Accepté pour publication dans Speech Communication 1999

0

Generating prosodic attitudes in French: Data, model and evaluation

YANN MORLEC, GÉRARD BAILLY and VÉRONIQUE AUBERGÉ

0.1 Introduction

The majority of research in the analysis and generation of prosody for use in speech synthesis systems has focused on prosodic features that ease the syntactic parsing of an utterance or highlight certain parts of it. It is well-known that prosody - and especially final lengthening in French – may reflect very fine and complex attributes of the syntactic structure (Gee and Grosjean 1981) and also indicates short and long-term relations between constituents.

This paper focuses on the latter property of prosodic signals and proposes an encoding of long-term relations via global contours that should not be confused with local salient events connected by phonological constructs. To demonstrate, we consider here situations involving interaction between speaker and listener where prosodic contrasts are coextensive with the whole message. Numerous descriptive studies (Fónagy 1984; Bolinger 1989) have in fact proposed that speakers use global prosodic patterns to convey their attitude regarding the message and the interlocutor or their emotional state (expressive function). Our aim here is to study prosodic attitudes and to identify and model their prosodic correlates. The reader will find more data on production, analysis and perception of emotional states in Scherer (1996) and recent progress in synthesis is described in Murray, Arnott and Rohwer (1996). Following Ohala (1996: 1815), we, however, distinguish attitudes from other fundamental communicative expressions more connected with the physiological state of the signaler: *"(Attitudes are) probably acquired, i.e., learned. Thus they are likely to vary considerably from culture to culture and perhaps even from one individual to another.*".

On a continuous axis ordering the expression of inner states from “unintentional” to “intentional”, we place attitudes towards the “intentional” end of the scale, which corresponds to the coding of deliberately-provided information placing the interlocutors in a “mutual belief space” (Clark and Marshall 1981). Evidently this notion of “intentional” information is very complex and difficult to define, due to the social capacity for feigning spontaneous emotions (Damasio 1994). These limitations are however compatible with the final objective of this study: synthesising prosody for one speaker in a given situation.

We propose here a model for the automatic generation of six prosodic attitudes for French. The article describes the three development phases of the model: the design and analysis of a corpus, the implementation of the training algorithm for the model, and the acoustic and perceptual evaluation of the model's behaviour. Our model is directly inspired by a morphological approach of intonation that accounts for the global nature of prosodic contours.

0.2 Theoretical model: a morphological approach

In our study, we envisage the production, perception, and comprehension of language as operations acting on mental representations. Prosodic configurations, which produce as much information about the phonological, lexical, and syntactic content of a discourse as about the speaker's involvement in the process of communication, could form a part of these mental representations, and could serve as reference patterns to which the current vocalisations are compared. This notion of configurations that structure the perception of prosody should be compared with the *morphological paradigm* put forward by Petitot (1986): morphological representations are organised on "natural" frontiers or *catastrophes* of the mapping between the physical and perceived world. Phonological objects and constructs (Petitot 1985; Petitot 1990) are then seen as an emergent organisation constrained by such initial frontiers and a necessary top-down enrichment of contrasts. Phonological contrasts are thus grounded in the acoustic material. We postulate here that they are encoded via variations of acoustic parameters referring to mental representations and stored in memory in the form of acoustic patterns.

During speech acts, listeners should be able to capture phonological contrast in a similar way to the TRACE model of lexical access (McClelland and Elman 1986): a contrast is triggered as soon as the dynamic comparison between the input prosodic contour and competing men-

tal candidates points to a unique solution. Evidence for these expectation mechanisms have been provided by experiments showing that listeners are capable of early prediction of the duration of an utterance (Grosjean 1983), as well as its modality (Boë and Contini, 1975; Thorsen 1980; van Heuven, Haan, Janse and vander Torre 1997). In section 0.5.1, we propose a gating test that shows the existence of the same capability for six attitudes in French. This holistic analysis of prosodic phenomena favours the use of global phonetic representations. As early as 1966, Delattre proposed an inventory of 10 basic contours obtained through semantic oppositions based on intonation alone. Fónagy, Bérard and Fónagy (1984) later introduced the notion of "clichés mélodiques" to define certain stereotype melodic contours for attitudes in French.

Following this morphological paradigm, we put forward the idea that prosodic information is encoded into global contours that may be immediately captured by the listener by comparison with a lexicon of prototypes. We thus assume that prosody can be described as the superposition of independent multi-parameter prosodic contours which belong to hierarchical linguistic levels (Aubergé 1992): sentence, clause, group, subgroup… These prototypical movements are stored in a prosodic lexicon and dynamically used to segment (boundaries), highlight (salience) and enrich (attitudes...) the linguistic structure of the discourse. In our approach, each syllable participates in the encoding of each linguistic level, and higher levels can use either melodic or rhythmic contours to express linguistic representations.

This top-down analysis requires the careful construction of large corpora exhibiting the necessary attributes at a given linguistic level. We then extract the prosodic correlates corresponding to the linguistic attributes being tested by statistical analysis. In the following section, we apply this methodology to the design of a corpus destined for the study of prosodic morphology at the sentence level.

0.3 Data

0.3.1 A corpus-based approach

Our methodology is based on the analysis of corpora that shed light on the structural "rendez-vous" between linguistic levels and prosody. The corpus design methodology generally consists in building sentences with minimal oppositions (Delattre 1969) in phonotactic, morpho-syntactic and lexical attributes. In this paper we focus on the sentence level. In our framework the

opposition at the sentence level is the speaker's attitude. At the group level, the nature of the group (noun group (NG), verb group (VG)…), its function in the utterance (for a NG: subject, object), its syntagmatic position and its length in terms of syllables are systematically varied. This ensures balanced occurrences between each attribute. The analysis of a given attitude consists then of fixing the sentence attribute and distributing the other remaining attributes (group, length…).

0.3.2 A corpus of attitudes

Following this approach, we developed a corpus designed to reveal the existence of global prosodic prototypes associated with given attitudes at the sentence level. Our purpose is not to propose a complete inventory of prosodic attitudes in French. We rather want to demonstrate that on a common set of sentences pronounced with a limited set of attitudes, a speaker produces well-identified prosodic contours, and that these contours can be generated with our model.

