



# SFC: A trainable prosodic model

G erard Bailly \*, Bleicke Holm

*Institut de la Communication Parl ee, UMR CNRS no. 5009, INPG/Universit  Stendhal, 46, av. F elix Viallet,  
38031 Grenoble Cedex, France*

Received 21 July 2004; received in revised form 31 March 2005; accepted 3 April 2005

---

## Abstract

This paper introduces a new model-constrained and data-driven system to generate prosody from metalinguistic information. This system considers the prosodic continuum as the superposition of multiple elementary overlapping multiparametric contours. These contours encode specific metalinguistic functions associated with various discourse units. We describe the phonological model underlying the system and the specific implementation made of that model by the trainable prosodic model described here. The way prosody is analyzed, decomposed and modelled is illustrated by experimental work. In particular, we describe the original training procedure that enables the system to identify the elementary contours and to separate out their contributions to the prosodic contours of the training data.

  2005 Elsevier B.V. All rights reserved.

*Keywords:* Intonation; Prosodic modelling; Automatic generation of prosody

---

## 1. Introduction

It is a commonly accepted view that prosody crucially shapes the speech signal in order to ease the listener's task of decoding linguistic and paralinguistic information. In the framework of automatic prosody generation, we aim to compute adequate prosodic parameters carrying that information. We thus consider here prosodic models

that are able to automatically compute prosodic parameters from linguistic (more precisely from syntactic, phonological and phonotactic) specifications in the context of speech synthesis. While a few systems dispense with a phonetic description of prosody by incorporating the linguistic specifications directly into the selection process (Taylor and Black, 1999), most speech synthesis systems use specific prosodic models that compute fundamental frequency ( $f_0$ ), phoneme durations or energy profiles that are used to alter prosody of selected original speech units and also to select them (Fujisawa and Campbell, 1998).

---

\* Corresponding author.

E-mail addresses: [bailly@icp.inpg.fr](mailto:bailly@icp.inpg.fr) (G. Bailly), [holm@icp.inpg.fr](mailto:holm@icp.inpg.fr) (B. Holm).

These prosodic models are generally built using machine learning with annotated corpora. Syntactic, phonological, phonotactic and phonetic descriptors are collected for each unit (generally the phoneme or the syllable). Model-based (e.g., regression trees, HMMs, neural networks) or sample-based (e.g., vector quantization, contour selection) mapping tools are then used to achieve the best phonetic prediction according to a distance metric, generally root mean square (RMS) error. Prediction is generally performed with separate trainable models for  $f_0$  (Ljolje and Fallside, 1986; Scordilis and Gowdy, 1989; Sagisaka, 1990; Traber, 1992), for phoneme durations (Klatt, 1979; O’Shaughnessy, 1981; Bartkova and Sorin, 1987; Riley, 1992; van Santen, 1992) and, more recently, for intensity profiles (Trouvain et al., 1998). With the development of corpus-based synthesis techniques and powerful mapping tools (Campbell, 1992; van Santen, 2002), multiparametric prosodic models (Mixdorff and Jokisch, 2001; Tesser et al., 2004) now tend to use general-purpose and theory-neutral tools. Most trainable prosodic models consider syntactic, phonological, phonotactic and phonetic descriptors to simply be possible factors influencing the prosodic realization of a certain phoneme given the speaker and the communication situation (which is often reading). The mapping tools are therefore responsible for evaluating the contributions of these factors and their interactions. Models of interaction range from additive, multiplicative, sum-of-products (van Santen, 1992) models to more complex non-linear models such as neural networks. When processing an utterance sequentially, some models also sometimes incorporate intermediate predictions made for earlier units (e.g. through recurrent connections as in Traber, 1992). These models are in a sense theory-neutral since the interaction is solved implicitly at a quantitative level by intensive training of the mapping tools and not by high-level comprehensive models of intonation (Bailly, 1997) that specify which factors influence the prosodic realization of each phoneme and what is effectively the scope of their action.

We present here a trainable prosodic model, the SFC (*superposition of functional contours*) model

(Holm and Bailly, 2002; Bailly and Holm, 2002), which implements a theoretical model of intonation. This model, initiated by Aubergé (1992, 1993), 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. Our strong hypotheses are that (1) these functions are directly implemented as prototypical prosodic contours that are coextensive with the unit(s) to which they apply,<sup>1</sup> (2) the prosody of the message is obtained by superposing and adding all the contributing contours.

The SFC is presented in the following sections. Section 2 describes the phonological model that specifies which metalinguistic functions are to be realized by prosody and on which phonological units they apply. This deep phonological model provides symbolic inputs. Section 3 presents the phonetic model used to describe observable prosodic continuums. The phonetic model provides a constant number of parametric targets for each syllable of the analyzed utterances. Section 4 describes the model that maps symbolic inputs delivered by the phonological model to parametric outputs delivered by the phonetic model via a superposition of prototypical parametric contours. The mapping model essentially consists of training several contour generators (one per metalinguistic function) using an original analysis-by-synthesis loop. A block diagram showing the different steps of the training process and the generation procedure is presented in Fig. 1. Section 4.3 summarizes the experimental results obtained with different linguistic material.

<sup>1</sup> Note that the domain of action of the prototypical prosodic contour—and thus of the function it implements—is not restrained to the focus of action but comprises its domain of influence. Anchor points further specify key points such as the focus and span.

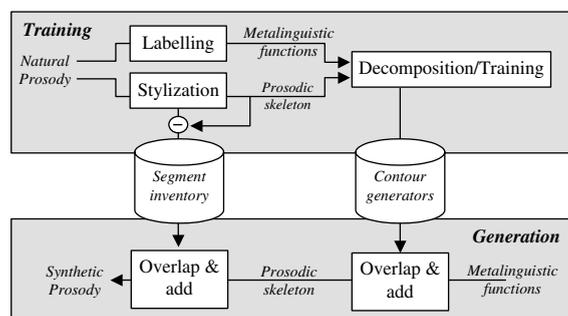


Fig. 1. Training the SFC consists in decomposing prosodic skeletons (stylized prosodic contours) of a set of training utterances into overlapping multiparametric elementary contours associated with the metalinguistic functions they recruit. One contour generator is responsible for generating multiparametric elementary contours associated with one metalinguistic function given its scope (the phonological units the function is applied to). The prosodic flesh obtained by subtracting the prosodic skeleton to the original prosody is stored into the characteristics of the acoustic segments used by the concatenative speech synthesizer. The generation process consists simply in the selection of the appropriate contour generators implementing the various metalinguistic functions recruited by the discourse and overlap-and-add them with the prosodic flesh of the segments used. Note that the generation process is also used during the training phase for decomposing prosodic skeletons into multiparametric elementary contours (see Fig. 6 and Section 4.2).

## 2. The phonological model

As stated by Cutler and Norris, “prosody is as much involved as any other aspect of linguistic structure in speakers’ efforts to do their part in achieving this goal [maximizing the successful transmission of the message]... both salience and segmentation figure in prosodic contributions to realization of the speaker-listener contract” (Cutler and Norris, 1991, p. 267). Other contributions of prosody to the ease of discourse interpretation include, of course, communicative values associated with each salient or segmented unit such as contrastive emphasis on phonemes or syllables, lexical stress, emphasis on words,<sup>2</sup> sentence modality or speaker’s attitude. Prosody is also a means of

<sup>2</sup> The most salient feature of these last three different focus is a pitch accent on the stressed syllable. Brichet and Aubergé (2004) have shown however that pitch contours around this landmark are quite different according to the focus type.

voluntarily or involuntarily signaling our cognitive activity, our psychological and emotional states as well as idiosyncratic and sociolinguistic information.

But, as stated by Hirst (2003), most current accounts of prosody function within prosodic annotation systems deal with prominence and boundaries (Wightman et al., 2000) aggregating often under identical symbols very different functions. Within the framework of non-linear phonology, prominence and boundaries apply and delimit embedded constituents such as rhythmic, tonal and intonation units. This strict layer hypothesis (Selkirk, 1984; Nespor and Vogel, 1986) is however questioned by a number of studies that claim the necessity of adding scopes/domains to the labeling of prominences and boundaries in order to account for the embedded (Marsi et al., 1997) and possibly recursive phonological hierarchy (Hirst et al., 2000; Schreuder and Gilbers, 2004).

Instead of considering a posteriori the mapping between linguistic units and constructs of prosodic phonology (both being determined separately), we consider on the contrary that the general ability of prosody to highlight and segment speech units is used mainly to encode discourse structure. The domain of action of prosody within the SFC model is *the linguistic domain* and the linguistic structure provides prosody with the specification of its tasks as *triplets* (metalinguistic function, units, importance). We consider in the following that all metalinguistic functions involved in the discourse have the same importance but will comment on the possible use of the third member of the triplet as a term of *gradience* in Section 6.

