PARAMETER SELECTION FOR PROSODIC MODELLING IN A RESTRICTED-DOMAIN SPANISH TEXT-TO-SPEECH SYSTEM

Juan M. Montero, Ricardo de Córdoba, Javier Macías-Guarasa, Rubén San-Segundo, Juana Gutiérrez-Arriola, José M. Pardo
Speech Technology Group. Electronic Engineering Dept. Universidad Politécnica de Madrid. E.T.S.I. Telecomunicación. Ciudad Universitaria, 28040-Madrid, Spain
juancho@die.upm.es

ABSTRACT

The prosodic modeling is one of the most important tasks for developing a new text-to-speech synthesizer, especially in a female-voice high-quality restricted-domain system. Our double objective is to get accurate predictors for both the F0 curve and phoneme duration by minimizing the model estimation error in a Spanish text-to-speech system. To achieve these complementary aims we needed to find the factors that most influence prosodic values in a given language. We have used neural networks and experimented with the different combinations of parameters that yield the minimum error in the estimation. In the restricted-domain environment the variation in the different patterns is reduced, and there are more instances of each parameter vector in the database. This way, the neural network proves to be an excellent tool for the modeling.

The resulting system predicts prosody with very good results (for duration: 15.5 ms in RMS and a correlation factor of 0.8975; for F0: 19.80 Hz in RMS and a relative RMS error of 0.43) that clearly improves our previous rule-based system.

KEYWORDS: Prosody, F0 modeling, duration modeling, text-to-speech, artificial neural networks, parameter selection, parameter coding

1. INTRODUCTION

The primary goal of this study was to develop an automatic system to model prosody for a Spanish text-to-speech system (TTS) in a restricted-domain environment for a female voice. This work is the continuation of [1] and [2] which were dedicated to a general-domain database for a male voice and [3], that included the first version of the restricted-domain modeling, achieving better results than our original rule-based system. For modeling duration this rule-based approach follows a classic multiplicative Klatt model; for F0, it models the curve in a parametric way as a series of text-dependent F0 peaks and F0 valleys [4].

Although a domain-specific application does not require as many sentence structures as a general one (the delivered messages are syntactically constraint), there can be many words embedded in them (e.g., more than 40,000 family names, more than 30,000 village names, etc.). A message is typically a sentence with two different parts: one of them, that is fixed, is a template for the other, which is composed of one or more slots (Variable Fields) containing the relevant information that the user is looking for in the message. Current prosodic patterns are judged as too monotonous to allow a great diversity of services, but in restricted-domain applications and by mixing female natural speech and diphone-concatenation synthesis (from the same speaker), we can provide high quality services if we mimic the natural prosody exhibited by the speaker.

Many studies have been successfully carried out lately in the field of automatic estimation of the prosodic values, using different techniques and input parameters to obtain the model. For duration, 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. For F0 modeling, the dominant techniques are artificial neural networks and k-nearest-neighbour, combined with a parametric model of the F0 curve [5].

In all the systems, regardless of the modeling technique, it is crucial to find the parameters (or features) that are most significant for prosodic modeling. So, we can take advantage of previous studies dedicated to prosody prediction, but using other techniques to decide the parameter set. Neural networks have previously been used with success. In [6] a neural network was trained to predict syllable timing. In [7] 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. Regarding the application of these techniques to Spanish, there are very little references and none is dedicated to neural networks or CART approaches. We have considered some of them but only to decide the parameters to be used as input. See in [1] a summary of references for Spanish.

2. DATABASE USED FOR THE MODELING

The database used in this paper is described in [3]. We extracted a set of 19 Carrier Sentences (CS) from two real services in banking and traffic information domains, provided by the IVR design company. The CS contained 24 Variable Fields (VF) and each VF conveys the most important information in the CS and must be surrounded by compulsory pauses. Prosodic values are only computed for the VFs. We classified the CS into 3 classes or groups:

- *Proper Names*: surnames (both compound and simple ones), cities, villages, etc
- *Questions*: bank-related information such as currency, check status, etc.
- *Noun Phrases*: regarding accounts, credit cards, names of transactions and banks...