The whole corpus consists of 1932 utterances pronounced by one reference speaker. It contains a set of 322 sentences with various syntactic structures and a small number of syllables (between 1 and 8 syllables) in order to minimise the number of carried contours. Six prosodic attitudes for each sentence are suggested to the speaker (during the recording sessions, short texts introduce the situations in which these sentences should be uttered): declarative (DC), question (QS), exclamation (EX), incredulous question (DI), suspicious irony (SC) and obviousness (EV) (see pragmatics and didactics of French (Callamand 1973) for the more detailed descriptions of these attitudes).

The distribution of the number of syllables for the 322 sentences is given in Table 0.1.

The table shows that the technique of minimal oppositions is applied up to 6 syllables. Seven and eight syllable utterances will be used to test the capacity for generalisation of our model (see section 0.4.2). Sentences containing up to 6 syllables have the following structure:

- Single words with mono- and poly-morphemic structures.
- Isolated Noun Groups (NG) (elliptic sentences).

Table 0.1. Distribution of sentences in the corpus of attitudes.

Number of syllables	1	2	3	4	5	6	7	8
Number of sentences	5	6	28	58	93	122	5	5

- Isolated Verb Group (VG). The VG is reduced here to the verb and its components (Aubergé 92).
- NG followed by a VG (subject+verb).
- NG followed by a verb with an adverb in initial, internal or final position.
- VG followed by a NG (verb+object).
- NG followed by a VG followed by a NG (subject+verb+object).

0.3.3 Prosodic parameters stylisation

0.3.3.1 Extraction of macrorhythm and phonemic durations

Our method for predicting phonemic durations is similar to that of Barbosa (1994). Barbosa adapted Campbell's hybrid model (Campbell 1992) to the IPCG unit (Inter Perceptual Centre Group) for which the elasticity hypothesis is better respected. We proceed in 2 steps:

1. Each IPCG (it coincides here with the interval between two consecutive vocalic onsets) is characterised by a shortening/lengthening factor called the *IPCG_Ratio* calculated as the ratio between the actual IPCG duration and a "reference" IPCG duration computed as the sum of the mean characteristic durations for each segment:

$$IPCG\_Ratio = \frac{d_{GIPC} - \sum_{p_i \in GIPC} dmoy_{p_i}}{\sum_{p_i \in GIPC} dmoy_{p_i}} \quad (1)$$

   The IPCG_Ratio will be the macrorhythmic unit of our model.

2. Segmental durations including emergence of pauses are obtained using a distribution algorithm distributing the predicted IPCG duration among its phonemic constituents. Phoneme durations belonging to each IPCG are obtained as follows:
   - The current IPCG duration is extracted from the *IPCG_Ratio*.
   - The elasticity factor *k* (Campbell and Isard 1991) associated with the current IPCG is calculated using the following formula:

$$IPCG\ duration = (IPCG\_Ratio + 1)\sum_{i=1}^{n} exp(\mu_i) = \sum_{i=1}^{n} exp(\mu_i + k\sigma_i) \quad (2)$$

   In this equation, μ_i and σ_i are the mean and standard deviation of the log-transformed durations (in milliseconds) of the speaker's realisations of the phoneme *i*.

Phonemic durations are then obtained from:

$$Phoneme\ duration = \exp(\mu_{phonème} + k\sigma_{phonème}) \quad (3)$$

0.3.3.2 F0 contour stylisation

For each utterance, the F0 curve is characterised by three values for the IPCG. These values are measured at the vocalic nucleus, as we consider that in French the vowel receives a melodic movement characterised by a second-degree polynomial function. These triplets are expressed in quarter tones:

$$F0_{1/4tons} = 24\log_2(F0_{Hz} / F0_{moyHz}) \quad (4)$$

0.3.4 Corpus analysis

0.3.4.1 From prosodic contours to prosodic movements

The analysis that follows aims at defining the notion of prosodic movement - a prototypical stretch defined according to the number of syllables within an utterance. The characterisation of these movements is a crucial step for proposing an adequate input parameterisation for our model.

This analysis begins with the description of Figure 0.7 and Figure 0.8 (see appendix) representing the melodic (left column) and rhythmic (right column) contours of single-word utterances. These utterances contain between 1 and 6 syllables and correspond to the 6 attitudes. Single words give a good representation of sentence prosodic contours by preventing contribution from the prosodic organisation of lower levels (except from the morphological level).

- Melodic analysis:
  - The left columns of Figure 0.7 and Figure 0.8 show that global melodic contours associated with the 6 attitudes are highly contrastive.
  - For each attitude, the final melodic movement is very precise. Standard deviations are also small in the first part of the contours, except for exclamations, during which morphological frontiers in polymorphemic single words may exhibit large melodic excursions.
  - Even at the sentence level, melodic contours can vary quickly. This is the case for incredulous questions with a large prominence on the penultimate IPCG and a rising movement on the last one.

- Rhythmic analysis

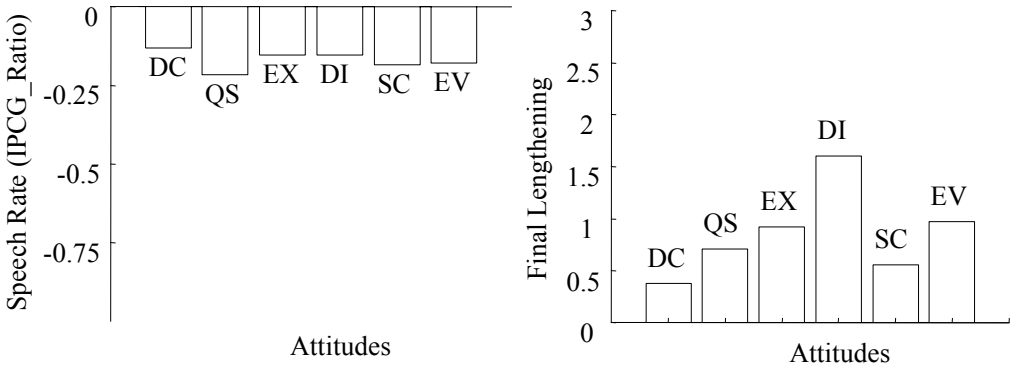

Figure 0.1. Means of non-final IPCG_Ratio (speech rate) (left) and means of final lengthening (right) for single-words and the 6 attitudes.