Note finally that SFC does not make any use of an intermediate representation of prosody. Prosody is assumed to *directly* encode deep phonological structure by phonetic events—here overlapping multiparametric contours—without any intermediate surface representation (see Hirst, 2003, for a further consideration of these levels of prosodic analysis).

### 2.1. Metalinguistic functions

As discussed above, metalinguistic functions, including segmentation, hierarchisation, emphasis

and attitude—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 Bolinger, 1989) is quite open: the examples given in the following give possible candidates for such functions and provides the reader with concrete examples from the current implementation of the model.

The functions we have considered so far are: prosodic attitudes applied to sentences (Morlec et al., 2001), dependency relations applied to syntactic constituents of read text (Morlec et al., 1998) or operands/operators of spoken math (Holm et al., 1999), cliticization typically applied to determiners and auxiliaries (Bailly and Holm, 2002), narrow focus applied to words (Brichet and Aubergé, 2004) and, more recently, tones applied to syllables in Mandarin Chinese (Chen et al., 2004).

In our work, metalinguistic functions responsible for giving cues to the syntactic structure of sentences in the discourse signal dependency relations between chunks (Bailly, 1989; see also Bachenko and Fitzpatrick, 1990; Pynte and Prieur, 1996, for dependency structure analysis and prosodic correlates of attachment/branching of syntactic constituents). We consider four kinds of dependency relations that may link constituents (words, groups, phrases, clauses): left dependency (DG, *dépendance à gauche*) linking the head of a sub-tree (the “governor” or “mother”) with its immediately preceding dependent unit (“sister”), right dependency (DD, *dépendance à droite*) 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. The syntactic parse we use is thus very simplified and can be accomplished using a chunk and chunk technique (Balfourier et al., 2002; see also Di Cristo et al., 2000, for the use for syntax/prosody mapping). For instance, the sentence labelled in Fig. 2 is parsed as in (1.a), is rewritten as in (1.b) and is finally processed (once the first two right dependencies linking clitics to chunks have been erased) as in (1.c).

(1.a) [s[<sub>CP</sub> Dès[<sub>CP</sub> que [v<sub>P</sub> [n<sub>NP</sub> le tambour] bat]],  
[v<sub>P</sub> [n<sub>NP</sub> les gens] accourent]]

(1.b) [[[[Dès]<sub>DD</sub>[[que]<sub>DD</sub>[[le tambour]<sub>DG</sub>[bat]]]]]<sub>IT</sub>  
[[les gens]<sub>DG</sub>[accourent]]]

(1.c) [[Dès que le tambour]<sub>DG</sub>[bat]]<sub>IT</sub>[[les gens]<sub>DG</sub>  
[accourent]]

Translation: ‘As soon as the drum starts, the people rush.’

The suggested list of metalinguistic functions is not extensive and the SFC is not tied to a particular linguistic theory of discourse: the SFC is tied to a phonological model that supposes that the selected metalinguistic functions are directly implemented as global multiparametric contours. These functions may apply to a single unit (e.g., modality operates at the level of the sentence, narrow focus can operate at the level of the word) but they are usually applied to two units (e.g. dependency relations between syntactic constituents, determiner and determined for cliticization, operator and operand for spoken math). For now, the number of units is limited to two, with the restriction that the units should be adjacent. This reduces to three the number of anchor points of the functional contour that will be applied to the units: beginning of unit 1, boundary between unit 1 and unit 2 and end of unit 2 (see Section 4.1 and Fig. 4 for more information on how these landmarks are used as inputs for contour generators). There is, however, no technical reason for these restrictions, and the number of units could be increased by adding additional anchor points.

## 2.2. Annotating corpora

SFC is a trainable prosodic model. The phonological model specifies qualitatively which metalinguistic functions are considered and on which discourse units they act. Training of SFC consists of quantitatively determining how these specifications are implemented by prosodic contours. Corpora should be designed and recorded in such a way that sufficient material is available to gather enough tokens of each metalinguistic function applied to discourse units of various sizes. The corpora are often situation-specific and always speaker-specific. Note however that prototypical functional contours provided as a by-product of

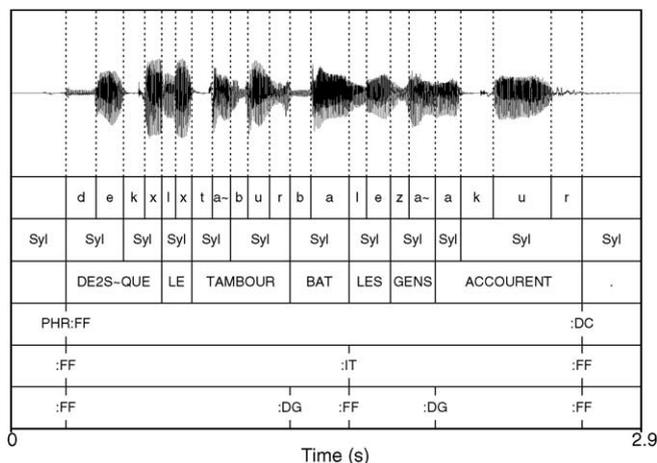


Fig. 2. A speech sample ('As soon as the drum starts, the people rush.') for the SFC, labelled using Praat. The [:FF] tag delimits the beginning and end of units. Other tags are used to label the metalinguistic functions [:DC] for declarative sentence, [:IT] between clauses, [:DG] between nominal and verbal groups.

SFC training (see Section 4) may constitute a valuable tool (see Section 6) for cross-speaker and cross-language studies.

*Building corpora.* We constructed several corpora, some of which were constructed by sentence-generators to exclusively study the implementation of specific metalinguistic functions (as in Morlec et al., 2001; Holm et al., 1999), and others of which were constructed by greedy algorithms (Chen et al., 2004) to ensure optimal coverage of the samples. The number of free parameters of the selection problem is small: typically a few metalinguistic functions applied to units from one to several syllables in length. The generalization abilities of the SFC model are quite good (see Fig. 5) and a few dozen sentences are generally sufficient to have comparable prediction errors between training and test sets: 100 sentences in Chen et al.'s experiment (2004), 104 spoken maths in Raidt et al.'s comparative study (2004).

*Technical issues.* Praat (Boersma and Weenink, 1996) is used to edit phonological scores (functions and units) and phonetic content (see Fig. 2). Label files are first generated using automatic segmentation, syllabification and syntactic parsing using a dependency grammar (Bailly, 1989) and then, if required, corrected by hand.

### 3. The phonetic model

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 on the vocalic nucleus as initially proposed by Tournemire, 1997) and a lengthening factor (that will stretch or compress all phonemic segments of that syllable using z-scoring, see below). These four values gathered for all syllables of the utterance build melodic and rhythmical contours, which constitute a sort of prosodic *skeleton* of the utterance.<sup>3</sup> Pursuing this metaphor of the body, this prosodic skeleton is *articulated*, i.e. built by elementary prosodic segments that are superposed. The phonetic model is responsible for giving *flesh* to this prosodic

<sup>3</sup> Note that this prosodic *skeleton* can be enriched within the SFC framework by other multiparametric contours as long as they characterize each syllable by a constant number of parameters. Candidate parameters are loudness and spectral tilt. We also currently working on the prediction of head movements sampled at syllable onsets.

skeleton, i.e. computing final prosodic parameters ( $f_0$  contour, phone durations) from the skeleton.

The generation process is thus a two-step procedure that first shapes the prosodic skeleton by computing the prosodic characteristics of each syllable and then computes the prosodic parameters according to parameter-specific phonetic models. Conversely, the analysis process first computes the prosodic skeleton from observable prosodic contours. This operation is straightforward. The analysis process then further decomposes the prosodic skeleton into contributing elementary prosodic segments. This operation is not straightforward and is described and illustrated in Section 4.2.

