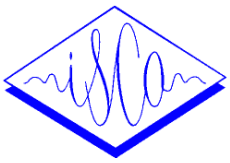

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A HIERARCHICAL INTONATION MODEL FOR SYNTHESISING F0 CONTOURS IN GALICIAN LANGUAGE

Xavier Fernández-Salgado and Eduardo R. Banga
Dpto. Tecnoloxías das Comunicacións. ETSE Telecomunicación.
Campus Universitario. Universidade de Vigo. E-36200. Vigo. SPAIN
xsalgado@tsc.uvigo.es, erbanga@tsc.uvigo.es

ABSTRACT

In this contribution we propose a hierarchical intonation model for synthesising f0 contours with application to text-to-speech synthesis in Galician language. This model makes use of the implicit knowledge that resides in a database of natural f0 contours obtained from a read corpus. The novelty of this method lies on the way the f0 contour is generated. First, no phonological description in terms of a sequence of tones is needed prior to f0 generation. The phrasing obtained from previous stages of the TTS system is enough for this task. Second, the final f0 contour is built through several steps that assign patterns at the phonic group level (intonational phrase), the tonic group level and the segmental level following a hierarchical method. The proposed algorithm guarantees a coherent concatenation of the patterns that belong to different levels, and it seems to work properly as a general intonation model for a wide range of sentence modalities.

Keywords: prosodic model, f0 contour, data-driven, text-to-speech.

1. INTRODUCTION

During the past few years we have been working on our TTS system for the Galician language. It is a Romance language spoken by three million people in the northwest of Spain, similar in a certain degree to its neighbour languages, Spanish and Portuguese. At the present time, we are particularly involved in intonation and duration modelling [1]. The first intonation module implemented in our synthesiser was based on the methodology proposed in [2] for Spanish, which makes use of a small set of average tonic-group patterns.

The previous model provides a quiet simple and fast implementation and renders reasonable good results for many applications. Nevertheless, when applied to unrestricted text-to-speech synthesis, the use of average patterns may lead to predictable and quiet monotonous prosody. In other to overcome this problem we propose a new model that generates the f0 contour in several steps, according to a defined hierarchy.

A traditional approach to f0 generation for English is established upon a rich phonological description (in terms of H and L tones in [3] or ToBI labels [4]. Unlike these approaches, the model proposed in this paper is set up upon syntactic and phonetic constituents for generating f0 values. This fact avoids the necessity of an intermediate phonological description of prosody, which makes this framework suitable for developing intonation modules for languages lacking in such kind of description.

The paper is organised as follows: in section 2 we briefly describe the main characteristics of the prosodic corpus employed in this research; in section 3 we introduce the basis of our model; and, finally, in sections 4 and 5 we depict the different levels of our model and its application to the synthesis of f0 contours.

2. CORPUS DESCRIPTION AND DATABASE GENERATION

The corpus we have employed in this research is made up of 1000 read sentences that include the usual and basic sentence modalities, i.e., simple statements, statements interrupted by another clause, statements finishing in enumeration or suspensive tone, wh- and yes-no questions, alternative questions, commands and exclamations, covering different lengths (from 3 to 10 tonic groups). These sentences have been segmented by means of an automatic alignment algorithm, and then the labels were revised manually. A set of pitch (f0) marks was obtained for each sentence using an algorithm based on a high-resolution pitch estimation [5].

The corpus of sentences was then analysed by the linguistic module of our TTS system to obtain the following features for each allophone: stressed/non stressed, type of syntagma, position of the allophone inside the syllable, position of the syllable within the tonic group, type of tonic group (oxitone, paroxitone, proparoxitone), position of the tonic group within the sentence, syllabic structure and some other additional information.

An automatic mechanism has been implemented in order to create a prosodic database by joining the linguistic features and the physical properties (duration and pitch marks) for every allophone. The labour is not hard providing there are not ellipses or insertions of allophones, so there is a one-to-one mapping between the theoretical and pronounced allophones. Otherwise a synchronisation method is applied by searching common sequences of allophones between the two sources of data. Finally, the pitch periods (or instantaneous T_0's) were postprocessed to obtain a robust estimation of a single f0 value per allophone.

