A MULTILINGUAL PROSODIC DATABASE

Estelle Campione, Jean Véronis

Laboratoire Parole et Langage
Université de Provence & CNRS
29, Av. Robert Schuman, 13621 Aix-en-Provence Cedex 1, France
Estelle.Campione@lpl.univ-aix.fr

ABSTRACT

We present a prosodic corpus in five languages (French, English, Italian, German and Spanish) comprising 4 hours and 20 minutes of speech and involving 50 different speakers (5 male and 5 female per language). The recordings on which the corpus is based are extracted from the EUROM 1 speech database and consists of passages of about five sentences. The corpus was stylized automatically by an algorithm which factors out microprosodic effects and represents the intonation contour of utterances by a series of target points. Once interpolated by a smooth curve (spline), these points produce a contour undistinguishable from the original when re-synthesized, apart from a few detection errors. A symbolic coding of the 50000 pitch movements of the corpus is also provided, along with the time-alignment of orthographic transcription to signal at word-level. The entire corpus was verified and manually corrected by experts for each language. It will be made available at production cost for research through the European Language Resource Association (ELRA).

1. INTRODUCTION

Large multilingual databases of prosodic information can be useful for theoretical research and for practical applications such as speech synthesis and speech recognition. However, little data is available at the moment (or none at all for some languages), due to the lack of generally available analysis tools and methods, and to the cost of manual verification and correction. In addition, the work done so far on different languages has in general relied on different theoretical frameworks and methodologies, which makes data and results difficult to compare.

We present in this paper a multilingual prosodic corpus that was developed in the context of the MULTEXT project [18]. In addition to the original recordings, the corpus consists of files for:

- the raw F_0;

- a stylization of intonation contours;

- a symbolic coding of pitch movements;

- the alignment of orthographic transcription to signal at word level.

The corpus is reasonably large since it comprises 4 hours 20 minutes of speech (ca. 50000 pitch movements) in five languages (French, English, Italian, German and Spanish). It has already been used in a number of studies on intonation (e.g. [17] [2] [16]).

2. RECORDINGS

The recordings of the multi-lingual corpus are drawn from the EUROM 1 speech database, developed within the Esprit SAM project) "Multi-lingual Speech Input/output Assessment, Methodology and Standardisation") [4]. We used only a subset of the database, consisting of passages read by ten speakers in each language (the "Few talker set").

For each language, the portion of the EUROM 1 corpus used contains 40 different passages of five sentences connected thematically (Figure 1). The translation among the various languages is rather free and often constitutes an adaptation to the local culture (for proper names, food, etc.). Every speaker was asked to read a subset of the passages and to try to have an intonation as natural as possible. The acoustic quality of the recordings is high (sampling speed at 20 kHz, 16 bits, recording in an anechoic room). The recorded material was controlled during acquisition so that bad quality recordings (noisy or misread sentences) were directly cancelled and repeated.

I have a problem with my water softener. The water level is too high, and the overflow keeps dripping. Could you arrange to send an engineer on Tuesday morning, please? It's the only day I can manage this week. I'd be grateful if you could confirm the arrangement in writing.

Figure 1. Example of recorded passage.

Table 1 gives the number of passages read by speaker and duration per language.

Language	Passages per speaker	Total duration (h:m:s)	Average duration per passage (s)
English	15	00:43:55	17.6
French	10	00:36:30	21.9
German	20	01:13:09	21.9
Italian	15	00:54:18	21.7
Spanish	15	00:52:21	20.9
Total	**-**	**04:20:13**	**20.8**

Table 1: Duration per language.

3. STYLIZATION

3.1. Principle

Stylization consists in extracting the macroprosodic component from the F_0, which reflects intonative intention of the speaker. The microprosodic component, which is entirely dependent on the choice of phonemes in the utterance (lowering of F_0 for voiced obstruents etc.) is factored out. Various stylization methods have been proposed since the sixties ([5] [14] [6] [7] [9] [15] etc.), and rely on more or less complex models.

