Pitch Range in spontaneous speech:semi-automatic approach versus subjective judgement

Cristel Portes† and Albert Di Cristo†

† Laboratoire Parole et langage, Université de Provence, Aix en Provence, France

E-mail: cristel.portes@lpl.univ-aix.fr, albert.dicristo@lpl.univ-aix.fr

ABSTRACT

Our multilinear conception of prosody leads us to take into account "pitch range" two dimensions: register level (ReLe) and register span (ReSp) which, in our opinion, assume linguistic functions in discourse structure. We propose two different models for ReLe and one for ReSp. They are implemented using an automatic clustering derived from the target values given by a phonetic/phonological model of intonation (MOMEL/INTSINT). Comparing them to a subjective classification of ReLe and ReSp, these models and their implementation are evaluated as an approximate but usable classification.

1. INTRODUCTION

Prosodic models which belong to the present day intonational phonology paradigm [1] propose to code the tonal organisation of uterrances as a sequence of tonal segments (H and L tones following the classical "bitonal approach").aligned with phonemes and syllabes in the segmental chain. In Pierrehumbert's model [2], the phonological tonal organisation is directly coded from the acoustic representation of the F0 curve. A more recent approach argues that several intermediate levels are needed between physical and phonological representations.[3]. Both approaches have in common the assumption that a linear coding with discrete symbols (intrinsic representation) is able to capture the basic intonation system of a language. However, it appears that an extrinsic (or orthogonal) component cannot be neglected. This refers to phenomena known as "pitch range" and "downtrends". The first notion refers to pitch variations on the vertical scale of height, the second notion deals with declination, downstep and downdrift effects [4].

The issues of the functional role and the linguistic status of orthogonal phenomena gave rise to numerous questions and conflicting explanations that we cannot resume here (see [5] for a review). Most of the authors present the orthogonal component as a gradient one according to what Ladd called the "Free gradient variability hypothesis" [6]. It is said to convey paralinguistic information and thus does not concern phonological analyses. Nevertheless, some authors have a more moderate opinion, claiming that orthogonal phenomena are also needed the full linguistic description of intonation (see [6] and [7]). We also think that the orthogonal component of intonation cannot be neglected since we show that it plays an important role in the organisation of spontanuous oral discourse which is our current research theme (see [8]; [9]).

The purpose of the present paper is to test a method allowing the categorical coding of the two dimensions of "pitch range": register level and register span [1]. The first dimension refers to local and global variations of the globally percieved voice height; the second dimension deals with what is sometimes called voice dynamic, namely the distance between maximum and minimum values of fundamental frequency used by one speaker.

2. INTSINT APPROACH

The INTSINT model on which we base our approach is made up of two stages respectively corresponding to a modeling of the F0 curve and a symbolic coding of significant values of the modeled curve. First, the MOMEL algorithm represents the F0 curve as a sequence of targets interpolated by a monotonic quadratic spline function. Target coding is automatically or manually performed with the INTSINT alphabet of symbols. Three symbols code absolute values of the speaker's pitch range: T(op), M(id) and B(ottom). Five other symbols are used to code relative pitch variations within this range: L(ower), H(igher), S(ame), U(pstepped) and D(ownstepped). INTSINT coding is said to be a "surface phonological representation" of intonation since it allows the conversion of a continuous representation of height variation into a representation made up of sequences of discrete symbols [10]. Although this symbolic representation succeeds in capturing fundamental aspects of tonal organisation, syntagmatic as well as paradigmatic, its current version does not quantify the orthogonal component. We still think that INTSINT coding may provide a reference frame.

3. ANALYSES

3.1 MATERIAL

The corpus chosen here is a 3 minute spontaneous speech sample corresponding to one conversational turn. It is performed by a unique speaker during a conversational exchange between 6 participants.

Our choice of spontanous speech (versus laboratory speech) corresponds to our conviction that particularly spontaneous speech uses register level and register span contrasts for different discourse purposes.

Our choice of a unique speaker is justified first by the

ISBN 1-876346-48-5

assumed prospective nature of the study. Second it is appropriate to the possible application of our results for speech synthesis, where the challenge is to generate the prosody of a particular speaker.

