Learning the Hidden Structure of Intonation : Implementing Various Functions of Prosody

B. Holm & G. Bailly

Institut de la Communication Parlée UMR CNRS n°5009 INPG/Univ. Stendhal
46, av. Félix Viallet 38031 Grenoble CEDEX France
{holm,bailly}@icp.inpg.fr

Abstract

This paper introduces a new model-constrained, data-driven method to generate prosody from metalinguistic information. We refer here to the general ability of intonation to demarcate speech units and convey information about the propositional and interactional functions of these units within the discourse. Our strong hypothesis are that (1) these functions are directly implemented as prototypical prosodic contours that are coextensive to the unit(s) they apply to, (2) the prosody of the message is obtained by superposing and adding all the contributing contours [2]. We describe here an analysis-by-synthesis scheme that consists in both identifying these prototypical contours and separating out their contributions in the prosodic contours of the training data. The scheme is applied to databases designed to evidence various functions of intonation. Experimental results show that the model generates faithful prosodic contours with very few prototypical movements.

1. Introduction

It is a commonly accepted view that prosody crucially shapes the speech signal in order to ease the decoding of linguistic and paralinguistic information by the listener. In the framework of automatic prosody generation, we aim at computing adequate prosodic parameters carrying that information. In order to automatically learn the mapping between discursive functions and prosody and eventually sketch a comprehensive model of intonation, we have to answer two main questions: *what* information is transmitted and *how* this information is encoded?

2. A morphogenetic model

Encoding discourse structure – supposed to be discrete -- by means of continuously varying prosodic parameters is described by a large variety of tentative approaches. A phonological interface is usually promoted that translates discourse structure in a multi-level - potentially infinite [15]- phonological structure. Phonological units are typically delimited by salient prosodic events, typically accents, tones or breaks such as pauses [12, 21]. This step of phonological transfer is followed by the generation of the prosodic continuum thanks to a specific phonetic model e.g. targets connected by interpolation functions [11, 19], series of syllable–sized contours [22, 23] or superposition of contours with variable size [1, 8-10].

The morphogenetic model developed at ICP [1, 4] contrasts with most the models developed so far on two main points: (a) functions of discourse units are directly encoded as *global multiparametric prosodic contours* (b) the encoding of the multiple functions acting at different scopes for structuring the message is simply done by *overlapping and adding* contributions of the different contours. Our phonetic model is thus clearly global and superpositional, but contrastively with Fujisaki et al [8], the phonetic model is not motivated by any production mechanism – although this mechanism may have acted as a bootstrap - but by communication needs, i.e. maintaining perceptual contrasts that ensure optimal decoding of the functions.

Note that we have added in the current implementation of the model another strong hypothesis to the point (a): the global contours are only parameterized by the *scope* – or domain - of the function, i.e. the size of the units the function is applied to… and does not depend on the nature and internal organization of the units.

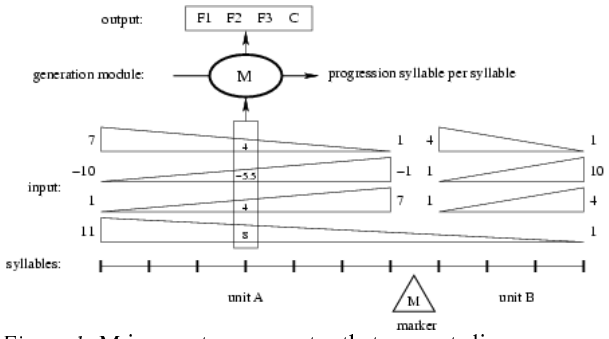


Figure 1: M is a contour generator that converts linear ramps - anchored on the boundaries of units A and B - into prosodic trajectories: for each syllable of the units, it delivers three F0 values (F0 values at 10,50 and 90% of the vocalic nucleus of each syllable[1]) and a lengthening factor (phoneme durations are further computed together with pause generation using the procedure described in *[5]*).

2.1. Contour generators

Each discourse function may be applied to diverse discourse units. We define the *scope* of a function as the continuous set of words which are concerned with this function. These functions typically assign a communicative value to a unit or qualify the link between units within the discourse. The *segmentation* function can for example indifferently demark a word, a group or a clause off the utterance. The same *qualification* function is applied indifferently to an adjective, a noun complement or a clause qualifying a preceding noun or nominal group. Similarly an *emphasis* function could be indifferently applied to any constituent of the discourse.