### 3.1. Rhythm

*Skeleton.* Barbosa and Bailly (1994) propose a multilevel timing generation process similar to that in Campbell (1992) but use a different Rhythmic Unit (RU). Instead of considering onsets of the phonological syllable as delimiting the RU, we consider that speech is paced by beats at the perceptual center (Marcus, 1981) of each syllable. This distinction, which concerns only the definition of a landmark for each syllable, is not important here, and the SFC can operate with any definition of RU. Each RU is characterized by a lengthening/shortening factor equal to the quotient between the duration of the RU and the expected RU duration. This expected duration  $D_{RU} = (1 - \alpha) \sum d_{Si} + \alpha D_0$  is a weighted sum of (a) the sum of the mean durations  $d_{Si}$  of its constituent segments  $Si$  and (b) an average RU duration  $D_0$  reflecting a tendency to isochrony. This weighting scheme is compatible with the so-called quantal effects introduced by Fant and Kruckenberg (1996) that exhibit a quasi-linear relationship between the average syllable duration and the number of phones, with the slope and intersect being language- and speaker-specific. The parameters  $\alpha$  and  $D_0$  can be optimized using an extensive search procedure minimizing the reconstruction error of the entire training database by the SFC. Typical values are  $\alpha = 0.6$  and  $D_0 = 190$  ms.

*Flesh.* A  $z$ -score procedure is then applied in order to distribute the duration of each RU among its constituent segments. Pause insertion is obtained by saturating the lengthening factor of the RU. The pause duration is computed as the duration loss due to this saturation (for further details please refer to Barbosa and Bailly, 1997). Pause is here an *emergent* phenomenon that results from large lengthening factors; it does not influence a priori the determination of the constituents of the phonological structure.

### 3.2. Melody

*Skeleton.* A first decomposition of the  $f_0$  curve is performed using a stylization procedure similar to MOMEL (Hirst et al., 1991) that factors a smooth macromelodic component and a microprosodic component consisting of residual deviations due to the segmental substrate. Contrary to MOMEL stylization, the procedure incorporates an a priori synchronization with the segment string: we stylize the macromelodic component by sampling the logarithm of  $f_0$  at 10%, 50% and 90% of the duration of the vocalic nucleus of each syllable.<sup>4</sup> The mapping model is charged with the prediction of this crude approximation of the melodic component, i.e. the melodic *skeleton* of the utterance.

*Flesh.* Concatenative synthesis provides a way of giving flesh to this skeleton. The same stylization process is in fact performed for the utterances from which the segments of the dictionary are extracted and the residual component (stylization errors + microprosodic component) is stored, retrieved and added at synthesis time (see initial proposal in Monaghan, 1992). Note that this generation process is entirely compatible with a superpositional model. Fig. 3 shows how differences between the original  $f_0$  curve and the stylization are memorized by the segment dictionary, warped and added at synthesis time to the melodic skeleton. All phonetic details such as jitter or

<sup>4</sup> In fact the raw  $f_0$  contour of each vocalic nucleus is first smoothed by a parabolic approximation that is then sampled.

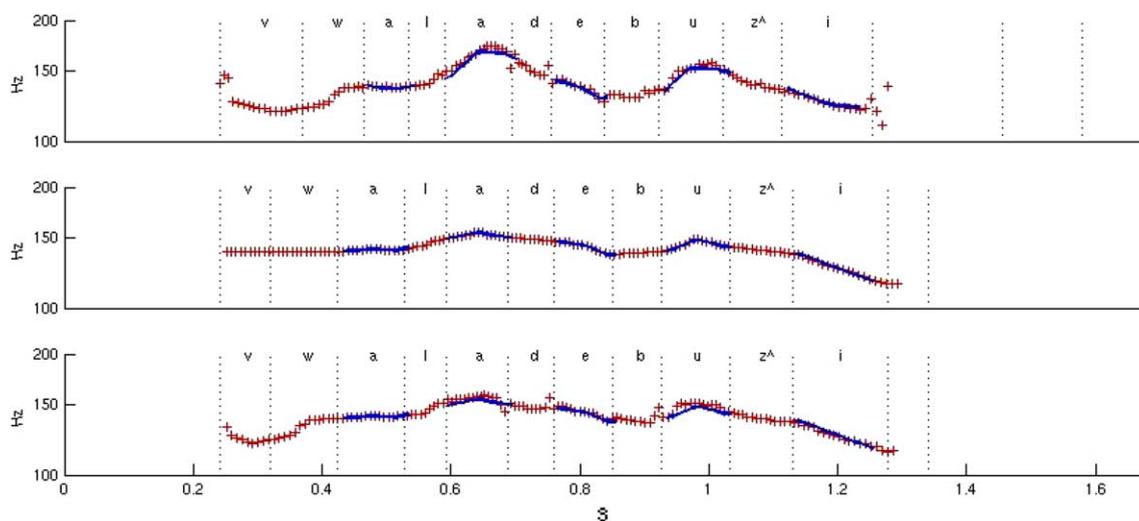


Fig. 3. Illustration of giving flesh to the melodic skeleton of a French utterance (*Voilà, des bougies!* ‘Look—candles!’). Top: original  $f_0$  curve with stylization superposed. Middle: melodic skeleton predicted by the SFC model trained using 235 such utterances (first training loop). Bottom: final predicted  $f_0$  curve with melodic skeleton also superposed. Microprosodic details such as bursts in plosives (/d/ and /b/ here) or  $f_0$  dips in the initial /v/ are reproduced by adding deviations between original  $f_0$  curve and stylization to the predicted melodic skeleton (displayed in the top panel).

microprosodic  $f_0$  movements due to consonantal perturbations (e.g., bursts,  $f_0$  dips due to drops in supraglottal pressure in case of constrictions) are restored by this simple additive scheme. A similar proposal was also introduced for synthesis-by-rule by adding a microprosodic component to the Fujisaki model (Bailly, 1989).

#### 4. The mapping model

Considering prosodic contours as the superposition of elementary contours is a many-to-one ill-posed problem that requires regularization schemes. The Fujisaki model (Fujisaki and Sudo, 1971), for example, imposes constraints on the shape of these elementary contours (exponential responses of second-order filters to impulses and square waves). The SFC model does not impose such low-level constraints, but relies only on the consistency between different instantiations of the same discourse function within the corpus. These instantiations are supposed to be performed by so-called *contour generators*.

##### 4.1. Contour generators

Each metalinguistic function is encoded by a specific prototypical contour anchored to the scope of that function (i.e. the extent of the units to which the function applies) by a few *landmarks*, i.e. the beginning and end of the unit(s) concerned with this function. As the metalinguistic function can be applied to different scopes, it is characterized by a family of contours, a set of prosodic *clichés* (Fónagy et al., 1984). 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 for instance implemented as simple feedforward neural networks (a) receiving as input linear ramps giving the absolute and relative distance of the current syllable from the closest landmarks of the scope and (b) delivering as output prosodic characteristics for the current syllable (see Fig. 4). Each network has very few parameters (typically four input, 15 hidden and four output units =  $4 * (15 + 1) + 15 * (4 + 1) = 139$  parameters) to

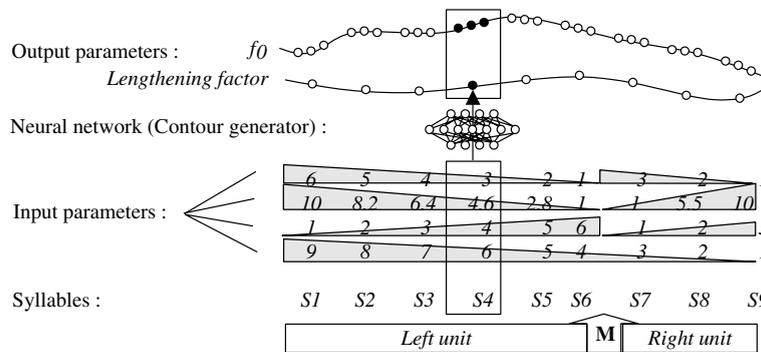


Fig. 4. Example of the implementation of a function  $M$  by a contour generator on two units of respectively six and three syllables. The generator computes four output parameters per syllable (three  $f_0$  values and a lengthening factor; see Section 3). Each syllable is characterized by four input parameters that specify the relative and absolute position of that syllable within the scope. The input parameters are thus series of linear ramps varying between the extreme values 1 and 10. From bottom to top: relative distance to the end of the scope; relative and absolute distance to the beginning/end of the unit to which it belongs.

be compared with the thousands of parameters generally necessary to learn the complex mapping between richer and more heterogeneous phonological and phonotactic inputs and prosodic parameters such as those in (Traber, 1992; Mixdorff and Jokisch, 2001). 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 PMEMs of different metalinguistic functions in Fig. 5): the PMEM defines how a prosodic cliché develops on a unit according to the size of the unit. We update here the concept of dynamic lexicons as proposed by Aubergé (1992). Note that the choice of the neural network implementation of the PMEM is not the only choice, but it offers an efficient learning paradigm (see next section). The final multiparametric prosody is thus obtained by superposing and adding the many functional contours produced by a few independent contour generators (typically three or four). Those inputs are parameterized according the variable scopes of the metalinguistic functions involved in the utterance.

