Argentine Spanish segmental duration prediction

H. M. Torres and J. A. Gurlekian

Laboratorio de Investigaciones Sensoriales, INIGEM, CONICET-UBA
hmtorres@conicet.gov.ar, jag@fmed.uba.ar

Abstract. In this paper we model the segmental duration of Spanish spoken in Buenos Aires, considering its application in a text-to-speech system. The work was performed on two hand labeled databases. We use artificial neural networks as predictor, and all the input features can be extracted automatically from the speech text. We experimented with a neural network for all phonemes and one neural network for phoneme. In both cases the results are very promising for the two databases used. The order of importance of input features revealed to be different for each of the methods tested and different according to the speaker style.

Keywords: Phone duration prediction, Prosody prediction, Text-To-Speech.

1 Introduction

Segmental duration refers to the time period in which a given segment of speech is produced. Duration is the second most important parameter in speech naturalness, after fundamental frequency. If we want to perceive a sentence as natural, all different segment durations must bear some relationship. Both segmental duration and pause locations determine the measure of speech rate. Duration is also the main acoustic correlate of perceived rhythm.

Different speech segments have been proposed as units for analysis: phonemes [10], diphonemes [27], triphonemes [15], syllables [7], group inter-perceptual center [1], and words [6].

Besides, some authors emphasize the sparse character of features affecting segmental duration [30]. For this reason most models identify the feature with more influence, and then try to use this reduced set as input parameters [36].

In a pioneering work, Klatt [17] presents a model of speech production. From this approach, we can identify several factors that affect segmental duration:

1. Extralinguistic factors, such as moods or regionalism. For example, in the Spanish spoken in the province of Cordoba, Argentina, there is an increased duration in the pre-accented syllable.
2. Speech-related factors, such as the position of the segment within a paragraph.
3. Semantic factors. In this sense, the same word has different lengths depending on the meaning it holds in the sentence.

4. Syntactic factors, such as pre pausal lengthening.
5. Factors related to the words, as the elongation of the phones at the end of the word.
6. Phonological/phonetic factors, which include the inherent duration of each phoneme, and the accented character or not.
7. Physiological factors such as inherent duration of each sound and its incompressibility. For example, a phoneme with a greater opening will correspond to a longer duration.

Similar factors have been postulated and accepted for Spanish. The first work dealing with segmental duration dates from 1918, with Navarro Tomas's work [33]. Later on, we can find publications about duration and timing [5], and others works that analyze large databases to extract rules that determine the duration of the phonemes [22]. In [12] they proposed that Spanish consonants change their durations for three reasons: the accent, syllable position, and position in the audio unit. Similarly, in [23] it is stated that a vowel in a prepausal position increases its duration compared to the same vowel at a position not prepausal. It is also proposed that a vowel in an open syllable also increases its duration with respect to the same vowel in a closed syllable. In [24] the results of analysis of the duration of some phonemes in Spanish are presented, and their results underline the effect of phonetic context on the duration change, as well as the inter-speaker variability. Later, in [25] they claim that in Spanish phones at the end of an intonation group suffer a systematic lengthening, and that duration increment depends on sentence type.

As mentioned before, one of the first duration models most widely used, was designed by Klatt for English [17]. This model could be represented by Eq 1, which is successively applied from an initial period, for different values of k_i representing the contributions of different factors that influence the duration of the segment *seg*. In Eq 1, $D_{min,seg}$ is the minimum duration for the segment *seg*, and $D_{seg,i}$ is the length of the segment obtained in the i-th application of the equation.

$$D_{seg,i} = D_{min,seg} + k_i \left(D_{seg,i-1} - D_{min,seg}\right) \tag{1}$$

Features or characteristics that are proposed as factors influencing the duration include: phonetic context; position in the sentence; accent; speech rate; word size; and syllable type.

A set of models, known under the name of *Sum of Products* [30], combines features from Eq 2, where $S_{i,j}$ represents the joint contribution of the factors i and j to the length of segment *seg*.

$$D_{seg} = \sum_i \prod_j S_{i,j}(seg) \tag{2}$$

As can be seen, for a given set of factors or features, we will have a set of potential models. For example, in the model presented in [30] it first carried out a grouping of the segments and then a model sum of products for each group is

applied. As a result, this generates a model for a series of vowels and consonants models. Among the features used we can include: tonal accent; syllabic accent; phonetic identity and context; number of consonants before the vowel in the word; number of syllables to the end of the word; and position in the sentence.

