A STOCHASTIC MODEL OF INTONATION FOR FRENCH TEXT-TO-SPEECH SYNTHESIS

Jean Véronis, Philippe Di Cristo, Fabienne Courtois, Benoît Lagrue
Laboratoire Parole et Langage,
Université de Provence & CNRS
29 Av. Robert Schuman, 13621 Aix-en-Provence Cedex 1, France
Tel. +33 4 42 95 36 33, FAX +33 4 42 59 50 96, E-mail: Jean.Veronis@lpl.univ-aix.fr

ABSTRACT

This paper presents a stochastic model of French intonation contours for use in text-to-speech synthesis. The model has two modules, a linguistic module that generates abstract prosodic labels from text, and a phonetic module that generates an F_0 curve from the abstract prosodic labels. This model differs from previous work in the abstract prosodic labels used, which can be automatically derived from the training corpus. This feature makes it possible to use large corpora or several corpora of different speech styles, in addition to making it easy to adapt to new languages. The present paper focuses on the linguistic module, which does not require full syntactic analysis of the text but simply relies on a part-of-speech tagging technique. The results were validated by means of a perception test which showed that listeners did not perceive a significant difference in quality between the sentences synthesized with the original F_0 curve (from a recording), and those synthesized with the model-generated curve. The proposed model thus appears to capture a large part of the grammatical information needed to generate F_0.

1. INTRODUCTION

Generating acceptable prosody is currently one of the most challenging tasks in the development of text-to-speech synthesis systems [4]. Many studies have shown that although semantic and pragmatic factors enter into play, syntax is a major determining factor of the prosody of utterances. However, while the existence of a relationship between prosody and syntax is undeniable, no one has ever been able to reduce that relationship to simple rules (hence the term "congruence" of prosody and syntax). Moreover, in today's state of the art, no one is capable of producing automatic syntactic parsing of arbitrary texts with an acceptable success rate.

This paper presents a probabilistic model that generates French intonation contours based on part-of-speech tagging, a step we are currently able to achieve with an acceptable degree of success. Probabilistic models have already been applied to F_0 synthesis [8], but unlike past approaches, the method presented here uses a system of prosodic labels that can be automatically derived from the training corpus. This saves time, so large corpora or multiple corpora of different speech styles can be used, and the system can be quickly adapted to new languages. The model is composed of two modules:

- a **linguistic module** that predicts a set of abstract prosodic labels from the text
- a **phonetic module** that predicts an F_0 curve from the prosodic labels generated by the linguistic module.

This paper focuses on the linguistic module. We shall show that the method proposed for this module produces excellent results, as validated by a perception test. These positive results suggest that the linguistic module captures a large part of the grammatical information needed to generate F_0.

2. LINGUISTIC MODULE

Most probabilistic models can be stated in information-theory terms. More specifically, it is assumed that input message I is supplied to a noisy channel that converts it into deformed output message O. Automatically finding input message I when only the output is known amounts to examining all possible input messages and selecting the message $\hat{I}$ that maximizes the probability of getting output message O. Applied to prosodic tagging, this model assumes that the noisy channel outputs a sequence of grammatical classes, C, corresponding to an unknown input sequence of prosodic labels, P. Retrieving P thus means finding the most probable sequence $\hat{P}$ capable of producing output sequence C, i.e.:

$$\hat{P} = \arg\max_P Pr(P|C) = \arg\max_P Pr(P)\,Pr(C|P)$$

The prosodic labels were taken from the INTSINT system [7]. They constitute a simple formal encoding of intonation contours that can be automatically derived from the F_0 curve, unlike systems such as ToBI [2] which encode events with a linguistic value. This feature makes it easier to build a training corpus than in ToBI-based probabilistic approaches to F_0 generation, where hand marking of the training corpus by one or more experts is necessary [8]. In addition, the INTSINT system is language-independent, so our model can be easily adapted to other languages.

The labels fall into two categories:

- *absolute labels*, which are defined relative to the speaker's voice range:

  ⇒ T (top),
  ⇒ B (bottom),
  ⇒ M (mid: initial label with a mean value);

- *relative labels*, which are defined relative to the context:

  ⇒ U (upstepped),
  ⇒ D (downstepped),
  ⇒ H (high: local maximum),
  ⇒ L (low: local minimum),
  ⇒ S (same as preceding).

