Intonational schemas, perceived grouping, and distortions of perceived duration

Alejna Brugos and Jonathan Barnes

Boston University
abrugos@bu.edu, jabarnes@bu.edu

ABSTRACT

Perception of duration is known to be affected by a variety of contextual factors, including pitch. It has also long been observed that rhythmic grouping can affect perceived duration such that intervals between perceived groups are inflated in perception. Pitch and timing cues both play key roles in prosodic grouping. This paper explores the hypothesis that certain pitch-based distortions of time perception are in fact due to perceived grouping effects, and that such interactions can affect speech timing perception. A pair of perception experiments eliciting judgments of perceived grouping and perceived timing used the same set of stimuli, resynthesized with crossed continua of pitch and timing manipulations. Results support a correlation between perceived grouping and distortion of perceived duration of between-group silent intervals.

Keywords: prosody, intonation, prosodic grouping, duration perception, pitch-timing interaction

1. INTRODUCTION

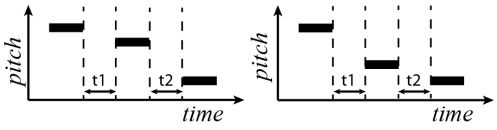
Measures of duration are central to much speech science research, from phoneme categorization to phrase-level prosody. However, human perception of duration can differ from objective duration, and has been found to be systematically distorted in a variety of contexts [8]. Among factors that have been shown to distort perceived duration in the auditory domain are several based on pitch [19]. Pitch factors have been shown to distort perceived duration in both speech [14, 26, 34] and non-speech [17] contexts, with both filled and silent intervals. We continue to know fairly little about how pitch and timing interact in speech perception, in spite of the importance of both to sentence level prosody. This paper examines a set of cross-phrase tonal patterns, and asks how these pitch patterns affect both perceived grouping of an ambiguous multi-phrase utterance, and perception of pause duration. The specific work undertaken here is a pair of perception experiments that investigate the hypotheses 1) that certain cross-phrase pitch patterns play a role in perceived prosodic grouping and 2) that perceived duration can be influenced by perceived grouping.

2. BACKGROUND

2.1. Distortions of perceived duration

A well-documented pitch-cued distortion of duration is the auditory kappa effect [12], an illusion whereby perceived duration of silent intervals is distorted by the relative pitch of filled intervals bounding them: pauses between intervals further apart in pitch sound longer than those of equal duration bounded by closer-pitch intervals (See Fig. 1). This effect has been shown widely in non-speech ([31, 28, 18] *inter alia*) and recently with speech [11].

Figure 1: A schematic example of the auditory kappa effect. Time intervals t1 and t2 are equal, but t2 is perceived as longer at left, shorter at right.



One proposed explanation for the kappa effect is the auditory motion hypothesis; i.e. That pitch-based distortions to perceived duration happen via imputed pitch velocity, based on analogies to perceived motion in physical space [18, 28]. While imputed movement may well explain some pitch-based time distortions, e.g. changes in perceived speed of downward or upward pitch movement, it is less clear how this would apply to speech. While subject to physiological limitations on pitch change, speech has quite variable pitch movement. It is also less straightforward how imputed pitch velocity would explain the results of [13], where a distortion in perceived duration occurred in stimuli that contained a change in direction of pitch movement.

An alternate hypothesis to explain the kappa effect is "the auditory grouping hypothesis," (so-called by MacKenzie [28] and attributed to Bregman [7]), by which items grouped together perceptually (e.g. by a shared feature, such as pitch) will be perceived as closer together in time. This hypothesis squares with results of research on the kappa effect with speech materials [10, 11]; not only did pitch manipulations trigger perceived duration distortion, they more strongly cued prosodic grouping.

Whereas MacKenzie [28] more-or-less rejects the auditory grouping hypothesis based on experiments

with simple tones, a wider literature showing grouping-based distortions to perceived duration can be found in the auditory perception and psychology literature. In fact, the effects of rhythmic grouping on perceived duration in non-speech stimuli have been observed at least since the turn of the last century [6, 29], as well as more recently [16]. In particular, the dilation of perceived duration of between-group silence is a phenomenon called the "duration illusion," and it may occur even in infancy [32]. This all suggests that the auditory grouping hypothesis bears revisiting with respect to pitch-based duration distortions, with a wider range of stimuli, including speech materials.

