10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Perceptual dynamics in the processing of tonal alignment

Leonardo Lancia[1], *Cristel Portes*[2]

[1]Laboratoire de Phonétique et Phonologie (CNRS / Sorbonne Nouvelle), Paris
[2]Aix-Marseille Université, CNRS, LPL, Aix-en-Provence, France
leonardo.lancia@sorbonne-nouvelle.fr, cristel.portes@lpl-aix.fr

Abstract

The perception of intonational meaning is affected by multiple features of the f0 curves. It remains unclear however how these features interact and how they are integrated in the perception of intonational categories. In this study we investigate how the shape of the f0 curves and the lag between f0 targets and tone bearing syllables (i.e. tonal alignment) affect the perception of high pitch accents. To this aim we conceived a minimal dynamical model accounting for the perception of abstract tonal categories from the continuous f0 curve. In the model the internal representation of the f0 curve is continuously updated by a law of change that depends on features of the actual f0 curve but also on the current state of its internal representation. Through this model, we could successfully simulate the main findings observed so far in the literature concerning the interaction between f0 shape and tonal alignment in the perception of High tones. The success of the model demonstrates the role of perceptual dynamics in the processing of intonational categories. Moreover, it permits explaining the partial success of previous accounts of the same facts as well as their shortcomings.

Index Terms: perceptual dynamics, tonal alignment, intonation categories, Drift-Diffusion model

1. Introduction

The effects of the shape of f0 trajectories on the identification of intonational categories are often presented as problematic for level-based theories in which the characterization of intonational events relies exclusively on target levels and on their time anchoring to the segmental string. An additional issue for level-based theories is the frequent lack of clear turning points in regions of the f0 curve where intonational targets would be expected. This is problematic if the implementation of intonation is conceived as the interpolation of H and L f0 target values. Often, evidence for a role of the shape of the f0 curve on the perception on intonation categories comes from studies on the perception of tonal alignment, which is the lag between segmental events (e.g. the syllable or the vowel onsets) and f0 targets [1,2]. In several languages, this feature is used to distinguish intonational meaning. So for example, in Neapolitan Italian the same segmental content can be attributed a statement meaning if the pitch accent occurs early in the tone bearing syllable, or a question meaning if the accent occurs later during that syllable [3]. Several studies show that the perception of the segmental anchoring of an f0 peak is affected by the presence of a plateau in the f0 curve, by differences in the slope of the rising or falling movement, by differences in the scaling of the starting or ending f0 values or by differences in the curvature of the rising or falling portions [4-10]. As observed by [5], taken together these findings seem to indicate that the perceived location in time of the tonal events depends on the average location of high f0 values. Consistent with this idea, the Tonal Center of Gravity (TCoG) model [11] postulates that the location of the f0 peak (TCoG) corresponds to:

$$TCoG = \frac{\sum_{i=1}^{i=n} F0_i t_i}{\sum_{i=1}^{i=n} F0_i} \quad (1)$$

Where i=1 may correspond to the onset of the pitch accented syllable and i=n may correspond to its offset. According to this model, the time point at which the High pitch accent is perceived to occur corresponds to the average location of high f0 values. Within this account, the perception of a tonal event depends on some knowledge of the location in time of the onset and offset of the current syllable and of its position in the metrical structure. A more critical problem of this conception of intonation perception comes from the observed asymmetries in the effects of f0 shapes. Indeed, the perception of tonal alignment is less sensitive to changes in the rise preceding a high tone than to changes in the following fall [3]. This behavior is not reproduced with the TCoG model. In the present study we overcome these limitations while maintaining the core intuition that the effects of the shape of the f0 curve on the perception of tonal alignment are due to the perceptual integration of f0 values. More precisely, we propose that they are due to the way the perceptual system integrates information over time. Our hypothesis is that features of the f0 curve affect the unfolding over time of the perceptual process and, as a consequence, anticipate or delay the perception of the f0 peak. To test this hypothesis, we propose a minimal model of intonation perception that, based on very general and largely shared hypotheses on the functioning of perceptual processes, naturally reproduces the behaviors observed in the literature.

