Perception of Pseudoswedish Tonal Contrasts by Native Speakers of American English: Implications for Models of Intonation Perception

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

[1] Boston University, Boston, Massachusetts, USA
[2] Simmons College, Boston, Massachusetts, USA
[3] Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
jabarnes@bu.edu, abrugos@bu.edu, veilleux@simmons.edu, sshuf@mit.edu

ABSTRACT

Results from an experiment on perception of lexical pitch accents in Gothenburg Swedish (11) showed that both higher initial F0 plateaux and later falls biased native listeners toward the Accent 2 category, while lower plateaux and earlier falls biased toward Accent 1. Segerup and Nolan interpret this as evidence for a perceptual mechanism integrating scaling and timing information, which they model using a measure of Area under the F0 Curve. It is not clear, however, that this result generalizes beyond native listeners. An analogous experiment on American English listeners demonstrates the same perceptual biases, suggesting that the mechanism responsible is indeed a more general phenomenon, and that it involves the integration of pitch information over time during some region of interest. An alternative to the AUC model, based on the Tonal Center-of-Gravity, could also account for these results, though current data do not distinguish between the two models.

Keywords: tone, perception, pitch-timing interactions, cue-trading, Swedish pitch accents.

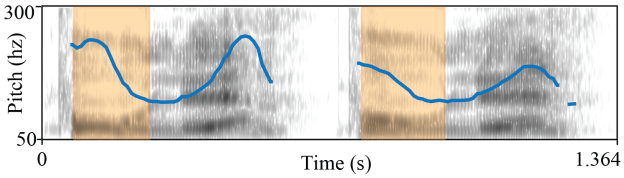
1. INTRODUCTION

Standard phonological analyses of tone in the autosegmental era (4, 9) treat tonal timing and tone scaling in the frequency domain as orthogonal aspects of tonal representation and implementation. That is, the identity of tone specifications (H, M, L, etc.) is logically and formally distinct from the mapping of those specifications to Tone Bearing Units in the segmental string. The notion that, e.g., a High tone is in some sense representationally the same, whether it is associated with one or with more host TBUs has yielded many valuable insights into the structure of tone systems over the years[1].

From a phonetic point of view, by contrast, pitch and time are not so easily separated. A range of well-known phenomena illustrate this interplay, including the influence of duration on the so-called "glissando threshold" (6), the *tau* and *kappa* effects (12)[2], and the connection between plateau duration for a High pitch accent and perceived scaling of same (6, 8).

How these facts should be understood in the context of phonological contrast and its representation remains unclear. Perhaps the most dramatic demonstration of a link to the phonological is the contrast documented by Segerup & Nolan (hereafter SN [11]) between traditionally designated Accent 1 and Accent 2 in the Gothenburg dialect of Swedish. In Stockholm Swedish, the distinction between these two lexical accent patterns can be understood in the Autosegmental-Metrical tradition as fundamentally a contrast in tonal timing, such that the same HL sequence is associated either earlier or later with respect to the accented syllable.[3] In Gothenburg, on the other hand, a scaling dimension has developed as well: the accent with the later peak (Accent 2) tends to be realized higher, while the accent with the earlier peak (Accent 1) tends to be realized lower. (10) documents a degree of individual variation in the realization of the Gothenburg contrast that suggests a trading relation between the timing and the scaling cues, such that addition of the one may be seen to compensate for absence or weakness of the other. A characteristic contrasting pair from a single speaker, displaying differences in both cues, can be seen in Figure 1.

Figure 1: Gothenburg Swedish Accent 1 (lower, earlier F0 peak on the accented initial syllable) and Accent 2 (higher, later F0 peak on the same syllable). (Accented vowels in orange.)



SN address this question with a perception study, in which subjects are presented with a range of contours realized on the disyllabic tonal minimal pair *Polen 'Poland'/pålen 'the pole'*, all beginning with a high plateau of some duration, followed by a fall somewhere in the accented initial syllable, and then a rise-fall, modelled after the patterns seen in the natural productions given in Figure 1. Six

distinct alignments of the initial fall were created, from early to late, and each alignment was realized at three different scalings for the plateau preceding the fall, from higher pitch to lower. Subjects were asked to identify each stimulus as either *Polen* (Accent 1) or *pålen* (Accent 2). Results showed that both later falls and higher initial plateaux biased listeners toward Accent 2, while earlier falls and lower plateaux biased them toward Accent 1.