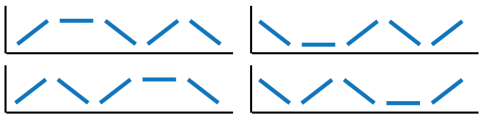
2.3. Prosodic grouping, gestalt principles and schemas

Growing numbers of researchers have suggested that several types of cross-phrase tone and timing cues to prosodic grouping can be characterized by cognitive grouping principles akin to the gestalt principles [33], such as proximity, similarity and continuity [9, 21, 22, 23]. Gestalt grouping principles have long been viewed as playing an important role in auditory scene analysis [7] and music grouping [27]. While results of [9, 10] suggest that tone proximity (e.g. pitch range) and continuity (e.g. phrase-initial reset) play a strong role in grouping, not all tonal contributions to grouping may be reducible to gestalt-like principles.

In work on auditory scene analysis and stream segregation, Bregman [7] maintains a distinction in mechanisms by which grouping can occur as either 1) pre-attentive processes, including gestalt-like principles or 2) more attentive recognition of learned patterns, i.e., schemas. Schemas may likewise play a role in sequential grouping in speech: intonational schemas, potentially language-specific pitch patterns, are here hypothesized to cue prosodic grouping without making direct reference to gestalt-like principles such as proximity and continuity.

In an exploration of pitch cues to grouping in Swedish, House [20] manipulated the pitch of sequences of digits in Swedish, leaving timing neutral: listeners indicated whether a sequence of repetitions of a digit was grouped as 55-555 ("2-3") or 555-55 ("3-2"). While patterns that reliably cued grouping included those that can be attributed to pitch proximity and continuity, several patterns strongly cued grouping that are not straightforwardly reducible to these. These groupings appear to have been demarcated by tonal shapes forming coherent contours across the grouped digits. Figure 2 shows tonal patterns which cued grouping: Patterns at left have in common a generally domed shape (i.e. rise-fall) across each of the perceived groups of digits.

Figure 2: 4 pitch patterns that cued grouping from [20]: top row cued 3-2 grouping, bottom the 2-3 .



Patterns at right have the flipped picture, with scooped shapes over the groups (i.e. fall-rise).

Given that the rise-fall and fall-rise are also both commonly attested shapes of single intonation phrases in American English (e.g. ToBI [4] L+H* L-L% and H+!H* L-H%), these are seen as candidates for cross-phrase tonal schemas in American English.

3. METHODS: TWO EXPERIMENTS

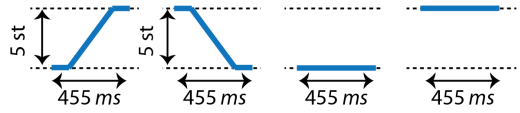
To investigate the two hypotheses that 1) intonational schemas can cue prosodic grouping, and 2) perceived duration of silent intervals can be distorted by means of pitch-based grouping in speech, a pair of experiments was conducted using the same experimental materials with 2 different tasks: 1) linguistic judgment of grouping in an ambiguous string based on cue interpretation (Experiment A) and 2) psychoacoustic judgment of perceived duration (Experiment B). It is expected that, as with the kappa effect, silent intervals between pitch-delineated groups will be magnified in perception compared to those within a pitch-delineated group. It is therefore hypothesized that silences occurring between schema-delineated groups will sound systematically longer than those occurring across digits within those groups. (This work makes the assumption that intonational phrases can be prosodically grouped ([5, 15, 23, 24, 25,] *inter alia*), but will not commit to the phonological status of such groups, i.e. whether they reflect recursive phrasing or higher prosodic categories).

3.1. Stimuli

The stimuli were 5 repetitions of the word *nine*, with manipulations to intervening silence duration and the f0 of the digits. All stimuli used resyntheses and concatenations of the same base file: an isolated full intonational phrase token of the word *nine*, ~455 *ms* long, naturally produced with roughly level pitch (ToBI H* H-L%, at ~190 *hz*) by a female native speaker of American English (the first author).

