

# *Automatic analysis of prosody for multi-lingual speech corpora.*

*Daniel Hirst*

CNRS Laboratoire Parole et Langage, Université de Provence,  
13621 AIX EN PROVENCE, France

This chapter outlines a general approach and describes a set of tools for the automatic analysis of multilingual speech corpora. Two levels of representation can be derived automatically: a phonetic representation, which provides an extremely close copy of the original speech signal, and a surface phonological representation, which reduces the variability to a small number of discrete values without great loss of information. The tools have been used to describe the prosody of a number of European and non-European languages.

## **Background**

It is generally agreed today that the single most important progress which needs to be made in order to improve the quality and naturalness of synthetic speech is towards a better understanding and control of prosody. This is true even for those languages which have been the object of considerable research (e.g. English, French, German, Japanese, ...) - it is obviously still more true for the vast majority of the world's languages for which such research is either completely inexistant or is only in a fairly preliminary stage. For a survey of studies on the intonation of twenty languages see Hirst and Di Cristo, (1998). Even for the best-researched

languages there is still very little reliable and robust data available on the prosodic characteristics of dialectal and/or stylistic variability.

It seems inevitable that the need for prosodic analysis of large speech data-bases from a great variety of languages and dialects as well as from different speech styles will increase exponentially over the next two or three decades, in particular with the increasing availability via internet of speech processing tools and data resources.

In this paper I outline a general approach and describe a set of tools for the automatic analysis of multi-lingual speech corpora based on research carried out in the *Laboratoire Parole et Langage* in Aix-en-Provence. The tools have recently been evaluated for a number of European and non-European languages (Hirst *et al.*, 1993, Astésano *et al.*, 1997, Courtois *et al.*, 1997, Mora *et al.*, 1997).

## The General Approach.

The number of factors which contribute to the prosodic characteristics of a particular utterance are quite considerable. These may be universal, language-specific, dialectal, individual, syntactic, phonological, semantic, pragmatic, discursive, attitudinal, emotional... and the list is obviously not at all complete.

Many approaches to the study of prosody attempt to link such factors directly to the acoustic characteristics of utterances. The approach I outline here is rather different. Following Hirst, Di Cristo and Espesser (2000), I propose to distinguish four distinct levels of representation: the *physical* level, the *phonetic* level, the *surface phonological* level and the *underlying phonological* level. Each level of representation needs to be interpretable in terms of adjacent levels of representation.

The underlying phonological level is conceived of as the interface between the representation of phonological form and syntactic/semantic interpretation. Although it is this underlying phonological level which is ultimately conditioned by the different factors listed above, this level is obviously very theory dependent and I shall not attempt to describe it any further here. Instead, my main concern in this paper is to characterise the phonetic and surface phonological levels of representation for prosody.

I assume, following Trubetzkoy, a fundamental distinction between phonology - the domain of abstract qualitative distinctions and phonetics/acoustics the domain of quantitative distinctions. I further assume that phonetics is a level of analysis which provides an interface between the phonological level and the physical acoustic/physiological level of analysis. For more discussion see Hirst and Di Cristo (1998), Hirst, Di Cristo and Espesser (2000).

The aim of the research programme characterised by this approach is to develop automatic procedures defining a reversible mapping between acoustic data and phonetic representations on the one hand and between phonetic representations and surface phonological (or at least 'quasi-phonological') representations on the other hand.

This programme aims, consequently, (at least as a first measure) not to *predict* the prosodic characteristics of utterances but rather to *reproduce* these characteristics in a robust way from appropriate representations.

The first step, *phonetic representation*, consists in reducing the acoustic data to a small set of quantitative values from which it is possible to reproduce the original data without significant loss of information. The second step, *surface phonological representation*, reduces the quantitative values to qualitative ones, again, in so far as possible without losing significant information. In the rest of this paper I describe some specific tools which have been developed in the application of this general research programme.