2. Dynamics of intonation perception

The perception of any external stimulation is not an immediate event but is the outcome of a complex process unfolding over time. This consideration is made explicit in several models of perceptual choice, in which the state of the perceptual system is represented by the value(s) of one or more variables that start changing in response to a stimulation and eventually reach states that correspond to different choices. Both the stimulus feature and the initial value of the variable condition the path it follows over time [12, 13]. The simplest version of such models is represented by the Drift-Diffusion Model (DDM), originally introduced to study the link between response probabilities and response times in forced choice tasks [14]:

$$dy(t) = wI + c\xi(t) \quad (2)$$

In the DDM, the current state of the perceptual system corresponds to the value of the variable y. Its amount of change

10.21437/SpeechProsody.2020-31

from instant t to $t + dt$, $(dy(t))$ is equal to the difference between the amounts of evidence supporting the two categories (I) weighted by the parameter w and added to a random value with variance c^2 $(c\xi(t))$, which represents random fluctuations affecting the choice process. In this model, the difference between the amounts of evidence in support of the two categories (I, henceforth *evidence bias*) depends on the input signal features. Negative values represent evidence supporting one category while positive values represent evidence supporting the alternative one. Evidence supporting one or the other categories is accumulated over time by the variable y (which will be referred to as the *accumulated evidence*). When y crosses either a positive or a negative threshold, enough evidence is collected in favor of one category, triggering the perceptual choice. In the application of this modelling framework to the perception of intonation, we map the output of the model to the perception of H and L tones (or alternatively of rising and falling movements) and make the hypothesis that, at each instant, the evidence bias depends on the current value of the f0 curve. DDMs have been conceived to model the perception of stimuli presented in isolation. When modeling the perception of a sequence of tonal events, we need to model explicitly the effect of the perception of an intonation category on the perception of the following one. As shown in [12], this can be done by making the accumulation rate (dy) depend on the accumulated evidence (y). Therefore, the choice process should become:

$$dy = wf(I) + \lambda g\big(y(t)\big) + \xi(t), \qquad (3)$$

where $g\big(y(t)\big)$ is a function of the accumulated evidence. The role of $g\big(y(t)\big)$ is that of providing the system with its own internal dynamics that interact with the effect of external stimuli. When $g\big(y(t)\big)$ is the identity function $(g(y(t) = y(t))$, the sign of λ determines if the systems adopt a stable or an unstable behavior. For negative values of λ, in the presence of a non zero external stimulation y will converge toward one between two possible values and stop changing when that value is reached. These values, corresponding to stable system states, are termed *attractors of the system* and are mapped on the possible perceptual categories. Importantly, when no stimulus is present (or more generally, with a null evidence bias), a negative value for λ will gradually bring y back to 0 (its neutral attractor), therefore modulating the effect of the identification of one perceptual event on that of the following ones.

The last choice we need to make to fully define our model concerns the dependency between the input signal I and the current value of the f0 signal. In order to motivate our choice, let's note that the results collected so far on the effect of the f0 shape on the perception of tonal alignment with the contours displayed in Figure 1 can be summarized by making reference to variations in the slope of the f0 movements and to their locations in time. Indeed, early peak perception is correlated: i) with a weaker slope during the rise (due to a high starting point as in Fig 1f, or to a longer rise as in Fig. 1b); ii) with the anticipation of the maximum positive slope (due to the presence of an early plateau as in Fig. 1d, or of a doomed rise as in Fig. 1h); and iii) with the anticipation of the negative slope (due to the presence of a scooped fall as in Fig. 1m). A late peak perception is observed: i) in presence of a scooped rise (Fig. 1l, in which the location of maximum slope is delayed); ii) in presence of a weaker negative slope during the fall (due to a partial fall as in Fig. 1g, or to the longer fall duration as in Fig. 1c); or iii) in presence of a delay in the location of a maximum negative slope (due to the presence of a late plateau as in Fig. 1e, or of a doomed fall as in Fig. 1i).

This interpretation of the data suggests to make the evidence bias of our model depend on the local slope of f0: $df0(t)$. This would lead to the following formulation:

$$dy(t) = wdf0(t) - \lambda y(t) + \xi(t). \qquad (4)$$

At a given instant t, y changes by a quantity equal to the current rate of change of f0 weighted by the sensitivity parameter w, minus the current value of y, weighted by the damping parameter λ, plus ξ which represents perceptual noise. In presence of constant f0, the system resides in its neutral attractor (y=0). When f0 starts changing (i.e., rising), the accumulated evidence starts increasing due to the presence of a positive evidence bias. During the rise, y crosses a positive threshold triggering the perception of the first tonal category.

