Perceptual Foundations for Naturalistic Variability in the Prosody of Synthetic Speech

Nanette Veilleux[1], *Jonathan Barnes*[2], *Alejna Brugos*[2], *Stefanie Shattuck-Hufnagel*[3]

[1]Computer Science and Informatics, Simmons College, Boston, MA, USA
[2]Romance Studies & Applied Linguistics, Boston University, Boston, MA, USA
[3]Research Lab of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA

veilleux@simmons.edu, jabarnes@bu.edu, abrugos@bu.edu, sshuf@mit.edu

Abstract

Recent studies have shown that the Tonal Center of Gravity is a better classifier than F0 Turning Points for at least two contrastively timed pitch accents in American English intonation contours. Within this framework, a binary F0 weighting function derived from the F0 contour can be used instead of the natural F0 contour without a degradation in discrimination performance. This success has important implications for speech synthesis. Just as we can capture the functional equivalence of a multitude of auditorily distinct F0 contour shapes in terms of their mapping to a single parameter (the TCoG) via a set of binary weighting functions, this same mapping could be run in reverse as a source to generate natural-sounding variability in speech synthesis.

Index Terms: Tonal Center of Gravity, F0 alignment, pitch accent classification, prosody, speech synthesis

1. Introduction

Prosody, or the suprasegmental aspects of human speech, serves to convey (among other things) emphasis and phrasing to listeners. According to the Autosegmental-Metrical theory, intonational contrasts are represented as sequences of high and low tone specifications, implemented as pitch level targets precisely timed with respect to the segmental content of the utterance ([1],[2],[3]) . This is true equally of the pitch accents associated with perceptually salient syllables, and the boundary tones that contribute to the partitioning of the utterance into prosodic phrases. Alignment of F0 events in spoken utterances allows listeners not only to determine which word is accented, but also to categorize what type of accent is intended by the speaker. For example, a high pitch realized during a primary stressed syllable might implement a High Pitch Accent (H* in the ToBI annotation system), and convey the perceptual impression of emphasis on the target word. Given the limited inventory of tonal elements, the vocabulary of pitch accents is small, consisting of combinations of high and low F0 targets. For example, both the L+H* and L*+H bitonal accents are produced with an F0 relatively low in the speaker's range that rises to a relatively high F0 level somewhere in the vicinity of the accented syllable. In this case, the alignment of the F0 targets with respect to that accented syllable is crucial to discriminating the pitch accent type, with the F0 rise for L*+H occurring significantly later than the rise for L+H*.

The line of research described here began as an inquiry into how differences in alignment of the nuclear pitch accent that contribute to linguistically contrasting meanings in a rise-fall-rise intonation contour in American English could be reliably recovered by listeners in natural speech. Much earlier work in this area takes the alignment of turning (inflection) points in the F0 contour to be a critical phonetic manifestation of this kind of phonological contrast ([4],[3],[5]). However, the precise location of F0 turning points can be quite unreliable, owing to segmental perturbations, or other ambiguities typically found in the speech signal [6]. This led to the conjecture that a more holistic or global parameter could prove both more reliably recoverable, and potentially perceptually relevant as well. Subsequent experiments, described in more detail below, showed that a point in time representing the tonal center of gravity (TCoG), that is, an F0 weighted average over the duration of the accented syllable (See Equation 1 and Figure 1) was a better classifier of production data, more robust to Gaussian noise [7] and had greater explanatory power for perceptual data [6], than either the peak location or the beginning of the rise (elbow) location (or both). The TCoG, therefore, is a compact and perceptually salient parameter that consolidates various mutually enhancing factors that contribute to perceived tonal alignment, including not just turning point locations, but the broader shape (e.g., domed vs. scooped, among other things) of the intonational contour as well.