For Spanish, in [10] it has been proposed a neural network, multilayer perceptron type, to predict the duration of phonemes. Inputs are: phoneme identity and context; accent information in a 5 phonemes window; tonal accent of the syllable; function or content word; type of sentence; position of the phoneme within the syllable; position of the syllable within the word; position of the word in the sentence; number of phonemes within the syllable; number of syllables within a word; number of words in the sentence; and distance at the beginning and end of the sentence.

In [16] it is proposed to model the phonemes duration using a technique of case-based reasoning. As input to the system, the identity of the phoneme in question and its context, and the position of the phoneme within the accent group, were used.

In the remainder of this paper we focus on the prediction of phone durations for Spanish, according to the SAMPA alphabet [13]. Also it is included in this task the duration of pauses. This paper is organized as follows: First, we present the databases. Second, the predictor and the input features are detailed. Then we present the results, discussion and finally the conclusions and future works.

2 Databases Definition

We use for our experiments two databases in Spanish, recorded by two professional female announcers, natives of Buenos Aires.

The first one, which we call DB1, was created with the aim to study the prosody [14]. Its text corpus consists of 741 declarative sentences extracted from Buenos Aires Argentine newspapers. The sentences contain 97% of all Spanish syllables, in both stress conditions and all possible syllabic positions within the word. Recordings were made in a sound proof chamber, with an AKG dynamic microphone and 16 Khz/16bit sampling rate conversion. The speaker was instructed to read the sentences with natural tonal variations. The speech material collect was of approximately 40 minutes.

The second database was created to be used in a text-to-speech system, and we call it DB2. The DB1 text corpus was used as basis, supplemented with new sentences with coverage's purpose: follow the distribution of diphonemes, with a minimum of five occurrences. Also, we included 200 interrogatives sentences. The corpus contains 1593 sentences, about 90 minutes of speech.

For the two databases, each sound file was manually labelled twice, by musically trained speech therapists who distinguished prosodic occurrences as intonational groups and accents. The files were labelled in different tiers: phonetic, orthographic, break levels between words, and tonal marks according to an extended ToBI method for Argentine Spanish. Parts of speech and syntactic layers were also indicated.

3 Prediction of phone duration

3.1 Predictor

We use an Artificial Neural Network (ANN)[1] as predictor, and we make experiments with two network topologies: using one network for all phonemes and one network for each phoneme. The networks inputs were coded according to Cordoba et al. (2002) proposal. For the logic inputs we used a binary encoding. For the categories we use a coding *one of n*. For entries with ordinal values we implement a percentage coding. From the available data, 60% was used to train the network, 20% to find the peak of generalization by validation, and the remaining 20% to test the network. We use five set cross-validation, and we test the predictor performance with the Root Mean Square Error (RMSE) and the Mean Absolute Error (AE), two commonly used measures in this task. Neural networks used had one hidden layer with sigmoid activation function, and an output layer with linear activation function. The training of each of the networks was performed using the Levenberg-Marquardt algorithm with a learning parameter $\mu = 0.1$, $\sigma = 10$, and the stopping criterion was a minimum gradient of 10^{-20} or 10000 iterations. The number of neurons in the hidden layer was optimized for each of the models proposed.

3.2 Input features

In Table 1 we list the inputs features to predict the phone and pause durations. We also indicate the dimension of each of the inputs for DB1 and DB2

Table 1. Input features description.