The prosodic database has two main applications. The first one is to verify whether our approach to intonation modelling is appropriate for TTS synthesis, i.e., a proper f0 contour can be derived from the propositional and syntagmatic structure. To accomplish this task we have employed visual tests of aligned contours at phonic and tonic group levels. As an example, in figure 1 we can see the aligned phonic group shapes of several statements with the same length and syntactic structure. They all follow a similar pattern and the verb position seems to be an important factor for determining this pattern [6]. The second application is to serve as a prosodic database to employ in the intonation module of a TTS system.

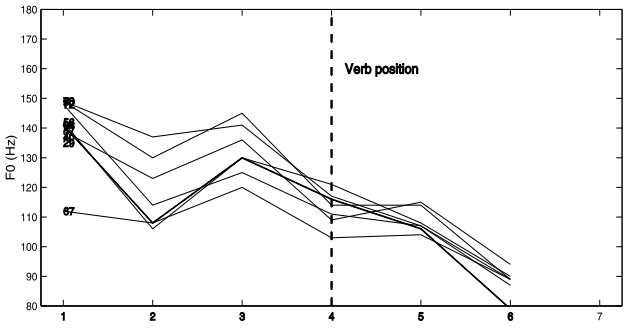

Figure 1: Aligned phonic group patterns (6 tonic group statements with verb in 4th tonic group position)

3. BASIS OF THE INTONATION MODEL

Our intonation model is built upon the following principles:

- It is a hierarchical model regarding the selection of f0 patterns at different levels and the construction of the f0 contour.
- It is a data-driven model. All the knowledge applied to f0 synthesis resides on a database of natural f0 contours whose allophones are featured according to syntactic-phonological units. No parameterisation of the natural contour is needed. Just one f0 value per allophone is used for the description of this contour.
- The model has an important linguistic background, using the structures of the phonic group linked to the clausal unit, the tonic group related to "syntagma", and the syllable.

4. LEVELS IN THE INTONATION MODEL

The model assigns f0 shapes to four different levels:

- At the sentence level, we assign an initial f0 pattern, which consists of a sequence of f0 values that represent the mean f0 values of the different phonic-groups. This strategy helps to maintain a proper relation among the mean f0 levels of the phonic groups that compound the different sentence structures.
- At the phonic group level, the f0 shape will be determined by the f0 values of the stressed vowels of the tonic groups. These values are considered as H targets and they determine a supra-line that will be the skeleton of the f0 contour. This supra-line contributes to model the declination along the sentence.
- The shape at the tonic group level is given by the array of f0 values of the syllabic nuclei. The aim of this level is to model the movements caused by pitch accents.
- At the segmental level, an f0 value will be specified for each allophone in the close neighbourhood of the most important prosodic events, i.e., accents and boundaries. We consider that the effect of these prosodic events extends over the tonic syllables, in the case of an accent, and over the last syllable in a boundary.

Using this methodology it is easy to mimic several intonation phenomena like declination, accents, downstep and final lowering, which extend, by nature, over units of different lengths. Of course, the different prosodic levels determine the phrasing to be generated by the Natural Language Processing module of our TTS.

5. F0 CONTOUR SYNTHESIS

The generation of the f0 contour includes two main operations: the selection of the different level patterns and the assembling of these pieces. There will be a set of minimal requirements for the selection of contours at each different level. In the case of several candidates matching the requisites, the selected pattern is the one that best matches the specifications of the lower levels. If there are no candidates, we apply a relaxation of the requisites or an alternative procedure that will be discussed later on.