For the design of the database we used a greedy algorithm that is described in [3]. We aimed at selecting a small database with the same probability distributions of certain phonetic and prosodic features as in a very big database (about 6600 phonemes and 2800 syllables per class)

3. NEURAL NETWORK SYSTEM DEVELOPMENT METHODOLOGY

3.1. Topology of the neural network

For both duration and F0 modeling, we have used a multilayer perceptron (MLP), using the sigmoid as the activation function and the backpropagation algorithm for training. For each phoneme (or syllable), we compute a series of parameters (features), which we code and use their values as inputs to the neural network. There is one output in our networks: the duration of the phoneme (or the F0 of the syllable). For duration experiments we used 2 sets (training and testing) and we divided the training into three phases of 300 iterations each (for over-fitting detection). For F0, we used a ten-fold cross-validation strategy with 3 non-overlapping sets (one for training, one for over-fitting detection and one for the final evaluation). As it is very difficult to know the optimum number of neurons and layers that the net should have, a set of experiments were carried out in order to optimize the system without overtraining.

In this restricted-domain system we had the option to use a single network for the 3 classes of sentences or 3 different networks for each class. Using the best configuration of parameters of [1] we compared both approaches. The 3-networks option improved the results in 6% for duration so we decided to use 3 different networks in our duration experiments.

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.

1. **Binary coding**: this is the standard coding for true/false parameters.
2. **One-of-N coding:** to code N classes, we use N neurons and only 1 of them is active.

In **ordinal** values we have more possibilities, as these values can be ordered:

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.

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. We developed an algorithm to obtain a uniform distribution of all the classes.
5. **Z-Score mapping**: we normalize the floating-point value by accounting for the average and the standard deviation of the variable (a good coding for very variable parameters).

3.3. Network evaluation

To evaluate the error of the networks (difference between the prediction from the network and the optimum value), we have considered different metrics. The most important one is the RMSE (sqrt[MSE]). Another one is the *relative RMSE* (RMSE / sqrt [Σ [t-t_i]^2]), that it is adimensional and independent of the way we code the target values (t_i), and it does not have an offset.

3.4. Modeling the output

We obtained in [1] that phoneme durations should be normalized by the duration of the phrase (to be less affected by changes of speed in the database recordings). After the normalization, we use the standard deviation of the logarithm of the duration (to balance the distribution of the values and to minimize the error, as it includes the characteristic duration of each phoneme in the prediction) and a Z-Score codification. For F0, we just used Z-Score.

4. PARAMETERS TO BE USED

4.1. Base experiment for duration

In our base experiment for duration (first row of Table 1) we have decided to include just the phoneme identity (with a set of 38 phonemes and a windowing of three values, described in next section), and the stress, which are the most relevant parameters according to our previous work and to our own statistical studies. 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.

In Table 1 we can see the relative RMSE and the average improvement obtained for the test set with individual parameters, using a 10-neurons network. The last column shows the results of applying a T-Student test to compare the base experiment and the experiment considered (when "2-tail-sig" is below 0.05 the difference between both systems is statistically significant).

4.2. Contextual phonemes

In our previous studies, the duration of a phoneme was significantly affected by the phoneme to the right and to the left. As the number of phonemes is too high, we made 14 clusters of phonemes according to its type. Using a two-phonemes context (a window of five values) we obtained an improvement of 5% for the test set (Table 1, experiment 1). This result is really remarkable, as it shows the importance of contextual information. But for a 7-values window the results were slightly worse.

4.3. Parameters related to position and binary parameters

In [1] we found that "Position in phrase in relation to first/last stress" was a especially relevant parameter, as it explicitly includes the "lengthening before pause" effect. We coded each syllable in 5 possible classes with very good results (Table 1, experiment 3).

We have also obtained new significant improvements over the base experiment by considering several binary parameters (experiments 4-6 in Table 1):

- Syllable structure: syllables ending with a vowel (open syllables) are generally longer.
- Vowel in 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.
- Phoneme in a function word. Syllables in a function word are shorter.

In [1] we considered different alternatives for parameters related to position and decided to use: phoneme in the syllable, syllable in the word, and word in the phrase, as they provide

different information to the network (not redundant), their range of values is smaller, and, so, fewer neurons and classes are needed. We carry out the following steps for the coding:

1. To normalize the value of position by the total length of the higher-order unit
2. This value is coded using 3 classes, and their intervals are computed automatically.
3. The 3 classes use a thermometer-type coding with 2 inputs (number of classes minus 1).