Feature	Phones	Pauses	# DB1	# DB2
Identity of two phones before and two after	x	x	30	30
Articulation type of two phones before and two after	x	x	5	5
Sound type of two phones before and two after	x	x	5	5
Part-of-speech to which it belongs [34,35]	x		15	26
Part-of-speech of words before and after		x	15	26
Does it belong to a lexical stressed syllable?	x		1	1
Does it belong to a prepause syllable?	x		1	1
Does it belong to a monosyllable word?	x		1	1
Number of syllables from the sentence beginning	x		1	1
Number of syllables from the sentence end	x		1	1
Number of syllables containing the word to which it belongs	x		1	1
Number of pauses before and after		x	2	2

When we used a single network for all phonemes, the input dimensions were 213 for DB1 and 222 for DB2. In the case of using a one network for each

[1] For the experiments we used the library MatLab Neural Network ©The MathWorks, Inc., v. 7. http://www.mathworks.com

phoneme, the input dimensions were 182 for DB1 and 192 for DB2. For pause durations prediction, the input dimensions were 192 for DB1 and 214 for DB2.

3.3 Experiments and results

Features selection As we presented in the introduction, there is a wide variety of studies that try to identify and quantify factors affecting segmental duration. Even some of them contradict each other [21]. These studies generally try to shed light on the factors affecting the duration of groups of phones, for example, separated into vowels and consonants, or by some phonetic, phonological or articulatory classification. Instead, we propose to focus on each of the phonemes [34].

An interesting work appears in [11], where they try to discover what factors affect the phone durations. An ANN as predictor and input features used by these authors are similar and/or equivalent to the setup used in present work. The experiments consisted on feeding the ANNs with different subsets of possible inputs, in order to find a group of features that provide the best performance. Paradoxically, the best result was obtained using all the factors proposed. An exhaustive search of this type has a high computational cost, having to make 2^n (where n is the number of features to be analyzed) experiments to find the optimal set. In our case, would be $2^{19} = 524,288$ experiments for each phoneme, i.e. a total of 15,728,640 tests. Not counting the repetitions to eliminate the effect of initializing the training algorithm and experiments required to find the most suitable network topology for each situation.

Another way to solve this problem is to use a predictive tool to select automatically the best features set, whether intrinsically during training process, or a posteriori using some pruning method. CARTs (Classification and Regression Trees) are well known for their ability to perform the input selection [28], but have proved to be less effective than the ANNs to predict segmental duration [29], which we could confirm in preliminary tests in our database. For example, for phoneme [i] from DB1 database, RMS errors of 11 ms versus 25 ms were obtained for ANNs and CARTs respectively.

Using an ANN, we can try any of the following way [18]:

1. Make a selection of significant features based on some analysis of dataset [2,31].
2. Use a training algorithm that eliminates the input neurons do not contribute to solve the problem [8].
3. To train the network with all inputs, and then perform a pruning of input nodes, or directly, based on the weights of the network to determine the inputs that are not useful [19]. In all cases, after these operations we must retrain the network.

At a quick glance, options number two and three are the most interesting, since everything is limited to work on the neural network. Nevertheless, these techniques have great difficulties, such as having a high computational cost.