What this means in terms of the nature of the contrast, whether it is fundamentally timing, or scaling, or somehow both, is not entirely clear. SN suggest an interpretation predicated on the notion that, instead of extracting independent targets in the frequency and time dimensions for the pitch accents in question, listeners may instead rely on an integration of F0 information gathered over time. In particular, SN argue that Accent 2 is cued by maximizing the area under the F0 curve during the accented syllable (achieved either by raising the plateau's maximum F0, or extending the plateau's duration). Accent 1 would then involve minimizing this same AUC value, by realizing lower peaks and earlier falls.

This interpretation is intriguing, and has wide-ranging ramifications for our understanding of tonal contrast and tone perception, but alternative accounts of the result nonetheless seem possible. For example, given the inherited phonological distinction between HL* and H*L, coupled with a waning salience of the timing difference in this dialect, some Gothenburg speakers might have begun to enhance the contrast by exaggerating a scaling distinction implied by the identity of the starred tone (High for Accent 2, Low for Accent 1). Listeners in the SN perception test would then identify both higher and later accents with Accent 2 words not because these two dimensions of the signal were interacting perceptually (as an IPP, à la [7], for example), but rather just because their experience of this lexical contrast in their native dialect happened to involve both these features: Accent 2 words would indeed tend to have peaks that are either higher, or later, or both. The SN result would then just reflect perceptual strategies adopted by speakers of a particular dialect at a particular stage in its development, rather than anything more general about the nature of tone perception. We might predict, then, that the SN result would fail to generalize beyond Gothenburgers to listeners with different linguistic backgrounds (i.e. whose perceptual experience did not include this particular confluence of cues).

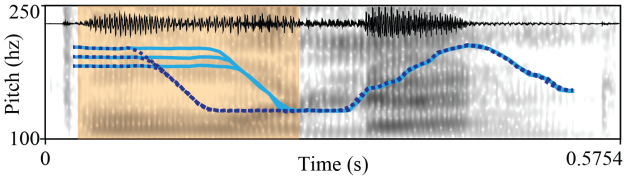
The current paper presents an argument, based on the perception of a Gothenburg-like tonal contrast by speakers of American English, in favour of the SN analysis, suggesting that the timing/scaling relationship identified by them is indeed more basic and more general than alternative analyses predict.

2. METHODS

Direct replication of SN with English speakers was obviously not possible, given the reliance in that study on speakers' ability to identify each synthetic stimulus with a natively-acquired lexical category. We chose instead to investigate perception of this contrast by testing English speakers' ability to discriminate between synthetic F0 contours similar to those of the Gothenburg Swedish pair. The study thus employed an AXBX matching-to-sample design, in which subjects were presented with two pairs of contours, one consisting of two identical contours, and the other with contours that were subtly different. Subjects were instructed to identify which pair (AX or BX) was the "matching" one.

Stimuli were created from natural recordings of a male speaker of Gothenburg Swedish uttering the words *Polen ('Poland',* Accent 1*)* and *pålen ('the pole',* Accent 2*)*.[4] Our initial intention was to create synthetic versions of these contours that were identical to those used by SN in their perception study. Piloting made it immediately clear, however, that the contours used by SN, designed to be close piecewise linear approximations of natural renditions of the Swedish words in question, were so similar perceptually that American listeners could not discriminate them at all. In order to make the task less difficult, the following changes were made to SN's model for the stimuli: 1) The vocalic portion of the base recordings was synthetically lengthened by a factor of 1.3, making it 215 ms. in duration, 2) the magnitude of each fall was increased, to 70, 60, and 50 Hz respectively for the High, Mid, and Low plateau contours, and 3) the temporal distance between the contrasting falls was increased, such that the High early-aligned fall (Accent 1) began 50 ms. beyond the onset of the accented vowel, while the late-aligned High fall (Accent 2) began 80 ms. later. (Mid and Low falls were separated from one another by the same interval, but began 10 ms. and 20 ms. later than the High fall respectively.)

