INTEGRATING CHANGES OF REGISTER INTO AUTOMATIC INTONATION ANALYSIS

Céline De Looze & Daniel Hirst

Laboratoire Parole et Langage, CNRS et Université de Provence, Aix-en-Provence, France
{celine.delooze; daniel.hirst}@lpl-aix.fr

Abstract

While current tools for the automatic analysis and modeling of intonation are satisfactory for laboratory or isolated sentences, they appear insufficient for the study of longer stretches of authentic speech, which are in general marked by systematic changes of register. This study shows that implementing automatically detected register changes significantly improves the accuracy of the automatic coding of intonation patterns with the INTSINT algorithm. This implies, upstream, to define a reliable measurement of register and a way to detect its changes automatically.

Index Terms: register changes, intonation systems, pitch scale

1. Introduction

The study of authentic speech (i.e. speech, rather read or spontaneous, with a communicative intention), unlike that of laboratory speech, needs to take into account the fact that fundamental frequency patterns can be of two types. They can be local pitch characteristics which correspond to the phonological representation of intonation patterns and longer term characteristics which correspond to register changes in key (or level) and range (or span). Their overlapping and interaction make their separation difficult.

Bolinger [1] pointed out many years ago that, in a system like that of Trager & Smith [2], with four distinct pitch levels, it was not possible, for example, to distinguish a high falling pitch movement /41/ in a narrow pitch range from a low falling movement /31/ or /21/ in a neutral and expanded pitch range. This argument was later taken up by Janet Pierrehumbert in her influential study of American English intonation [3] and the assumption that there is only a binary phonological distinction between H and L tones became a cornerstone of both ToBI [4] as well as many variant models within the general Autosegmental-Metrical framework of Intonational Phonology (eg [5], [6], [7], [8]).

In INTSINT ([9], [10], [11]), an explicit multilingual alphabet for the representation of surface phonological contrasts for intonation, high and low levels can be accounted for by a set of 8 possible tonal values for each significant point in a given pitch pattern. These are either interpreted globally, i.e. with respect to the speaker's global register, or locally, i.e. with respect to the previous tone. However, any linear coding with discrete symbols which neglects long-term variations cannot adequately account the intonation patterns of a language.

While current tools for the automatic analysis and modeling of intonation are satisfactory for laboratory or isolated sentences, they are insufficient for the study of longer stretches of authentic speech, which are in general marked by systematic changes of register. Yet, implementing register changes in intonation analysis is not an easy task: first, it implies a reliable measurement for register, for which there is no current consensus in the literature; next, it implies, upstream, the definition of the domain within which register changes take place. However, register changes, while they are considered to span longer-term domains, do not seem to belong to a specific domain. Rather, they depend on the hierarchical and organisational structure of discourse, on the intentions and mood of the speaker; and they may therefore span different domains.

Defining a reliable measurement of register and detecting its changes automatically, therefore becomes a first step before implementing them into intonation analysis. Two types of measurements have been used in the literature for register: acoustic measurements (A-measures), i.e. based on the long term distributional properties of f0, on the one hand, and linguistic measurements (L-measures), i.e. based on linguistically-motivated targets, on the other hand. The most commonly used A-measures are the median and the mean for key and the difference between the minimum and the maximum, or the difference between percentiles (95^{th}-5^{th}; 90^{th}-10^{th}) or again standard deviations around the mean, for range. However, these measurements have been criticized ([12], [13]) because they are often affected by pitch tracking errors and are not adapted to non-normal distributions of f0. The authors prefer L-measures, such as the average of final lows (for key) and the difference between the average of non-phrase-initial accent peaks and the average of post-accent valleys (for range), justifying their use by the claim that they are more correlated with listeners judgments than A-measures. We show, however, that the dichotomy A-L is artificial and that using "controlled" A-measures gives as good results as using hand-labelled L-measures, at least for the measurement of register key and range.

The A-measures defined were integrated into a clustering algorithm (ADoReVA) we developed for the automatic detection of register changes. We then show that integrating automatically detected register changes significantly improves the accuracy of the automatic coding of intonation patterns with the INTSINT algorithm.

2. Corpora

Four corpora were used in this study:

AIX-MARSEC *([14])* **–** A total of 54 minutes of recording (13 female and 38 male speakers of standard British English) were selected from the AIX-MARSEC corpus. Mainly prepared monologues, the recordings correspond to commentaries, new broadcasts, lectures, religious broadcasts, magazine-style reporting, fiction, poetry, dialogues and propaganda.

PAC *(Phonologie de l'Anglais Contemporain, [15])* – A total of 30 minutes of newspaper article-like readings were selected from the PAC (5 female and 3 male speakers from Northern England).