#### 4.2. Training contour generators

The problem is now to feed our contour generators with samples of elementary multiparametric

contours extracted from raw training data. In the case of a superpositional model, the problem is often ill-posed since each observation is in general the sum of several contributions, i.e. here the outputs of contributing contour generators. We thus need extra constraints to regularize the inversion problem, eventually complemented by restrictions on the number and positioning of contours using a priori linguistic information. For example, inversion of the Fujisaki model (Narusawa et al., 2002) is facilitated by the very different characteristics of the phrase and accent command filters. In our phonetic model, the shapes of the contributing contours are unconstrained a priori, an important characteristic since we have shown that contours may potentially have complex shapes (e.g. those encoding attitudes at the sentence level as in Fig. 5 or tones as in Fig. 11). Note, however, that nothing in the following framework forbids us from adding more constraints (such as imposing exponential shapes as in the Fujisaki model) on those contours. Imposing also the shape of some contours can also ease the emergence of other contours. Here the shapes of the contributing contours simply emerge as a by-product of an inversion procedure. This inversion procedure tunes contour generators so that the prosodic contours predicted by overlapping and adding their contributions in the discourse best predicts observed realizations. The main analysis-by-synthesis loop

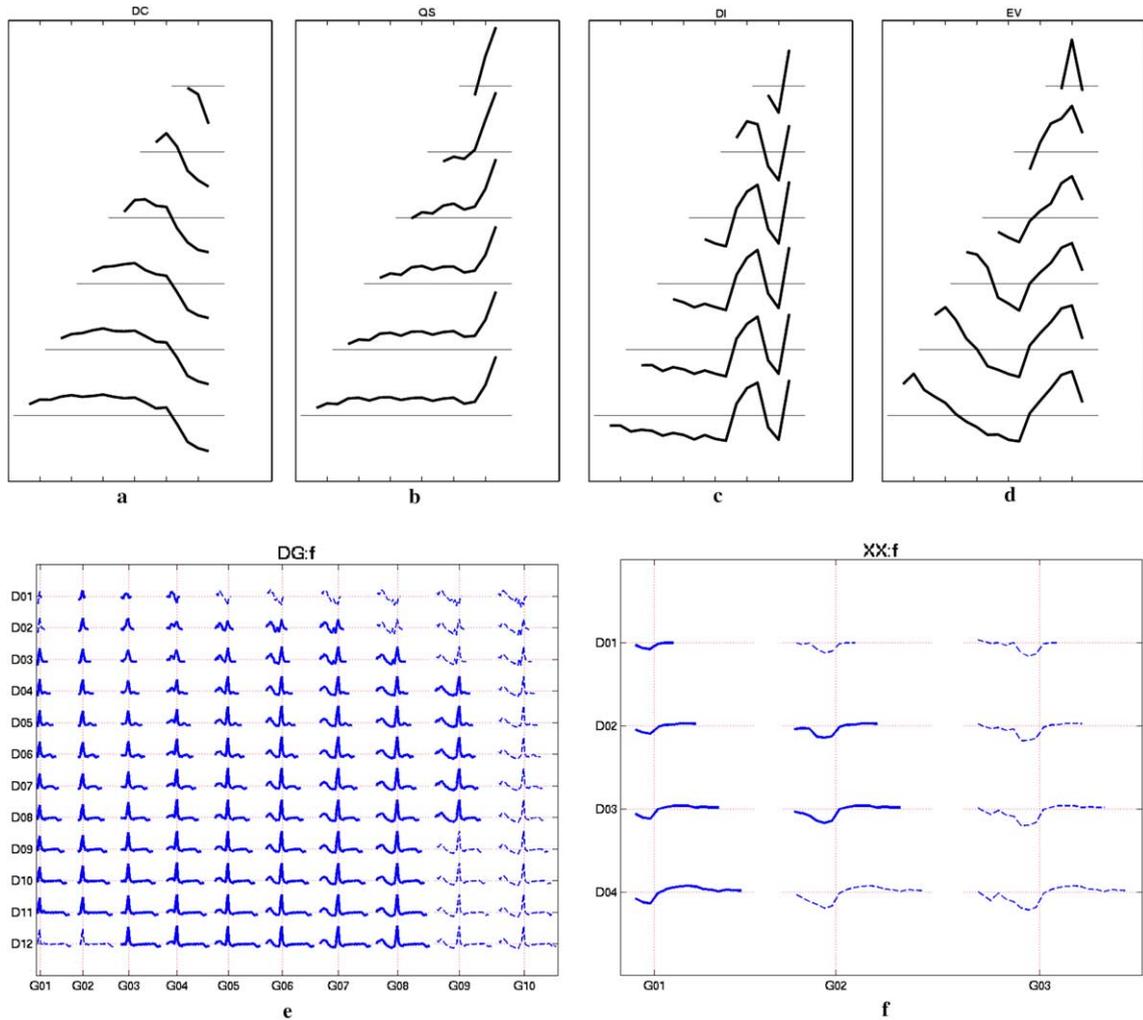


Fig. 5. Expansion of  $f_0$  movement for different metalinguistic functions applied to units with increasing size. These expansions are synthesized by trained contour generators. Contours with dotted lines are extrapolated, i.e. no exemplars are present in the training corpus. Top: modalities and prosodic attitudes in French: (a) assertion, (b) question, (c) incredulous question, (d) statement of an obvious fact. Please note that contour generators learn the laws governing the variation of the global slope of the contours and final movements of that contours thanks to the absolute and relative linear ramps (see Fig. 4): the relationship between global slope and amplitude of the final melodic movement versus sentence length may be non-linear. Bottom: linking two constituents (abscissa/ordinate: size of first/second unit): (e) between a governor and its left dependent when considering a dependency tree (see Section 2.1) and (f) between a clitic word and the following content word (e.g.  $f_0$  dip on the determiner preceding a noun).

is described in Fig. 6a. SFC generators are trained iteratively. At iteration  $n$ ,

1. Generators predict functional contours for all units of the corpus using the parameterization learnt at the previous iteration  $n - 1$ . Generators at first iteration predict null contours, i.e. output zero values.
2. For each utterance, synthetic prosodic contours are then computed by superposing the functional contours associated to all units of the utterance.
3. For each utterance, these predicted contours are compared to the observed contours. The prediction error is computed and distributed among the contributing functional contours. In the

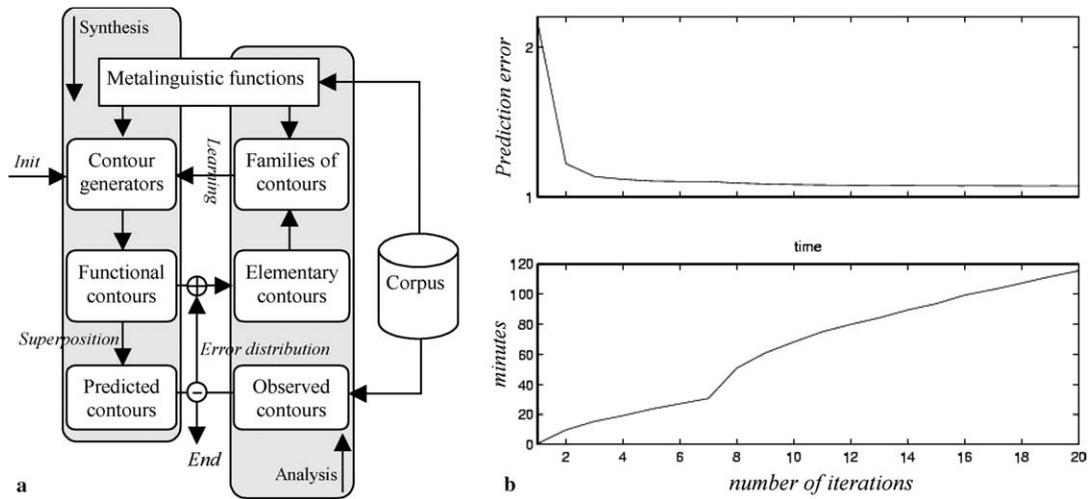


Fig. 6. (a) Analysis-by-synthesis loop. SFC generators are trained using patterns built by adding to what they already predict a proportion of what they together still do not predict, i.e. the difference between observed and predicted contours at the iteration considered (see Section 4.2 for detailed comments on the training loop). The learning loop stops when this difference no longer diminishes significantly. (b) Typical evolution of the reconstruction error (top) and training time on a standard Pentium III computer (bottom) for a sample corpus (1000 utterances; six contour generators) as a function of the number of iterations. Convergence is obtained within a few dozen iterations.