- For 6 syllable single-words, 2 rhythmic parameters are represented in Figure 0.1: speech rate on the left (means of IPCG Ratios except the final one) and final lengthening on the right (means of last vowel durations). We note that for single words and for every attitude speech rates are equivalent, and higher than the reference value (IPCG Ratio < 0). Rhythmic contours essentially differ by their final lengthening (IPCG Ratios are greater for attitudes with a non-monotonic final melodic pattern (DI, EX, EV)).

0.3.4.2 Prosodic movement expansion

The description given above enables us to propose a few hypotheses about the way prosodic movement expansion is performed when the number of syllables in the utterance increases:

- The hypothesis of global prosodic contours encoding speaker attitudes seems to be acoustically verified. In section 0.5.1, we present a gating experiment that perceptually validates this hypothesis.
- When the number of syllables increases, rhythmic contours remain monotonic and increasing. Final lengthening is always present and its magnitude depends on the attitude.
- Final melodic movements are even more attitude-specific. They seem to cover either the last IPCG (DC,QS) or the last two IPCGs (DI). This fi-

nal prototypical movement or *capture* is preceded by a movement spreading over the first part of the utterance (*preparation*).

- The preparation movement is also attitude-specific: the contour is however warped according to the length of the movement. This warping is linear at a first approximation.

The model we present in the next section must be able to predict both phases of prosodic realisations at the sentence level.

0.3.4.3 Discussion

The description given above shows that those prosodic contours conveying speaker attitudes sometimes exhibit large melodic variations at the sentence level. The purpose of this discussion is to give a brief inventory of current phonetic representations that should be applied to deal with such melodic contours.

The classical approach for describing prosodic realisations at the sentence level usually implies linear constructions such as declination lines (Pierrehumbert 1981; ’t Hart 1973) or tonal grids (Gårding 1991). These rigid templates can be replaced by more flexible structures, such as downstepping and pitch reset (Kohler 1997). Sentence contours can also be generated with response filter models (Öhman 1967; Fujisaki and Sudo 1971).

These descriptions of sentence contours are not sufficient to characterise precisely the prosodic prototypes we encountered during our analysis. Most of them cannot be modelled with a single sentence component of response filter models, as their sentence command generates slow melodic movements. An accent command must be added that does not correspond to any morpho-syntactic boundary. Nor can they be completely described with a simple declination (or "raising") line. Local salient events have to be added: for instance a pitch target on the penultimate IPCG of incredulous questions.

This has given rise to an important debate about the way local events and global constructs of phonetic representations are connected to a given meaning (Hirst 1991; Gussenhoven 1991). We do not participate in this debate, since we consider that prosodic contours convey speaker attitude globally, and believe that they should therefore be modelled holistically. The next section presents the current implementation of our model generating these global and coextensive prosodic contours.

0.4 Modelling

0.4.1 Overview of the complete model

The implementation of our prosodic model (Morlec 1997) is a direct application of our superpositional approach. It consists of a collection of specialised modules. Each module is in charge of predicting the expansion of contrastive prosodic movements (melody and macrorhythm) for a given linguistic level. The resulting prosody is the weighted sum of predictor outputs: each output is weighted by global factors in order to focus or reduce the contribution of any given structural level to the actual intonation.

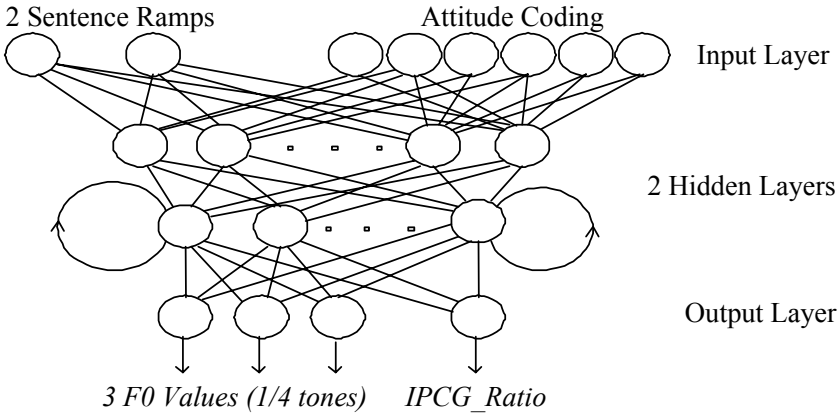

Figure 0.2. Architecture of the sentence module.

Learning is carried out progressively starting from the sentence level. Once the model has learned sentence movements correctly, the parameters of the sentence module are frozen, learning of the immediate lower level may then occur, and so on. In other words, the prosodic performance of the model improves little by little, in several stages. This modular architecture has the great advantage that it enables sequential training from appropriate corpora associated to each linguistic level. The expansion model of sentence prosodic movements will be trained with short single-word utterances of the corpus described in the preceding section. Then the typology of group contours will be obtained from more syntactically structured sentences. In this section, we focus on the training of the sentence module.

0.4.2 The sentence module

0.4.2.1 Architecture

The sentence module consists of a single Elman-like (1988) recurrent neural network (RNN) devoted to the generation of the 6 prosodic attitudes. RNNs enable the prediction of complex time-varying multiparametric movements (Jordan 1988). Our RNN has the following architecture (see Figure 0.2):

- Input:
  - A binary coding of the attitude and two IPCG linear ramps:
  - One ramp decreasing from one to zero gives proportional timing to the network and thus enables it to stretch a prototypical melodic movement.
  - One ramp decreasing from the number of syllables of the sentence down to zero enables the network to trigger the final pattern.