Such an approach can easily explain why an anticipation of the maximum f0 slope during the rising portion of the contour anticipates the perception of the f0 peak, as the input values that maximally support the perception of the High tone arrive earlier to the perceptual system. However, in order to explain the role of the features of the falling portion of the contour, we need to postulate that the pitch accent is perceived when both an upward f0 movement and a downward f0 movement have been detected (or alternatively after the detection of both the High tone and the following Low tone).

While effects of the right portions of the contour are straightforward consequences of this adjustment (as both late negative slope values and smaller slope magnitude delay the identification of the fall), those of the left portion rely on the systems internal dynamics (characterized by the $-\lambda y(t)$ term) mediating the effect of the rise on the perception of the fall.

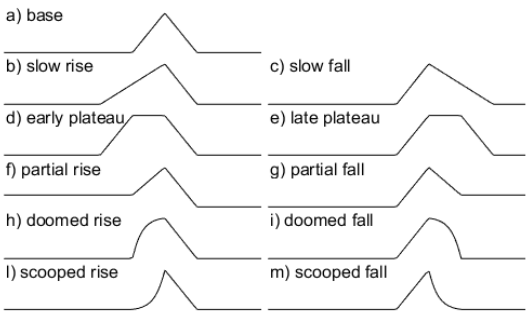

Figure 1. *Stylized symmetric high pitch accent contour (curve labelled a) and variants used in perceptual experiments. All modifications of the rise, made exception for the scooped rise, result in an early pitch accent perception. All modifications of the fall, made exception for the scooped fall, result in a late pitch accent perception,*

3. Simulations

Figures 2 and 3 display the results of several simulations based on the model in Eq. 4. In each simulation, we provided as input to the model variants of the stylized f0 contour represented by the dotted line in the upper panel of Fig. 2a (see figure legend for details). Each variant was meant to reproduce a modification of the f0 curve whose effect has been reported in the literature. In the simulations, each time step on the x axis corresponds to an update cycle of y, whose values displayed in red are obtained

by feeding Eq. 4 with the current value of the evidence bias (as displayed by the bottommost panel of each labeled pair).