CID *(Corpus of Interactional Data, [16])* **–** A total of 30 minutes of dialogues recorded in a sound-proof room (3 female and 3 male speakers of regional French - Marseille) were selected from the CID corpus.

PFC *(Phonologie du Français Contemporain, [17])* **–** A total of 30 minutes of recording (6 female and 4 male speakers of regional French - Marseille) were selected from the PFC

corpus. The recordings consisted of newspaper article-like readings.

3. Acoustic vs. Linguistic measurements

In order to evaluate which measurements capture at best a speaker's register and its changes, different acoustic and linguistic measurements were tested and compared.

3.1 Acoustic measurements

A-measures used for key and range were the median (the mean being too sensitive to erroneous values) and the log2(maximum/minimum) respectively. Log-transformation brought the distribution of values closer to a Gaussian distribution, as can be seen in *Figure 1*, and therefore justifies taking maximum and minimum values for the measurement of register range.

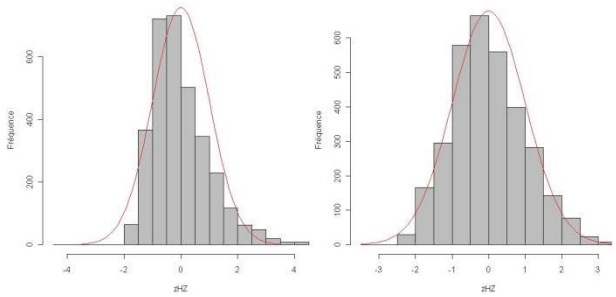

Figure 1: Distribution of f0 samples after a z-score transformation on the left, and after a z-score and log transformation on the right.

Measurements were obtained using the Praat [18] software. In order to avoid possible pitch tracking errors, pitch floor and pitch ceiling were set to the values q15*0.83 (where 'q' stands for percentile) and q65*1.92. These formulae (as well as the formulae q25*0.75 – q75*1.5 and q35*0.72 – q65*1.90) have been shown [19, 20, 21] to give a better estimation of pitch extrema, therefore to exclude more octave errors at the extreme of the f0 distribution, than setting pitch floor and ceiling parameters to the default values (60 – 600) or to default values depending on the gender of the speaker (female: 100 – 500; male: 75 – 300).

3.2 Linguistic measurements

L-measures for key and range were obtained automatically using the absolute values of the INTSINT [22] alphabet T(op), M(id) and B(ottom). Key was measured in terms of the mean of M targets, and the mean of the B targets. Range was measured in terms of the interval between the mean of T targets and the mean of B targets. Automatic measurements of key and range obtained by the MOMEL-INTSINT algorithm were also used.
Key is given in Hz, range in octaves (log2(Hz)).

3.3 Statistical analyses

A-measures and L-measures were compared using linear regression. We tested the hypothesis of a possible linear relation between these two variables. If the hypothesis proves to be true, then A-measures and L-measures may be considered equivalent for measuring register.

First, the correlation between the A-measure **median** and the L-measure mean of M targets (**MEAN-M**) is observed. The linear model shows a coefficient of determination (R^2) of 0.9259. The correlation between the median and the mean of B targets (**MEAN-B**) and the median and automatic measure of key, as obtained with the INTSINT-MOMEL algorithm (**INTSINT-Key**), were evaluated. Regression models show high correlations between median and MEAN-B (R^2=0.9218) and median and INTSINT-Key (R^2=0.9475). It can therefore be concluded that A-measures and L-measures are equivalent for the measure of *register key*.
To test whether this is also the case for the measurement of *register range*, the correlation between, on the one hand, the A-measure log2(maximum/minimum) (**max-min**) and the L-measure difference between the mean of T targets and the mean of B targets (**T-B**), and on the other hand, the correlation between max-min and the automatic L-measure range, i.e. as obtained with the INTSINT-MOMEL algorithm (**INTSINT-Range**), are observed. Contrary to key measures, regression models show that both max-min and T-B, and max-min and INTSINT-Range are not highly correlated (R^2=0.1393 & R^2=0.08627). This may be explained by the fact that whereas max-min is based on *one* maximum value and on *one* minimum value of the f0 distribution, T-B and INTSINT-Range are based on *a set of* high and low tonal targets.