The method used in this work has some appealing features compared to other methods:

- it is language-independent;

- it does not require any pre-segmentation of the signal (e.g. in syllables);

- it does not require any training on the data;

- it performs automatically with a very good success rate;

- the stylized curve is undistinguishable from the original.

It was originally proposed by Hirst ([10] [11]) and consists in reducing the intonation contour to a series of target points, which represent the relevant pitch movements (Figure 4). Once interpolated by a quadratic spline curve (unvoiced segments are interpolated so that the resulting curve presents no discontinuities), the series of target points produces an F_0 contour undistinguishable from the original, apart from a few detection errors that must be corrected by hand. Other authors have used smooth curves to model contours ((7] [15]), as opposed to straight lines ([5] [6] [14]). However, the underlying model of Hirst's method is particularly simple, and the representation by means of target points very economical.

3.2. Algorithm

An efficient algorithm for computing target points (MOMEL, standing for *MOdélisation de MELodie*) was proposed by Hirst and Espesser ([9] see also [8]).

The MOMEL algorithm consists in four stages (we summarize the description given by [9]).

1. Aberrant values due to F_0 detection errors are eliminated.

2. A quadratic regression technique is applied in a moving analysis window, which computes a candidate target for all F_0 values in that window (Figure 1). The regression is computed incrementally in an asymmetrical way, based on the hypothesis that the only effect of the microprosodic component of the F_0, is to locally lower the values of the underlying macroprosodic curve.

3. The last stage reduces the set of candidate targets by clustering targets together when their distance does not exceed a fixed threshold.
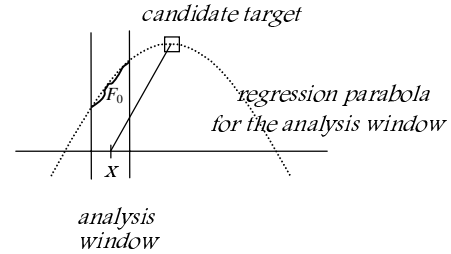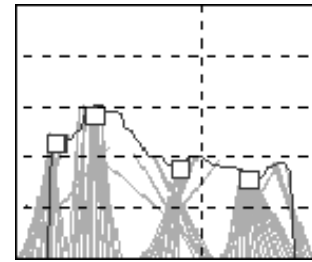

Figure 1. Calculation of a candidate target point.

Figure 2. Clustering of candidate targets.

3.3. Manual correction

The stylization algorithm described above presents minor defects leading to a few errors. The corpus was therefore manually verified and corrected by experts. All the experts have used the same correction strategy, which was minimalist and consisted in correcting the smallest possible number of target points. Only those leading to an audible difference between the original F_0 curve and the re-synthesis obtained by stylization were corrected, so that the re-synthesis was judged similar to the original. A quantitative assessment showed that the algorithm produces about 5% of errors. A large part of these errors (approximately 3%) were moreover systematically of two or three different types, in particular missing targets in transitions from voiced to voiceless segments of speech, which suggests that an improved algorithm could probably eliminate the majority of them [3].

4. SYMBOLIC CODING

Various systems for symbolic coding of intonation have been proposed. They can be categorized in two types: linguistic systems, such as ToBI [13], which encode events of a linguistic nature, and phonetic systems, such as HLCB [15] or INTSINT [8], which aim only at providing a purely configurational description of the macroprosodic curve without interpretation. Systems of the first category are probably more desirable, but (1) are not straightforward to adapt to new languages; (2) pose difficult problems for automatic labeling (see [1] [12]). The

second category, easier to implement, is of course less interesting in linguistic terms, but can still contribute to the development of labeled corpora useful for many applications. In addition, it can be seen as a first step towards automatic labeling of systems like ToBI, and can be a help in the development of such systems for languages other than American English.

In this work, melodic movements have been coded using a statistical model described in Véronis and Campione [16]. Seven categories are used:

- **L+**: large descending movement;
- **L**: medium descending movement;
- **L-**: small descending movement;
- **S**: very small or null movement;
- **H-** : small ascending movement;
- **H**: medium ascending movement;
- **H+**: large ascending movement.