$$TCoG = \frac{\sum_i f0_i \times t_i}{\sum_i f0_i} \quad (1)$$

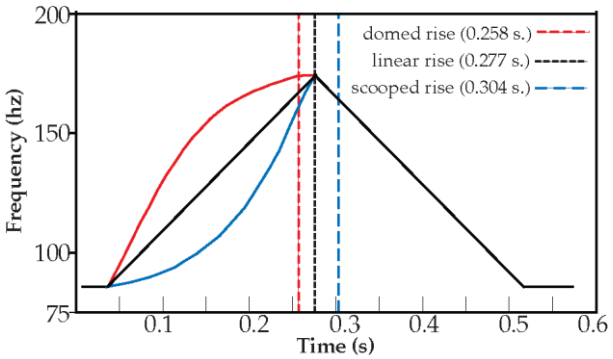

Figure 1*: Illustration of TCoG for scooped, straight line and domed F0 contours for the same turning point alignments.*

While these TCoG results for both the production and perception experiments have been discussed elsewhere ([6],[7],[8]), a secondary finding is worth highlighting and is the subject of this paper. That is, the TCoG model not only does not degrade, but in fact performs slightly better in discriminating contrasting F0 contours when a sigmoid shaped warping function is applied to the F0 contour before calculating the TCoG (Equations 2 and 3). As shown in Figure 2, the resulting simplified F0 contour is essentially binary, such that all high F0 values contribute equally in locating the TCoG (factor ≈1), while low F0 values contribute barely at all (factor ≈ 0). Critically for our purposes here, note that any F0 contour that shares the same F0 range and midpoint intercepts, regardless of its precise shape or turning point locations, will

be warped to the same sigmoid function. This mapping of a family of auditorily distinct F0 contours to a single sigmoid for purposes of the derivation of TCoG makes strong, but ultimately correct predictions regarding the functional or linguistic equivalence of the contours in question.

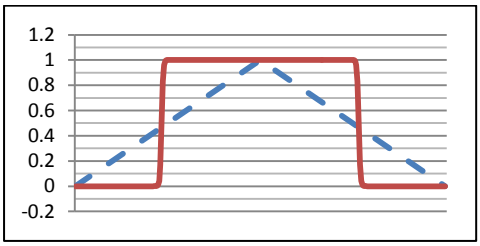

Figure 2: *The effect of applying a sigmoid warping function (solid line) to a linear rise-fall F0 (dashed). The original F0 is scaled for comparison.*

The paper is organized as follows: a short description of a production and a perception experiment used to investigate the TCoG (described in greater detail in Barnes et al., forthcoming) is followed by a description of the sigmoid filter and a discussion of the performance benefits resulting from the use of the filter. The paper concludes with a discussion of the implications of this result.

2. Background

2.1. Experiments

In order to test the perceptual salience of the TCoG, it is useful to find two contours that differ primarily by the alignment of a rise-fall shaped pitch accent. Contours of this type have been studied extensively in numerous languages, including English [9], Italian [10] and German [11],[12], where categorically distinct timing patterns correlate with a categorical shift in meaning. In American English, Ward and Hirschberg ([13] *et seq.*) have identified two contours they called Incredulity and Uncertainty, both implemented with a rise-fall-rise contour. Although Ward and Hirschberg describe the Incredulity contour as simply a more strongly cued Uncertainty contour (greater F0 excursion and amplitude) with the same alignment (L*+H), Brugos and colleagues [14] found that at least some speakers use a contrast of alignment: L*+H pitch accents for Uncertainty and L+H* for Incredulity (see Figure 3).

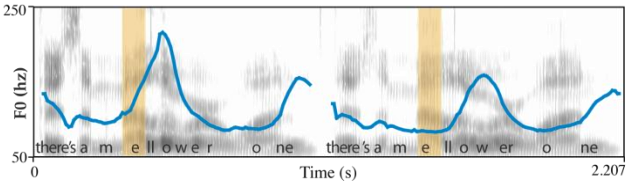

Figure 3: *L+H* (Incredulity) and L*+H (Uncertainty) respectively, on two renditions of /There's a mellower one/. The pitch-accented vowel is highlighted.*

Using this contrastive pair, the discrimination power of the TCoG parameter was compared to that of F0 turning points in two experiments. In a production experiment, the TCoG and turning points were used to categorize utterances elicited in two discourse contexts; in a perception experiment, the predictive power of the TCoG parameter in regression analysis was examined with respect to subjects' category judgments of F0 contours that spanned the continuum between the two contours.