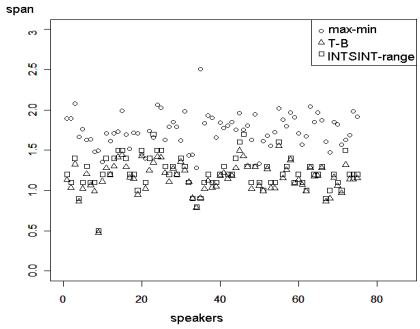

*Figure 2: Graphical representation of the span values obtained with **max-min** (log2(maximum/minimum)), **T-B** (the difference between the mean of T targets and the mean of B targets) and **INTSINT-Range** (as obtained with the INTSINT-MOMEL algorithm). Values are given in a logarithmic scale (log2).*

As can be seen in *Figure 2*, the values obtained with max-min are mainly higher than the T-B and INTSINT-Range values.
Whether speakers actually rely on extreme values or on a set of tonal targets, only a perception test may answer. However, we feel that something more than just two extreme values are probably necessary to judge a speaker's register. We can consequently wonder how we might obtain similar values as those obtained with L-measures from the long-term distributional properties of f0. If we look back to the correlation between the median and MEAN-B, an interesting point may be raised. If the median is also strongly correlated with the mean of T targets (**MEAN-T**), then we would have a way to measure the register range in the way L-measures do.
The correlation between median and MEAN-T is calculated. The regression model shows a high correlation between these two variables, with a R^2 of 0.9152. If MEAN-B and MEAN-T can both be predicted from median, then, this can be used for the measurement of register range. Regression models give the following linear relations (the intercept is not included since it is not significant; p-val = 0.161):
*MEANB=0.706*median*
*MEANT=1.561*median*

As we are interested in giving a measurement of register range on a log scale, we looked at the linear relations on an octave (log2) scale. Relations are as followed (again, the intercept is not significant):
*LOG2(MEANB)=0.928*log2(median)*
*LOG2(MEANT)=1.089*log2(median)*
Therefore, A-measures of register key and range would be:
KEY = median (Hz)
*RANGE = 0.161*log2(median).*

3.4 Log2(Hz/median): a natural scale for pitch?

If we look back at the formulae *MEAN-B=0.706*median & MEAN-T=1.561*median*, it's interesting to see that both coefficient 0.706 and coefficient 1.561 are very close to half an octave (log2(0.706) = -0.5 & log2(1.561) = 0.6). And in fact, if we plot the interval of MEAN-B and MEAN-T with respect to the median, we obtain the following graph (*Fig. 3*).

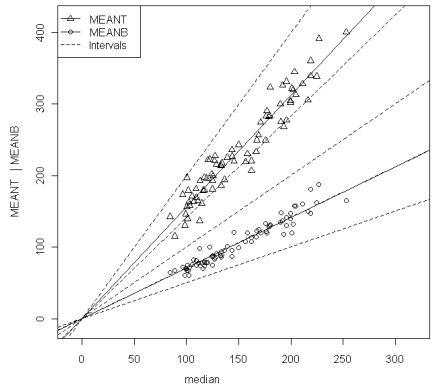

Figure 3: Graphical representation of MEAN-B and MEAN-T with respect to the median. Linear regressions corresponding to MEAN-B and MEAN-T are traced in continuous lines and the dotted lines represent, from top to bottom, the intervals +octave, +half-octave, unison, -half-octave and –octave.

As can be seen, MEAN-B regression line is indistinguishable from the –half-octave interval. We can say that the values of MEAN-B is clearly situated at one half-octave below the median while the values of MEAN-T are bounded by the octave and half-octave above the median. In other words, it seems that a speaker rarely goes higher than an octave above the median of his/her median voice, and hardly ever goes lower the half-octave below the median. These musical intervals could consequently be used as default min and max values for pitch floor and ceiling in measuring pitch. It also suggests that log2(Hz/median) is a transformation which provides a natural scale for the representation of pitch.

4. ADoReVA: a Praat plugin for the automatic detection of register changes

ADoReVA [20, 21] is a clustering algorithm developed for the automatic detection of register changes. Freely available, it may be installed in the Praat Preference folder so that the user may run it from the Praat menus. The algorithm includes 4 main steps: first, it calculates the register key and range, according to the A-measures defined in **3.3**, for each unit of speech (previously annotated in a TextGrid object). Then, it calculates the difference in key and range between two consecutive units. Next, after obtaining consecutive differences between units, the clustering algorithm groups the units together according to their difference in key and range. The smaller the difference between two units, the sooner these units are clustered together. The output generated by the algorithm is a binary tree structure in the form of a layered icicle diagram (*Figure 4*).