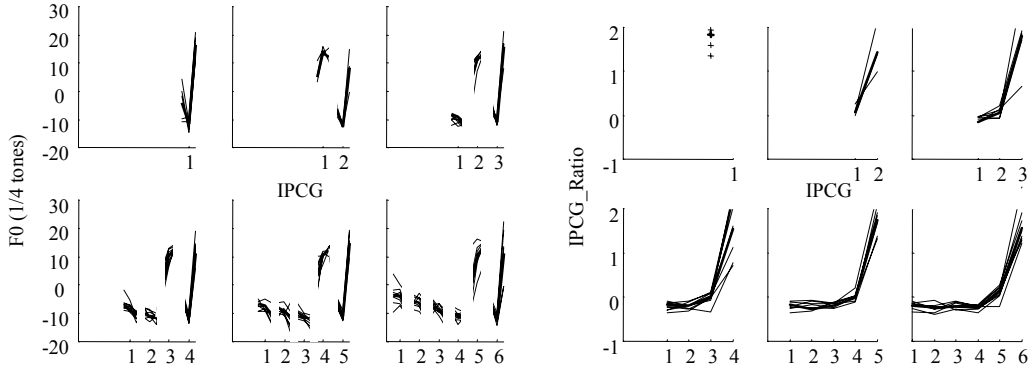

Figure 0.3. Predictions (thick lines) on the single-word training set (thin lines). F0 (top) and IPCG Ratio (bottom) movement expansion for incredulous questions.

- The input layer is fully connected to a series of two hidden layers with non-linear activation functions (*atan*). The first hidden layer receives the delayed activations of the second layer.
- Output: A stylised F0 melodic movement for each IPCG and the IPCG Ratio.

0.4.2.2 Training

The training set for this RNN consists of all single word sentences between 1 and 6 syllables. The distribution of the number of syllables for the 38 utterances is given in Table 0.2. Sentences containing 7 and 8 syllables or more are kept for the generalisation tests.

0.4.2.3 Predictions on the training set

On the training set, predictions are very close to the means of the training contours (see for example incredulous questions in Figure 0.3). The global error rate per predicted F0 is around 6 % (see Table 0.3), except for exclamations, for which large melodic variations at the word level occur. Relative error rates for IPCG durations are around 10.3 %.

Table 0.2. Distribution of sentences in the training set of the sentence module.

Number of syllables	1	2	3	4	5	6
Number of sentences	5	3	6	9	7	8

0.4.2.4 Movement expansion and extrapolation

The generalisation abilities of the RNN guarantee that the main features of each attitude are respected even with up to 10 syllable sentences. For incredulous questions, the capture phase is correctly predicted (see Figure 0.3 and Figure 0.4). In the preparation phase, the model also adequately generates the progressive increase of the initial melodic value when the number of syllables within the utterance is increased.

0.4.3 Discussion

An exhaustive analysis of the predictions shows that the sentence module is able to generate and extrapolate acoustically adequate prosodic contours for both rhythm and F0. This does not guarantee that these predictions adequately carry the original speaker's attitude. Listeners are indeed not equally sensitive to all aspects of the prosodic continuum. In the next section we propose two perceptual experiments:

- The first is designed to reveal the global nature of prosodic attitude perception, to quantify the distribution of prosodic information, and thus to assess our theoretical approach.

- The second aims at validating the sentence module with a functional test comparing natural and synthetic contour performance in a common identification task.

Table 0.3. Relative errors on F0 predicted values (in Hz) and IPCG durations (except the last IPCG of each sentence) on the training set.

Nb syllables	DC	QS	EX	DI	SC	EV
1	5.4 - *	8.6 - *	21.6 - *	10.3 - *	8.5 - *	15.6 - *
2	6.6 – 8.9	9.0 – 6.7	16.7 – 4.7	6.5 – 7.9	6.9 – 15.4	6.2 – 10.4
3	5.9 – 6.8	6.1 – 7.0	10.6 – 6.3	6.0 – 8.7	3.8 – 8.0	6.4 – 7.1

4	5.7 – 10.0	6.0 – 10.6	18.4 – 9.4	4.7 – 9.1	5.2 – 15.4	8.8 – 11.2
5	6.2 – 10.4	5.4 – 11.7	19.2 – 9.2	4.6 – 8.4	4.9 – 10.0	8.0 – 10.2
6	6.7 – 12.7	7.7 – 11.2	17.7 – 9.8	6.2 – 8.4	5.5 – 12.7	7.9 – 10.2
% total	6.2 – 10.7	6.6 – 10 .6	17.5 – 9.0	5.6 – 8.6	5.2 – 12.2	8.2 – 10.1

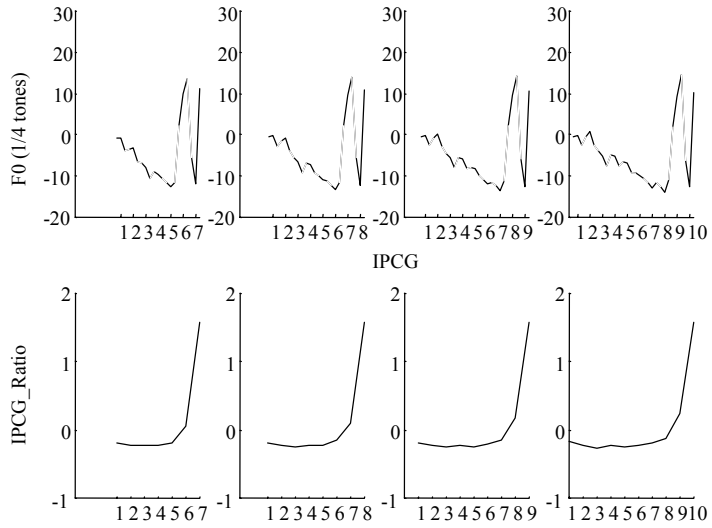

Figure 0.4. Extrapolation of prosodic movement expansion for incredulous questions between 7 and 10 syllables.

0.5 Evaluation

0.5.1 The gating experiment

The main challenge in this first perceptual experiment is to evaluate to what extent global prosodic contours at the sentence level are an actual "monnaie d'échange" between speakers and listeners. Following Grosjean (1983), we want to demonstrate that listeners use past and current prosodic features not only to process current linguistic information but also to anticipate incoming ones. In order to test this anticipatory behaviour, and to locate the moment when listeners will be able to identify the speaker's attitude, we apply the gating paradigm to sentence prosodic contours.

0.5.1.1 Stimuli

For this gating experiment, 3 sentences of 2 syllables and 6 sentences of 5 syllables are selected from the corpus described in section 0.3.2. These sentences have a [NG, VG] structure with a phrase boundary moving from fi-

nal syllable position (isolated NG) to the first syllable position (isolated VG). Each sentence is pronounced with the 6 attitudes.