3.2 OBSERVATION WINDOW

The main hypothesis of the research is that variations of register level and register span contribute to distinguish discourse segments of different length. This plays an important role in the interpretation of global discourse and its utterances. The length of the segments cannot be specified in advance. We choose the Intonational Unit (IU) because of the relative consensus around its definition: its boundary corresponds to a specific nuclear contour with a continuating or concluding function (non conclusive IU versus conclusive IU).

3.3 SUBJECTIVE ANALYSIS

It was performed by the two authors of the present paper. They listened to the auditory signal and had the non punctuated orthographic transcription of the text to note their observations.

First they were required to segment the corpus into IUs.

Then they had to note register level and register span variations using a set of symbols (in bold below) attributed to each IU:

- Register Level: **Rai**(sed), **Low**(ered), **N**(ormal)
- Register Span: **Exp**(anded), **Com**(pressed), **N**(ormal)

3.4 SEMI AUTOMATIC ANALYSIS

We propose a semi automatic analysis of pitch range based on measurements realised only with MOMEL targets. First we design two models for register level and one for register span. Then we try to transpose these models into an automatic classification in order to compare the result with subjective classification.

- Register Level

On the one hand, the register level of each IU may be represented by the onset target of the IU. Following the literature, this onset target is often considered as a key value where a "resetting" may occur. So we take it as a good reference point among the whole values of F0 targets.

On the other hand, the valleys of the F0 curve also may give a good account of register level : they vary less than F0 peaks, but they seem to follow general variations on discourse relevant domains, as Figure 1 shows.

L targets then appear as a better value to account for reference level *variations* [11], than bottom targets (B in the INTSINT alphabet) which values are not reliable [1].

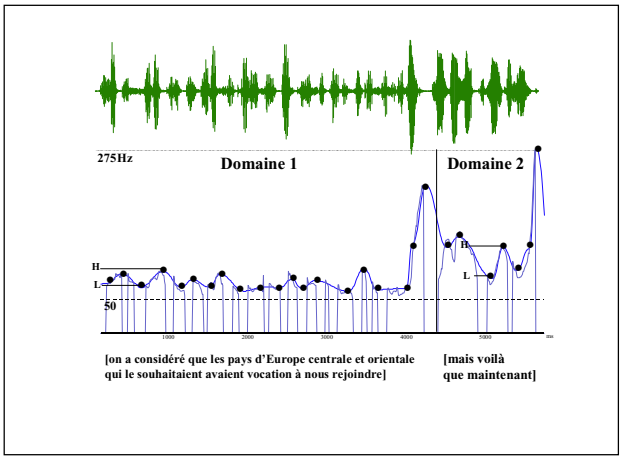

Figure 1: orthogonal (vertical) variations of L and H MOMEL targets following discourse relevant domains

- Register span

Here we follow Ladd and Patterson [5] in choosing the distance between F0 peaks and F0 valleys, i.e. between L targets and H targets, as a good model for register span. As Figure 1 shows, we see clearly that the difference of height between L and H is greater in domain II than in domain I; two domains which are distinguished in the discourse structure.

- Automatic classification

The statistical method of *clustering* gives us a good way to approach automatically the kind of classification on a gradient scale that the cognitive system is supposed to perform with F0 values on the orthogonal dimension of intonation.

We use the k-mean algorithm [12], carried out with the R software [13]. We program each clustering to obtain three clusters, corresponding to the three classes required for the subjective analysis (see 3.3 below).

ReLe: To improve our first model for register level, we performed a clustering on the whole of the F0 target values. We may thus situate the onset F0 target of each IU in one of the three clusters corresponding to its automatic classification.

We do the same for our second model: first we perform a clustering on L target values only. Then we may decide to which automatic class one IU belongs, given the mean of its L target values.

ReSp: Here we confront the mean of H-L distance values of each IU to the three clusters obtained on all the H-L differences of the corpus.

4. RESULTS

4.1 RESULTS OF SUBJECTIVE ANALYSIS

- Segmentation into Intonation Units

Among the 76 IUs segmented by the two authors of the present paper on the 3 minute corpus, 7 IUs are identified by only one transcriber whereas 69 IUs were common to both transcriptions. These are the IUs investigated here.

