Generating Spanish intonation with a trainable prosodic model

Gérard Bailly and Alex Bartroli

GIPSA-Lab, Speech & Cognition dpt., 46, av. Félix Viallet, 38031 Grenoble Cedex, France
{gerard.bailly,alex.bartroli}@gipsa-lab.inpg.fr

Abstract

A trainable prosodic model called SFC (Superposition of Functional Contours), proposed by Holm and Bailly, is here applied to Castillan intonation. Training material has been kindly provided by the Speech Processing Group at UPC. We describe the labeling framework and first evaluation results that compares the original prosody of test sentences of this corpus with their prosodic rendering by the proposed model and state-of-the-art systems available on-line on the web.

Introduction

The trainable prosodic model SFC (Superposition of Functional Contours) has been developed by Holm and Bailly [5, 14, 15]. It implements a theoretical model of intonation initially sketched by Aubergé [3] that promotes an intimate link between phonetic forms and linguistic functions: metalinguistic functions acting on different discourse units (thus at different scopes) are directly implemented as global multiparametric contours. These metalinguistic functions refer to the general ability of intonation to demarcate phonological units and convey information about the propositional and interactional functions of these units within the discourse. This trainable prosodic model has been confronted to speech styles and different languages including French, Galician or more recently Chinese [8] and German [4]. Spanish is of most interest because of its rich morphology and its potentially deep recursive syntactic embedding. Most quantitative models of Spanish intonation that have been so far applied to speech synthesis use a phonological representation with few levels when not limited to prosodic phrases [11, 16]. This approach favors data-driven techniques using concatenation of prosodic contours [9], more straightforward mapping between breaks and prosody [12] or both [1].

We describe here our first efforts in confronting the SFC to Spanish intonation. Our first parameterization of the SFC using limited training material is evaluated against state-of-the-art text-to-speech systems available on the web.

1. The SFC prosodic model

The SFC trainable prosodic model *directly* encodes metalinguistic functions by phonetic events – i.e. overlapping multiparametric contours – without any intermediate surface representation. For more details please refer to Bailly et al [5].

Input. These metalinguistic functions refer to the general ability of prosody to segment, structure, emphasize or encode semantic or pragmatic cues associated with speech units. These metalinguistic functions apply to units of variable sizes (discourse, sentence, clause, group, word, syllable, phoneme). The set of these metalinguistic functions [see intonation and its uses in 7] is quite open and parameterizing the SFC mainly consists in choosing the metalinguistic functions used in the training material and the speech units they apply to.

Prosodic contours. The SFC postulates that these metalinguistic functions are encoded via multiparametric contours. The contours are coextensive to the speech units carrying the functions – namely the scope. Since the same metalinguistic function (e.g. segmentation) may apply to speech units of various sizes - potentially embedded – the elementary multiparametric contours associated with each unit and each function overlap. The parallel encoding of these several metalinguistic functions by overlapping contours is simply done by superposing and adding these elementary contributions by parameter-specific operators (see illustration in ***Figure 1***) e.g. addition in the log-domain for f0 or addition of z-scores of rhythmic units for duration.

Mapping functions to contours. Recovering elementary prosodic contours from their superposition is a many-to-one ill-posed problem that requires regularization schemes. The SFC model does not impose low-level constraints such as contour shapes [such as in 10], but relies only on the consistency between all instantiations of the same discourse function on different units of different sizes within the corpus.

Contour generators The instantiation of a given function on a unit - the calculation of one elementary contour - is performed by so-called *contour generators*. A *contour generator* generates thus a family of contours that develop on units of different size but encode the same metalinguistic function. General-purpose contour generators have been developed in order to be able to generate a coherent family of contours indexed only by their scopes. These contour generators are implemented as simple feed-forward neural networks and an analysis-by-synthesis method has been developed to train these networks [5].

2. Training the model

2.1. The prosodic database

We used speech data from the female speaker MARTA from UPC [13]. The parts of the corpus we used consists of 50 sentences designed to get prototypical dialog messages for bank accounting and 516 sentences extracted from newspapers. All utterances of the corpus were segmented automatically and labeled phonemically (SAM-PA) and prosodically (borders+ accents). We corrected by hand all segmentations and checked carefully all phonetic transcriptions. Assimilations, elisions or idiosyncratic variations have been taken into account in order to have a clean phonetic training material. We used 556 utterances as training material. The 10 remaining sentences were used for evaluating the trained prosodic model (see §3).

