A quantitative investigation of the prosody of *Verum Focus* in Italian

Giuseppina Turco[1], *Michele Gubian*[2], *Jessamyn Schertz*[2]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[2] Centre for Language & Speech Technology, Radboud University, Nijmegen, The Netherlands
giusy.turco@mpi.nl, m.gubian@let.ru.nl, jschertz@u.arizona.edu

Abstract

In this study we present a preliminary investigation of the prosodic marking of Verum focus (VF) in Italian, which is said to be realized with a pitch accent on the finite verb (e.g. A: *Paul has not eaten the banana* - B: *(No), Paul HAS eaten the banana!*). We tried to discover whether and how Italian speakers prosodically mark VF when producing full-fledged sentences using a semi-spontaneous production experiment on 27 speakers. Speech rate and *f0* contours were extracted using automatic data processing tools and were subsequently analysed using Functional Data Analysis (FDA), which allowed for automatic visualization of patterns in the contour shapes. Our results show that the postfocal region of VF sentences exhibit faster speech rate and lower *f0* compared to non-VF cases. However, an expected consistent difference of *f0* effect on the focal region of the VF sentence was not found in this analysis.

Index Terms: Italian *Verum focus*, prosody, functional data analysis, principal component analysis, forced alignment

1. Introduction

Verum focus (VF) [1] is generally described as a phenomenon by which a pitch accent is realized on the finite verb, i.e. *Paul HAS eaten the banana* as a contrastive answer to a contextually preceding negative sentence *Paul has not eaten the banana*. In this way, as [1] noticed, the pitch accent on the finite verb is used to emphasize the positive polarity of the proposition (thus the assertion part of the sentence), rather than the semantic content of a particular constituent (e.g. A: *John has lost his key* B: *(No), John has BROKEN his key*).

In a recent cross-linguistic investigation on Germanic and Romance languages [2], it was found that in German and Dutch such a focus is expressed by using intonation (VF) or special particles expressing a similar assertion-contrast function (i.e. the Dutch particle *wel,* roughly meaning 'indeed' in English). French and Italian, on the contrary, were not shown to have equivalent linguistic means such as a 'scope' specifically affecting the assertion part of the sentence (i.e. Verum focus marking was never found to be expressed by Romance speakers in the expected contexts). This hypothesis was further investigated in an experimental study on prosodic marking of VF in German and French [3] where it was found that, in contrast to French speakers, Germans systematically marked VF by realising a nuclear falling accent (H*L-) on the auxiliary verb accompanied by post-focal deaccentuation. In the Romance linguistic literature [4], the phenomenon of VF is mainly addressed from a syntactic viewpoint (i.e. the Spanish particle *sí que* 'yes') since word order is regarded as the most optimal linguistic strategy conveying such type of focus, unless a clear marked intonation is used by speakers. This experimental study tries to address the question of whether Italian speakers mark VF when speakers are encouraged to produce full-fledged sentences and, if so, how they mark it. It represents a preliminary prosodic investigation on this phenomenon which has so far received very little attention in the prosodic literature.

The analysis of data was carried out relying entirely on automatic data processing tools. Two quantitative features were studied, *f0* and relative speech rate, where the latter was inferred by an automatic phonetic alignment performed with an automatic speech recogniser (ASR). Speech rate and *f0* contours were jointly analysed applying an advanced technique called functional data analysis (FDA). FDA allowed us to automatically visualise and assess the presence of weak yet consistent patterns in the contour shapes of the whole sentence. FDA does not require specifying in advance what type of pattern (e.g. a peak shift) nor in which part of the sentence we expect something to emerge.

