SYNTHESISING ATTITUDES WITH GLOBAL RHYTHMIC AND INTONATION CONTOURS

Yann Morlec, Gérard Bailly et Véronique Aubergé
Institut de la Communication Parlée
46, av. Félix Viallet 38031 Grenoble CEDEX FRANCE
e-mail: (morlec,bailly,auberge)@icp.grenet.fr

ABSTRACT

We present here a trainable generative model of French prosody. We focus on the sentence level and design SNNs able to generate both rhythmic and intonation contours for diverse attitudes. First results of a perceptual test show that listeners are able to retrieve the right definition of attitudes by listening to synthetic PSOLA stimuli.

1. THEORETICAL FRAMEWORK

In our theoretical framework prosody can be described as the superposition of independent multiparametric prosodic contours belonging to diverse linguistic levels [1]: sentence, clause, group, subgroup... These prototypical movements are progressively stored in a prosodic lexicon and dynamically used by the speaker to mark (segmentation...), enlight (salience) and enrich (attitudes...) the linguistic structuration of his discourse. In our approach, each syllable participates in the encoding of each linguistic level and higher levels can use whatever melodic or rhythmic variations to express linguistic representations. This theoretical framework contrasts with most popular models described in the literature:

– Tonal approaches such as promoted by prosodic phonology where intonation is described with local events such as tones and breaks, the function of which are described [9] by higher phonological constructs such as the intonational, phonological phrase or word.

– Superpositional models only based on physical or geometric [7] parameters such as cut-off frequencies or declination lines.

– Data fitting or purely lexicon-based approaches, where synthesis is reduced to adequate and accurate labelling [4].

The model proposed here makes strong assumptions on the way linguistic and paralinguistic attributes are encoded in prosody. The main challenge of our work is to demonstrate that parameters of this model may be learned in order to adequately and accurately predict a multiparametric prosodic continuum.

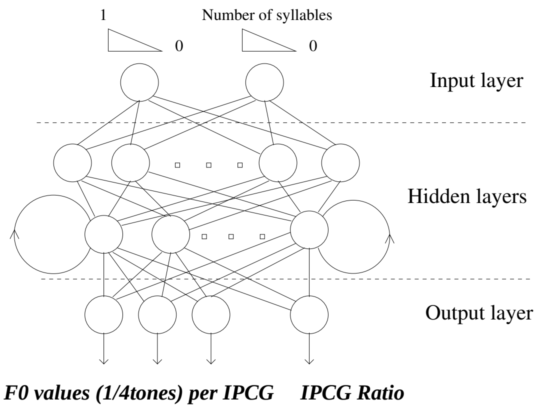

Figure 1: The sentence module architecture

2. THE MOVEMENT EXPANSION MODEL

Our current implementation of this model is a modular association of sequential neural networks (SNNs) . Each SNN is in charge of the prosodic prediction of a specific linguistic level (F0(dB) and macrorhythm). The resulting prosody is the weighted sum of SNNs outputs: each output is weighted by global factors in order to focus or reduce the contribution of given structural level in the actual intonation.

Learning is progressively carried out starting from the sentence level. Once the model has learned sentence prosodic contours correctly, the parameters of the sentence module are frozen and learning of the lower level (the clause) may occur and so on. In other words, the prosodic performance of the model is growing little by little, in several stages. This modular architecture has thus the advantage to enable a separate training by using several appropriate corpora for the generation of each linguistic level. For instance, the expansion model of sentence melodic gestures can be generated with short single-word utterances whereas the typology of group contours can be achieved with larger and more syntactically complex sentences.

3. MULTIPARAMETRIC LEARNING PROCEDURE OF THE SENTENCE LEVEL

3.1. Corpus design

We developed a corpus designed to reveal the existence of global prosodic prototypes associated with given communication needs at the sentence level [6]. This corpus contains a set of 322 utterances with various syntactic structures. The lengths are kept short (between 1 and 8 syllables) in order to minimise the number of carried contours. Six variants of each sentence were recorded: assertion, question, exclamation, incredulous question, suspicious irony and evidence.

3.2. Fondamental frequency stylisation

The 1932 melodic curves were stylised by three values per inter-perceptual-center group (IPCG) as we consider that, in French, the IPCG receives a single melodic movement characterised by a second order polynomial function.

3.3. Rhythmic patterns extraction

Plinio Barbosa's work on the choice of the rhythmic programation unit (syllable or IPCG ?) revealed the existence of rhythmic contours structured by the group final accentuation in French. Following this analysis on our corpus, we chose the IPCG unit in which phonetic realisations have a quasi linear elasticity.