2.2. Annotating utterances with metalinguistic functions

The first metalinguistic function is sentence modality (we have shown elsewhere [17] that prosodic attitudes in general are characterized/encoded by prosodic clichés). Its scope is the whole sentence. Some utterances of the corpus contain more than one sentence: we end up with 550 declarative

Table 1: Names and web sites of the systems used in the evaluation experiment (output gathered on July 6th 2007).

UPC	*Transmitted by e-mail (Marta, adult female)*
AT&T	*http://public.research.att.com/~ttsweb/tts/demo.php -(Rosa, latin American Spanish)*
UPM	*http://www-gth.die.upm.es/research/synthesis/synth-form-concat.html (adult male)*
VIGO University	*http://www.gts.tsc.uvigo.es/cotovia/ (Freire, adult male)*
IBM	*http://wizzardsoftware.com/ibm_demo.php (adult female)/*
Cepstral	*http://www.cepstral.com/demos/ (Marta, American Spanish)*
Loquendo	*http://actor.loquendo.com/actordemo/default.asp?language=es (Carmen, adult female)*

sentences, 17 exclamatives and 17 interrogatives. Markers thus cue beginning and ending of each sentence. For instance, the sentence shown in ***Figure 1*** is first annotated with the mark DC (for declaration) as below:

[llevaba el típico conjunto de blusas de quitaipón]$_{DC}$
Wearing a typical set of in- and-out blouses)

In our work, we always consider metalinguistic functions responsible for giving cues to the syntactic structure of sentences in the discourse. We thus consider dependency relations between chunks. Four kinds of dependency relations between constituents (words, groups, phrases, clauses) are considered: left dependency (DG, *dépendence à gauche*) linking the head of a sub-tree (so-called governor) with its immediately linearly preceding dependent unit (so-called sister), adjunct (AD) linking the governor with its immediately following dependent unit, interdependency (IT) linking two adjacent units headed by the same governor and independency (ID) when none of the preceding simple relations can be identified. For instance, the sentence shown in ***Figure 1*** should be further parsed as below:

[[llevaba]$_{AD}$[[el típico conjunto]$_{AD}$[[de blusas]$_{AD}$[de quitaipón]]]]$_{DC}$

Spanish has lexical stress. So markers are also added to signal lexical stress position within each word – when this position is not word final. In the example, the words "llevaba" and "conjunto" are further marked as below:

[[lleva]$_{AC}$[ba]] and [[conjun]$_{AC}$[to]]

2.3. Automatic annotation

The syntactic parse given above does not coincide with the one displayed on Figure 1 since the parsing is done fully automatically from raw text using the FreeLing software developed [2] at the TALP Research Center at UPC. The version 1.3 provides morphological analysis and PoS tagging, incorporates chart-based chunking/parsing, quantity recognition (currency, ratios, physical magnitudes...) as well as dependency parsing.

The dependency parsing was simply corrected to cope with coordinated phrases and incomplete sentences. Results were then used to place marks and determine their scopes. No attempts were made to improve this parsing. The aim of this work was to demonstrate that the SFC was able to discover meaningful prosodic regularities even in case of noisy annotations. We will show that SFC does not require so much expertise to place marks and scopes nor so much precision in order to generate adequate prosody.

2.4. Prosodic stylization

We analyze and generate *multiparametric prosodic contours*, i.e. we model the melody *and* rhythmic organization of the utterance. These contours capture the prosodic characteristics of the syllables of each utterance. Each syllable is characterized by a melodic movement (stylized by three F0 values) and a lengthening factor (that will stretch /compress all phonemic segments of a syllable using z-scoring, see [6]).

Table 2: *RMS errors (correlation coefficients) for training and test material. F0 mean is set to 155 Hz for that speaker.*

Parameter	Training	Test
F0 (Hz)	18.3 (0.71)	14.1 (0.76)
F0 (semitones)	1.94 (0.71)	1.51 (0.76)
Lengthening factor	0.29 (0.46)	0.29 (0.55)
Durations(ms)	20.4 (0.55)	20.5 (0.57)
Nb. of phonemes	20500	409