2. Methods

1.1. A semi-controlled production experiment

A picture-difference task in the form of a mini-dialogue between two speakers (a confederate speaker and the participant) was designed for the semi-spontaneous elicitation of full–fledged sentences in two conditions: the Verum Focus (VF) condition and the non-Verum Focus condition (non-VF) [5]. The experimental set-up for the VF target condition is based on three pictures: a *Baseline* picture, accessible to both speakers, in which a certain situation is illustrated (e.g. *a dog eating a bone*); a *Negation* picture where the opposite situation is depicted (*the dog is not eating the bone*), only accessible to the confederate; an *Affirmation* picture that is similar to the *Baseline* picture (*the dog is eating the bone*), only accessible to the participant. The confederate speaker (a female native speaker from Rome) described the difference between her (*Negation*) picture and the *Baseline* picture, whereas the participant had to describe his (*Affirmation*) picture in relation to the confederate speaker's preceding negative description. The confederate speaker provided the participant with a negation statement (1.a below) creating the conditions for VF to appear (1.b). In the non-VF condition, the relations between the three pictures do not provide the conditions for VF contrast. In this case, the confederate described her picture with no use of negation (2.a). An example is reported below.

1. VF condition:
    [baseline picture showing "a dog eating a bone"]
(a) confederate with negation picture: *In my picture the dog is* <u>*not*</u> *eating the bone*
(b) participant with affirmation picture: *In my picture the dog* $[is]_{focus}$ *eating the bone*

2. non-VF condition:
    [baseline picture showing "a man" only]
(a) confederate with description picture: *In my picture the man is drinking a beer*

(b) participant with description picture: *In my picture the man [is eating an apple]$_{focus}$*

By providing context utterances with and without negation (respectively, in the VF and in the non-VF condition), this interactive task allowed us to elicit semi-spontaneous comparable sentence pairs and to test in which respects the analysed prosodic features were specific to the structures elicited in the VF condition.

1.2. Material and Participants

Twenty-seven Italian native speakers (male = 8; female = 19, average age = 23.7, SD = 3.6) were recorded in a quiet room at the University "La Sapienza" in Rome. Each subject underwent a session comprising 110 trials that lasted approximately 20 minutes. There were 30 VF condition trials, 30 non-VF condition trials, and 50 filler trials. Out of the 30 VF trials, 12 pictures illustrated simple-present actions (focus on lexical verb items, LEX); 12 pictures illustrated present-perfect actions (focus on auxiliary items, AUX), and eight trials used emotional state pictures (focus on copula items, COP). The 30 non-VF trials contained the same target verbs as the corresponding VF ones. The 50 fillers had the focus on other parts of the sentence. All the elicited sentences were S(ubject)-V(erb)-O(bject) with non-branching S/O. Although for Germanic languages it is claimed that only the auxiliary verb is affected by the focus construction, we wanted to capture effects that might extend beyond the very short auxiliary verb.

In non-VF trials the description of the event depicted on the picture was more up to participants' interpretation than in the VF trials, where only a change in the VF domain had to be expressed (participant: X *HAS done* Y) with respect to the confederate's sentence (ex. confederate: X *has not done* Y). For this reason, many non-VF trials did not contain the target word (i.e. the verb phrase) as in the VF cases. Since the techniques used in this work require the rhythmic structures of the sentences to be as similar as possible, we were forced to exclude those trials from the analysis, which created an unwanted imbalance in the data-set.

1.3. Data Pre-processing: Forced Alignment and *f0* extraction

In order to obtain detailed and consistent landmarks for our analysis, we created a broad phonetic transcription from the orthographically transcribed utterances based on a lexicon of canonical pronunciations. We then used the Hidden Markov Model Toolkit HTK [6] to automatically align the transcription with the speech signal. The utterances were used to train 3-state monophone acoustic models for each of the 35 phones in the transcription (diphthongs such as /je/ or /wo/ were considered as independent phones), while simultaneously aligning them with the speech signal using a bootstrapping process consisting of a train-test-evaluate-retrain loop. Since the 3-state models were trained using a 10 ms frame shift, the minimum length for each phone was 30 ms. This resulted in a phone-level transcription of all of the utterances in the data set which was used both as the basis for extracting speech rate and for creating landmarks for FDA analysis (Section 1.4).

