EVALUATION OF AUTOMATIC GENERATION OF PROSODY WITH A SUPERPOSITION MODEL

Y. Morlec, V. Aubergé and G. Bailly
Institut de la Communication Parlée, INPG & Université Stendhal
46, av. Félix Viallet 38031 Grenoble Cedex 01, France
e-mail: (morlec, auberge, bailly)@icp.grenet.fr

ABSTRACT

A new paradigm for modelling prosody is introduced. We assume that global melodic prototypes are built and stored in a "prosodic lexicon". The actual generation of adequate prosodic contours is achieved by retrieving and combining these elementary global contours accessed by linguistic keys. Two automatic F0 generation procedures have been used: The first consists of a structured lexicon, the second uses a recurrent neural network. Preliminary results show that both methods provide F0 contours which can compete with natural ones.

THEORETICAL FRAMEWORK

The intonation of an utterance is classically described in terms of tone units regarded as the primary units of intonational structure [9] [10]. So-called pitch targets are the phonetic realisations of a limited set of phonologically distinct tone segments, typically less than ten. The dynamics of tones is often constrained by an utterance template consisting of upper-lines and base-lines.

Within this framework the structural coherence of pitch movements is ensured by higher phonological components. Our approach aims to associate these higher phonological units more directly with their prosodic instanciations via a superposition model. For each phonological level, global prosodic movements achieve the necessary contrasts: phonologically-relevant information is thus distributed and enables priming [7]. The prototypic prosodic movements signal level-specific contrasts such as modality of the utterance within the discourse or strength of linguistic boundaries between groups of words within the utterance.

The actual prosodic contour results from a superposition of prototypes where upper-level ones are minimally anchored

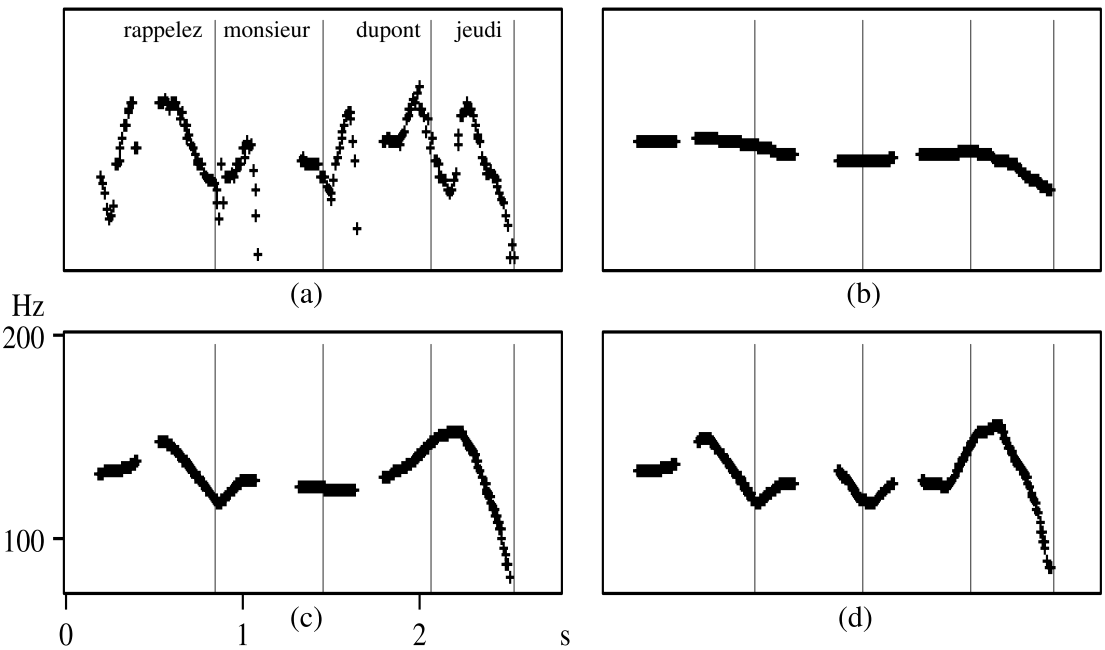

Figure 1. Comparison of original and synthetic contours obtained by the lexicon approach for the sentence: "Rappeler Mr Dupont Jeudi ! ". (a) the original curve, then successive superposition of (b) the sentence, (c) the syntagm and (d) the prosodic group. Note the lack of micromelody which could be easily produced with an additional level.

onto lower-level units.

In the next section, we test two superposition models: a simple additive model using a structured lexicon and a non-linear superposition model using a sequential neural network.

LEARNING PROCEDURE

Two automatic learning procedures have been applied to a corpus of 88 imperative utterances produced by one male speaker at a comfortable speech rate. The melodic curves were stylised with three values per vowel [5].

Three phonological levels directly related to the linguistic structure are considered here: (a) the sentence, (b) the syntagm, (c) the prosodic group (grouping each content word together with its function words).

The contours related to (b) and (c) are supposed to mark the degree of cohesion of the adjacent units as proposed in [2].

