. The module is the same as the one used in our TTS system.

3. Neural network system development methodology

In order to develop a system based on neural networks, we have to make several decisions: Which topology should we use for our network? Which is the best way to code the parameters? Which are the best parameters that we should use (the most representative for our modelling)?

3.1. Topology of the neural network

We have used a multilayer perceptron (MLP) and the sigmoid as the activation function. The training algorithm is backpropagation. The experiments in this section were for the purpose of fine-tuning the network. We worked with what we call the "base experiment", that includes as parameters just the phoneme identity and the stress.

Our basic unit is the phoneme. For each phoneme, we compute a series of parameters, which we code and use their values as inputs to the neural network. There is one output in our network: the duration of the phoneme.

In our experiments, we divided the training into three phases of 200 iterations each. This way we could analyse the evolution of the error for the test set as a function of the number of iterations and detect a condition of overfitting (network too adapted to the training data).

It is very difficult to know the optimum number of neurons and layers that the net should have. A large number usually causes overtraining (results decrease for the test set although they improve for the training set) whilst a small number may be insufficient. Our approach has been to begin from the very beginning: just three neurons. The best procedure is to increase the number of neurons one by one and observe the test results (training results are not significant for our problem, as the greater the number of neurons, the greater the improvement), and stop when there is an overtraining symptom (e.g. there is an increase in RMS on the development test set). As we will see, in some cases we got better results with just three or four neurons. So, we have to be careful in the design, as using two layers and many neurons is rarely the best approach.

3.2. Coding of the parameters

We have considered different ways of presenting the parameters to the neural network, i.e. the way they are coded, as we have different kinds of parameters. A good explanation of all possible codings can be found in Masters (1993).

1. Binary coding: this is the standard coding for true/false parameters. A binary "one" is assigned to the true condition and a binary "zero" to the false condition.
2. One-of-n coding: we use n neurons and only one of them is active, that is, the one that corresponds to the class or category.
   In ordinal values we have more possibilities, because there is an order relationship between the different values. For example, the position of a unit inside a higher-order unit. For these values we have considered three codings:
3. Percentage transformation: we divide the current value by the maximum value to obtain a percentage. We obtain a floating-point value between 0 and 1 as input. This transformation presents many problems when the distribution of the parameter is not uniform. If there are values very far from the average (the maximum value is too high), the secondary effect is a compression of the information that is close to the average, which means a decrease in the performance of the transformation.
4. Thermometer: we divide all the possible values into different classes (intervals). We activate all the neurons until we get to the current class and leave the remaining neurons inactive. If we use equidistant classes, we have the same problem that we had with the percentage transformation: if the distribution of the parameter is not uniform, the classes will be unbalanced, which has a bad effect on the training of a neural network. So, to decide the size of each class we developed an algorithm to obtain a uniform distribution of all the classes. This algorithm considers the whole database and the histogram of all the values of the parameter.
5. Z-score mapping: we apply a normalization to the floating-point value that takes into account the average and the standard deviation of the variable. It is a good coding for high variance parameters, as is the case for all the "number of units" parameters.

3.3. Network evaluation

To evaluate and compare the networks we have considered different metrics for the error (the difference between the prediction from the network and the optimum value) (Masters, 1993).

1. The RMS is equal to the square root of the mean square error (MSE). It is the most important metric, and is more reliable than the average absolute error.
2. To make our comparisons it is better to use an adimensional metric, because it will be independent of the way we code the duration. We decided to use the following metric because it does not include any offset and is independent of the average value of the magnitude compared. This way, we can use it to compare results obtained with different-size databases.

$$\text{Relative RMS error} = \frac{\text{RMS}}{\sqrt{\sum (t_i - \bar{t})^2}}$$

   where t_i are the target values.

3.4. Parameters to be used

3.4.1. Background

The different parameters considered in our system are the result of thorough statistical studies applied to Spanish that were carried out previously in our laboratory (Santos, 1984) using a multiplicative model (Allen *et al.*, 1987): each phoneme is assigned an initial inherent duration based on the phoneme identity, and then this value is multiplied by a series of factors. The following factors were determined as relevant for duration in Spanish:

- Poly-syllabic shortening: depending on the number of syllables in the word, the initial duration of the vowel is reduced accordingly.
- Stress, syllable structure (syllable ending with/without vowel) and position of the syllable in the word. All these three parameters were relevant for duration, but their efforts to model them independently using their multiplicative model were useless. That was the reason to model them together, i.e. they obtained one single value for each combination of the three parameters.
- Post-vocalic context: considering the type of the consonant that followed the vowel (plosives, voiceless/voiced fricatives, voiceless affricate, nasal, lateral) they computed the different factors for the multiplicative model.
- Final lengthening: they found that the vowel in a syllable which is before a pause is lengthened by a factor between 11 and 47%, depending on the number of syllables in the word and the syllable structure.

In a later work by Berrojo (1994), significance tests were made to guarantee that these factors were significant in the database used in this paper. All of them were successful, and the following factors were found significant too:

- Voiced/unvoiced post-vocalic context. In Spanish, unvoiced contexts lengthen the vowels (5% over voiced contexts).
- Function words: duration of vowels in function words is shortened 30% compared to vowels in non-function words. The difference is significant.
- Number of phonemes in the syllable: there are significant differences between the three classes: one phoneme, two phonemes, three or more phonemes.
- Position in the phrase: considering initial, medium and final syllable, there are significant differences.
- Size of the phrase: there were no significant differences for the classes considered.

These studies are the basis we have followed for the election of our parameters. We have considered all of them (the only difference is the way to code them for the neural network) and introduced new variations for type of phrase, position and “number of units” parameters.