We then extracted pitch contours for the phone-aligned utterances using the pitch algorithm in the Praat toolkit [7]. We automatically extracted the contours using a default range of 70-350 Hz for males and 100-500 Hz for females, then adjusted these ranges for specific speakers in order to minimize obvious errors such as octave jumps in the contours.

1.4. Functional Data Analysis

As said before, this work is entirely based on automatic data analysis procedures. In this section we briefly introduce the reader to Functional Data Analysis (FDA), the technique that allowed us to study *f0* and relative speech rate contours. FDA is a family of statistical tools that extend well-known tools so that the input elements consist of curves or trajectories, as opposed to vectors of numbers [8]. For example, it is possible to carry out (functional) Principal Component Analysis (FPCA) or to apply a (functional) linear model directly to a set of *f0* contours, as opposed to (manually) extracting a fixed number of descriptive features from them (e.g. peak and valley coordinates) and then applying ordinary multivariate statistics on those features.

All FDA tools represent sampled contours, like *f0* sequences measured with Praat, in the form of continuous functions of time. This is achieved by applying standard smoothing techniques [8]. The FDA statistical machinery is based on performing comparisons among the values of the input functions at every instant in time. It means that what happens at, say, 0.3s from the beginning of each signal is compared across all signals and used to infer stable trends. Since we know that even the production of the same sentence by the same speaker is not synchronised across repetitions, we need a way to adjust for the misalignment of syllables, phones or any unit that we deem as containing information that repeats across a given set of utterances. If we did not do that, random alignment of mismatched units would blur the final analysis. To avoid this, all contours are warped in order to synchronise similar events, like syllable boundaries, across the dataset. This operation, called *landmark registration,* is carried out automatically once the time location of landmarks is known [8]. In our case, landmarks were taken from the phone-level transcription provided by the automatic segmenter (Sec. 1.3). We made use of different sets of landmarks in our analyses; for example, one set of landmarks consisted of all stressed syllables in the sentence.

While applying FDA to *f0* contours is a relatively straightforward operation, integrating the analysis of speech rate requires a further step. Relative speech rate is encoded in the differences in duration of each interval between landmarks across utterances. Since those differences are eliminated by landmark registration, we need to 'save' them and introduce them in the analysis in a form suitable for FDA. The second author proposed in [9] a way to overcome this problem, which will be applied here. Landmark registration operates for each curve a warping of the original time axis to a new 'registered' or 'normalised' time axis, where each landmark occurs at the same instant for all curves, usually corresponding to its average location in the dataset. This warping is represented by a function *h(t)* that contains information about relative speech rate, because it records all the local rate variations necessary to make landmark positions coincide with their reference (average) position. Since FDA allows for the analysis of multidimensional trajectories, we will composed each input element by coupling a (registered) *f0* contour *f(t)* with its corresponding relative speech rate *r(t),* the latter being a convenient transformation of *h(t)* (see [9] for details). By doing so, the information about speech rate was preserved in the analysis.

In this work we applied functional Principal Component Analysis (FPCA) to the set of *(f(t),r(t))* function pairs described above. Given a set of (two-dimensional) curves, FPCA extracts the main independent shape variations (*principal components*, PCs) found across the curve set. Those variations are expressed in terms of alterations of the shape of the average curve, i.e. the curve $(mean_f(t),mean_r(t))$ obtained by averaging all input curves at every instant *t* (the solid curves in Fig. 2). Each input curve is assigned a set of so-called *PC scores*, one for each PC. PC scores are numbers that quantify the deviance from the mean for a specific curve (more details can be found in [9] [10]).

We chose to apply FPCA because, like ordinary PCA, it allows us to remain as theory-neutral as possible. The reason is that PCs are computed on a set of curves *without* making use of the labeling information, which in our case is the binary condition VF vs. non-VF. Only after PCs are found and each input curve is described in terms of PC scores, the results are matched with the labels in order to find correlations between 'natural' dynamic trends in the signals and the linguistic category at study.

