# A Gestalt Morphology of Prosody Directed by Functions: the Example of a Step by Step Model Developed at ICP.

*Véronique Aubergé*

Institut de la Communication Parlée, UMR CNRS 5009, Grenoble, France
auberge@icp.inpg.fr

## Abstract

This paper proposes a point of view consisting in showing prosody as an emergent form perceived via Gestalt processing. Contours carry some function values at different levels through the superposition of independent contours. In this view, tonal phonology can be interpreted as a bottom-up sub-processing approach, and global form modelling as supported by a top-down approach. The model can integrate a large scale of functions from the linguistic domain to the expression of emotion. The founding principles of the model are explained and the most significant steps developed at ICP are recalled to illustrate this theoretical background.

## 1. Introduction

This paper summarises a complete model of prosody, which has been successively implemented and validated for French. The principles, on which this model is based on, are first recalled, and then the main steps of development are traced in order to give a global and coherent view of this long period of proposals (from 4 to [2]). The methodology followed for this whole work is hypothetico-inductive (see figure 1):

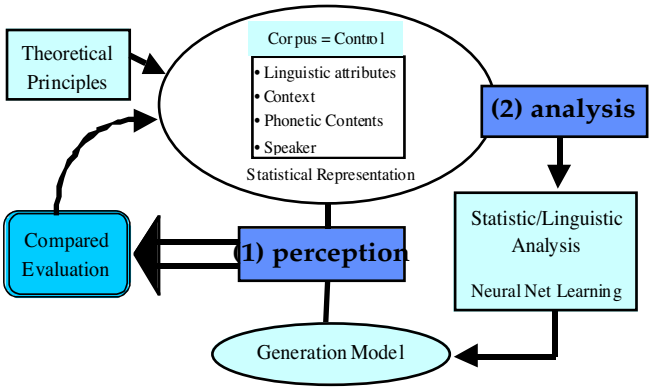


***Figure 1:*** *The methodology followed to develop the ICP prosody generation model (from Morlec et al, 99)*

This is why from the theoretical principles (described in §2.1.) some large corpora are first built, dense in terms of principle representativity. Then (1) perception experiments are held to validate/evaluate these principles (2) some simulation models are built from the data, driven by the principles, finally (3) the compared evaluation returns to (1).

## 2. The model developed at ICP

### 2.1. The model principles

The theoretical hypotheses on which the model is based are listed in Table 1.

*Table 1: The five hypotheses of the ICP model*

| **Principle 1** |
|---|
| • linguistic, pragmatic and emotional functions drive the communication system, the agents – prosody being one of them – *co-operate* following meaningful strategies to carry out the function values;<br>=> structural *rendezvous* between the agents |
| **Principle 2** |
| • cognitive processing of prosody is based on global movements in a given linguistic domain (the segment which carries functions values), in perception and production;<br>=> the phonological unit of prosody is a global emergent contour, which accesses an associated function value. |
| **Principle 3** |
| • each level delimited by the segmentation/hierarchisation function, and which receives some functional values (demarcation, modalisation, focalisation, attitude), gets a morphology independent of the level height and of other levels;<br>=> superposition of contours, independence of the contours between levels |
| **Principle 4** |
| • several function values (e.g. demarcation + focalisation) can be given to the same segment, the associated morphology is the superposition of associated contours on the same segment at the same level. ;<br>=> superposition of multiple independent function value contours |
| **Principle 5** |
| • the emotional function is shared by prosody and other agents (like facial gestures); these function values are associated with prosodic patterns which are not controlled on a linguistic segment domain, but constrained by the segment domains that do not destroy the linguistic values of prosody. |

### 2.2. Principle 1

This first hypothesis derives from several sub-hypotheses (Figure 2).