The amplitude of movements for a given label, e.g. **L**, depends on the height of the initial point of the movement in the speaker's range (Figure 3). The model assumes a normal distribution of pitch target points in order to minimize the error. Such a distribution constitutes a reasonable approximation of the speakers' distributions despite large individual variations [2].
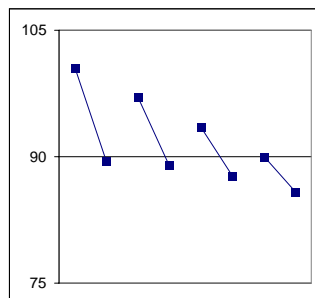

Figure 3. Amplitude predicted by the model for **L** movements starting at various frequencies (speaker with a mean frequency of 90 ST and a standard deviation of 3.5 ST).

The proposed model enables a re-generation of the original target points from the symbolic coding with a mean squared error of 0.74 ST, and 99% of re-generated points are placed closer than 2 ST from the original. It was verified by listening to randomly selected passages of the corpus that when points are generated at less than 2 ST from the original, there was no change of linguistic intention. The largest differences can be audible under careful listening conditions, but are not perceived in the normal speech flow and in any case have no linguistic impact. The 1% of points that are badly predicted by the model are points very far in the extremes of the speakers' range, especially in the infra grave. Labels that do not enable a close regeneration have been marked in the corpus.

5. CONCLUSION

We have presented a prosodic database in five languages, composed of 4 hours 20 minutes of speech and involving 50 different speakers. The corpus has been entirely stylized by means of target points interpolated by a spline curve, and manually verified and corrected by experts. A symbolic coding of the 50000 pitch movements of the corpus is also provided, along with the time-alignment of orthographic transcription to signal at word-level. We believe that this database could be a useful resource for intonation studies and practical applications.

6. ACKNOWLEDGEMENTS

7. REFERENCES

1. Black, A. and Hunt, A. "Generating F_o contours from ToBI labels using a linear regression." *ICSLP'96*, Philadelphia, 1996.

2. Campione, E. and Véronis, J. "A statistical study of pitch target points in five languages." *ICSLP'98,* Sidney, Australia (these proceedings), 1998.

3. Campione, E. and Véronis, J. "Une évaluation de l'algorithme de stylisation mélodique MOMEL." *Travaux de l'Institut de Phonétique d'Aix-en-Provence,* forthcoming.

4. Chan, D., Fourcin, A., Gibbon, D., Grandström, B., Huckvale, M, Kokkinakis, G., Kvale, K., Lamel, L., lindberg, B., Moreno, A., Mouropoulos, J., Senia, F., Transcoso, I., Velt, C. and Zeiliger, J. (1995). "EUROM – A Spoken Language Ressource for the EU." *In Proceedings of Eurospeech'95,* Madrid, 1995.

5. Cohen, A. and t'Hart, J. "Perceptual Analysis of Intonation Pattern." *5ème Congrès International d'Acoustique*, Liège, 1-4, 1965.

6. D'Alessandro, C. and Mertens, P. "Automatic Pitch Contour Stylisation Using a Model of Tonal Perception." *Computer, Speech and Language*, 9, 257-288, 1995.

7. Fujisaki, H. and Hirose, K. "Modelling the dynamic characteristics of voice fundamental frequency with application to analysis and synthesis of intonation." In *Proceedings of 13th International Congress of Linguists*, 57-70, 1982.

8. Hirst, D.J., Di Cristo, A. and Espesser, R. "Levels of representation and levels of analysis for the description of intonation systems." In Horne, M. (Ed.), *Prosody: Theory and Experiment,* Kluwer Academic Publishers, Dordrecht, forthcoming.

9. Hirst, D. and Espesser, R. "Automatic Modelling of Fundamental Frequency Using a Quadratic Spline Function." *Travaux de l'Institut de Phonétique d'Aix-en-Provence,* 15, 75-85, 1993.

