GENERATING INTONATION BY SUPERPOSING GESTURES

Y. Morlec, G. Bailly and V. Aubergé

Institut de la Communication Parlée, INPG & Université Stendhal
46, av. Félix Viallet 38031 Grenoble Cedex 01, France
e-mail: (morlec, bailly, auberge)@icp.grenet.fr
http://www.icp.grenet.fr/~morlec/

ABSTRACT

A dynamic model for synthesizing intonation is presented. This model is based on the following assumptions: intonation is the result of superposed and independent prototypical gestures belonging to diverse linguistic levels: sentence, clause, group, sub-group ... Prototypical movements are progressively stored in a prosodic lexicon and used by the speaker in given communication tasks. Our recent application of this model is an association of sequential neural networks (SNNs). Each dynamic module is in charge of the melodic prediction of a specific linguistic level. The resulting melody is the weighted sum of SNNs outputs. We presently focused on the sentence level. We built a corpus consisting of various length utterances pronounced with 6 attitudes. We then designed SNNs able to perform the expansion of prosodic sentence movement. Preliminary results show that these simple SNNs can give acceptable F0 prediction and keep essential features of each attitude whatever the syllabic length of the sentence.

1. INTRODUCTION

Superposition of prosodic contours to generate intonation has already been used : Thorsen [1] and Gårding [2] proposed a superposition of declinaison lines whereas Fujisaki adopted a superposition of filter responses [3]. Although being superpositional, our model does not a priori impose any constraint on the global shapes nor on cut-off frequencies. Each syllable participate to the encoding of each linguistic level and higher levels (such as the sentence level) can use whatever melodic or rhythmic variations to encode linguistic representations. In this approach, contours emerge from concurrent learning: they must be distinctive and prototypical. We claim that linguistic structure is partially encoded via contours specific to each linguistic level and which encode level-specific linguistic attributes: sentence contours signal the speaker's position *vis-a-vis* his own discourse such as attitudes (see below), movements at group level signal at least the degree of cohesion of adjacent units whereas micromelody helps segmentation and identification of consonants.

We previously proposed two automatic F0 generation methods:
- An additive superposition of melodic prototypes organized in a structured lexicon - indexed by phonotactical constraints [4] or stored as polynomials [5]. None of these methods satisfactorily takes into account the syllabic size of the cued linguistic unit in the actual synthesized F0 contours.
- A SNN trained to predict F0 sequences from elementary ramps [6] triggered by onsets of linguistic units. In this former model, the melodic contours cueing each linguistic level were simultaneously generated by a single SNN. The encoding process of a given