Syllabic gates are used during this test: the (N-1) syllables up to the last syllable of an N-syllable utterance are hidden with white noise. The duration of this noise is calculated to produce a 2-second stimulus. This ensures that listeners are prevented from having access to information about the total duration of each utterance.

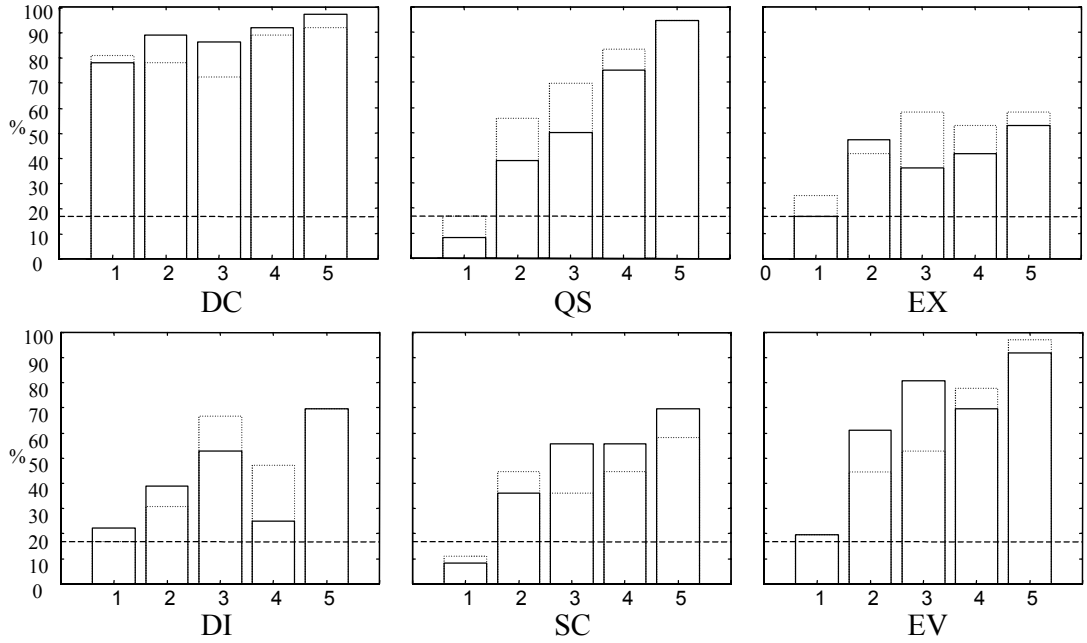

Figure 0.5. Identification rates for each gate with a normal order (solid lines) and a reverse order (dotted lines) for 5-syllable truncated utterances and the 6 attitudes.

0.5.1.2 Protocol

Twelve listeners participated in this experiment. 6 listeners first listened to the 2-syllable truncated utterances – i.e. 3x2x6=36 stimuli - and then the 5-syllable ones – 6x5x6=180 stimuli -(“normal order”).

The 6 other subjects listened to stimuli in the reverse order (“reverse order”). They were asked to associate each stimulus with one of the 6 attitudes.

0.5.1.3 Results

We focus here on 2 main results. An exhaustive analysis is given in Aubergé, Grépillat and Rilliard (1997).

- Figure 0.5 shows the identification rates for each gate and the 6 prosodic attitudes. We notice that when the entire utterance is presented (gate 5), both categories of listeners are very accurate in identifying declaratives

(DC), questions (QS) and obviousness (EV) but far less so for exclamations (EX), incredulous questions (DI) and suspicious irony (SC). DC is a preferential category in cases where there is indecision. This strong discrepancy between attitude identifications leads us to organise a survey on the relevance of chosen attitudes and their definitions (Moroni 1997).

- The most interesting result is on gate 2, for which the identification rate for each attitude is already above that for random choice (17 % is represented by a horizontal dotted line). The anticipatory behaviour also pointed out by van Heuven et al. (1997) with declaratives and declarative questions is clearly revealed. This result is consistent with our hypothesis of a mental lexicon of prosodic prototypes. Moreover it shows that even the preparation movement carries information about speaker attitude.

0.5.2 Sentence module evaluation

The sentence module can be considered as a system that converts a set of symbolic descriptors (*SD*) characterising attitudes (in our corpus [DC, QS, EX, DI, SC, EV]) into time-varying prosodic contours (*PC*). We trained a statistical predictor that learns the mapping associating *SD* and *PC*. The major challenge during this second evaluation procedure is to quantify the *functional equivalence* between natural and predicted mappings by evaluating both natural and predicted prosodic contours: they should give access to the same code and produce the same confusions.

0.5.2.1 Stimuli

Each attitude is represented by 8 “natural” utterances between 1 and 8 syllables in length: for each original utterance, stylised F0 and phoneme duration values are applied via TD-PSOLA on the corresponding utterance. Predicted versions are obtained with the same procedure by replacing original prosodic contours with predicted ones. The degradation of natural productions by applying stylised prosodic values produces a homogenous test corpus. Nevertheless, prosodic parameters other than duration and F0 contribute to the decoding of the message. Applying the computed parameters to the original utterances maintains these prosodic indices and may improve global identification results for these stimuli. In a recent perceptual experiment (Morlec, Rilliard, Bailly and Aubergé 1998), we avoided such a bias by applying stylised and predicted prosodic contours on monotonous and isochronous utterances.

0.5.2.2 Protocol

Twenty subjects participated in this experiment.

1. A training stage was used in order to familiarise subjects with the definitions of the six attitudes: 24 natural single-word utterances between 3 and 6 syllables were presented during this test.
   Subjects had to associate each utterance (presented only once) with one of the six definitions. Once they had given their choice, the correct answer was displayed. Subjects also had the possibility to enrich the given definitions with their own keywords.

2. The real identification task was then performed by the 20 subjects:
   48 natural re-synthesised utterances and the 48 corresponding synthetic ones were then presented in a random order. As for the training phase, subjects chose between the six definitions. This test was performed twice by each listener. These two runs will be referred to as *t1* and *t2*.