The results of these experiments (7 to 9 in Table 1) have improved the base experiment again. The best parameter is 'position of the word in the phrase', one conclusion that we did not obtain in the unrestricted-domain system, where all parameters related to phrase were worse. The reason is that the range of values is much more uniform in this restricted-domain system.

4.4. Parameters related to the "Number of units"

In a similar way as for parameters related to position, we decided to use the number of phonemes in the syllable, the number of syllables in the word, and the number of words in the phrase. Because of their different distribution, we needed a different coding:

1. To normalize the value by the maximum one: a floating point value between 0 and 1.
2. To apply Z-score (using average and standard deviation): this way, we can restrict at our will the operating range of the parameter, which is too variable.

The improvements (experiments 10-12 of Table 1) were significant and very similar to those of position parameters (the number of words in the phrase is the best parameter). In order to check the suitability of this floating point coding, we tested the thermometer-type coding (as for position-related parameters), but the results were always below.

4.5. Summary of results for duration

The summary in Table 1 correspond to the best network (10 neurons). We have obtained the best results for: window of 5 phonemes, number of words in the phrase, position of the word in the phrase and position in phrase in relation to first/last stress.(Stress is important too, but it is included in the base experiment); almost all the improvements are significant (not as in [1]).

Experiment	Test set	Improvement	2-tail-sig
Base experiment	0.5580	-	-
1- Base + window of 5 phonemes	0.5318	4.98 %	0.000
2- Base + window of 7 phonemes	0.5350	4.81 %	0.000
3- Base + position in phrase	0.5450	2.48 %	0.001
4- Base + vowel in diphthong	0.5515	1.53 %	0.045
5- Base + syllable structure	0.5462	2.43 %	0.001
6- Base + function word	0.5451	2.35 %	0.000
7- Base + position of Phoneme in Sentence	0.5523	1.03 %	0.419
8- Base + position of Sentence in Word	0.5462	2.29 %	0.006
9- Base + position of Word in Phrase	0.5427	2.49 %	0.001
10- Base + number of Phoneme in Sentence	0.5494	2.07 %	0.010
11- Base + number of Sentence in Word	0.5501	2.20 %	0.048
12- Base + number of Word in Phrase	0.5403	3.43 %	0.000

Table 1. Summary of results in average relative RMS (for duration)

4.6. Final experiments for duration

The next set of experiments was dedicated to including all the parameters together. This is the crucial step in neural networks, because many times the improvements combining parameters are not additive, because the parameters are closely correlated (do not offer additional information), or the topology of the network needs to be tuned (a larger number of neurons may be needed).

In Table 2 we can see the summary of results. The numbers in the description of the experiments refer to the experiments specified in Table 1. The T-Student test is now applied to the comparison of an experiment with the previous one.

- Experiment 13: it is the base experiment using now a window of 5 phonemes and position in phrase in relation to first/last stress. The improvement was remarkable.
- Experiment 14: we added the binary parameters: vowel in diphthong, syllable structure and function word. The improvement is reduced and not significant
- Experiment 15: with position parameters. The improvement is significant.
- Experiment 16: including the 'no. of units' parameters with significant improvements.

The results are really good, and the system keeps improving for both the train and the test set as we increase the number of parameters, which shows the correct learning of the networks.

Experiment	Test	Improvement	2-tail-sig
Base experiment	0.5580	-	-
13- Base + 1 + 3	0.5214	**6.58 %**	0.000
14- 13 + 4 + 5 + 6	0.5206	6.83 %	0.512
15- 14 + 7 + 8 + 9	0.5121	**8.09 %**	0.039
16- 15 + 10 + 11 + 12	0.4927	**11.12 %**	0.002

Table 2. Results for duration including all parameters.

In the unrestricted-domain system [1], there were symptoms of overtraining with very few neurons, which impeded the improvement of the global system. In this system, the best results correspond to the topology with 20 neurons. The improvement over the base experiment is 18.71%, which shows that our solutions improved this system drastically. The relative RMS is 0.4536, the average absolute error is 11.79 ms, and the absolute RMS is 15.5 ms. The Pearson correlation coefficient between estimated and measured durations is 0.8975, a very good figure.

4.7. Comparison to previous systems

As could be anticipated, the results are much better than those obtained with the unrestricted-domain database: an absolute RMS equal to 19.1 ms. The relative RMS was equal to 0.76428, clearly worse than the 0.4536 obtained in this domain.