Each IPCG was then characterised by a shortening/lengthening factor k as proposed in [3]. This k factor is the ratio between the actual IPCG duration and a "reference" IPCG duration computed as the sum of the mean characteristic duration of each segment[1]. Segmental durations including emergence of pauses are obtained using a repartition algorithm distributing the predicted IPCG duration among its phonemic constituents [3]. This k factor is the macrorhythm output of our sentence module.

The first step consists in computing the means and standard deviations of the durations (in milliseconds) of the phonemic realisations of our speaker. Phoneme durations belonging to each IPCG are obtained as followed:

– The current IPCG duration is extracted from the k factor.

– *z-score* associated to the current IPCG is calculated using the following formula:

$$Dur_{IPCG} = \sum_{i=1}^{n} \exp(\mu_i + z\sigma_i) = (k+1)\sum_{i=1}^{n} \exp(\mu_i)$$

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

[1]Recent studies [10, 5] showed that the independence between macrorhythm and the nature or the number of phonemes was not verified: our solution to deal with intrinsic characteristic and number of phonemes in the unit is to replace the initial internal clock with the sequence of "reference" units.

– Phonemic durations are then obtained from:

$$dur_{phoneme} = \exp(\mu_{phoneme} + z\sigma_{phoneme})$$

3.4. Architecture and learning of the group module

The sentence module is a single SNN (cf. fig. 1). Its weights and initial activities depend on the attitude to be generated. This perceptron has the following architecture :

– Two IPCG linear ramps

– Two hidden layers with non-linear activation functions (*atan*). The second hidden layer receives its own delayed activations.

– Four linear outputs: Three outputs provide the stylisation of the F0 melodic movement for each IPCG. The fourth one gives the IPCG Ratio.

The training set of this SNN consists of all single word sentences between 1 and 6 syllables: 38 sentences per attitude. Single words enable to reduce the modulation of the sentence level by the carried contours. Sentences containing 7 and 8 syllables or more are kept for the generalisation tests.

3.5. Results

On the training set, predictions made by the sentence module rest within the statistics done on the training sentence patterns (see an example in fig. 2). The global relative rate per F0 predicted value is around 6 % (cf. table 1) except for exclamations for which larger modulations at the word level occur. Relative errors for segmental durations are around 13.6 % and have to be compared to the 10.3 % for the IPCG durations: perceptual experiments [3] showed that listeners are more sensitive to relative PC locations than exact segmental matching. Predicted F0 and rhythmic structure are almost indistinguishable for the original utterance exept for exclamations. The generalisation abilities of the SNN show that the main features of each attitude were respected even with longer sentences.

length	AS	QU	EX	IQ	SI	EV
1	5.5:12.0	8.1:10.4	22.9:12.5	10.2:8.6	9.8:12.8	15.3:16.6
2	6.6:10.3	9.3:13.7	19.9:22.4	6.9:15.2	7.3:10.8	6.3:9.0
3	5.9:14.1	6.2:9.4	13.3:10.7	5.9:12.6	3.7:10.5	6.5:9.8
4	5.7:12.9	6.1:12.2	20.2:12.6	4.8:12.8	5.3:16.1	8.9:11.0
5	6.2:13.0	5.3:12.1	19.1:11.8	4.7:12.2	4.9:13.7	8.0:10.4
6	6.9:17.3	7.5:16.4	17.4:16.2	6.1:11.2	5.8:16.8	8.0:15.3
% sum	6.3:14.3	6.6:13.2	18.3:13.7	5.6:12.1	5.2:14.8	8.2:12.2

Table 1: Relative errors on F0 values (in Hertz) and phoneme durations (in ms) for the training set sentences. AS (assertions), QU (questions), EX (exclamations), IQ (incredulous questions), SI (suspicious irony), EV (evidence)

4. PERCEPTUAL EVALUATION

An experiment was designed to evaluate the perceptual relevance of the predicted prosodic prototypes generated by the sentence module.