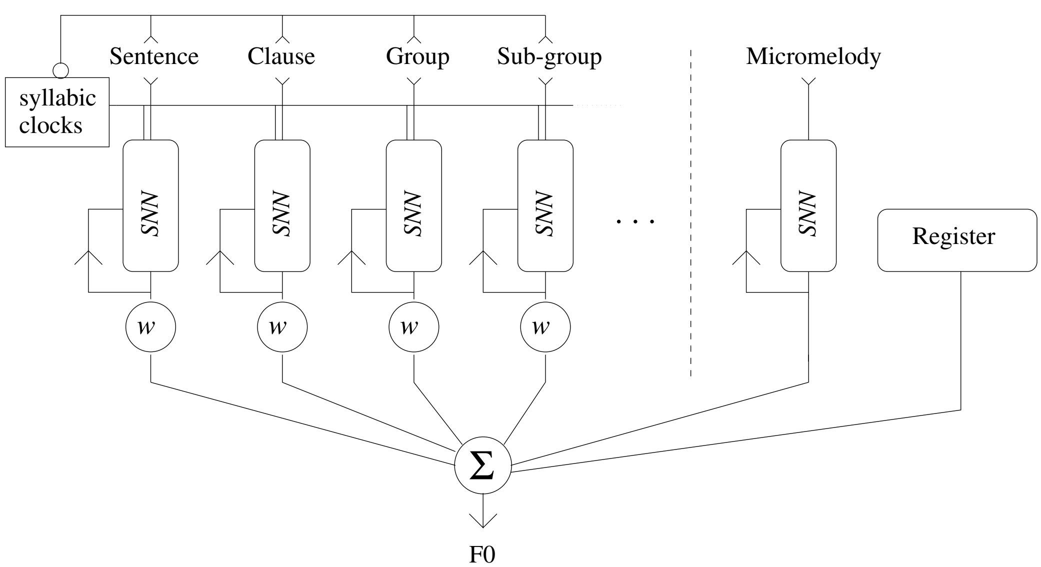

Figure 1: Full architecture of the intonation generation model. Each SNN output is in quarter tones. The register gives the mean laryngeal frequency. the weights *w* enables the hierarchical training and the tuning of the carried contours during the generation process.

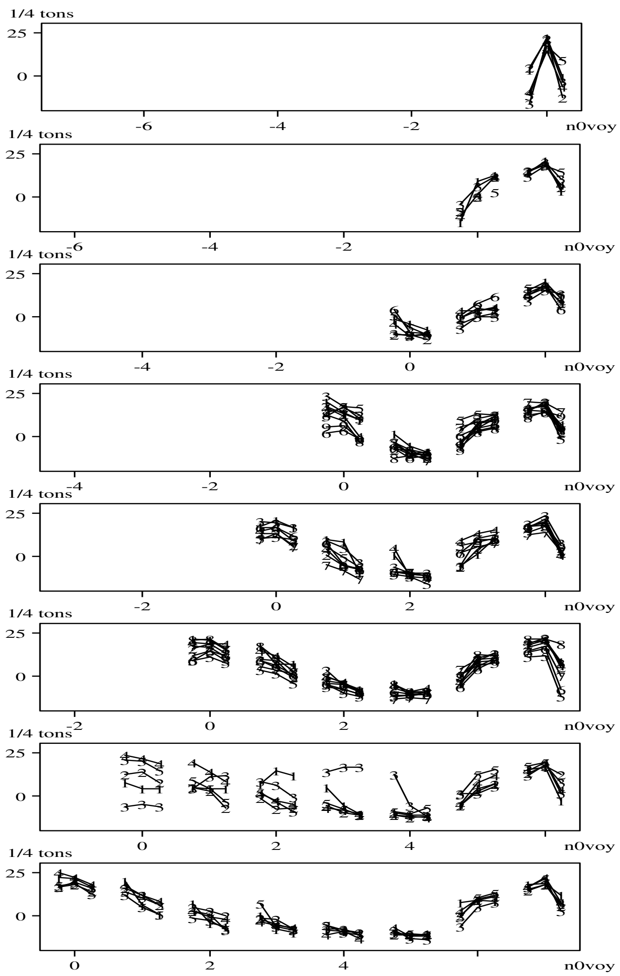

Figure 2: Melodic contours of 1 to 8 syllables sentences uttered with the attitude of evidence.

linguistic level was thus highly non-linear and dependent of the other levels. Moreover the lack of an explicit prosodic model has major drawbacks: all linguistic levels have to be present in the training corpus to ensure statistical significance; inclusion of a new level or a new linguistic attribute implies the full retraining of the SNN. We thus included the superpositional prosodic model in a neural network structure: the F0 output is produced by a linear neuron which receives the output activities of several specialized SNNs. Each of these SNNs handles the encoding of a given linguistic level. This enable to control a step-by-step learning of independent melodic generation modules.

2. THE MOVEMENT EXPANSION MODEL

Our current implementation of this model is a modular association of sequential neural networks (SNNs) (Figure 1). Each SNN is in charge of the melodic prediction of a specific linguistic level. The resulting melody is the weighted sum of SNNs outputs: each output is weighted by global factors in order to focus or reduce the contribution of any given structural level in the actual intonation. The automatic learning of such superposition model favors a top-down approach : the system learns how language-specific prosodic contours carry given linguistic contrasts. Learning is progressively done starting from the sentence level. Once the model has learned sentence gestures correctly (all weights of the sum are set to zero except for the sentence module), the parameters of the sentence module are frozen and learning of the immediate lower level (the clause) may occur (by setting its weight to one) 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 great 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 as described below whereas the typology of group contours can be achieved with larger and more syntactically complex sentences.