Each discourse function is then encoded by a specific prototypical contour anchored to the function's scope by so-called *landmarks*, i.e. beginning and end of the units concerned with this function. As the discourse function can be applied to different scopes, it is characterized by a family of contours - some sort of prosodic "clichés" [7].

General-purpose *contour generators* have been developed in order to be able to generate a coherent family of contours given only their scope. These contour generators are actually implemented as simple feedforward neural networks [13] receiving as input linear ramps giving the absolute and relative

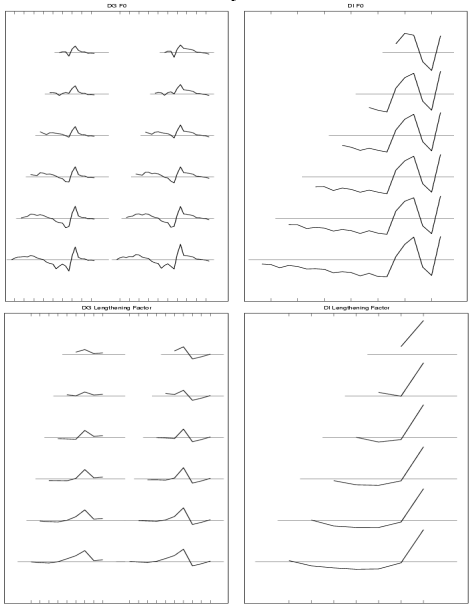
[1] This simple strategy explains the oscillations exhibited by the prosodic contours due to the adjacent consonantal dips that a smoothing procedure (e.g. [11]) could easily wipe out.

distance of the current syllable from the closest landmarks and delivering as output the prosodic characteristics for the current syllable (see *Figure 1*). Each network have very few parameters - typically 4 input, 15 hidden and 4 output units = 4*(15+1)+15*(4+1) = 139 parameters - to be compared to the thousands parameters necessary to learn a "blind" mapping between phonological inputs and prosodic parameters such as in [6, 24]. We have shown that our contour generators implement a so-called Prosodic Movement Expansion Model (PMEM) that describes how prototypical contours develop according to the scope (see for example *Figure 2*). Note that the choice of the neural networks implementation of the PMEM is not exclusive, but offers an efficient learning paradigm as described below.

2.2. Analyzing prosody

The mapping between discourse structure and the phonological structure is usually not straightforward: a direct mapping between these two structures is highly problematic [16, 20]. Most authors thus rely on a specific analysis technique – often requiring expertise - for constructing the phonological tree from raw acoustic data, not to speak of a further mapping between this surface phonological tree and communicative functions.

In the case of a superpositional model, the problem is often ill-posed since each observation may be the sum of several contributions, i.e. here the outputs of contributing contour generators. We thus need extra constraints to regularize the inversion problem, e.g. shapes/equations of the superposed components as in [17]. In our phonetic model, shapes of the contributing contours are unconstrained – e.g. we have shown that contours encoding attitudes at the sentence level may have complex shape [18]. Note however that nothing forbids in the following framework adding further constraints on the shapes of the contours such as imposing exponential shapes as in the Fujisaki's model.



(a) (b)

Figure 2: Expansion of the melodic contour produced by a contour generator encoding different functions: (a) a presupposition relationship between two units. First the length of the first unit is varied from 2 to 7 syllables. Second unit has 2 (left column) and 3 (right column) syllables; (b) an incredulous question on a sentence of 2 to 7 syllables. Top: melodic prototypes; bottom: lengthening factor profiles.

The shapes of the contributing contours emerge here as a by-product of an inversion procedure that parameterize contour generators in such a way that the prosodic contours predicted by overlapping and adding their contributions in the discourse best predict observed realizations. The analysis procedure is by essence reversible and our phonological model - implemented as dynamical prototypes - emerges from an iterative analysis-by-synthesis process as follows:

1. we generate the assumed contribution of each discourse function at each supposed scope in the corpus with the generators. In their initial state, they produced a null output.
2. we compute a prediction error by subtracting the sum of these elementary overlapping contours to the original prosodic contours observed in the corpus.
3. this prediction error is then distributed and partial contributions are added to the contributing contours. These new contours are used as targets during a classical learning procedure for neural networks.

These three steps are iterated until the prediction error of a test set reaches a minimum. This scheme relies on three hypothesis: (a) the prediction error contains the information that is contained in natural prosody but not (yet) captured by the contributing generation modules; (b) step 3 provides a filter capturing regularities within each target set, i.e. if a contribution of the prediction error is attributed to the "wrong" module, it should have no systematic relation to the associated input values and will thus be flattened in step 3; (c) the family of contours that contour generators are able to produce is finite i.e. the simple phonotactic information provided to the contour generators constrains the topology of the mapping implemented by each network.

