GENERATING PROSODY BY SUPERPOSING MULTI-PARAMETRIC OVERLAPPING CONTOURS

Bleicke Holm and Gérard Bailly

Institut de la Communication Parlée - CNRS/INPG/U3
46, avenue Félix Viallet, 38031 Grenoble Cedex 1, France

ABSTRACT

We present here a model for generating prosody by superposing overlapping multi-parametric contours. These contours are associated with high-level communication tasks such as segmentation, hierarchisation or emphasis of discourse units. We propose a analysis-by-synthesis scheme for automatically learning these contours and apply this new paradigm to the enunciation of mathematical formulae and utterances carrying various attitudes.

1. INTRODUCTION

It is a commonly accepted that prosody takes part in the transmission of linguistic information during the speech act. One of the most studied linguistic functions that prosody assumes is hierarchisation and segmentation of units. If there is a large consensus in literature concerning the major role of prosody in language acquisition and structuring of speech, authors diverge on the way prosody actually encodes information. In the framework of automatic prosody generation the aim is to establish an acceptable evolution of prosodic parameters depending on the structure of the utterance. The major questions to answer are: *What* information is transmitted? – and: *How* is this information encoded?

Encoding linguistic information – supposed to be discrete – by means of continuously varying prosodic parameters is described by a large variety of approaches. In most cases a phonological interface is used that aims to translate linguistic structure in a phonological structure by organising salient prosodic events into tonal and accentual structures. This step of phonological transfer is followed by the generation of the prosodic continuum thanks to a specific phonetic model.

Phonological units. Automatic generation systems may be differentiated by means of the richness of the phonological structures they employ: Black and Hunt [3] use the ToBI-transcription [13], Di Christo et al. IntSyn [7] – whereas other authors prefer a more economic description of the phonological structure mainly in terms of accentual groups [15, 9, 10]. This last group of studies is very heterogeneous: Malfrère et al. [9] use a crude segmentation and indexing procedure concentrating on an elegant concatenation of contours, whereas Pasdeloup [12] focusses on establishing a well suited segmentation into accentual groups.

Phonetic Models. Differences in the phonetic models are partly induced by the differences in phonological models: If the phonological description allows a determination of target points, a simple interpolation between these points may be sufficient [7]. If the output of the phonological model consists in indexed domains, prosodic contours are generally obtained by concatenation and/or superposition of sub-contours living on the domains. Subcontours maybe generated by various manners ranging from rule-based to statistical methods.

Targets vs. Contours. Gating experiments seem to indicate that prosodic contours contain more information than a series of target points. Our believe is that such targets (if they exist) are a secondary product of a more global and implicit process. Salient prosodic events may be used to describe specific prosodic realisations and to compare natural and synthetic prosody, but they should not be at the basis of a model. An extreme example of such an implicit modelling is the work of Traber [15] who uses a neural network to map the information supposed to determine prosody directly to the phonetic realisation. The important contribution is *not* the use of a neural network, but the proposition of an encoding procedure on a sub-symbolic level. On the other hand, this model operates as a black box and hence provide no comprehensive inside on prosody at all.

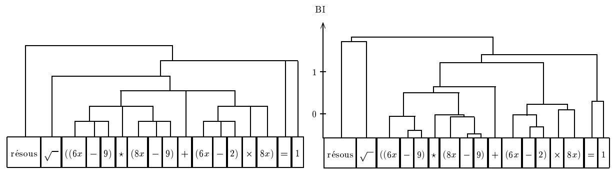
Linear vs. superpositional encoding. In order to communicate we have to be able to compose and decompose the information in the prosodic signal. If we reject the decomposition into target points for the reasons given above, we have to adopt another encoding/decoding procedure. Since Fujisaki [5], superposition of contours has been widely used. Many models use at least two superposing levels: phrase and accentual group. Superposition may be very explicit as in Möbius et al. [10] or more implicit as in Black et al. [3] where accent realisations are expressed relative to a declination line. Following Thorsen [14], Aubergé [1] uses up to four levels. The main characteristic of her proposal is the priority given to the functions that may be encoded by prosody and the direct process of phonetic encoding.