Figure 2: The 6 f0 contours, overlayed on the spectrogram of the base file (accented vowel in orange). Early-aligned falls are shown in dark blue dashed, and late-aligned in solid light blue.



In order to assess the influence of pitch-accent scaling on the perception of our Pseudoswedish contrast, five distinct trial types were included in the experiment. For the first three, the two contours making up the "different" pair in each AXBX trial were distinguished by timing of the pitch fall alone. In these trials, all contours, including those in the "same" pair, were taken from a single scaling level, yielding for example the following pairings[5]:

Table 1: Example AXBX trials in which the "different" pair is distinguished only by timing.

(HighLate–HighLate) (HighEarly–HighLate)
(MidLate–MidLate) (MidEarly–MidLate)
(LowLate–LowLate) (LowEarly–LowLate)[6]

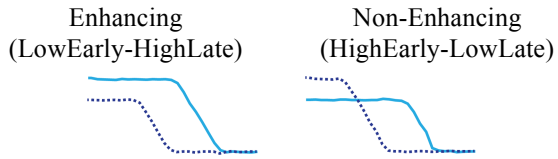
In what follows, these three trial types will be referred to as "timing-only", to distinguish them from the last two trial types, in which the contours represented in the "different" pair differed by scaling of the plateau as well as by timing of the fall. These trial types involving scaling differences are the critical ones for testing the SN hypothesis regarding the interaction of timing and scaling in tone perception. Recall that for SN, what listeners are actually sensitive to in the realization of contrasts such as those of Gothenburg Swedish (or by extension, Bostonian Pseudoswedish) is not pitch movement timing or scaling per se, but instead an integration of these two dimensions, as captured by differences in the area under the F0 curve (AUC) for the two contour types. Given this, our Accent 1 and Accent 2 stand-ins should differ maximally from one another in perception when the scaling difference enhances the AUC difference between the two contours already created by the timing contrast. For example, a late fall will have its already-large AUC increased when realized from a high plateau, while an early fall will have its already-small AUC decreased when realized from a low plateau. This combination, which we will hereafter refer to as "enhancing", serves to maximize the difference between the AUC of the two contours. If, by contrast, the late-fall pitch accent were realized lower, and the early-fall pitch accent higher, then timing and scaling would effectively be working at cross-purposes, to the extent that whatever AUC difference is created by timing is counteracted by maladaptive differences in scaling. We call this later trial type "non-enhancing". Example of both these trial types are given in Table 2, and a schematic of the Enhancing and Non-Enhancing comparison contours is shown in Figure 3.

Our prediction, then, is that if SN are correct in their analysis of the interaction of timing and scaling, then English speakers should be able to distinguish trials of our enhancing type more readily than they can trials of the time-only type. An identical amount of scaling difference, however, when deployed in a non-enhancing fashion, should be at best unhelpful in terms of listeners' ability to discriminate.

Table 2: Examples of AXBX trials incorporating both timing and scaling differences.

Enhancing:
(LowEarly–LowEarly) (LowEarly–HighLate)
Non-Enhancing:
(HighEarly–HighEarly) (HighEarly–LowLate)

Figure 3: Schematic of contour comparisons in enhancing and non-enhancing scaling combination: Dashed dark lines are early aligned, and solid lighter blue lines are late high aligned.



15 college-aged native speakers of American English were participants in this study. The total number of trials per subject was 108. Results of the experiment were as follows.

3. RESULTS AND DISCUSSION

Figure 4 shows the mean proportion of trials in which subjects correctly distinguished the "same" pair of contours from the "different" for the three timing-only trial types. Subjects performed marginally better on trials with high pitch accents than with mid or low, though this difference failed to reach statistical significance. Mixed-effects logistic regression (lme4 package [2]) in R) with trial type as a fixed factor and subject as a random effect shows no difference between high and mid (Wald Z = -0.772, p =0.44). It must also be noted that, on the whole, American listeners performed abysmally on this discrimination task, with percent correct hovering around 60 for all three conditions.