The base file was resynthesized with 5 contours: *high* level, *low* level, *rise* and *fall* (Fig. 3). The *high* f0 was arbitrarily set to 200 *hz*, and the *low* 5 *st* lower (~150 *hz*). The *rise* and *fall* each had f0 glides of 5 *st* between 200 *hz* and ~150 *hz*. Because the goal was to have pitch across adjacent digits sound continuous (to control for continuity), *rise* and *fall*

Figure 3: Individual pitch contours used in concatenations: *rise, fall, low and high.*



were given a slight sigmoid shape, starting and ending with 100 *ms* of level f0. (Listeners less accurately perceive pitch at glissando starts and ends (Cf. Rossi's 2/3 rule [30]), and appear to discount pitch in less sonorant regions (cf. [2]).) Files were concatenated in 5 pitch sequences. Two were predicted to cue 3-2 grouping (*rise-high-fall-rise-fall* = "rise-fall-3-2" & *fall-low-rise-fall-rise* = "fall-rise-3-2"), and two the 2-3 grouping (*rise-fall-rise-high-fall* = "rise-fall-2-3" & *fall-rise-fall-low-rise* = "fall-rise-2-3"). The 5th pattern ("neutral"), a steady fall across all 5 digits, was included as a baseline.

The base timing pattern had a 200 *ms* pause between each of the 5 digits. Duration was increased for either the 2nd or 3rd pause (and never for both at the same time, and never the 1st or 4th pause). Pause 2 and 3 were each increased by 25, 50, 75, 100, 150 and 200 *ms*, giving 13 timing steps. Timing was expected to cue 2-3 grouping when pause 2 was increased, and 3-2 grouping when pause 3 was increased. The 13 timing patterns were crossed with the 5 pitch patterns (See Fig. 4 for schematic of stimuli types, and Fig. 5 for a sample stimulus).

Figure 4: Schematic of the 5 pitch patterns with sample timing patterns. Left column has longer pause 2; middle column shows all pauses equal; right column shows longer pause 3.

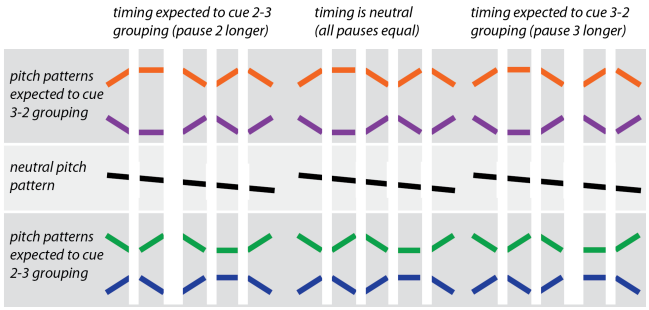
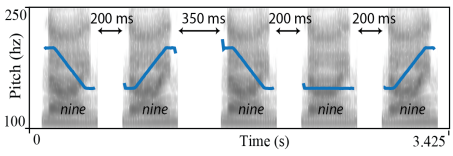


Figure 5: A sample stimulus showing the "fall-rise-2-3" pitch pattern and 150 *ms* increase to pause 2.



3.2 Subjects and presentation

33 native speakers of American English (age 18-23, 8 m) participated in both experiments in separate paid sessions. For Experiment A, subjects indicated whether each string was "grouped" as 99-999 ("2-3 grouping") or 999-99 ("3-2 grouping"). For Experiment B, subjects indicated whether pause 2 or pause 3 was longer. 16 subjects completed Experiment A first, the other 17 completed B first.

Subjects listened to 260 randomized trials (4 repetitions x 5 pitch patterns x 13 time patterns) over headphones in a quiet room. Experiments were forced-choice, with responses indicated via laptop keys. Text was displayed on the screen to represent the 2 choices: for Experiment A, digits grouped by dashes (i.e. 999-99 & 99-999); for Experiment B, the word *nine* was repeated orthographically, with extra space between words to indicate the location of the larger pause (i.e. *nine nine nine nine nine* for longer pause 3). Left/right position of text on the screen matched the orientation of designated response keys.

