9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Using periodic energy to enrich acoustic representations of pitch in speech: A demonstration

Aviad Albert, Francesco Cangemi, Martine Grice

University of Cologne, Germany
a.albert@uni-koeln.de, fcangemi@uni-koeln.de, martine.grice@uni-koeln.de

Abstract

This paper aims to strengthen the link between acoustic and perceptual representations of intonation, a link that has been weakened by the over-reliance on the F0 trajectory, which can only be interpreted in relation to landmarks in the segmental string, placed manually or semi-automatically at a separate stage in the analysis. Only then can F0 events be identified as linguistically relevant (e.g. early, medial or late peaks, accentual tones or edge tones etc.).

We provide an analysis and visualization of two acoustic dimensions contributing towards the perceived pitch contour, F0 over time and, crucially, periodic energy. Periodic energy reflects the degree to which pitch is intelligible, a higher value representing a stronger F0 signal that is consequently more easily perceived. A representation of F0 that includes periodic energy is thus able to flag portions of the speech signal that are relevant for the analysis of intonation, without the need for a separate segmentation of the signal into phones and syllables.

Index Terms: intonation, pitch perception, periodic energy, tonal alignment, segmentation, data visualization, sonority

1. Introduction

While articulation and perception relate directly to the human capacity for speech, there is only an indirect relation to acoustics. Although research into intonation has involved numerous studies on perception (and very few on articulation), the main bulk of research has relied on acoustic data. The latter is easier to collect, maintain, process and quantify. This leap from acoustics to perception is plausible thanks to some well-established correlations between acoustic and perceptual phenomena. For example, while it is relatively difficult to collect data on the perception of pitch, one can easily and quite reliably measure an established correlate of pitch from acoustic signals – the *fundamental frequency* (F0) [1,2].

This paper focuses on the acoustics-perception link in the phonetics and phonology of intonation. We use measurements of acoustic periodic energy to enrich standard F0 representations with a view to gaining new insights informing our research methodologies. The goal of this paper is to demonstrate the advantages of these enriched representations in the analysis of intonation.

2. Background

The acoustic toolbox of most phonetic-phonological research has barely changed over the last few decades. Its main tools remain the products of the *Fourier Transform*, that measure acoustic power at different frequencies, alongside various techniques for F0 detection. In segmental phonology, which also enjoys a large body of articulatory-based research, the products of the standard Fourier Transform are usually deemed enough for a satisfying acoustic description. Measurements of acoustic energy at different frequencies over time, have thus far proven to be an effective way to measure the quality and quantity of most linguistically distinctive segmental phenomena (e.g. formant structure, spectral dispersion, transiency etc.).

Since the study of intonation does not lend itself to straightforward articulatory investigation, it relies almost exclusively on perception, using mostly acoustic data in order to quantify and analyze prosodically distinctive phenomena. In intonation research, the F0 trajectory is perhaps the single most crucial acoustic dimension used in the description of any model. As such, it is surprisingly poor, consisting only of a single vector of continuous data reflecting the F0 value over time.

In intonational phonology, F0 trajectories have to be mapped on to "meaningful" abstract categories, such as tones and tonal clusters (henceforth tonal events), serving to mark *accents* and *boundaries*. Locating the acoustic reflex of these tonal events is not a straight-forward task using a standard representation of F0. Trajectories are 2-dimensional (F0 over time) and binary in terms of strength/intensity (F0 data is either present or absent from the analysis at any time point). To enrich this binary aspect of F0 representation, researchers employ assumptions that refer to other phonological abstract entities, namely *syllables* and possibly *segments*. For example, it is widely assumed that pitch accents are associated with stressed syllables and that tonal information is more salient on vowels than on consonants (as well as more on sonorants than on obstruents). With current tools, we need these discrete segmental/syllabic landmarks to make sense of F0 trajectories and to analyze them in comparable ways (e.g. by using landmarks within or in the vicinity of stressed syllables as anchors for comparison of different accent types).

An F0 trajectory without segmental landmarks is difficult to interpret. Changes in the shape of the trajectory can only be taken as indicative of tonal events if they can be identified as categories within the language under investigation. Much work within the Autosegmental-Metrical model [3,4] has been concerned with tune-text association, reflected in the synchronization between aspects of the F0 trajectory (usually turning points or elbows) and landmarks in the segmental string (in relation to stressed syllables or edges of constituents) obtained through a separate segmentation process.

