Detecting changes in key and range for the automatic modelling and coding of intonation

Céline De Looze and Daniel Hirst

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

Abstract

The analysis of authentic speech, unlike that of laboratory speech, needs to take into account the fact that the fundamental frequency patterns corresponding to the intonation of utterances can be of two types - local pitch characteristics determined by the surface phonological representation of the intonation and longer term characteristics corresponding to less well understood changes in pitch *key* and *range*. In this paper a number of acoustic correlates of changes in pitch *key* and *range* are examined and compared to subjective annotations and a preliminary attempt is made to estimate these changes automatically.

1. Introduction

Bolinger pointed out many years ago [4] that a major drawback of most scalar systems for the annotation of intonation patterns was the difficulty in separating local pitch changes corresponding to a phonological distinction from more global changes of register, which tend to be related to more general discourse or extralinguistic factors. Specifically, he argued that in a system like that of Trager & Smith [29], with four distinct pitch levels, it would not be possible to distinguish a high falling pitch movement /41/ in a narrow pitch range from a low falling movement /31/ or /21/ in an neutral or expanded pitch range. Since a speaker can also modify the overall mean pitch of his voice, his key, it would also not be possible to distinguish, for example, an isolated /43/ from /32/ or /21/ or /42/ from /31/ unless there was some independent way of establishing the speaker's *key*.

This argument was later taken up by Janet Pierrehumbert in her influential study of American English intonation [26] and the assumption that there is only a binary phonological distinction between H and L tones became a cornerstone of both ToBI [28] as well as many variant models within the general Autosegmental-Metrical framework of Intonational Phonology (eg [20], [11], [7] [17]).

It is, of course, obvious that on the phonetic or surface phonological level, more than two distinctive levels need to be distinguished. In INTSINT ([13], [14], [15]), for example, an explicit multilingual system for the representation of surface phonological contrasts for intonation, high and low can be interpreted either globally with respect to the current pitch range as T(op) or B(ottom), or locally with respect to the previous tone as H(igher) or L(ower). A second type of relative tone, often occurring in an iterative sequence, is defined by a smaller relative pitch change which is either U(pstepped) or D(ownstepped). Together with the possibility of an intermediate initial pitch M(id) and an intermediate relative pitch S(ame), this makes a total of 8 possible tonal values for each significant point in a given pitch pattern.

The fact that an isolated falling pitch pattern can be analysed in this framework as either /xD/ /xL/ or /xB/ (where the value of x can be any of the 8 possible tones, depending on what precedes it) means that Bolinger's original argument against multiple scalar values needs to be answered here, too.

The implementation of INTSINT as an automatic algorithm, coding the output of the Momel pitch modelling algorithm [15], [16], makes the crucial assumption that, in the portion of speech analysed, the speaker's *key* and *range* remain constant. The algorithm estimates both the optimal sequence of tonal symbols to code the Momel target points and the two optimal values of *key* and *range* which together, with the value of the preceding target point, are sufficient to derive a phonetic interpretation of the tonal targets.

This algorithm has proved fairly satisfactory for the analysis of laboratory speech or fairly neutral read speech. It is obvious, however, that the analysis of more authentic speech cannot make assume that the speaker's *key* and *range* remain unchanged. It is thus crucial to be able to identify changes in *key* and *range* before applying the coding algorithm.

2. Experiments

2.1. Material

The English data used were taken from the AIX-MARSEC database [2], [18] and the French data from the PFC, Phonologie du Français Contemporain [10].

The Aix-Marsec Corpus

The recordings correspond to commentaries, new broadcasts, lectures, religious broadcasts, magazine-style reporting, fiction, poetry, dialogues and propaganda, and are mainly prepared monologues. We selected from this corpus 57 sound files, a total of 2 hours 30 minutes of recordings.