3. THE SENTENCE LEVEL

3.1. Corpus design and analysis

The corpus design methodology generally consists in building sentences with minimal pairs taken in phonotactic, morpho-syntactic and lexical attributes. Following this, we developed a corpus designed to reveal the existence of global prosodic prototypes associated with given communication needs at the sentence level. This corpus contains a set of 322 utterances with various syntactic structures and short lengths (between 1 and 8 syllables) in order to minimize the number of carried contours. Short texts introduce the situations in which these sentences should be uttered. Six variants of each sentence were suggested: 3 modalities - assertion, question and exclamation - and 3 attitudes - incredulous question (as described by Fónagy in [7], suspicious irony and evidence (see [8] for the explanations of these attitudes). A perceptual test attested the ability of our speaker to actually perform this contrastive task.

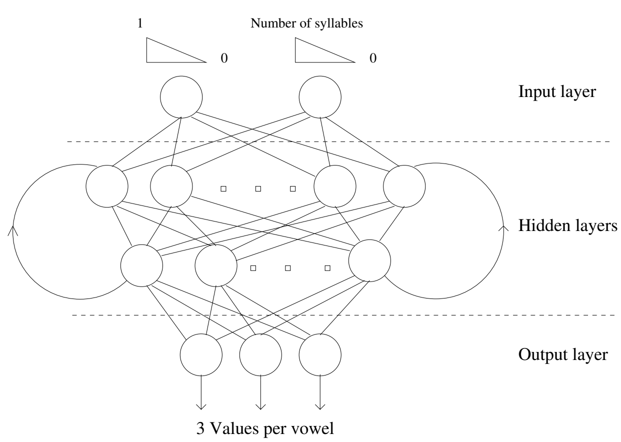

Figure 3: The sentence module architecture.

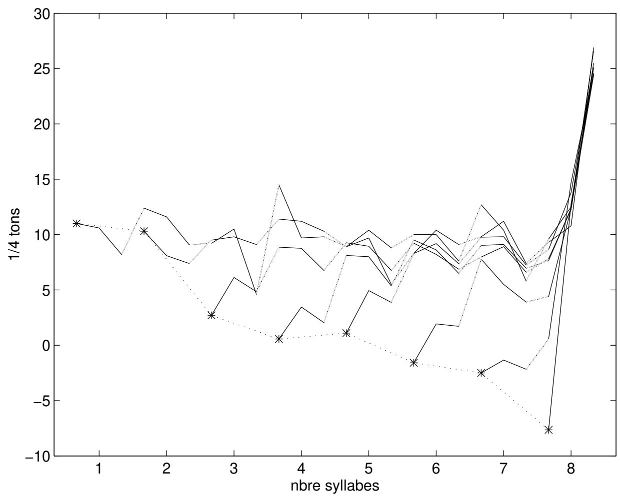

Figure 4: Means of stylized F0 for single-word sentences (between 1 and 8 syllables) pronounced as questions. Sentences from 1 to 6 syllables represent the training set of the network. 7 and 8 syllables sentences are used for generalization tests.