## The Tools

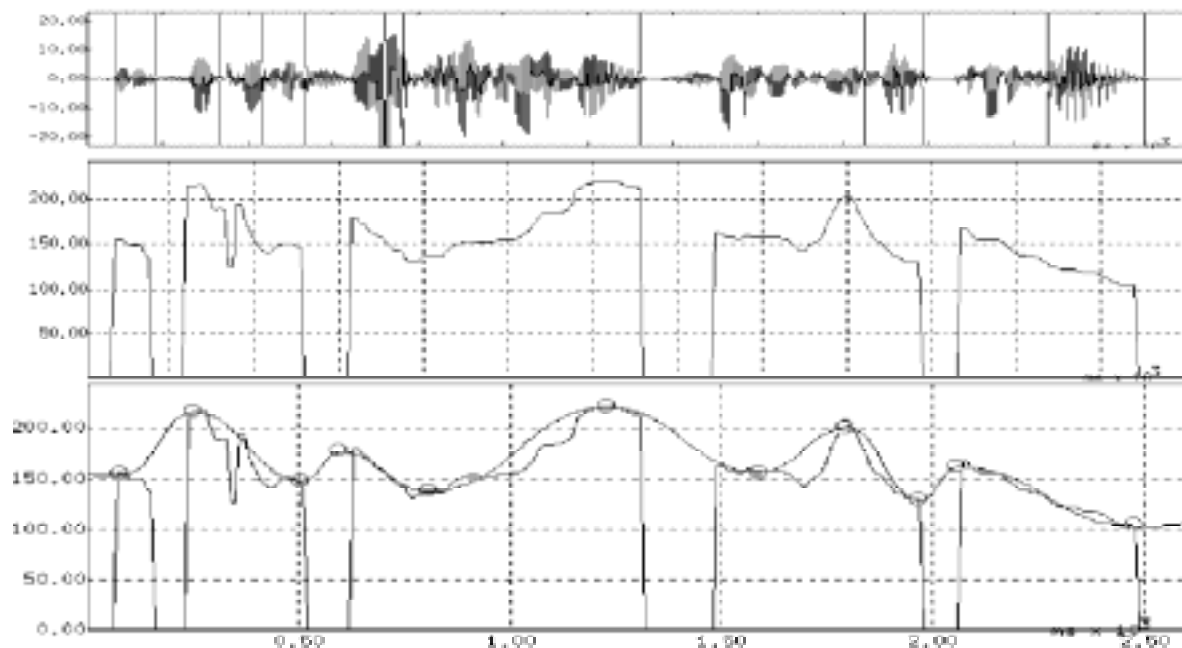
The prosodic characteristics of speech concern the three dimensions of time, frequency and intensity. Our research up until quite recently has been centred mainly on the representation of fundamental frequency although we are currently also working on the integration of durational and rhythmic factors into the representation.

## Phonetic representation

### *Duration*

A *phonetic* representation of duration is obtained simply by the alignment of a phonological label (phoneme, syllable, word etc.) with the corresponding acoustic signal. Usually such alignments have been carried out manually. This task is very labour-intensive and extremely error-prone. It has been estimated that it generally takes an experienced aligner more than fifteen hours to align phoneme labels for one minute of speech (nearly 1000 times real-time).

Recently, software has been developed to carry out this task (or at least a first approximation) automatically (Dalsgaard *et al.*, 1991, Talkin and Wightman 1994, Vorsternan, Martens, and Van Coile, 1996). Such software, which generally uses the technique of Hidden Markov modelling, requires a large hand-labelled training corpus. Recent



**Figure 1.** Wave form (top), F0 trace (middle) and quadratic spline stylisation (bottom) for the French sentence "Il faut que je sois à Grenoble Samedi vers quinze heures" (I have to be in Grenoble on Saturday around 3 p.m.). The stylised curve is entirely defined by the target points, represented by the small circles in the bottom figure.

experiments, however, (Di Cristo & Hirst 1997, Malfrère & Dutoit 1997) have shown that a fairly accurate alignment of phonemic labels can be obtained without prior training by using a diphone synthesis system such as *Mbrola* (Dutoit 1997). Once the corpus to be labelled has been transcribed phonemically, a synthetic version can be generated with a fixed duration for each phoneme and with a constant  $f_0$  value. A dynamic time warping algorithm is then used to transfer the phoneme labels from the synthetic speech to the original signal.

