GENERATING THE PROSODY OF ATTITUDES

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

This paper presents a superpositional model of prosody where linguistic structures are directly encoded into the prosodic parameters via global melodic and rhythmic contours. This model is applied here to the generation of prosody specific to six common attitudes in French. Perceptual tests using high-quality TD-PSOLA re-synthesis show that predicted contours yield the same identification scores as natural stimuli.

1. INTRODUCTION

The basic functions supported by synthetic prosody in current text-to-speech systems are salience and segmentation. However, besides its intrinsic ability to help the listener in chunking utterances in linguistically-relevant parts-of-speech, intonation is an essential means for signaling how we feel about what we say (attitudes), or how we feel when we say (emotions) [5]. The need of an extented repertoire of emotions and attitudes is highly desirable for blind people as well as for designing person-system interfaces.

While salience and segmentation may be described with a limited set of syllabic-sized prosodic events, we will demonstrate here that prosodic encoding of attitudes is achieved via global contours spreading over the entire utterance. This observation assesses our model of prosody which generalises this direct prosodic encoding of linguistic functions: prosodic information is encoded into global contours which are 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 multiparametric prosodic contours which belong to hierarchical linguistic levels: sentence, group, subgroup [1]... These prototypical movements are stored into a prosodic lexicon and dynamically used to mark, enlight and enrich the linguistic structuration of the 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 [12] by higher phonological constructs such as the intonational, phonological phrase or word.

– Superpositional models only based on physical or geometric [9] 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 [6].

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.

2. A CORPUS OF ATTITUDES

We recorded a corpus designed to reveal the existence of global prosodic prototypes associated with six given attitudes in French [8]: *assertion*, *question*, *exclamation*, *incredulous question*, *suspicious irony* and *evidence*.

2.1. The IPCGs

The 1932 melodic curves (322 utterances times 6 attitudes) were stylised by three values per inter-perceptual-center group (IPCG). The IPCG coincides with major energy increase in the 300-2000 Hz bandwith within the syllable.

2.2. Macrorhythm and phonemic durations

Each IPCG was also characterised by a shortening/lengthening factor k (IPCG Ratio). This k factor is the ratio between the actual IPCG duration and a "reference" IPCG duration computed as the sum of the mean characteristic durations for each segment. Segmental durations including emergence of pauses are obtained using a repartition algorithm distributing the predicted IPCG duration among its phonemic constituents [4]. This k factor is the macrorhythmic 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 follows:

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

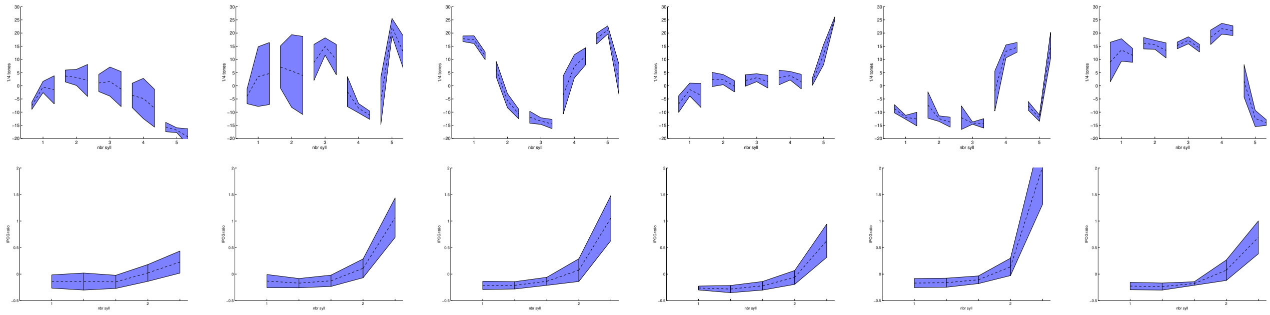

Figure 1: F0 (top) and IPCG Ratio (bottom) means and standard deviations for *assertion*, *exclamation*, *evidence*, *question*, *incredulous question* and *suspicious irony* for 5 syllable utterances with various syntactic structures. As evidenced for F0 , prototypical contours emerge from both durational and melodic parameters .