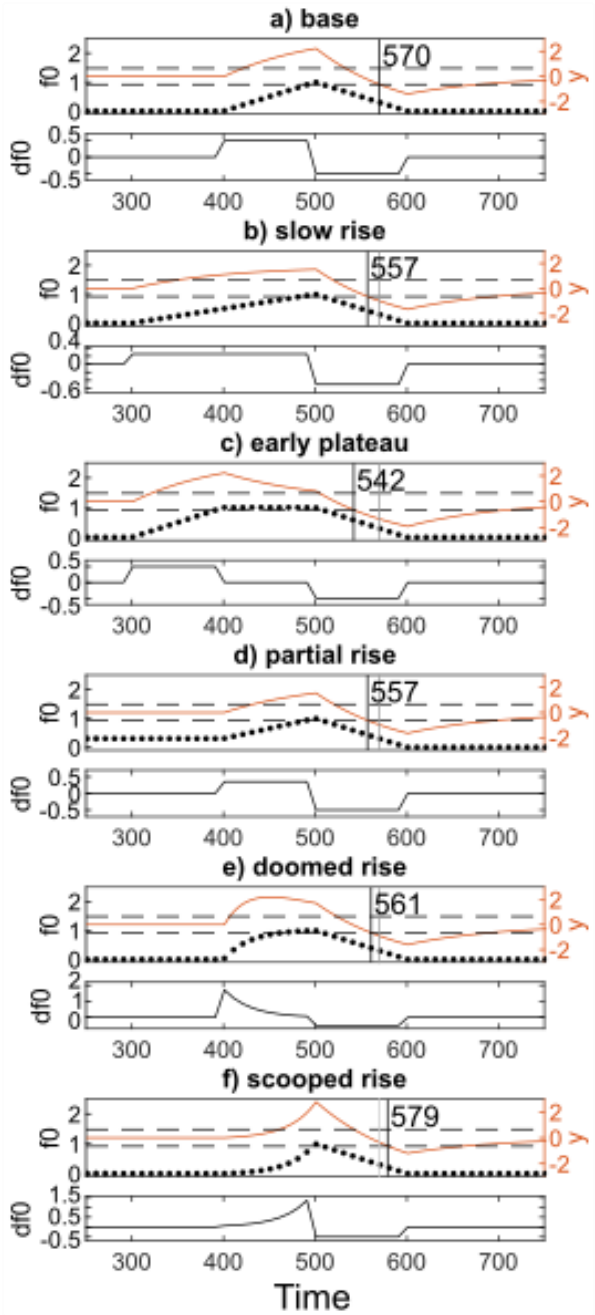

Figure 2 *Simulations conducted with f0 contours showing different rise shapes. Panels are grouped in pairs (one pair per simulation). Topmost panel in each pair displays the f0 curve (dotted line, values in arbitrary units on the left y axis) and the accumulated evidence (continuous red line, values in arbitrary units on the right y axis). Horizontal dashed lines represent the perceptual thresholds. Black vertical lines (and numbers at their side) indicate the time-point when the fall is detected and the f0 peak perceived (that obtained with the baseline f0 contour a) is duplicated in gray in all plots). The bottommost panel displays evidence bias.*

The following values of the models' free parameters were kept constant across simulations: $w = 35$, $\lambda = 0.2$, $\xi = 0$. The perceptual thresholds that, once crossed, trigger the perception of the abstract categories are set to ± 0.4. At each time step the r.h.s. of Eq. 4 is multiplied by the rate parameter $dt = 0.1$. While in all simulations the f0 peak is located at frame 500, the location of the detection point changes due to the shape of the f0 curve.

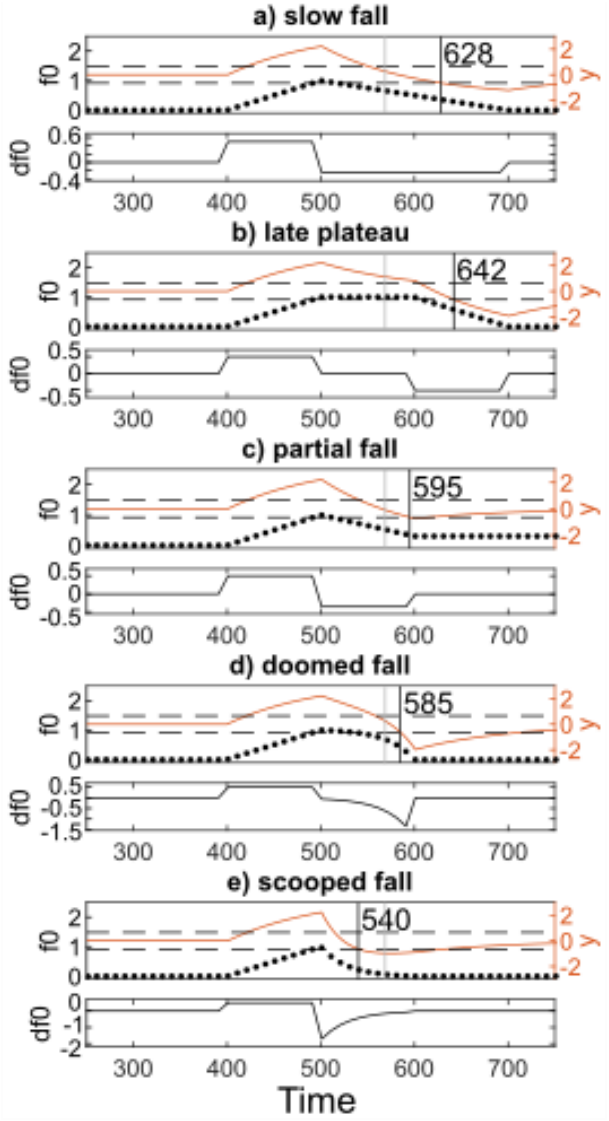

Figure 3. *Behaviour of the model during the presentation of several versions of the basic f0 contour showing variations in the shape of the fall.*

