Modelling intonation: Beyond segments and tonal targets

Francesco Cangemi, Aviad Albert, Martine Grice

I*f*L Phonetics – University of Cologne
fcangemi@uni-koeln.de

ABSTRACT

A widespread approach to research on intonation categories involves extracting relevant landmarks in the tune (e.g. F0 turning points) and relating them to relevant landmarks in the text (e.g. phone boundaries). This approach is problematic, both in theoretical terms (e.g. no consensus as to what is a relevant landmark in either domain) and in practical terms (e.g. locating segments and F0 turning points with either error-prone automatic annotations or time-consuming manual annotations).

We propose and test an alternative approach that overcomes both types of problem. The tune is modelled taking into account the shape of F0 contours, and the text is modelled by means of periodic energy curves. We exploit the interaction between these two dynamic trajectories to observe and quantify intonation. Compared to the standard approach, our method requires less manual work and makes fewer theoretical assumptions, but nonetheless has similar descriptive power.

Keywords: Alignment, Synchrony, Periogram

1. INTRODUCTION

It is notoriously difficult to adequately describe intonational categories and to distinguish between them. To these ends, intonation analysts often rely on the extraction of acoustic cues from both the tune and the text. For example, in order to characterise the difference between the pitch accents employed in questions vs. statements in some languages, researchers annotate the temporal boundaries of the stressed vowel or syllable (the text) and locate a high tonal target in the F0 contour (the tune). These landmarks are then used to calculate the position of the F0 peak with respect to the chosen segmental unit (i.e. the tonal alignment) [1]. If questions and statements showed consistent differences in their tonal alignment, there are reportedly (at least) two distinct intonational categories in the language under investigation. This approach (henceforth standard approach), often under the umbrella of autosegmental-metrical (AM) phonology [2], has informed a large part of intonation research in the last two decades, and has yielded a considerable number of insights for a great number of languages. However, it is problematic in at least two respects.

First, on practical grounds, it requires the location and annotation of landmarks in the tune (e.g. tonal targets) and landmarks in the text (e.g. segmental boundaries). These operations are not trivial, since manual annotation is known to be time-consuming, and automatic annotation error-prone [3].

Second, in theoretical terms, this approach requires defining which landmarks are relevant, both for the tune and for the text. However, in most cases, such choices are tied to a specific theoretical framework, either explicitly or implicitly. Take for example the choice of high turning points as relevant landmarks for the characterisation of the tune. This is a reasonable choice for models within the AM framework, relying on a sequential organisation of level tonal targets. However, this choice is also intrinsically incompatible with models approaching the perception of intonation in terms of configurations [4]. Note that the notion of a tonal target has also been treated as problematic even within sequential models [5,6], as in work on the perception of high plateaux [7], or on F0 contour shapes as captured by the Tonal Center of Gravity [8]. Similarly, the use of segments as relevant landmarks for the text implicitly assumes a strictly sequential representation of speech. This view has been challenged by models that emphasise the temporal overlap in the activation of different articulators [9], the continuous nature of the acoustic signal [10], the integrated nature of speech perception [11] and the epiphenomenal nature of phoneme-sized units [12].

In this paper, we propose and test an alternative approach, which overcomes both the methodological and the theoretical shortfalls mentioned above. The approach relies on modelling both the text and the tune without discretising them into sequential units. Rather than reducing F0 contours to turning points, the tune is modelled by taking into account the full details of the shape of the contours in the relevant regions of the signal. Rather than segmenting the speech stream into phones and syllables, the relevant regions of the text are modelled by extracting periodic energy curves. Periodic energy fluctuations show peaks in correspondence with vocalic sounds, and troughs in correspondence with non-vocalic sounds. As such, they can be used as a proxy for syllabic cycles, thus reducing the necessity for segmenting the text. Importantly, periodic energy is also crucial to adequately model the perception of tonal events, since sensitivity to F0 is known to be greater when periodic energy is stronger [13].

To test this new approach, we compare its performance with a classical analysis in terms of tonal alignment, using a corpus of read Neapolitan Italian speech. In the following, we detail the metrics we employ when applying both the standard approach based on the *Alignment* of high turning points with respect to syllabic boundaries (§2.1.1), and our own proposal based on the *Synchrony* of tonal centers of gravity with centers of periodic mass (§2.1.2). Then we introduce the corpus (§2.2) and detail the logic behind the statistical modelling (§2.3), before we report on our results in both visual (§3.1) and quantitative (§3.2) terms.