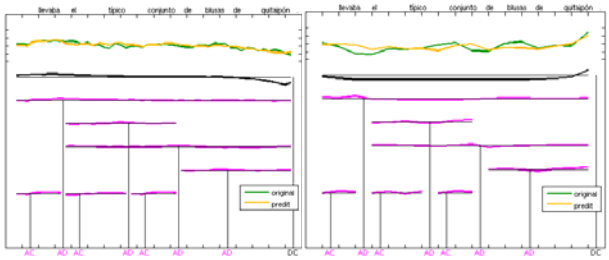

Figure 1. *Comparing original and predicted prosodic contours for the sentence "llevaba el típico conjunto de blusas de quitaipón ». Left: f0 contours. Right: syllable lengthening. For each caption: top: superposition of predicted (yellow) and original (green) contours; bottom: elementary contours predicted for each discourse function; the prediction is obtained by superposing and adding these elementary contours. Horizontal axis represents the syllable count.*

2.5. Mapping metalinguistic functions to elementary multiparametric contours

The SFC (i.e. the eight contour generators) is then trained using phonetic and associated annotations from 1556 utterances. As stated above, the iterative training consists in adjusting the contour generators so that their combined outputs best predict observed multiparametric contours. Prosodic stylization of 10 test utterances (see Appendix) is then predicted using the trained SFC. The ***Figure 1*** displays the elementary melodic and rhythmic contours generated by the 8 contour generators (responsible for generating elementary contours encoding DC, EX, QS, DG, AD, IT, ID, AC) on the different scopes. It also displays the results of the superposition and addition of these elementary contours in comparison with the original training material. Figure 4 displays movement expansion of the lengthening factor produced by the two contour generators DC and AD for various scopes. For DC the length of the sentence is varied from 7 to 23 syllables while both left and right parts of the scope are varied for AD. We demonstrate that our generators are quite consistent and generate coherent contours even for

unseen scopes. Prediction performance on training and test material is given in Table 2. Note that these numbers are not very high, largely due to the automatic annotation.

Figure 2. *Java interface for subjective evaluation (www.icp.inpg.fr/~bailly/Evalpro/Castillan). Subjects should position stimuli by drag and drop (the natural utterance + synthetic renderings by 8 different prosodic models) on a MOS scale. Subjects can listen to stimuli as much as required (simple-click) for taking their decision.*

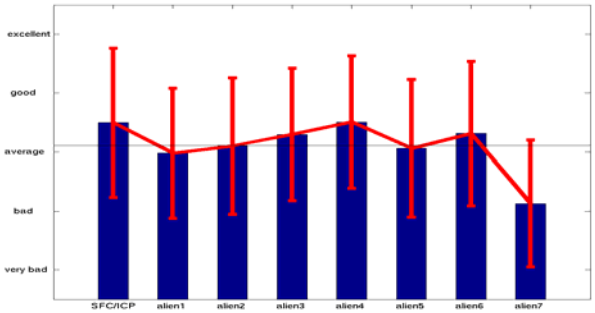

Figure 3. *Comparative results of the MOS test performed by 30 subjects. The proposed model lies in the heading set of state-of-the-art systems.*

3. Subjective evaluation

In order to compare this preliminary prosodic model with reference to available implementations, we collected prosodic characteristics of outputs of 7 state-of-the-art text-to-speech synthesizers (see Table 1): these systems will be anonymously named Alien1..7. The 10 test sentences were submitted to each online text-to-speech server and the synthetic audio files collected. Automatic f0 detection and phonemic alignment was then corrected by hand. The same prosodic stylization as used by SFC was finally performed to gather 7 alternative prosodic contours for each test sentence. Note that we compensate summarily for differences in voice registers of the different systems: we compute the mean f0 for each alien system and scale it to the mean f0 of our target speaker.

3.1. Evaluation procedure

We compared the synthetic prosody computed by our 8 different prosodic systems using TD-PSOLA resynthesis of the natural test signals. Prosody for all systems is characterized by segmental durations and three F0 values on the vocalic nucleus. The evaluation paradigm combines advantages of Mean Opinion Score (MOS) ratings and preference tests. We ask our subjects to position our synthetic stimuli – identified as colored icons – in a geometric plane whose abscissa is the MOS scale. As they can listen to stimuli as many times as they want, they can compare stimuli by pairs, position stimuli already ordered in some part of the plane and further refine their judgment (this procedure has been used by Pfitzinger [18] for studying perceived speech rate and by Bailly et Gorisch [4] for assessing German). Each subject ranks thus the ten sets of 8 stimuli arbitrarily thrown on 10 successive test planes.