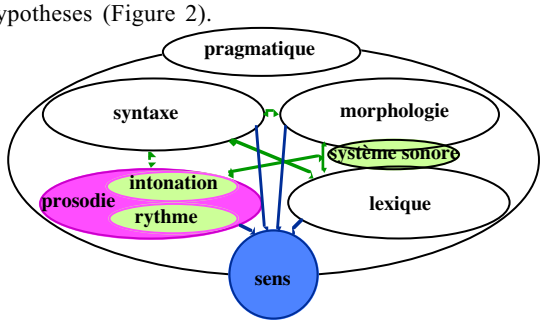


*Figure 2: some functions are global to the system, shared by the modular agents, following relevant strategies*

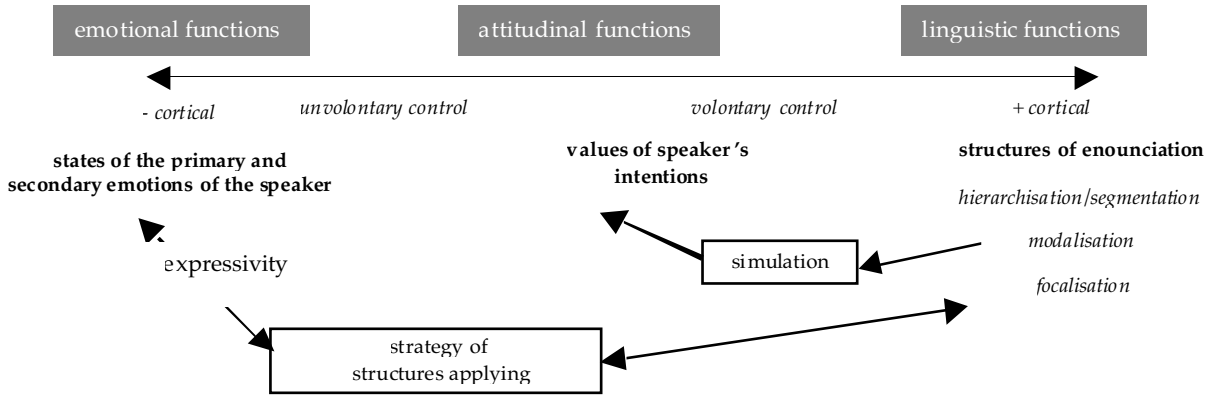


***Figure 3:*** *Which functions for prosody?*

First the communication system, driven by communication goals, uses a set of functions which are valued globally to the system. Second, the system is a set of modules, not in a Fodor organisation, but in interactive organisation based on co-operation between the modules, typically in a multi-agent architecture: the specific constraints and degrees of freedom of each module can consequently be respected. The coherence of several modules, when they encode the same function values, is made by a rendezvous between the different agent structures for a given function value.

It could explain for example the bootstrapping effect from prosody to syntax observed in developmental studies [17] in considering the segmentation/ hierarchisation function as shared by syntax and prosody. Third, the instantiation of the function, the emergent result of this co-operation, is both characterised by its values and by the strategy in it is shared between modules (the structure and the structure of the structure), which is usually rejected as a non meaningful style effect. Finally these functions are very basic (see figure 3). While these functions can trivially be separated into three groups (linguistic, pragmatic, emotional), there is surely no strict boundary between the three domains.

In the first group, the segmentation/hierarchisation function is essential since it determines the domains on which the other linguistic and pragmatic functions can be applied. This function ensures the basic linguistic intelligibility of the prosodic material, it has been noted for many languages [6]. The role of prosody in front of morpho-syntax, in taking in charge of this function, is dependent on many parameters of the communication situation. In read texts for example, typical material to be simulated in TTS, the redundancy with syntax is surely maximised. In all cases the coherence between both agents is over the specific instantiation of this function for each agent. This function was the first implemented in the model [1] and was later measured in perception in conditions of prosody alone [10]. In this study, it was observed in particular that some rendezvous seem to be obligatory, some others are never taken by prosody, some others optional. The location of some rendezvous in prosody can move (as already proposed by Campbell [3]).

Focalisation gives values to some segments delivered by the segmentation function. It can be sub-divided into the deixis function and the emphasis function, which are shared by all the agents, and surely over the language system (the deixis function is given by some authors as a primitive of communication – first by the eye, the finger). The strategy chosen to encode such a function between the different agents characterises more than social features and surely the expressivity function. A current study aims to show which kind of information is carried by encoding strategies.