2. METHOD

2.1. Metrics

2.1.1. Alignment (and Scaling)

To model our data using the standard approach, we applied state-of-the-art practices in intonation research. For each utterance in the dataset, we extracted the F0 contour, and manually corrected pitch-detection errors using a scripted procedure, which yields smoothed curves [14,15]. We used first and second derivatives to annotate the high tonal targets and turning points of the relevant pitch accents. We then used a publicly available forced-alignment tool to segment the utterances at the phone level [16]. The performance of this forced-aligner was evaluated against manual segmentation, yielding satisfactory results (e.g. 94% of phone boundaries are positioned within 20ms of the manual reference).

Once a turning point was located, its *Alignment* was calculated as the position of the tonal target relatively to the duration of the stressed syllable. In this way, tonal targets appearing at the midpoint of the stressed syllable receive an alignment value of 50%, and targets located in the post-stressed syllable have alignment values >100%. We also extracted peak Scaling, calculated in relative terms, by taking the F0 range of the relevant speaker into account, using all her productions available in the dataset.

2.1.2. Synchrony (and Excursion)

In our alternative approach, we obtain periodic energy data using the APP detector [17] and combine them with the corresponding F0 time-series. The periodic energy curves undergo a set of smoothing and log transform functions in R [18], and they are corrected for loss of data when periods in the signal change length too rapidly (i.e. when steep rises or falls occur in F0; see [19] for details).

Using the first and second derivatives of the periodic energy curves we automatically locate local minima that reflect the boundaries of periodic fluctuations, expecting the number of fluctuations to correspond to the number of syllables (always eight in this corpus, cf. §2.2). The Center of Periodic Mass (CoPM) is calculated as the average time point within fluctuation, weighted by periodic energy to designate the balance point of each fluctuation. Finally, an algorithm is designed to choose the single accented fluctuation which corresponds to the pitch-accented syllable (96% of CoPM in accented fluctuations fall within the region of the stressed vowel).

Using this procedure, we obtain three time points that define the accented fluctuation in time, alongside continuous data that reflect the strength of its vocalic content (area under the periodic energy curve), and, in conjunction with F0, reflects the relative strength of different F0 qualities over time (Periogram [19]).

The accented periodic fluctuation is taken as the locus of the tonal event in question and is used to determine the characteristics of this event. We measure the Tonal Center of Gravity (TCoG, [8]) in a similar way to the above-mentioned CoPM measurement, only now the average time point within the fluctuation is weighted by the F0 trajectory (subtracted by the local minimum F0 value). The TCoG within fluctuation is sensitive to continuous aspects of the contour shape (linear rise vs. fall, as well as concave vs. convex slope), thus the distance between the TCoG and the CoPM within the accented fluctuation is taken to reflect perception in terms of synchrony between the text and the tune. For example, the location of the TCoG further to the right of the CoPM signals the perception of a rising movement. The distance between the two points is expected to increase with a steeper rise and/or a more convex shape. The measurements presented under *Synchrony* are obtained by calculating the TCoG-to-CoPM distance, relative to the duration of the accented periodic fluctuation (zero denotes perfect synchrony, akin to an in-phase relation).

The accented fluctuation is also informative in relation to the relative scaling of the tonal event. To this end, we measure the F0 range within fluctuation and compare it with the speaker’s range to report scaling in terms of the relative F0 Excursion.

2.2. Material

These four metrics (*Alignment* and Scaling; *Synchrony* and Excursion) were extracted for the 756 utterances composing a corpus of read Neapolitan Italian speech ([20], §4.2.1.2). 21 native SPEAKERS read 3 REPETITIONS of one of 2 SENTENCES composed by subject, verb and indirect object. Sentences were presented together with a contextualisation paragraph, which induced one of six possible interpretations. These are given by the combination of two different sentence MODALITIES (question vs. statement) with three different FOCUS placements (subject vs. verb vs. object). A small number of items was discarded for technical reasons (N=25).

According to descriptions of the intonation of this variety of Italian, peaks in Questions are expected to be aligned later than peaks in Statements, making MODALITY the crucial predictor for the following analyses, and *Alignment* vs. *Synchrony* our test metrics [21]. Scaling and Excursion will be used merely to visualise the dataset in a familiar way (e.g. utterances as point clouds, with *Alignment* on the x-axis and Scaling on the y-axis), given their reportedly minor role in this contrast [20].