The model we present here is inspired by Aubergé. We propose a straightened relation between functions and phonetic realisation which allows for a simple and comprehensive description of the encoding process that operates by superposing overlapping and nested contours that are coextensive to the discourse units concerned by the encoded function. This model is able to generate highly structured prosody since we do not impose a restriction on the number of superposed levels. An important part of this paper is the description of how to obtain from a learning corpus the contours to be superposed. We will show and discuss the application of the model on the enunciation of mathematical formulae (MF) and utterances carrying various attitudes – demonstrating that our model is compatible with the expansion of prosodic contours described by Morlec [11].

2. ENUNCIATION OF MATHEMATICAL FORMULAE

Our corpus was established in order to study how prosody may encode dependency relations in an utterance. Read MF were chosen because they offer a deep syntactical structure and because they are – when spoken – often ambiguous, forcing thus speaker and 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 varying systematically the length and syntactic depth of constituents. We end up with 157 MF recorded by one male French speaker who was instructed not to use lexical structural markers – such as "open parenthesis" – but to make use of prosody.

As shown in figure 1 and more thoroughly analysed in [8], the prosodic structure of the spoken formulae is rich and deep. Moreover, the height of a node in the performance tree (cf. Grosjean [6]) grows with the syntactic depth of the associate constituent which is in good correspondence with a superpositional model.



(a) Syntax (b) Performance

Figure 1: Syntactic structure and performance structure of a formula

3. DESCRIPTION OF THE MODEL

In Aubergé's [1] proposal every linguistic level generates, depending on its nature, function and length, a set of multiparametric contours that are then concatenated and superposed. Morlec [11] proposed a modular connectionist model enabling a learning of contours with less a priori assumptions and a thorough analysis of movement expansion. This approach has been validated in analysis as well as synthesis.

Nevertheless this morphological model suffers from two drawbacks: (a) it follows syntax too closely, i.e. contours are indexed by detailed and often redundant syntactic characteristics – and (b) even if the implementation by means of recurrent neural networks copes better with coarticulation phenomena as a simple concatenation of contours [1, 9], it looses the notion of a lexicon of contours indexed by linguistic functions.

Our proposal aims to solve these problems by (a) establishing a real functional morphophonology – and (b) using a superposition of localised and overlapping contours.

3.1. A functional morphophonology

We suppose that prosody is part of an interacting system of linguistic agents that provide general functions. Some functions – such as encoding of attitudes and emotions – may be fulfilled by prosody in a more or less autonomous manner, others – such as segmentation and hierarchisation are accomplished more synergetically. In this study we test the hypothesis that prosody is able to encode segmentation and hierarchy of adjacent discourse units by a contour that is coextensive with the two units and that is anchored at the frontier between these units. Hierarchy is encoded by superposition of these localised contours that may overlap and that may be nested. The number of nesting levels is not constrained. This type of encoding offers an attractive compromise between information spreading (needed to explain anticipatory phenomena) and information localisation (to that part of the utterance it is meant to act on).

3.2. Learning of contours

In order to get the model to work, the crucial problem is its inversion, i.e. the extraction of contours from a natural speech corpus. Inversion becomes possible if we impose several constraints: (1) a contour is determined by the function it encodes, (2) it does *not* depend on the hierarchical level it is applied to and (3) contour shapes depend only on their lengths (in terms of number of syllables) and the position of the anchor point.

For *each* linguistic function to be encoded there is *one* generation module in charge with providing the mapping from phonotactic input information to prosodic parameters (cf. fig. 2). The modules are implemented as simple feed-forward neural networks – *not* in order to claim any neuromimetic properties but because they have proofed their ability to generate coherent contours with respect to the input information.