3.4.2. Methodology

As we found out in Córdoba *et al.* (1999), it is too difficult to decide which parameters are relevant and the best way to code them using a very large network with many parameters, because the differences in performance are too small and not always consistent. So, we have used a base experiment using only the phoneme identity and the stress, which are, without doubt, the most relevant ones, as all previous work related to duration estimation—and our own pilot experiments—has shown.

Then, we have added the different parameters one by one to see their effect and the significance of each of them. We have considered the introduction of contextual information in some of them, i.e. windowing information.

We have carried out the experiments in four directions:

- First, we used a fixed number of neurons (three) to test each possible parameter.
- As increasing the number of parameters demands more neurons in the network, we increased the number of neurons until we detected an overtraining symptom.
- We tested different codings for non-binary parameters.
- We included all the parameters to get the final network.

Now, we will describe in detail the procedures followed for all the parameters that we have considered.

3.4.3. Phoneme identity

This is the most obvious one. In our first approach, we considered 51 different phonemes with all the allophonic variations for Spanish, including four variants of each vowel. In Table II we can see these units. The first column shows the class of the phonemes; the third and fourth columns show examples of each phoneme in Spanish and English (if there is any correspondence). In Quilis and Fernández (1989) there is a complete description of Spanish phonetics and its relation to the English one.

We analysed the distribution of phonemes in the database and observed that many of them had too few samples, which is an undesirable situation for the training of the network. So, we decided to cluster the allophones with a very low number of repetitions in the database: the variants for each vowel and the variants for the vibrant (the "r"). The fifth column in Table II shows this reduction to 33 phonemes. The coding used is a one-of-n coding: a "1" in the input which corresponds to the phoneme and "0" for all the other inputs.

Contextual phonemes. As our previous studies in duration showed, the kind of phoneme adjacent to the phoneme considered significantly affects its duration. In our neural network, we decided to use the phonemes that are to the right and to the left of the current one. As the number of phonemes is too high (we need 99 inputs for the three phonemes in the window) the results obtained are low, showing that there are not enough examples in the database to train all the possible contexts.

The solution to this problem is to make clusters of phonemes of the same class, considering mainly its manner of articulation (just in two cases where the number of instances of the phoneme in the database was very low we have considered the similarity of their duration in the distribution). Then, we divide our set of phonemes into 13 classes and classify the left and right context phonemes in these classes. The result of the clustering can be seen in the sixth column in Table II.

This way we reduce the inputs to 59 (33+13+13). The results improve with this approach. So we will use it in our base system. The next step would be to consider a two-phoneme context (a window of five values, instead of three), but the results decreased slightly for the test set. The most probable reason is the excessive number of inputs—and, at the same time, of weights to be trained—needed for this window. So, we discarded this option at this point.

TABLE II. Phonemes considered in the system

	Complete set of phonemes (51)	Examples	Equivalent in English	Reduced set (33)	Clusters of phonemes
Vowels	a e i o u	p**a**labra reduc**e** real**i**dad teatr**o** j**u**sticia	**a**isle f**a**il **i**t n**o**vember g**oo**d	a a’ ∼a ∼a’ e e’ ∼e ∼e’ i i’ ∼i ∼i’ o o’ ∼o ∼o’ u u’ ∼u ∼u’	All vowels
Stressed vowels	a’ e’ i’ o’ u’	pal**a**bra r**e**to m**i**to l**o**gro j**ú**bilo			
Nasal vowels	∼a ∼e ∼i ∼o ∼u	**a**ndar **e**n m**i**nistro auton**o**mía **u**nido			
Stressed nasal vowels	∼a’ ∼e’ ∼i’ ∼o’ ∼u’	m**a**no fom**e**nto m**í**nimo t**o**nto m**u**ndo			
Diph-thongs	j (i before a, e, o) w (u before a, e, o)	p**i**e c**u**atro	**y**es **w**ine	j w	j w
	I (i after a, e, o) U (u after a, e, o)	a**i**re ra**u**do	— —	I U	I U
Lateral	l L (Spanish ll)	**l**ado ca**ll**e	**l**eaf —	l L	l L
Vibrant	r (between vocals)	pe**r**o	**r**ed	r	r R R/ R* R_
	R (begin of word) R/ (cons. + r) R* (final r) R_	**r**opa pat**r**ón juga**r** pe**rr**o	— — — —	R R/ R*	
Plosive	p t k	**p**ar **t**ope **c**ama	**p**ay **t**ea **c**old	p t k	p t k
	b d g	**b**ar **d**ar **g**ana	**b**ar **d**ay **g**o	b d g	b d g
Fricative	B D G	sa**b**er Co**d**o pa**g**a	— **th**en —	B D G	B D G
	s f T (Spanish “z”) X (Spanish “j”)	**s**ol **f**in **z**umo **j**ota	**s**ee **f**oot **th**in —	s f T X	s f T X

Table II—*continued*

	Complete set of phonemes (51)	Examples Examples	Equivalent in English	Reduced set (33)	Clusters of phonemes
Nasals	n m	**n**o **m**amá	**n**o **m**ake	n m	n m J
	N	ta**n**go	lo**n**g	N	N J/
	N∼ (Spanish "ñ")	ca**ñ**a	—	N∼	N∼ T/
Affricates	T/ (Spanish "ch")	**ch**ico	**ch**eap	T/	
	J (Spanish "y") J/	a**y**er cón**y**uge	— **j**ump	J J/	

3.4.4. The stress

The effect of this parameter is always relevant for duration modelling (the duration for stressed vowels is 20% longer on average than for unstressed vowels), so we have included it in our base experiment. We use the lexical stress, which could be defined as a phonological feature by which a syllable is heard as more prominent than others. The perceived prominence may be due to a longer duration, F0, intensity, or a combination of all of them, although usually it also includes a component of phonation. In Spanish, words only have one stressed syllable (with just one exception).