Figure 5, by contrast, shows that, while percent correct for trials with a non-enhancing scaling contrast fails to separate from that of the pooled set of timing-only trials, trials with an enhancing pitch accent scaling contrast yielded a significantly higher rate of correct responses (Wald Z = 5.644, p < .001 in a mixed-effects logistic regression with trial type as a fixed factor and subject as a random effect).

Figure 4: Proportion correct discrimination for trials in which only fall timing distinguished the two contours in the "different" pair.

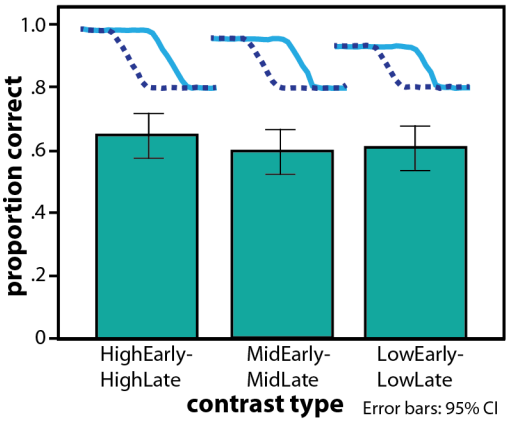
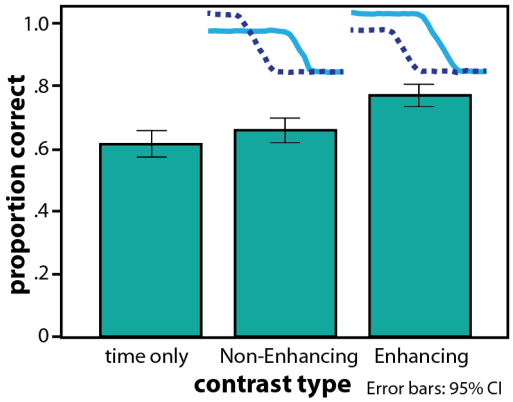


Figure 5: Proportion correct discrimination in timing-only trials compared to that of trials with enhancing and non-enhancing scaling differences.



We take this difference to confirm the conclusions of SN regarding the interaction of timing and scaling in the perception of Gothenburg Swedish. Native speakers of American English displayed the same patterns in perception of the Pseudoswedish tonal contrast as did Gothenburg Swedes in the perception of their native tongue. Because these English speakers had no prior experience of this (or similar) tone contrasts, this pattern of results cannot be attributed either to patterns present in previous language experience, or to the natural realization properties of contours with a particular, language-specific phonological representation. Instead, we would argue that these results can be best understood as the result of a more general property either of linguistic tone perception, or of auditory pitch perception more generally. In particular, categorization of pitch contours in these studies can be seen to involve not the parallel extraction of distinct tonal targets in the scaling and time dimensions, as standard AM phonological analyses might predict, but instead, the integration of pitch information over time during some region of interest (here, the accented syllable).

SN characterize this integration directly in terms of area under the F0 curve, an analysis that is clearly compatible with the data at hand. Another way of understanding these results would be through the conceptually-related mechanism of an average F0, weighted by various perceptually relevant factors, such as sonority of carrier segments, as suggested by (2), for the Tonal Center-of-Gravity approach to perceived tone scaling. In the example at hand, it should be clear that both raising the plateau, and lengthening it relative to the duration of the fall (i.e., making the fall later in the accented syllable) would have the effect of raising the mean F0 of the region in question, as compared with the lower, shorter plateau of the contrasting contour. This approach would also succeed in avoiding certain undesirable predictions of the AUC model (e.g., the predicted perceptual equivalence of 100 ms. of a 200 hz signal and 200 ms. of a 100 hz signal). Unfortunately, the present results do not allow us to distinguish between these two related approaches in terms of their suitability to the phenomena in question, leaving this as a challenge for future research.