The attitude function adds values about the intention of the speaker to the linguistically encoded content. Again prosody is not alone to encode this function. In figure 3, this function is clearly separated from the emotion expression, because it is supposed that this function is applied to the enunciation domain, that is on segments delimited by the demarcation function. This is contrary to emotion values which are, according to this hypothesis, carried on non-linguistically identified segments. This is what was noticed when modelling attitudes for French [8] *vs*. one emotional value [2]. Some other indices can be taken in the developmental field: it seems that some attitudinal contours are learned early (between 7 and 11 years)[8]. We could not find any data concerning the simulation of emotion which depends, like attitude, on voluntary control (and not on involuntary control like in the expression of emotion), in this particular case, following our hypothesis the expression of the simulated emotion is expected to map onto linguistic segments.

### 2.3. Principle 2

Our model belongs to the class of global approaches, following Delattre [10] or Fónagy [11] for French. Global approaches are generally opposed to tonal approach descriptions. On this point, we think that the controversy between tonal and global must first be related to the implicit questions answered by these two kinds of approaches [4].

If the aim is to answer the question "what is prosody used for?" (figure 1), the morphology of prosody is the organisation of the prosodic material (whatever the relevant signals studied) specifically relevant to carry out the given function values. If the aim is to answer"what is prosody built from?", the morphology of prosody is directly built from the signal.

There is a paradox, in this view, since a phonetic approach such as the one proposed here, is clearly top-down, directed by function, whereas the so called phonological approach, the tonal approach, is an early level of symbolisation, close to the signal level, precisely because the tonal representation is aimed at abstracting from the physical world. Most tonal models, in French (Hirst, Mertens) or other languages (Gussenhoven, Pierrehumbert and Beckman) propose anyway to access linguistic values by a construction of contours of tones, or/and by salient tones.

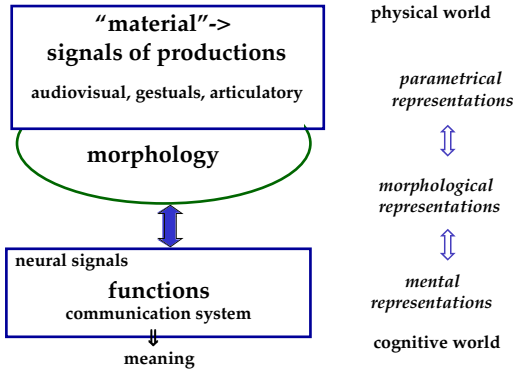


*Figure 4: bottom-up vs. top-down approaches to prosody.*

This is a "second articulation" access. In the global approach proposed here, the contour is the phonological unit, which is not processed as a concatenation of minimal contours (like for example the IPO approach developed for French by [4] or like Vaissière's model).

This does not mean that the contour could not be processed into sub-phonological units, like tones, but the tones could just be considered then as the "stones" from which a global contour *emerges* (in the sense of Morgan). Some tones can be more salient (like key stones in a bridge), but the contour is not the result of a combination, a symbolic calculation – a grammar – of tones. The tones are only strong indices in the Gestalt processing of perception involved by the global approach. Already in 1983, Grosjean showed that, on the basis of rhythmic indices, listeners could predict the length of truncated utterances. Thorsen [23], van Heuven et al. [13] and Grépillat et al. [11] got similar results for different languages. This is why listeners can predict its modality or attitude value early in an utterance in a gating paradigm (for French attitudes, the prediction is efficient at the second syllable of five-syllable utterances). It can explain for example why a "jigsaw-like" generation system like Chatr obtains such contrastive results in evaluation, when the selected or not selected salient pieces can imply or not imply access to the global pattern (like missing or not-missing ears can be strong predictors for access to the donkey vs. horse Gestalt).