2.1.1. The production experiment

In order to elicit the two test contours, Uncertainty and Incredulity, 6 subjects were asked to repeat target sentences embedded in short discourse scenarios. Target words composed of sonorant segments were used in order not to disadvantage the turning point model with missing F0 intervals. In a crossword completion context, subjects were asked to respond to two types of discourse prompts: first, where a confederate reported an outrageously inappropriate suggestion by a fictitious "Bob" and second, when the confederate asked for a suggestion given some partial list of constraints. In the first situation, the subject was prompted to respond with Incredulity; in the second the subject, aware of possible alternatives, would respond with Uncertainty.

a) Incredulity:
Speaker A: *When I asked for a 10-letter word for "sweeter", Bob said "There's 'lemonier'".*
Speaker B: *"There's 'lemonier'"!? That's absurd!*

b) Uncertainty:
Speaker A: *Can you think of an 8-letter word for "more citrusy"?*
Speaker B: *There's "lemonier". Would that work?*

The task elicited 116 recorded productions, with each of 29 targets repeated 4 times in counterbalanced order over two sessions, one for each context. The beginning of the F0 rise, the peak, the end of the fall and segmental boundaries were hand labelled. Praat [15] was used to extract the F0 contours and the TCoG in the region from the beginning of the rise to the end of the fall was calculated for each utterance (again, Equation 1). Interestingly, error analysis suggested that the contributions of long, low F0 intervals were skewing TCoG in perceptually unsupported directions. To address this problem, a modified TCoG ($TCoG_s$ here) was calculated by altering the weights from the F0 contour by a normalized sigmoid function:

$$f0_{si} = \frac{1000}{998\, e^{\left(-6.9N\left(2\left(\frac{f0_i - f0_{min}}{f0_{max} - f0_{min}}\right)\right)-1\right)}} \quad (2)$$

$F0_{max}$ and $F0_{min}$ are the maximum (peak) and minimum F0 values over the accented syllable and N=16. The modified F0 produces a square weighting function (Figure 2, above) where F0 values above the midpoint of the rise contribute a weight of approximately 1 and those below the midpoint value weigh approximately 0.

$TCoG_s$ is calculated in the same way, using the sigmoid warped F0 weights:

$$TCoG_s = \frac{\sum_i f0_{si} \times t_i}{\sum_i f0_{si}} \quad (3)$$

The TCoG, the $TCoG_s$ and a model using the Turning Points (peak and rise elbow) were used to classify the elicited productions. Correct categorization was based on the elicitation context (Incredulity vs. Uncertainty) and results are summarized in the section 3 (Results).

2.1.2. The perceptual experiment

A perceptual experiment was also conducted to determine the extent to which contour shape differences which the TCoG

model predicts to interact perceptually with alignment differences did in fact have the predicted effects on listeners' categorical judgments. A neutral (deaccented) production of the phrase (*There's a lemonier meringue*) was extracted from a longer production with a pitch accent before the target region (*BOB said there's a lemonier meringue*), and resynthesized to produce a series of manipulated F0 contours. The contours differed in two ways: 1) the shape of the rising portion of the rise-fall pitch accents was varied in 7 degrees, from scooped through a neutral straight line, to domed and 2) all seven shapes were presented in a 7 step sliding alignment continuum. (See Figures 4 and 5)