The coding is binary: the phoneme may or may not have stress. In this section, we studied the inclusion of contextual information about the stress (adjacent values). We experimented with different sizes for the window of stress values and obtained better results using five values.

In Table III, first row, we can see the results of the base experiment considering just the phoneme identity and the stress (both with the windowing mentioned earlier). In this table, the relative RMS (see Section 3.3) for the train and test set is presented for all experiments dedicated to the study of individual parameters. Each row corresponds to an experiment in which a new parameter has been added to the base experiment.

3.4.5. Binary parameters

Later, we show other binary parameters that have been considered in our experiments. As the coding is fixed, we have worked in their effect for different numbers of neurons. All of them have shown an improvement over the base experiment (see experiments 1–4 in Table III). That is what we expected, as all of them affect the duration.

- Stress in the syllable. It is similar to "stress in the phoneme", used in the base experiment, but now all the phonemes of the syllable are considered as "stressed" when the main vowel is stressed. In spite of being somewhat redundant, the inclusion of this parameter has been positive, showing again the relevance of stress-related factors. Analysing our database, stressed syllables were 25% longer on average than unstressed ones (Berrojo, 1994).
- Syllable structure (syllable ending with/without vowel). A syllable ending with vowel is called an open syllable, which affects duration significantly, as we have observed in the statistical analysis mentioned in Section 3.4.1. Vowels in an open syllable are longer: 13% for stressed vowels and 9% for unstressed vowels (on average).
- Diphthong ("i/u" before/after "a/e/o"). In Spanish, we differentiate both of them as different allophones, and they follow different rules for duration. In previous studies (Santos, 1984) we observed that there were different degrees of shortening for them.

TABLE III. Summary of results (relative RMS)

Experiment	*N*	Train	Test
Base experiment	4	0.7877	0.8261
1-Base + stress in syllable	6	0.7887	0.8215
2-Base + diphthong	6	0.7827	0.8200
3-Base + syllable structure	5	0.7767	0.8232
4-Base + function word	3	0.7870	0.8237
5-Base + type of phrase (8 types)	3	0.7877	0.8221
Base + type of phrase (4 types)	3	0.7850	0.8180
6-Base + pos. of P in S (3 classes)	5	0.7836	0.8189
Base + pos. of P in W (5 cl.)	4	0.7773	0.8204
Base + pos. of P in PHR (5 cl.)	4	0.7848	0.8215
7-Base + pos. of S in W (4 cl.)	3	0.7863	0.8190
Base + pos. of S in PHR (6 cl.)	5	0.7735	0.8116
8-Base + pos. of W in PHR (3 cl.)	5	0.7960	0.8251
9-Base + number of P in S	6	0.7882	0.8196
Base + number of P in W	6	0.7750	0.8109
Base + number of P in PHR	6	0.7881	0.8187
10-Base+ number of S in W	6	0.7755	0.8119
Base + number of S in PHR	6	0.7877	0.8183
11-Base+ number of W in PHR	6	0.7777	0.8172
12-Base+ position in the phrase ...	5	0.7755	0.8130

N = number of neurons, P = phoneme, S = syllable, W = word, PHR = phrase.

- Phoneme in a function word. In our database, syllables in a function word are 30% shorter on average.

The last two apply only to special cases, so they do not show a significant improvement in the overall system, but they improve the prediction in their specific cases.

3.4.6. Type of phrase

After a statistical analysis of duration values according to type of phrase we found that there were significant differences between them. So, we decided to include this parameter. In our text-to-speech system, it is the type of punctuation mark that is at the end of the phrase. We are using punctuation in place of explicit prosodic phrase prediction because punctuation is a primary factor in phrase prediction in Spanish. We have considered the following punctuation marks:

- Closing exclamation (**!**)
- Opening bracket (**(**)
- Closing bracket ( ))
- Comma (,)
- Period (.) and semicolon (;)
- Colon (:)
- Final question mark (?)
- Any other punctuation mark.

Our first experiment was to use a one-of-n coding (with eight inputs), but the results were not as good as expected. A thorough examination of the database showed that the distribution of phrases was not uniform, some types had significantly less examples than the others.

The solution was to reduce the number of types to four obtaining a more uniform distribution:

- Closing exclamation (**!**) and final question mark (?)
- Opening bracket (**(**), closing bracket ( )) and comma (,)
- Period (.), colon (:), and semicolon (;)
- Any other punctuation mark.

Using a one-of-n coding as before, the results were good (see Table III, item 5), showing an improvement of 1·0% for the test set.

3.4.7. Position parameters

We have considered different alternatives for parameters related to position:

- Position of the phoneme in the syllable, word and phrase.
- Position of the syllable in the word and phrase.
- Position of the word in the phrase.

Phrase boundaries are determined using the punctuation marks mentioned in the previous section. In our first approach we carried out the following steps for the coding:

1. Normalize the value of position dividing it by the total in the higher-order unit—we obtain a floating point value between 0 and 1. We need to subtract "1" prior to the division to obtain values from 0 to 1.
2. This value is coded using four classes. The intervals that define these classes are computed automatically. We consider all the values of the parameter in the training database and compute the boundary values looking for uniform distributions, i.e. that the number of examples in each class is balanced.
3. The four classes use a thermometer-type coding with three inputs (always the number of classes minus 1).

The first results using just three neurons showed very little improvement. After analysing the results we decided to increase the number of neurons and use a different number of classes for each parameter. It is difficult to find an exact rule for deciding the optimum number of classes. Our approach has been to use the average value of the parameter in the whole database as the first approximation of the number of classes, and experiment with different values around that number.