A structured lexicon

A methodology for developing a structured lexicon of prosodic contours has been already described in [1]. The data corpus is processed in a top-down hierarchy: upper-level prototypic contours are iteratively extracted by averaging and are then subtracted.

We developed a simplified version of this model which uses a normalised time-axis. A lexicon of prototypic melodic curves was built using a fixed 4th order polynomial interpolation where contours are scaled linearly to fit the considered linguistic boundaries. The figure 1.b shows the prototype for the sentence level. This prototype was then subtracted from the original contours and syntagmatic sub-contours are then grouped according to their relational marker and further processed.

The superposition model of generation then consists of a simple additive model (see Figure 1) which warps the prototypic melodic curves onto the actual syllables. Perceptual experiments described below show that this simple method leads to acceptable F0 generation. However, it is obviously too simple in its present form to adequately describe some important factors affecting melodic contours:

- It doesn't take account of the syllabic "weight" of the cued linguistic units.
- As f0 is the audible consequence of articulatory movements, undershoot can occur and thus speech rate, as well as stiffness of gestures, influence the actual realisations of intended targets.

Both restrictions mentioned above in addition to those imposed on the global shape of each melodic curve by polynomial interpolation could be solved by storing parameters of a dynamic equation. Sequential Neural Networks (SNNs) are known to model non-linear dynamics [8].

A neural network

In parallel with the structured lexicon, we performed simulations with SNNs. Although greatly inspired by the pioneering work of Traber [11], our approach differs in the characterisation of the input task: Traber uses a large window (13 symbols including syllables and phrase/word boundaries) on a linear

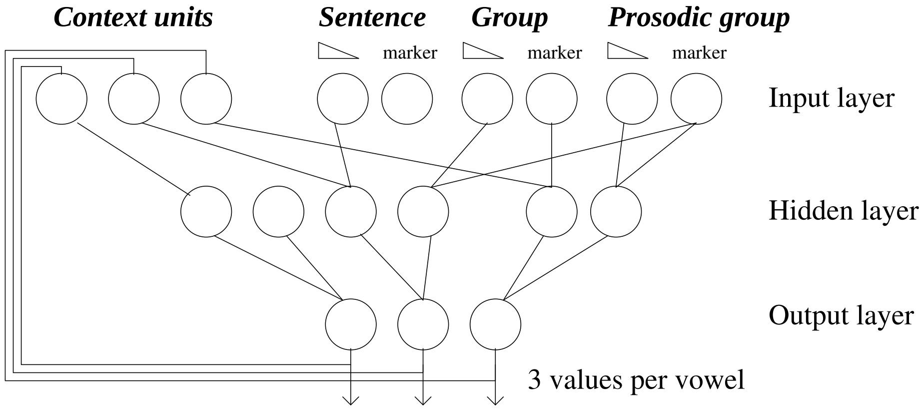

Figure 2. Structure of our SNN. Ramps signal the extent of associated units connected by limited sets of markers.

phonological representation of the input sentence where major and minor accentual positions are already given.

Our network (see Figure 2) is responsible for transforming simple linear movements (ramps) into more complex contours according to level-specific strategies, and for superposing them in order to mimic the original melodic curve.

Instead of a large window on a phonological description as used by Traber, our network input consists of only two parameters per level (sentence, syntagm, and prosodic group) :

Prosodic markers :

They indicate how subsequent units of the same level are linked. Linguistic function of prosody is thus restricted here to signal relations entertained by multi-layered linguistic units.

Syllabic ramps :

They signal the "syllabic distance" from the current syllable to the next marker.

- The output consists of 3 values per vowel.

The network was trained with half of the corpus. The other utterances were kept for prediction tests. Encouraging results were obtained with this basic recurrent network. Experiments described below showed that it was even able to learn the systematic initial emphatic stress used by our speaker for this specific task.

PERCEPTUAL EVALUATION

Method

A preference test was designed to evaluate the perceptual relevance of these two methods : 10 triplets of sentences (giving 60 presented pairs) were generated using a high-quality TD-PSOLA analysis-resynthesis technique [3].

The A version is the natural utterance only degraded by our F0 description (3 points per vowel). The B and C configurations are obtained by the structured lexicon and the SNN respectively.

Seven subjects participated in this perception experiment. During the preference test, the subjects were asked to choose the most natural utterance from each presented pair.

Results

Considering all subjects, results show:

- When the A version is presented against B or C, it is identified only 70% of the time. This demonstrates that the models have captured essential features of the original prosodic contours.

- The C version is only occasionally preferred to B demonstrating that the statistical distribution of linguistic structures within the corpus is adequate: the iterative analysis of the corpus produces similar results to the global learning strategy offered by SNNs.

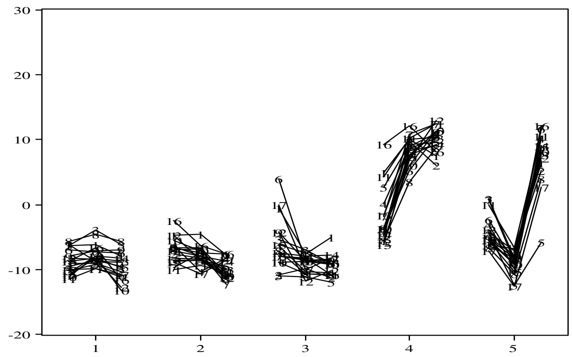

Figure 3. Superposition of syllabic f0 contours for 17 incredulous questions.