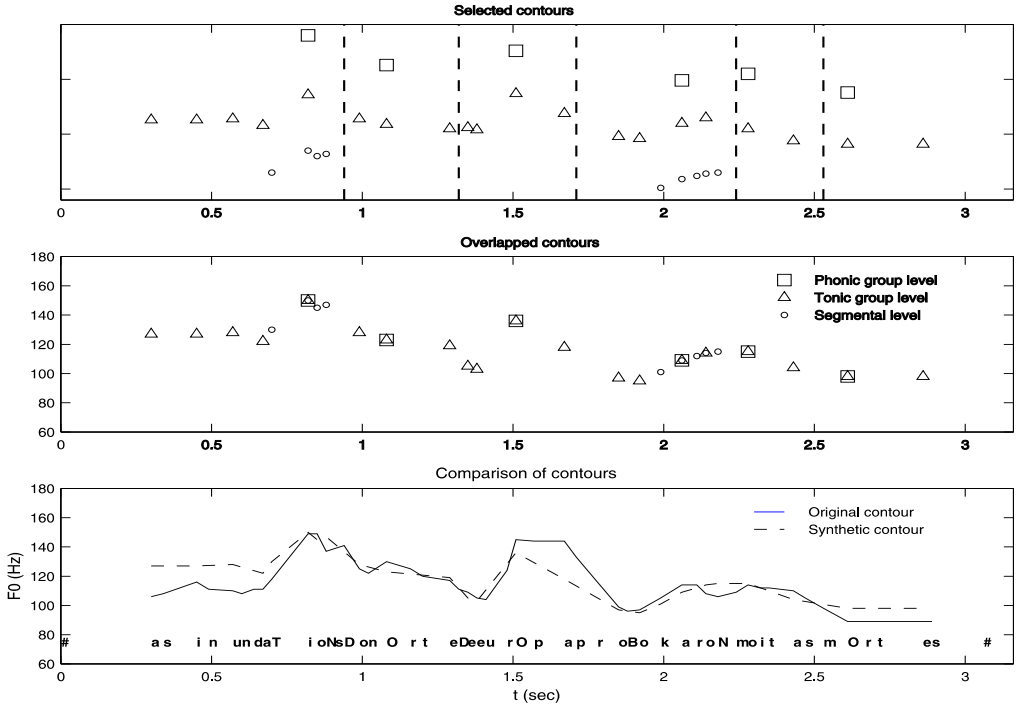

Figure 2: Selected patterns at the different levels, overlapped contours and comparison of the original and the synthetic contours

The algorithm starts selecting the proper phonic group pattern that will be the skeleton of the F0 contour. This phonic group has to fulfill the following demands: it must have the same sentence modality (statement, wh/yes-no question...) and the same syntactic structure. The syntactic structure is expressed in terms of the sequence of syntagmas (group of words composed by a content word and the previous function words). The types of syntagma considered are the nominal syntagma (which includes the adjectival syntagma), the prepositional syntagma, the verbal syntagma, the adverbial syntagma, etc. Given the syntagmatic structure of a phonic group, the f0 contours seem to follow a similar pattern, as depicted in figure 1. If we find several patterns in the prosodic database satisfying the search conditions, we choose the one whose tonic groups best accomplish the requirements of the immediate lower level.

The selected tonic group must have the same accentual structure according to the position of the stressed syllable within the word, and the same location within the phonic group (although a difference of +/-1 is admitted except for initial and final tonic groups). This tonic group will be assembled onto the phonic group pattern hanging of the target set in the upper level. In other words, the mean f0 value of the selected tonic group is increased or decreased in order to make its value equal to the target value of the phonic group.

At the segmental level we look for a segmental shape that matches the segmental structure of the stressed syllable, described in terms of the voiced/unvoiced feature of each allophone. This segmental contour must belong to the tonic group selected in the previous phase, so there is no process of selection, just a verification of this requirement. If the chosen tonic group does not match the segmental structure, then no contour is assigned at this level.

Finally, at the sentence level, the different synthesised contours rise or fall their mean f0 value to mimic a suitable sentence pattern stored in our database. This stage is not compulsory although it may improve the naturalness in some cases. A graphical explanation of the whole algorithm is shown in figure 2. The upper plot shows the selected contours (f0 values) at the phonic, tonic and segmental levels. The intermediate plot displays the overlapped contours according to our rules and, finally, the bottom plot depicts the comparison between the original and the resulting synthetic contour.

The phonic group level is the most important stage of the proposed intonation model since it has the greatest influence on the final shape of the f0 contour. Occasionally, depending on the coverage of the prosodic database, it is possible to find no suitable tonic groups. Such cases are solved by using the f0 value of the nuclei of the pre- and post-tonic syllable of the same tonic group

from which we had selected the phonic group target. In this way we are roughly modelling the vicinity of the pitch accent. The remaining points of the tonic group (one per syllable) are obtained by interpolation. An extreme case of this approach is shown in figure 3, where all the tonic group shapes are generated in this way. In that figure we can compare the original contour (solid line) and two synthetic contours. The phonic group pattern for the synthetic contours has been selected from two sentences with the same syntactic structure. The synthetic contours resemble the original contour quite reasonably.

