ANN F0 Modeling for Female-Voice Synthesis in Spanish: restricted and non-restricted domains

J.M. Montero†, L.F. D'Haro† ,R. Córdoba†, J.A. Vallejo‡, J. Gutiérrez-Arriola† and J.M. Pardo†

† Universidad Politécnica de Madrid, Spain

‡ Universidad de Oviedo, Spain

E-mail: juancho@die.upm.es, lfdharo@die.upm.es, cordoba@die.upm.es, jmga@die.upm.es, pardo@die.upm.es

ABSTRACT

In this paper, we describe the modeling of the F0 curve of a female voice in several Restricted-Domains and in a general-domain, aimed at developing a Speech Synthesis System for Spanish. For modeling F0, we have used Multi-Layer Perceptrons, based on our previous experience with a male voice.

For isolated speech and for continuous speech, the use of specialized MLPs is always preferred. The main difference between restricted and non-restricted domains is the relevance of the number of the recording carrier sentence for predicting F0 in a restricted domain. The most relevant predicting parameters are stress, the position in the intonation group and the type of the group.

Keywords: ANN, F0 Modeling, Multi-Layer Perceptron, and Restricted-Domain Synthesis.

1. INTRODUCTION

Recorded speech is generally preferred in automatic information systems, but certain situations make speech synthesis the only economically-acceptable solution.

In automatic speech applications such as telephone banking or traffic information systems, a general-purpose speech synthesis system would be economically preferred (in order to develope applications that are easy-to-maintain), but speech quality is better when only a restricted domain is modelled.

Restricted-domain synthesis comprises at least two kinds of systems:

- isolated speech, for synthesis of keyword that must be recorded and played at a certain point in a carrier sentence
- continuous speech, using just a small vocabulary (specialized for the domain) under severe syntactic restrictions.

In the first case, the most difficult aspect is the size of the vocabulary that can be arbitrarily large (for instance: surnames, first names, towns and villages, etcetera). In the second case, syntactic constraints limit the variability of general-domain continuous speech.

In spite of these restrictions imposed by the restricted domain, spontaneous speech effects are also present: pausing, resetting, diversity of prosodic patterns for similar linguistic phenomena, etcetera.

An important advantage of restricted domain synthesis over a general purpose one is the possibility of tuning the design of the database. Usually we can apply summarizing techniques in order to mimick the general prosodic and phonetic characteristics of a whole database using just a few carefully-selected examples. These techniques can be used not only for prosodic modelling but also for collecting the database for a unit-selection concatenation synthesis system.

Although knowledge-based parametric prosody models would be preferred (they make the development of new voices easier) [2], the variability that characterises the naturalness of human speech is more easily accomplished through the use of Artificial Neural Networks such as Multilayer Perceptrons. The main drawback of this alternative is the need of getting enough training samples for allowing the network to generalize [3].

This study is organized as follows: Section 2 describes the databases for restricted and general domains. Section 3 explains the input parameters for prosody prediction. Section 4 describes experiments with isolated speech, and with continuous speech. Finally, in Section 5 we review the main conclusions of our work.

2. THE SPEECH DATABASES

2.1 THE RESTRICTED DOMAIN DATABASE

From two current actual services (in banking and traffic restricted domains), we selected 19 Carrier Sentences (CS) with 24 Variable Fields (VF). We can classify the sentences into 2 classes:

ISBN 1-876346-48-5

- *Names*: 9 Carrier Sentences with 11 Variable Fields (surnames and names of cities, villages and mountains). It is isolated speech.
- *Noun Phrases*: 6 Carrier Sentences with 9 Variable Fields (including complex banking operations, mainly). It is continuous speech.

We recorded and processed 660 Proper Names and 458 Noun Phrases (containing 360 surnames, 250 village names, 172 Bank names, 254 banking operations, etc. in their Variable Fields).

The database was automatically designed for preserving the phonetic and prosodic characteristics of the complete original unrecorded data [1]. The criteria for selecting the data comprise phonetic, syllabic, stress and lexical factors.

2.2 A GENERAL-PURPOSE DATABASE

It contains 5 recording sessions:

- 2 real oral speeches,
- 1 real oral interview, including questions, answers and exclamative sentences
- 2 specially designed sets of sentences that are structurally rich, in order to improve the coverage of certain linguistic phenomena.

The recorded speaker was the same as in the previous database, a professional female speaker used to record for automatic telephone applications.

Both databases were hand-labelled by a phonetician using a semiautomatic epoch-extraction software [1].