3.2. Results

We launched the evaluation campaign on October 10th. By the time of the submission, 39 Spanish listeners participated in the MOS test. 5 listeners have experience with synthetic speech and prosody while 34 have none. The preliminary results are shown in ***Figure 3***. Four groups emerge from this first evaluation: the SFC has a statistically significant different mean from all alien systems except the commercial system alien4 (Anova analysis using anova1 and multcompare with hsd option ***M***atlab® procedures). Experience with synthesis/prosody has no significant influence on the results.
These comparison are not completely fair since the SFC model uses training and test sets from the same speaker. The results show however that the proposed system generates an acceptable prosody that lies in the heading set of state-of-the-art systems we were able to input.

4. Conclusions

We described a first confrontation of a trainable prosodic model, to Spanish intonation. The highly embedded syntax of the Spanish language fits quite well our theoretical framework. The first results are objectively and subjectively encouraging but the generated prosody is still far from natural prosody. This certainly due to the impoverished linguistic input: markers and scopes are determined automatically using a parser that has some difficulties to deal with newspaper articles and a manual correction of its output may certainly reduce training modeling errors. But despite its crude assumptions and potential refinements, the SFC is flexible enough to automatically capture essential prosodic regularities of the language it observes given general assumptions on metalinguistic functions of intonation.

5. Acknowledgements

We warmly thank Antonio Bonafonte and Pablo Daniel Agüero for providing their precious database to us.

6. Appendix: test sentences

„.se fue a madrid robando y ha vuelto de madrid ganando. aullar y maullar son cosas parecidas. esa pieza de zinc ofrece unos pocos ohms de resistencia. Zeus, Hercules, Minerva y Diana fueron unos dioses muchísimo más socorridos. el arroyuelo se adentraba en el bosque con un suave zigzag. estando en el vivac se escuchaban los aullidos de un lobo. se había adoptado en manlléu hacía tiempo y estaba compungido. el ñu pudo escapar con un asombroso zigzag, fue un giro ceñidísimo. acuña y josé ignacio participan en el referéndum. concepción y gonzález hicieron las labores de información“

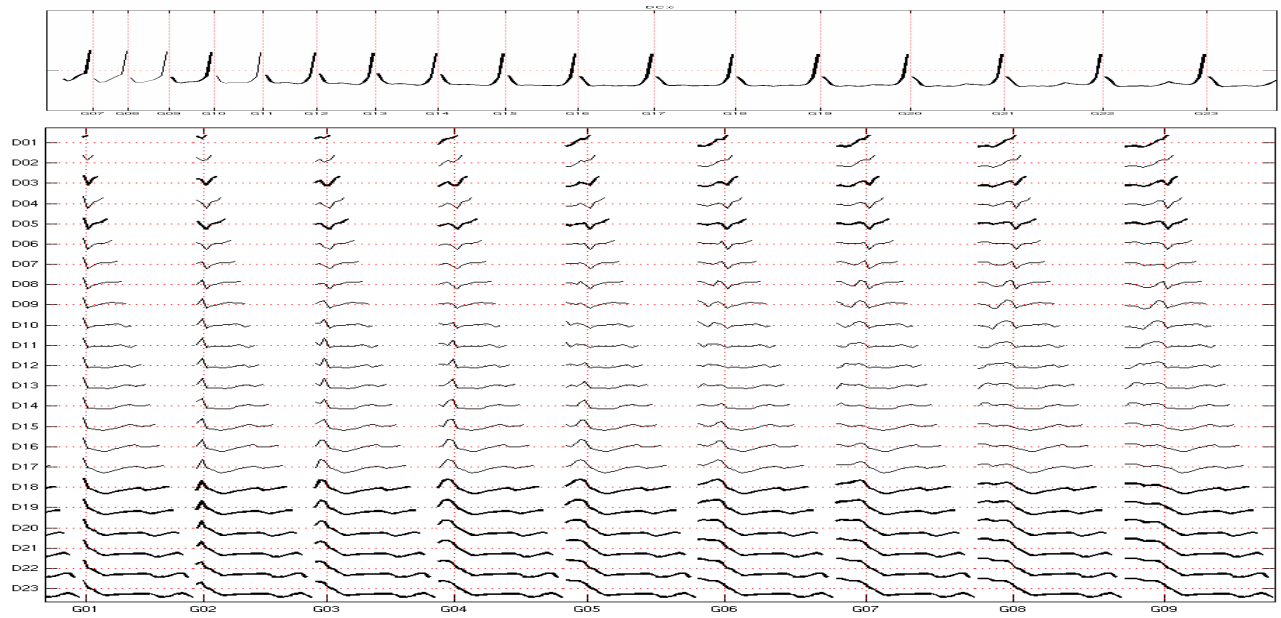

Figure 4. *Movement expansion of the lengthening factor produced by two contour generators. Top:DC generating contours for declarative sentences here from 7 to 23 syllables. Bottom: AD for adjuncts from 1 to 9 syllables following a group of 1 to 23 syllables. Dark contours are represented in the training corpus while prototypical contours displayed with light gray are not.*