These are the considerations made for each parameter:

- Position of the phoneme in the syllable: the average value is 2·27 and the optimum number of classes is three. The intervals obtained automatically are $x = 0$, $0 < x < 0{\cdot}75$, $x > 0{\cdot}75$.
- Position of the phoneme in the word: the average value is 4·5 and the optimum number of classes is five. The intervals obtained automatically are $x = 0$, $0 < x < 0{\cdot}3684$, $0{\cdot}3684 < x < 0{\cdot}6429$, $0{\cdot}6429 < x < 0{\cdot}9474$, $x > 0{\cdot}9474$.
- Position of the phoneme in the phrase: the average value is 21 and has a high variability. We experimented with five, 13, and 21 classes, but the results were not good for any of them. It was a foreseeable result, as this information is not relevant to duration modelling.
- Position of the syllable in the word: the average value is slightly lower than two and the optimum number of classes is four.

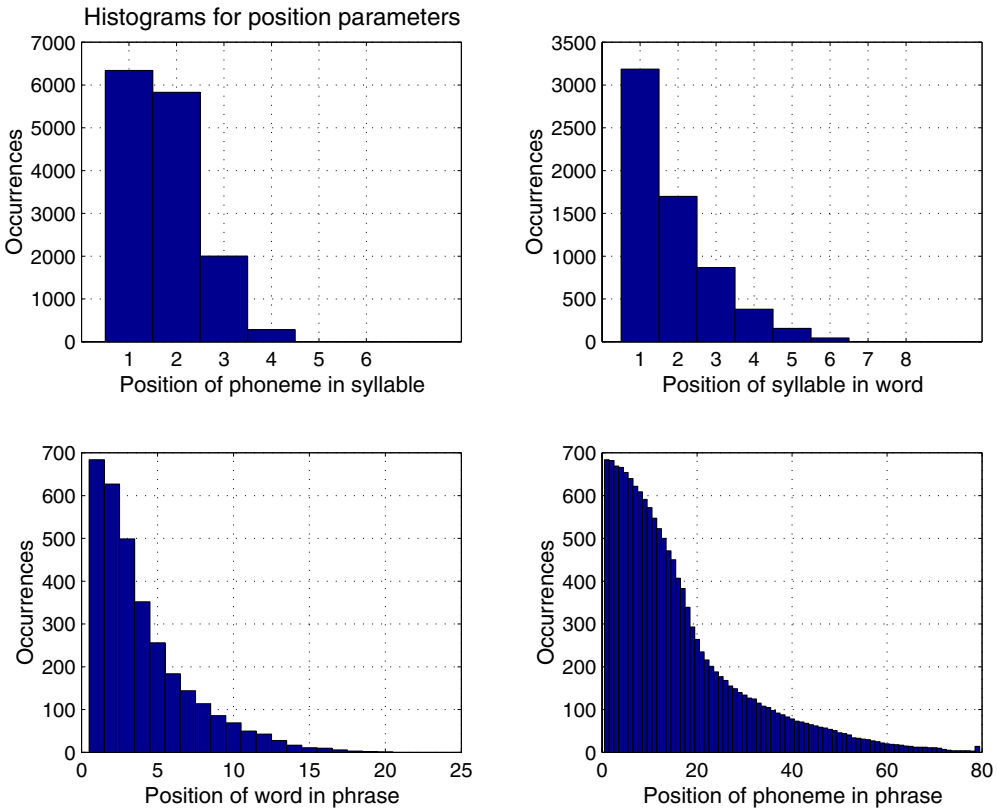

Figure 1. Histograms for position parameters.

- Position of the syllable in the phrase: the average value is nine and the optimum number of classes is six. The results are low again, showing the high variability of this parameter and its difficulty to generalize.
- Position of the word in the phrase: the average value is five and the optimum number of classes is three.

The best results for each parameter are shown in items 6–8 of Table III. All of them have improved the base experiment, and—although the results do not show significant differences between them—we have decided to use: phoneme in the syllable, syllable in the word, and word in the phrase. The reason is that they provide different information to the network (they are not redundant), their range of values is smaller, and, because of their smaller range, they need fewer neurons and classes to reach an optimum. Another advantage of this solution is that it automatically responds to effects like the lengthening before a pause, where all inputs to the network will be close to 1 for these position parameters.

In the first three histograms of Figure 1 we can see the distribution of the selected parameters. The fourth histogram of Figure 1 displays the distribution of 'position of the phoneme in the phrase", which shows its great variability compared to the three selected parameters. That is a good reason for the difficulties we had to model it.