2.3. Modelling

The quantitative analyses below (§3.2) are based on linear mixed effect modelling. Intonation in this variety is sensitive to the position in the utterance of the focal accent, since questions can also be characterised by the presence of a final F0 rise (which can further affect tonal alignment; see [22] for details). For this reason, tests were run separately for each FOCUS condition. We predicted the two test cues (*Alignment* and *Synchrony*) using four individual models, specifying MODALITY and REPETITION (and their interaction) as fixed factors, and fitting random intercepts and random by-MODALITY slopes for SENTENCE and SPEAKER.

The interaction between REPETITION and MODALITY did not contribute to the fit of most models, but since it did contribute in a few cases, we refrained from simplifying this term in any of our models. We then compared models predicting the old and new metrics, separately for each FOCUS condition (e.g. *Alignment* vs. *Synchrony* for data from FOCUS condition 1). Since the models predict different dependent variables, they could not be compared using Likelihood Ratio Testing. In cases like this, the usefulness of the Akaike Information Criterion is also disputed. We thus compared old and new cues by means of marginal R^2 [23,24], as implemented in [25].

3. RESULTS

3.1. Data visualisation

We compared the performance of our two metrics by plotting data separately for each FOCUS condition, but jointly for all SPEAKERS, SENTENCES and REPETITIONS. The two MODALITIES were symbol-coded (Question: empty black square, Statement: filled red circle). The Figures below show data for Object-FOCUS cases, separately for *Alignment* (**Figure 1**) and *Synchrony* (**Figure 2**). Vertical lines are added at x={0,100} in Figure 1 to indicate the boundaries of the stressed syllable, dividing the plot in three vertical panels (roughly corresponding to early, medial and late peaks). The vertical line x=0 in Figure 2 shows the point of perfect synchronisation between TCoG and CoPM, dividing the plot into two panels (roughly corresponding to falling and rising contours). The plots suggest that the two test metrics perform very similarly. This holds for the two other FOCUS conditions (omitted for brevity).

Figure 1: Standard approach (*Alignment*).

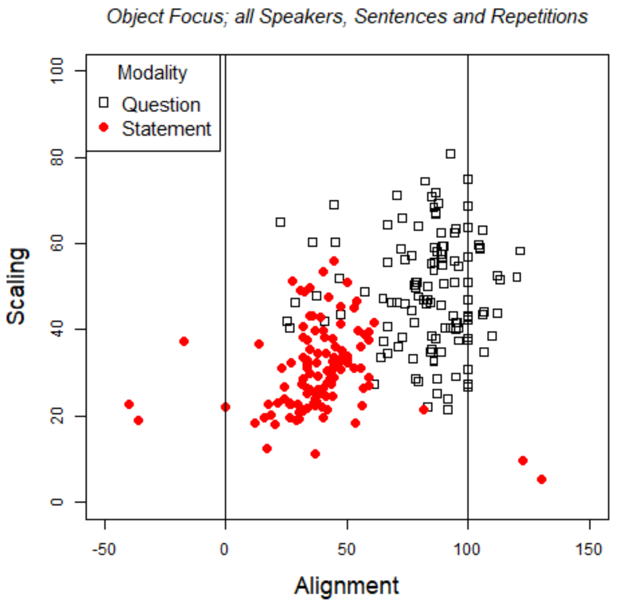
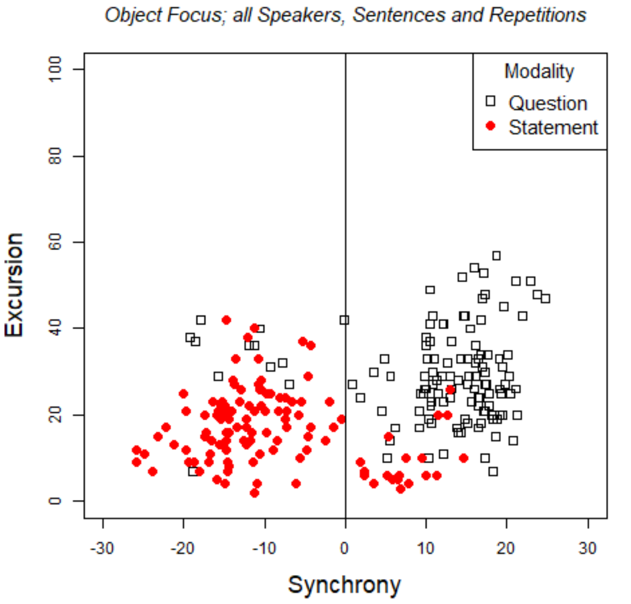


Figure 2: Proposed approach (*Synchrony*).