Table 1: RMS prediction errors (correlation coefficients) for different corpora. F0 errors are given in semitones, GIPC and phoneme durations in ms. The last column gives the number of syllables and phonemes considered. The last and first syllables of the sentences are excluded

	F0 [st]	Gipc [ms]	Phon. [ms]	Nsyl/Nphon
Math	2.29 (0.87)	105 (0.90)	31.2 (0.67)	2805/ 7557
DC	1.99 (0.90)	47 (0.72)	27.9 (0.61)	1702/ 4199
DI	1.53 (0.95)	36 (0.81)	43.5 (0.86)	870/ 2263
Text	1.27 (0.82)	21 (0.84)	15.6 (0.77)	10209/20267

3. Applying the morphogenetic model to diverse corpora

We summarize here the results obtained on different corpora using half of the corpus as learning data. The prediction statistics are given in *Table 1* using all available data.

3.1. Maths

The Math corpus [14] was established in order to study how prosody may encode highly embedded dependency relations between constituents of an utterance. Read Mathematical Formulae (MF) were chosen because they offer a deep syntactical structure and because they are - when spoken - often ambiguous, forcing thus the speaker and the listener to use prosodic cues. All formulae are algebraic equations such as proposed in 4th grade exercises. They involve classical operations on 2nd degree polynomials. The corpus was generated automatically by systematically varying the length and syntactic depth of constituents. We end up with 157 MF that were recorded by one male French speaker who was instructed not to use lexical structural markers - as "open parenthesis" - but to make use of prosody.

Each formula has been uttered twice. In order to describe the natural variability of our data we give here correlations/RMS-errors between the repetitions: phoneme

durations: 0.857/20.6 ms, syllable durations: 0.919/92.6 ms and F0 0.902/2 semi-tones (sm). The two versions have 579 – internal - pauses in common of a total of 616. Pause durations[2] are correlated by 0.917. Note that even in case of such a close repetition, we still have a large variance. These values serve as reference for the corresponding values between the model's predictions and the original variance given *Table 1*. The z-score repartition scheme [5] generates 558 pauses at a location common to either natural versions, omit 58 pauses and generates 128 extra pauses[3]. Durations are correlated by 0.794. The results for F0 and syllable durations are close to the natural variability. The rather big difference in phoneme durations is considered to be perceptually less crucial (the model predicts in fact only Inter-Perceptual Centres (IPCG) durations: phoneme and pause durations are obtained by a z-score- not optimized for that particular speaker.

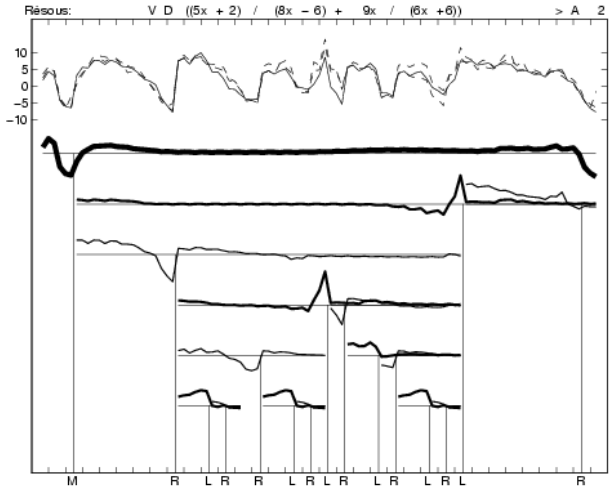


Figure 3: Predicting/analyzing the melody of a complex formula as the superposed contributions of three contour generators encoding an introduced assertion (M) and two dependency relations between the left operand and the operator (L) and between the operator and its right operand (R). Top are superposed the prediction (plain) and the original F0 stylization (dashed). M contribution (thick) is shown below with L (plain) and R (thin).

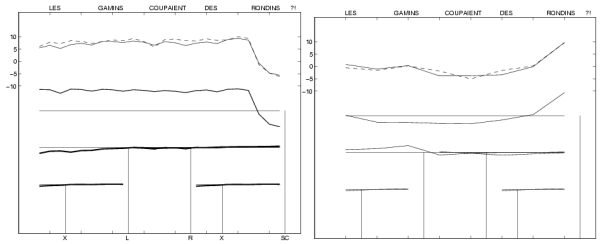


Figure 4: Predicting/analyzing the melody (left) and syllable lengthening (right) of a sentence uttered with a suspicious irony. Note that the amplitudes of the contours carrying phrasing structure are quite reduced.