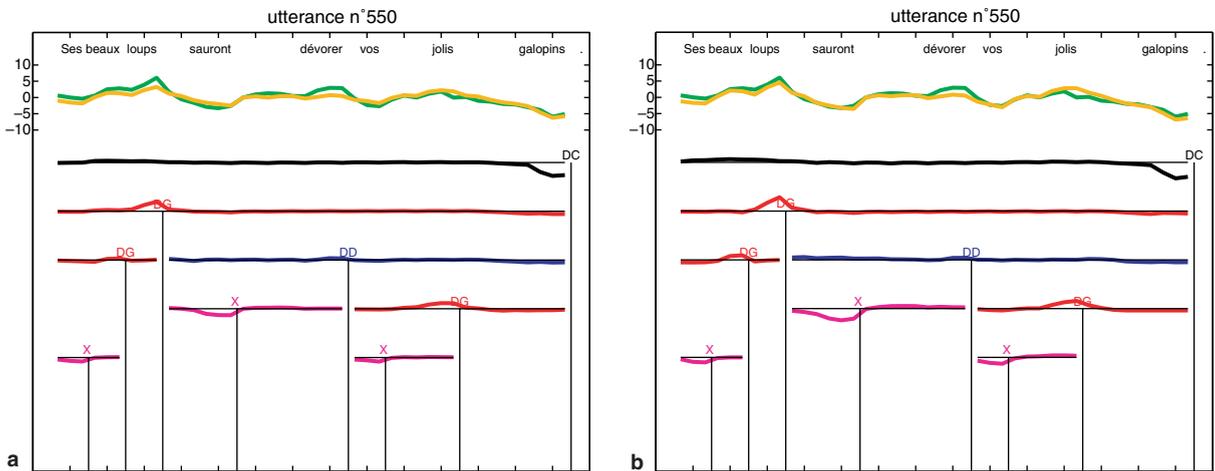


Fig. 7. Optimal decomposition predicted for the  $f_0$  skeleton for one utterance ('His/her pretty wolves will know how to devour your nice urchins.') at iterations 1 and 6. For each caption: top: predicted  $f_0$  skeleton superposed with original one; bottom: elementary contours predicted for each discourse function used to encode the linguistic structure of the utterance; the prediction is obtained by superposing and adding these elementary contours. Horizontal axis represents the syllable count and all  $f_0$  values are those sampled at 10%, 50% and 90% of the vocalic nucleus. These values are connected with plain lines for sake of readability. Although decomposition structure is almost determined at iteration 1, successive iterations refine amplitudes of elementary contours, e.g. the contours at the bottom encoding the discourse function X or second elementary contour encoding a DG discourse function. Optimal decomposition after 1 iteration (a) and 15 iterations (b).

most simple SFC implementation, this difference is simply distributed syllable-by-syllable among

functional contours that are effectively superposed at the considered syllable. As functional

contours have different scopes, this difference is not equally distributed. This distributed prediction error is then added to each predicted functional contour to form new target functional contours for the training of generators.

4. Target functional contours are collected for all utterances of the training corpus and sorted according to the discourse function they implement. These target functional contours are then considered as new target patterns for generator training. Here a classical neural network training procedure, error back-propagation, is used. Once trained, the generators, are further reconsidered starting back at step 1.

The learning loop stops when the prediction error no longer diminishes significantly. Holm (2003) demonstrates that convergence is guaranteed: in practice, asymptotic behavior is always obtained within a few dozen iterations. Fig. 6b shows the typical evolution of the reconstruction error and training time for a large training corpus. Fig. 7 shows that the first iteration already converges towards a correct placement of the main prosodic events which further iterations contribute to optimally shape. For further details on this original analysis-by-synthesis loop, see Holm and Bailly (2000, 2002).

#### 4.3. Examining prediction results: A case study

Figs. 8–10 summarize the different representations we have in hands for evaluating the quality of the training procedure and the generation process. We picked up one utterance of a corpus of spoken formulas studied by Holm et al. (1999). Spoken formulas were chosen because their recursive and highly embedded syntactic structure and because prosody must be recruited to encode this structure. The reader can convince himself by examining Fig. 9 that compares the original and predicted performance structures of the spoken formula. The computed performance structures (Gee and Grosjean, 1983) uses here the lengthening factor of the last syllable of each word as a cue for word grouping instead of the duration of its rime as proposed initially by Gee and Grosjean. Fig. 9 shows that SFC is able to predict accurately highly structured performance trees by simply superposing contours acting on different scopes.

Fig. 8 shows in more details how the  $f_0$  and rhythmic skeletons are decomposed into elementary  $f_0$  and rhythmic contours acting on different scopes and encoding a few metalinguistic functions. Only four functions are used here: one function for encoding the function “statement of an equation” and three for encoding the dependency

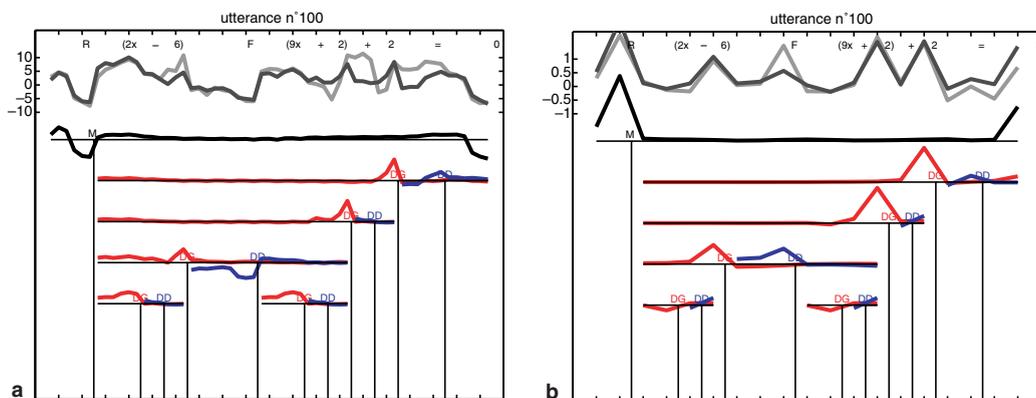


Fig. 8. Analyzing and predicting the prosodic skeleton of a spoken mathematical formula (“solve:  $(2x - 6)x(9x + 2) + 2 = 0$ ”). The  $f_0$  skeleton is figured on the left hand side using the same conventions as for Fig. 7. The rhythmic skeleton is figured on the right hand side using also the same conventions except that the lengthening factor is displayed per syllable instead of three  $f_0$  values. Three discourse functions are here recruited (M, DG and DD) and applied to different scopes in order to encode the hierarchical syntactic relations between the different terms of the formula. The SFC has been trained on 132 such spoken maths.  $f_0$  skeleton (a) and rhythmic skeleton (b).

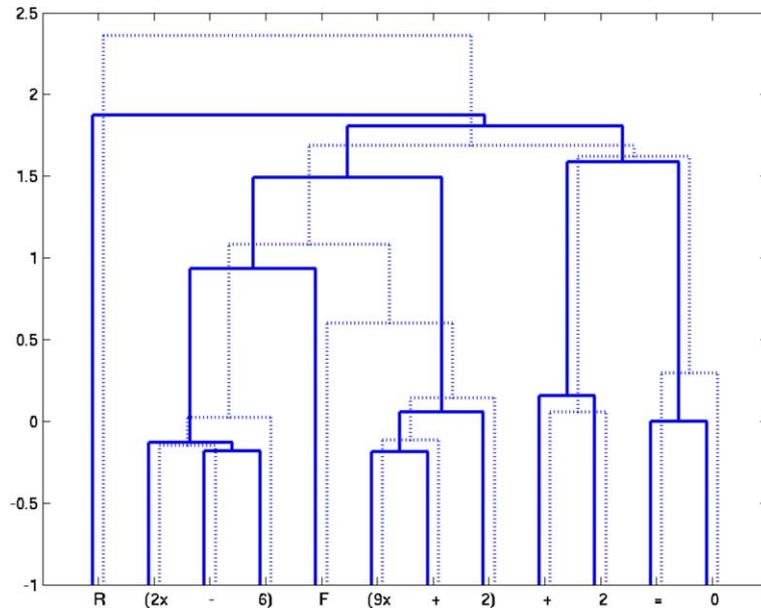


Fig. 9. Performance structures (Gee and Grosjean, 1983) of the original (plain) and predicted (dotted) prosody for the same formula as Fig. 8. We use here lengthening factors of the last syllable of each word as a cue for word grouping. We provide here a more structural view of the data already displayed in Fig. 8b. One can see that the SFC renders properly the hierarchy except the balance between the first two operands of the formula.

