PROSODY MODELLING OF SPANISH FOR EXPRESSIVE SPEECH SYNTHESIS

Ignasi Iriondo , Joan Claudi Socoró, Francesc Alías

Department of Communications and Signal Theory
Enginyeria i Arquitectura La Salle, Ramon Llull University, Barcelona, Spain
{iriondo;jclaudi;falias}@salle.url.edu

ABSTRACT

This paper presents the use of analogical learning, in particular case-based reasoning, for the automatic generation of prosody from text, which is automatically tagged with prosodic features. This is a corpus-based method for quantitative modelling of prosody to be used in a Spanish text to speech system. The main objective is the development of a method for predicting the three main prosodic parameters: the fundamental frequency (F0) contour, the segmental duration and energy. Both objective and subjective experiments have been conducted in order to evaluate the accuracy of our proposal.

Index Terms— Speech synthesis

1. INTRODUCTION

A text to speech (TTS) system converts a written text into spoken language. TTS research was initially centred on reaching the greatest degree of intelligibility. Later, the goal has been to improve the naturalness of synthetic speech, that is to say, the ability to emulate the complexity of the human speech, which is intrinsically expressive, since the spoken message does not only contain verbal content, but also the speaker's intention, attitude or emotional state. In this context, the improvement of the expressiveness of TTS systems has been possible thanks to the advances in both prosody modelling and speech signal processing.

The study of the acoustic correlates of expression is complex and it has been tackled from different approaches (see [1] for an extensive comparative study). We can distinguish between studies centred on the voice quality or the paralinguistic use of the prosody [2, 3]. Expressive speech synthesis from tagged text requires the automatic generation of prosodic parameters related to the emotion/style and a synthesis module able to generate high quality speech with the appropriate prosody and the voice quality. The predominant TTS technique is based on unit selection [4].

This paper presents case-based reasoning (CBR) [5] —a machine learning (ML) method— applied to the automatic generation of prosodic parameters (fundamental frequency, phone duration and *rms* energy) for expressive speech synthesis in Castilian Spanish. Therefore, the presented approach and results are focused on the prosodic component of expressive speech. With this purpose, an acted expressive female speech corpus has been developed, which is also used in the concatenative speech synthesis process. The objective measures used to evaluate the accuracy of the prosodic modelling are the root mean square error (RMSE) and the correlation coefficient (ρ). The subjective performance of the method has been evaluated by means of a perceptual test using the Comparative Mean Opinion Scale (CMOS). A time domain PSOLA-like technique [6] was used to modify waveforms according to the new pitch and duration values, while a gain function was used to set the correct energy values.

2. RELATED WORK

The parameters that determine the prosody of a spoken text are essentially the segmental duration and intensity, the pause placement and duration and the F0 contour [7]. In the TTS framework, the literature on prosodic modelling is very extensive. The intonation contour has been the most studied feature, distinguishing between quantitative (as TILT [8], Fujisaki [9] or Bezier [10]) and qualitative methods (ToBI [11] or Intsint [12]). The segmental duration modelling has been tackled by rule-based methods [13] or statistical methods such as neural networks [14] or classification and regression trees (CART) [15]. The intensity modelling is the least present in the literature although there are some specific works in this direction such as [16, 17].

Modelling segmental duration of speech requires the definition of a basic speech unit. In most studies (i.e. [18, 19]), the phone has been chosen as basic unit for the duration, although other basic units could be used such as syllables [14]. In natural speech, the segmental duration varies depending on the context where it is fitted in. The common used features that influence the segmental duration of speech are: *i)* the identity of the current, previous and next phoneme, being possible to use directly their identifiers or a finite set of characteristics such as vowel/consonant, the mode or the place of articulation and sonority; *ii)* information related to the stress; and *iii)* information about the position of the phoneme within

This work has been partially supported by the European Commission, project SALERO FP6 IST-4-027122-IP.

a superior unit (syllable or phrase).

The prediction of the intensity contour is usually carried out at phoneme or syllable level. The features to consider for its modelling [7] are also related to the segment identity, its stress and its placement.

In [10], Escudero presents a complete state-of-the-art of the most used units of intonation in Spanish and the features that characterize them. There are different kinds of units used to model the intonation contour: the syllable or smaller units (microintonation), the stress group (SG) -related to rhythm-, the intonation group (IG) and other superior units (i.e. planning of the speech). Common features used for pitch modelling are the type of IG, the placement of SG in the IG, the position of the stressed syllable and the number of syllables of the SG and IG. Moreover, the intonation contour of every unit can be quantitatively modeled by different functions (polynomials, Bezier functions [10], logarithmic).