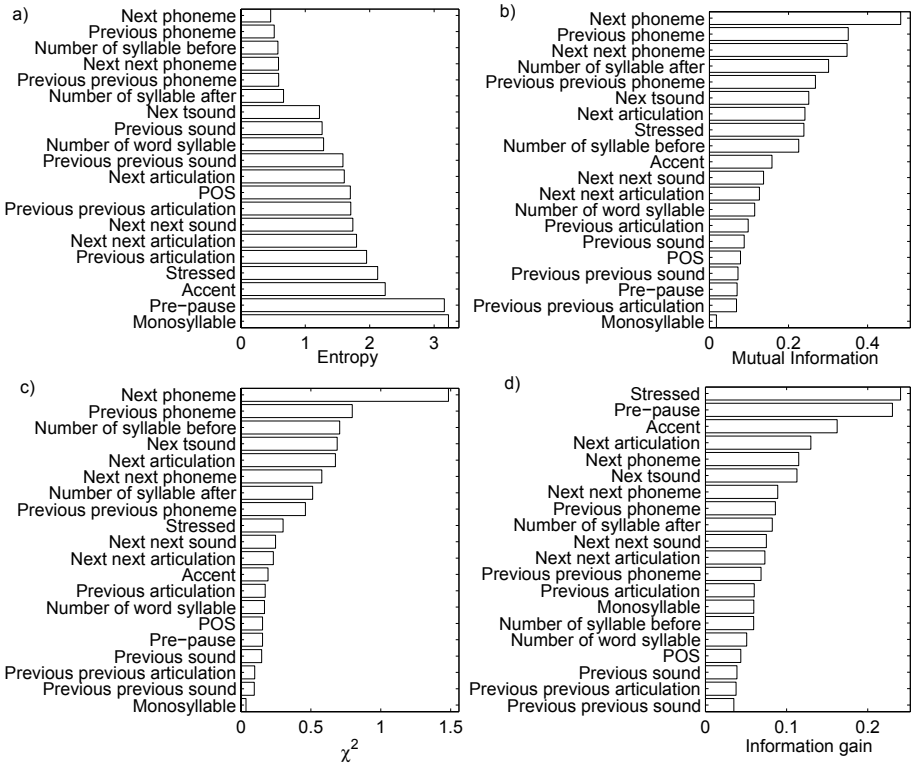

Fig. 1. Features ordering according to their importance, obtained by the methods described in a) [9], b) [2], c) [20], and d) [31], for phonene [i] extracted of DB1.

Moreover, in our case some of the features are encoded in more than one input node, and thus debunks many of the assumptions involved in the application of these techniques.

Option number one is the most practical. For a review of the proposed methods, see [4]. In different works, like [3], comparisons of different methods of sorting the entries by their importance have been presented. These works reflect one of the main disadvantages of this approach: not all evaluated methods give the same order, and their performance depends on the problem at hand. In addition, a set of input features may be optimal, according to some criteria, but fail when used in a particular forecasting system. That is, the features that best solve the problem when using CARTs, are not necessarily the same as when using neural networks. Another drawback is that these methods should infer the importance from training data set, which can be sparse or poorly conditioned [30], making it a difficult or impossible task.

For example, we implemented the methods of ordering the variables according to their importance, given by: a) [9], b) [2], c) [20], d) [31]. In Figures 1 and 2, we present the results obtained when applying these methods to phonemes [i], from

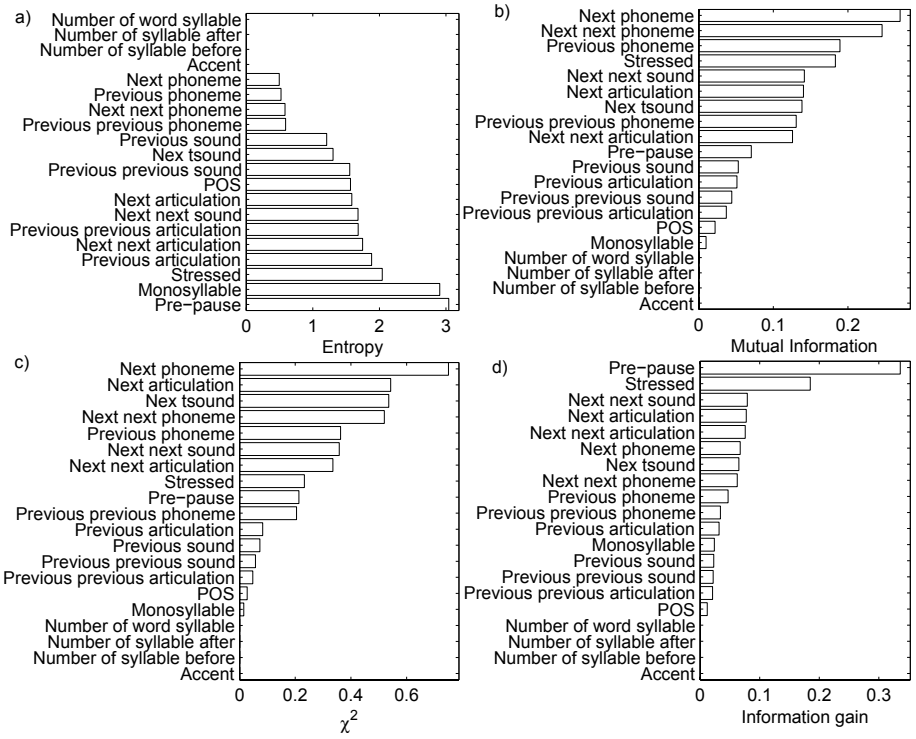

Fig. 2. Features ordering according to their importance, obtained by the methods described in a) [9], b) [2], c) [20], and d) [31], for phoneme [i] extracted of DB2.