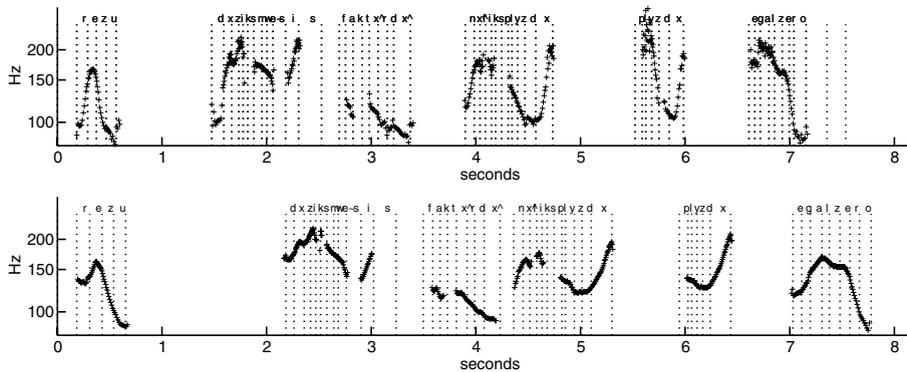


Fig. 10. Comparing original and predicted F0 contours for the same formula as Fig. 8. Top: original and stylized  $f_0$  curve. Bottom: stylized  $f_0$  curve predicted by the SFC.

relations between operands and operators. Some operators only require a “right dependency” with their following operand such as “square root of” or “absolute value of”, most require both “left dependency” and “right dependency” with their preceding and following operands such as “plus”, “minus” or “multiplied/divided by”, few require more complex organization of following operands

such as “integral from...to...of...” where an “interdependency” metalinguistic function is used to link subsequent operands. Despite this very few number of metalinguistic functions—and thus of trained contour generators—the SFC is able to predict complex and bumpy prosodic contours by overlapping and superposing simple and smooth elementary prosodic contours.

The multiparametric prosodic skeleton is then used to drive a concatenative speech synthesizer. Fig. 10 shows the predicted prosodic contour. Once the rhythmical skeleton has been computed and the  $z$ -scoring procedure has determined the phonemic durations of each segment, the  $f_0$  skeleton is injected at the 10%, 50% and 90% duration of the nucleus of each predicted syllable. Fig. 10 shows that this crude approximation can already be used to synthesize speech.

As described in Section 3.2, this  $f_0$  contour can be further enriched by the melodic flesh stored in the segment dictionary.

## 5. Performance

In Table 1, we summarize results obtained in some of the various languages and linguistic material the SFC has been confronted with. For each corpus we present the average result of four simulations in which half of the stimuli were randomly selected for training and half were randomly selected for testing. Mean RMS error for predicted  $f_0$  targets, lengthening factors for syllables and predicted phoneme durations are 1.8 semi-tones, 20% and 20 ms, respectively. Note that half of the average RMS error for phoneme durations may be due to the  $z$ -score procedure (see Section 3.1): the  $z$ -score procedure that distributes the syl-

lable duration among segments generates a mean modelling error of 11 ms.

Results are difficult to compare with those of other trainable models. We are clearly missing benchmarking procedures for assessing proposals (please refer to Raidt et al., 2004, for a tentative comparative evaluation of two trainable models). Our results for  $f_0$  could however be compared to the 34.7 Hz obtained by Ross and Ostendorf (1999) (2.33 semi-tones if we consider a baseline  $f_0$  of 240 Hz for their female speaker) for 48 min of training material (8841 words) and 11 min of testing material. Similar results were obtained by Mixdorff and Jokisch (2001): 18 Hz (2.41 semi-tones if we consider a baseline  $f_0$  of 120 Hz for their male speaker) for 5000 training and test syllables. Note, however, that these models, as with most of the models proposed in the literature, treat *only* the mapping between the surface phonological structure and the prosodic continuum: prosodic labels, such as ToBI (Tones and Break indices, see Silverman et al., 1992) labels in Ross and Ostendorf (1999) or those in (Black and Hunt, 1996), are considered as inputs. Despite inter-labeller disagreements with respect to accent and edge tone type (Syrdal and McGory, 2000), these hand-labelled data bias the evaluation of a fully automatic mapping. Secondly, a further mapping between the linguistic structure and the prosodic labels (see for example the prediction of accent locations in

Table 1  
SFC performance on various training and test materials

Language	French						German	Chinese
	Mal. A	Mal. B	Mal. A	Mal. C	Mal. D	Fem. A	Mal. E	Fem. B
Nb of Utt	1000	235	157	6 × 400	235	235	489	100
Nb of SFC	6	6	4	5	6	6	4	7
Type	Text	Text	Math	Attitude	Text	Cued Sp.	Math	Text
$f_0$ Training	1.51/0.84	1.44/0.72	2.23/0.88	1.47/0.94	1.26/0.86	0.88/0.83	1.54/0.77	1.75/0.92
$f_0$ Test	1.54/0.83	1.60/0.65	2.19/0.89	1.71/0.93	1.58/0.76	0.97/0.80	1.58/0.76	1.78/0.84
LF Training	0.15/0.68	0.26/0.42	0.23/0.95	0.14/0.82	0.20/0.52	0.25/0.77	0.20/0.89	0.18/0.83
LF Test	0.15/0.68	0.27/0.34	0.25/0.93	0.17/0.72	0.22/0.39	0.30/0.65	0.20/0.88	0.19/0.82
ms Training	16.1/0.75	27.9/0.50	24.3/0.72	19.6/0.81	27.9/0.63	45.0/0.56	33.1/0.77	22.7/0.86
ms Test	16.3/0.75	29.2/0.50	23.8/0.72	21.6/0.76	30.0/0.53	45.4/0.52	34.5/0.76	23.3/0.84

Prediction errors on  $f_0$  targets (semi-tones), lengthening factor (LF) and phoneme durations (ms) are characterized by RMS and the correlation factor between predicted patterns and ground-truth. The number of metalinguistic functions (and thus contour generators) used is also given. Results are presented according to language and speakers (Mal. for male speakers and Fem. for female speakers) as well as type of linguistic material and style (read texts, spoken math, prosodic attitudes or texts uttered with cued speech). Three other speakers tested for Chinese have shown similar performance (see Chen et al., 2004).

Dusterhoff et al., 1999) constitutes an additional source of errors that may have a drastic impact on the performance of the trainable models considered above. Systems that do not use an intermediate symbolic representation are rare (Strom, 2002; Buhmann et al., 2000) and their results are less encouraging than the ones discussed above.

## 6. Comments

*Analyzing prosody.* The analysis-by-synthesis procedure presented here gives access to the *hidden structure* of intonation (Holm and Bailly, 2002): the prosodic implementation of metalinguistic functions emerges from the automatic parameterization of contour 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. The SFC is thus not only a trainable model for *generating* prosody but also a valuable tool for *analyzing* prosody, i.e. testing different ways (e.g., metalinguistic functions recruited, scopes considered) of decomposing prosody into functional contours.

*Gradiance.* Contrary to most other trainable models of intonation, the training phase of the model presented here essentially learns the *shapes* of the contours associated with pre-defined metalinguistic functions. It does not learn, for example, how these metalinguistic functions are transmitted in parallel with the prosodic continuum; this is imposed by the superposition hypothesis. One can, however, imagine more sophisticated interaction models in which not only shapes but also parameter-specific weighting factors are computed in order to give more priority to given metalinguistic functions for specific prosodic parameters at certain positions of their scope. We refer here to the concept of *gradiance* (Gussenhoven, 1999). For example,  $f_0$  contours of non-modal attitudes (Morlec et al., 2001) are influenced neither by the size nor by the structure of the discourse units, while the organization of lengthening factors still reflects (though in a reduced way) the organization observed for the same sentences uttered in a declarative modality. Two interpretations are possible here: (a) in interactive speech, some parts of the discourse structure or content can be predicted from discourse history and so are already known for the listener. The non-modal attitudes studied here (statement of evident facts, incredulity,