3. INPUT CHARACTERISTICS

To model the syllable-F0, we used a 3-layer Perceptron with a variable number of input parameters taken from the text of the recordings.

3.1 WINDOWED CHARACTERISTICS

As it is generally known, **stress** is one of the best predicting characteristics for the F0 contour we can use. In Spanish the position in the sentence is also very important [4]. We need to know whether the syllable is **initial** (at the beginning of the sentence), or **final** (it is before a pause at the end of an intonation group). According to previous experiments [4], in Spanish the initial part of the intonation group starts at the first syllable and goes to the first stressed syllable; the final part of the group begins at the syllable just before the last stressed one, and goes to the final syllable.

But the use of these characteristics in an isolated way (without windowing) is not enough for a good prosodic modelling. We will use several bits per characteristic, instead of using just one input parameter for coding the stress of the target syllable, one for coding whether this syllable is at the beginning of the sentence and one for coding whether it is at the end. To determine the optimal size of the window (up to +- 5 syllables around the target one) will be one of the objectives of our experiments.

We have tried alternative ways of coding this piece of information, using up to seven input bits (instead of three) to code whether the syllable is stressed, at the end or at the beginning, but there were no improvement in these alternatives, so they were discarded. To treat these parameters as orthogonal but correlated parameters seems to be better for predicting than to decorrelate them through the use of more input bits.

Other windowed characteristics (up to +-1 syllable) will be:

- **To belong to a function word**: although these words are generally spoken as unstressed ones, it is still useful to know that a syllable is part of a function word, or just before or after a function word.
- **To be at the end of a word**: our speaker used to speak in a rather rhythmic way with euphonic purposes, with regular F0 risings at the end of a word that is at the end of a clause.

3.2 NON-WINDOWED CHARACTERISTICS

- **The number of the carrier sentence**: in restricted-domain database each target element was recorded in a certain carrier sentence from the domain application.
- **Number of syllables** in the intonation group: we have used a code with 5 bits in a thermometer way; preliminary experiments with several codifications confirmed that a good one was: >0, >5, >10, >15, and >20.
- **Final or initial punctuation mark**: we distinguished 5 options: comma, full stop, colon, semicolon and question mark. In restricted-domain experiments each mark has a certain meaning related to the F0 movements that are typical in the domain
  - *Comma*: it codes the insertion of a spontaneous break by the speaker.
  - *Full stop*: it codes a standard non-rising ending.
  - *Semicolon*: it codes a final F0 continuation rise in the intonation group.
  - *Colon*: it codes an origin – destination group for a traffic domain.
  - *Question mark*: standard questions.

ISBN 1-876346-48-5

4. EXPERIMENTS

In all the experiments we will used a 10-fold cross-validation leave-one-out strategy with eight subsets for training the perceptron, one sub-set for avoiding over-training and one subset for evaluation; for each combination of input characteristics we will performed ten subexperiments with non-overlapping evaluation subsets and with one final combined evaluation score.

The output will be coded using a z-scored function [3], in order to profit the whole output range.

The error is measured in terms of mean of absolute differences between real syllable F0 and predicted one. No significant differences were observed between absolute distance and a mean square one.

4.1 ISOLATED SPEECH EXPERIMENTS

The total number of syllables is 2099.

The size of the context in the windowed characteristics results in no significant score differences, except for the smallest one (no context). The optimal size is +-1. To increase the size of the window does not decrease the error rate consistently.

The best learning coefficient of the network is around 0,2. The optimal number of hidden neurons is around 20.

The best option is to use the main input parameters: stress, initial and final. When not using one of the groups of these parameters, the results are worse but not significantly; the optimal size of the context gets higher (+-4 syllables). When preserving just one of these groups, the results are significantly worse with a 95% confidence level.

There is no significant improvement when using the number of syllables (in spite of several codifications we tried), but the improvement is consistent through several experiments.

4.1 EXPERIMENTS WITH CONTINUOUS SPEECH IN RESTRICTED DOMAIN

The total number of syllables is similar: 2416.

The importance of the parameters initial, final and stress is confirmed by these new experiments, but the size of the context has to be increased (+-2 syllables). In spite of this, only the lack of context exhibits a significantly worse performance.

When we do not include the number of the carrier sentence or if we do not code the punctuation mark, there is no significant increase of the error, because they are highly correlated (all the recordings in a certain carrier sentence have the same final punctuation mark, only the spontaneous break make the difference). When we suppress both of them, the decrease is now significant.