A slow rise (Fig 2b) or a partial rise (Fig 2d) both produce weaker evidence supporting the accumulation of y with respect to a rise with the baseline slope (Fig 2a). As a consequence, the positive attractor moves closer to 0 and y reaches smaller peak values. Therefore, when the evidence bias becomes negative at frames 500, the change in y required to reach the negative threshold is smaller and the threshold is reached sooner (see upper panel of Fig 2b). The responses to shapes displaying an early plateau (Fig 2c) or a doomed rise (Fig 2e) are characterized by the presence of the damping term that brings to zero the accumulated evidence after the vanishing of the supporting stimulus. This term lowers the value of y during the

intervals of time characterized by weak changes in f0. Responses to stimuli characterized by these f0 shapes are then faster due to the shorter distance between accumulated evidence value at the onset of the f0 fall and the negative threshold that triggers the perception of the low target (or of the falling movement). A scooped rise has the opposite effect because the portion of stimulus supporting the presence of a rising movement (or of high f0 values) ends at frame 500, moreover the slope of the f0 is maximum at this time step and the value of y is maximally distant from the negative threshold triggering the perception of the pitch accent.

Panels of Fig. 3 display the effects of modifications of the falling portion of the pitch accent. The slow fall (Fig. 3a) provides a negative evidence bias relatively small in magnitude which slows down the accumulation of negative evidence supporting the detection of the fall, thus delaying the perception of the pitch accent with respect to the symmetric contour in Fig. 2a. The late plateau and the doomed rise delay the occurrence of the evidence supporting the presence of a negative movement, while the scooped fall anticipates it. By comparing the effects of the left and right versions of each change of the shape of the f0 curve, we can see that changes in the falling portion of the curve shift the detection time by a larger amount compared to changes of the rising portion. This occurs because the effects of the rise are mediated by the effect of the damping term and by the effect of the nonzero evidence bias during the falling portion.

4. Discussion and conclusion

In this work we explored the hypothesis that the effect of the stimuli acoustic on the perception of intonation is mediated by the internal dynamics of the perceptual processes. The model proposed implements this basic principle by considering speech perception as a leaky integration process: which is an accumulation process affected by a damping term that modulates the effect of the perception of a tonal category (e.g. a L or an H) on that of the following one. In this way we could successfully reproduce the effects of the shape of the f0 curve on the perception of tonal alignment observed in the literature.

The proposed account also explains the partial success of the TCoG model which results from the accumulative nature of the perceptual drift process (as averaging f0 values corresponds to a sum normalized by the duration of the pitch accent). Contrarily to the TCoG model which maps directly the f0 values on the perception of the pitch accent, the model proposed here relies on two intermediate abstract categories (may they correspond to level tones or to rising and falling movements), whose perception is a prerequisite for the identification of the pith accent. This difference permits capturing the dependency of the shape effects on their location in the pitch accent. Moreover, by modeling the perception of a pitch accent as due to the perception of its abstract features we provide a mean to link the perception of the high pitch accent to that of other tonal events.

The model presented in this study is fully defined by two simple hypotheses. The first hypothesis, implemented by the adoption of leaky integrator model, is that the perceptual system always tends to a stable state and that this is changed by evidence favoring the presence of a perceptual category. The function of the damping term in Eq. 4 is just that of providing the system with a stable neutral attractor. This hypothesis underlies the functioning of many if not all perceptual models that consider the temporal dimension. The second hypothesis is that the evidence bias favoring the perception of tonal categories corresponds to the local change in f0. This feature is required in order to have a null evidence bias during the plateau portions of the pitch accent. However, the fact that the input to the model is a slope does not characterize it as a configuration-based model, because its output is compatible with both level-based and configuration-based phonologies. Moreover, this feature is required only with a one dimensional description of perceptual dynamics (the model relies on a law of change that governs the behavior of one variable). Therefore, it depends on the level of abstraction adopted in our modeling endeavor[1]. It is therefore possible that the behavioral patterns described in this paper can be accounted for by a more complex model whose input is a function of the current f0 value and not of its slope. Still, our findings would remain valid: it is possible to account for the effects of the shape of the f0 curve on the perception of tonal alignment by modeling intonation perception as a process that tends toward an equilibrium state defined by the acoustic input. On the other hand, the fact that at the level of abstraction considered the local slope of f0 plays a major role in the perception of intonation is consistent with the finding that the f0 slope is more important than its scaling in the perception of contrasts based on the alignment of the f0 peak [16][2].