Once the labels have been aligned with the speech signal, a second synthetic version can be generated using the duration defined by the aligned labels and the fundamental frequency of the original signal. This second version is then re-aligned with the original signal using the same dynamic time-warping algorithm. This process, which corrects a number of errors in the original alignment (Di Cristo & Hirst *op. cit*) can be reiterated until no further improvement is made.

### *Fundamental frequency*

A number of different models have been used for modelling or stylising fundamental frequency curves. The MOMEL algorithm (Hirst & Espesser,

1993, Hirst *et al.*, 1997) factors the raw  $f_0$  curve into two components: a microprosodic component corresponding to short-term variations of  $f_0$  conditioned by the nature of the individual phoneme, and a macroprosodic component which corresponds to the longer term variations, independent of the nature of the phonemes. The macroprosodic curves are modelled using a quadratic spline function. The output of the MOMEL algorithm is a sequence of target points corresponding to the linguistically significant targets as seen in the lower panel of figure 1. These target points can be used for close-copy resynthesis of the original utterance with practically no loss of prosodic information by comparison with the original  $f_0$  curve. It would be quite straightforward to model the microprosodic component as a simple function of the type of phonematic segment, essentially as unvoiced consonant, voiced consonant, sonorant or vowel, see Di Cristo and Hirst (1986), and to add this back to the synthesised  $f_0$  curve, although this is not currently implemented in our system.

## Surface phonological representation

### *Duration*

The duration of each individual phoneme, as measured from the phonetic representation, can be reduced to one of a finite number of distinctions. The value for each phoneme is calculated using the z-transform of the raw values with respect to the mean and standard deviation of the phoneme. (Campbell 1992). Currently for French, we assume four phonologically relevant values of duration: normal, shortened, lengthened and extra-lengthened (cf Di Cristo *et al.* 1997, Hirst 1999). A lot more research is obviously needed in this area.

### *Fundamental frequency*

The target points modelled by the *MOMEL* algorithm described above could be interpreted in a number of ways. It has been shown, for example, (Mixdorff, 1999) that the quadratic spline stylisation provides a good first step for the automatic estimation of parameters for Fujisaki's superpositional model of  $f_0$  contours (Fujisaki, 2000) and that this can then

be used for the automatic recognition and characterisation of ToBI labels (Mixdorff & Fujisaki, 2000, this volume).

A different interpretation is to reduce the target points to "phonological-like" symbols using the *INTSINT* transcription system described in Hirst & Di Cristo (1998) and Hirst *et al.* (2000).

This system represents target points as values either globally defined relative to the speaker's pitch range: Top (**T**), Mid (**M**) and Bottom (**B**), or locally defined relative to the previous target-point. Relative target-points can be classified as Higher (**H**), Same (**S**) or Lower (**L**) with respect to the previous target. A further category consists of smaller pitch changes which are either slightly Upstepped (**U**) or Downstepped (**D**) with respect to the previous target.

Two versions of a text-to-speech system for French have been developed, one stochastic (Courtois *et al.*, 1997) and one rule-based (Di Cristo *et al.*, 1997) implementing these phonetic and surface phonological representations.

The software for deriving both the phonetic stylisation as a sequence of target points and the quasi-phonological coding with the *INTSINT* system are currently being integrated into a general-purpose prosody editor *ProZed* (Hirst 2000a).

