A model for the f0 reset in corpus-based intonation approaches

Francisco Campillo[*], *Jan van Santen*[†], *Eduardo R. Banga*[*]

[*] Signal Theory Group
Dpto. Teoría de la Señal y Comunicaciones
University of Vigo, SPAIN
[†] Center for Spoken Language Understanding
OGI School of Science & Engineering
20000 NW Walker Road, Beaverton, OR 97006, USA.
{campillo, erbanga}@gts.tsc.uvigo.es, {vansanten}@cslu.ogi.edu

Abstract

Concatenative intonation systems model the intonation contours as the concatenation of small natural units extracted from suitable contexts. The special characteristics of this type of models make it difficult to include some factors whose effect overcomes the phonic group domain. In this paper one of those factors, the f0 reset related to the sentence internal pauses, is addressed.

Index Terms: intonation unit selection, speech synthesis.

1. Introduction

Intonation is widely acknowledged as the most important factor for the perception of prosody, so much that not only the naturalness but the intelligibility of a speech synthesizer is highly dependent on the design of an appropriate intonation module. In the literature there are many different approaches to the problem of intonation, from the ones that model the phenomenon as a sequence of tones interrelated by a grammar ([1]), to the superpositional models that consider the frequency contour as the result of the addition of several components with different temporal scopes ([2]).

In the last years corpus–based intonation models have been proposed ([3], [4]). Under the same principles of the corpus–based acoustic unit selection ([5]), the global frequency contour is generated by the concatenation of smaller natural units selected from a suitable context. The intonation contours generated by these methods can be almost indistinguishable from the natural ones, but their performance depends heavily on the proper characterization of the context of the basic unit for concatenation. In this paper we will study an aspect that is not paid too much attention in these models, the f0 reset after a sentence internal pause, crucial for preserving the tonal coherence among the phonic groups at the sentence level.

The outline is as follows. Section 2 describes the main characteristics of the corpus employed in this research. Section 3 begins with a study of the influence of several factors on the f0 values at the beginning and the end of the phonic group, in order to justify the need for a model of f0 reset in concatenative intonation methods. In Section 4 two approaches for this model are considered, linear regression and neural networks, with the results presented in Section 5. Finally, Section 6 is dedicated to the conclusions and suggestions for future research.

2. Corpus description

In this study a corpus recorded by two male speakers was used. It consists of about 1300 isolated sentences, from which 800 were manually designed by an expert linguist to be a rich prosodic sample of the Galician language, and the remaining 500 were automatically collected in order to include more complex prosodic structures. The fundamental frequency contours were extracted using Praat (www.praat.org), with a further processing for removing those values corresponding to voiceless regions according to the underlying phoneme identity.

The prosodic corpus information was organized into accent groups, defined as a sequence of non accented words ending in an accented word, and phonic groups, defined as a sequence of words between pauses. The accent groups were clustered according to previous linguistic knowledge ([6]) into 48 different classes, taking into account the type of proposition (declaratives, interrogatives, exclamatories and ellipsis), the position within the phonic group (initial, final, medial, and initial and final) and the position of the accented syllable within the accent group (last, penultimate and antepenultimate).

Table 1 shows the distribution of the accent groups in both corpora, according to the type of proposition ($ProType$) and the position within the phonic group ($Position$). The average number of accent groups per phonic group is 3.05 for Speaker 1 and 2.87 for Speaker 2.

	$ProType$				$Position$			
	Dec	Int	Exc	Ell	In	Med	Fin	In&Fin
Speaker 1	4853	1044	342	221	1596	2693	1713	458
Speaker 2	4801	1067	330	168	1608	2455	1809	494

Table 1: Accent group distribution

3. The f0 reset in concatenative intonation models

A thorough evaluation of the performance of the system presented in [7] showed a general good quality, but also some sporadic flaws in sentences with several phonic groups. In those cases, although the intonation of every phonic group was natural, when concatenated the overall impression was bad, as a result of a lack of tonal coherence across consecutive phonic groups.