References

[1] Agüero, P.D., K. Wimmer, and A. Bonafonte. *Joint extraction and prediction of Fujisaki's intonation model parameters*. in *International Conference on Spoken Language Processing*. 2004. Jeju, Korea. p. 757-760.

[2] Atserias, J., et al. *FreeLing 1.3: Syntactic and semantic services in an open-source NLP library*. in *International Conference on Language Resources and Evaluation*. 2006. Genoa - Italy. p. 2281–2286.

[3] Aubergé, V. and G. Bailly. *Generation of intonation: a global approach*. in *Proceedings of the European Conference on Speech Communication and Technology*. 1995. Madrid. p. 2065-2068.

[4] Bailly, G. and I. Gorisch. *Generating German intonation with a trainable prosodic model*. in *InterSpeech*. 2006. Pittsburgh, PE. p. 2366-2369.

[5] Bailly, G. and B. Holm, *SFC: a trainable prosodic model*. Speech Communication, 2005. **46**(3-4): p. 348-364.

[6] Barbosa, P. and G. Bailly, *Generation of pauses within the z-score model*, in *Progress in Speech Synthesis*, J.P.H.V. Santen, et al., Editors. 1997, Springer Verlag: New York. p. 365-381.

[7] Bolinger, D., *Intonation and its Uses*. 1989, London: Edward Arnold.

[8] Chen, G.-P., et al. *A superposed prosodic model for Chinese text-to-speech synthesis*. in *International Conference of Chinese Spoken Language Processing*. 2004. Hong Kong. p. 177-180.

[9] Escudero, D. and V. Cardeñoso. *Optimized selection of intonation dictionaries in corpus based intonation modelling*. in *Interspeech*. 2005. Lisboa, Portugal. p. 3261-3264.

[10] Fujisaki, H. and H. Sudo, *A generative model for the prosody of connected speech in Japanese*. Annual Report of Engineering Research Institute, 1971. **30**: p. 75-80.

[11] Garrido, J.M., et al. *Prosodic markers at syntactic boundaries in Spanish*. in *Proceedings of the 13th International Congress of Phonetic Sciences*. 1995. Stockholm, Sweden. p. 370-373.

[12] Gutierrez, J.M., et al. *New rule-based and data-driven strategy to incorporate Fujisaki's F0 model to a text-to-speech system in Castillian Spanish*. in *ICASSP*. 2001. Salt Lake City, UT. p. 821-824.

[13] Hernández, I. and A. Moreno. *Diseño de un corpus para una base de síntesis de voz*. in *Actas del XV Simposium Nacional de la Unión Científica Internacional de Radio URSI*. 2000. Zaragoza, Spain: ISBN 84-600 9597-5. p. 67-68.

[14] Holm, B., *Implémentation d'un modèle morphogénétique de l'intonation. Application à l'énonciation de formules mathématiques*. 2003, Institut National Polytechnique: Grenoble - France. Thèse de Doctorat Signal-Image-Parole-Telécoms sous la direction de Gérard Bailly p. 239.

[15] Holm, B. and G. Bailly. *Learning the hidden structure of intonation: implementing various functions of prosody*. in *Speech Prosody*. 2002. Aix-en-Provence, France. p. 399-402.

[16] López-Gonzalo, E. and L.A. Hernández-Gómez, *Data-driven joint F0 and duration modeling in text-to-speech conversion for Spanish*. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994. **1**: p. 589-592.

[17] Morlec, Y., G. Bailly, and V. Aubergé, *Generating prosodic attitudes in French: data, model and evaluation*. Speech Communication, 2001. **33**(4): p. 357-371.

[18] Pfitzinger, H.R. *Local speech rate as a combination of syllable and phone rate*. in *International Conference on Spoken Language Processing*. 1998. Sydney, Australia. p. 1087-1090.