The PFC

10 speakers were selected. Six female and four male speakers of French, all from Marseille aged between 17 and 73, were asked (i) to read an article from a newspaper and (ii) to talk about their hobbies, job, plans or childhood. 10 sound files of reading, lasting about two minutes each and 10 guided conversations lasting about 3 minutes each were analysed, a total of 50 minutes of recording.

2.2 Detecting pitch minimum and maximum automatically

First, to detect intra-speaker variations in register automatically, a pitch scale, ie, detecting minimum and maximum pitch for each speaker, has to be determined. Minimum and maximum values can be obtained automatically in Praat by creating a Pitch object with the pitch floor and ceiling taking the default values of 75 and 600Hz.

Because fo detection is very sensitive to microprosodic effects and octave errors, however, results tend to be very error-prone. For this reason, when analysing fo, Praat users are usually advised to first determine the pitch floor and ceiling manually, before creating the Pitch object. Since manual interventions are very time-consuming and also very user-dependent and error-prone, we have proposed an algorithm for the automatic estimation of these values, which was implemented in the Momel-Intsint Praat plugin [16], [3].

2.2.1 Defining pitch ceiling and floor for the estimation of fo extrema

AIX MARSEC DATA BASE

The algorithm was determined in an empirical experiment [8] using the Aix-Marsec data.

The first author annotated manually the minimum and maximum values for each of the 59 recordings (MinRef/MaxRef).

Next, a script was used to obtain fo quantiles, with pitch floor and ceiling taking the default values of 60 and 750Hz.

A comparative study of the different quantiles (from q05 to q95) showed that q05 to q25 and q75 to q95 were best correlated with manual estimations of maximum and minimum pitch respectively. However, mean difference and standard deviation were too high. Various formulae were tested to predict optimal values of pitch floor and pitch ceiling (figure 1) and a comparison was made of the output min and max of the different scripts (figure 2).

Floor	Q05	Q10	Q15	Q20	Q25	Q25
	0.89	0.81	0.78	0.75	0.73	0.75
Ceiling	Q95	Q90	Q85	Q80	Q75	Q75
	1.39	1.55	1.65	1.74	1.82	1.75

Figure1. Formulae tested as pitch floor and ceiling

From these analyses we concluded that the formulae:

ceiling = 1.5*q75 floor = 0.75*q25

gave the best estimation of fo extrema (figure 2).

Floor(%)			Ceiling(%)		
Qi	mean	sd	Qi	mean	sd
Q05*0.89	-4.59	10.47	Q95*1.39	6.57	29.86
Q10*0.81	-5.73	9.28	Q90*1.55	6.21	27.72
Q15*0.78	-4.87	9.75	Q85*1.65	3.97	15.23
Q20*0.75	-4.41	10.15	Q80*1.74	4.00	15.45
Q25*0.73	-3.77	10.72	Q75*1.82	4.14	16.20
Q25*0.75	**-1.80**	12.65	Q75*1.75	-13.83	13.34

Figure 2. MinRef/MinMo(Qi) mean difference and standard deviation obtained with 12 different floors and ceilings.

Comparing the detection of MinRef and MaxRef of the 59 sound files and serving as reference, with min and max values obtained with the formulae integrated in the Momel-Intsint Praat plugin (MinMo/ MaxMo) and the min and max values obtained with Praat (Min Praat/ MaxPraat) (i.e. by using the default floor and ceiling and then the functions Get minimum... and Get maximum...), it appears that the algorithm greatly improves the detection of minimum and maximum fo values in Praat. For the 59 sound files, the MinRef/MinMo mean difference and standard deviation are -1,80% and 12,65% respectively against a MinRef/MinPraat mean difference of -13,83% and a standard deviation of 13,34% (Figure 3). The MaxRef/MaxMo mean difference and standard deviation are 4.14% and 16.20% respectively against a MaxRef/MaxPraat mean difference and standard deviation of 89,75% and 49,89% (Figure 4).