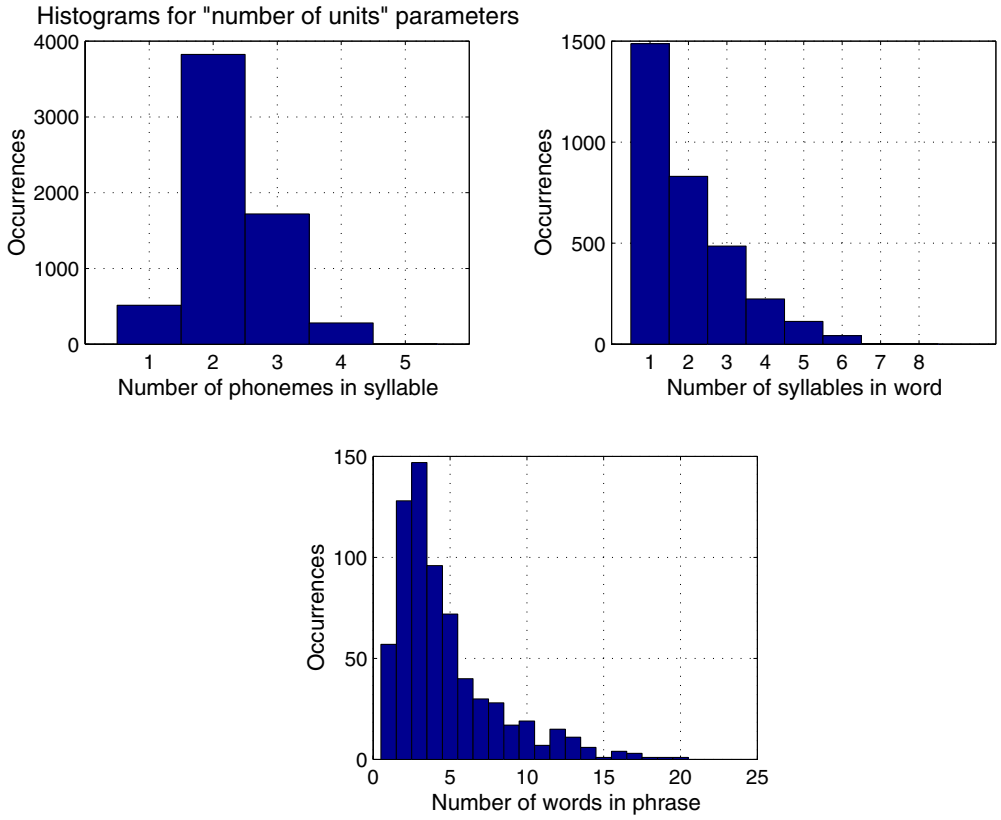

Figure 2. Histograms for "number of units" parameters.

3.4.8. "Number of units" parameters

In a similar way as for parameters related to position, we have considered the number of phonemes in the syllable, word and phrase; syllables in the word and phrase; and words in the phrase. Because of its different distribution, we needed a different coding for this kind of parameter.

We decided to follow the following steps for their coding:

1. Normalize the value of position by the maximum value—we obtain a floating point value between 0 and 1.
2. Apply Z-score (using average and standard deviation) as it is the usual recommendation in neural network literature (Masters, 1993): we can restrict at our will the operating range of the parameter, which is too variable.

The results can be seen in items 9–11 of Table III. The conclusions that can be extracted are very similar to those of position parameters. All parameters referring to phrase provide worse results, which is due to their broad range of values. As can be seen in Figure 2, the range of values for "number of words in phrase" is bigger.

In order to check the suitability of this coding we tested the thermometer-type coding instead of the floating point one (as for position-related parameters). We applied it to the

number of phonemes in the word. The average value is 4·5 and the variability is low. We considered three, four, five, and nine classes, obtaining the optimum three classes, but in all cases the results were not as good as those obtained with the floating point coding.

At this point, we wondered whether the floating point coding was better for position-related parameters than the thermometer-type coding. We carried out a similar experiment but the results were slightly lower. In any case, when results are similar, in a neural network it is always safer to use binary inputs that are more suitable for the training (considering the final network, where the training is more difficult because of the large number of inputs and neurons).

3.4.9. Position in the phrase in relation to first/last stress

The motivation of this parameter is the explicit inclusion of the "lengthening before pause" effect, which is relevant for all languages. We code each syllable in five possible classes: beginning of phrase till first stress; first stressed syllable; after the first stress and before the last stress; last stressed syllable; from the last stress till the end of the phrase. The improvement obtained (1·6%) was remarkably good (see Table III, item 12). In Subection 3.5 we will check if this improvement can be additive, combining in the same model this parameter with the position parameters.

We compared the five classes coding with using just two classes: beginning of phrase till last stress, from the last stress till the end of the phrase. We thought that it could capture better the "lengthening before pause" effect, but the improvement was only 0·6%, so we decided to use the coding with five classes.

3.4.10. Summary of results for parameter evaluation

The summary of most relevant results is shown in Table III. We have been able to find a good coding for all the parameters, as there is an improvement in all of them. We have obtained the best results for: type of phrase, position in the phrase in relation to first/last stress, and the "number of units" parameters. In any case, the maximum improvement with a single parameter has been 1·8% for the development test set.

We have applied the same methodology to a female speaker in a restricted-domain environment (Córdoba *et al.*, 2001). The most remarkable differences with the unrestricted-domain database presented in this paper are the following: the improvements are bigger, up to 5%; all position and "number of units" parameters give consistent improvements (between 2 and 3·5%); syllable structure and function word mean an improvement close to 2·5%; the parameters related to phrase provide better results; the best parameter is "position of the word in the phrase"; and the window of five phonemes for the phonemes identity is 5% better than the window of three phonemes.

3.4.11. Modelling of the duration

To model the duration the first decision to make is whether it should be normalized. As the results in Table IV show (experiment B), we found that it is better to normalize the duration of the phoneme by the average phoneme duration in the phrase; this way the system is less affected by changes in speed in the database recordings. The improvement is more than 3% in the development test set and 4% for the train set.

After that conclusion, we considered several transformations for the duration. The objective of these transformations is to balance the distribution and to have a magnitude that

TABLE IV. Results for different ways to code the duration (relative RMS)

Experiment	Train	Test
A—Duration not normalized	0.7877	0.8261
B—Normalized duration	0.7539	0.7993
C—Log of normalized duration	0.7891	0.8359
D—Standard deviation	0.7495	0.7991
E—Standard deviation of the logarithm	0.7493	0.7980

can be more easily trained. These are the options considered (all references are to results in Table IV):

- The duration itself (experiment B).
- The logarithm of duration (experiment C). The objective of taking the logarithm is to balance the distribution of the duration. The results are bad and this option is discarded.
- Find the average duration for each phoneme and model the standard deviation, i.e. the normalized difference of duration from a phoneme-dependent mean (experiment D, Z-score mapping, see Section 3.2). The objective is to minimize the error made by the network, as it includes the characteristic duration of each phoneme in the final prediction. The improvement over option A is close to 3·3% for the test set.
- The standard deviation of the logarithm of the duration: it tries to cover both objectives (experiment E). The improvement over option A is 3·4% for the test set.