0.5.2.3 Results

The main results of this identification task are as follows:

- The global identification rate for the training stage is 72.9 %, with a strong discrepancy among attitudes (DC rate is 93 % whereas SC rate is 36 %).
- The average identification rate for the natural stimuli for *t1* and *t2* is 72.8 %. The confusion matrix for these utterances is given in Table 0.4.a.
- The average identification rate for the synthetic versions is 68.6 %. The confusion matrix for utterances with predicted prosody is shown in Table 0.4.b.
- The *t2* identification rate is 3.4 % higher than the *t1* rate for natural stimuli, and 5.6 % higher for synthetic ones.

Table 0.4. Confusion matrix for natural (a) vs. synthetic (b) utterances for t1 and t2.

(a)

	DC	QS	EX	DI	SC	EV
DC	88.7	0.3	0.0	0.6	2.5	7.8
QS	2.2	81.4	4.8	7.1	3.2	1.3
EX	1.6	1.6	72.9	15.0	4.4	4.5
DI	5.7	1.9	8.0	58.7	17.0	8.7
QS	9.6	3.9	3.5	25.0	48.5	9.7
SC	5.8	0.3	2.9	2.9	1.9	86.3

(b)

	DC	QS	EX	DI	SC	EV
DC	90.2	0.0	0.0	1.9	5.0	2.9
QS	2.9	83.5	3.8	5.7	3.2	0.9
EX	2.2	9.0	54.6	19.5	5.1	9.6
DI	8.0	5.2	8.0	57.0	15.7	6.1
QS	15.3	4.8	3.5	22.8	45.6	8.1
SC	10.2	1.3	1.6	2.9	3.2	80.8

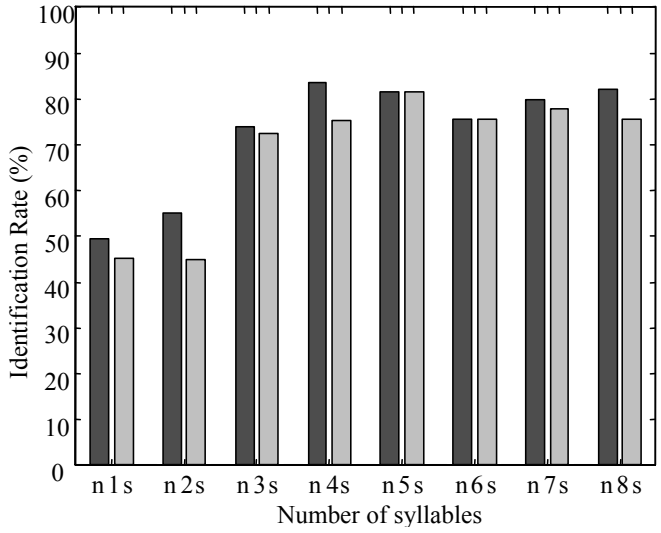

Figure 0.6. Identification rates for natural (n) and synthetic (s) versions of utterances between 1 and 8 syllables. Seven and eight syllable synthetic utterances are obtained with the generalisation abilities of the network.

Going further in the analysis of the results, we notice several interesting phenomena:

- Incredulous Question (DI) and Suspicious Irony (SC) are often confused, despite the clear difference in their prosodic features. This could be due to the definitions of these attitudes, since the situations of communication in which they may occur are very similar.

- Identification rates increase for 1, 2, 3 syllable utterances and they remain stable beyond 3 syllables (see Figure 0.6).
- Global identification rates show that natural utterances are better identified than synthetic ones, but the difference between the two rates remains small, whatever the length of the proposed utterances. Moreover, if we omit the identification rate for Exclamation, for which the lack of intra-word (morphological) modulation in the synthetic version is clearly perceptible, the five remaining synthetic attitudes are recognised as well as the natural ones.

0.6 Conclusions and future work

The analysis and modelling of a number of attitudes in French has allowed us to establish a model capable of generating the expansion of prosodic movements at the sentence level. Two perceptual evaluation tests have shown, on the one hand, the necessity of modelling contours covering the whole of a sentence (gating experiment) and, on the other hand, that the synthetic contours predicted by the model permitted the listener to access the linguistic code almost as efficiently as the original training contours.

Further experiments need to be conducted to analyse more precisely the respective roles of the different prosodic parameters in decoding the message. In addition, the present version of our multi-parameter generative model only generates a single prototypical contour for a given length and class of attitude. Future work ought to be directed towards the modelling of the topology of intra-class prosodic variants in the perceptual-acoustic space. The connectionist approach should allow us to generate these variants by controlling the initial conditions of our recurrent networks. Finally, we have already applied the same type of dynamic model to the generation of group contours conveying the syntactic structure of the utterance. Preliminary experiments on a corpus of sentences comprising three elementary syntactic structures demonstrate that prosodic contours at the level of the group can be described just as well in terms of global prototypical movements that are superposed on the sentence movements described here.

Acknowledgements

This work was supported by COST258 "Naturalness of Synthetic Speech" and AUPELF ARC B3.

References

Aubergé, V. 1992. Developing a structured lexicon for synthesis of prosody, in G.Bailly and C.Benoît (eds), *Talking Machines: Theories, Models and Designs*, Elsevier B.V., 307-321.

Aubergé, V., Grépillat, T. and Rilliard, A. 1997. Can we perceive attitudes before the end of sentences? The gating paradigm for prosodic contours*, Proceedings of EUROSPEECH'97*, Vol.2, Rhodes, Greece, 871-874.

Barbosa, P. and Bailly, G. 1994. Characterisation of rhythmic patterns for text-to-speech synthesis*, Speech Communication* 15:127-137.

Boë, L.-J. and Contini, M. 1975. Etude de l'intonation de la phrase interrogative en français (question totale). Premiers résultats, *Bulletin de l'Institut de Phonétique de Grenoble*, Université de Grenoble III, Grenoble, France.

Bolinger, D. 1989. *Intonation and its uses*, Edward Arnold, London.

Callamand, M. 1973. *L'intonation expressive*, Collection le français dans le monde, B.E.L.C. Librairies Hachette et Larousse.

Campbell, W. 1992. Syllable-based segmental duration, in G.Bailly and C.Benoît (eds*), Talking Machines: Theories, Models and Designs*, Elsevier B.V., 211-224.

Campbell, W.and Isard, S.D. 1991. Segment durations in a syllable frame, *Journal of Phonetics*, 19:37-47.