3. Current proposal

Segmentation of speech data requires a separate process in which symbolic discrete entities are concatenated linearly, with no overlap. It is therefore a theoretically limiting and methodologically time-consuming process. A richer acoustic representation of F0 trajectories could highlight the salient

10.21437/SpeechProsody.2018-162

portions of F0 trajectory, with a potential for uncovering the relevant parts (i.e. tonal events), independently of any additional layer of segmentation. We report here on promising attempts to achieve this goal with the addition of periodic energy measurements to our acoustic analyses.

Measurements of periodic energy may be understood as reflecting a certain dimension of acoustic intensity. Unlike general acoustic intensity, or any type of frequency-filtered intensity, periodic energy directly correlates with the strength of F0 [5,6], which is expected to be higher in intonationally relevant portions of the signal. Periodic energy is also an acoustic dimension with extreme importance for speech research, as it provides the medium for transmission of pitch events, modulating the relation between tune and text. Periodic energy trajectories tend to align their local peaks with the location of syllable nuclei, especially vowels. We exploit this tendency to automatically segment the speech data in accordance with the macro-fluctuations in the periodic energy curve, in a manner that resonates with previous uses of frequency-filtered intensity curves (see [7,8,9,10,11] for a few prominent examples).

Incorporating periodic energy in analyses of F0 is motivated by general auditory perceptual and cognitive principles, according to which periodic energy correlates with pitch intelligibility [5,6]. There are also good linguistic motivations (even if more arguable) to incorporate periodic energy in phonological analyses, as periodic energy seems to correlate well with sonority [12,13], as well as with syllable-sized units.

Taken together, it should be possible to obtain analyses of F0 trajectories that reflect the relative strength of the tonal information, as it is transmitted by the underlying segmental makeup, yet in a continuous fashion, independently of any prior annotation of segmental and syllabic landmarks. Furthermore, the periodic energy trajectory should be capable of segmenting the speech stream into units that are comparable to the underlying sequence of syllables.

4. Pre-processing

We obtain continuous measurements of periodic energy using the *APP Detector*, a computer code that was introduced in [14] and developed in subsequent publications up to 2008 [15,16], with the ability to measure periodic energy in audio signals. In what follows we describe how we use these periodic energy measurements alongside more typical acoustic speech data obtained from *Praat* [17]. The end result at this stage requires the combination of data from different sources in order to display, calculate and analyze interactions between acoustic data types. Furthermore, an ad-hoc patch was required in this task to compensate for some technical problems that we encountered with the specific set of tools at our disposal (see 4.3). It is therefore important to view this work not as a finished product, but, rather, as a preliminary demonstration of possibilities that await better technological solutions. In the current paper we combine all data within *R* [18] where data manipulation, visualization, quantization and statistics are all extensively supported, easily accessible and freely available.

4.1. The APP detector

The APP Detector is capable of measuring (among other things) a spectral distribution of periodic energy, i.e. the amount of periodic energy at different frequency rates over time. Although the typical spectral range for human hearing is between approx. 20 and 20,000 Hz, our pitch perception range is limited to periods between approx. 30 and 4,000 Hz [19,20,21]. The typical range of F0 in speech is well within that range, in a pitch-privileged bandwidth between approx. 50 and 600 Hz. We therefore sum over the different frequencies that the APP Detector measures to obtain one vector of the sum of periodic energy (over different frequencies) at each time point, under the assumption that different pitch heights are equally intelligible within the typical human speech range (male–female, young–old). This sum of periodic energy is log transformed and divided by a value that corresponds to energy beyond the threshold of effective pitch transmission, much like in the standard equation for dB SPL, where the log variable is divided by a constant which represents the threshold of hearing, thus obtaining a meaningful zero. To obtain this value and adjust the floor of the periodic energy vector we measure purely voiceless portions and set the maximal periodic energy value of those voiceless portions as the constant floor denominator of the log variable.

4.2. Specialized Praat programs

We use *mausmooth* [22] to extract manually inspected and automatically smoothed and interpolated F0 trajectories using Praat. We also use *Prosogram* [23] to extract the corresponding band-pass filtered intensity data from Praat. Text-grid annotations with segmental and/or syllabic data are also readily incorporated within the same data frame in R.