The results are shown in Table IV. Clearly, options B, D and E are the best possibilities. Although the differences between them are not significant, the best behaviour in general (considering several experiments not shown in this paper which use different topologies) corresponds to the last option, which we will use from now on.

3.5. Putting everything together

The next set of experiments was dedicated to including all the parameters together. This is the crucial step, because often the improvements obtained for one parameter are not reflected when this parameter is used in conjunction with others. There are two possible reasons for this behaviour:

- The parameters are closely correlated, so one of them does not offer additional information.
- The topology of the network needs to be tuned: including more parameters, the number of inputs also increases, so a larger number of neurons can help to discriminate the information.

It is difficult to determine which of the two reasons is responsible for a bad result when many parameters are included at the same time. Our solution has been to try different topologies each time a new parameter has been added to the network.

In Table V we can see the summary of results using the parameters together. The numbers in the description of the experiments refer to the experiments specified in Table III. In all experiments the optimum was obtained Using eight or 10 neurons in the network.

The first experiments in Table V are new windowing experiments that show an improvement:

TABLE V. Results with the inclusion of all parameters (relative RMS)

Experiment	Train	Test
Base	0.7493	0.7980
13—Base + window of 5 phonemes	0.7293	0.7842
14—Base + window of 7 phonemes	0.7170	0.7844
15—13 + window of 5 in stress	0.7505	0.8055
16—13 + 1 + 4	0.7447	0.7824
17—13 + 1 + 4 + 6 + 7 + 8	0.7238	0.7791
18—13 + 1 + 4 + 6 + 7 + 8 + 9 + 10 + 11	0.7071	0.7774
19—13 + 1 + 4 + 6 + 7 + 8 + 9 + 10 + 11 + 5	0.7120	0.7814
21—13 + 1 + 4 + 6 + 7 + 8 + 9 + 10 + 11 + 14	0.7036	0.7730
22—13 + 1 + 4 + 6 + 7 + 8 + 9 + 10 + 11 + 14 + 2 + 3	0.7059	0.7719

- Experiment 13: it is the base experiment now using a window of five phonemes (clustering into 13 classes the adjacent phonemes, as we did in Section 3.4.3). The improvement was 1·7% for the test set and 2·7% for the train set.
- Experiment 14: we experimented with a window of seven phonemes, but the results were similar for the test set, which was probably due to an excess of inputs to the network.

So, from that point we used the window five phonemes in all the experiments.

- Experiment 15: we used a window of five values for the stress, but with worse results, so we returned to the window of three values.

The last set of experiments (16–22) shows the inclusion in successive steps of the different parameters studied in Section 3.4. The main conclusion is that all inclusions provide an improvement (except the type of phrase), but as could be expected the improvements are not additive. One important aspect is that the inclusion of the parameter "position in the phrase in relation to first/last stress" improved the global system (already using the position parameters), which means that it provides complementary information.

The global improvement obtained in the inclusion of all parameters is 3·3%. In the restricted-domain environment with a female speaker the inclusion of all the parameters meant an improvement of 11·2% for a similar number of neurons, and 18·7% after increasing the number of neurons to 20. This is another advantage of the better coverage of that database: we could increase the number of neurons and still improve the results for the test set.

3.6. Final results

In our best experiment the average absolute error is 230 samples, equivalent to 14·3 ms, which is really close to the maximum accuracy of the segmentation of our database (we were limited to 10 ms steps). The absolute RMS for that experiment is equal to 19·3 ms.

We applied the same topology to the validation set. These are the results: average absolute error 14·2 ms, absolute RMS 19·1 ms and relative RMS 0·7667. So, the results are very similar in both sets. There is a slight improvement in the validation set, but it is obviously not significant. In any case, it shows the robustness of the selected topology.

In our restricted-domain environment with a female speaker the relative RMS was 0·4536, the average absolute error was 11·8 ms, and the absolute RMS was 15·5 ms. Using the rule-based system, the absolute RMS was 28·5 ms. Again, the better coverage of that database guaranteed better results.

We can compare ourselves to Fernández-Salgado and Banga (1999) in Galician, where the RMS was 19·6 ms, but they used a database of 300 short sentences, much more uniform than ours and similar in size (as they admit in the paper). In Riley (1992), the RMS was 23 ms in a system based on CART. In any case, the results are not really comparable as they use very different databases.

3.7. Comparison to the rule-based system

Our previous rule-based system had a relative RMS equal to 0·9055, which is clearly worse than our best result (17·3% worse).

To check the significance of the comparison of both systems, we applied the Student's *t*-test to the comparison of the normalized duration obtained with both systems. We obtained a "2-tail Sig" $= 0{\cdot}000 \ll 0{\cdot}05$, so the means are clearly different. The 95% confidence band for the difference in means is $-0{\cdot}138$, $-0{\cdot}077$.

We have visually observed that the results provided by the net follow the peak values of the duration very accurately, which shows the correct training of our system. Maybe sometimes the neural network cannot reach the value of the peaks, which is the main factor in the error. But in these cases, we have observed that they are mostly due to changes made by the speaker in an emotional way, because the speaker is using special emphasis and/or greater speaking rate variation. In these experiments, we did not want to model these kinds of variations, but it is another line of investigation that we are following, and it will be reported later when it is finished.

This effect can be seen in Figure 3, which shows the comparison between the predicted duration value and the manually labelled value for a typical part (40 phonemes) of the recognition set.