As usual in probabilistic methods, it is impossible to estimate the parameters of the model directly from the above equation, so two approximations will be used. We will assume first of all that it is possible to synchronously associate a prosodic label to each word without losing the ability to reconstruct high-quality F_0 contours. Sequences *P* and *C* thus have the same number of elements, *n*. We will also assume that the current prosodic label depends solely on the two labels that precede, and that the grammatical class depends solely on the current prosodic label and the preceding grammatical class:

$$Pr(P) = \prod_{i=1}^{n} Pr(P_i | P_{i-1} P_{i-2})$$

$$Pr(C|P) = \prod_{i=1}^{n} Pr(C_i | C_{i-1} P_i)$$

3. PHONETIC MODULE

As stated in the introduction, this paper deals mainly with the linguistic module. Although it produces acceptable output, the phonetic module used here is very simple. It is now under improvement, but even in its current state, it is sufficient for testing the linguistic module.

The F_0 curve corresponding to the sequence of labels P_i is a quadratic spline curve that goes through the set of target points F_i, in one-to-one-correspondence with the labels (P_i). Each target point is placed at 2/3 of the duration of the corresponding word (which is usually the final stressed syllable in French). Frequencies F_T, F_M, and F_B, which correspond to the absolute prosodic labels T, M, and B, respectively, are assumed to be fixed. Let F_i be the frequency of the current target point. The frequency of the next target point is calculated by the following linear law:

$$F_{i+1} = F_i + \alpha(F_T - F_i) \qquad \text{if } P_{i+1} = \text{U, H}$$
$$F_{i+1} = F_i + \beta(F_B - F_i) \qquad \text{if } P_{i+1} = \text{D, L}$$

An ascending sequence of target points thus converges towards F_T, and a descending sequence converges towards F_B, which is consistent with the tendency observed by Hirst *et al.* [6]. The spline curve that yields the final F_0 is made up of parabola arcs whose extrema are the target points (F_i). The arcs are connected by a common tangent to the median point between two consecutive target points.

4. PARAMETER ESTIMATION

The model parameters were estimated on 500 sentences taken from the EUROM 1 corpus [3], totaling 35 minutes of speech or 7022 words. The 200 sentences in this corpus are grouped together into 40 five-sentence passages pronounced by different speakers. Example of passage:

> *La semaine dernière, mon amie est allée chez le médecin se faire faire des piqûres. Elle doit partir en vacances en Extrême-Orient, et elle est obligée de se faire vacciner contre le choléra, la typhoïde, l'hépatite A, la polio et le tétanos. Je pense qu'après ça elle aura vraiment besoin d'un médecin ! D'autant plus qu'elle veut se faire faire tout ça en une seule fois. Moi, je la plains pas. Tant pis pour elle !*

The corpus was prepared using the tools developed in the MULTEXT project [9], as follows:

- Grammatical tagging of sentences was achieved using an HMM tagger which performs at a correct tagging rate of approximately 95 %. The categories used are the basic parts of speech (noun, verb, adjective, etc.) along with a few punctuation categories. The tagging was verified and corrected manually.

- The recordings of the 500 sentences in the training corpus were manually segmented into words using a fast, computer-assisted method. Segmentation accuracy is not critical, so this phase will be performed quasi-automatically in the future, using tools now being developed.

- The recorded F_0 was stylized with quadratic spline functions [7]. The stylized F_0 was checked and corrected by an expert using the PSOLA re-synthesis system, accuracy of the method is around 95 % [1]. Various trials showed that the errors corrected were in fact minor and that the manual validation phase could have been skipped without any drastic effects on parameter estimation (figure 1).

- The INTSINT prosodic labels associated to each word and corresponding to the movements in the stylized curve were automatically derived from the stylized F_0 (figure 1).

The linguistic module probabilities were estimated from the relative frequencies of the various prosodic-label and grammatical-class sequences in the corpus, corrected by a distribution smoothing in order to take unobserved cases into account. Parameters α, β, F_T, F_M, and F_B of the phonetic module were estimated by taking the mean of the observed values for a given speaker (T_L test set; see below).

5. EVALUATION

An original set, T_L, consisting of 20 sentences of comparable length and style to the ones in the training corpus, was recorded by a male speaker. Examples:

> *Qu'est-ce que tu penses de la réception donnée à l'occasion des Victoires de la musique? — Le problème majeur de la jeunesse, c'est de ne pas trouver de travail. — La fête nationale qui a lieu en France le quatorze juillet est clôturée par un feu d'artifice. — Pourquoi les hommes ne comprennent-ils pas que les oiseaux ne doivent pas être mis en cage?*

T_L was segmented into words, and then grammatical tagging, F_0 stylization, and INTSINT labeling were added.