In [8] the intonation differences between consecutive terminal sentences and the same sentences presented as coordinated

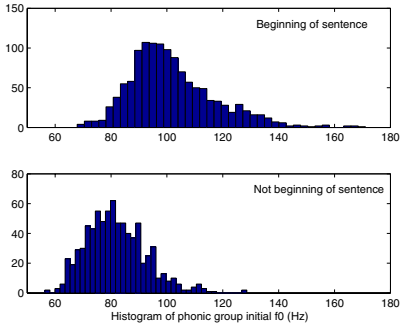

Figure 1: *Histograms of the phonic group initial f0. Beginning of sentence (top), and not beginning of sentence (bottom)*

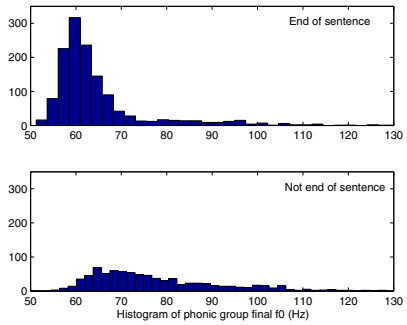

Figure 2: *Histograms of the phonic group final f0. End of sentence (top), and not end of sentence (bottom)*

clauses were studied, finding that the intonation contours were steeper and presented greater amounts of resetting in the former ones. The intonation model in [7] selects the best sequence of accent groups from the prosodic corpus taking into account only the available candidates in each cluster according to the classification shown in Section 2, this is, initial accent groups for the first accent group of a target phonic group, and so on, which implies that the internal boundaries are treated in a very similar manner than the beginning and the end of the sentence. Although the position of the phonic group within the sentence and the phonic group type of final pause are considered during selection, it does not seem to be enough to obtain always a f0 reset after the internal boundaries of the sentence that respects the tonal coherence of the natural intonation.

In order to solve this problem, a preliminary study of the variation of the phonic group initial and final f0 values was carried. All the results presented in this section were obtained for Speaker 1, but it is important to note that exactly the same tendencies existed in the other Speaker's data. Figure 1 shows two histograms of the phonic group initial f0, being it sentence initial (top), and not sentence initial (bottom). Both distributions are fairly different, being their average values 101.96 Hz (top) and 81.88 Hz (bottom). A two-tailed t-test was performed, finding that the differences were highly statistically significant ($p - value < 0.001$). A parallel test was carried excluding the interrogative sentences, usually with much higher initial f0 values at the beginning of the sentence, but the results were still highly statistically significant, with an average f0 value at the beginning of sentence of 99.41 Hz. With respect to the final f0 value, Figure 2 shows two similar histograms taking into account whether the phonic group is at the end of the sentence (top) or not (bottom). In this case the differences, highly statistically significant as well, are even more obvious than before, being the average values 64.71 and 77.28 Hz respectively.

These results agree with [9], and confirm the need to introduce a model of the f0 reset after internal pauses within the sentence. With this problem in mind, a study of the influence of several features into the phonic group initial and final f0 values was conducted. Figure 3 shows the initial and final f0 values as a function of the phonic group position. As it can be seen, the non initial phonic groups have an average initial value close to 80 Hz, and their differences were found to be not statistically significant. In the special case of the first phonic group, the values from sentences with only one phonic group and more than one were clustered, as a separate test showed no statistically significant differences between both cases ($p - value = 1$).

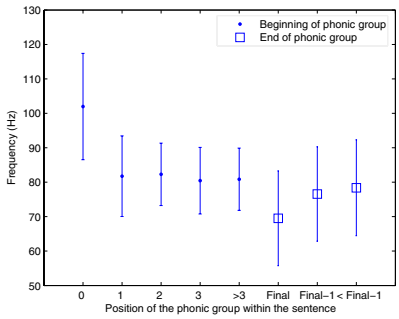

Figure 3: *Initial and final f0 values as a function of the phonic group position within the sentence*

With respect to the final f0, the difference is also highly statistically significant between the last phonic group and the others, but not in the remaining cases. The influence of the position of the phonic group within the sentence with respect to the initial and final f0 values was also reported in [8] and [10].