– *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*.

– Phonemic durations are then obtained from:

$$dur_i = \exp(\mu_i + z\sigma_i)$$

3. A CONNEXIONIST IMPLEMENTATION

The implementation of our superpositional model is an association of sequential neural networks (SNNs) modelling the expansion of contrastive prosodic "gestures" (Melody (dB) and macrorhythm) associated with each linguistic level (cf. fig. 3) [10]. The resulting prosody is the weighted sum of SNNs outputs: each output is weighted by global factors in order to focus on 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 correctly learned sentence prosodic contours, the parameters of the sentence module are frozen and learning of the lower level - the group - may occur and so on.

We will first focus on the sentence level. We designed SNNs able to generate both rhythmic and melodic contours for diverse attitudes. A perceptual test showed that listeners were able to retrieve the right definition of these attitudes by listening to predicted versions.

We then added a group module generating the modulations conveying the syntactic structurations of sentences. These modulations clearly emerge only for some of these attitudes.

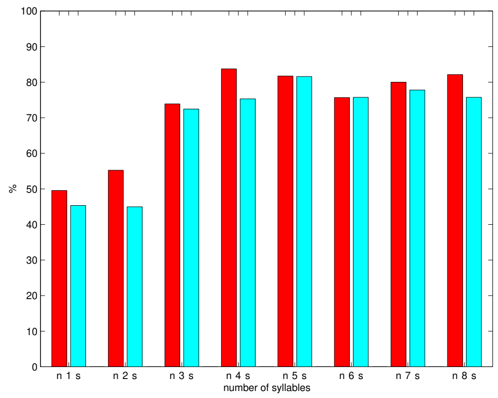

Figure 2: Identification rates of the second experiment for natural (n) and synthetic (s) versions of utterances between 1 and 8 syllables. 7 and 8 syllable synthetic utterances obtained with the generalisation abilities of the network compete very well with natural ones.

4. TRAINING AND TESTING THE SENTENCE MODULE

This first SNN (cf. fig. 3) achieves a non-linear mapping between two sequences:

– Input: A binary coding of the attitude and two IPCG linear ramps.

– Two hidden layers with non-linear activation functions (*atan*). The second hidden layer receives its own delayed activations [7].

– Output: The stylisation of the F0 melodic movement for each IPCG and the IPCG stretching ratio.

The training set of this SNN consists of all single word sentences between 1 and 6 syllables. Sentences containing 7 and 8 syllables or more are kept for the generalisation tests.

On the training set, the global error rate per predicted F0 is around 6 % except for exclamations for which larger modulations at the word level occur. Relative error rates for IPCG durations are around 10.3 %. The generalisation

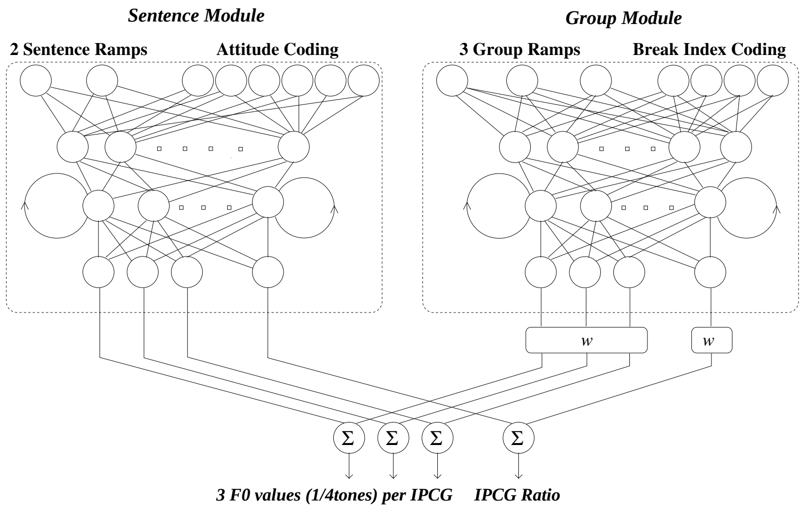

Figure 3: The sentence module and group module architectures

abilities of the SNN show that the main features of each attitude were respected even with long sentences.