The test material included four test sets, T_O, T_P, T_M, and T_A, each composed of the same 20 sentences as in T_L, but synthesized by the MBROLA synthesizer [5]. To avoid potential duration-related biases, the sentences were synthesized with a constant phoneme duration of 90 ms. The test sets were assigned the following F_0 curves:

- T_O: stylized F_0 of original set T_L, mapped onto the new durations.
- T_P: F_0 generated by the phonetic module from INTSINT labeling of T_L.
- T_M: F_0 generated by the full model from the grammatical classes of the words.
- T_A: F_0 generated from random target points between F_B and F_T.

The sentences in the four test sets were presented in random order to 13 judges (graduate students in phonetics) in two 40-sentence sessions separated by a free-length break. Subjects were instructed to rate each sentence for intonation quality on a scale ranging from 0 to 10. The task proper was preceded by a five-sentence practice session.

The test sets obtained the following mean ratings:

Set	T_O	T_P	T_M	T_A
Rating	6.014	5.446	5.696	4.319

An analysis of variance revealed a significant difference between the four sets: $F(3,1036) = 27.030$, $p < 10^{-16}$. Set T_A was rated significantly below the other sets: comparison with T_P (the closest set to T_A) yielded $F(1,518) = 31.135$, $p < 10^{-7}$. In contrast, sets T_O, T_M, and T_P obtained similar ratings (in that order). The T_O - T_M and T_M - T_P differences were not very significant [$F(1,518) = 2.564$, $p = 0.11$ and $F(1,518) = 1.494$, p = 0.22, respectively]. However, note that the difference between T_O and T_P was significant $F(1,518) = 8.328$, $p = 0.004$.

In summary, the T_O, T_M, and T_P ratings were quite similar and were far better than that of the random set, T_A. Set T_M, synthesized directly from the text using the full model, obtained a rating very close to T_O, which carried the original F_0 curve (nonsignificant difference).

The proposed model thus seems to capture a large part of the grammatical information needed to generate F_0. The relatively low rating obtained by T_P synthesized from the INTSINT labeling of T_O can be explained by the overly simplified state of the phonetic module used in this experiment.

6. CONCLUSION

Despite its simplicity, the model proposed in this study generates realistic F_0 contours. Clearly, the linguistic module captures a large part of the grammatical information that would ideally be needed to generate F_0, without requiring a thorough or complex syntactic analysis of the text. This module can nevertheless be enhanced in various ways since it now only supports a small set of grammatical classes and a limited grammatical context (bigrams). The phonetic module, which was not the focus of this study, is already being improved. In particular, estimating parameters by the mean is probably not an ideal solution, and the linearity hypothesis is perhaps too strong. The virtually automatic feature of this method will facilitate the implementation of a much larger training corpus that will support trigrams and a broader set of classes, and will serve as a basis for making finer-grained parameter estimates in the phonetic module.

7. ACKNOWLEDGEMENTS

The authors would like to thank Thierry Dutoit for the MBROLA synthesizer, Daniel Hirst for our many discussions and his careful reviewing, Corine Astésano for her help with the corpus, Christian Cavé for his advice on the perception test, Emmanuel Flachaire for the statistical processing, and Robert Espesser for his technical assistance. Robert Espesser wrote the signal editing software and the re-synthesis program used in this experiment. The perception test was conducted using the ASTEC software developed at the laboratory as part of the European project OSCAR.

8. REFERENCES

[1] Astésano, C., Espesser, R., Hirst, D., Llisterri, J. (1997). Stylisation automatique de la fréquence fondamentale : une évaluation multilingue. *4ème Congrès Français d'Acoustique*, Marseille, 441-444.

[2] Beckman, M, Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, 3, 255-309.

[3] Chan, D., Fourcin, A., Gibbon, D., Granström, B., Hucvale, M., Kokkinakis, G., Kvale, K., Lamel, L., Lindberg, B., Moreno, A., Mouropoulos, J., Senia, F., Trancoso, I., Veld, C., Zeiliger, J. (1995). EUROM- A spoken language resource for the EU. *Proceedings of the 4th European Conference on Speech Communication and Speech Technology, Eurospeech'95*. Madrid. vol. 1, 867-870.

[4] Collier, R. (1991). Multi-language intonation synthesis. *Journal of Phonetics*, 19, 61-73.

[5] Dutoit, T., Leich, H. (1993). MBR-PSOLA : Text-to-speech synthesis based on an MBE re-synthesis of the segments database, *Speech Communication*, December 1993.

[6] Hirst, D., Nicolas, P., Espesser, R. (1991). Coding the F0 of a continuous text in French : an Experimental Approach. *Proceedings of 12th International Congress of Phonetic Sciences*, Aix-en-Provence, 5, 234-237.