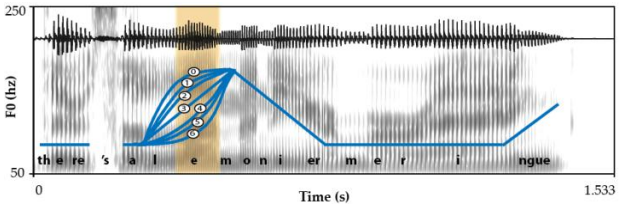

Figure 4: *Continuum of 7 shapes of the rise, from most domed (0) to most scooped (6), with intermediate straight-line rise (3). The pitch-accented vowel is highlighted in yellow.*

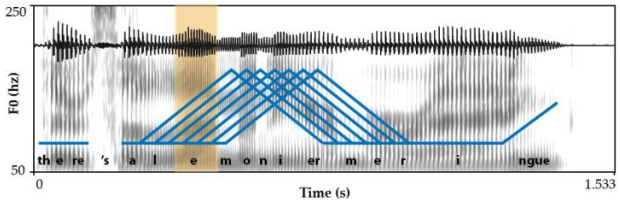

Figure 5: *Continuum of 7 alignments, shown for the straight-line rise shape. All 7 shapes were shifted on this continuum.*

This method produced 49 stimuli with a range of TCoG and TP alignment values varying from early to late, with respect to the pitch-accented syllable. Nineteen subjects took part in an ABX experiment where A and B were unambiguous Uncertainty (L*+H) and Incredulity (L+H*) exemplars, balanced in order of presentation. (See [8] for details.)

As illustrated in Figure 1, the shape of the contour shifts the position of the TCoG (and the $TCoG_s$, see Figure 6) for each stimulus. Peak location, however, is the same for each shape and shifts only with the 7 alignments. If turning point locations alone determine listeners' perceptions, all shapes contrast, TCoG positions will be distinct for each shape within a given alignment, since domed rises naturally pull $TCoG_s$ should sound functionally the same within each alignment. In further to the left (earlier), while scooped rises analogously push it rightward. Figure 6 illustrates the effect of applying the sigmoid warping function to the F0 contours.

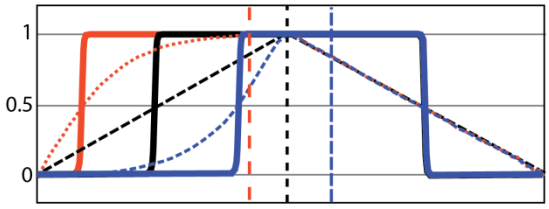

Figure 6: *F0 weights used in the calculation of the $TCoG_s$ for 3 shapes of the continuum: scoop 0 (blue), line 3 (black), and dome 6 (red)., Actual F0 contours shown in dashed lines (displayed normalized for comparison), sigmoid-warped F0 in solid lines, and $TCoG_s$ locations shown as vertical bars.*

3. Results

3.1. Production study results

The experiments discussed here were designed to compare the TCoG with TP models and results have been presented elsewhere [6]. To summarize, in logistic regression models using various acoustic/auditory cues predict pitch accent category membership, both TCoG models were better classifiers of elicited contours of than a TP-only model. Now, since both turning point alignment, and a range of aspects of contour shape have been shown to be critical in determining the position of TCoG in the utterance, it might be expected that a sigmoid warping function such as that used in $TCoG_s$, that essentially removes all trace of both original contour shape and TP locations, would be strongly detrimental to the performance of the model.. Suprisingly, however, Table 1 shows a slight enhancement of the $TCoG_s$ model over the TCoG model. From this and other evaluations, it appears that the simplified binary weights, derived from the elicited utterance's F0, are sufficient to capture the relevant differences between the two contours.

	% correct	
Observed Contour	TCoG	$TCoG_s$
L*+H	94.5	96.1
L+H*	90.7	92.5
total	92.5	94.3

Table 1: *Production experiment results. The $TCoG_s$ performed slightly better despite simplified binary F0 weighting function.*

3.2. Perception study results

In the perceptual study, the location of the peak (which is fixed with respect to other turning points here) and both TCoG values were used as predictors in a logistic regression model to predict subject responses to the ABX task. Both TCoG models performed better than a turning-point-only model. Using $TCoG_s$, derived using the simplified binary F0 weights, as a predictor in place of standard TCoG again did not degrade performance, but instead improved it slightly. Table 2 shows the results of the logistic regression analysis when the two TCoG parameters were used to model subjects' perceptions of the 49 presentations (7 contour shapes, aligned at 7 positions with respect to the accented syllable).