As a final remark, we had to accommodate for the fact that the *f0* signal is absent in silences and voiceless phones. Since FDA requires continuous input functions *f(t)*, smoothing was carried out by connecting missing *f0* intervals with smooth flat lines.

2. Results

In a preliminary phase, FPCA was applied separately to each of the three sentence types, namely AUX, LEX and COP (see Sec. 1.2). Since the syllabic structure varied across sentences but the lexical stress count does not, we used the latter to locate landmarks for time registration of contours (Sec. 1.4). This analysis did not reveal any convincing pattern that we could safely associate with the VF condition. We imputed this to the random variation of both *f0* and speech rate introduced by the differences in syllabic structure. We resorted to analysing a selection of the data, concentrating only on subsets with exactly the same syllabic count and lexical stress position. This was possible only for AUX and COP. In this paper we focus on AUX in order to compare our data with previous findings from German [3].

We selected a subset of items in AUX containing the S(tressed) U(nstressed) syllabic sequence SU#S#USU#U#SU#, where # marks a word boundary. This corresponds to the last part of every sentence, like in "uomo ha mangiato la mela" *man has eaten the apple.* This reduced the data to 78 VF and 13 non-VF items from a total of 23 speakers. In this way, we could place 9 landmarks, one at the onset of every syllable nucleus (drawn from the automatic alignment).

FPCA results for AUX are shown in Figg. 1 and 2. In Fig. 2, solid curves show the mean of the input curves (where the PC score is zero) (see Sec. 1.4) while the '+' and '-' curves represent input curves scoring one standard deviation above or below the mean of the corresponding PC score. These graphical descriptions are matched with the PC scores scatterplot in Fig. 1, allowing us to visualize the relationship between PC scores and the corresponding curves.

PC1 describes a variation in the range of the whole *f0* contour (Fig. 2.a), which is associated with a global variation in speech rate (Fig. 2.b): for wider *f0* excursions ('-' curve) speech rate is globally slower (speech rate r(t) is logarithmic, so zero means average rate, +/- 0.7 means double/half rate). Looking at the distribution of PC1 scores in Fig. 1, we noted that no correlation is shown between this shape variation and VF condition. To confirm this, by running a *t*-test on the PC1 scores grouped by the two VF conditions we got a *p*-value of 0.48.

PC2 shows a less marked effect on *f0*, mostly concentrated in the post-focal region of the sentence (marked as 'obj', last 3-4 landmarks in Fig. 2c). At the same time, we note a marked effect on speech rate in the last word of the sentence (last 2 landmarks in Fig. 2d). Looking at Fig. 1 we notice a clear correlation with VF condition.

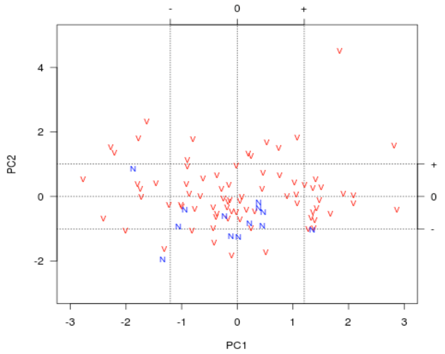
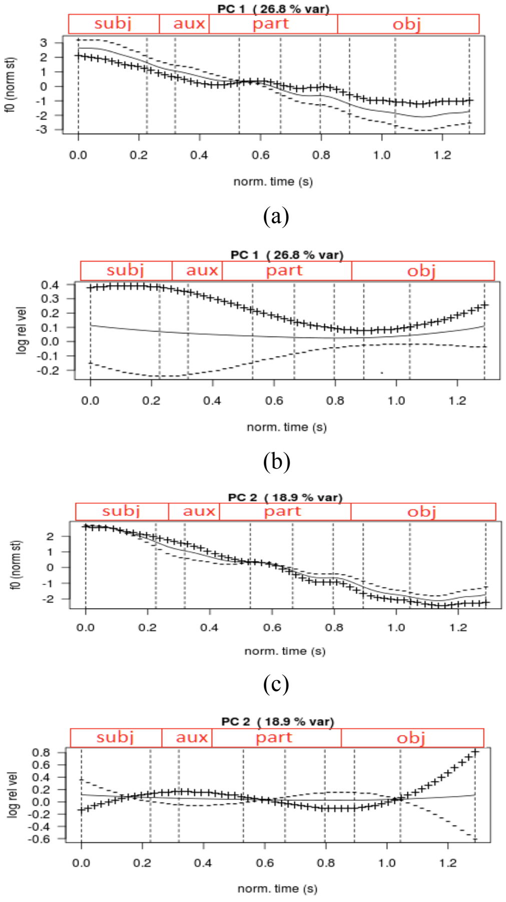