[2] A pause which is not realized in either stimuli is considered with null duration.

[3] These locations and associated pause durations tend globally to enhance the phrasing structure of the utterance such as disconnecting the right operand from the major operand "equals to".

3.2. Prosodic attitudes

The corpus of prosodic attitudes [18] reveals the existence of statistically significant global prosodic contours that encode communicative functions at the utterance-level: 322 syntactically balanced unmarked sentences were uttered by one speaker with six different prosodic attitudes: declarative (DC), question (QS), exclamation (EX), incredulous question (DI), suspicious irony (SC) and obviousness (EV). The morpho-syntactic structure of the sentences and their lengths (between 1 and 8 syllables) were systematically varied in order to eliminate coincidental covariations between the contours encoding the communicative function at the utterance-level and the morpho-syntactic structure of the sentence. The decomposition of utterances into sentential and phrasal intonation is performed for each prosodic attitude separately. A further analysis demonstrates that contours carrying morpho-syntactic information are quite reduced especially for non modal attitudes (see *Figure 4*). In this case, the speaker is supposed to doubt, be ironical or suspicious about a previous assertion of his interlocutor, who does not require phrasing to be returned back to him.

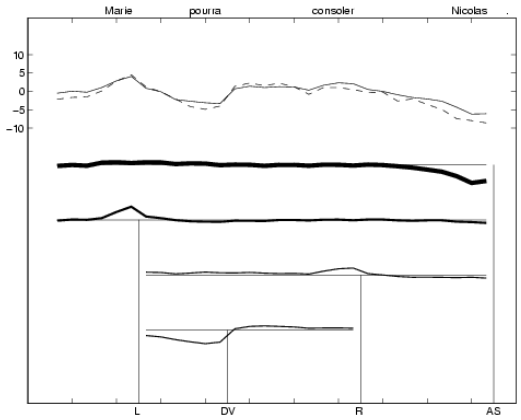


Figure 5: Predicting/analyzing the melody of a read sentence "Marie pourra consoler Nicolas". The DV contour is often observed and is generally used to segment determinants or auxiliaries, here "pourra" from "consoler"

3.3. Text reading

This eclectic studies of highly dedicated materials assess some properties of the natural intonation and evidence some important features of the morphogenetic model such as (a) the existence of global contours that encapsulate co-occurring salient events and have been confirmed by gating experiments [3] and (b) the possibility of intonation – with syntax – of carrying structural information with very few contour generators.

This should however not obscure the main technological grail of speech synthesis: being able to read texts. A corpus of 1000 sentences was designed to cover extensively the standard declarative form of French sentences Gn Gv, while extending Gn from a simple pronoun to a complex nominal group with adjectives, noun complements and simple qualificative clauses, and Gv with adverbs or verb complements. When a simple decomposition of utterances into sentential and phrasal intonation is performed, the systematic opposition between a full name and a determinant+noun nominal group as well as between a full verb and a modal auxiliary+infinitive reveals the necessity to introduce an additional function DV used to segment between a function word and its related content word (see for example Figure 5).

4. Conclusion & perspectives

The analysis-by-synthesis procedure presented here gives access to the *hidden structure* of intonation: the phonetic implementation of discourse functions emerges from the automatic parameterization of contours generators. This procedure is data-driven but also model-constrained and thus converges towards optimal prototypical contours that satisfy *both* bottom-up (close-copy synthesis) and top-down (coherent phonological description) constraints.

Such a phonology of prototypes can easily include a paradigm for learning automatically *alloprosodic* variations i.e. privileged directions of variations around the prototypes and implement a model of phonological *gradience* able to encode and modulate the degree of importance of the information carried by the contour in the discourse.

By applying the model to different communicative functions we have demonstrated that this model can actually capture statistically significant prosodic variations with a rather few number of prototypical movements, generates *faithful* and varied prosodic contours. This model provides a useful tool for analyzing the "hidden" structure of intonation i.e. decomposing a surface prosodic contour into overlapping contours that actually implement a given communicative function in a statistically-significant way. We plan to exploit this model for analyzing multilingual corpora and implementing new functions. For instance, we are currently working on Galician, a language with lexical stress.

We have also tried to demonstrate that this model-based comprehensive generation scheme may be compatible with a certain technological efficiency: confronting data-driven models against such thematic databases used here should provide an interesting basis of comparison between models and approaches that we are still looking for.