To code the number of syllables decreases the error but on in a significant way. The same applies to “be part of a function word”.

4.3 EXPERIMENTS WITH A GENERAL PURPOSE DATABASE

The total number of syllables is similar: 6166. This time, a greater size of context is preferred (3 or 4). There is no way of using the “carrier sentence” parameter.

New parameters such as “the initial punctuation mark” are useful but not significantly.

Observations about the main parameters (final punctuation mark, stress, position in the group and number of syllables in the group) are confirmed.

5. CONCLUSIONS

The most important parameters for the prediction of f0 for both continuous and isolated speech are: stress, the position in the intonation group and the type of the group.

For restricted-domain, the use of specialized MLPs is always preferred for different sub-domains. The main difference between restricted and non-restricted domains is the relevance of the number of the carrier sentence for predicting F0 in a restricted domain: as these sentences were recorded in the same recording session, they are more similar to one another. This is specially important in isolated speech, more conditioned by the surrounding sentence.

REFERENCES

[1] J.M. Montero, R. Córdoba, J.A. Vallejo, J. Gutiérrez-Arriola, E. Enríquez, J.M. Pardo, “Restricted-Domain Female-Voice Synthesis in Spanish: from Database Design to ANN Prosodic Modelling” in *Proceedings of the International Conference on Spoken Language Processing*, 2000.

[2] J.M. Montero, J. Gutiérrez-Arriola, R. de Córdoba, E. Enríquez, J.M. Pardo, “The role of pitch and tempo in emotional speech” in *Improvements in speech synthesis. Ed. Wiley & Sons,* 2002.

[3] R. de Córdoba, J.M. Montero, J. Gutiérrez-Arriola, J.A. Vallejo, E. Enríquez, J.M. Pardo, “Selection of the Most Significant Parameters for Duration Modeling in a Spanish Text-To-Speech System Using Neural Networks” in *Computer Speech & Language*, vol. 16, 2002.

[4] J.A. Vallejo “Improvement of the fundamental frequency in text-to-speech conversion”. Doctoral Thesis, ETSIT, Madrid, UPM, 1998.

ISBN 1-876346-48-5

¿Initial, Final and stress?	Size of context	Punctuation marks	Number of carrier sentence	Number of syllables	End of word	Absolute error
All	2	Yes	No	Yes	No	13.275
All	1	Yes	No	Yes	No	13.284
All	3	Yes	No	Yes	No	13.315
Without Stress	3	Yes	No	Yes	No	13.307
Without Initial	1	Yes	No	Yes	No	13.346
Without Final	4	Yes	No	Yes	No	13.630
Only Final	3	Yes	No	Yes	No	18.496
Only Stress	5	Yes	No	Yes	No	14.036
Only Initial	3	Yes	No	Yes	No	18.056
All	2	No	No	Yes	No	22.787
All	1	Yes	Yes	Yes	No	12.278
All	2	Yes	No	No	No	13.346
All	1	Yes	No	Yes	Yes	12.151

Table 1: Summary of results for isolated speech

¿Initial, Final And stress?	Size of context	Punctuation marks	Number of Carrier sentence	Number of syllables	Function word	End of word	Absolute error
All	2	Yes	Yes	No	No	No	16.663
All	4	Yes	Yes	No	No	No	16.714
All	5	Yes	Yes	No	No	No	16.732
All	1	Yes	Yes	No	No	No	17.906
All	4	Yes	No	No	No	No	17.702
All	2	Yes	Yes	Yes	No	No	16.634
All	2	Yes	Yes	Yes	Yes	No	16.553
All	2	Yes	Yes	Yes	No	Yes	16.565
All	2	Yes	Yes	Yes	Yes	Yes	16,322

Table 2: Summary of results for continuous speech in restricted domain

¿Initial, Final And stress?	Size of context	Final punctuation marks	Initial Punctuation mark	Number of syllables	Function word	End of word	Absolute Error
All	4	Yes	Yes	Yes	No	Yes	22.037
All	3	Yes	Yes	Yes	Yes	No	22.059
All	3	Yes	Yes	Yes	Yes	Yes	22.060
All	3	Yes	No	Yes	No	No	22.753
All	3	No	No	Yes	No	No	23.725
All	3	Yes	No	No	No	No	22.791
Without stress	3	Yes	No	Yes	No	No	23.235
Without initial	5	Yes	No	Yes	No	No	23.213
Without final	5	Yes	No	Yes	No	No	23.160

Table 3: Summary of results for continuous speech in a non-restricted domain

ISBN 1-876346-48-5