	MinRef/MinMo		MinRef/MinPraat	
	Mean %	Sd %	Mean %	Sd %
Total	-1,80	12,65	-13,83	13,34

Figure 3. Mean difference and standard deviation between MinRef and MinMo on the one hand and MinRef and MinPraat on the other hand.

	MaxRef/MaxMo		MaxRef/MaxPraat	
	Mean %	Sd %	Mean %	Sd %
Total	4.14	16.20	89,75	49,89

Figure 4. Mean difference and standard deviation between MinRef and MinMo on the one hand and MinRef and MinPraat on the other hand.

PFC DATABASE

To check whether these results were corpus dependent, we decided to run the script on our corpus: the PFC.

Min and max values were firstly annotated manually (MinRef/MaxRef). Then, MinRef and MaxRef were compared to the MinMo and MaxMo and to the Min Praat and MaxPraat of the 20 sound files. The results showed again that the algorithm greatly improves the detection of these extrema in Praat (Figure 5 and Figure 6).

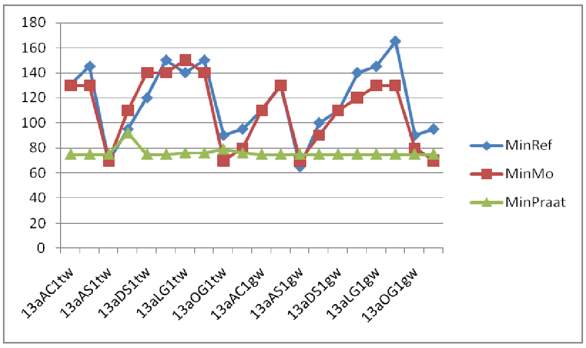

Figure 5. *Graphic representation of MinRef, MinMo and MinPraat (Hz).*

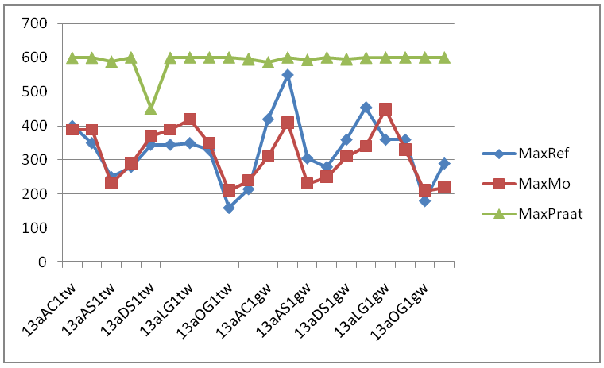

Figure 6. *Graphic representation of MaxRef, MaxMo and MaxPraat (Hz).*

For the 20 sound files, the MinRef/MinMo mean difference and standard deviation are -5,38% and 11,86% respectively against a MinRef/MinPraat mean difference of -30,19% and a standard deviation of 20,24% (Figure 7). The MaxRef/MaxMo mean difference and standard deviation are -1,16% and 18,35% respectively against a MaxRef/MaxPraat mean difference and standard deviation of 95,41% and 67,15% (Figure 8).

The algorithm proposed seems more robust to detect the minima than the maxima (-2,21% and -8,56% for the minima vs 9,37 and -11,69 for the maxima) and gives better results for read speech than dialogue speech.

	MinRef/MinMo		MinRef/MinPraat	
	Mean%	Sd%	Mean%	Sd%
Total	-5,38	11,86	-30,19	20,24
Reading	-2,21	12,77	-30,14	21,29
Conversation	-8,56	10,56%	-30,25	20,29

Figure 7. *Mean difference and standard deviation between MinRef and MinMo on the one hand and MinRef and MinPraat on the other hand.*

	MaxRef/MaxMo		MaxRef/MaxPraat	
	Mean	Sd	Mean	Sd
Total	-1,16	18,35	95,41	67,15
Reading	9,37	11,09	108,02	72,56
Conversation	-11,69	18,48	82,81	62,46

Figure 8. *Mean difference and standard deviation between MaxRef and MaxMo on the one hand and MaxRef and MaxPraat on the other hand.*

We are currently working on an evaluation of the algorithm using more extensive data in order to check the validity of these preliminary results and in particular to improve the detection of maximum values.