4.3. Pitch-patch

A quick survey of speech data revealed one major systematic problem with the measurements of periodic energy that we obtained from the APP Detector. Whenever there is a relatively sharp F0 change over time, which is often the case in accented portions of speech, the periodic energy curve drops and appears as irregularly low when it is actually expected to be higher than average. We concluded that the APP Detector fails to reliably detect periodic energy when the rate of change in the length of periods exceeds a certain threshold. To overcome this problem in an ad-hoc manner we designed a patch by computing the *growth rate* (or *first derivative*) of the F0 curve. This new data vector was designed to operate at around the same time points and velocity thresholds in which the APP Detector fails, thus adding back the missing parts to the periodic energy curve, based on the amount of change in F0 (see Figure 1). To further control this measurement of F0 growth rate we multiplied its values by the corresponding values of the general intensity curve. Thus, the growth rate of the F0 curve is limited by the general intensity level of the signal (so, for example, the growth rate of a sharp rise in F0 will be diminished when the intensity curve is at the same time falling, but it will be enhanced if intensity is at the same time rising). Clearly, this patch is not desirable, but given its relative success and the rationale behind it, we believe that future periodic energy detectors will be able to dramatically improve on that.

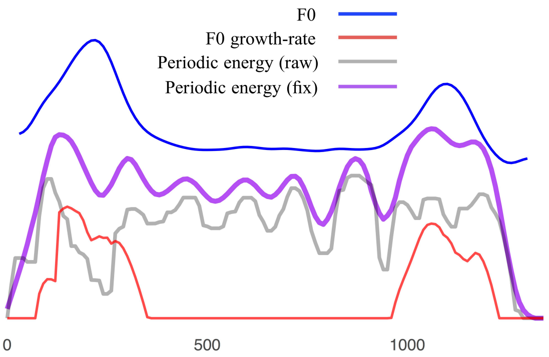

Figure 1: *Raw periodic energy data (gray) results in fixed and smoothed curve (purple), augmented by the controlled growth-rate levels (red), based on changes in the F0 curve (blue).*

Examples in Figures 1-3 are taken from elicited speech for intonation research in Italian. All figures exhibit the sentence *Danilo vola da Roma* ('Danilo flies from Rome'). Figures 1-2 feature an object-focus interrogative (main accent on **RO**-ma**)**, while Figure 3 features a subject-focus interrogative (accent on da-**NI**-lo).

4.4. Smoothing and cycling

Finally, we fit a LOESS smooth [24] to the adjusted combination of periodic energy and the F0 growth-rate patch. The resulting final smoothed periodic energy curve is now informative in various ways. It typically exhibits a sequence of fluctuations with local minima (onsets and offsets) and peak points between them. We refer to each interval between two local minima as a *periodic cycle*. In clear careful speech, each local peak along the periodic energy curve aligns with a sonority peak, most often associated with an underlying vowel which would also be considered the syllabic nucleus. This 1:1 mapping between periodic cycles and syllables becomes fuzzier with spontaneous rapid speech, as even whole syllables often undergo a phonological reduction [25,26]. In such cases, one periodic cycle may include more than one underlying syllable.

Given the low resolution of the periodic energy measurements at 10 ms intervals, there is a notable trade-off between amounts of smoothing and periodic cycle detection. If not sufficiently smoothed, transient fluctuations along the periodic energy curve may appear as local peaks, resulting in cases where there is more than one periodic cycle for one underlying syllable. Smoothing too much, on the other hand, results in more cases where 2 syllables are collapsed into one periodic cycle. We tend to prefer the latter result when adjusting the smoothing parameters, as it is, indeed, reflective of natural reduction processes, while the former result is more reflective of the technical shortcomings of the machinery. Once again, we believe that future periodic energy detectors could allow much higher resolutions that would reduce the need for smoothing while at the same time increase the reliability of cycle detection.