In this study we aimed at showing the relevance of perceptual dynamics in the explanation of the mechanisms underlying intonation perception. Further work must be conducted to understand how perceptual dynamics are affected by additional bottom up factors, as for example the scaling of f0, the amount of spectral change or the amplitude of the periodic component of the speech signal [19]; or by top down factors such as speaker or contextual and social information [20]. One advantage in adopting a low dimensional perspective on modeling is the possibility to enrich the models with terms reflecting in an explicit fashion the effect of meaningful quantities. For example, in order to test potential effects of the amount of the amplitude of the periodic component of the signal, this could be integrated into the model as a multiplicative term (accelerating the accumulation process during portions of signals with high amplitude) or as an additive term (biasing evidence toward the perception of H tones when amplitude is high). Likewise, additive and multiplicative terms can account for the dependency of intonation perception on top-down information.

[1] For example it has been shown that the behavior of several versions of the DDM can emerge in dual choice tasks from that of more complex architectures based on competition dynamics (see e.g.: [15]).

[2] The finding that by introducing a high plateau in a f0 contour and keeping unvaried the target levels induces the perception of a higher f0 (see [17]) may seem a counter example for our account. Indeed in our account an f0 plateau correspond to a portion of input during which the accumulation process can only tend toward the neutral attractor. However, the early plateaus adopted in [17] always followed steep rises, whose effect is that of increasing the accumulated evidence in favor of a high tone.

5. References

[1] Prieto, P. (2011). Tonal alignment. *The Blackwell companion to phonology*, 1-19.

[2] D'Imperio, M. (2011). Prosodic representations. *Handbook of Laboratory Phonology*. Oxford: Oxford University Press (section on tonal alignment).

[3] D'Imperio, M., & House, D. (1997). Perception of questions and statements in Neapolitan Italian. In *Fifth European Conference on Speech Communication and Technology*.

[4] Gósy, M., & Terken, J. (1994). Question marking in Hungarian: timing and height of pitch peaks. *Journal of Phonetics*, *22*(3), 269-281.

[5] D'Imperio, M. (2000). *The role of perception in defining tonal targets and their alignment* (Doctoral dissertation, The Ohio State University).

[6] Niebuhr, O. (2007). The Signalling of German Rising-Falling Intonation Categories-The Interplay of Synchronization, Shape, and Height. *Phonetica*, *64*(2-3), 174-193.

[7] Niebuhr, O. (2003). Perceptual study of timing variables in F0 peaks. In *Proc. 15th ICPhS, Barcelona* (pp. 1225-1228).

[8] Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. (2010). The effect of global F0 contour shape on the perception of tonal timing contrasts in American English intonation. In *Speech Prosody 2010-Fifth International Conference*.

[9] Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. (2019). The interaction of timing and scaling in a lexical tone system: an example from shilluk. Proceedings of ICPHS 2019.

[10] Dorokhova, L., & d'Imperio, M. (2019). Rise dynamics determines tune perception in french: the case of questions and continuations. In *International Congres of Phonetic Science ICPhS2019*.

[11] 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.

[12] Tuller, B., Case, P., Ding, M., & Kelso, J. A. (1994). The nonlinear dynamics of speech categorization. *Journal of Experimental Psychology: Human perception and performance*, *20*(1), 3.

[13] Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, *113*(4), 700.

[14] Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological review*, *92*(2), 212.

[15] Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, *108*(3), 550.

[16] Rathcke, T. (2006). A perceptual study on Russian questions and statements. *AIPUK*, *37*, 51-62.

[17] Knight, R. A. (2008). The shape of nuclear falls and their effect on the perception of pitch and prominence: peaks vs. plateaux. *Language and Speech*, *51*(3), 223-244.

[18] Warren, P. (2017). The interpretation of prosodic variability in the context of accompanying sociophonetic cues. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *8*(1).

[19] House, D. (1996). Differential perception of tonal contours through the syllable. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96* (Vol. 4, pp. 2048-2051). IEEE.

[20] Portes, C. & German, J. S. (2019). Implicit effects of regional cues on the interpretation of intonation by Corsican French listeners. *Laboratory Phonology: Journal of the Association for Laboratory Phonology 10(1).*