# Modelling intonational variation with GIMEL

**Stefan Werner**[*], **Eric Keller**[†] and **Brigitte Zellner Keller**[†]

[*]Joensuun yliopisto
Language Technology, P.O. Box 111, 80101 Joensuu, Finland
stefan.werner@joensuu.fi

[†]Université de Lausanne
LAIP, Faculté des Lettres, CH-1015 Lausanne, Switzerland
{eric.keller,brigitte.zellnerkeller}@imm.unil.ch

## ABSTRACT

GIMEL is a new system for the modelling of fundamental frequency in speech. In its final form, it will automatically assign intonation labels to text, providing all information necessary for the prediction of an appropriate $F_0$ curve. Presently, GIMEL contains modules to employ an adaptable information reduction algorithm that identifies macro-prosodic turning points in an $F_0$ contour after approximating it with tension splines.

The first speech corpus whose intonation contours were modelled with GIMEL consists of elicited speech from French speakers. For every speaker, the spline approximation was adjusted such that microprosodic $F_0$ variation was eliminated, and that the average distance between original and modelled curve was identical for all speakers. Informal listening evaluation with MBROLA-synthesised voices indicates that spline-approximated $F_0$ is indistinguishable from the original contour. Now the turning points in the corpus can be labelled according to their statistical properties and be used for synthesis.

## 1 INTRODUCTION

Modelling of fundamental frequency in speech is an endeavour that many researchers have undertaken, employing a huge variety of methodologies. Phonological models usually are based on a linguistically motivated a priori definition of what constitutes an intonationally relevant $F_0$ event, e.g. target points (as in Pierrehumbert/ToBI-style models, see [1, 2]) or contour shapes (as in the British tradition, see [3]). Other models, often conceived by experimental phoneticians or engineers, intend to keep closer to the empirical reality of observed $F_0$ curves. But they, too, may have to go back on this goal in order to satisfy competing objectives like reflecting articulatory processes ([4]) or building a more abstract intonation grammar ([5]). Consequently, a large proportion of potentially important $F_0$ data is routinely left unused.

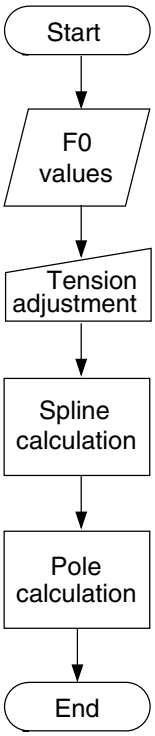


**Figure 1:** Preprocessing intonational information.

The "General Intonational Modelling Environment for Labelling" (GIMEL) that we propose is designed to extract and analyse as much information as possible from observed $F_0$ curves before any generalisations are made. In particular, timing, intensity and $F_0$ data related to the automatically located macro-prosodic $F_0$ turning is collected and statistically evaluated. The results are then used to define descriptive labels for the turning points and rules for their distribution. Finally, $F_0$ for new synthetic utterances can be controlled on the basis of these labels and rules.


ISBN 1-876346-48-5 © 2003 UAB

## 2 SYSTEM OVERVIEW

In GIMEL, intonational information is processed in several steps. In order to find the relevant $F_0$ turning points in an utterance, the extracted and interpolated $F_0$ values are approximated with a quadratic tension spline. As shown in Figure 1, the resulting set of turning points ("poles") depends on the choice of tension parameter: the greater the spline's tension, the closer the approximated curve is to the original, and the higher is the number of poles in the spline curve. Figure 3 shows a spline-approximated curve together with straight lines connecting the turning points of the curve. The effect of adjusting the tension parameter is shown in Figures 4–6 where the degree of smoothing and flattening clearly increases with the spline tension.

To help the researcher find the appropriate tension value for a given task, GIMEL produces also MBROLA `.pho` files for synthetic generation of the utterance with the new approximated $F_0$ values. This in addition to the option for PSOLA resynthesis of the original with replaced $F_0$ values ensures that adequate perception tests assessing the validity of a spline stylisation can be carried out.