Following this Gestalt hypothesis, the question is then whether these contours are organised into categories, which can be phonologically represented by prosodemes, whether a "good form" can be described in terms of production/perception parameters, whether the opposition between two categories can be described in the same terms, and whether the space of the variants of the same category can be characterised.

### 2.4. Principle 3

The Gestalt hypothesis is not enough to describe the complexity of the function values to be carried. Since these belong to different levels (as given by the segmentation function), each level is associated with a set of Gestalts and is described independently of the others. A classical superposition process (addition in time) is applied to all contours of all levels to analyse/generate the utterance pattern. This notion of independence between levels is quite original compared to other superposition models: the same contour can appear at several levels. At a given level, a carrying contour carries some carried contours which may include the same morphology as the carrying.contour, but applied on an lower domain. The "function driven" superposition makes such a model very different, for example, from Fujisaki's model, since the physiological characteristics appear in our model only as constraints in the superposition process.

The model was first implemented using a first order statistical treatment, carried out on sets of "mean-contours" (see[1]). A more sophisticated calculation method (second order function in a connectionist network) carried out a set of prototypes (see [17]). Both models simulated efficiently the utterances of the original corpus by applying the first three principles. Recently, Holm and Bailly [15] have applied this to a corpus of read mathematical formulae, in utterances where the segmentation function (mathematical symbols) exactly coincides with the semantic content, and where this content can occur at whatever level. Their results confirm the independence hypothesis.

For now, the implementation is simply the superposition of stored prototypical contours. But in observing data and its perceptual efficiency (for lexical focalisation [6]), it was noted that some variants are not the consequence of variants inside a given class, but the consequence of a varying "weight" associated with the superposition. It would be necessary to complete the model with such a parameter, as an input parameter.

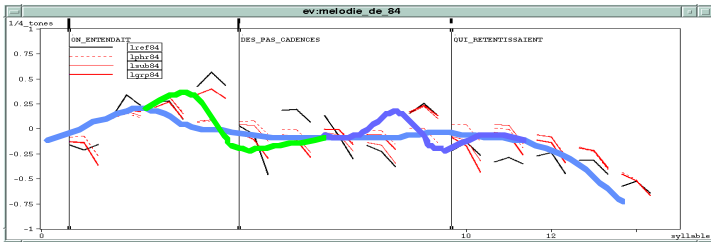


*Figure 5: the superposition process (from [Morlec, 98])*

### 2.5. Principle 4

As proposed by the previous hypotheses, a contour is global to a given domain (defined by the segmentation function). This means that for a given function value, (domain), the contour is applied as a whole to the segment, even if it can be considered as local at a higher level. Since different functions are valued in the same time, they can coincide on the same domain (e.g., some segmentation/hierarchisation, deixis, and attitude values can be associated on a nominal group). In such a case, the superposition function is generalised to the same domain. In generation, the processing can be deduced from the superposition processing, but for analysis, the solution can be ambiguous.

### 2.6. Principle 5

Emotions are carried in two ways: directly via emotional expressions and indirectly via expressivity, that is using linguistic structures. While the role of prosody in emotional expression is still not well known ([23][20]), these can be supposed to be timed by emotional events and constrained by "linguistic prosody". This was verified for one specific emotion [2], and some current studies

confirm this hypothesis. The problem is to understand how linguistically emotionally timed contours are integrated: do they just use different parameters? (e.g. it seems that amusement for French speakers uses mainly intensity which is poorly used for linguistic encoding [2]), but this would be contradictory with current hypotheses of innate/universal emotion expression), or is it contour discrimination processing?

## 3. Conclusion

This paper spells out the theoretical hypotheses of a model developed in different ways (like TTS) at ICP. The only experiment which could validate the hypothesis of global contour processing, and make the role of tones in this processing apparent, would be to demonstrate a clear categorial perception effect. The generation module is under implementation in a multi-agent architecture, which would carry out a new kind of TTS able to model the strategies of speakers. Adding a weight parameter will certainly increase the sophistication of the superposition model.

Finally the main progress will surely come from ongoing studies on emotional speech, which should help us understand, following Damasio, how emotion can condition the linguistic processing of prosody.