The influence of the number of accent groups within the phonic group with respect to its initial and final f0 values (excluding sentence initial and final boundaries) was also studied, finding that the initial values were in all cases around 80 Hz, without statistical differences, as well as the final values, slightly lower. These results agree with [10], although there the length of the phonic group was measured in number of syllables. Other features were studied, such as the effect of the number of phonic groups within the sentence, with very small differences in the average values (around 3 Hz in both cases), but with a great variance in the final f0 data.

These results show the variability of the initial and final f0 values according to several features. In the context of a concatenative intonation model, this means that no every phonic group initial accent group is equally suitable after an internal pause. Therefore, as mentioned before, it is necessary a model of f0 reset taking into account the variability at both sides of the internal boundary. Next section will be dedicated to this task.

4. Models of f0 reset

Table 2 displays the input parameters that were taken into account in the different tests. The type of proposition ($ProType$) included the same classification as mentioned in Section 2, as well as the accent parameters ($Ac1$, $Ac2$). The duration of the

pause (Dur) was measured in seconds, and the initial and final f0 values ($Freq_i$, $Freq_{i-1}$), in Hertz.

Table 2: *Parameters considered in the models*

Parameter	Meaning
$ProType$	Type of proposition
$PgPos$	Index of the phonic group within the sentence
$PgNum$	Number of phonic groups within the sentence
$Ag1$	Number of accent groups (preceding phonic group)
$Ag2$	Number of accent groups (next phonic group)
$Ac1$	Last accent position (preceding phonic group)
$Ac2$	First accent position (next phonic group)
Dur	Duration of the pause
$Freq_{i-1}$	F0 value (end of preceding phonic group)
$Freq_i$	F0 value (beginning of next phonic group)

The f0 at the beginning of the next group ($Freq_i$) was modeled both in the logarithmic and natural domains, but it seemed to behave better in the latter one, so from now on we will only consider this one. The following sections are dedicated to the two types of models studied in this work: linear regression and neural networks. A rule based method was also considered at first, but it was discarded as it would probably need more linguistic information. The proposed methods have the advantages of allowing a completely automatic, language independent training and avoiding the need for expert knowledge.

4.1. Linear regression model

A stepwise linear regression model was trained for the f0 reset after an internal pause. Taking a set of features as input, the stepwise regression chooses a suitable combination of them for explaining the behavior of the dependent variable. In this case, we used a forward stepwise regression, beginning with an empty subset of input features, and adding incrementally the most statistically significant feature in each step, until no statistically significant terms remained. The minimum p-value for an input parameter to be removed from the model was set to 0.10 (being the null hypothesis that coefficient to be zero).

Equation (1) shows the model resulting from this training. Although there were not significant differences in the results, the inputs were normalized ($z - scores$) so that the model coefficients can give a hint about the importance of each term.

$$\begin{aligned} Freq_i = \quad & 94.21 - 0.73 \times PgNum + 0.92 \times PgPos \\ & +0.67 \times Ag1 + 1.05 \times Ag2 - 0.51 \times Ac2 \\ & +1.32 \times Dur + 13.06 \times Freq_{i-1} \end{aligned} \tag{1}$$

Two factors were excluded from the model. The first one, the type of proposition ($ProType$), is probably excluded as a consequence of the considered classification, perhaps too naive for this phenomenon. A finer classification taking into account cases such as enumerations or explanations ("The boy, *who came the other day*, ...") would probably improve the model. With respect to the second one, the position of the preceding phonic group last accent ($Ac1$), it does not seem to add important information to the last frequency parameter ($Freq_{i-1}$). On the other hand, the most important feature is $Freq_{i-1}$, to which most of the explained variance can be ascribed. With regards to the rest of the parameters, the most relevant ones seem to be the pause duration (Dur), the number of accent groups in the second phonic group ($Ac2$) and the position of the phonic group within the sentence ($PgPos$), in this order.