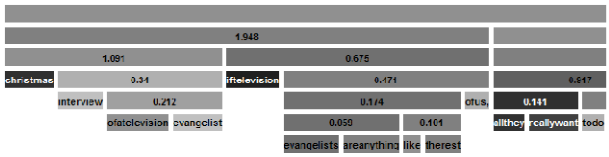

*Figure 4: Extract of a layered icicle diagram representation as obtained from the algorithm. The representation suggests that units "*Christmas*" to "*evangelist*" belong to the same group and are separated from the group of units starting with "*if television*". In fact, the distance between the nodes "*evangelist*" and "*if television*" indicates the presence of a strong boundary. A colour scale indicates register key for each unit. The darker the colour, the higher the key.*

Finally, once groups of units are distinguished, the algorithm calculates the distance between the leaf nodes according to key and range parameters. This allows boundary strength measurements between units, and hence the detection of register changes. Indeed, the larger the distance, the stronger the boundary between two groups. On the contrary, a short distance suggests that two consecutive units belong to the same group of units, in terms of register.

5. Integrating register changes with the INTSINT algorithm

The calculation of the node distances and a visual inspection of register changes from the binary tree representation as obtained with ADoReVA allowed us to set different thresholds for which the automatic coding of tonal targets in INTSINT might be improved. Delimiting the temporal span of register variations from the thresholds 0.5, 1, 2, 3, 4, 5 and 6 appeared possible choices and were tested. An algorithm was developed to extract from the table of node distances the different distances obtained between the leaf nodes or units, according to the specified threshold. If the value of the node distance is superior to the given threshold, then the algorithm stops running through the table and the MOMEL-INTSINT algorithm [22] is applied from the first unit to the one indicated with a leaf node superior to the threshold. The process is repeated until the end of the table and thereby until the end of the created PitchTier object. The output generated by the MOMEL-INTSINT algorithm is a TextGrid object containing 3 main tiers (*Figure 5*).

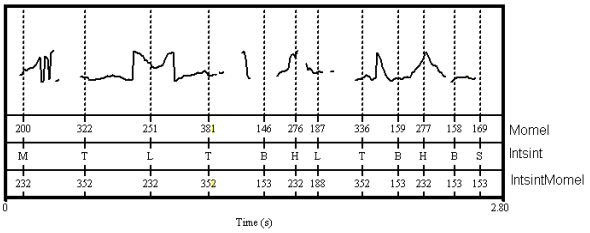

Figure 5: Example of a TextGrid object obtained with the MOMEL-INTSINT algorithm. The first tier Momel indicates the value in Hz for each tonal targets; the second tier Intsint gives the INTSINT coding obtained for each targets; the third tier InstintMomel indicates the recalculated value of the tonal targets according to the INTSINT coding.

In order to evaluate which threshold allows the improvement of the INTSINT coding, the degree of correlation between the

values Momel (from the first tier) and the values IntsintMomel (from the third tier) was observed. Therefore, for each threshold, a coefficient of determination is obtained and is compared to the one calculated when no register changes are taken into account (WRC). *Table 2* indicates the different R^2 obtained for each threshold and WRC for the four corpora.

Corpora	0.5	1	2	3	4	5	6	WRC
PFC	0.97	**0.97**	0.96	0.96	0.96	0.96	0.95	**0.94**
PAC	0.98	**0.98**	0.97	0.97	0.97	0.97	0.97	**0.96**
CID	0.97	**0.97**	0.96	0.96	0.95	0.95	0.95	**0.94**
AM	0.98	**0.98**	0.98	0.97	0.97	0.97	0.97	**0.96**

Table 2: Coefficients of determination between Momel and IntsintMomel values obtained according to thresholds 0.5, 1, 2, 3, 4, 5 and 6 and without taking into account register changes (WRC) for the four corpora.

As can be seen, integrating register changes improve the coding of the INTSINT system since all the R^2 obtained with thresholds are superior to the one obtained WRC. The threshold which obtains the larger R^2 is threshold 1.
More specifically, the implementation of register changes into intonation systems becomes clearer when we look at the correlations between Momel and IntsintMomel values only for T targets on the one hand, and for B targets on the other hand. *Table 3* indicates the R^2 obtained when integrating register changes with a threshold of 1 and without taking register changes into account (WRC).

	R^2 for T targets		R^2 for B targets	
Corpora	Threshold1	WRC	Threshold1	WRC
PFC	0.95	0.91	0.92	0.86
PAC	0.96	0.9	0.95	0.85
CID	0.92	0.83	0.9	0.76
AM	0.96	0.9	0.97	0.9

Table 3: Coefficient of determinations (R^2) between Momel and IntsintMomel values obtained according to threshold 1 and without taking into account register changes (WRC) for T targets and B targets, using the four corpora.