2.2.2 *Defining speakers' register*

The Momel-Intsint Praat plugin provides an estimation of the register or a recording: its bottom line and top line (MinMo and MaxMo), its level (Key) and span (Range). MinMo, MaxMo and Key are given in a linear scale (Hertz), range in a logarithmic scale (octaves). These four parameters allow the comparison of different speakers' register. As expected, the average *key* for male speakers is lower than that for female speakers, whereas *range* is about the same, slightly greater for female speakers of French (Figures 9&10). Reading productions show higher *key* but narrower *range* compared to conversation productions (Figure 11).

Speakers	Min	Max	Key	Range
Female	125	412	202	1.2
Male	78	272	132	1.28

Figure 9. Average fo extrema, key and range extracted from the Aix-Marsec Corpus.

Speakers	Min	Max	Key	Range
Female	136	385	213	1.28
Male	87	245	129	1.14

Figure 10. Average fo extrema, key and range extracted from the PFC corpus.

Task	Min	Max	Key	Range
Reading	118	302	190	1.1
Conversing	115	356	168	1.35

Figure 11. Average fo extrema, key and range extracted from the PFC corpus.

2.3 Detecting variations in pitch *range* and *key* automatically

Once the register had been estimated automatically, we used two methods to detect intra-speaker variations, ie variations in pitch *range* and *key*.

Material

We used the PFC sound files for these experiments. As we were interested in intra-speaker variations, each file was split into inter-pausal units separated by an silent pause of at least 250 ms [22], under the assumption that *key* and *range* are more likely to vary *between* interpausal units rather than *within* them. Final lengthenings and linking hesitations inside units were kept. Non-linguistic sounds such as laughs and coughs were excluded from the analysis.

First Method

First, a manual annotation of variations in *key* and *range* was made. Variations were annotated in interpausal units. The labels used were N(arrower), W(ider) or S(ame) for *range* and H(igher), L(ower) or S(ame) for *key*, these categories being those generally used in perception experiments.

Then, for each unit an estimation of its *key* and *range* was made, using the Momel-Intsint algorithm. A script was run to obtain the absolute difference in *key* and *range* between a given unit (n) and the preceding one (n-1). Different thresholds were used in the script in order to evaluate the best match with manually annotated variations in *range* and *key*.

Setting the threshold at 0,15 allows the detection of 56% and 57% of variations and non-variations in *key* for the readings and 62% and 57% for the guided conversations. A threshold at 0,33 allows the detection of 51% and 49% of variations and non-variations in *range* for the readings and 54% and 48% for the guided conversations.

Second method

In a second experiment, we proposed to recursively reduce the Euclidian distance between units in a space defined by *key* and *range* parameters. First, values of *key,* given in a logarithmic scale, and values of *range* were normalized. Then, the script was run and gave an output in the form of a binary tree.

Results from the first experiment are of course very preliminary and not robust enough to say that variations in *range* and *key* can be accurately detected automatically. The second experiment is at present only in the stage of testing, however, we are confident that these two methods are the first steps towards an automatic detection of intra-speaker variations in register and towards a better analysis of linguistic, extralinguistic and paralinguistic characteristics of speech.

3. Discussion

It is nowadays agreed that long term variations, such as variations in register, tempo and intensity, have to be taken into account for the study of shorter term variations, such as tone, quantity and loudness for the improvement of TTS systems. Notably, in recent years, different investigations have been carried out for the estimation and modelling of register (*range* and *key*) ([17], [25], [22], [23], [5], [6], [8], [12], [9], [19], [21], [24], [27], [30], [1]). In this presentation we propose (i) formulae to obtain the minimum and maximum values of a speaker's register automatically, (ii) two thresholds to detect automatically intra-speaker variations in *key* and *range*, and (iii) a method to obtain a binary tree as representative of register variations. There is, of course, still a considerable amount of research required to improve in particular the automatic detection of variations in *range* and *key* since at the moment only half of the variations are detected correctly automatically.