In the next step of the process, when the suitable degree of tension has been determined, another GIMEL module generates a large variety of acoustic and linguistic measures and categories related to the turning points' timing and frequencies. These include positions and distances within segments, syllables and other possibly relevant domains, speech rate and intensity, as well as information about lexical classes and utterance type. A concise summary of GIMEL's input and output is depicted in Figure 2.

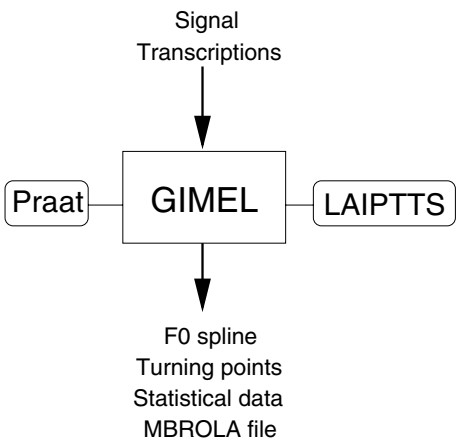


**Figure 2:** Inputs to and outputs from the present version of GIMEL. For every signal file with its accompanying transcriptions, GIMEL and the programmes it calls produce the types of data shown at the bottom.

In its present implementation, GIMEL consists of a package of JAVA classes and several Praat[1] scripts. The system runs Praat's $F_0$ extraction and intensity calculation on the input sound files and also utilizes Praat TextGrids containing the phonemic transcriptions. An additional input to the statistical analysis is the enriched phonemic annotation of an utterance's text produced by the text-to-speech system LAIPTTS-SpeechMill[2]. It provides information about minor and major phrase boundaries, syllable boundaries, lexical class (function vs. content word) and phonological processes like liaison and schwa deletion.

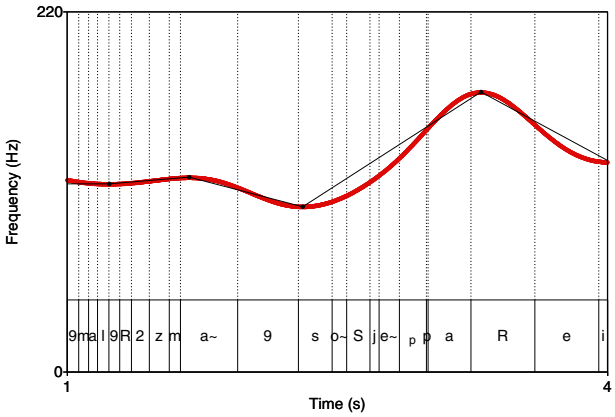


**Figure 3:** Spline curve and poles (tension $10^{-8}$, see Figure 6. Beginning of sentence *Euh malheureusement euh son chien part et il est tout mouillé.*

## 3 SPEECH MATERIAL

GIMEL's first application was to a speech corpus with 6.5 minutes of elicited speech from five French speakers (1 female, 4 males). The speakers describe in their own words the contents of a pictorial story, producing a total of 56 sentences. The material was digitally recorded in 16 bit, 16 KHz RIFF files. Canonical orthographic transcriptions were converted to enriched phonemic transcriptions with LAIPTTS-SpeechMill. A trained phonetician segmented and labelled the material with Praat. Afterwards, the label files were checked by another phonetician. Because the corpus utterances are free speech, not lab speech sentences read from paper, a considerable amount of manual correction was needed for the LAIPTTS transcription files (see [6] for a general account of grapheme-tophoneme conversion problems in French).

Figures 4–6 show $F_0$ spline poles for part of an ut-

[1] `http://www.praat.org/`
[2] `http://www.unil.ch/imm/docs/LAIP/LAIPTTS.html`

ISBN 1-876346-48-5 © 2003 UAB 

terance from the corpus. Different numbers of poles have been extracted by adjusting the tension parameter of GIMEL's spline routine. For each chosen tension value, the automatically generated data files contain 3046 lines of observations for the currently extracted 40 parameters.