4.5. Periodic energy masses and their center

With periodic cycles that are generally equivalent to syllables in size, it is possible to measure the duration and overall intensity of each periodic cycle, as is standard in prosodic research. However, this results in 2 mutually exclusive values that are only partially sensitive to the tonal cycle — duration is completely indifferent to the acoustic content and while (band-pass) intensity roughly correlates with periodic energy, it lumps together periodic and aperiodic components of speech, making it less reliable for the task. Once a periodic energy vector is available, it is possible to measure the *area under the curve* of periodic energy. This gives one value which is the sum integral of the duration and intensity of the periodic component of the cycle. We refer to this type of measure as *periodic energy mass*, which we interpret as reflecting the relative strength of each periodic cycle.

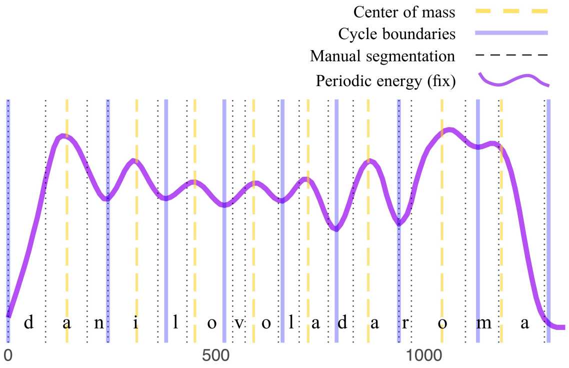

Figure 2: *Periodic cycles (purple curve), their center of mass (yellow dashed vertical line), their boundaries (blue vertical lines) and corresponding manual segmentation (black dotted vertical lines).*

Periodic cycles have one peak, typically around the middle of the cycle. We also calculate the *center of mass* (often referred to as *center of gravity* in the phonetics literature) of the periodic energy curve within each cycle. This enables us to calculate the average time point within cycles, weighted by periodic energy, as in (1),

$$\frac{\sum_i per_i\, t_i}{\sum_i per_i} \tag{1}$$

where we sum over the product of periodic energy (*per*) and time (*t*) at each time point (*i*), and divide that by the sum of periodic energy at those same time points.

The center of mass takes the shape of the periodic energy curve into account as it finds the point of equilibrium, rather than simply locating the peak within each cycle. It is therefore a good estimate for the perceptual tonal focal point.

It is also possible to obtain the weighted average F0 for each periodic cycle, using periodic energy as weights, in a similar procedure to the one used to find the center of mass in the time domain. We simply replace time values with F0 values, as in (2) below.

$$\frac{\sum_i per_i\, F0_i}{\sum_i per_i} \tag{2}$$

5. F0 representations

Periodic energy is extremely useful when analyzed in conjunction with F0. F0 alone reflects pitch height, while periodic energy estimates the strength of the signal producing the F0 (reflecting pitch intelligibility). Together, they can be used to achieve richer and more informative visual representations as well as novel quantification possibilities of F0 trajectories.

5.1. F0 data visualization: Periograms

We plot the F0 curve in R using ggplot [27]. As in the standard case, time is on the x-axis and F0 is on the y-axis of the 2-dimensional representation. In a similar way to a spectrogram display, we add a third dimension that reflects the strength of each F0 time point within the x,y matrix. We do this by letting the periodic energy vector control two variables of the line appearance in ggplot — transparency and width (via 'alpha' and 'size' variables). The resulting 3-dimensional representation displays an F0 curve with dynamically changing width and transparency. It is wide and solid at the most periodic portions and it becomes gradually narrower and more transparent as periodic energy drops. In portions that have zero periodicity, nothing will be displayed, even though the derived F0 data is continuous, having undergone *mausmooth* interpolation. We refer to this type of F0 display as a *periogram*. Periograms are information-rich alternatives for visual inspection of F0 data.