Clark H.H. and Marshall C.R 1981. Definite reference and mutual knowledge, in Joshi A.K., Webber B.L and Sag I.A (eds), *Elements of discourse understanding*, Cambridge University Press, 11-63.

Damasio, A. 1994. *Descartes' error: emotion, reason and human brain.* C. Grosset/Putnam Books.

Delattre, P.C. 1969. L'intonation par les oppositions, *Le français dans le monde* 64:6-13.

Elman, J.L. 1988. Finding structure in time, *CRL Technical Report 8801*, University of California, San Diego, Center for Research in Language, La Jolla, CA.

Fónagy, I., Bérard, E. and Fónagy, J. (1984). Clichés mélodiques, *Folia Linguistica* 17:153-185.

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

Gee, J.P. and Grosjean, F. 1981. Performance structures: A psycholinguistic and linguistic appraisal, *Cognitive Psychology* 15:411-458.

Gårding, E. 1991. Intonation parameters in production and perception*, Proceedings of the XIIth International Conference on Phonetic Sciences*, Vol.1, Aix-en-Provence, France, 300-304.

Grosjean, F. 1983. How long is the sentence? Prediction and prosody in the on-line processing of language*, Linguistica* 21:501-529.

Gussenhoven, C. 1991. Tone segments in the intonation of Dutch, in T.Shannon and J.Snapper (eds*), The Berkley conference on Dutch linguistics*, University Press of America, Lanham, MD, 139-155.

Hirst, D. 1991. Intonation models: towards a third generation*, Proceedings of the XIIth International Conference on Phonetic Sciences*, Vol.1, Aix-en-Provence, France, 305-310.

Jordan, M.I. 1988. Supervised learning and systems with excess degrees of freedom, *COINS Tech. Rep. 88-27*, University of Massachussetts, Computer and Information Sciences, Amherst - MA, USA.

McClelland, J.L. and Elman, J.L. 1986. The TRACE model of speech perception, *Cognitive Psychology* 18: 1-86.

Morlec, Y. 1997. *Génération multiparamétrique de la prosodie du français par apprentissage automatique*. PhD thesis. Institut National Polytechnique de Grenoble.

Morlec, Y., Rilliard, A., Bailly, G. and Aubergé, V. 1998. Evaluating the adequacy of synthetic prosody in signalling synthetic boundaries: methodology and first results*. 1st International Conference on Language resources and Evaluation*, Granada, Spain, 1:647-650.

Moroni, V. 1997. Enquête sur quelques attitudes prosodiques du français : définitions et interprétations, T.E.R. sciences du langage, Université Stendhal, Grenoble, France.

Murray, I.R., Arnott, J.L. and Rohwer, E.A. 1996. Emotional stress in synthetic speech: Progress and future directions. *Speech Communication* 20:85-91.

Ohala, J.J. 1996. Ethological theory and the expression of emotion in the voice*, Proceedings of the International Conference on Speech and Language Processing*, Vol.3, Philadelphia, USA, 1812-1815.

Öhman, S. E.G. 1967. Word and sentence intonation: a quantitative model, *Technical Report 2-3*, Speech Transmission Laboratory - Departement of Speech Communication and Music Acoustics - KTH, Stockholm - Sweden.

Petitot, J. 1985. *Les Catastrophes de la parole. De Roman Jakobson à René Thom*, Maloine, Paris, France.

Petitot, J. 1986. Le "morphological turn" de la phénoménologie*. Chapitre I, II et III de morphogénèse du sens II*, Centre d'analyse et de Mathématique Sociales, EHESS-CNRS.

Petitot, J. 1990. Le physique, le morphologique, le symbolique - remarques sur la vision, *Revue de Synthèse* IV(1-2): 139-183.

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

Scherer, K.R. 1996. Adding the affective dimension: a new look in speech analysis and synthesis*, Proceedings of the International Conference on Speech and Language Processing*, Philadelphia, USA.

't Hart, J. and Collier, R. 1973. Intonation by rule: a perceptual quest, *Journal of Phonetics* 1:309-327.

van Heuven, V.J., Haan, J., Janse, E. and vander Torre, E.J. 1997. Perceptual identification of sentence type and the time-distribution of prosodic interrogativity marker in Dutch, *ESCA Workshop on Intonation*: Theory, Models and Applications, Athens, Greece, 317-320

Thorsen, N.G. 1980. A study of the perception of sentence intonation - Evidence from Danish*, Journal of the Acoustical Society of America* 67(3):1014-1030.