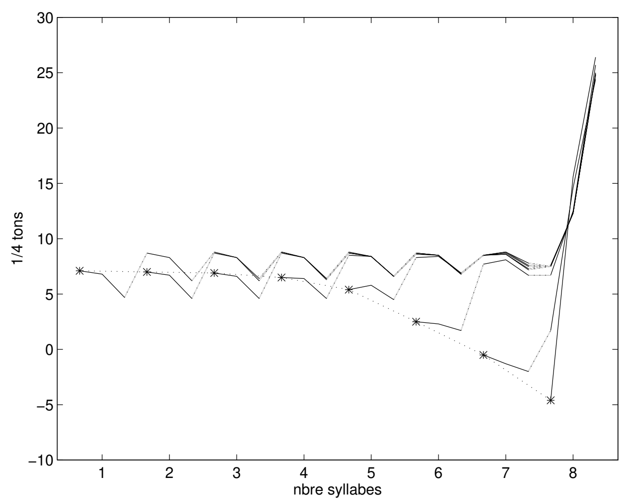

Figure 5: Melodic predictions of the sentence module for questions (1 to 6 syllables) and generalization to 7 and 8 syllables.

Figure 2 shows the melodic superposition of sentences between 1 and 8 syllables (mainly single-word) pronounced as evidence. This figure as those representing the other modalities or attitudes [9] gives the essential features of the Prosodic Movement Expansion (PME). The contours can be divided into three parts: (a) a final prototypical pattern covering typically one or two syllables preceded by both (b) an initial gesture (starting with a relatively stable initial F0 value) and (c) a relaxation. This observation leads to apply the input parameters of the sentence module described in section 3.2.

The 1932 melodic realizations are stylized by three F0 values per syllable as we consider that, in French, the syllable receives a single melodic movement characterized by a second degree polynomial function.

3.2. Sentence module architecture

The sentence module consists of a set of 6 SNNs devoted to the generation of the 6 types of melodic realizations. The perceptrons have a similar architecture (see Figure 3) :

- Two syllabic linear ramps: (a) a ramp decreasing from one to zero gives a proportional timing to the network and thus enables it to stretch a prototypical melodic gesture, (b) a ramp decreasing from the number of syllables of the sentence down to zero enables the network to trigger the final pattern.
- 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.
- Then three linear outputs provide the stylization of the F0 melodic movement for each syllable.

3.3 Training and prediction results

The training sets of these SNNs consists of all the single-word sentences between 1 and 8 syllables of our corpus: 18 sentences per modality or attitude. Single-words enable to reduce the modulation of the sentence level by the carried contours: they prevent from the clause, group and sub-group organization. Sentences containing 7 and 8 syllables (test set) are kept for the generalization tests. In order to avoid overtraining, the number of training steps (number of melodic sentence patterns presented to the network) is fixed to 1000 whatever the modality or attitude.

On the training set, predictions made by the sentence module are within the statistics done on the training sentence patterns (errors between μ-σ et μ+σ). Table 1 gives the precision of these predictions. Note that the error percentage for exclamation is higher than the others. This is due to the larger modulation of some items at the word level (secondary accents often occur when the single-word typically exceeds 5 syllables). Figure 4 and 5 enable to compare the means of the training contours and the network predictions for questions between 1 and 6 syllables. These figures lead to the following remarks :

- The sentence module is able to adequately mimic the final movement on the last syllable.
- More efforts must be done to obtain better initial F0 values.

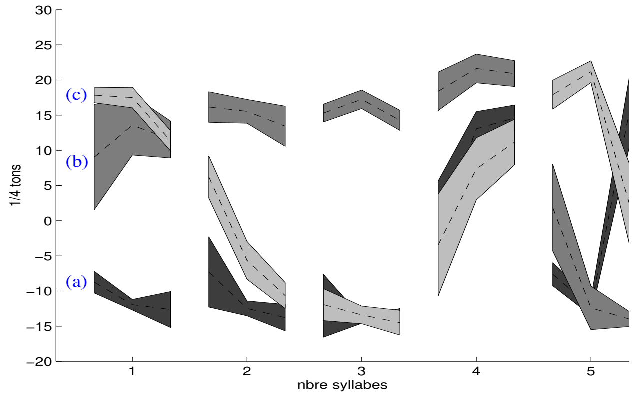

Figure 6: Means (in dotted lines) and standard deviation (solid lines) of stylized F0 for 5 syllable sentences pronounced as incredulous questions (a), suspicious irony (b) and evidence (c).