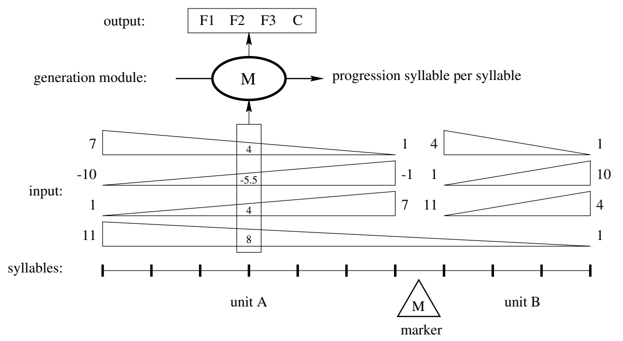


Figure 2: Generation of a contour encoding the relation "M" between two adjacent units. For each syllable the module provides based on a 4-component input vector four prosodic parameters (three F_0-targets and a lengthening factor that measures the deviation of the syllable duration with respect to an expected duration).

Since each module may be used to generate contours on any hierarchical level, we cannot use a simple hierarchical learning scheme as done by Morlec. We propose an iterative *analysis-by-synthesis* method as follows (cf. fig. 3): (1) generating synthetic contours with the modules in a preliminary state, (2) using the error of the predicted contours with respect to the observed ones in order to calculate a better set of contours for each module that (3) will be used as targets during a classical learning procedure for neural networks. – This scheme relies on two hypothesis: (a) The prediction error contains the information that is contained in natural prosody but not (yet) captured by the contributing generation modules – and (b) the learning 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 (3). This filter property allows for a certain freedom in the implementation of (2). The main constraint for "better" target contours is that their superposition equals the observed contour. The easiest way to do this is to divide for each syllable and for each prosodic parameter the error by the number of contributing contours and to add that to these contours. While this procedure provides a good approximation of the training corpus after a reasonable number of learning cycles, we can expect better generalisation properties if we impose further constraints during (2): Prototypical contours should be reasonable simple – this can be achieved by flattening the contributing contours before calculating the preliminary prediction. Properties of the superposing contours may be used to weight the correction terms (reinforcing existing maxima e.g.). Knowledge about salient prosodic events associated with some function could be used to attribute a privilege to a contour in (2).

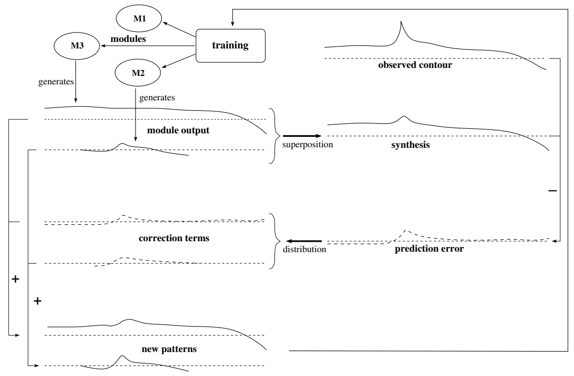


Figure 3: analysis-by-synthesis-scheme

3.3. Enunciation of mathematical formulae

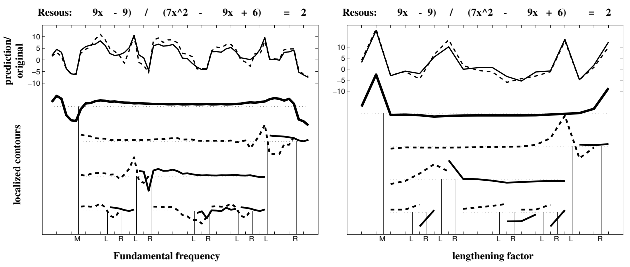
The predominant role of prosody in the enunciation of MF is segmentation/hierarchisation, i.e. to signal the nesting of operators. There are two kind of operators in our corpus: unary ones (such as $\sqrt{\ldots}$) or binary (such as $+$).

Analysis of the observed performance structures suggests that the intonation follows syntactic hierarchy. Thus we distinguish two types of dependency relations: left (L) between an operator and its left-hand operand and right (R) between an operator and its operand on the right. Inspired by syntactic markers used in literature (e.g. [4, 2]) we put a marker L or R at each frontier between two leaves of the syntactic tree. Two generation modules are charged to produce all instances of elementary contours depending on the size of the associate units and the position of the marker.

We provide a third module (M) in order to encode the relation between the introductory word "Résous" ("Solve") and the equation itself. The corresponding contour is coextensive to the entire utterance.