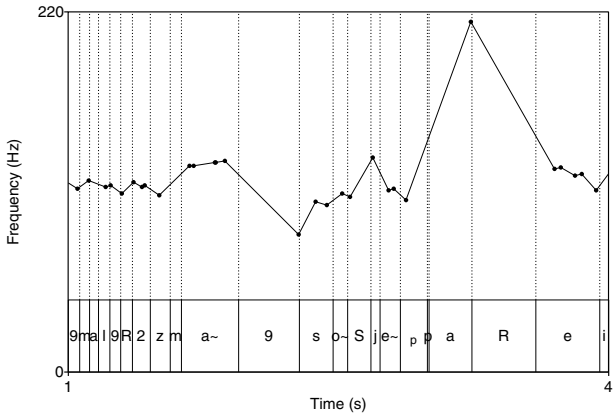


**Figure 4:** Turning points at a spline tension of $10^{-4}$ (same signal as in Figure 3).

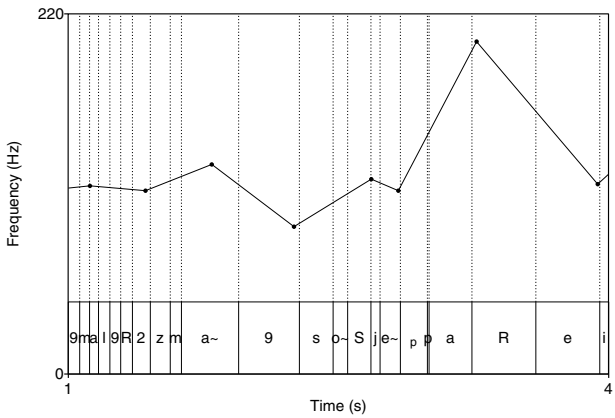


**Figure 5:** Turning points at a spline tension of $10^{-6}$.

## 4 ANALYSIS

Informal listening tests with two native speakers showed that for a tension parameter of $10^{-4}$ (Fig. 4) the auditive impression evoked by the resynthesised signals with spline-approximated fundamental frequency is identity with the original. For a tension of $10^{-8}$ (see the overlay of original and approximated $F_0$ curve in Fig. 7), the signal is still judged semantically equivalent.

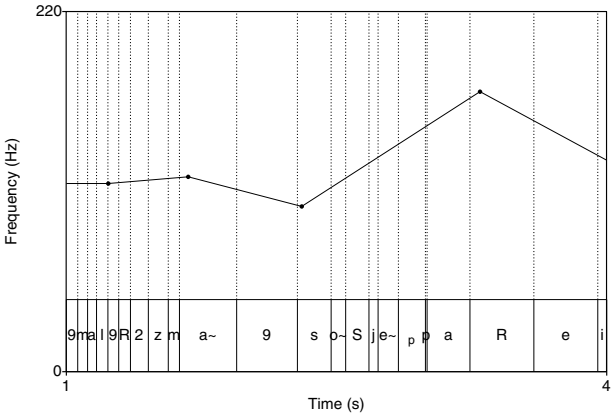


**Figure 6:** Turning points at a spline tension of $10^{-8}$.

Our statistical evaluation is not complete but we have found some evidence that relevant parameters for turning point placement include at least segment position in the syllable, number of syllables in the word, word position in the phrase and lexical class. We also found tentative support for recent findings with Swedish speech ([7]) showing that a considerable part of inter-speaker variability can be explained in terms of speakers' mean $F_0$ and segment durations.