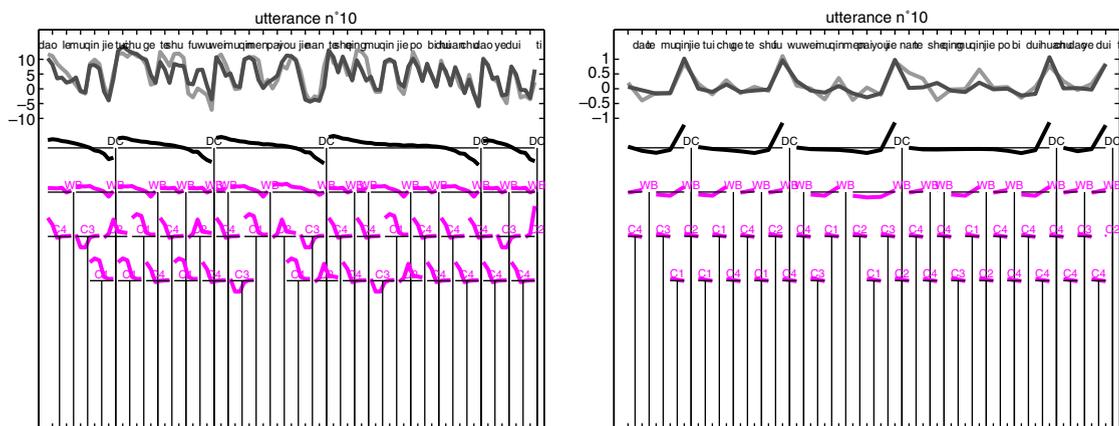


Fig. 11. Decomposing the melodic (left) and rhythmic (right) contours of a recording consisting of five Chinese sentences. Same conventions as for Fig. 7. In each sub-figure, from top to bottom: superposition of original (light gray) and predicted (dark gray) prosodic skeleton; sentence-specific contours; word-specific contours and tone-specific contours. The tone contours are displayed in two rows for sake of clarity since they overlap. The scope of each tone contour is two syllables: the syllable carrying the tone and the following one (except for the last syllable of each sentence, where only the tone-bearing syllable is considered). No group- or clause-specific contours have been yet introduced.

suspicious irony) can be used by a speaker to moderate or cast doubt on an affirmation just uttered by his or her interlocutor. Internal structure of the discourse (wording, syntax, etc.) is thus given and does not need to be given back. (b) *f0* functional contours encoding non-modal attitudes are complex and their salience should not be spurred by superposing other functional contours. Less important functional contours are thus modulated by the salience of the more important functional contours. These interpretations may be implemented and tested using an additional mechanism responsible for weighting functional contours either using higher-level pragmatic information—scaling contours by an additional factor *importance* associated with the specification of each metalinguistic function—or breaking the strict independence between functional contours.

## 7. Conclusion

We have described and illustrated here the core principles and the basic properties of the trainable model SFC. The SFC has been applied to various discourse types and to different languages. We demonstrated that this model-based generation scheme is compatible with a certain technological efficiency. The comparative evaluation between the SFC and other trainable models performed by Raidt et al. (2004) should be promoted and would benefit the development of alternative approaches to the modelling of prosody.

We look forward to confronting the SFC with an increasing variety of challenges. Tone languages are of most interest because prosodic structure, especially intonation, is expected to be *hidden* by syllabic melodic contours used as phonetic features. We have been working on Cantonese with Gaopeng Chen of University of Science and Technology of China (USTC). Preliminary results (Chen et al., 2004) of a simple three-level decomposition distinguishing tone (using five contour generators: one for each of the four tones + neutral tone), word-domain (one contour generator for word segmentation) and sentence-domain (one contour generator for declarative sentences) contributions are illustrated in Fig. 11. Despite massive

and coherent contributions of the tonal sequences, smooth and coherent functional contours are predicted for encoding Cantonese declarative sentences. Adding more linguistic structuring, notably syntactic bracketing, will of course improve an already low prediction error (see Table 1).

There is, of course, still room for improvement. The most important point is surely the two-stage phonetic model that separate microprosodic, segment-dependent prosodic events from prosodic phenomena encoding the discourse structure, referenced to above as *flesh* and *skeleton*. Despite satisfactory results, the interaction models (additive) and the deconvolution framework (removing microprosody first) are certainly too simple and can be improved. For example, the simple *z*-score procedure used here neglects complex interactions between the different timing constraints. We look forward to setting up a training framework that will also cope with microprosody in a more homogenous way.

Finally, the SFC analysis-by-synthesis scheme may be used for other trainable models: the resulting trained model is almost never used to look back to input/output training data. These data are nevertheless often noisy: input phonological characterization of the training stimuli is often done automatically and analysis of output parameters characterizing prosody is often not guided by linguistic constraints (see for example Hirst et al., 1991). A feedback loop that automatically or semi-automatically adjusts constraints with observations in light of the prediction errors made by the trained model may be of interest (see for example for recent work of Agüero et al., 2004).

## Acknowledgments

We are greatly indebted to Véronique Aubergé, Plinio Barbosa and Yann Morlec for their major contributions to the SFC proposal. Gaopeng Chen and Stefan Raidt helped us to confront the SFC with new languages. The content of the paper has been greatly improved by the judicious comments and recommendations of Pauline Welby, Keikichi Hirose and two anonymous reviewers.

## References

- Agüero, P.D., Wimmer, K., Bonafonte, A., 2004. Joint extraction and prediction of Fujisaki's intonation model parameters. In: *Internat. Conf. on Spoken Language Processing*, Jeju, Korea.
- Aubergé, V., 1992. Developing a structured lexicon for synthesis of prosody. In: Bailly, G., Benoît, C. (Eds.), *Talking Machines: Theories, Models and Designs*. Elsevier, pp. 307–321.
- Aubergé, V., 1993. Prosody modeling with a dynamic lexicon of intonative forms: Application for text-to-speech synthesis. *Working Papers of Lund University*, Vol. 41, pp. 62–66.
- Bachenko, J., Fitzpatrick, E., 1990. A computational grammar of discourse-neutral prosodic phrasing in English. *Comput. Linguist.* 16, 155–167.
- Bailly, G., 1989. Integration of rhythmic and syntactic constraints in a model of generation of French prosody. *Speech Comm.* 8, 137–146.
- Bailly, G., 1997. No future for comprehensive models of intonation? In: Sagisaka, Y., Campbell, N., Higuchi, N. (Eds.), *Computing Prosody: Computational Models for Processing Spontaneous Speech*. Springer-Verlag, pp. 157–164.
- Bailly, G., Holm, B., 2002. Learning the hidden structure of speech: From communicative functions to prosody. *Cadernos Estudos Linguist.* 43, 37–54.
- Balfourier, J.-M., Blache, P., van Rullen, T., 2002. From shallow to deep parsing using constraint satisfaction. In: *Coling, Taipei, Taiwan*, pp. 36–42.
- Barbosa, P., Bailly, G., 1994. Characterisation of rhythmic patterns for text-to-speech synthesis. *Speech Comm.* 15, 127–137.
- Barbosa, P., Bailly, G., 1997. Generation of pauses within the z-score model. In: Santen, J.P.H.V., Sproat, R.W., Olive, J.P., Hirschberg, J. (Eds.), *Progress in Speech Synthesis*. Springer-Verlag, New York, pp. 365–381.
- Bartkova, K., Sorin, C., 1987. A model of segmental duration for speech synthesis in French. *Speech Comm.* 6, 245–260.
- Black, A.W., Hunt, A.J., 1996. Generating F0 contours from ToBI labels using linear regression. In: *Proc. Internat. Conf. on Speech and Language Processing*, pp. 1385–1388.
- Boersma, P., Weenink, D., 1996. Praat, a System for doing Phonetics by Computer, version 3.4. Institute of Phonetic Sciences of the University of Amsterdam, Report 132, 182 p.
- Bolinger, D., 1989. *Intonation and Its Uses*. Edward Arnold, London.
- Brichet, C., Aubergé, V., 2004. La prosodie de la focalisation en français: faits perceptifs et morphogénétiques. In: *Journées d'Etudes sur la Parole, Nancy-France*.
- Buhmann, J., Vereecken, H., Fackrell, J., Martens, J., Van Coile, B., 2000. Data driven intonation modelling of 6 languages. In: *Internat. Conf. on Spoken Language Processing, Beijing, China*, pp. 179–182.
- Campbell, N., 1992. Multi-level timing in speech. PhD Thesis, University of Sussex, Brighton, UK, 300 p.
- Chen, G.-P., Bailly, G., Liu, Q.-F., Wang, R.-H., 2004. A superposed prosodic model for Chinese text-to-speech synthesis. In: *Internat. Conf. of Chinese Spoken Language Processing, Hong Kong*, pp. 177–180.
- Cutler, A., Norris, D., 1991. Prosody in situations of communication: Saliency and segmentation. In: *Proc. Internat. Congress of Phonetic Sciences, Aix-en-Provence, France*, pp. 264–270.
- Di Cristo, A., Di Cristo, P., Campione, E., Veronis, J., 2000. A prosodic model for text to speech synthesis in French. In: Botinis, A. (Ed.), *Intonation: Analysis, Modelling and Technology*. Kluwer, Amsterdam, pp. 321–355.
- Dusterhoff, K.E., Black, A.W., Taylor, P., 1999. Using decision trees within the tilt intonation model to predict F0 contours. In: *EuroSpeech, Budapest, Hungary*, pp. 1627–1630.
- Fant, G., Kruckenberg, A., 1996. On the quantal nature of speech timing. In: *Proc. Internat. Conf. on Speech and Language Processing, Philadelphia—USA*, pp. 2044–2047.
- Fónagy, I., Bérard, E., Fónagy, J., 1984. Clichés mélodiques. *Folia Linguist.* 17, 153–185.
- Fujisaki, H., Sudo, H., 1971. A generative model for the prosody of connected speech in Japanese. *Annual Report of Engineering Research Institute*, Vol. 30, pp. 75–80.
- Fujisawa, K., Campbell, N., 1998. Prosody-based unit selection for Japanese speech synthesis. In: *ESCA/COCOSDA Internat. Workshop on Speech Synthesis*.
- Gee, J.-P., Grosjean, F., 1983. Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychol.* 15, 411–458.
- Gussenhoven, C., 1999. Discreteness and gradience in intonational contrasts. *Lang. Speech* 42, 283–305.
- Hirst, D.J., 2003. The phonology and phonetics of speech prosody: Between acoustics and interpretation. In: *Internat. Conf. on Speech Prosody, Nara, Japan*, pp. 163–169.
- Hirst, D., Nicolas, P., Espesser, R., 1991. Coding the F0 of a continuous text in French: An experimental approach. In: *Proc. Internat. Congress of Phonetic Sciences, Aix-en-Provence, France*, pp. 234–237.
- Hirst, D.J., Di Cristo, A., Espesser, R., 2000. Levels of representation and levels of analysis for the description of intonation systems. In: Horne, M. (Ed.), *Prosody: Theory and Experiment*. Kluwer Academic Publishers, Dordrecht—the Netherlands, pp. 51–87.
- Holm, B., 2003. Implémentation d'un modèle morphogénétique de l'intonation. Application à l'énonciation de formules mathématiques. PhD Thesis. Institut National Polytechnique, Grenoble—France, 239 p.
- Holm, B., Bailly, G., 2000. Generating prosody by superposing multi-parametric overlapping contours. In: *Proc. Internat. Conf. on Speech and Language Processing, Beijing, China*, pp. 203–206.
- Holm, B., Bailly, G., 2002. Learning the hidden structure of intonation: Implementing various functions of prosody. In: *Speech Prosody, Aix-en-Provence, France*, pp. 399–402.
- Holm, B., Bailly, G., Laborde, C., 1999. Performance structures of mathematical formulae. In: *Internat. Congress of Phonetic Sciences, San Francisco, USA*, pp. 1297–1300.