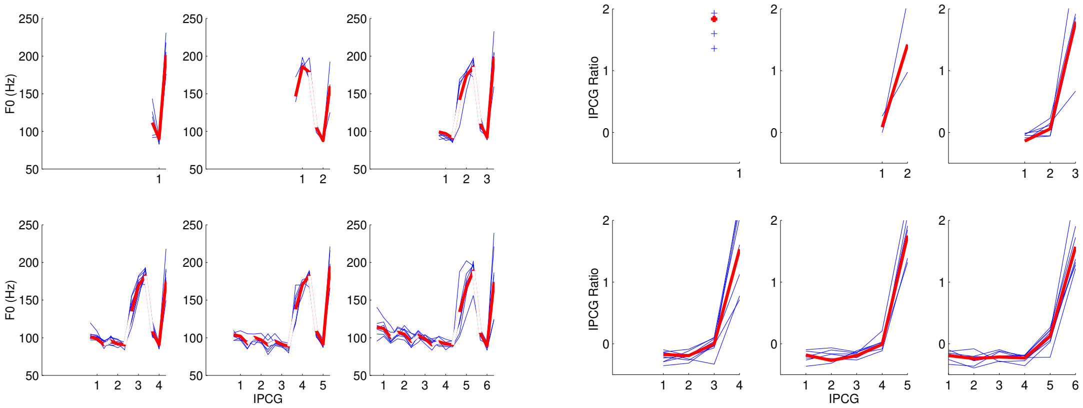

Figure 2: Superposition of F0 (left) and IPCG Ratio (right) predictions (thick lines) with the training set (thin lines) for incredulous questions: as evidenced for F0 [8], prototypical contours emerge from both durational and melodic parameters.

4.1. Method

Twenty subjects participated in this experiment:
– A training stage is used in order to get people accustomed to the definitions of the six attitudes. 24 natural single-word utterances between 3 and 6 syllables are presented during this test. Subjects have to associate each utterance (presented only once) with one of the six definitions. Once they have given their choice, the correct answer is displayed. Subjects also have the possibility to enrich given definitions with their own keywords.
– Then, 48 natural utterances (six attitudes, 1 to 8 syllabes) and the 48 corresponding synthetic versions (generated using a high-quality TD-PSOLA analysis-resynthesis technique) with predicted F0 and phoneme duration values are presented in a random order. As for the training phase, subjects choose between the six definitions. This test is performed twice by each listener. These two runs will be refered to *t1* and *t2*.

4.2. Results

The main results of this identification task follow:
– The global identification rate for the training stage is 72.9% with a strong discrepancy among attitudes (AS rate is 93 % whereas SI rate is 36 %).
– The average identification rate of natural stimuli for *t1* and *t2* is 72.8%. The confusion matrix for these utterances is given in table 2.a.
– The average identification rate of synthetic versions is 68.6%. The confusion matrix for utterances with predicted prosody is proposed in table 2.b.
– *t2* identification rate is 3.4 % higher than *t1* rate for natural stimuli and 5.6 % higher for synthetic ones.

4.3. Discussion

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

(a)

	AS	QU	EX	IQ	SI	EV
AS	88.7	0.3	0.0	0.6	2.5	7.8
QU	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
IQ	5.7	1.9	8.0	58.7	17.0	8.7
SI	9.6	3.9	3.5	25.0	48.5	9.7
EV	5.8	0.3	2.9	2.9	1.9	86.3

(b)

	AS	QU	EX	IQ	SI	EV
AS	90.2	0.0	0.0	1.9	5.0	2.9
QU	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
IQ	8.0	5.2	8.0	57.0	15.7	6.1
SI	15.3	4.8	3.5	22.8	45.6	8.1
EV	10.2	1.3	1.6	2.9	3.2	80.8

Table 2: Confusion matrix for (a) natural versus (b) synthetic utterances for *t1* and *t2*.

– During the first experiment, the association of prosody with its linguistic code is very quickly carried out as it was shown before in a similar identification task (see [2] in these proceedings).
– *Incredulous Question* and *Suspicious Irony* 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 are increasing for 1,2 ,3 syllable utterances and they remain stable beyond 3 syllables (see fig. 3).
– Global identification rates show that natural utterances are better idendified than synthetic ones but the difference between the two rates remains small wathever the length of proposed utterances. Moreover, if we omit identification rate of *Exclamation* for which the lack of intra-word (morphological) modulation is clearly perceptible, the five remaining synthetic attitudes are as well recognised as natural ones.

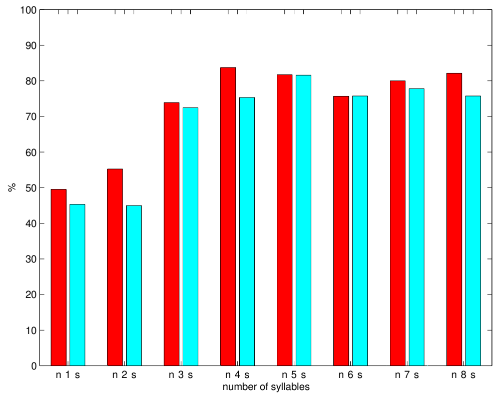

Figure 3: Identification rates of the second experiment for natural (n) and synthetic (s) versions of utterances between 1 and 8 syllables. 7 and 8 syllable synthetic utterances obtained with the generalisation abilities of the network compete very well with natural ones.