5. Acknowledgements

This work was supported by Cost258. We thank H. Loevenbruck and G. Rolland for providing us the text reading corpus and fruitful comments on the first version of this paper. A special thank also to our proofreader L. Ménard.

6. References

[1] Aubergé, V., 1992. Developing a structured lexicon for synthesis of prosody. In *Talking Machines: Theories, Models and Designs*, G. Bailly; C. Benoît, Editors: Elsevier B.V., 307-321.

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

[3] Aubergé, V.; Grépillat, T.; Rilliard, A., 1997. Can we perceive attitudes before the end of sentences? The gating paradigm for prosodic contours. In *Proceedings of the European Conference on Speech Communication and Technology*. Rhodes - Greece, 871-874.

[4] Bailly, G.; Aubergé, V., 1997. Phonetic and phonological representations for intonation. In *Progress in Speech Synthesis*, J.P.H.V. Santen, et al., Editors. New York: Springer Verlag, 435-441.

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

[6] Chen, S.-H.; Hwang, S.-H.; Wang, Y.-R., 1998. An RNN-based prosodic information synthesizer for Mandarin text-to-speech. *Speech and Audio Processing*, 6(3):226-239.

[7] Fónagy, I.; Bérard, E.; Fónagy, J., 1984. Clichés mélodiques. *Folia Linguistica*, 17:153-185.

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

[9] Gårding, E., 1991. Intonation parameters in production and perception. In *Proceedings of the International Congress of Phonetic Sciences*. Aix-en-Provence, France, 300-304.

[10] Grønnum, N., 1992. *The ground-works of Danish intonation*. Copenhagen: Museum Tusculanum Press - Univ. Copenhagen.

[11] Hirst, D.; Nicolas, P.; Espesser, R., 1991. Coding the F0 of a continuous text in French: an experimental approach. In *Proceedings of the International Congress of Phonetic Sciences*. Aix-en-Provence, France, 234-237.

[12] Hirst, D.J.; Di Cristo, A.; Espesser, R., 2000. Levels of representation and levels of analysis for the description of intonation systems. In *Prosody: Theory and Experiment*, M. Horne, Editor. Dordrecht - the Netherlands: Kluwer Academic Publishers, 51-87.

[13] Holm, B.; Bailly, G., 2000. Generating prosody by superposing multi-parametric overlapping contours. In *Proceedings of the International Conference on Speech and Language Processing*. Beijing, China, 203-206.

[14] Holm, B.; Bailly, G.; Laborde, C., 1999. Performance structures of mathematical formulae. In *Proceedings of the International Congress of Phonetic Sciences*. San Francisco, USA, 1297-1300.

[15] Ladd, D.R., 1986. Intonational phrasing: the case for recursive prosodic structure. *Phonology Yearbook*, 3:311-340.

[16] Marsi, E.C.; Coppen, P.-A.J.M.; Gussenhoven, C.H.M.; Rietveld, T.C.M., 1997. Prosodic and intonational domains in speech synthesis. In *Progress in Speech Synthesis*, J.P.H. van Santen, et al., Editors. New York: Springer-Verlag, 477-493.

[17] Mixdorff, H., 2000. A novel approach to the fully automatic extraction of Fujisaki model parameters. In *International Conference on Acoustics, Speech and Signal Processing*. Istanbul - Turkey, 1281-1284.

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

[19] Pierrehumbert, J., 1981. Synthetizing intonation. *Journal of the Acoustical Society of America*, 70(4):985-995.

[20] Pierrehumbert, J.; Hirschberg, J., 1990. The meaning of intonational contours in the interpretation of discourse. In *Intentions in Communication*, P.R. Cohen; J. Morgan; M.E. Pollak, Editors. Cambridge, MA: MIT Press, 271-311.

[21] Silverman, K.; Beckman, M.; Pitrelli, J.; Ostendorf, M.; Wightman, C.; Price, P.; Pierrehumbert, J.; Hirschberg, J., 1992. TOBI: a standard for labeling English prosody. *International Conference on Speech and Language Processing*, 2:867-870.

[22] t' Hart, J.; Collier, R.; Cohen, A., 1990. *A perceptual study of intonation: an experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.

[23] Taylor, P., 2000. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, 107(3):1697-1714.

[24] Traber, C., 1992. F0 generation with a database of natural F0 patterns and with a neural network. In *Talking Machines: Theories, Models and Designs*, G. Bailly; C. Benoît, Editors: Elsevier B.V., 287-304.