3.2. Quantitative analysis

Statistical modelling confirms the outcome of data visualisation. As expected, predictions for both *Alignment* and *Synchrony* have significantly higher accuracy when MODALITY is included in the model ($p<0.001$ for both metrics and for all focus types). More importantly, the new metrics appear to be as reliable as the old ones in separating the two MODALITIES.

For example, in the case of Object-FOCUS data, the model predicting *Alignment* has a marginal R^2 (*R^2m*) of 0.56, while the model for *Synchrony* has 0.58. This small difference in favour of the new metrics is not attested for Verb-FOCUS data, and is inversed for Subject-FOCUS data. This pattern suggests that neither *Alignment* nor *Synchrony* perform clearly better than the other, indicating that the old and new metrics have similar descriptive power. For completeness, **Table 1** also include values for conditional R^2 (*R^2c*, which seem to favour *Alignment* for all FOCUS conditions) and for AIC (which seem to favour *Synchrony* for all FOCUS conditions), thus confirming that *Alignment* and *Synchrony* show comparable performance.

Table 1: Marginal R^2 comparisons, by FOCUS.

FOCUS	Metrics	R^2m	R^2c	AIC
Subject	**Alignment**	**0.60**	0.74	2043
	Synchrony	0.44	0.66	1696
Verb	Alignment	0.57	0.74	1904
	Synchrony	0.57	0.70	1588
Object	Alignment	0.56	0.76	2055
	Synchrony	**0.58**	0.74	1703

4. DISCUSSION

Our results suggest that, compared to the standard metric of *Alignment*, the proposed metric of *Synchrony* is equally reliable in separating Question from Statement focal pitch accents in read Neapolitan Italian. We interpret these findings as suggesting that no descriptive power is lost when using metrics that rely neither on segmenting the speech stream, nor on reducing intonation contours to a sequence of tonal targets.

More importantly, by removing the need for the laborious and delicate procedure of segmenting the text (into segments or syllables) and the tune (into tonal targets), the new metrics also contribute to reducing both the workload required for manual annotation and the problems encountered owing to the reliability of automatic annotation of segments and tonal targets. These difficulties in extracting and parametrising intonation are an important reason for the relative overrepresentation of experiments using carefully scripted read speech, as opposed to the relatively less frequent use of ecologically valid spontaneous recordings. Most intonation research relies on stimuli avoiding intervocalic geminates, voiceless sounds, and other features that make the discretisation of tune and text less manageable. These stimuli might indeed be “ideal” for the extraction of *Alignment* due to the lack of F0 discontinuities and of ambiguous segmental boundaries, but they require the use of tasks often devoid of communicational plausibility, such as reading tasks. Using periodic energy to model the text (with emphasis on the strength of the vocalic content), allows us to shed light on the relevant portions of the F0 contour from within the acoustic signal itself. By removing the need of discretising and reducing the signal, we remove the need to use stimuli that minimise these problems, with the consequence of broadening the spectrum of speech materials viable for intonation research.

Beyond having similar descriptive power, the crucial contribution of this new approach lies in reducing the amount of theoretical assumptions required to model intonation. Standard autosegmental-metrical modelling rests on the assumption that the text can be segmented into units and the tune can be discretised into a sequence of tonal targets. As such, it is only with great effort that it can be made compatible with models of holistic pitch perception. The proposed novel approach removes this hurdle, while retaining a crucial insight of the AM approach, namely the relevance of the tune-text synchronisation.

5. CONCLUSION

We introduced and tested an approach to intonation modelling that relies on the parametrisation of continuous signals (F0 contours and periodicity profiles) rather than on the discretisation of the tune (into tonal targets) and of the text (into segments). The model makes fewer theoretical assumptions and can be applied to stimuli traditionally considered as sub-optimal in intonation research. As such, the proposed method holds promise for the unification of intonation research, being compatible with a variety of theoretical frameworks and testable on a variety of speech materials.

6. ACKNOWLEDGEMENTS

Parts of this work were funded by the German Research Foundation (through the SFB 1252 “Prominence in Language”) and by the German Academic Exchange Service.

7. REFERENCES

[1] Arvaniti, A., Ladd, D.R., Mennen, I. 2006. Tonal association and tonal alignment: evidence from Greek polar questions and contrastive statements. *Language and Speech* 49, 421-450.