3. OUR APPROACH

3.1. CBR applied to prosody modelling

In this work, we present a new approach based on CBR since this analogical learning method allows a simple treatment of discrete and numerical attributes (without discretization) and numeric array classes (the parameters to be predicted). CBR has been formalized as a four-step process, named the 4R CBR cycle —Retrieve, Reuse, Revise and Retain— [5]. In the following paragraphs, the adaptation of these steps to the prosodic modelling is explained.

Initialization of the system is not properly a phase of the 4R CBR cycle, but it will be essential to obtain the memory of cases. Data compression and data fidelity should be properly balanced in order to generate the database (memory of cases). First of all, it is necessary to identify the attributes (or features) that define the cases for each one of the three system's tasks (the prediction of phone duration, phone energy and the intonation contour). Then, the training set is generated by joining the prosodic parameters annotated in the speech corpus with the prosodic features extracted from the linguistic analysis of the text. The reduction of cases is achieved through a clustering of the classes that are represented by the same attributes.

The aim of the Retrieve step is to map the solution from the previous cases to the target problem. The most similar case (or *k* cases) is recovered from the database using an adequate metric to the selected attributes. Reuse tries to solve the new case by reusing the information stored in the database. First of all, phoneme durations are predicted, since F0 contours have been stored after time normalization. Once the durations of the phonemes have been computed, the temporary axis is expanded and thus it is possible to associate the predicted mean F0 of every phoneme as the evaluation of the polynomial in the middle of the phoneme. If more than one sample ($k > 1$) is retrieved, it will be necessary to select only one solution. In the current implementation, the CBR Revise is not required since the storage (Retain) is already realized only in the initialization. Therefore, the system does not have the possibility of adding new cases when it is running.

3.2. Prosody representation

The automatic extraction of prosodic features from text is achieved by means of our linguistic analysis tool that carries out the phonetic transcription of the text (SAMPA), annotating intonation groups (IG), stress groups (SG), words and syllables. The IG in Spanish is defined as a structure of coherent intonation that does not include any major prosodic break. Prosodic breaks take place due to pauses or significant inflections of the F0 contour. Up to now, we are only considering the breaks defined by the signs of punctuation. The SG is defined as a stressed word preceded, if appearing, by one or more unstressed words. After evaluating different configurations of attributes, the best results with objective measures have been achieved with the set showed in Table 1, that depicts the label, a brief description and the type[1] of attribute. For segmental duration and energy modelling, the phone has been chosen the basic acoustic unit (as [18, 19]). The duration of the phone depends on basically its identity and the context where it is placed. Similar attributes have been used for energy (see Table 1).

Table 1. *Prosodic features for duration, energy and F0.*

Label	Features for duration prediction	Type
PHON0	Previous phoneme	D
PHON1	Current phoneme	D
PHON2	Next phoneme	D
STRESS	Stressed phoneme	B
SG-in-IG	Position of SG into IG	D
PHON-in-IG	Position of PHON1 in IG	D
DURATION	Phone duration in ms	N

Label	Features for energy prediction	Type
PHON	Current phoneme	D
STRESS	Stressed phoneme	B
SG-in-IG	Position of SG into IG	D
PHON-in-IG	Position of PHON in IG	D
PHON-in-SG	Position of PHON in SG	D
ENERGY	Phone energy in *rms*	N

Label	Features for F0 contour prediction	Type
IG-TYPE:	Type of IG	D
SG-in-IG	Position of SG into IG	D
STRESS	Position of the stressed syllable	D
IG-in-SEN	Position of SG in the sentence	D
SYL-NUM	Number of syllables of SG	N
F0	Polynomial coefficients of F0 contour	A

[1] (D) discret, (B) binary, (N) numeric, (A) numeric array

For the F0 contour modelling, the SG has been chosen following the proposal of [10]. The SG incorporates the influence of the syllable (it includes one stressed syllable plus some unstressed ones) and the pitch structure at IG level is achieved by the concatenation of SG contours. However, this model lacks variations due to microintonation. Up to now, we only differentiate between declarative, exclamatory and interrogative IGs, which are easy to annotate from punctuation signs. STRESS indicates the placement of the tonic syllable in the SG. The number of syllables is related to the length of the SG (see table 1).

A quantitative representation of the intonation has been used, by means of the coefficients of the polynomial that minimizes the error between the original set of points and the polynomial. Therefore, the class of the intonation parameter consists of the coefficients of the polynomial that are adjusted to minimize the distance between the polynomial and a collection of points that represent the value of the average F0 of every phoneme. This mean value of F0 is referenced to the centre of the phoneme.