Figure 1: *Scatterplot showing PC1 and PC2 scores obtained by applying FPCA on the AUX subset. V stands for VF condition, N stands for non-VF condition. For each axis, '-', '0' and '+' levels mark PC scores corresponding to '-', solid and '+' curves in Fig. 2. PC1 axis matches with panels (a) and (b), PC2 with panels (c) and (d) of Fig. 2.*



(a)

(b)

(c)

(d)

Figure 2: *Principal components of FPCA applied to f0 and speech rate contours jointly. Panels (a,c) show f0 and panels (b,d) show speech rate. Vertical dashed lines mark the nine syllabic boundaries used to align the contours. The sentence structure (subject, auxiliary, participle, object) is also marked.*

Only VF items have PC2 scores greater than zero, whereas all non-VF items have scores less than zero. Correspondingly, only VF curves exhibit shapes resembling the '+' curves in Fig. 2c and d. By running a *t*-test on PC2 scores like we did for PC1 we obtained a p-value $<10^{-3}$, which allowed us to safely state that the two groups come from different distributions.

3. Discussion

From a first *f0* and relative speech rate inspection, we have seen that the most remarkable aspect in VF marking is a strong speech rate increase accompanied by a slightly lower *f0* affecting the post-focal region of the sentence (i.e. the syntactic object, see last 3-4 landmarks shown in Fig.2, panel (d)). As reported in other experimental works on narrow focus marking [11] [12] and on rate effects on phrasing [13], this seems to suggest that under VF condition the post-focal element undergoes a process of pitch compression and possibly of sentence phrasing reorganization [14] [15]. The pitch compression and rate increase can be explained by the fact that the syntactic object carries 'given' information in the context of our mini-dialogue (i.e. confederate*: In my picture the dog is not eating* <u>*the bone*</u> vs. participant: *In my picture the dog is eating* <u>*the bone*</u>).

As far as the focal region of the sentence (i.e. the verb) is concerned, at this stage of the analysis, we found no remarkable effect of *f0* across conditions, as shown in Fig.2 (landmarks 3 and 5 in panels (a), (c)). This suggests that VF and non-VF cases are only distinguishable via post-focal deaccentuation. This finding is quite unexpected if we consider that a) in the same experiment Germans were systematically using a falling nuclear accent on the auxiliary for marking VF with respect to non-VF sentences [3]; b) a previous FDA run on Italian read speech found a marked difference in *f0* between contrastive and non-contrastive sentences [9]. This certainly implies that future studies should take into account other potential prosodic cues to VF in Italian (e.g. presence or absence of phrase breaks, glottalization, etc.) when powerful automatic methodologies such as FDA are implemented. Moreover, a phonological analysis is currently run by the first author [16] in order to gain a full understanding of the realisation of this focus construction.

4. Conclusions

In this paper we have shown preliminary results on *Verum focus* marking in Italian by relying entirely on automatic data processing tools. This study is part of an on-going research project investigating *Verum Focus* in Romance and Germanic languages and posing cross-linguistic questions regarding the accentability status of function words (e.g. auxiliary and copula) vs. content words (full lexical verbs). Results show consistent effects of speech rate and *f0* in the post-focal region, and are only partly in line with other findings based on impressionistic analysis. Further investigations require both an increase of the data set and other measures besides *f0* and speech rate. Finally, this paper represents a first attempt to carry out a data-driven analysis of prosodic phenomena by using advanced statistical tools (*f0* extraction, automatic segmentation, and functional data analysis) for the inspection of contours. Our work sheds light on advantages as well as limitations of such an approach when dealing with semi-spontaneous speech, and our preliminary findings suggest that automatic quantitative analysis provides a promising direction for investigation of naturalistic linguistic data.