Appendix

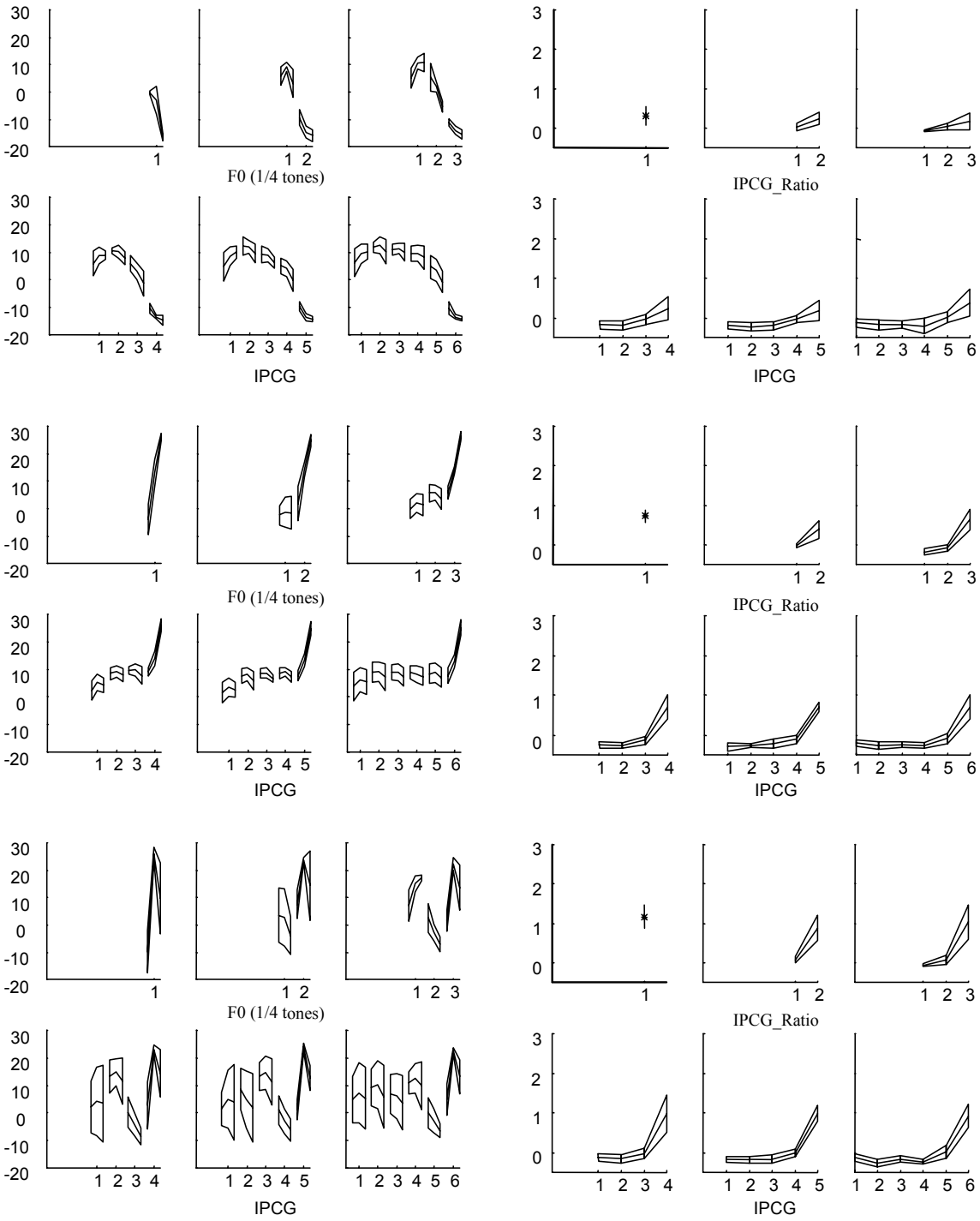

Figure 0.7. Means (dotted lines) and standard deviations (solid lines) of F0 (left column) and IPCG Ratio (right column) for single-words pronounced as declaratives (top), questions (middle) and exclamations (bottom).

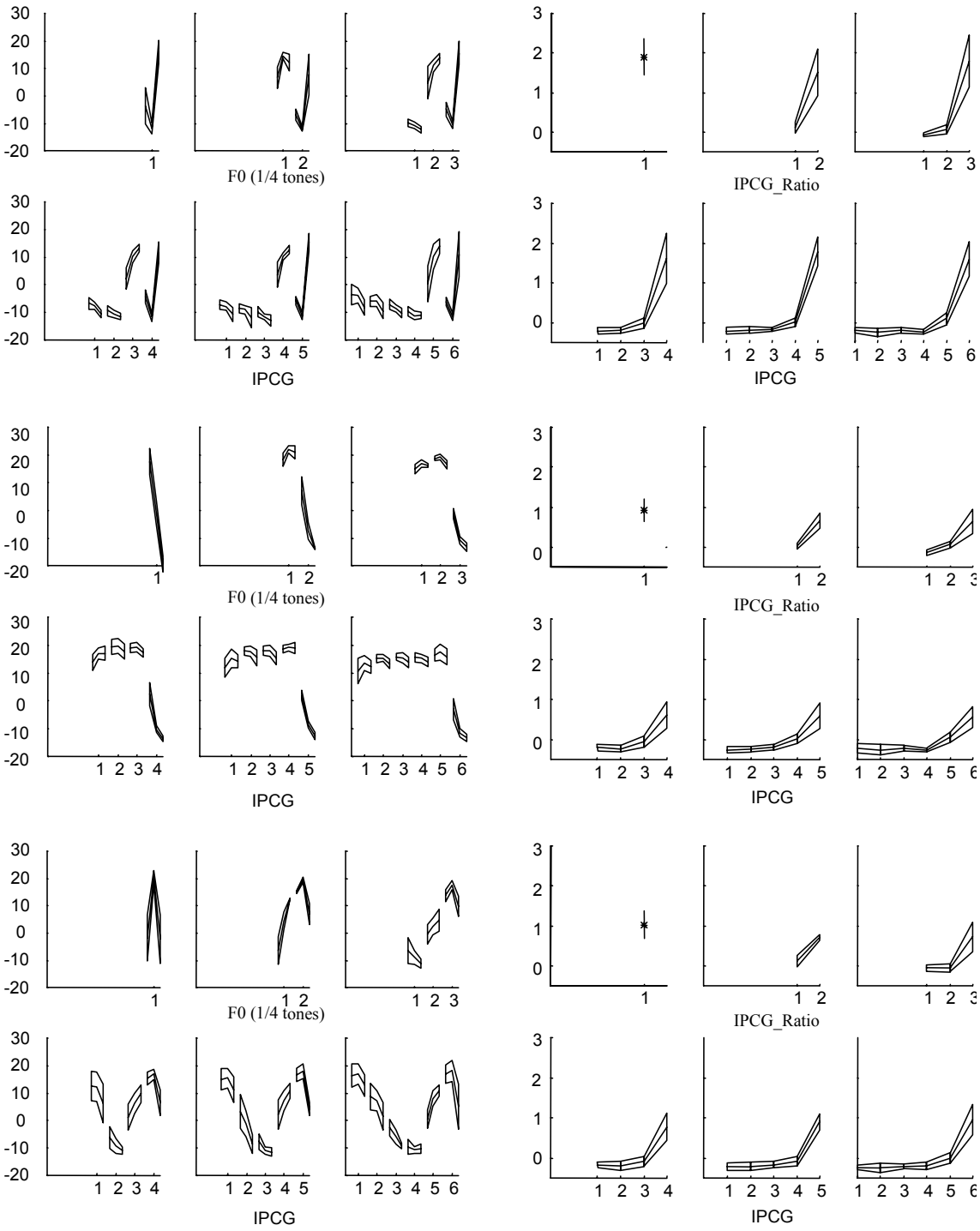

Figure 0.8. Means (dotted lines) and standard deviations (solid lines) of F0 (left column) et IPCG Ratio (right column) for single-words pronounced as incredulous questions (top), suspicious irony (middle) and obviousness (bottom).