## 4. Acknowledgement

Those collaborating in implementing and validating the model were very numerous but special thanks are due to Gérard Bailly, Yann Morlec and Albert Rilliard, and to many, many students.

## 5. References

[1] Aubergé, V., 1992. Developing a structured lexicon for synthesis of prosody. In Bailly & Benoit, eds., *Talking machines*. Elsevier.

[2] Aubergé, V.; Cathiard, 2001. Can we hear smile? *Speech Communication*, submitted.

[3] Aubergé V.; Grépillat, T.; Rilliard, A., 1997. Can we perceive attitudes before the end of sentences? A gating paradigm for prosodic contours. *Proceedings Eurospeech*, Rhodes.

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

[5] Beaugendre, F,, 1994. *Une étude perceptive de l'intonation du français. Développement d'un modèle et d'une application à la génération automatique de l'intonation pour un système de synthèse à partir du texte*. Doctoral thesis, Univ. Paris XI, Paris France.

[6] Brichet; Aubergé, V., 2001. La focalisation en français : morphologie de la prosodie. *Actes des Journées Prosodie*. Grenoble, France.

[7] Campbell, N., 1993. Automatic Detection of prosodic boundaries in speech. *Speech Communication*, 13, 343-354.

[8] Clément, J., 1999. *Structure des représentations prosodiques. Développement normal et pathologique du traitement de la prosodie*. Doctoral thesis, Univ. de Paris V.

[9] Couper-Kuhlen, E.; Selting, M., 1996. *Prosody in conversation*. Cambridge: Cambridge University Press.

[10] Delattre, 1969. L'intonation par les oppositions. *Le français dans le monde*, 64, 6-12.

[11] Fónagy, I.; Bérard, E.; Fónagy, J., 1983, Clichés mélodiques . *Folia Linguistica*, 17, 153-185.

[12] Aubergé, V.; Grépillat, T.; Rilliard A., 1997. Can we perceive attitudes before the end of sentences? The gating paradigm for prosodic contours. *Proceedings of EuroSpeech'97*, Rhodos, Greece, 2, 871-877.

[13] van Heuven V.; Haan J.: Janse, E.; van der Torre, E., 1997. Perceptual Identification of sentence type and the time-distribution of prosodic interrrogativity markers in Dutch. *Proceedings of ESCA workshop on intonation*, 317-320, Athens.

[14] Hirst, D.J.; Di Cristo, A. (eds), 1998. *Intonation systems. A survey of twenty languages*. Cambridge: Cambridge University Press.

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

[16] Lacheret, A.; Beaugendre, F., 1999. *La prosodie du français*. Paris: Editions du CNRS.

[17] Morgan, J.L.; Demuth, K. (eds.), 1996. *Signal to Syntax: Bootstrapping from speech to grammar in early acquisition*. Mahwah: L. Erlbaum.

[18] Morlec, Y., 1998. *Génération multiparamétrique de la prosodie du français*, Doctoral thesis, INPG Grenoble.

[19] Morlec, Y.; Rilliard, A.; Bailly, G.; Aubergé, V., 1998. Evaluating the adequacy of synthetic prosody in signaling syntactic boundaries: methodology and first results, *Proceedings 1st LREC,* Grenade.

[20] Mozziconnaci, S., 1998. *Speech Variability and emotion: Production and Perception*, Doctoral Thesis, Eindhoven University,

[21] Rilliard, A., 1999. Prosody diagnostic using reiterant speech, *Proceedings ICPhS*, 37-40.

[22] Rilliard, A.; Aubergé, V., 2001. Prosody evaluation as a diagnostic process: subjective vs. objective measurements, *Proceedings 4th ISCA Workshop on Speech Synthesis*, Atholl, Scotland.

[23] Banse, R.; Scherer, K. R., 1996. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70, 614-636.

[24] Thorsen, N., 1980. A study of perception of sentence intonation - evidence from Danish. *J. Acoust. Soc. Am.* 67 (3), 1014-1030.