The generalization abilities of our SNNs were tested with 7 and 8 syllables length sentences (figure 4 and 5). Global results shows that the main features of each modality or attitude were respected even with long sentences. However, we can notice that the melody with such long utterances becomes relatively monotone because of the lack of lower level prosodic modulation.

Modality or attitude	Relative error per F0 predicted value (%)
Assertion	8.9
Question	7.8
Exclamation	15.3
Incredulous question	8.9
Suspicious irony	7.9
Evidence	10.1

Table 1: Relative errors committed by our SNN's when predicting F0 values (in Hertz) for the training set sentences.

4. ASSESSMENTS

Perceptual experiments are presently designed to assess our prosodic model and will be presented at the conference. A first test will be performed with a large number of listeners to evaluate the performance of our speaker. A second perceptual test is currently designed to show that prosodic information is distributed along the utterance and that speakers actually use global prosodic movements as « exchange money ». In that way, we are using the gating paradigm (also used in [1]) where syllables are progressively masked with white noise. Figure 6 shows the statistical analysis of the stimuli corresponding to the 3 attitudes and utterances with 5 syllables. We can assume that the listening of the first syllable may be sufficient to recognize incredulous question (a) as its initial F0 values are very low (-10 1/4tones) whereas suspicious irony (b) and evidence (c) may be confused. By listening the first two syllables, the listener may be able to distinguish the 3 attitudes. These hypotheses should be confirmed in the near future and correlation between perceptual confusion matrix and physical information should be discussed.

5. CONCLUSIONS

We aim at demonstrating that a morphologic model can emerge from the hierarchical training of global prosodic contours. A training strategy focusing iteratively on several linguistic levels can be compared to the spotlight theory [10]: prosody aims at enlarging the linguistic competence of a child thanks to successive focus. We show that sentence contours may be adequately described by prototypical movements and specific PME.

The a priori use of an explicit prosodic model together with automatic learning enables an iterative and progressive construction of the prosodic modules of current speech synthesizers.

REFERENCES

1. Thorsen, N. A study of the perception of sentence intonation - evidence from Danish. *Journal of the Acoustical Society of America*, 67(3): 1014-1030, 1980.
2. Gårding, E. Intonation parameters in production and perception. *International Congress of Phonetic Sciences*, Aix-en-Provence, France, 1: 300-304, 1991.
3. Fujisaki, H. & Sudo, H. A generative model for the prosody of connected speech in Japanese. *Annual Report of engineering Research Institute*, 30: 75-80, 1971.
4. Aubergé, V. Developing a structured lexicon for synthesis of prosody. In Bailly, G., Benoît, C., editors, *Talking Machines: Theories, Models and Design*, 307-321, Elsevier B.V., 1992.
5. Morlec Y., Aubergé V., and Bailly G. Evaluation of automatic generation of prosody with a superposition model. *International Congress of Phonetic Sciences*, 4: 224-227, Stockholm, Sweden, 1995.
6. Morlec, Y., Bailly, G. & Aubergé, V. Synthesis and evaluation of intonation with a superposition model. *EUROSPEECH'95*, 3: 2043-2046, Madrid, Spain, 1995.
7. Fónagy, I., Bérard, E. and Fónagy, J. Clichés mélodiques. *Folia Linguistica,* 17: 153-185, 1984.
8. Calbris, G. & Montredon, J. *Oh là là. Expression intonative et mimique. Livre du professeur*. CLE international, 1981.
9. Morlec, Y., Bailly, G. & Aubergé, V. Un modèle connexioniste modulaire pour l'apprentissage des « gestes » intonatifs. *XXXIes Journés d'Etude sur la Parole,* to appear.
10. Strömqvist, S., Peters, A., and Ragnarsd'ottir, H. Particles and prepositions in Scandinavian child language development :effects of prosodic spotlight *? International Congress of Phonetic Sciences*, 4: 38-45, Stockholm, Sweden, 1995.