4. EXPERIMENTS AND RESULTS

The experiments were performed using a *tiering* multi-domain Spanish speech corpus (2.5h) [20] recorded by a female professional speaker from an advertising database including enuntiative, interrogative and exclamative sentences. This corpus consists of 2590 sentences, which are grouped into three different domains: education (916 sentences), technology (833 sentences) and cosmetics (841 sentences). Each domain has been recorded using a predefined speaking style: happy (HAP), neutral (NEU) and sensual (SEN) respectively. Recently, we have added two new styles: anger and sadness, which are being now in process of segmentation and annotation.

4.1. Objective evaluation

The evaluation of the presented method has been conducted by means of objective measures as the root mean squared error (RMSE) and the correlation coefficient (ρ). The 75% of the speech corpus has been used for training the model and the rest for test. The mean values of both measures for the three styles (NEU, HAP and SEN) are shown in Table 2. Energy and duration present good results for the three styles, being SEN the worst rated style. For F0 prediction, the method seems to fail in HAP due to it is the style that presents the highest variability (NEU: μ=167 Hz, std=40.9; SEN: μ=134 Hz, std=26.1; HAP: 271 Hz, std=89.1).

4.2. Subjective evaluation

The experiment was set up as a Comparative Mean Opinion Score (CMOS) test. Nine subjects were asked to listen to the pairs of sentences (14 for NEU, 13 for SEN and 16 for HAP) and rate the similarity of both prosodies (natural and synthetic) on a five-point scale. The subject had to choose one answer from: (5) Very High, (4) High, (3) Certain, (2) Little, (1) No similarity. During the experiment, subjects were asked to pay attention mainly in the prosody and they could repeat the stimuli for many times until a selection was made. Both utterances were re-synthesized using a TD-PSOLA technique [6] to modify waveforms according to the input pitch and duration values, while a gain function was used to set the correct energy values. Input values for natural prosody (NP) utterances were extracted from the mean F0, duration and energy annotated for each phone in the test speech corpus. Synthetic prosody (SP) was computed by the proposed CBR method. The total time for the full test was about 25 minutes.

Table 2. *Mean values of the objective measures for the test corpus separted by styles.*

	F0 (Hz)		Duration (msec)		Energy (rms)	
Style	RMSE	ρ	RMSE	ρ	RMSE	ρ
NEU	30,55	0,71	21,91	0,70	0,022	0,85
HAP	73,25	0,51	26,48	0,75	0,026	0,77
SEN	22,33	0,40	29,00	0,64	0,030	0,68

Figure 1 shows the percentage of CMOS punctuation for the three styles. Notice that adding the best scores (Very high and High similarity) the results are very good for NEU ($>$ 55%) and SEN ($>$ 72%), and acceptable for HAP ($>$ 38%). The worst scores (Little or no similarity) are low rated: NEU ($<$ 14%), SEN ($<$ 6%) and HAP ($<$ 16%).

Figure 2 shows the resulting box plots of the CMOS puntuation for each style (also is indicated the mean and standard deviation). From the mean values of the CMOS test, SEN is the best rated style, followed by NEU and HAP the worst. The analysis of variance (ANOVA) for the 3 styles showed statistical significance of these mean CMOS results with $F(2, 386) = 27.61, p < 0.0001$. Note the correlation of this subjective measure with the RMSE for F0 (objective measure) shown in Table 2.

5. CONCLUSION AND FUTURE WORK

An adaptation of CBR has been done for predicting the duration, the F0 and the energy of the phonemes transcribed from text, that are the input to the synthesis module in our TTS system. The conducted objective and subjective experiments show good preliminary results for the major part of the task, but fails in the prediction of FO contour for happyness.

As future work, we want to explore the use of new prosodic features and the representation of F0 in order to improve the results, especially for happy style. Moreover, the influence of database clustering will be investigated. It also requires the improvement of the Retrieve step (recovering k solutions) and the development of a selection method to choose the final

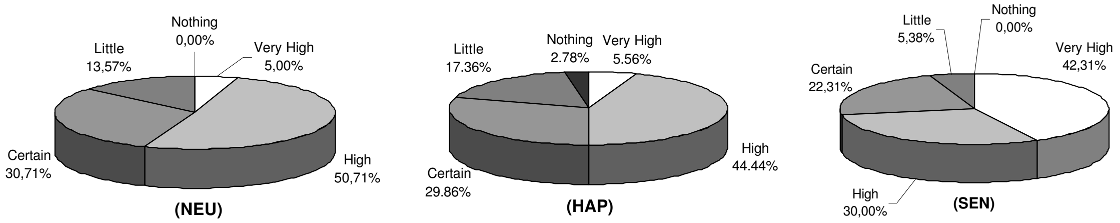

Fig. 1. *Similarity percentage between natural and synthetic prosody obtained with the perceptual test for the three styles.*

solution from the best candidates list.

Finally, two new emotions (sadness and anger) will be included in the modelling and, therefore, a more complete set of emotions will be tested.