- Subjective coding of pitch range (Register level and Register span)

First, subjective coding shows great variability according to the transcriber. But a fine analysis reveals more agreement between the two judges. Table 1 distinguishes 4 different types of responses :

Case1	Case 2	Case 3	Case 4
Same response	One tier in common, the other different	Both tier different without contradiction	Contradiction
29	29	8	3
42%	42%	12%	4%

Table 1: Population and percentage of responses by type, from agreement to contradiction. One response per IU; two tiers in each case: one for register level, the other for register span.

In case 1, "Same response" indicates that both transcribers code the same for both tiers (level and span).

Analysis of case 2 shows that each time both tiers are marked (coded with another symbol than N(ormal)=14 times=50% of case 2), these codes "go in the same direction", i.e. you find Rai with Exp and Low with Com. This is consistent with a claim in the literature that level and span often go together and are difficult to distinguish perceptually [1].

We therefore assume there is a contradiction when one transcriber codes "in the high direction", i.e. Rai or Exp and the other codes "in the low direction", i.e. Low or Com. This may happen on different tiers or in the same tier. We find only 3 such occurences (case 4).

Moreover, differences between judges is sometimes explained because one of them perceives fine contrasts than the other (more IUs in contrast=10 occurences). But most of the time, even when the coding is different, the direction of the contrast with the former IU is the same for both transcribers (only 6 cases of different directions).

We note then that the apparent diversity of coding hides quite a high degree of agreement on pitch range contrasts between transcribers.

- What to compare with automatic classification?

In order to compare the subjective coding with the semi-automatic classification presented below, we proceed as follows: we keep responses of case 1, as well as common responses of case 2; where the other tier "goes in the same direction", we keep it as well. In other situations, we keep the direction of contrast as it appears and both authors-coders find an agreement for recoding.

4.2 RESULTS OF SEMI-AUTOMATIC ANALYSIS

- Clustering on the whole target space/ onset target location: first model for register level

Cluster 1 = **Low**	Cluster 2 = **N**	Cluster 3 = **Rai**
Min = 50 Mean = 79.05 Max = 99	Min = 100 Mean = 119.7 Max = 155	Min = 160 Mean = 199 Max = 296

Table 2: Limits and means of the 3 clusters of the whole F0 targets and their corresponding code.

The categories defined in table 2 allow the location of the onset target of each IU, i.e. its coding with the appropriate symbol (Low, Rai, N).

- Clustering on L targets only: second model for register level

Cluster 1 = **Low**	Cluster 2 = **N**	Cluster 3 = **Rai**
Min = 50 Mean = 73.55 Max = 89	Min = 90 Mean = 105.3 Max = 131	Min = 135 Mean = 162.2 Max = 200

Table 3: Limits and means of the 3 clusters of L F0 targets and their corresponding code.

Each IU is coded according to the location of the mean of its L targets.

- Clustering on H-L distances (differences): model for register span

Cluster 1 = **Low**	Cluster 2 = **N**	Cluster 3 = **Rai**
Min = 0 Mean = 24.85 Max = 52	Min = 53 Mean = 80.03 Max = 112	Min = 117 Mean = 147.2 Max = 200

Table 3: Limits and means of the 3 clusters of H-L differences and their corresponding code.

In this case we code each IU locating the mean of the differences between H and L targets that belong to the IU.

4.3 SUBJECTIVE JUDGEMENT VS. CLUSTERING

- First model for register level

Automatic coding	Subjective coding
33 L	13 L mean=80,3 16 N mean=85,6 **4 R**
26 N	1 L 14 N mean=112,4 11 R mean=121,5
2 R	2 N 7 R

Table 4: Register level 1: comparison of semi-automatic coding versus subjective coding (in number of IUs coded L(ow), N or R(ai)). In bold: inconsistent values.

ISBN 1-876346-48-5

- Second model for register level

Automatic coding	Subjective coding
38 L	14 L mean=71,8 20 N mean=78,3 **4 R**
26 N	13 N mean=100,5 11 R mean=106,3
2 R	2 R

Table 5: Register level 2: comparison of semi-automatic coding versus subjective coding (in number of IU coded L(ow), N or R(ai)). In bold: inconsistent values.