	Model Chi-square	Nagelkerke r^2	
TCoG	1428	.436	$p < .001$
$TCoG_s$	1459	.443	$p < .001$

Table 2: *Perceptual experiment results: Again, $TCoG_s$ results in slightly better performance compared to TCoG.*

4. Discussion

The sigmoid filter was originally motivated by the observation that long, low F0 movements preceding or following the peak of an accented syllable tended to pull the TCoG to the "left" or "right" respectively, despite a complete absence of perceptual evidence for this shift. This furthermore resulted in frequent miscategorizations of contours by the regression model described above. While several potential explanations for this phenomenon are currently under investigation (including the nature of the perceptually salient interval and the phonetic characteristics of the sound segments of the words [16], a

variety of observations from the literature on tone perception suggest that listeners may not be attuned to low F0 in the same way that they are to high F0. For example, d'Alessandro et. al review a range of sources for the idea that high F0 is more salient perceptually than lower F0, and find in their own results that, in modelling listeners' judgments regarding the scaling of tonal glissandos realized on synthetic renditions of the vowel /a/, it was necessary to assume that the higher portions of each glissando had more effect on listener perceptions of glissando endpoint scaling than lower portions, regardless of the direction of the glissando's movement (ascending or descending) [17].

Likewise, our results suggest that, at least for purposes of the perception of tonal alignment, subjects are less concerned with the exact positioning of the turning points, or even with the precise overall shape of the F0 contour, but instead, appear sensitive primarily just to the interval over which F0 is raised above a certain midline. It was sufficient, in other words, just for high F0 to be sustained long enough, or positioned correctly, for TCoG to be centered within a timing window appropriate for the intended contrasting alignment pattern (medial or late).

The small improvements observed in classification performance do not seem, on the face of it, particularly useful. However, if one considers this result in terms of parameter parsimony, it suggests that only one parameter is necessary to recover the intended alignment (but not, obviously, the F0 contour itself): the TCoG. Moreover, equivalent TCoG values could be derived from any F0 contour that turns "on" (rises above the midpoint) and "off" (falls below) at times that place the TCoG point appropriately, even though the original F0 contours are auditorily distinct. This flexibility generates a family of F0 contours that equivalently map to the same intended intonational category. With the same TCoG, and constrained by factors including physiological constraints on production and desired scaling effects, the F0 contour can be realized with a relatively free range of shapes and turning point locations. For example, the stylized F0 shapes shown in Figure 7 illustrate just a few of the variety of possible F0 contours that have the same $TCoG_s$. These F0 contours would be auditorily discriminable from one another, but, by generating the same weighting function and therefore the same $TCoG_s$, would be functionally equivalent in terms of their phonological category membership.

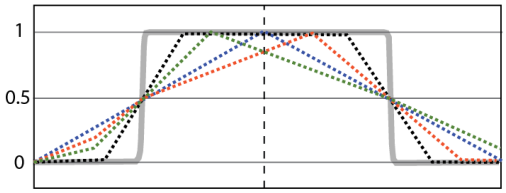

Figure 7: *Stylized F0 contours (dotted lines) that map to the same binary weighting function and $TCoG_s$ (vertical dashed line).*

This aspect of the results is significant because, despite enormous improvements in the quality of synthetic speech, listeners still detect a lack of naturalness in the prosody (and occasionally in its segmental aspects, many of which could be arguably linked to prosody). Although concatenative approaches currently produce more natural-sounding synthetic speech, it is not practical for synthesis applications to retrieve pre-existing F0 contours for every possible accent type on any arbitrary syllable (e. g. [18]). Moreover, the repetition of even a faithful reproduction of a canonical F0 contour across all accented syllables in synthesized utterances will be perceived as unnaturally regular. Humans clearly exhibit great variability in implementing pitch accents and natural sounding synthetic speech should as well. The results presented here suggest which aspects of this variation are perceptually fixed and which can be freely implemented with little impact on perception of accent type. Armed with this information, it is possible to inject variability into F0 contour synthesis by adding random but functionally equivalent variation.