[7] Hirst, D., Ide, N., Véronis, J. (1994). Coding fundamental frequency patterns for multi-lingual synthesis with INTSINT in the MULTEXT project. *Proceedings of the 2nd ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, NY, September 1994, 77-81.

[8] Ross, K. N. (1995). *Modeling intonation for speech synthesis.* PhD thesis. Boston University.

[9] Véronis, J., Hirst, D., Espesser, R., Ide, N. (1994). NL and speech in the MULTEXT project. *AAAI'94 Workshop on Integration of Natural Language and Speech*, Seattle, 72-78.

9. SOUND EXAMPLES

The accompanying CD-ROM contains two examples, resynthetised from original recordings with the TD-PSOLA technique, using the F_0 generated by our model :

- [sound A0429S01.WAV] : sample from the training corpus ;
- [sound A0429S02.WAV] : new passage, not from the training corpus.

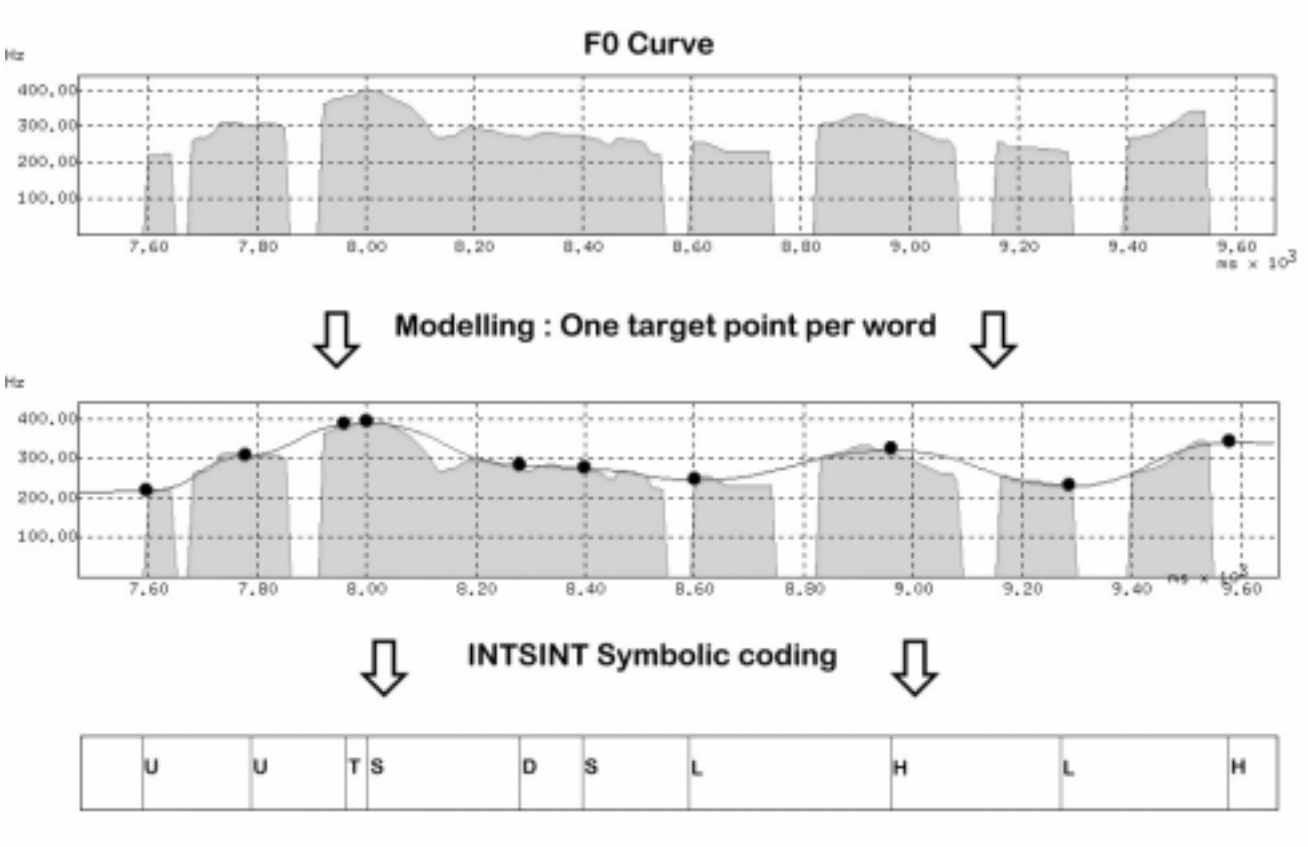

Figure 1. F0 stylisation and symbolic coding