4. Conclusions

Compared to our previous rule-based system, the results are much better, even when using a limited number of parameters. We have a system that can be easily adapted to specific contexts and/or new databases. We have applied the same methodology to a database with restricted prosody and have obtained a major improvement in these circumstances, as the database is more uniform and homogeneous, with many instances of each type of phrase.

With our environment, the generation of the network for a new speaker can be fast, semi-automatic and inexpensive. Two approaches can be taken:

1. The cheapest way is to adopt the existing configuration and generate directly the new network. We have done that for the female speaker with very similar results.
2. If there are no time constraints, some of the experiments can be repeated to fine tune the configuration.

Regarding the topology, it is difficult to find the optimum of the network. It is better to begin with a low number of neurons and increase it step by step. The same applies to the inclusion of parameters: it is better to decide their best coding in small networks. We have found that in most cases a second hidden layer is not necessary. It is good for the training set but not for the test set, so we decided not to use it. The "Z-score" normalization used with numeric parameters shows an optimum behaviour: it adjusts the margin of accepted values automatically and rejects the out-of-range values.

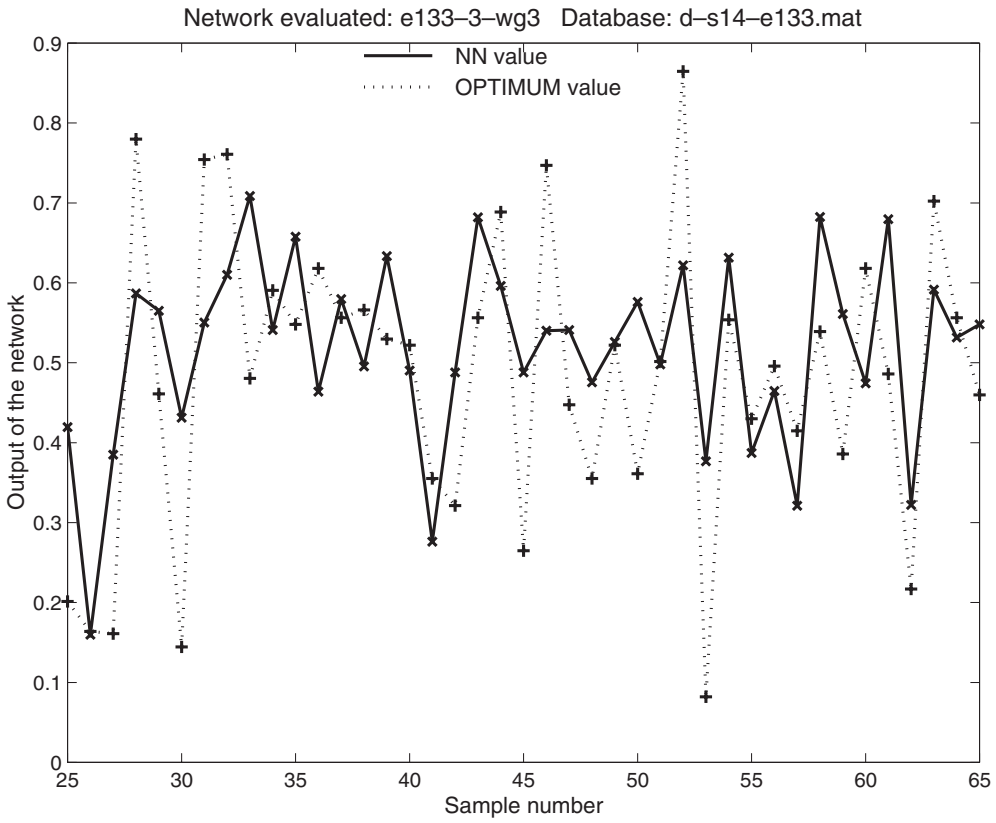

Figure 3. Performance of the neural network.

Regarding the parameters, the most important ones are the phoneme identity and the stress, which is just as we expected. The inclusion of contextual information for them has been positive. For phoneme identity we needed to cluster the phonemes to reduce the number of inputs to the network. Another important aspect is the way to code the duration: we have found better results by normalizing it and modelling the standard deviation of the logarithm. The effect of the other parameters has been positive in general, although they have had more impact on the training set, which shows not only their relevance but also the difficulty to generalize.

In general, we can say that we have found a good compromise between network topology and the parameters considered, with good results that are stable.

The system will be included in a high quality text-to-speech system in Spanish, Boris-GTH (Pardo *et al.*, 1995). This system has been commercially distributed and can be tested on-line at the web address: `http://www-gth.die.upm.es/index.html`.

This work has been financed by the Interministerial Commission of Science and Technology projects Sicovox 194/89 and Demóstenes TIC 95-0147. We acknowledge the help of M. A. Berrojo and J. Ferreiros in the development of the database used in this work, and Miguel A. López-Carmona in the preparation of the environment and the experiments.

References

Allen, J., Hunnicut, S. & Klatt, D. H. (1987). *From Text to Speech: The MITalk System*. Cambridge University Press, Cambridge.

Berrojo, M. A. (1994). Preliminary work in duration modeling in Spanish. Internal Report TR/GTH-DIE-ETSIT-UPM/2-94. Departamento de Ingeniería Electrónica, Universidad Politécnica de Madrid.

Campbell, W. N. (1992). Syllable-based segmental duration. In *Talking Machines: Theories, Models and Designs*. (G. Bailly, C. Benoit and T. R. Sawallis, eds), pp. 211–224. Elsevier, Oxford.

Córdoba, R., Vallejo, J. A., Montero, J. M., Gutierrez-Arriola, J., López, M. A. & Pardo, J. M. (1999). Automatic modeling of duration in a Spanish text-to-speech system using neural networks. *European Conference on Speech Communication and Technology*, 1619–1622.