DB1 and DB2 database, respectively. In all these graphs, the features are sorted in descending importance, from top to bottom. By comparing Fig. 1 and 2, we see that the methods do not assign the same importance to features extracted of each database. As an example, look at the feature labeled as *Pre-pause*, which indicates that the phone under analysis belongs or not to a word that is found before a pause. The work of [23] shows that it strongly influences the phone durations, but among all the methods tested, only the measure of Information Gain (Fig. 1. D and 2. D) [31] assigned it a high importance. On the other hand, in this method the size of the word is not an important feature, which contradicts previous work [17].

Segmental duration prediction In Table 2, we show the results obtained to predict the phone durations, using one network for all phonemes and one network for each phoneme, for the databases DB1 and DB2. This table also included the number of occurrences of each phoneme. We use all input features listed in Table 1. The number of neurons in the network hidden layer used to predict all phonemes was set empirically at 12 and 10, for databases DB1 and DB2 respectively, and the value for each phoneme is detailed in Table 2.

Table 2. Results of model phone durations, in milliseconds.

SAMPA	# phones		One net for all				One net for each phoneme					
			DB1		DB2		DB1			DB2		
	DB1	DB2	RMSE	AE	RMSE	AE	# Neurons	RMSE	AE	# Neurons	RMSE	AE
All	27050	65769	26	18	20	15		-	-	-	-	-
i	1284	3159	23	17	19	14	9	22	17	10	20	15
e	3418	8638	23	17	18	14	30	25	19	8	18	12
a	3580	8346	27	19	22	16	10	26	19	8	21	15
o	2395	6132	27	20	23	17	21	28	20	3	21	14
u	744	1744	22	17	20	15	8	25	14	8	21	15
j	847	1862	33	21	23	17	10	26	17	9	21	16
w	395	795	24	18	22	18	7	27	21	23	21	16
l	1776	3724	27	19	20	13	7	26	19	14	14	10
m	738	1844	27	18	19	13	10	32	22	3	15	10
n	1742	4473	31	22	25	18	14	35	25	15	15	9
N	131	258	40	26	24	18	3	35	28	27	14	9
B	372	1239	17	12	14	10	8	20	12	12	15	11
D	482	2154	20	13	17	12	7	20	14	30	16	12
G	223	579	19	14	20	15	6	20	13	19	19	14
b	305	350	23	18	19	13	27	26	19	8	18	12
d	657	803	22	14	17	12	22	15	12	4	15	11
g	101	84	21	18	10	8	8	23	18	5	27	18
r	1532	3730	16	11	15	11	14	22	15	5	9	6
R	318	462	34	24	20	15	12	33	30	5	22	16
Z	101	254	15	11	16	12	7	21	17	23	18	14
h	639	1870	19	15	19	15	21	21	16	3	9	4
p	731	1764	22	16	16	12	28	26	19	5	17	12
t	1293	3161	20	16	15	11	3	21	15	25	16	12
k	1028	2582	19	15	14	11	18	22	17	10	14	11
H	106	198	20	17	15	12	4	20	17	8	19	15
s	1505	4364	35	27	27	18	11	37	26	5	16	10
f	308	705	31	23	17	13	28	34	26	14	20	15
x	203	223	21	17	19	15	9	25	29	21	24	18
C	96	272	32	23	17	13	14	36	24	18	18	14
Min	96	84	16	11	10	8	3	15	12	3	9	4
Max	3580	8638	40	27	27	19	30	37	30	30	27	18
Mean	933	2268	25	18	19	14	13	26	20	12	18	13

For pause durations prediction we use one network with 18 and 19 hidden neurons, for DB1 and DB2 respectively. For DB1 we obtained a RMSE of 33 ms and a AE of 25 ms, and for DB2 we obtained a RMSE of 29 ms and a AE of 22 ms.