5. Acknowledgments

The research of the first author is framed in the ANR-DFG project 'LANGACROSS' (DFG Grant and the Geschäftszeichen DI 808/1-1) awarded to Christine Dimroth. The research of the second and the third authors was supported by the Marie Curie Research Training Network *Sound2Sense* (www.sound2sense.eu).

References

[1] T. Höhle, Über Verum-Fokus im Deutschen. *Linguistische Berichte, Sonderheft*, 4, pp. 112-141, 1992.

[2] C. Dimroth, C. Andorno, S. Benazzo, and J. Verhagen, "Given claims about new topics. How Romance and Germanic speakers link changed and maintained information in narrative discourse". *Journal of Pragmatics*, 42(12), pp. 3328-3344. doi:10.1016/j.pragma.2010.05.009, 2010.

[3] G. Turco, B. Braun, "They ARE different: intonational means to mark polarity contrast in German and French", (in prep.)

[4] A. Dufter, D. Jacob (Eds.), *Focus and Background in Romance Languages*, pp. 155-204, Amsterdam: John Benjamins, 2009.

[5] G. Turco, *The Polarity-Switch Dialogue*. Max Planck Institute for Psycholinguistics, Browsable Corpus Archive [online].Available: https://www.mpi.nl. 2009

[6] S. Young, S., G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book 3.2. Cambridge: Entropic, 2002.

[7] P. Boersma, D. Weenink, Praat: doing phonetics by computer [Computer program]. Version 5.2.20, 2011.

[8] J. O. Ramsay, B. W. Silverman, Functional Data Analysis. 2nd Ed., Springer, 2005.

[9] M. Gubian, F. Cangemi and L. Boves, "Joint analysis of f0 and speech rate with functional data analysis", in Proceedings of the 2011 International Conference on Acoustics, Speech and Signal Processing, ICASSP2011, May 22-27, 2011 Prague, Czech Republic, pp. 4972-4975.

[10] F. Cangemi, L. Boves, "Automatic and data driven pitch contour manipulation with functional data analysis," Proceedings of Speech Prosody, 11–14 May 2010, Chicago, USA, pp. 100954:1–4, 2010.

[11] M., D'Imperio "Focus and tonal structure in Neapolitan Italian." *Speech Communication*, 33, pp. 339–356, 2001.

[12] M., D'Imperio "Italian intonation: An overview and some questions." *Probus,* 14, pp. 37-69, 2002.

[13] C., Fougeron, S.-A., Jun "Rate Effects on French Intonation: Prosodic Organization and Phonetic Realization". *Journal of Phonetics,* 26, 45-69, 1998.

[14] M. D'Imperio, G. Elordieta, S. Frota, P. Prieto, M. Vigarion, "Intonational phrasing and constituent length in Romance". In S. Frota and M. Vigàrio, M.J. Freitas, *Prosodies* (Selected papers from the Phonetics and Phonology in Iberia Conference, 2003). [Phonetics and Phonology Series]. The Hague: Mouton de Gruyter, pp. 59-98, 2005.

[15] M. D'Imperio, B. Gili Fivela "How many levels of phrasing? Evidence from two varieties of Italian", in Phonetic Interpretation, Papers in Laboratory Phonology VI, J. Local, R. Ogden, R. Temple (eds.). Cambridge University Press, Cambridge, pp.130-1.2004.

[16] G., Turco "The expression of Verum focus in Romance and Germanic languages: Prosody and particles in native speakers and advanced L2 learners." MPI, Nijmegen (in prep).