RESTRICTED-DOMAIN FEMALE-VOICE SYNTHESIS IN SPANISH: FROM DATABASE DESIGN TO ANN PROSODIC MODELING

J.M. Montero, R. Córdoba*, J.A. Vallejo*, J. Gutiérrez-Arriola*, E. Enríquez**, J.M. Pardo**
*Grupo de Tecnología del Habla. Dpto. de Ingeniería Electrónica. Universidad Politécnica de Madrid
E.T.S.I. Telecomunicación. Ciudad Universitaria s/n, 28040 Madrid, Spain
**Grupo de Tecnología del Habla-Departamento de Lengua Española-Universidad Nacional de Educación a Distancia-Ciudad Universitaria s/n, 28040 Madrid, Spain
e-mail: juancho@die.upm.es
http://www-gth.die.upm.es

ABSTRACT

In this paper, we describe the development of a female voice in a Restricted-Domain Speech Synthesis System for Spanish.

For the design of the database, we have used a greedy-algorithm approach that focus not only on covering a set of target phonemes, but also on mimicking the histogram of prosodic features from a larger database.

For modeling the prosody, both duration and F0, we have used two Multi-Layer Perceptrons, based on our previous experience in unrestricted-domain modeling. The error normalised by the deviation is always below 0.7.

Keywords: ANN, Prosody Modeling, Greedy Algorithm, Multi-Layer Perceptron, and Restricted-Domain Synthesis.

1. INTRODUCTION

The primary goal of this study was to develop a new female voice for our Spanish text-to-speech system [1], specially designed for restricted domain applications.

Although recorded speech is generally preferred in automatic information systems, certain situations (i.e. surnames) make speech synthesis the only economically acceptable solution. Thus TTS is commonly used when continuous updates are required, or when there are too many items to record.

A domain-specific application does not require so many sentence structures, but there are many words embedded in them. Although the delivered messages are syntactically constraint, the vocabulary size is potentially huge (i.e. more than 40,000 family names, more than 30,000 village names, etc.). A message is typically a sentence composed of 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. However, this contrast of human and artificial voices forces synthetic speech to be as close as possible to natural voice. The objective of such a system is to learn as well as possible the prosody of one specific speaker uttering messages of the specific application. The speech synthesis obtained tends to mimic the natural prosody exhibited by the speaker [2].

The need of an optimal carefully-designed database for training (and the benefits of this approach) has been underlined and proved in the recent literature [2] [3], under both objective and subjective evaluation. Not only for the training of the prosodic component, but also for having a variety of segments when we apply unit-selection techniques.

For F0 contour generation, many studies have been carried out lately using neural networks in a successful way [4][5][6]. For duration modeling, the use of neural networks has also been a matter of recent research [7][8]. In this paper, we propose the application of the same kind of approach to the generation of application specific prosody.

This study is organised as follows. Section 2 describes the database for two restricted domains. In Section 3, we describe the algorithm for the selection of recording items. Sections 4 and 5 present the parameters and experiments in prosodic modeling that have been carried out. Finally, in Section 6 we review the main conclusions of our work.

2. THE DATABASE FOR TWO RESTRICTED DOMAINS

We extracted an initial set of 22 sentences from two real services in banking and traffic information domains, provided by the IVR company that made the design of the dialogue. We discarded three of them because the Variable Fields (VF) were hours, telephone numbers or combinations of characters and numbers. These kind of items can be played with high quality using word concatenation, due to the limited vocabulary involved (although recording several examples per item is necessary when you aim at providing the highest quality, because you must account for a certain number of different prosodic situations).

The remaining 19 Carrier Sentences (CS) contained 24 Variable Fields (VF). As each VF conveys the most important information in the sentence, and for further restricting the prosody, the professional speaker had to utter each VF between 2 compulsory pauses.

We can classify the sentences into 3 classes:

- *Proper Names* (PN): 9 CS with 11 VF, that include surnames (both compound and simple ones), names of cities and villages, and names of mountain roads.
- *Questions* (Q): 4 interrogative CS with 4 VF containing bank-related information: currency, cheque status, etc.
- *Noun Phrases* (NP): 6 CS with 9 VF, also regarding to banking information: accounts, credit cards, names and types of financial transactions, names of banks. We include these later items in the NP class because they are syntactically related to NP, as in *Caja de Ahorros y Monte de Piedad de ...* or in *Banco de Crédito Local de ...*, where the names of these banks include a typical Noun Phrase structure with one or more Prepositional Phrases attached to it).

Due to restrictions in time and money, we decided to record about 600 sentences of each class. We recorded and processed 660 Proper Name sentences, 307 NP sentences (we recorded less examples because they were considerably longer) and 600 Q sentences. The embedded variable fields contain 360 surnames, 250 village names, 172 Bank names, 254 banking operations, etc.