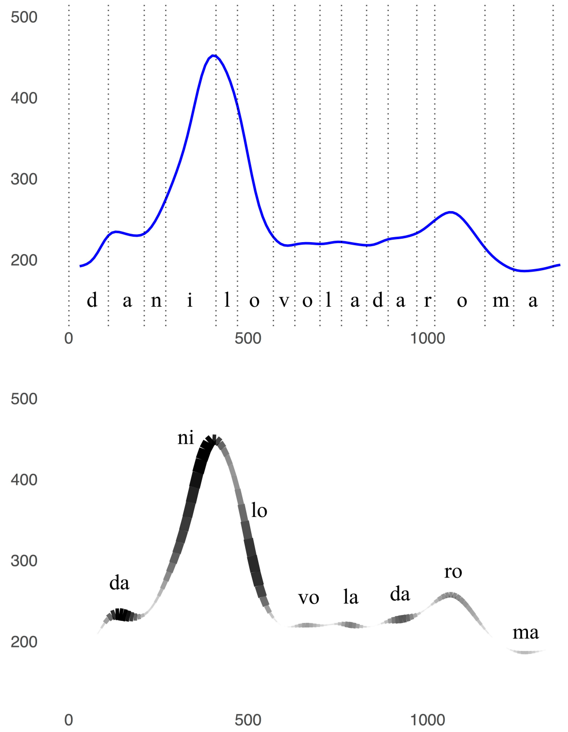

Figure 3: *Binary F0 display with manual segmentation (top) vs. periogram F0 display with annotated syllables at the center of mass locations (bottom).*

6. Discussion

6.1. Periodic energy and phonetic representations

Periodic energy can thus be used to enrich phonetic representations, as in the case of periograms. By modulating F0 trajectories with periodic energy, the periogram integrates the relevant acoustic cues into a perceptually motivated representation of the pitch contour of an utterance. Compared with the F0 trajectories traditionally employed in intonation research, periograms offer both a methodological and a theoretical advantage:

(i) Minima (and maxima) in the periodic energy function are indices of boundaries (and peaks) of periodic cycles. Periograms thus provide an automatic segmentation of the speech stream into syllable-sized units, thus removing the need for time-consuming segmental annotation, and yielding a significant methodological benefit. In this respect, it builds on earlier work on automatic syllabification [10,11,23] but focusses on periodic energy (rather than frequency-filtered intensity curves) as the main cue to syllabicity.

(ii) From a theoretical point of view, by integrating information on the dynamics of both F0 and periodic energy, periograms offer a representation of the acoustic signal which is richer and more informative than traditional F0 trajectories. Periograms thus further the line of work on perceptually-based pitch contours by providing a representation which is continuous in nature, and thus allows (but does not *require*) categorical accounts such as the use of glissando thresholds for the distinction between level vs. dynamic tones [23,28,29].

6.2. Quantification and application

In the previous sections we showed how intonation research can benefit from richer phonetic representations by focussing on the *visual representation* of pitch contours. While acoustically-rich, perceptually-motivated and graphically-intuitive pitch representations might already be seen as an improvement on the techniques currently in use, we also plan to use periograms for *quantitative analyses* and hypothesis testing.

Notably, by providing a representation of pitch that is based on two vectors (F0 and periodic energy) for separate periodic cycles, periograms contain all the information needed to model phonetic differences between tonal events, without the need for further data annotation. In most work carried out within the autosegmental-metrical approach, the establishment of an inventory of tonal events starts with a characterisation of phonetic differences in F0 trajectories. For example, two different pitch accents might be described as showing differences in tonal alignment: given two different rising-falling accents, an F0 peak aligned early in the syllable triggers the perception of a falling accent, while an F0 peak aligned late in the syllable triggers the perception of a rising accent. This approach thus requires prior detection both of syllable boundaries and of tonal targets. These are two non-trivial operations, especially when working on tonal targets other than F0 peaks (viz. troughs or elbows), on casual speech (for which segmentation poses additional challenges) or on understudied languages (for which annotators and forced alignment tools are not readily available).

Crucially, these two operations (syllable segmentation and target location) are not required in a periogram analysis. The differences in tonal alignment mentioned above would surface here as distinct movements (rising vs. falling) in the region of sufficiently high periodic energy within a given periodic cycle. For example, the second syllable of *Danilo* (proper name) in Figure 3 is clearly rising in its strongest portion. This is consistent with what we know from phonetics (the alignment of F0 peak is late in the syllable in questions) and from autosegmental-metrical phonology (questions have rising L*+H accents) [30]. The periogram illustrates this very clearly without the need to perform any segmental annotation.

7. References

[1] J. F. Schouten, "The residue and the mechanism of hearing," In *Proc. K. Ned. Akad. Wet.*, vol. 43, pp. 991-999. 1940.