4. OBSERVATIONS AND EVALUATION

Figure 4 shows an example of generated contours for a rather simple formula. The predicted contours are only slightly flatter than the original ones, all accents and final lengthenings are well-predicted. The four lines in the bottom half of the plots show all contributing contours. Looking closer at the contours generated by the left dependency module (L, dashed lines) illustrates how the model captures movement expansion[1]: the F_0-accent marking the corresponding frontier appears and grows with the length of the preceding unit. A similar behaviour can be observed in the contours for the lengthening factor. RMS-errors and correlations for the F_0-contours are given in figure 4. We divided each corpus into two equally sized parts A and B. Training was performed either on the entire corpus (T) or on one of the parts. No effort has yet been put on the optimisation of the stop criterion which might explain the slight over-learning with the training set A.



test	training		
	T	A	B
T	2.02	1.97	2.07
A	2.11	1.85	2.34
B	2.07	2.14	2.01

(a) RMS-error

test	training		
	T	A	B
T	0.90	0.91	0.90
A	0.90	0.92	0.87
B	0.90	0.89	0.90

(b) correlation

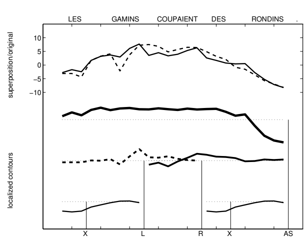
Figure 4: Top part of the plots: Predicted(full line) and original(dashed) contours (F_0 in half-tons with respect to 126Hz); bottom part: contours contributing to the prediction – anchor points and the associated generation modules are given (M, L, R), dotted lines indicate zero for each contour. The table shows RMS-errors and correlations for different training/test-configurations

The strength of the proposed model is its ability to exploit the recursivity of the syntactic structure by using the same generation modules on all hierarchical levels which is particularly efficient for the enunciation of MF. Nevertheless, it provides good prosody predictions when trained on ordinary speech corpora. In figure 5 we show an example for a French sentence uttered in four different attitudes (assertion, evidence, incredulity and yes/no-question). The RMS-errors show – as for the maths corpus – fairly good generalisation properties (e.g. training on A – test on B). Best scores are obtained for yes/no-questions due to the fact that the contours are rather flat. Nevertheless, results for the intonationally more structured assertions are satisfying. Figure 6 shows the movement expansion of the IQ-contours. Our observations are fully compatible with Morlec [11].

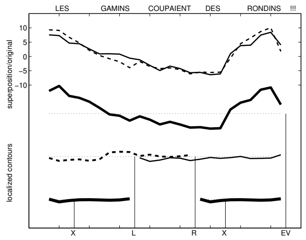
5. CONCLUSIONS AND OUTLOOK

Compared to the prior work of Aubergé/Morlec the model presented here offers several advantages: (1) It is now possible to exploit naturally the deep recursivity of dependency relations in the maths corpus – the model architecture does *not* depend anymore on the number of nesting levels aimed to capture. (2) The relation between linguistic functions to encode and the generation of contours is straightened – *one* module is charged to encode *one* function. The model offers thus a *dynamic lexicon*: contours are indexed by the linguistic function and generated dynamically in order to fit phonotactic needs. (3) Contours are localised where they are meant to act i.e. on the discourse units they relate.

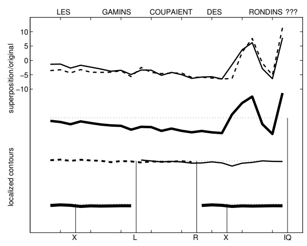
[1] see also figure 6 and explanation below



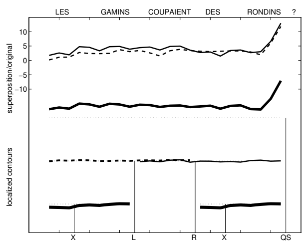
(a) assertion (AS)



(b) evidence (EV)



(c) incredulity (IQ)



(d) yes/no-question (QS)

AS	T	A	B
T	1.68	1.67	1.69
A	1.82	1.59	2.01
B	1.81	1.96	1.64

EV	T	A	B
T	1.76	1.76	1.76
A	2.00	1.68	2.27
B	1.91	2.17	1.63

DI	T	A	B
T	1.31	1.34	1.29
A	1.39	1.32	1.45
B	1.35	1.49	1.20

QS	T	A	B
T	1.20	1.20	1.19
A	1.28	1.15	1.39
B	1.31	1.47	1.14

Figure 5: (a–d): F_0-contours for the sentence "Les gamins coupaient des rondins" in four different attitudes. Bottom: RMS-errors for the different corpora.