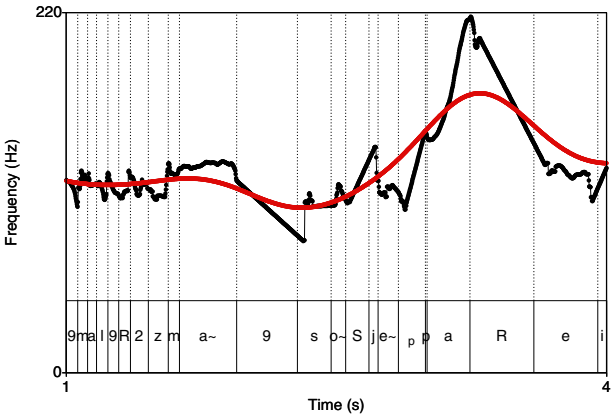


**Figure 7:** Comparison of original and spline-approximated $F_0$ curve.


ISBN 1-876346-48-5 © 2003 UAB

## 5 CONCLUSION

Studying intonation with GIMEL gives the researcher the possibility to try out different degrees of data reduction and immediately obtain a wealth of statistical data linking the $F_0$ approximation's turning points to acoustic events and linguistic categories. This combination of an exploratory modelling tool with comprehensive empirical analysis distinguishes GIMEL from other methods to approximate $F_0$ with splines, like MOMEL ([8]). Another difference is that GIMEL is not conceptually coupled to any specific representation scheme for intonational events (as MOMEL is to INTSINT[3]). Thus, the GIMEL system lends itself naturally to any kind of data-based intonation research, whether within or outside the framework of an existing paradigm.

## REFERENCES

[1] J. Pierrehumbert, *The Phonology and Phonetics of English Intonation*, Ph.D. thesis, MIT, 1980, Indiana University Linguistics Club 188.

[2] K. Silverman, M. E. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: a standard for labeling English prosody," in *Proceedings of the Second International Conference on Spoken Language Processing*, Banff, 1992, vol. 2, pp. 867–870.

[3] D. Crystal, *Prosodic Systems and Intonation in English*, Cambridge University Press, Cambridge, 1969.

[4] H. Fujisaki, "Modeling the process of fundamental frequency control of speech for synthesis of tonal features of various languages," in *Proceedings of the 1997 China-Japan Symposium on Advanced Information Technology*, 1997, pp. 1–12.

[5] J. T. 't Hart, R. Collier, and A. Cohen, *A perceptual study of intonation: an experimental-phonetic approach to speech melody*, Cambridge University Press, Cambridge, 1990.

[6] F. Yvon, P. Boula de Mareüil, C. d'Alessandro, V. Aubergé, M. Bagein, G. Bailly, F. Béchet, S. Foukia, J.-F. Goldman, E. Keller, D. O'Shaughnessy, V. Pagel, F. Sannier, J. Véronis, and B. Zellner, "Objective evaluation of grapheme to phoneme conversion for text-to-speech synthesis in French," *Computer Speech and Language*, pp. 393–410, 1998.

[7] G. Fant, A. Kruckenberg, K. Gustafson, and J. Liljencrants, "A new approach to intonation analysis and synthesis of swedish," in *Proceedings of the Speech Prosody 2002 conference, 11-13 April 2002. Aix-en-Provence*, B. Bel and I. Marlien, Eds., 2002.

[8] D.J. Hirst, A. Di Cristo, and R. Espesser, "Levels of representation and levels of analysis for intonation.," in *Prosody : Theory and Experiment*, M. Horne, Ed., pp. 51–87. Kluwer, Dordrecht, 2000.

[9] A. Dobnikar, "Improvement in modelling the f0 contour for different types of intonation units in slovene," in *Improvements in Speech Synthesis*, E. Keller, G. Bailly, A. Monaghan, J. Terken, and M. Huckvale, Eds., pp. 144–153. Wiley, Chichester, 2001.

[10] E. Keller, "La vérification d'hypothèses linguistiques au moyen de la synthèse de la parole.," in *Cahiers de l'institut de linguistique*, vol. 2. Université de Louvain, in press.

[3] But see [9] for an interesting new hybrid approach.