It can be seen that the R^2 becomes larger for a threshold of 1 and that the improvement in the coding of targets is greater for B targets than it is for T targets. This can be explained by the fact that R^2 are already large WRC for T targets.

6. Discussion & Conclusion

The issue of the temporal span of register changes is far from being solved. The difficulty of such an issue lies in the fact that register changes depend on the hierarchical and organisational structure of discourse, and on the intentions and the mood of the speaker. It is difficult, if not impossible, to define *a single* domain over which register changes operate. That is why, we propose to detect register changes automatically, through the development of a clustering algorithm, ADoReVA. It has then been shown that integrating register changes with the INTSINT system greatly improves the coding of tonal targets, something which might also be the case with other intonation systems such as ToBI.
The elaboration of an algorithm detecting register changes implied, upstream, defining a reliable measurement of register. We have shown that the dichotomy Acoustic vs. Linguistic measurements was artificial and that using "controlled" Acoustic measures allows the measurement of register and its changes. Besides, A-measures, contrary to L-measures allow the treatment of a large amount of data since they are obtained automatically. A-measures also let us to propose log2(Hz/median) as a new natural scale for pitch, based on the octave interval with respect to the median pitch.

7. References

[1] Bolinger, D. 1951. Intonation: levels versus configurations. Word Vol 7 N3 Part1, 199-211.

[2] Trager, G.L. & Smith, H.L. 1951. An Outline of English Structure (Battenburg Press; Norman, Oklahoma).

[3] Pierrehumbert, JB. 1980. The Phonology and Phonetics of English Intonation. Doctoral thesis, MIT Cambridge MA.

[4] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J. 1992. ToBI : a Standard for Labelling English Prosody. Proceedings ICSLP92 (2) 867- 870, Banff, Canada.

[5] Ladd, D.R. 1997. Intonational Phonology. Cambridge University Press, Cambridge

[6] Hirst, D.J. 1998. Intonation in British English in Hirst & Di Cristo (eds) 1998.

[7] Gussenhoven, C. 2004. The phonology of tone and intonation. Cambridge University Press, Cambridge.

[8] Jun, Sun-Ah (ed.). 2006. Prosodic Typology and Transcription: A Unified Approach. Oxford University Press.

[9] Hirst, D.J. & Di Cristo, A. 1998. A survey of intonation systems. in Hirst & Di Cristo (eds) 1998.

[10] Hirst, D.J., Di Cristo, A. & Espesser, R. 2000. Levels of representation and levels of analysis for intonation. in M. Horne (ed) Prosody : Theory and Experiment. Kluwer Academic Publishers, Dordrecht. 51-87.

[11] Hirst, D.J. 2005. Form and function in the representation of speech prosody. Speech Communication 46 (3-4), 334-347.

[12] Patterson, D & Ladd, R. Pitch Range Modelling: Linguistic dimensions of variation in ICPHS99. 1169-72., 1999.

[13] Patterson, D. A linguistic approach to pitch range modeling. PhD dissertation, University of Edinburgh, 2000.

[14] Auran, C., Bouzon, C. & Hirst, D. The Aix-MARSEC Project: An Evolutive Database of Spoken British English, in Speech Prosody 2004, ISCA.

[15] Carr, P. Durand, J. La Phonologie de l'Anglais Contemporain : usages, variétés et structure : The Phonology of Contemporary English: usages, varieties and structure., 2003.

[16] Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B. and Rauzy, S. Le CID-Corpus of Interactional Data-: protocoles, conventions, annotations In Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix en Provence, 25, 25-55., 2007.

[17] Delais-Roussarie, E. & Durand, J. Corpus et variation en phonologie du français: méthodes et analyses, Presses Universitaires du Mirail., 2003.

[18] Boersma, P. & Weenink, D. 2007. Praat: doing phonetics by computer (Version 4.6.35) [Computer program]. Downloadable from http://www.praat.org/.

[19] De Looze,C. and Hirst, DJ., Detecting changes in key and range for the automatic modelling and coding of intonation., In Speech Prosody 2008, Campinas, Brazil.

[20] De Looze, C. and Rauzy, S. Automatic Detection and Prediction of Topic Changes Through Automatic Detection of Register variations and Pause Duration. In INTERSPEECH 2009, Brighton, England.

[21] De Looze, C. In progress. Analyse et interprétation de l'empan temporel des variations prosodiques en français et en anglais contemporain. Doctoral thesis, Université de Provence.

[22] Hirst, D.J., "A Praat Plugin for MOMEL and INTSINT with improved algorithms for modelling and coding of intonation.", In *Proc*. Int. Conf. Phonetic Sci. XVI, Saarbrucken., 2007.