Further work is needed to refine the error-distribution step of the model-inversion-algorithm and to optimise the learning of individual modules (number of hidden units, stop criterion) in order to obtain even better generalisation properties.

Our main interest is to furnish a perceptual evaluation of the model. We are working on an experimental protocol aiming to compare different prosodic realisations with respect to the online processing of a listener. We will analyse the lag between natural vs synthetic speech and handwriting events in a dictation task.

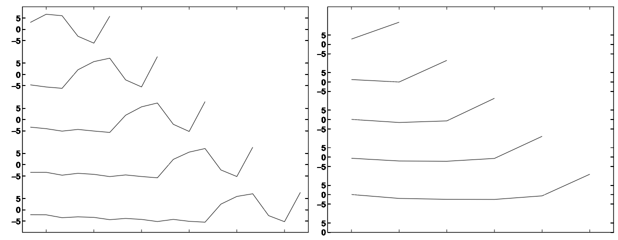


Figure 6: Movement expansion for contours encoding incredulity for sentence lengths varying from 2 to 6 syllables (left: F_0, right: lengthening factor).

REFERENCES

1. Aubergé, V. Developing a structured lexicon for synthesis of prosody. In Bailly, G. and Benoît, C., editors, *Talking Machines: Theories, Models and Designs*, pages 307–321. Elsevier B.V., 1992.

2. Bailly, G. Integration of rhythmic and syntactic constraints in a model of generation of French prosody. *Speech Communication*, 8:137–146, 1989.

3. Black, A.W. and Hunt, A.J. Generating f0 contours from tobi labels using linear regression. In *Proceedings of the International Conference on Speech and Language Processing*, pages 1385–1388, 1996.

4. Emerard, F. Synthèse par diphones et traitement de la prosodie. Thèse de troisième cycle, Université de Grenoble III, Grenoble, France, 1977.

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

6. Gee, J.P. and Grosjean, F. Performance structures: a psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15:411–458, 1983.

7. Hirst, D.J. and Di Cristo, A. *Intonation systems: a survey of twenty languages*. Cambridge University Press, Cambridge, 1998.

8. Holm, B., Bailly, G., and Laborde, C. Performance structures of mathematical formulae. In *Proceedings of the International Congress of Phonetic Sciences*, volume 2, pages 1297–1300, San Francisco, USA, 1999.

9. Malfrère, F., Dutoit, T., and Mertens, P. Automatic prosody generation using suprasegmental unit selection. In *ESCA/COCOSDA Workshop on Speech Synthesis*, pages 323–328, 1998.

10. Möbius, B., Pätzold, M., and Hess, W. Analysis and synthesis of german f0 contours by means of fujisaki's model. *Speech Communication*, 13:53–61, 1993.

11. Morlec, Y., Bailly, G., and Aubergé, V. Synthesising attitudes with global rhythmic and intonation contours. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 1, pages 219–222, Rhodes - Greece, 1997.

12. Pasdeloup, V. A prosodic model for french text-to-speech synthesis: A psycholinguistic approach. In Bailly, G., Benoît, C., and Sawallis, T., editors, *Talking Machines: Theories, Models and Designs*, pages 335–348. Esevier B.V., 1992.

13. Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. Tobi: a standard for labeling english prosody. *Proceedings of the International Conference on Speech and Language Processing*, 2:867–870, 1992.

14. Thorsen, N.G. Standard Danish sentence intonation – Phonetic data and their representation. *Folia Linguistica*, 17:187–220, 1983.

15. Traber, C. Fo generation with a database of natural fo patterns and with a neural network. In Bailly, G. and Benoît, C., editors, *Talking Machines: Theories, Models and Designs*, pages 287–304. Elsevier B.V., 1992.