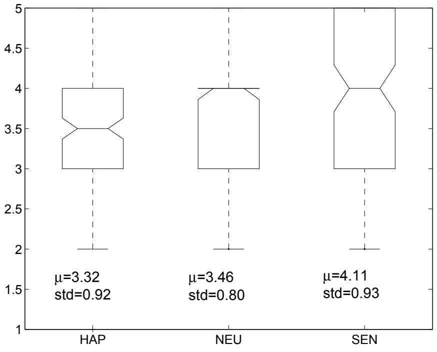

Fig. 2. *ANOVA Box plots and the mean and standard deviation of the CMOS scores for each style in the perceptual test.*

6. REFERENCES

[1] R.Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human computer interaction," *IEEE Signal Processing*, vol. 18, no. 1, pp. 33–80, January 2001.

[2] J. E. Cahn, "Generating Expression in Synthesized Speech," M.S. thesis, Massachusetts Institute of Technology, 1989.

[3] Montero J.M., J. Gutiérrez Arriola, J. Colás, E. Enríquez, and J.M. Pardo, "Analysis and modelling of emotional speech in Spanish," in *Proceedings of 14th International Conference of Phonetic Sciences*, San Francisco, USA, 1999, pp. 957–960.

[4] Alan Black, "Unit selection and emotional speech," in *the 8th European Conference on Speech Communication and Technology (EUROSPEECH)*, Geneve, Switzerland, 2003.

[5] A. Aamodt and E. Plaza, "Case-based reasoning: foundational issues, methodological variations, and system approaches," *AI Commun.*, vol. 7, no. 1, pp. 39–59, 1994.

[6] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for TTS synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, 1990.

[7] J. Llisterri, M. J. Machuca, C.de la Mota, M. Riera, and A. Ríos, *Entonación y tecnologías del habla*, Tecnologías del texto y del habla. Prieto, P. (Ed.) Teorías de la entonación. Ariel (Lingüística), Barcelona, 2003.

[8] Paul Taylor, "Analysis and Synthesis of Intonation using the Tilt Model," *Journal of Acoustical Society of America*, 2000.

[9] H. Fujisaki, S. Ohno, K. Nakamura, M. Guirao, and J. Gurlekian, "Analysis of accent and intonation in Spanish based on a quantitative model," in *Proc. ICSLP*, 1994.

[10] D. Escudero, *Modelado Estadístico de Entonación con Funciones de Bézier: Aplicaciones a la Conversion Texto-Voz en Español*, Ph.D. thesis, Universidad de Valladolid, 2003.

[11] K. Silverman, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labelling English prosody," in *Proceedings of ICSLP92*, 1992.

[12] D.J. Hirst, N. Ide, and Veronis J., "Coding fundamental frequency patterns for multilingual synthesis with INTSINT in the MULTEXT project," in *2nd ESCA/IEEE Workshop on Intonation*, 1994.

[13] D.H. Klatt, *Synthesis by rule of segmental durations in English sentences*, B. Lindblom and S. Öhman (Ed.), Frontiers of Speech Communication. New York: Academic, 1979.

[14] N.W. Campbell, "Analog I/O nets for syllable timing," *Speech Communication*, vol. 9, pp. 56–61, 1990.

[15] B. Möbius and J. van Santen, "Modelling segmental duration in German TTS synthesis," in *Proc. of ICSLP*, 1996.

[16] J. Trouvain, W. J. Barry, C. Nielsen, and O. Andersen, "Implications of energy declinations for speech synthesis," in *3rd ESCA/ COCOSDA Workshop on Speech Synthesis, November*, Jenolan Caves, Australia, 1998, pp. 47–52.

[17] I. Iriondo, R. Guaus, A. Rodríguez, P. Lázaro, N. Montoya, J. Blanco, D. Bernadas, J. Oliver, D. Tena, and L. Longhi, "Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques," *Proc. of the I. W. on Speech and Emotion*, pp. 161–166, Sept. 2000.

[18] E. Navas, I. Hernáez, and J. M. Sánchez, "Modelo de duración para conversión texto a voz en euskera," *Procesamiento del Llenguaje Natural*, vol. 1, no. 3, 2002.

[19] J. P. Teixeira and D. Freitas, *Evaluation of a Segmental Durations Model for TTS*, Computational Processing of the Portuguese Language - 6th Int. Workshop. N. Mamede, J. Baptista, I. Trancoso, M.G. Nunes (Eds), Springer, 2003.

[20] F. Alías, J.C. Socoró, X. Sevillano, I. Iriondo, and X. Gonzalvo, "Multi-domain text-to-speech synthesis by automatic text classification," in *In Proc. of ICSLP*, Pittsburg (USA), 2006.