4.2. Neural network model

Training a neural network is a much harder task than training a linear model, as there are many decisions that can affect drastically the final performance of the model, such as the number of hidden layers or the number of neurons in each layer. With regards to the input layer, the addition of new parameters can even lead to worse results if these are found to be noisy, as the network does not make any assumption on the nature of the dependent variable, and simply tries to find the best combination of weights for explaining it as a function of the input.

In this case, we decided to use a multi–layer perceptron with a hidden layer of nodes, as it is known to be a universal approximator ([11]). After a preliminary set of tests where several configurations were considered, a network with one hidden layer of 20 neurons fully connected to the input nodes was found to produce the better results. The algorithm for adjusting the weights was backpropagation ([11]), with the considered set of parameters as input (see Table 2) and the next phonic group initial f0 ($Freq_i$) as output. The algorithm for finding the best combination of weights was cross-validation, dividing the data into three different sets: training, validation and test. Moreover, as the performance of the network may be very dependent on the distribution of the data along these sets, the training was repeated n (n = 30) times with different seeds for their random generation. This way, the performance of a certain configuration was considered to be the average value of the root mean square error ($RMSE$) and explained variance (R^2) over the test data after the n repetitions of the training.

Finally, for finding the best subset of input parameters, every possible combination was considered. So, with 9 input parameters (see Table 2), 511 different models were trained according to the steps described in the preceding paragraph. The whole process was computationally expensive, less than three days in an Athlon XP 3 GHz, but feasible as it was completely automatic.

After the experiment there was a set of combinations yielding quite similar results, but the best one included only four input parameters: the position of the phonic group ($PgPos$), the number of accent groups in the second phonic group ($Ag2$), the duration of the pause (Dur) and the final frequency before the pause ($Freq_{i-1}$), being this last parameter the most important one in the model. Although in [10] it was found that there was no relationship between the length of the phonic group and its initial and final f0 values, in this case we are referring to the f0 reset after a pause, and it seems to be important both in the neural network and linear regression models, where it has the third bigger coefficient (see Equation (1)). In the neural network model it is even the second most important parameter.

5. Results

Table 3 shows the root mean square error ($RMSE$), the absolute error average ($|Mean|$), the variance (Var) and the explained variance (R^2) of both models. The neural network with the same parameters of the linear regression model had a similar performance, but the variance of the error was larger after the n different trainings (see Section 4.2). Finally, the neural network one was preferred, given its smaller error variance over the test data, as shown in Figure 4. Obviously, a linear regres-

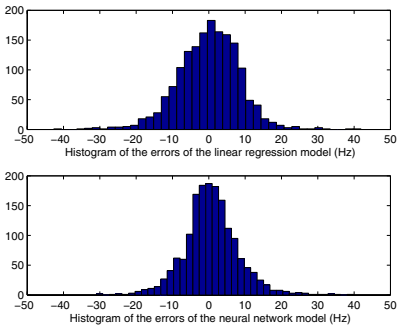

Figure 4: *Histograms of the errors in the linear regression (top) and neural network (bottom) models*

sion method can not capture some non linear effects such as the initial versus non initial distinction in Figure 3.

	$RMSE$	$\|Mean\|$	Var	R^2
Linear regression	8.59	6.57	73.63	0.70
Neural network	7.92	5.62	57.46	0.74

Table 3: Models performance (units are Hz)

The neural network model was applied to our bilingual Galician and Spanish synthesizer Cotovía ([7], [12]), with very encouraging results after the first informal listening tests, although a more thorough evaluation is needed. The model was included into the intonation concatenation cost function, and it was implemented as the difference between the initial f0 value of the next accent group and the desired value generated by the model. Finally, as more variability can be allowed across pauses, a threshold around the desired value was enabled, such that the accent groups with a close enough initial f0 are given a null concatenation cost. The informal tests showed a threshold of 10 Hz as a reasonable choice.