4. Conclusions

This study showed that it is possible (i) to detect fo min and max values automatically with an algorithm implemented in the Momel-Intsint Praat plugin [16] and (ii) to detect variations in register (*range* and *key*) automatically with an algorithm setting a threshold at 0,15 and 0,33 for variations in *key* and *range* automatically. Further work is obviously needed to assess the validity of such a threshold. The detection of such long term variations will allow a better analysis of the functions of short term and long term variations of pitch which will of course have immediate applications for the analysis of discourse, and extralinguistic, or paralinguistic characteristics of authentic speech.

5. References

[1] Arons, B. 1994 . Pitch-based Emphasis Detection for Segmenting Speech Recordings, dans Proceedings of international Conference on Spoken Language Processing (September 18-22 Yokohama, Japan), vol. 4, 1931-1934

[2] Auran, C., Bouzon, C. & Hirst, DJ. 2004. The AIX-MARSEC Project: an evolutive database of spoken British English in Speech Prosody, March 23-26, Nara, Japan, 561-564.

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

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

[5] Brazil, D., Coulthard, M. & Johns, C. 1980. Discourse Intonational and language Teaching: Longman 1-82

[6] Clark, RAJ. 1999. Using prosodic structure to improve pitch range variation in text to speech synthesis in ICPhS99

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

[8] 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.

[9] Dilley, LC. 2007. Pitch range variation in english tonal contrasts : continuous or categorical ? in ICPHs XVI . Saarbrücken, 6-10 August 2007. 1153-6

[10] Durand, J.; Laks, B.; & Lyche, C. 2002. La phonologie du français contemporain: usages, variétés et structure. In: C. Pusch & W. Raible (eds.) Romanistische Korpuslinguistik- Korpora und gesprochene Sprache/Romance Corpus Linguistics - Corpora and Spoken Language. Tübingen: Gunter Narr Verlag, pp. 93-106. http://www.projet-pfc.net/.

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

[12] Hirst & Di Cristo (eds) 1998. Intonation Systems. A survey of Twenty Languages. Cambridge University Press, Cambridge.

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

[14] 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.

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

[16] Hirst, D.J. 2007. A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation. in Proceedings International Conference on Phonetic Sciences, Saarbrücken, paper 1443.

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

[18] Knowles, G. 1996. Working with Speech: Addison Wesley Longman Limited.

[19] Ladd, DR & Terken, J.1995. Modelling intra-and inter-speaker pitch range in Proceedings ICPhS 95: Stockholm Vol 2, 386-9

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

[21] Liberman, M. & Pierrehumbert, J 1984 Intonational under Changes in Pitch Range and Length in Language Sound Stucture edited by M Aronoff and RT Oehrle: MIT Press, 157-232

[22] Miller, J., Grosjean, F., Lomanto, C. 1984. Articulation rate and its variability in spontaneous speech: A reanalysis and some applications. Phonetica, 41. 215-225.

[23] Nicolas, P & Hirst, D.J. 1995. Symbolic coding of higher-level characteristics of fundamental frequency curves in ESCA Eurospeech'95. 4th European Conference on Speech Communication and Technology. Madrid, September 1995. ISSN 1018-4074

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

[25] Patterson, D.2000. A linguistic approcha to pitch range modelling: Doctoral thesis.

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

[27] Shriberg, E., Ladd, DR., Terken, J. & Stockle, A. 1996. Modelling pitch range variation within and across speakers: predicting F0 targets when "speaking up" in ICSLP96.

[28] 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.

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

[30] Xu, Y., Xu, C. & Sun, X. 2004. On the Temporal Domain of Focus in Speech Prosody 2004. 81-84.