Each experimental session took ~30 minutes, including breaks. Subjects read a brief introduction, then proceeded to a training section. Experiment A training used recorded examples of prosodic grouping of 3 repeated digits produced by a naïve speaker. Subjects indicated groupings of two digits followed by one, or one digit followed by two (e.g., 44-4 or 4-44). For Experiment B, training used only "neutral" pitch versions of the experimental stimuli (5 *nines*), with timing increases of 100, 150 and 200 *ms*, (i.e., the largest differences) to pause 2 or 3. Subjects indicated whether pause 2 or 3 was longer.

4. RESULTS & ANALYSIS

Figure 6 shows results for Experiment A ("grouping"), from 8484 trials: percent responses "3-2 grouping" is graphed by time step, with lines to indicate the 5 pitch contours. The x-axis shows duration difference in *ms* between pauses 2 and 3: Positive values indicate that pause 3 is longer than 2; negative values that pause 2 is longer; time 0 that pauses 2 and 3 were equal. The general increasing diagonal trend of the lines reflects that time step was a strong cue to grouping: longer pause 3 cued more responses of "3-2 grouping"; longer pause 2 cued more of "2-3 grouping". The neutral pitch line (solid black) runs more-or-less through the center diagonal, suggesting no bias to either grouping. Lines for both fall-rise patterns (green and purple dashed) overlap with the neutral, showing that neither strongly cued grouping for the data set as a whole (though see discussion for more details). However, both rise-fall patterns (orange and blue dashed) show lines that separate distinctly, both from each other and from the neutral, and in the direction predicted: The "rise-fall-3-2" pattern indeed cued more "3-2 grouping" responses, and "rise-fall-2-3" cued more "2-3-grouping", at all time steps.

Figure 6: Results from Expt. A: responses "3-2-grouping" by time difference between pause 2 and 3, with separate lines for the 5 pitch patterns.

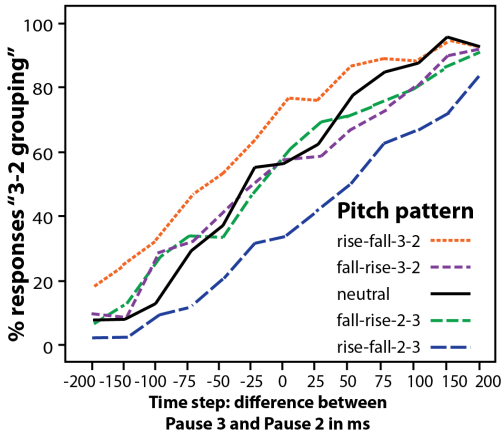


Figure 7: Results from Expt. B: responses "pause 3 longer" by time difference between pause 2 and 3, with separate lines for the 5 pitch patterns.

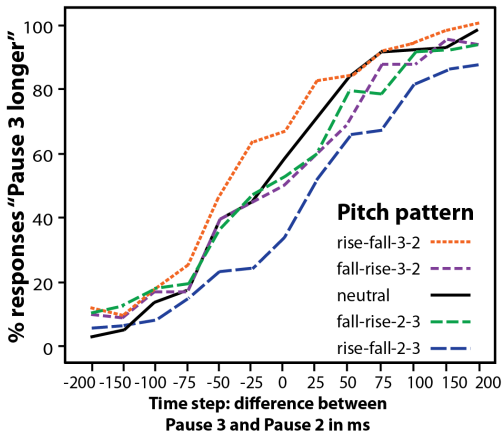


Figure 7 shows results from 8527 trials from Experiment B ("timing"): percent responses "pause 3 longer" ("3-2 timing") graphed by time step, as in Figure 6, with lines for pitch patterns. Again, both fall-rise patterns (green and purple dashed lines) largely overlap with the neutral (black solid), but both rise-fall patterns (orange and blue dashed) do distinctly separate from both the neutral and each other, especially for middle time steps (i.e., where pause 2 and 3 have a difference under 100 *ms*).