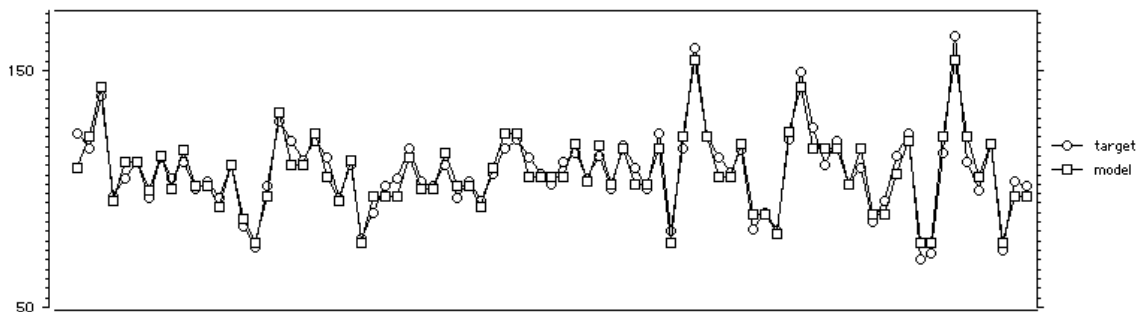
## **Evaluation and improvements.**

Evaluation of the phonetic level of representation of  $f_0$  for English, French, German, Italian, Spanish and Swedish has been carried out on the EUROM1 corpus (Chan *et al.*, 1995) within the MULTEXT project (Véronis *et al.*, 1994, Astésano *et al.*, 1997). The results show that the algorithm is quite robust. On the English and French recordings (all together about one and a half hours of speech) around 5% of the target points needed manual correction. The majority of these corrections involved systematic errors (in particular before pauses) which an improvement of the algorithm should eliminate.

Evaluation of the surface phonological representation has also been undertaken (Campione *et al.*, 1997). Results for the French and Italian versions of the EUROM1 corpus show that while the algorithm described in Hirst *et al.* (2000) seems to preserve most of the original linguistic information it does not provide a very close copy of the original data and it also contains too many rather arbitrary degrees of freedom. A more highly constrained version of the algorithm, however (Hirst, 2000a, 2000b), assumes that the relationship between the symbolic coding and the actual target points can be defined on a speaker-independent and perhaps even language-independent basis with only two speaker dependent variables corresponding to the speaker's *key* (approximately his mean fundamental frequency) and his overall pitch *range*. For the

passages analysed the values of *key* and *range* were optimised within the parameter space:  $key = \text{mean} \pm 20 \text{ Hz}$ ,  $range \in [0.5-2.5]$  octaves. The mean optimal *range* parameter resulting from this analysis was not significantly different from 1.0 octave. It remains to be seen, however, how far this result is due to the nature of the EUROM1 corpus which was analysed (40 passages consisting each of 5 semantically connected sentences) and whether it can be generalised to other speech styles and other (particularly non-European) languages.

Figure 2 shows the sequence of target points derived from the raw f0 curve compared with those generated from the INTSINT coding. The actual MOMEL targets, INTSINT coding and derived targets for this 5 sentence passage are available in the accompanying example file (HI00E01.txt, CD-ROM).



**Figure 2.** Target points output from the MOMEL algorithm and those generated via the optimised INTSINT coding algorithm using the two parameters  $key = 109 \text{ Hz}$ ,  $range = 1.0$  octave for passage fa0 of the Eurom1 (English) corpus.

## Perspectives.

The *ProZed* software described in this chapter will be made freely available for non-commercial research and will be interfaced with other currently available non-commercial speech processing software such as *Praat* (Boersma, 1995-2000) and *Mbrola* (Dutoit, 1997). Information on these and other developments will be made regularly available on the web page and mailing list of *SProSIG*, the Special Interest Group on Speech Prosody recently created within the framework of the International Speech Communication Association *ISCA*.

<http://www.lpl.univ-aix.fr/projects/sprosig>

It is hoped that this will encourage the development of comparable speech databases, knowledge bases and research paradigms for a large number of languages and dialects and that this in turn will lead to a significant

increase in our knowledge of the way in which prosodic characteristics vary across languages, dialects and different speech styles.