[2] J. C. R. Licklider, "'Periodicity' pitch and 'place' pitch," *The Journal of the Acoustical Society of America* 26 (5): 945-945, 1954

[3] J. Pierrehumbert, *The phonetics and phonology of english intonation*, Doctoral dissertation, 1980.

[4] D. R. Ladd, *Intonational Phonology*. Cambridge: Cambridge University Press, 2008

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

[6] A. De Cheveigne, "Pitch perception models. In *Pitch: Neural Coding and Perception*," Ed. Christopher J Plack, Andrew J Oxenham, and Richard R Fay. New York: Springer, 2005.

[7] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *The Journal of the Acoustical Society of America* 58 (4): 880-883, 1975.

[8] P. Mertens, "Automatic segmentation of speech into syllables," In *Proceedings of the European Conference on Speech Technology*, 1987.

[9] D. House, *Tonal Perception in Speech*. Lund: Lund university press, 1990.

[10] M. Petrillo and F. Cutugno, "A syllable segmentation algorithm for English and Italian," In *Eighth European Conference on Speech Communication and Technology*, 2003.

[11] N. Obin, F. Lamare and A. Roebel, "Syll-O-Matic: An adaptive time-frequency representation for the automatic segmentation of speech into syllables," In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[12] P. Ladefoged, "Linguistic phonetic descriptions," In *The Handbook of Phonetic Sciences*. Ed. William J Hardcastle and John Laver. Oxford; Cambridge, Mass.: Blackwell, 1997.

[13] B. Heselwood, "An unusual kind of sonority and its implications for phonetic theory," *Working Papers in Linguistics and Phonetics* 6:68-80, 1998.

[14] O. Deshmukh and C. Espy-Wilson, "Detection of periodicity and aperiodicity in speech signal based on temporal information," In *The 15th International Congress of Phonetic Sciences,* Barcelona, Spain, 2003.

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

[16] S. Vishnubhotla and C. Y. Espy-Wilson, "Detection of irregular phonation in speech," In *ICPhS CVI*, 2007.

[17] P. Boersma, and D. Weenink, *Praat: doing phonetics by computer*. Computer program, 2018.

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

[19] G. E. Wever, and C. W. Bray, "The perception of low tones and the resonance-volley theory," *The Journal of Psychology* 3 (1): 101-114, 1937.

[20] F. Attneave, and R. K. Olson, "Pitch as a medium: A new approach to psychophysical scaling," *The American Journal of Psychology* 147-166, 1971.

[21] D. Pressnitzer, R. D. Patterson, and K. Krumbholz, "The lower limit of melodic pitch," *The Journal of the Acoustical Society of America* 109 (5): 2074-2084, 2001.

[22] F. Cangemi, *mausmooth.* Computer program. Version 1.0, retrieved 6 June 2017 from http://phonetik.phil-fak.uni-koeln.de/fcangemi.html, 2015.

[23] P. Mertens, "The prosogram: Semi-automatic transcription of prosody based on a tonal perception model," In *Speech Prosody* 2004, International Conference, 2004.

[24] W. S. Cleveland, E. Grosse and W. M. Shyu, "Local regression models," In *Statistical models in S*, eds. John M Chambers, Trevor J Hastie Wadsworth & Brooks/Cole, 1992.

[25] M. Ernestus, and N. Warner, "An introduction to reduced pronunciation variants," *Journal of Phonetics* 39 (3): 253-260, 2011.

[26] F. Cangemi, M. Clayards, O. Niebuhr, B. Schuppler and M. Zellers (Eds.), *Rethinking reduction: Interdisciplinary perspectives on conditions, mechanisms, and domains for phonetic variation*. Berlin: De Gruyter Mouton, (in press).

[27] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

[28] J. 'tHart, "Psychoacoustic backgrounds of pitch contour stylisation," *I.P.O. Annual Progress Report* 11, 11-19, 1976.

[29] M. Rossi, "La perception des glissandos descendants dans les contours prosodiques," *Phonetica* 35(1), 11-40, 1978.

[30] M. Grice, M. D'Imperio, M. Savino and C. Avesani, "Towards a strategy for labelling varieties of italian," In Sun-Ah Jun (ed.), *Prosodic typology: The phonology of intonation and phrasing*, 362–389. Oxford: Oxford University Press, 2005.