These results do, however, make strong predictions about the ways in which we would expect to find scaling and timing cues deployed in the service of rendering tonal contrasts cross-linguistically. To the extent that the Gothenburg pattern is perceptually enhancing, sound changes bringing about this pattern could be seen as adaptive, and should be relatively commonly attested. Sound changes that move contrasting categories in the Anti-Gothenburg, non-enhancing direction, on the other hand, would be maladaptive, resulting in less robust, and hence presumably less commonly attested, patterns of contrast.

3. CONCLUSIONS

While American English listeners without knowledge of Swedish are exceedingly bad at perceiving even exaggerated versions of Gothenburg Swedish tonal contrasts, the mechanisms by which their perception of these contrasts is enhanced, through interaction of tonal target scaling and timing, appear to be the same as those of native speakers of the Gothenburg dialect: Discrimination is improved when these theoretically independent cues conspire to maximize contrasts in both area under the f0 curve and scaling of Tonal Center-of-Gravity. Distinguishing between the explanatory powers of these two models, however, remains a task for future research.

4. REFERENCES

[1] Arvaniti, A., Ladd, D. R. & Mennen, I. 2000. What is a starred tone? Evidence from Greek. In *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, M. B. Broe & J. B. Pierrehumbert (Eds.), Cambridge: Cambridge University Press, pp. 119-131.

[2] Barnes, J.A., Brugos, A., Veilleux, N. & Shattuck-Hufnagel, S. 2014. Segmental Influences on the Perception of Pitch Accent Scaling in English. In *Proceedings of Speech Prosody 7*, Campbell, Gibbon, and Hirst (eds.), pp. 1125-1129.

[3] Bates, D. & Maechler, M. 2009. Lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-32.

[4] Goldsmith, J. 1976. *Autosegmental Phonology*. PhD dissertation, MIT.

[5] Gussenhoven, C. 2004. *The phonology of tone and intonation.* Cambridge: Cambridge University Press

[6] 'Hart, J. 't., 1976, Psychoacoustic backgrounds of pitch contour stylisation. I.P.O. Annual Progress Report 11,11-19.

[7] Kingston, J. and Diehl, R. 1995. Intermediate properties in the perception of distinctive feature values. In A. Arvaniti, & B. Connell (eds), *Papers in laboratory phonology IV*. Cambridge: CUP, pp. 7–27.

[8] Knight R. 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.

[9] Leben, W.R. 1978. The representation of tone. In V. Fromkin, ed., Tone: A Linguistic Survey. New York: Academic Press, 177-219.

[10] Segerup, M. 2004. Gothenburg Swedish word accents: a fine distinction. In Proceedings, FONETIK 2004, Dept. of Linguistics, Stockholm University

[11] Segerup, M. & Nolan, F. 2006. Gothenburg Swedish word accents: a case of cue trading? In: Bruce, G., Horne, M. (eds), *Nordic Prosody: Proceedings of the IXth Conference.* Frankfurt am Main: Peter Lang, pp. 225-233.

[12] Shigeno, S. 1986. The auditory tau and kappa effects for speech and nonspeech stimuli. *Perception & Psychophysics*, *40*(1), 9–19.

---

[1] Leben's famous 1978 analysis of Mende nominal tone melodies, for example, is impossible without this formal distinction between what the tones are and over what domains they are realized.

[2] The *tau effect* is perhaps better known among linguists as a putative manifestation of the "effort code" [5], which connection, to our knowledge, remains unexplored.

[3] Depending, of course, on how we understand "starredness" in this phonological tradition [1]. Since Accent 1 is analyzed as HL*, and Accent 2 as H*L, we could conceivably also see this as a scaling contrast, insofar as these are basically a Low tone (with a High on-glide), or a High tone (with a Low off-glide).

[4] These recordings were created by My Segerup, and formed the basis of the synthetic contours used in SN as well. We are extremely grateful to Francis Nolan and My Segerup for their generosity in sharing these materials with us.

[5] Here and throughout, stimuli will be referred to as, for example, "high late", meaning a high plateau with a late-timed fall, "mid early", meaning a mid-level plateau with an early-timed fall, etc.