- Klatt, D.H., 1979. Synthesis by rule of segmental durations in English sentences. In: Lindblom, B., Ohlman, S. (Eds.), *Frontiers of Speech Communication Research*. Academic Press, London, pp. 287–300.
- Ljolje, A., Fallside, F., 1986. Synthesis of natural sounding pitch contours in isolated utterances using hidden Markov models. *TrASSP* 34, 1074–1080.
- Marcus, S.M., 1981. Acoustic determinants of Perceptual center (p-center) location. *Percept. Psychophys.* 30 (3), 247–256.
- 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: van Santen, J.P.H., Sproat, R.W., Olive, J.P., Hirschberg, J. (Eds.), *Progress in Speech Synthesis*. Springer-Verlag, New York, pp. 477–493.
- Mixdorff, H., Jokisch, O., 2001. Building an integrated prosodic model of German. In: *Eur. Conf. on Speech Communication and Technology*, Aalborg, Denmark, pp. 947–950.
- Monaghan, A.I.C., 1992. Extracting microprosodic information from diphones—a simple way to model segmental effects on prosody for synthetic speech. In: *Internat. Conf. on Speech and Language Processing*, Banff, Canada, pp. 1159–1162.
- Morlec, Y., Rilliard, A., Bailly, G., Aubergé, V., 1998. Evaluating the adequacy of synthetic prosody in signaling syntactic boundaries: Methodology and first results. In: *First Internat. Conf. on Language Resources and Evaluation*, Granada, Spain.
- Morlec, Y., Bailly, G., Aubergé, V., 2001. Generating prosodic attitudes in French: Data, model and evaluation. *Speech Comm.* 33 (4), 357–371.
- Narusawa, S., Minematsu, N., Hirose, K., Fujisaki, H., 2002. A method for automatic extraction of model parameters from fundamental frequency contours of speech. In: *Internat. Conf. on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, pp. 1281–1284.
- Nespor, M., Vogel, I., 1986. *Prosodic Phonology*. Dordrecht, Foris.
- O'Shaughnessy, D., 1981. A study of French vowel and consonant durations. *J. Phonetics* 9, 385–406.
- Pynte, J., Prieur, B., 1996. Prosodic breaks and attachment decisions in sentence parsing. *Lang. Cognitive Process.* 11 (1), 165–191.
- Raidt, S., Bailly, G., Holm, B., Mixdorff, H., 2004. Automatic generation of prosody: Comparing two superpositional systems. In: *Internat. Conf. on Speech Prosody*, Nara, Japan, pp. 417–420.
- Riley, M., 1992. Tree-based modelling of segmental durations. In: Bailly, G., Benoit, C. (Eds.), *Talking Machines: Theories, Models and Designs*. Elsevier, pp. 265–274.
- Ross, K.N., Ostendorf, M., 1999. A dynamical system model for generating fundamental frequency for speech synthesis. *IEEE Trans. Speech Audio Process.* 7 (3), 295–309.
- Sagisaka, Y., 1990. On the prediction of global F0 shapes for Japanese text-to-speech. In: *IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 325–328.
- Schreuder, M., Gilbers, D., 2004. Recursive patterns in phonological phrases. In: *Internat. Conf. on Speech Prosody*, Nara, Japan, pp. 341–344.
- Scordilis, M., Gowdy, J., 1989. Neural network based generation of fundamental frequency contours. In: *IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing*, pp. 219–222.
- Selkirk, E.O., 1984. *Phonology and Syntax*. MIT Press, Cambridge, MA.
- Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., 1992. TOBI: A standard for labeling English prosody. In: *Internat. Conf. on Speech and Language Processing*, Vol. 2, pp. 867–870.
- Strom, V., 2002. From text to prosody without ToBI. In: *Internat. Conf. on Spoken Language Processing*, Denver, CO, pp. 2081–2084.
- Syrdal, A.K., McGory, J., 2000. Inter-transcriber reliability of ToBI prosodic labeling. In: *Internat. Conf. on Spoken Language Processing*, Beijing, China, pp. 235–238.
- Taylor, P., Black, A.W., 1999. Speech synthesis by phonological structure matching. In: *EuroSpeech*, Budapest, Hungary, pp. 1531–1534.
- Tesser, F., Cosi, P., Drioli, C., Tisato, G., 2004. Prosodic data-driven modelling of narrative style in Festival TTS. In: *Workshop on Speech Synthesis*, Pittsburgh, USA, pp. 185–190.
- Tournemire, S.D., 1997. Identification and automatic generation of prosodic contours for a text-to-speech synthesis system in French. In: *Proc. Eur. Conf. on Speech Communication and Technology*, Rhodes, Greece, pp. 191–194.
- Traber, C., 1992. F0 generation with a database of natural F0 patterns and with a neural network. In: Bailly, G., Benoit, C. (Eds.), *Talking Machines: Theories, Models and Designs*. Elsevier, pp. 287–304.
- Trouvain, J., Barry, W.J., Nielsen, C., Andersen, O., 1998. Implications of energy declination for speech synthesis. In: *ETRW Workshop on Speech Synthesis*, Jenolan Caves, Australia, pp. 47–52.
- van Santen, J.P.H., 1992. Deriving text-to-speech durations from natural speech. In: Bailly, G., Benoit, C. (Eds.), *Talking Machines: Theories, Models and Designs*. Elsevier, pp. 275–285.
- van Santen, J.P.H., 2002. Quantitative modeling of pitch accent alignment. In: *Internat. Conf. on Speech Prosody*, Aix-en-Provence, France, pp. 107–112.
- Wightman, C.W., Syrdal, A.K., Stemmer, G., Conkie A., Beutnagel, M., 2000. Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative text-to-speech synthesis. In: *Internat. Conf. on Spoken Language Processing*, Beijing, China, pp. 71–74.