4 Discussion

Speaker use of distinctive features also contributes to have different results on duration. For example speaker from DB1 uses duration increase more often to

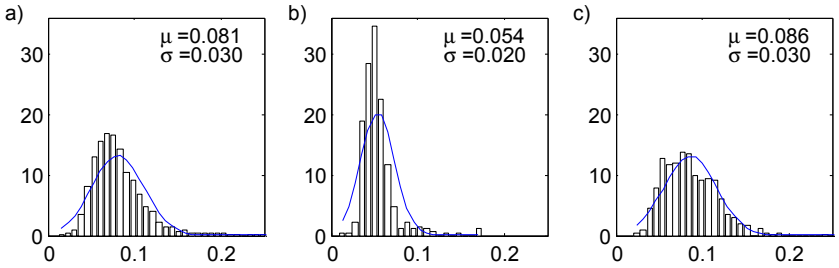

Fig. 3. Phoneme durations histogram: a) [e], b) [B], and c) [u], in seconds. The solid line represents the lognormal distribution associated.

indicate accent than speaker from DB2. This can be seen comparing the feature accent for the same method in Figures 1 and 2.

We can say that our RMSE results are close to the resolution of a human labeler, i.e. 20 ms [37]. It is difficult to compare our results with other approaches given the differences in the databases used for the experiments, as well as the proportions of the train/validation/test data used. For example, Cordoba et al. [11] obtain a RMSE of 19 ms, over a Castillan Spanish database with neutral speech, which suggests a lower degree of difficulty than the databases used here. Iriondo et al. [16] show a RMSE of 22 ms for a database of 2.5 h and a data partition of 75/25 for train/test, respectively. For American English, Webster et al. [36] show a RMSE of 23 ms for a database of 5.5 h and a data partition of 80/10/10 for train/validation/test, respectively. Besides, the only one prosodic feature that we used is pause locations, unlike other approaches where features as pitch accent position are employed [10][30].

The results of our experiments did not show a significant advantage when using one network for all phonemes vs. one network for each phoneme. Nevertheless when using one network for each phoneme some further analysis could be made.

In Table 2, we can see a large scatter in the number of neurons in the hidden layer. This can be attributed to different causes, for example, there are phonemes that have a greater variability in duration than others. Factors that influence the duration are not the same for all phonemes and their relationships may vary in complexity, and so on. For database DB1, the extremes are the number of neurons in the hidden layer for [e] (30 neurons) and [t] (3 neurons), especially if we note that the latter has better performance than the first. In Figure 3 we can see a big difference between the distribution dispersion of the durations of [e] (Fig. 3. a)) and [B] (Fig. 3. b)). This partly explains why the neural network used to predict the duration of [e] must be more complex. At the same time, this example justifies the use of an ANN by phoneme. In contrast, the distribution of [u] (Fig. 3. C)) has a greater deviation than for [e], but it only takes an ANN with 8 neurons in the hidden layer to predict its duration. This may be because the factors that influence the duration of [u] are less than those that affect [s] or that the relationships among these factors are simpler in the case of [u]. In the same

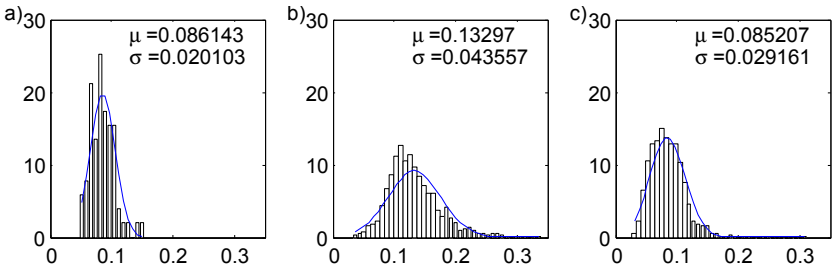

Fig. 4. Phoneme durations histogram: a) [Z], b) [s], and c) [t], in seconds. The solid line represents the lognormal distribution associated.

way, we can explain that the best performance is achieved for [Z] (Fig. 4. A)) and worst for [s] (Fig. 4. a)). The sound [s] frequently ends a sentence and some times it is well pronounced or produced in a very relaxed way or it can be absent. Additionally when we use one net for each phoneme the number of neurons and errors give us an insight of the complexity of each speaker production. See [s] in Table 2.

We obtained acceptable results in the task of pause duration predictions, and even much better than other approaches, for example, for portuguese in [32] they obtained a RMSE of 95 ms, and for Basque in [26] they obtained a RMSE of 80 ms.