When selecting the list of items to record, we have faced 4 situations:

- The number of available/possible items was the same as the number of recording items: we just used the available items for recording. Most of the banking information fell into this category.
- The number of available/possible items was smaller that the number of recording items: we took items from another VF in order to get a certain number of recorded examples to train the prosody generator. This was the case of Questions: just a few possible VF were obtained from our applications but, for extensibility, we added Proper Names and names of transactions and banks to get a richer and more general database.
- The number of possible items was greater than the number of recording items: we summarised the available database using the algorithm described in next section. This was the case of the names of villages and simple surnames.
- The available and recorded items were not homogeneous: we had no list for compound surnames. We created new names by mixing the most frequent 80 Spanish surnames and 80 surnames obtained by summarisation. The items that were selected because of their high probability of occurrence came from the results of the Onomastica Project [9], as in the previous case.

The recorded database was then phonetically labelled in a semiautomatic way. We used an automatic pitch epoch extraction software, and manually revised the outcome using a GUI adjusting programme. The same audio editor was used for the segmentation and labelling of the phonemes.

3. A NEW ALGORITHM FOR DATABASE DESIGN

As we mentioned previously, for those VF where the number of synthesisable items (i.e. surnames) was much larger than the number of recordable ones, we have used a new algorithm for item selection, based on a greedy strategy [10] with some variants to the basic distance. This kind of algorithm is based on the assumption that the sequential selection of locally optimum VF will lead us to a list of VF that is close to the global optimum, if we use a good distance to guide the greedy search.

We aimed at selecting a small database with the same probability distributions of certain phonetic and prosodic features as the distribution in the total database, not just covering the features that were included in this one. We were looking for the same percentage of each phoneme, syllable-type, stress-type and VF length, under a certain tolerance (e.g. a five per cent).

From our previous experience in unrestricted-domain prosodic modeling [5], we decided to impose up to 131 restrictions to the search. We can classify them under 4 categories:

- *Phonetic criteria*: we aim at designing small databases with the same phonetic histograms as the total ones for villages, surnames, etc.
- *Syllabic criteria*: the probability of each kind of syllable has to be preserved: stress level/position in the phrase/first phoneme in the syllable, etc.
- *Stress-type criteria*: in Spanish the stress of a word can be placed in any syllable of the word, but most often they are in the last three syllables.
- *Lexical criteria*: we aim at having the same probability distributions for the number of words in

each variable field, and for the number of syllables in each word.

As we have decided to record a certain number of items, we do not look for the minimum number of words that are a 95 per cent similar to the original ones. We need, for example, a set of less than 250 village names with almost the same characteristics as the database of 30232 Spanish village names.

At each step, the algorithm chooses one item, trying to minimise a distance between the cumulative distribution of selected items and the remaining target distribution at this step.

The best distance that we have experimented with moderates the greedy behaviour of the algorithm by setting a partial target distribution at each step that is not equal to the final one (*dynamic approach*). This partial target is the proportional part of the final one that should be covered at this step, taking into account the total number of steps that we accept (that is, the total number of allowed recording items). We compared this dynamic strategy with a *static approach*, where the target considered in each step of the algorithm is always our final target distribution.

In **Table 1** we show that, as the algorithm knows the initial and final number of items, it prefers to get a cumulative distribution that grows in a regular way at each step (dynamic approach). In contrast, the static approach grows initially faster but very slow in the final steps (some elements of the distribution cannot be covered). This is especially true when we must select fewer items.

Selected items from a database with 30232 items	**Absolute mean for the static objective**	**Absolute mean for the dynamic objective**
100	**0.028155**	**0.012760**
150	**0.016572**	**0.011882**
250	**0.009005**	**0.008255**

Table 1 Comparison of the absolute mean error for the static and dynamic strategies

In order to get the target accuracy in the distribution, we penalise the items that bring the cumulative distribution above the estimated partial target. As we were looking for a 95 per cent similitude in each condition (not just a 0.95 Pearson's correlation coefficient), we put a global penalty on those words that put the cumulative count beyond our partial objective for any of the conditions tested. If we put just a penalty on the component of the distribution that breaks our aims, the effect of the penalty is divided by the total number of components, and the algorithm tends to stop at a point with more gross errors, but better global correlation.

When summing up more than 32000 village names in just 250 balanced examples, the distance between the distributions of the target database and the selected one was less than a 1 % in absolute value. Only 4 out of 131 components were more than a 5% distant from the target, with a global correlation above 0.9995. When the number of items was increased (100, 150, 250 items), the Pearson's correlation was also increased, and the mean absolute error (MAE) and the number of gross errors (more than 5 % away from the final target value) were reduced in a consistent way as shown in **Table 1**.

Selected items	**MAE**	**Gross errors**
100	**0.01373**	**18**
300	**0.00871**	**7**
600	**0.004558**	**6**
1000	**0.003161**	**5**
5000	**0.001288**	**5**

Table 2 Summarisation experiment for a database with 79663 village and family names.