[2] Pierrehumbert, J. 1980. *The phonology and phonetics of English intonation*. PhD thesis, MIT. Distributed 1988, Indiana University Linguistics Club.

[3] del Giudice, A., Shosted, R., Davidson, K., Salihie, M., Arvaniti, A. 2007. Comparing methods for locating pitch "elbows." *Proceedings of the 16th International Congress of Phonetic Sciences* Saarbrücken, 117-120.

[4] 't Hart, J., Collier, R., Cohen, A. 1990. *A perceptual study of intonation: An experimental-phonetic approach*. Cambridge: Cambridge University Press.

[5] Ladd, D.R. 2008. *Intonational phonology* (2nd edition). Cambridge: Cambridge University Press.

[6] Niebuhr, O., D'Imperio, M., Gili Fivela, B., Cangemi, F. 2011. Are there “shapers” and "aligners"? Individual differences in signalling pitch accent category. *Proceedings of 17th International Congress of Phonetic Sciences* Hong Kong, 120-123.

[7] Knight, R.A., Nolan, F. 2006. The effect of pitch span on intonational plateaux. *Journal of the International Phonetic Association* 36(1), 21-38.

[8] Barnes, J., Veilleux, N. Brugos, A., Shattuck-Hufnagel, S. 2012. Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology* 3(2), 337-383.

[9] Browman, C.P., Goldstein, L. 1992. Articulatory phonology: an overview. *Phonetica* 49 (3–4), 155–180.

[10] Firth, J. R. 1948. Sounds and prosodies. *Transactions of the Philological Society* 47(1), 127-152.

[11] Nguyen, N., Hawkins, S. 2003. Temporal integration in the perception of speech. *Journal of Phonetics* 31(3/4), 279-287.

[12] Cangemi, F., Niebuhr. O. 2018. Rethinking canonical forms. In: Cangemi, F., Clayards, M., Niebuhr, O., Schuppler, B., Zellers, M. (eds). *Rethinking reduction: Interdisciplinary perspectives on condition, mechanisms, and domains for phonetic variation*. Berlin: De Gruyter Mouton.

[13] Oxenham, A.J. 2012. Pitch perception. *The Journal of Neuroscience* 32 (39), 13335-13338.

[14] Cangemi, F. 2015. *mausmooth*. [Computer program]. Retrieved from http://ifl.phil-fak.uni-koeln.de/fcangemi.html

[15] Boersma, P., Weenink, D. 2018. *Praat: doing phonetics by computer* [Computer program]. Retrieved from http://www.praat.org/

[16] Cangemi, F., Cutugno, F., Ludusan, B., Seppi, D., Van Compernolle, D. 2011. ASSI - Automatic Speech Segmentation for Italian: tools models, evaluation and applications. In: Gili Fivela, B., Stella, A., Garrapa, L., Grimaldi, M. (eds). *Contesto comunicativo e variabilità nella produzione e percezione della lingua* (Atti del VII Convegno Nazionale AISV. Lecce, January 2011). Roma: Bulzoni Editore, 337-344.

[17] Deshmukh, O., Espy-Wilson, C. Y., Salomon, A., Singh, J. 2005. Use of temporal information: Detection of periodicity, aperiodicity, and pitch in speech. *IEEE Transactions on Speech and Audio Processing* 13 (5): 776-786.

[18] R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Computer program.

[19] Albert, A., Cangemi, F., Grice, M. 2018. Using periodic energy to enrich acoustic representations of pitch in speech: A demonstration. In *Proc. 9th International Conference on Speech Prosody*, 804-808.

[20] Cangemi, F. 2014. *Prosodic detail in Neapolitan Italian*. Berlin: Language Science Press.

[21] D’Imperio, M. 2001. Focus and tonal structure in Neapolitan Italian. *Speech Communication* 33(4), 339–356.

[22] Cangemi, F., Grice, M. 2016. The importance of a distributional approach to categoriality in autosegmental-metrical accounts of intonation. *Journal of the Association for Laboratory Phonology*, 7(1):9, 1-20

[23] Nakagawa, S., H. Schielzeth. 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in Ecology and Evolution 4(2), 133-142.

[24] Johnson, P. 2014. Extension of Nakagawa & Schielzeth’s R2GLMM to random slopes models. Methods in Ecology and Evolution 5(9), 944-946.

[25] Lefcheck, J. 2016. piecewiseSEM: Piecewise structural equation modeling in R for ecology, evolution, and systematics. Methods in Ecology & Evolution 7(5), 573-579.