5 Conclusions and Future works

A phone based duration model has been presented based on ANN's. The selected ANN is a forward feeded net trained by the Levenberg-Marquardt algorithm. Unlike other propositions training one net per phone produced similar results than training only one net for all phones. The results presented here show the high performance of the proposed model that captured factor interrelations that are considered to convey segmental duration information. Results also show that relevant input features depends upon speaker style. These results encourage its application on an Argentine Spanish Text-To-Speech system.

By using one neural network for phoneme, we can optimize the number of hidden layer neurons, which in turn gives us an idea of the degree of difficulty of the task of prediction.

Future work should further investigate the input features and their relationships to see the way they influence the duration of individual phonemes. This would require applying a priori knowledge of different areas such as physiology and anatomy of the vocal apparatus, acoustic phonetics and linguistics.

6 Acknowledgements

This research has been carried out with the support of Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina.

References

1. Barbosa, P., Bailly, G.: Characterisation of rhythmic patterns for text-to-speech synthesis. Speech Comunication 15, 127–137 (1994)
2. Battiti, R.: Using mutual information for selecting features in supervised neural net learning. IEEE Trans. on Neural Networks 5(4), 537–550 (1994)
3. Biesiada, J., Duch, W., Kachel, A., Maczka, K., Palucha, S.: Feature ranking methods based on information entropy with parzen windows. In: Proc. of the 9th International Conference on Research in Electrotechnology and Applied Informatics. pp. 109–119. Katowice, Poland (August 2005)
4. Bonnlander, B.: Nonparametric selection of input variables for connectionist learning. Phd, University of Colorado (1996)
5. Borzone, A., Signorini, A.: Segmental duration and rythm in Spanish. Journal of Phonetics 11, 117128 (1983)
6. Bouzon, C., Hirst, D.: The influence of prosodic factors on the duration of words in british english. In: Proc. of Sspeech Prosody 2002. pp. 191–194. Aix-en-Provence, France (April 2002)
7. Cambell, W., Isard, S.: Segment durations in a syllable frame. Journal of Phonetics 19, 37–47 (1991)
8. Castellano, G., Fanelli, A.M.: Variable selection using neural-network models. Neurocomputing 31, 1–13 (2000)
9. Chi, Z., Jabri, M.: An entropy based feature evaluation and selection technique. In: Proc. of 4th Australian Conference on Neural Networks. pp. 193–196. Melbourne, Australia (February 1993)
10. Córdoba, R., Vallejos, J., Montero, J., Gutierrez-Arriola, J., Lopez, M., Pardo, J.: Automatic modelling of duration in Spanish Text-To-Speech System using Neural netorks. In: Proc. of 6th European Conference on Speech Communication and Technology. vol. 4, pp. 1619–1622. Budapest, Hungary (September 1999)
11. Córdoba, R., Montero, J.M., Gutiérrez, J.M., Vallejos, J.A., Enriquez, E., Pardo, J.M.: Selection of the most significant parameters for duration modelling in a Spanish text-to-speech system using neural networks. Computer Speech & Language 16(2), 183–203 (2002)
12. Estévez, L.D.B., Castells, S.T.: La duración consonántica en castellano. Lingüística Española Actual XXI(1), 99–126 (1999)
13. Gurlekian, J.A., Colantoni, L., Torres, H.M.: El alfabeto fonético SAMPA y el diseño de córpora fonéticamente balanceados. Fonoaudiológica 47(3), 58–70 (2001)
14. Gurlekian, J.A., Rodriguez, H., Colantoni, L., Torres, H.M.: Development of a prosodic database for an argentine spanish text to speech system. In: Bird, B., Liberman (eds.) Proc. of the IRCS Workshop on Linguistic Databases. pp. 99–104. SIAM, University of Pennsylvania, Philadelphia, USA (December 2001)
15. Huber, K.: A statistical model of duration control for speech synthesis. In: Proc. of the 5th European Signal Processing Conference. pp. 1127–1130. Barcelona (Septiembre 1990)
16. Iriondo, I., Socoró, J.C., Formiga, L., Gonzalvo, X., Alías, F., Miralles, P.: Modelado y estimación de la prosodia mediante razonamiento basado en casos. In: Actas de las IV Jornadas en Tecnologia del Habla. pp. 183–188. Zaragoza (Noviembre 2006)
17. Klatt, D.: Linguistic uses of segmental duration in english: Acustical and perceptual evidence. J. Acoust. Soc. Am. 59(5), 1208–1220 (May 1976)