Results were analyzed with mixed-effects logistic regression, implemented in R with the lme4 package [3] with response ("3-2-grouping"/"2-3-grouping" for Expt. A; "pause 3 longer"/"pause 2 longer" for Expt. B) as dependent variables, and time step and pitch pattern as fixed factors. Subject was included as a random effect [1]. The result for Expt. A was a model (N=8345, log-likelihood=-4172) showing an expected significant main effect of timing (Wald Z= 43.85, p<.001), and main effects for pitch pattern, with "rise-fall-2-3" and "rise-fall-3-2" differing from "neutral" (rise-fall-2-3, Wald Z=-11.85, p<.001; rise-fall-3-2, Wald Z=7.69, p<.001). The result for Expt. B (N=8345, log-likelihood=-3734) showed significant main effects of timing (Wald Z= 46.41, p<.001) and pitch pattern, with "rise-fall-2-3" and "rise-fall-3-2" differing from "neutral" (rise-fall-2-3, Wald Z=-9.15, p<.001; rise-fall-3-2, Wald Z =4.97, p<.001).

5. DISCUSSION & CONCLUSIONS

Predicted intonational schemas did cue grouping, at least for rise-fall patterns; pitch had an effect even when timing supported the opposite grouping. The picture for fall-rise patterns is more complex: results of individual subjects reveal that fall-rise patterns did cue grouping in the predicted manner (as in [20]) for a substantial subset of subjects (N=11). For another subset (N=14), however, these same patterns cued groupings *opposite* to those predicted: "fall-rise-2-3" cued more "3-2 grouping" responses, and "fall-rise-3-2" more "2-3 grouping". (The other 8 subjects showed no tendencies in either direction for fall-rises.) This suggests considerable variation in how subjects interpreted pitch cues.

As for the time distortion, the similarity of the results graphs (Figures 6 & 7) supports the hypothesis that distortion of perceived duration occurred between perceived groups: subjects tended to hear longer pauses between rise-fall delineated groups. It is also the case that for fall-rise patterns, individual subjects' perception of duration appeared to overlap/correspond with their individual grouping perception. A rough-and-ready index of the relation between subjects' grouping and timing responses was calculated: the difference in overall "3-2-grouping" responses for each schema pair (rise-fall 3-2 vs. 2-3 and fall-rise, 3-2 vs. 2-3) in Expt. A, was compared to those same calculations for the "pause 3 longer" responses in Expt. B. A two-tailed test of significance resulted in a Pearson co-efficient of .723 for fall-rise patterns, and of .549 for rise-fall patterns, both significant at the .01 level, supporting a correlation between individual grouping perception patterns and distortions to perceived duration.

Overall, it does appear that pitch can cue grouping in ways that are not easily characterized by proximity and continuity: results support a role for intonational schemas in prosodic grouping. Further, grouping cued by such patterns can lead to inflation of perceived duration of between-group pauses. Results thus support the auditory grouping hypothesis. While the auditory motion hypothesis may indeed explain some pitch-based distortions to perceived duration, it is less likely than the auditory grouping hypothesis to explain the results of these two experiments.

6. REFERENCES

[1] Baayen, R., Davidson, D. & Bates, D. 2008. Mixed-effects modeling with crossed random effects for subjects and items, *Journal of Memory and Language*, 59: 390-412, 2008.

[2] Barnes, J., 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] Beckman, M., & Ayers-Elam, G., 1997. *Guidelines for ToBI Labelling* (version 3, March 1997).

[5] van den Berg, R., Gussenhoven, C., & Rietveld, T. 1992. Downstep in Dutch: Implications for a model. *Papers in laboratory phonology II: Gesture, segment, prosody*, *335*, 359.

[6] Bolton, T. L. 1894. Rhythm. *The American Journal of Psychology*, *6*(2), 145–238.

[7] Bregman, A. S. 1994. *Auditory scene analysis: The perceptual organization of sound*. MIT Press, Cambridge, MA.