5. LEARNING THE GROUP MODULE

Statistical and perceptual evaluations of prosodic predictions of the sentence module show that our SNN is able to generate adequate prosodic contours for both rhythm and fundamental frequency. The next step will be to freeze this module and to go further into our hierarchical training.

The main challenge of our present work is to demonstrate that group prosodic contours resulting from the subtraction of sentence prosodic prototypes from the actual prosodic realisation are similar for a given syntactic structure and whatever the attitude.

As group modulations do not emerge from *Question*, *Incredulous Question*, *Suspicious Irony* and *Evidence* prosodic contours, we just keep *Assertion* and *Exclamation* to train the group module. Figure 4 represents the residual group modulation for 6 syllable *exclamative* and *assertive* sentences with a common syntactic structure composed of a 4 syllable Noun Group followed by a 2 syllable Verb. Note the similar variations of these two prosodic movements.

The Group module is fed with such utterances and the prosodic generation resulting from its predictions combined with sentence prototypes will be presented at the conference.

6. CONCLUSIONS

We aim at demonstrating that a morphologic model can emerge from the hierarchical training of global prosodic contours focusing iteratively on several linguistic levels. We show that sentence contours may be adequately described by prototypical movements and specific Prosodic Movement Expansion Models. Recent improvements of our model enable the generation of global rhythmic and intonation contours in a single architecture. Perceptual experiments are and will always be achieved at each level

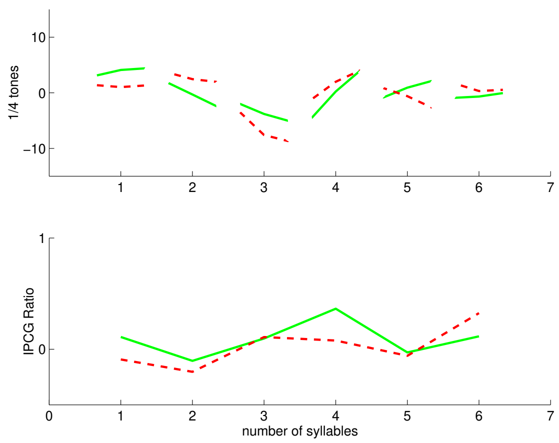

Figure 4: Group prosodic contours for 6 syllable assertions (plain lines) and exclamations (dotted lines) with a 4 syllable noun group and a 2 syllable verb.

of this iterative training to show that listeners are able to retrieve linguistic information from our synthetic prosody.

REFERENCES

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

[2] Aubergé, V. and Grépillat, T. Can we perceive intonation attitudes before the end of sentences? the gating paradigm for prosodic contours. In *Proceedings of the European Conference on Speech Communication and Technology*, Rhodes - Greece, (to appear).

[3] Barbosa, P. and Bailly, G. Generation of pauses within the z-score model. In van Santen, J. P. H., Sproat, R. W., Olive, J. P., and Hirschberg, J., editors, *Progress in Speech Synthesis*, pages 365–381. Springer Verlag, New York, 1997.

[4] Campbell, W. Synthesizing spontaneous speech. In Sagisaka, Y., Campbell, N., and Higuchi, N., editors, *Computing prosody: Computational models for processing spontaneous speech*, pages 165–186. Springer Verlag, 1997.

[5] Fant, G. and Kruckenberg, A. On the quantal nature of speech timing. In *International Conference on Speech and Language Processing*, volume 3, pages 2044–2047, Philadelphia - USA, 1996.

[6] Fónagy, I., Bérard, E., and Fónagy, J. Clichés mélodiques. *Folia Linguistica*, 17:153–185, 1984.

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

[8] Morlec, Y., Bailly, G., and Aubergé, V. Generating intonation by superposing gestures. In *International Conference on Speech and Language Processing*, volume 1, pages 283–286, Philadelphia - USA, 1996.

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

[10] van Santen, J. P. H. Segmental duration and speech timing. In Sagisaka, Y., Campbell, N., and Higuchi, N., editors, *Computing prosody: Computational models for processing spontaneous speech*, pages 225–249. Springer Verlag, 1997.