5. PERCEPTUAL EVALUATION

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

5.1. 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 referred as *t1* and *t2*.

5.2. 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 1.a.
– The average identification rate of synthetic versions is 68.6%. The confusion matrix for utterances with predicted prosody is proposed in table 1.b.
– *t2* identification rate is 3.4 % higher than *t1* rate for natural stimuli and 5.6 % higher for synthetic ones.

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

5.3. Discussion

Going further in the analysis of the results, we can notice several interesting phenomena:
– 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 [2].
– *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 ut-

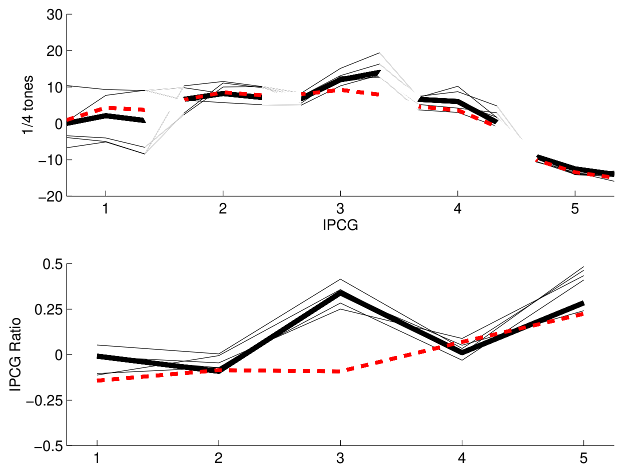

Figure 4: Prosodic prediction (thick line) resulting from the superposition of the sentence (dotted line) and group contours for 5 syllable assertions (thin lines) with a 3 syllable NG and a 2 syllable VG.

terances and they remain stable beyond 3 syllables (see fig. 2).

– 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 in the synthetic version is clearly perceptible, the five remaining synthetic attitudes are as well recognised as natural ones.

6. TRAINING 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 is to freeze this module and to go further into our hierarchical training by setting the group module weigths w to 1. As both rhythmic and melodic group modulations essentially emerge from *Assertions* and *Exclamations* (cf. standard deviations in fig. 1), we just keep the utterances associated with these two attitudes to train the group module. Syntactic structures of utterances of the training set are here limited to NG + VG with any combination of syllable numbers for NG & VG, up to 5 syllables. The group module has an internal architecture similar to the sentence module (cf. fig. 3) but inputs are different:

– Two IPCG linear ramps giving the distance between the current IPCG and the end of the current group. One IPCG linear ramp has been added to indicate the distance between the current IPCG and the end of the next group in order to manage the merging or deletion of the prosodic frontier associated with the current group.

– Four binary inputs coding the break index of the current group [3].

First tests on *assertions* give very natural sounding prosodic (sentence + group prototypes) predictions. We are now testing the prosodic movement expansion at the group level by predicting the prosody of utterances which have group units longer than five syllables. We are also carrying out a perceptual evaluation test using the "mamama" delexicalisation method [11].

7. CONCLUSION

A superpositional model of prosody has the advantage of expliciting the contributions of the different linguistic levels to the overall prosodic parameters. This approach is essential for preserving features of natural prosody such as prosodic priming [2]. We show here that it is also efficient for statistical prosodic generation.

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., Grépillat, T., and Rilliard, A. Can we perceive 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] Bailly, G. Integration of rhythmic and syntactic constraints in a model of generation of French prosody. *Speech Communication*, 8:137–146, 1989.

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

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

[6] 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.

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

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

[9] 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.

[10] Morlec, Y., Bailly, G., and Aubergé, V. Synthesising attitudes with global rhythmic and intonation contours. In *Proceedings of the European Conference on Speech Communication and Technology*, Rhodes - Greece, to appear.

[11] Rilliard, A., Aubergé, V., Bailly, G., and Morlec, Y. Vers une mesure de l'information linguistique véhiculée par la prosodie. In *1ères JST FRANCIL*, pages 481–487, Avignon–France, 1997.

[12] 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.