6. Conclusions

In this paper a novel method that takes into account the f0 reset of the intonation contour across pauses within a sentence was introduced. First, a study of the intonation contours behavior as a function of several features like the position of the phonic group within the sentence or the number of accent groups was presented, showing some evidence of their influence into the phonic groups initial and final f0 values. After that, two different approaches based on linear regression and neural networks were proposed for modeling the fundamental frequency difference at both sides of the pause.

The obtained model was integrated into a unit selection speech synthesizer. Although it is obvious that a closer evaluation is needed, the preliminary tests are promising. From the authors' point of view, the intonation modeling has reached a state such that those aspects of the natural intonation that are not currently considered are clearly audible and affect the quality of the synthetic contours. In this case, the model here introduced avoids not only the excessively large excursions of the intonation contour after a pause, but the extreme continuity that can also be unnatural depending on the context. As a future line, it would be interesting to include a finer linguistic classification for the type of sentence, as it would likely improve the performance of the model considerably. Finally, although the model was designed in the context of a concatenative intonation model, it is important to note that this problem exists in every intonation model, and the method proposed here could be easily applied to other approaches such as the superpositional models.

7. Acknowledgments

The work reported here was carried out while a visiting research post doc at Center for Spoken Language Understanding of the first author, with funds from Dirección Xeral de Investigación, Desenvolvemento e Innovación, Consellería de Innovación e Industria, Xunta de Galicia, support from NSF grant 0205731, "ITR: Prosody Generation for Child Oriented Speech Synthesis" to Jan van Santen, and Xunta de Galicia under the project PGIDTIT05TIC322202PR.

8. References

[1] Pierrehumbert, J. B., "The phonology and phonetics of English intonation". PhD thesis, MIT, 1980.

[2] van Santen, J., and Möbius, B., "A quantitative model of f0 generation and alignment". In Botinis, editor, Intonation Analysis, Modelling and Technology, chapter 12, p269–288, Kluwer Academic Publishers, Netherlands, 1999.

[3] Malfrere, F., Dutoit, T. and Mertens, P., "Automatic prosody generation using supra–segmental unit selection". In 3rd ESCA Workshop on Speech Synthesis, Jenolan Caves, Australia, p323–328, 1998.

[4] Campillo, F. and R. Banga, E., "Combined prosody and unit selection for Corpus-based text-to-speech systems". In Proceedings of ICSLP'02, p141–144, Denver, 2002.

[5] Hunt, A. and Black, A., "Unit selection in a concatenative speech synthesis system using a large speech database". In ICASSP'96, p373–376, Philadelphia, 1996.

[6] López, E., "Estudio de técnicas de procesado lingüístico y acústico para sistemas de conversión texto voz en Español basados en concatenación de unidades". PhD thesis, E.T.S.I. de Telecomunicaciones, Universidad Politécnica de Madrid, España, 1993.

[7] Campillo, F. and R. Banga, E., "A method for combining intonation modelling and speech unit selection in corpus-based speech synthesis systems". Accepted for publication in Speech Communication, 2006.

[8] Gronum-Thorsen, N., "Intonation and text in standard Danish". Journal of the Acoustical Society of America, 77, 3, 1205–1216, 1985.

[9] Gronum-Thorsen, N., "Sentence intonation in textual context–Supplementary data". Journal of the Acoustical Society of America, 80, 4, 1041–1047, 1986.

[10] Garrido, J.M., "Spanish intonation for text-to-speech applications". Ph.D. Thesis, Factultad de Lletres, Universitat de Barcelona, España, 1996.

[11] Haykin, S., "Neural networks. A comprehensive foundation", MacMillan, 1994.

[12] Campillo, F. and Banga, E. R., "On the design of the cost functions for a unit selection speech synthesis". In Proceedings of EUROSPEECH'03, p289–292, Geneva, 2003.