18. Lemaire, V., Féraud, R.: Driven forward features selection: A comparative study on neural networks. In: King, I., Wang, J., Chan, L., Wang, D.L. (eds.) Proc. of the 13th International Conference on Neural Information Processing. vol. Part II, pp. 693–702. Springer-Verlag, Hong Kong, China (October 2006)
19. Li, K., Peng, J.X.: Neural input selection—a fast model-based approach. Neurocomputing 70, 762—-69 (2007)
20. Liu, H., Setiono, R.: Chi2: feature selection and discretization of numeric attributes. In: Proc. of 7th IEEE Int. Conference on Tools with Artificial Intelligence. pp. 388–391 (1995)
21. Llisterri, J., Aguilar, L., Garrido, J., Machuca, M., de la Mota, C., Ríos, A.: Fonética y tecnologías del habla, pp. 449–479. Editorial Milenio, Barcelona (1999)
22. Macarron, A., Escalada, G., Rodriguez, M.A.: Generation of duration rules for a Spanish text-to-speech synthesizer. In: Proc. of the 2nd European Conference on Speech Communication and Technology. pp. 617–620. Genova, Italy (September 1991)
23. Marín, R.: La duración vocálica en español. Estudios de Lingüística de la Universidad de Alicante 10, 213–226 (1994)
24. Mendoza, E., Carballo, G., Cruz, A., Fresneda, M., Muñoz, J., Marrero, V.: Temporal variability in speech segments of Spanish: context and speaker related differences. Speech Communication 40, 431–447 (2003)
25. Murillo, A.A.M.: Alargamiento final en el español. Signos Lingüísticos 1, 43–59 (Enero 2005)
26. Navas, E.: Modelado prosódico del Euskera Batúa para conversión de texto a habla. Ph.D. thesis, Universidad del País Vasco (2003)
27. O'Shaughnessy, D., Barbeu, L., Bernardi, D., Archambault, D.: Diphone speech synthesis. Speech Comunication 7, 55–65 (1988)
28. Questier, F., Put, R., Coomans, D., Walczak, B., Heyden, Y.V.: The use of CART and multivariate regression trees for supervised unsupervised feature selection. Chemometrics and Intelligent Laboratory Systems 76, 45–54 (2005)
29. Riedi, M.P.: Controlling segmental suration in speech synthesis systems. Phd, Swiss Federal Institute of Technology, Zurich (1998)
30. Santen, J.V.: Assignment of segmental duration in text-to-speech synthesis. Computer Speech & Language 8, 95–128 (1994)
31. Setiono, R., Liu, H.: Improving backpropagation learning with feature selection. Applied Intellig.: The Int. Journal of Artif. Intellig., NNs,and Complex Problem-Solving Technologies 6(2), 129–139 (1996)
32. Teixeira, J.P.R.: A prosody Model to TTS Systems. Phd, Faculdade de Engenharia da Universidade do Porto (May 2004)
33. Tomás, T.N.: Manual de pronunciación española. Consejo Superior de Investigaciones Superiores, Madrid (1918)
34. Torres, H.M.: Generación automática de la prosodia para un sistema de conversión de texto a habla. Ph.D. thesis, Universidad de Buenos Aires, Buenos Aires, Argentina (Agosto 2008), (PhD. Thesis)
35. Torres, H.M.: Etiquetado de clase de palabras. Informe técnico (2010)
36. Webster, G., Buchholz, S., Latorre, J.: Automatic feature selection from a large number of features for phone duration prediction. In: Proc. of Speech Prosody. Chicago, USA (May 2010)
37. Wesenick, M.B., Kipp, A.: Estimating the quality of phonetic transcriptions and segmentations of speech signals. In: Proc. of Fourth International Conference on Spoken Language Processing (ICSLP'96). vol. 1, pp. 129–132. Philadelphia, PA (October 1996)