[8] Brown, S. W. 2008. Time and attention: Review of the literature. *Psychology of Time*, 111–138.

[9]Brugos, A. & Barnes, J. 2014. Effects of dynamic pitch and relative scaling on the perception of duration and prosodic grouping in American English. In *Proceedings of Speech Prosody 7, Campbell, Gibbon, and Hirst (eds.), pp. 388-392.*

[10] Brugos, A. & Barnes, J. 2012. Pitch trumps duration in a grouping perception task, 25th Annual CUNY Conference on Human Sentence Processing, New York, NY. March, 2012.

[11] Brugos, A. & Barnes, J. 2012. The auditory kappa effect in a speech context. Speech Prosody, Shanghai, China.

[12] Cohen, J., Hansel, C., & Sylvester, J. 1953. A new phenomenon in time judgment. *Nature*, *172*: p. 901, 1953.

[13] Crowder, R. G., & Neath, I. 1995. The influence of pitch on time perception in short melodies. *Music Perception*, *12*(4), 379–386.

[14] Cumming, R. 2011. The effect of dynamic fundamental frequency on the perception of duration. *Journal of Phonetics*, *39*(3), 375–387.

[15] Féry, C., & Truckenbrodt, H. 2005. Sisterhood and tonal scaling. *Studia Linguistica*, *59*(3).

[16] Geiser, E., & Gabrieli, J. D. 2013. Influence of rhythmic grouping on duration perception: a novel auditory illusion. *PLoS ONE*, *8*(1), e54273.

[17] Henry, M. J. 2011. *A Test of an Auditory Motion Hypothesis for Continuous and Discrete Sounds Moving in Pitch Space*. PhD dissertation, Bowling Green State University.

[18] Henry, M. J., & McAuley, J. D. 2009. Evaluation of an imputed pitch velocity model of the auditory kappa effect. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 551–564.

[19] Hoopen, G. T. 2008. Classic Illusions of Auditory Time Perception. *Journal of the Human-Environmental System*, *11*(1), 27–35.

[20] House, D. 1990. *Tonal Perception in Speech*. Lund: Lund University Press.

[21] Hunyadi, L. 2006. Grouping, the cognitive basis of recursion in language. *Argumentum*, 2, 67–114.

[22] Jeon, H.-S., & Nolan, F. 2013. The role of pitch and timing cues in the perception of phrasal grouping in Seoul Korean. *The Journal of the Acoustical Society of America*, *133*(5), 3039–3049.

[23] Kentner, G., & Féry, C. 2013. A new approach to prosodic grouping. *The Linguistic Review*, *30*(2).

[24] Ladd, D. R. 1986. Intonational phrasing: The case for recursive prosodic structure. *Phonology* 3. 311–340.

[25] Ladd, D. 1988. Declination "reset" and the hierarchical organization of utterances. *The Journal of the Acoustical Society of America*, *84*, 530.

[26] Lehiste, I. 1976. Influence of fundamental frequency pattern on the perception of duration. *Journal of Phonetics*, 4, 113–117.

[27] Lerdahl, F., & Jackendoff, R. 1983. *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press.

[28] MacKenzie, N. 2007. *The kappa effect in pitch/time context*. Dissertation. Ohio State University.

[29] McDougall, R. 1903. The structure of simple rhythm forms. *Psychological Monographs,* **4**: 309–416.

[30] Rossi, M. 1971. Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole, *Phonetica*, **23**, 1- 33.

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

[32] Thorpe, L. A., & Trehub, S. E. 2004. Duration Illusion and Auditory Grouping in Infancy. *Developmental Psychology*, *25*(1), 122–127.

[33] Wertheimer M. 1938 Laws of organization in perceptual forms. In *A source book of Gestalt Psychology,* Ellis W. (ed.). London: Routledge and Kegan Paul. 71–88.

[34] Yu, A. C. L. 2010. Tonal effects on perceived vowel duration. In *Papers in Laboratory Phonology,* Fougeron, C., Kühnert, B., D'Imperio, M & Vallée, N. (eds.), Vol. 10. Berlin: Mouton de Gruyter.