10. Hirst, D. J. "Structures and categories in prosodic representations." In Cutler and Ladd (Eds), *Prosody: Models and Measurements*. Berlin, Springer. 93-109, 1983.

11. Hirst, D. J. "Un modéle de production de l'intonation." *Travaux de l'Institut de Phonétique d'Aix*, 7, 297-315, 1980.

12. Ostendorf, M. and Ross, K. "A multi-level model for recognition of intonation labels." In Sagisaka, Campbell and Higuchi (Eds), *Computing Prosody.* Springer, Berlin, 291-308, 1997.

13. Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. "ToBI: a standard for labelling English prosody." *Proceedings of ICSLP'92,* 2, Banff, Canada, 867-870, 1992.

14. t'Hart, J., Collier, R. and Cohen, A. *A Perceptual Study of Intonation, an experimental-phonetic approach to speech melody*, Cambridge University Press, Grande-Bretagne, 212 p, 1990.

15. Taylor, P. "The Rise/Fall/Connection Model of Intonation." *Speech Communication*, 15, 1&2, 169-186, 1994.

16. Véronis, J. and Campione, E. "Towards a reversible symbolic coding of intonation". *ICSLP'98,* Sydney, Australia (these proceedings), 1998.

17. Véronis, J., Di Cristo, P., Courtois, F. and Lagrue, B. "A stochastic model of intonation for French text-to-speech synthesis." *Proceedings of the 5th European Conference on Speech Communication and Technology, Eurospeech'97,* Rhodes (Greece), 2643-2646, 1997.

18. Véronis, J., Hirst, D. J. and Ide, N. "NL and speech in the MULTEXT project". *AAAI'94 Workshop on Integration of Natural Language and Speech,* Seattle, 72-78, 1994.
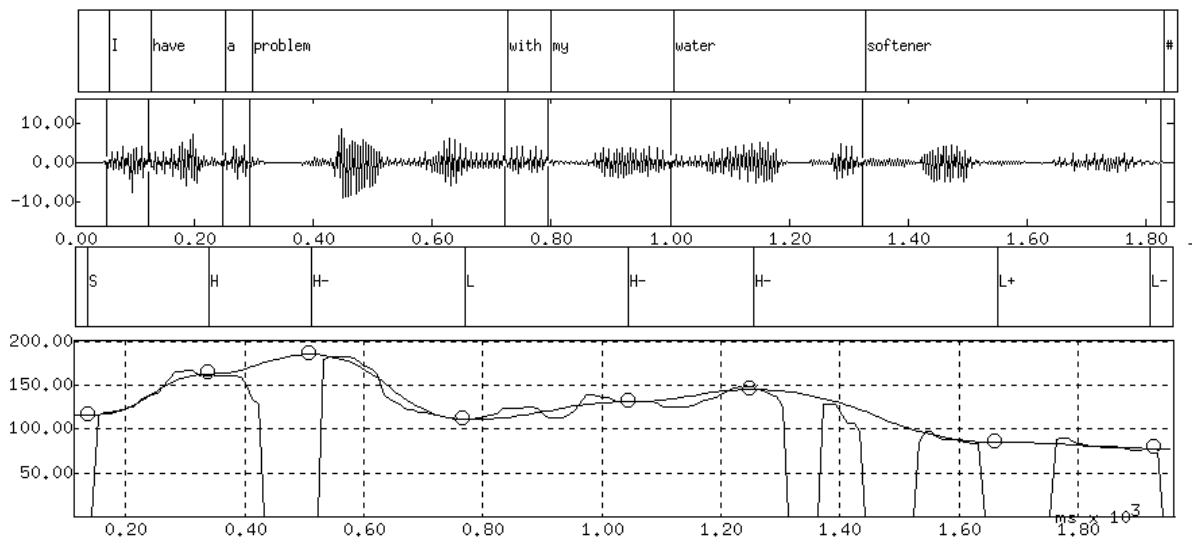
Figure 4. Example of sentence from the English corpus. From top to bottom: orthographic transcription and signal (aligned at word level), symbolic coding and stylized F_0 curve (superimposed to the original F_0).