4. RESTRICTED-DOMAIN PROSODIC MODELING

To generate an F0 contour, our basic unit is the syllable. For modeling the syllable pitch, we have used a 3-layer Perceptron (one hidden layer) and 43 binary-coded input parameters:

- *VF type*: according to the prosody of our database, there are 5 kinds of phrases: interrogative, declarative with a final rising contour, declarative with a final falling contour, phrases ending at a spontaneous break and origin-destination structures. In a systematic way, the speaker used the falling/rising contrast to signal that the break was natural (falling contour) or forced by our recording conditions (rising contour).
- The *stress and position* in a window of 11 syllables. To code the position we have distinguished between initial syllables (before the first stressed one), final syllables (after the last stressed one) and intermediate syllables as in [4].
- The *number of syllables* in the phrase (5 bits)

and the F0 output parameter used a Z-score codification.

For phoneme duration modeling, we have also used a 3-layer Perceptron with 117 input parameters (as in [7]):

- *Target phoneme and its phonetic context*: a window of 5 phonemes coded as one central phoneme (38 bits) and the broad classes of the two previous and the two next phonemes (14 bits per position).

- *Stress level* of the syllable and the phoneme.
- *Syllable type*: stressed or non stressed, with or without an initial vowel, with or without a diphthong.
- *Position*: of the phoneme in the syllable, in the word and in the phrase, with a thermometer coding.
- *Number of*: phonemes in the syllable, syllables in the word and words in the phrase, with a thermometer coding too.
- *Phrase type*: the same 5 classes as in F0 modeling.

and also a logarithmic Z-scored output (normalised by the average elocution speed, and by the mean duration of each phoneme).

5. EXPERIMENTS

We can see some results from our experiments in **Table 3** and **Table 4**.

Syllable F0	
Size of the training set	**3926**
Size of evaluation set	**1656**
RMSE/standard deviation of evaluation data	**0.6748**
Evaluation Absolute Mean Error	**14.55 Hz**

Table 3 Results for the best experiment in ANN F0 modeling.

Phoneme duration	
Size of the training set	**15673**
Size of evaluation set	**4671**
RMSE/standard deviation of evaluation data	**0.6363**
Evaluation Absolute Mean Error	**17.5 ms**

Table 4 Results for the best experiment in ANN duration modeling

Using a multiplicative model for duration [11], with the best parameter coding of the ANN experiments, the absolute error was 19.8 ms, which is clearly worse than the result obtained with our neural network.

6. CONCLUSIONS

A new algorithm has been developed to design small speech databases with the same phonetic and prosodic characteristics of much larger ones.

An ANN is a good modeling tool for restricted-domain prosody, when compared to previous rule-based methods.

Due to its high quality, this new voice (both diphones and prosody) has been introduced in public real-time services.

7. REFERENCES

1. J.M. Pardo, F. Giménez de los Galanes, J.A. Vallejo, M.A. Berrojo, J.M. Montero, E. Enríquez, A. Romero, “Spanish text-to-speech: from prosody to acoustics”. International Congress on Acoustics, volume III, pp. 133-136, 1995.
2. O. Boëffard, F. Emerard, “Application-dependent prosodic models for text-to-speech synthesis and automatic design of learning database corpus using genetic algorithm”. Proceedings of Eurospeech Conference, volume IV, pp. 2507-2510, 1997.
3. J. Van Santen, A. Buchsbaum “Methods for optimal text selection”. Proceedings of Eurospeech Conference, volume II, pp. 553-556, 1997.
4. J.A. Vallejo “Improvement of the fundamental frequency in text-to-speech conversion”. Doctoral Thesis, ETSIT, Madrid, UPM, 1998.
5. C. Traber. “F0 generation with a database of natural F0 patterns and with a neural network”, in *Talking machines: theories, models and designs*, Elsevier Science Publishers B.V. pp. 287-304, 1992.
6. S. Tournemire. “Identification and automatic generation of prosodic contours for a text-to-speech synthesis system in French”. Proceedings of Eurospeech. pp. 191-194, 1997.
7. R. Córdoba, J.A. Vallejo, J.M. Montero, J. Gutiérrez-Arriola, M.A. López, J.M. Pardo, "Automatic Modeling of Duration in a Spanish Text-to-Speech System Using Neural Networks". Proceedings of Eurospeech, volume IV, pp. 1619-1622, 1999.
8. Y. Morlec, G. Bailly, V. Aubergé. “Synthesising attitudes with global rhythmic and intonation contours”. Proceedings of Eurospeech, 1997.
9. The Onomastica Consortium, “The ONOMASTICA interlanguage pronunciation lexicon”. Proceedings of Eurospeech, volume I, pp. 829-832, 1995.
10. G. Brassard, P. Bratley *Algorithmics: Theory and Practice* Prentice Hall, 1996.
11. J.M. Montero, J. Gutiérrez-Arriola, J. Colás, E. Enríquez, J.M. Pardo, “Analysis and Modeling of Emotional Speech in Spanish”. Proceedings of the XIVth International Congress of Phonetic Science, volume II pp. 957-960, 1999.