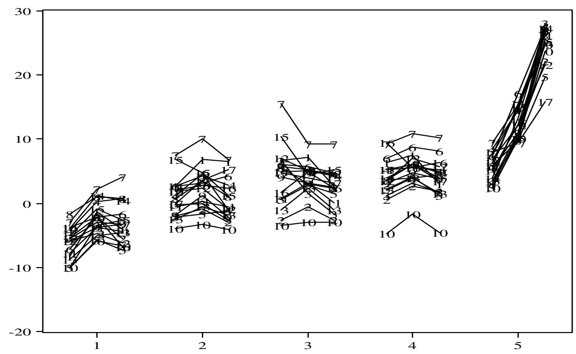

Figure 4. Superposition of syllabic f0 contours for the same set of sentences as Fig.3 but uttered as full questions.

FURTHER WORK

A major challenge to our approach is to demonstrate that prosodic information is distributed along the phonetic string via global contours which could be extracted and compared to melodic prototypes. We try here, to extend the proposal made by Fónagy [6] using the term "clichés mélodiques" for large pieces of prototypic melodic shapes associated to given communication needs.

We are thus looking for prosodic events associated with sentence level that can not be explained by the linguistic sub-unit configurations, e.g. an utterance template more elaborate than the two basic declination lines currently used. We thus recorded a second corpus designed especially for revealing the existence of global prototypic contours associated with sentence level: this corpus consists of short utterances (from 4 to 6 syllables) in order to limit the influence of carried sub-units. The unmarked syntax allowed several modalities such as assertion, question, exclamation... with various affects such as incredulity, doubt, surprise...

Early results with 5 syllables sentences seem to confirm our predictions. Figures 3 and 4 respectively show the superposition of 17 incredulous vs full questions ("Question incrédule" mentioned by Fónagy [6]) with various syntactic and phonotactic structures[1]. Note in Figure 3 the emergence of a global melodic prototype at the sentence level with an "accent" on the penultimate syllable.

CONCLUSION , PERSPECTIVES

Preliminary results show that both generation methods presented in this article are well appropriated for automatic generation of F0 contours using our theoretical framework.

The structured lexicon enables the main differences between prosodic movements of the same phonological level to be captured and allows one to observe the way in which prosodic contours of different phonological levels are combined.

The SNN approach seems the most promising, since it achieves a high quality of synthetic F0 curves with fewer assumptions on the shapes of elementary patterns. Moreover, this approach is a very versatile one: such a strategy has been efficiently applied to rhythmic control [4] and will enable coherent multi-parametric prosodic generation. The next step will be the training of a new sequential neural network with both rhythmic and melodic information patterns.

The automatic learning capacities of SNNs have to be guided by a strong hypothesis for the way linguistic units and affect are encoded onto the prosodic signals. Our assumptions is that this encoding is done via global patterns which could be quasi directly perceived and identified. This "direct perception" of intonation has immediate applications in the field of speech recognition.

[1] The number of words goes from 1 to 5. The relative size of each word was also varied within each syntactic structure.

REFERENCES

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

[2] Bailly, G. (1989), Integration of rhythmic and syntactic constraints in a model of generation of French prosody. *Speech Communication*, vol. 8, pp. 137-146.

[3] Bailly, G. (1992), Barbe, T. and Wang, H. Automatic labelling of large prosodic databases: tools, methodology and links with text-to-speech system. In Bailly, G.Benoît, C., editors, *Talking Machines: Theories, Models and Design*, pp. 323-333. Elsevier B.V.

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

[5] Emerard, F. and Benoît, C. (1987), De la production à l'extraction, l'état d'un chantier, *16èmes journées d'Etudes sur la Parole*, pp. 224-228.

[6] Fónagy, I., Bérard, E. and Fónagy, J. (1984), Clichés mélodiques. *Folia Linguistica,* vol. 17, pp. 153-185.

[7] Grosjean, F. (1983), How long is the sentence ? Prediction and prosody in the on-line processing of language, *Linguistica*, vol. 21, pp. 501-529.

[8] Jordan, M.I. (1989), Serial order: A parallel, distributed processing approach. In Elman, J.L. and Rumelhart, D.E., editors, *Advances in connectionist Theory: Speech.* Lawrence Erlbaum, Hillsdale, NJ.

[9] Ladd, D.R. (1983), Phonological features of intonation peaks. *Language*, vol. 59(4), pp. 721-759.

[10] Silverman, K.E.A. & al (1992), TOBI: A standard for labelling English prosody, *Proc. International Conference on Spoken Language Processing*, vol. 2, pp. 867-870.

[11] Traber, C. (1992), F0 generation with a database of natural F0 patterns and with neural network. In Bailly, G.Benoît, C., editors, *Talking Machines: Theories, Models and Design*, pp. 287-304. Elsevier B.V.