5. Acknowledgements

This work was supported by NSF grant numbers 0842912, 0842782 and 0843181.

6. References

[1] Pierrehumbert, J.B., *The Phonetics and Phonology of English Intonation*, PhD thesis, MIT, 1980.

[2] Pierrehumbert, J. & Beckman, M., *Japanese Tone Structure* (Linguistic Inquiry Monograph 15), Cambridge: MIT Press, 1988.

[3] Ladd, D. R., *Intonational Phonology* (2nd ed.). Cambridge: Cambridge University Press, 2008. Original edition, Cambridge: Cambridge University Press, 1996.

[4] Bruce, G., *Swedish word accents in a sentence perspective*. (Travaux de l'Institut de Linguistique de Lund 12.) Lund, Sweden: CWK Gleerup, 1977.

[5] Arvaniti, A., Ladd, D. R. & Mennen, I., "Stability of tonal alignment: the case of Greek prenuclear accents", *Journal of Phonetics* 26(1): 3-25, 1998.

[6] Barnes, J., Veilleux, N., Brugos, A. & Shattuck-Hufnagel, S., "Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology", *Laboratory Phonology*, Forthcoming.

[7] Veilleux, N., Barnes, J., Shattuck-Hufnagel, S. & Brugos, A., "Perceptual Robustness of the Tonal Center of Gravity for Contour Classification", Workshop on Prosody and Meaning, Barcelona, 2009.

[8] Barnes, J., Veilleux, N., Brugos, A. & Shattuck-Hufnagel, S., "The effect of global F0 contour shape on the perception of tonal timing contrasts in American English intonation", *Speech Prosody 2010*, 100445: 1-4, 2010.

[9] Pierrehumbert, J., & Steele, S., "Categories of tonal alignment in English", *Phonetica* 46: 181-196, 1989.

[10] D'Imperio, M., *The Role of Perception in Defining Tonal Targets and their Alignment*, PhD thesis, Ohio State University, 2000.

[11] Kohler, K. J., "Categorical pitch perception", *Proceedings of the 11th International Congress of Phonetic Sciences*, 5: 331-333. Tallinn, 1987.

[12] Niebuhr, O., "The signalling of German rising-falling intonation categories —the interplay of synchronization, shape, and height", *Phonetica* 64: 174-193, 2007.

[13] Ward, G., & Hirschberg, J., "Implicating uncertainty: The pragmatics of fall-rise intonation", *Language* 61(4):747-776, 1985.

[14] Brugos, A., Barnes, J., Shattuck-Hufnagel, S. & Veilleux,N., "(At least) two members of the rise-fall-rise family", Poster presented at Experimental and Theoretical Advances in Prosody, Ithaca, NY, 2008.

[15] Boersma, P. & Weenink, D. 2009. Praat: doing phonetics by computer. (Version 5.1). http://www.praat.org

[16] Barnes, J., Brugos, A., Veilleux, N. & Shattuck-Hufnagel, S. "Modelling the perception of English F0 scaling in a segmental context", Presentation at Experimental and Theoretical Advances in Prosody 2, Montreal, Canada, 2011.

[17] d'Alessandro, C., Rosset, S. & Rossi, J.-P., "The pitch of short-duration fundamental frequency glissandos", *Journal of the Acoustical Society of America*, 104(4): 2339-2348, 1998.

[18] Campbell, W. N., Wightman, C. "Prosodic encoding of syntactic structure for speech synthesis", In *International Conference on Spoken Language Processing-1992*: 1167-1170, 1992.