## References

- Astésano, C., Espesser, R., Hirst, D.J., & Llisterri, J. (1997). Stylistisation automatique de la fréquence fondamentale : une évaluation multilingue. *4e Congrès Français d'Acoustique*, (14-18 avril 1997, Marseille), 441-443.
- Boersma, P., & Weenink, D. (1995-2000). *Praat: a system for doing phonetics by computer*. <http://www.fon.hum.uva.nl/praat/>
- Campbell, W.N. (1992). *Multi-level Timing in Speech*, PhD Thesis, University of Sussex.
- Campione, E., Flachaire, E., Hirst, D.J., & Véronis, J. (1997). Stylistisation and symbolic coding of  $f_0$ , a quantitative approach. *ESCA Tutorial and Research Workshop on Intonation*. (18.-20.September. Athens).
- Chan, D., Fourcin, A., Gibbon, D., Granström, B., Huckvale, M., Kokkinas, G., Kvale, L., Lamel, L., Lindberg, L., 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) 1, 867-880.
- Courtois, F., Di Cristo, Ph., Lagrue, B., & Véronis, J. (1997). Un modèle stochastique des contours intonatifs en français pour la synthèse à partir des textes. *4ème Congrès Français d'Acoustique*. (Marseille, avril 1997), 373-376.
- Dalsgaard, P. Andersen, O., & Barry, W. (1991). Multi-lingual alignment using acoustic-phonetic features derived by neural-network technique. *ICASSP-91*, 197-200.
- Di Cristo, A., Di Cristo P., & Véronis, J. (1997). A metrical model of rhythm and intonation for French text-to-speech. *ESCA Workshop on Intonation : Theory, Models and Applications*. (Athens, September 1997).
- Di Cristo, Ph. & Hirst, D.J., (1997). Un procédé d'alignement automatique de transcriptions phonétiques sans apprentissage préalable. *4e Congrès Français d'Acoustique*, (Marseille, April 1997), 425-428
- Dutoit, T. (1997). *An Introduction to Text-to-Speech synthesis*. Dordrecht: Kluwer Academic Press.
- Fujisaki, H. (2000). The physiological and physical mechanisms for controlling the tonal features of speech in various languages. In *Proceedings Prosody 2000: Speech recognition and Synthesis* (Kraków, October 2000).

- Hirst, D.J. (1999). The symbolic coding of segmental duration and tonal alignment. An extension to the INTSINT system. In *Proceedings Eurospeech* (Budapest 1999).
- Hirst, D.J. (2000a). ProZed. A multilingual prosody editor for speech synthesis. In *Proceedings IEE Colloquium on State-of-the-Art in Speech Synthesis*. (London, April 2000).
- Hirst, D.J. (2000b). Optimising the INTSINT coding of F0 targets for multi-lingual speech synthesis. in *Proceedings ISCA Workshop: Prosody 2000 Speech recognition and Synthesis* (Kraków, October 2000).
- Hirst, D.J. & Di Cristo, A. (eds) (1998). A survey of intonation systems. in Hirst & Di Cristo (eds) *Intonation Systems : a Survey of Twenty Languages*. Cambridge: Cambridge University Press, 1-44.
- Hirst, D.J.; Di Cristo, A., Le Besnerais, M., Najim, Z., Nicolas, P., & Roméas, P. (1993). Multi-lingual modelling of intonation patterns. *Proceedings ESCA Workshop on Prosody*. (Lund, September 1993), 204-207.
- Hirst, D.J., Di Cristo, A., & Espesser, R. (2000). Levels of representation and levels of analysis for the description of intonation systems. in M. Horne (ed) *Prosody : Theory and Experiment*. Dordrecht: Kluwer Academic Publishers.
- Malfrère, F., & Dutoit, T. (1997). High quality speech synthesis for phonetic speech segmentation, *Proceedings EuroSpeech 97*, (Rhodes, September 1997).
- Mixdorff, H. (1999). A novel approach to the fully automatic extraction of Fujisaki model parameters. *ICASSP 1999*.
- Mixdorff, H., & Fujisaki, H. (2000). Symbolic versus quantitative descriptions of f0 contours in German: Quantitative modelling can provide both. In *Proceedings Prosody 2000: Speech recognition and Synthesis* (Kraków, October 2000).
- Mora, E., Hirst, D. & Di Cristo, A. (1997). Intonation features as a form of dialectal distinction in Venezuelan Spanish. *ESCA Workshop on Intonation : Theory, Models and Applications*. (Athens, September 1997).
- Talkin & Wightman, C. (1994). The aligner. *Proceedings ICASSP 1994*.
- Véronis, J., Hirst, D.J., Espesser, R., & Ide, N. (1994). NL and speech in the MULTEXT project. *AAAI '94 Workshop on Integration of Natural Language and Speech*, 72-78.
- Vorsterman, A. Martens, J.P., & Van Coile, B. (1996) Automatic segmentation and labelling of multi-lingual speech data. *Speech Communication* 19, 271-293.