Using our previous multiplicative rule-based system, with the best parameter coding of the ANN experiments, the absolute error was 19.8 ms and the RMS was 28.5 ms, which is clearly worse than the result obtained with our neural network.

4.8. F0 experiments

For F0, we performed similar experiments with a different set of parameters. Our previous rule-based system used features such as: whether the syllable is stressed, whether the following syllable is stressed, the type of punctuation mark at the end of the intonation group (this parameter is related to the shape of the F0 curve at the end of the group) and the number of stressed syllables and the position of the syllable in the group. The F0-curve obtained this way is acceptable but unnatural in human perception tests [2].

In addition to these general parameters, we tried several ways of coding the influence of the carrier sentences from the restricted-domain. The best results obtained correspond to a one-of-N coding of the carrier sentences (we grouped sentences according to 3 classes as defined in section 2; with only a 1% improvement, that is not significant). No significant improvement was obtained through parameters related to position, to function words or to the number of units.

The summary in Table 3 correspond to the best network (20 neurons). We have obtained the best results for: a one-of-N coding for the carrier sentence and the final punctuation mark, a window of 11 syllables for stress and for the position of the syllable in the phrase (in relation to

first and last stressed syllable). All the improvements are significant when compared to the previous one except for experiments 5 and 6.

F0 Experiment	Test	Improvement
Base experiment: stress	0.7378	-
1- stress in a 3-syllables window	0.6815	7.63 %
2- stress in a 11-syllables window	0.6326	14.26 %
3- 2 +final punctuation mark	0.5500	25.45 %
4- 3 + identifier of the carrier sentence	0.4554	38.28 %
5- 4 + position of the syllable in the group	0.4360	40.91 %
6- 5 + 3-neural-networks option	0.4312	41.56 %

Table 3. Results in average relative RMSE

5. CONCLUSIONS

Compared to our previous rule-based systems, the results are much better, even when using a limited number of parameters. As we expected, the results obtained in the restricted-prosody domain show improvements that are much more significant than in [1] (because the database is more homogeneous) and than in [3] (due to a better parameter selection). For a new female voice, we have demonstrated that our prosodic model can be easily adapted to specific contexts and/or new databases in a very short time. For duration another important aspect is that the results improve when we include all the parameters and increase the number of neurons, a tendency we did not observe in the unrestricted-domain system.

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 a second hidden layer is not necessary. The "Z-score" normalization for numeric parameters shows a good behavior: it adjusts the margin of accepted values automatically rejecting the out-of-range values.

In general, we can say that we have found a good compromise between network topology and parameters considered, with good results that are stable. The system has been included in a commercial high quality TTS system in Spanish [3] (http://www-gth.die.upm.es/index.html)

6. REFERENCES

1. Córdoba, R., J.M. Montero, J. Gutiérrez-Arriola, J.A. Vallejo, E. Enríquez, J.M. Pardo (2002). Selection of the most significant parameters for duration modeling in a Spanish text-to-speech system using neural networks. *Computer Speech & Language*, Vol 16 Nº 2, pp. 183-203.
2. Vallejo, J.A. (1998). *Mejora de la frecuencia fundamental en la conversión de texto a voz.* PhD Thesis Universidad Politécnica de Madrid.
3. Montero, J.M., Córdoba, R., Vallejo, J.A., Gutiérrez-Arriola, J., Enríquez, E., Pardo, J.M. (2000). Restricted-Domain Female-Voice Synthesis in Spanish: from Database Design to ANN Prosodic Modeling. *Proceedings of ICSLP,*- pp. 621-624.
4. Allen, J., Hunnicut, S. & Klatt, D.H. (1987). *From Text to Speech: The MITalk System*. Cambridge University Press, Cambridge.
5. Tournemire, S. (1997). Identification and automatic generation of prosodic contours for a text-to-speech synthesis system in French. *Proceedings of Eurospeech*, pp. 191-194.
6. Campbell, W.N. (1992). Syllable-based segmental duration. In Bailly, G., Benoit, C., and Sawallis, T.R. (Eds.) *Talking machines: theories, models and designs* (pp. 211-224). Elsevier.
7. Fackrell, J.W., Vereecken, H., Martens, J.P., Van Coile, B. (1999). Multilingual prosody modeling using cascades of regression trees and neural networks. *Proceedings of Eurospeech*, pp. 1835-1838.