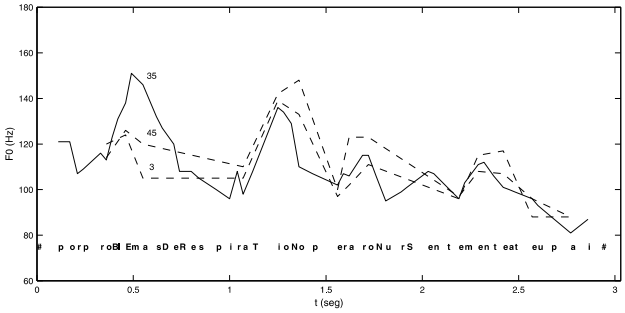

Figure 3: Original and synthetic contours obtained by the default method for generating the tonic group pattern.

The current implementation of the prosodic module is based on SQL queries over a table (stored in a file) which gathers all the sentences of the corpus. This makes up an amount of about 38000 records. Although this method allows a rapid development, it is not working on real time yet. To accomplish this goal, first we must reduce the size of the database by pruning the less useful patterns with the aim of loading the database in memory, and second we must optimise the search method rather than employing standard SQL software.

6.CONCLUSIONS

The proposed hierarchical model shows some interesting advantages over other prosodic models:

- The use of the hierarchical model for synthesising f0 contours by means of a database of natural f0 contours enables a coherent concatenation of the different intrinsic mean levels of the tonic group shapes. This seems to be adequate to preserve the global shape and the declination phenomenon.
- Flexibility. In case of unavailable patterns in the prosodic database we can relax the requirements demanded for a given level. This fact avoids the construction of complete databases at the beginning of the experimentation with the prosodic model.
- Easy upgrading. This property is linked to the flexibility of the model. Its performance can be easily improved by incorporating new sentences to the f0 contour database.
- The implementation of the proposed model avoids the necessity of a phonological description previous to f0 generation. It only needs an adequate phrasing and categorisation of the input sentence.

The new prosodic model has been perceptually tested by means of a sinusoidal algorithm [7] with very promising results.

7. ACKNOWLEDGEMENTS

This work has been partially supported by the "Centro Ramón Piñeiro para a Investigación en Humanidades", the Spanish CICYT under the projects 1FD97-0077-C02-C01 and TIC1999-1116, and the COST Action 258 "The naturalness of synthetic speech".

8. REFERENCES

1. Fernández Salgado Xavier, Banga Eduardo R., "Segmental Duration Modelling in a Text-to-Speech System for the Galician.Language". *Proceedings Eurospeech'99. Volume 4:* pp. 1635-1638. *Budapest, Hungary,1999*
2. Eduardo López Gonzalo. "Estudio de técnicas de procesado lingüístico y scústico para sistemas de conversión texto-voz en español basados en concatenación de unidades". *PhD thesis, Universidad Politécnica de Madrid. 1993*
3. Huang X.D., Acero A., Adcock J., Hon H.W., Goldsmith J., Liu J. Plumpe M. "Whistler: A Trainable Text-to-Speech System". *Proceedings of ICSLP 96, pp. 2387-2390, 1996*
4. Black, A. and Lenzo, K. "Building Voices in the Festival Speech Synthesis System" (DRAFT) Documentation and Scripts. 1999 http://www.cstr.ed.ac.uk/projects/festival/docs/festvox/
5. Medan Y, Yair E. and Chazan D. "Super Resolution Pitch Determination of Speech Signals". *IEEE Transactions on Signal Processing, vol 39, nº 1, pp. 40-48, 1991*
6. Fernández Salgado Xavier, Banga Eduardo R. "Proposición de un marco adecuado para el estudio de contornos de F0 para síntesis de voz". *XVI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural. September 2000, Vigo, Spain*
7. Banga Eduardo R., García Mateo Carmen, Fernandez Salgado Xavier. 1997. "Shape-Invariant Prosodic Modification Algorithm for Concatenative Text-to-Speech Synthesis". *Proc. Eurospeech'97. Vol 2, pp. 545-548. Rhodes (Greece), 1997*