Córdoba, R., Montero, J. M., Gutierrez-Arriola, J. & Pardo, J. M. (2001). Duration modeling in a restricted-domain female-voice synthesis in Spanish using neural networks. *International Conference on Acoustics, Speech and Signal Processing*, II-793-796.

Corrigan, G. & Massey, N. (1997*a*). Generating segment durations in a text-to-speech system: a hybrid rule-based/neural network approach. *European Conference on Speech Communication and Technology*, 2675–2678.

Corrigan, G. & Massey, N. (1997*b*). Text-to-speech conversion with neural networks: a recurrent TDNN approach. *European Conference on Speech Communication and Technology*, 561–564.

Deans, P., Breen, A. & Jackson, P. (1999). CART-based duration modeling using a novel method of extracting prosodic features. *European Conference on Speech Communication and Technology*, 1823–1826.

Fackrell, J. W. A., Vereecken, H., Martens, J. P. & Van Coile, B. (1999). Multilingual prosody modeling using cascades of regression trees and neural networks. *European Conference on Speech Communication and Technology*, 1835–1838.

Febrer, A., Padrell, J. & Bonafonte, A. (1998). Modeling phone duration: application to Catalan TTS. *3rd ESCA/COCOSDA Workshop on Speech Synthesis*, 43–46.

Fernández-Salgado, X. & Banga, E. R. (1999). Segmental duration modeling in a text-to-speech system for the Galician language. *European Conference on Speech Communication and Technology*, 1635–1638.

Ferreiros, J., Córdoba, R., Savoji, M. H. & Pardo, J. M. (1995). Continuous speech HMM training system: application to speech recognition and phonetic label alignment. In *NATO ASI "Speech Recognition and Coding, New Advances and Trends"* (A. J. Rubio and J. M. López, eds), pp. 68–71. Springer, Berlin.

López-Gonzalo, E. & Rodríguez-García, J. M. (1996). Statistical methods in data-driven modeling of Spanish prosody for text-to-speech. *International Conference on Speech and Language Processing*, 1373–1376.

Macarrón, A., Escalada, G. & Rodríguez, M. A. (1991). Generation of duration rules for a Spanish text-to-speech synthesizer. *European Conference on Speech Communication and Technology*, 617–620.

Masters, T. (1993). *Practical Neural Network Recipes in C++*. Academic Press, Inc, San Diego, CA.

Möbius, B. & van Santen, J. P. H. (1996). Modeling segmental duration in German text-to-speech synthesis. *International Conference on Spoken Language Processing*, 2395–2398.

Morlec, Y., Bailly, G. & Aubergé, V. (1997). Synthesising attitudes with global rhythmic and intonation contours. *European Conference on Speech Communication and Technology*, 219–222.

Navarro Tomás, T. (1948). *Manual de entonación española (Manual of Spanish Intonation)* 2nd edition. Hispanic Institute, New York, NY.

Pardo, J. M., Martínez, M., Quilis, A. & Muñoz, E. (1987). Improving text-to-speech conversion in Spanish. Linguistic analysis and prosody. *Proceedings of the European Conference on Speech Technology*, volume 2, pp. 173–176. CEP Consultants LTD, Edinburgh.

Pardo, J. M., Giménez de los Galanes, F. M., Vallejo, J. A., Berrojo, M. A., Montero, J. M., Enríquez, E. & Romero, A. (1995). Spanish text-to-speech, from prosody to acoustics. *15th International Congress on Acoustics*, 133–136.

Quazza, S. & Heuvel, H. (1997). The use of lexica in text-to-speech systems. *Course at Nijmegen University*.

Quilis, A. & Fernández, J. A. (1989). Curso de fonética y fonología españolas. Ed. C.S.I.C.

Riedi, M. (1995). A neural-network-based model of segmental duration for speech synthesis. *European Conference on Speech Communication and Technology*, 599–602.

Riedi, M. (1997). Modeling segmental duration with multivariate adaptive regression splines. *European Conference on Speech Communication and Technology*, 2627–2630.

Riley, M. D. (1992). Tree-based modeling of segmental durations. In *Talking Machines: Theories, Models and Designs*. (G. Bailly, C. Benoit and T. R. Sawallis, eds), pp. 265–273. Elsevier, Oxford.

Rodríguez-Crespo, M. A., Escalada, J. G. & Torre, D. (1998). Conversor texto-voz multilingüe para español, catalán, gallego y euskera. *SEPLN (Spanish Society of Natural Language Processing)*, **Rev. 23**, 16–23.

Santos, A. (1984). *A system for multi-voice synthesis in real-time from a written text*. PhD Thesis, Universidad Politécnica de Madrid. E.T.S.I. Telecomunicación.

Tournemire, S. de. (1997). Identification and automatic generation of prosodic contours for a text-to-speech synthesis system in French. *European Conference on Speech Communication and Technology*, 191–194.

Vallejo, J. A. (1998). *Improvements to the fundamental frequency in the text-to-speech conversion*. PhD Thesis, Universidad Politécnica de Madrid. E.T.S.I. Telecomunicación.

van Santen, J. P. H. (1994). Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, **8**, 95–128.

van Santen, J. P. H., Shih, C., Möbius, B., Tzoukermann, E. & Tanenblatt, M. (1997). Multi-lingual duration modeling. *European Conference on Speech Communication and Technology*, 2651–2655.

van Son, R. J. J. H. & van Santen, J. P. H. (1997). Strong interaction between factors influencing consonant duration. *European Conference on Speech Communication and Technology*, 319–322.

Villar, J. M., López-Gonzalo, E. & Relaño, J. (1999). A mixed strategy approach to Spanish prosody. *European Conference on Speech Communication and Technology*, 1879–1882.

(Received 25 July 2000 and accepted for publication 25 November 2001)