- Register span

Automatic coding	Subjective coding
32 C	11 C mean=27,6 20 N mean=39,3 **1 E**
31 N	23 N mean=77,7 8E mean=85,5
1 E	1 E

Table 6: Register span: comparison of semi-automatic coding versus subjective coding (in number of IUs coded C(om), N or E(xp). In bold: inconsistent values.

4.4 REGISTER LEVEL VS. REGISTER SPAN

Comparison is made for model 1 and also for model 2. On the one hand we compare level to span in semi-automatic coding, on the other hand we do the same for subjective coding. Each time the result of the comparison is very similar: about 60% of convergent coding; contradiction being exceptional (1 or 2 cases on 69 IU).

4.5 DISCUSSION

We note that all clusterings tend to produce larger classes than subjective judgment does. But we also show that inconsistent values are very few so that the classification order is preserved. Comparing mean values of subjective subclasses into automatic classes tell the same: "higher" codes (E>N>C) have higher means, even if we don't have enough values to get statistical confirmation. It will be the subject of further investigation to understand why clustering have too high boundaries: perhaps by excluding some extreme values (focuses, strongest UI boundaries) from computation.

Semi-automatic vs. subjective comparison does nost say which model is better for register level, both behaving in a very similar way. More investigation is needed here too.

As claimed in the literature, we find a real convergence between register level and register span: here we find no difference between automatic and subjective classification, that could explain the deviation between them.

5. CONCLUSIONS

This work confirms that discourse structure uses prosodic contrasts based on pitch range variations that listeners are able to perceive: it adds new arguments in favour of linguistic use of the intonational orthogonal dimension. A first version of semi-automatic implementation of register level and register span variations related to discourse structure is proposed. It gives a reliable approximation of perceptual classification.

ACKNOWLEDGMENTS

We are very greatfull to A. El Ahmadi for the idea he gave us to use clustering in order to capture the orthogonal dispersion of F0 targets. We also deeply acknowledge R. Espesser for running statistical analysis.

REFERENCES

[1] D. R. Ladd, *Intonational Phonology*, Cambridge UK: Cambridge University Press, 1996.

[2] J. Pierrehumbert, *The Phonology and Phonetics of English Intonation*, PhD Thesis, MIT, 1980.

[3] D. J. Hirst, A. Di Cristo, R. Espesser, "Levels of representation and levels of analysis for the description of intonation systems", in *Prosody: Theory and Experiment*, M. Horne, Ed., pp. 51-88. Dordrecht: Kluwer, 2000.

[4] B. Connell and D. R. Ladd, "Aspects of pitch realisation in Yoruba", *Phonology*, vol. 7, pp. 1-30, 1990.

[5] D. Patterson, *A linguistic approach of pitch range modeling,* PhD Thesis, University of Edinburgh, 2000.

[6] D. R. Ladd, "Constraints on the gradient variability of pitch range, or Pitch level 4 lives!", *Papers in Laboratory Phonology*, vol. III, pp. 43-75. Cambridge University Press, 1994.

[7] G. N. Clements, "The status of register in intonation theory", *Papers in Laboratory Phonology*, vol. I, pp. 58-71. Cambridge University Press, 1990.

[8] C. Portes, E. Rami, C. Auran, A. Di Cristo, "Prosody and discourse :a multilinear analysis", *Proceedings of the first International Conference on Speech Prosody*, pp. 579-582. Aix-en-Provence, April 8-10, 2002.

[9] A. Di Cristo, C. Auran, R. Bertrand, C. Chanet, C. Portes, "An integrative approach of the relations of prosody to discourse: toward a multilinear representation and an interface network", to appear in *Proceedings of International AAI Workshop "Prosodic Interfaces"*. Nantes, March 27-29, 2003.

[10]D. J. Hirst, A. Di Cristo, *Intonation Systems*, Cambridge UK: Cambridge University Press, 1998.

[11]M. Liberman *et al*, "The phonetic interpretation of tone in Igbo", Phonetica, 50, pp. 147-160, 1993.

[12]J. A. Hartigan and M. A. Wong, "A K-means clustering algorithm", *Applied Statistics*, 28, pp. 100-108, 1979.

[13]R. Ihaka and R. Gentleman, "R: a Language for Data Analysis and Graphics", Journal of Computational and Graphical Statistics, 3, vol. 5, pp. 299-314, 1996.