Pitch cues to hierarchical metric structure in children's poetry

Mara Breen[1†]
Ahren Fitzroy[1,2]
[1] Mount Holyoke College, South Hadley, MA, USA
[2] University of Massachusetts Amherst, Amherst, MA, USA
[†] Corresponding author: mbreen@mtholyoke.edu
Published 16 December 2021; https://doi.org/10.18061/FDMC.2021.0038
Author video presentation and/or other conference material: https://doi.org/10.17605/OSF.IO/HZVJU

Abstract

We investigated whether speakers use pitch to signal hierarchical metric structure in productions of Dr. Seuss's *The Cat in the Hat*, by modeling fundamental frequency (F0) of monosyllabic words as a function of metric strength and a set of control parameters. We modeled maximum F0 of ~25000 words in a corpus of book productions from 17 speakers, comparing a 3-level musical metric model and a 5-level linguistic metric model. Results demonstrate that speakers consistently realized two levels of musical metric strength, as words corresponding with downbeats were produced with higher maximum F0 than all other beats. In addition, speakers simultaneously realized three levels of linguistic metric strength, as maximum F0 decreased linearly across the three highest linguistic metric levels. These results are consistent with previous work in both prose speech production and Western music composition, demonstrating that poetic speech uses pitch variation in ways that are consistent with both music and speech, and they complement prior demonstrations that duration and intensity variation signal musical and linguistic metric structure in the same corpus.

KEYWORDS: *speech, pitch, meter, rhyme, intonation*

Introduction

Children's poetry shares many structural features with music (Lerdahl, 2001), but there has been little empirical investigation of how these features are realized in production. This project assesses how speakers use pitch variation to signal both musical and linguistic metric structure in children's poetry by analyzing productions of *The Cat in the Hat* (Dr. Seuss, 1957), a quintessential example of children's poetry with regular hierarchical metric structure.

Prior work demonstrates that music performers signal hierarchical metric structure with duration (Palmer, 1996; Todd, 1985) and intensity variation (Drake & Palmer, 1993). Our previous work demonstrates that, in productions of *The Cat in the Hat,* speakers systematically signal hierarchical musical metric structure with intensity variation (Fitzroy & Breen, 2020) and hierarchical linguistic metric structure with duration variation (Breen, 2018). The goal of the current study is to investigate how a third acoustic variable – pitch – is manipulated by speakers to signal hierarchical musical and linguistic metric structure in productions of *The Cat in the Hat*.

There is considerable evidence from psycho-linguistics that speakers signal metric structure with pitch variation. Pitch accents, which are aligned with metrically strong syllables, are generally signaled by a local increase in pitch (Breen et al., 2010). Music cognition makes similar claims about how metric structure is realized in Western music; the Generative Theory of Tonal Music predicts that metrically strong positions are more likely to correspond with pitch excursions (Lerdahl & Jackendoff, 1983). Moreover, analyses of Western music reveal that pitch accents frequently occur at metrically strong positions (Huron & Royal, 1996). In addition, musical phrase boundaries, which typically coincide with metric unit boundaries (Temperley, 2003), are often signaled with falling pitch (Huron, 2006), and, in Western music, "late phrase compression" in which the pitch interval size tends to decline toward the end of a phrase (Shanahan & Huron, 2011).

The current paper investigates how speakers use pitch variation to signal *both* musical and linguistic hierarchical metric structure in productions of child-directed poetic speech. To do this, we modeled the pitch of each word using two hierarchical models of metric structure (Figure 1): a 3-level musical metric model based on a 6/8 measure structure, and a 5-level linguistic metric model based on cross-linguistic metrical poetry (e.g., Fabb & Halle, 2008). Based on prior work in music and speech production, we predict that speakers will signal metrically strong syllables with higher pitch than metrically weak syllables, and that speakers will signal ends of metric units with decreases in pitch.

Method

Participants

In the current study, we analyzed productions from the *The Cat in the Hat* corpus (Breen, 2018) from 17 female native speakers of American English.

Stimuli

Participants read aloud from a hardcover copy of *The Cat in the Hat* (Dr. Seuss, 1957) – a 61-page, illustrated children's book written in rhyming anapestic tetrameter, which is widely read by English-speaking caregivers to 0-3 year old children (Hudson Kam & Matthewson, 2017). The book consists of 1625 words (1576 monosyllabic words, 236 unique lexemes) organized primarily into 70 stanzas; each stanza contains two lines of four anapests each (as in (1)). The first line in each stanza ends with a rhyme prime and the second line ends with a phonologically predictable rhyme target.

(1)

"Put me down!" said the fish.
"This is no fun at all!
Put me down!" said the fish.
"I do NOT wish to fall!"

Acoustic Measures

Word and silence boundaries (Figure 2) were identified by automatic force-alignment of the audio productions with the text in Praat (Boersma & Weenink, 2018) using the Prosodylab-Aligner (Gorman et al., 2011), then manually adjusted as needed. F0 values were identified using Praat's auto-correlation algorithm. The maximum F0 value of each word was defined as the parabolically interpolated maximum pitch (Figure 2). Multisyllabic words were excluded from analysis, because the unstressed syllables have reduced pitch for reasons unrelated to metric structure (Fry, 1958). Limiting investigation to one-syllable words resulted in exclusion of 16 of 236 unique lexemes (49/1625 words). Disfluent and incorrect word productions were also excluded, resulting in 473 out of 26,792 possible monosyllabic word productions (1.77%). The remaining maximum F0 values of monosyllabic words were centered and scaled to standard deviation units (i.e., converted to z-scores) separately within each participant.

Text Annotation

All words were first annotated for a set of control factors: a) number of phonemes (M = 2.98, SD = 0.8), b) word class (542 open-class, 1034 closed-class), c) log lexical frequency (M = 5.82, SD = 2.15), d) syntactic dependency structure, e) text emphasis (26 words in SMALL CAPS), and f) intra-stanza repetition. Next, metric structure was annotated in two ways (Figure 1): with a 3-level musical metric structure based on 6/8 musical meter, where performers signal greatest prominence on beat 1, intermediate prominence on beat 4, and lowest prominence on beats 2, 3, 5, and 6 (Drake & Palmer, 1993); and a linguistic metric structure where metric feet are iteratively grouped in pairs, creating a 5-level hierarchical structure (Fabb & Halle, 2008).

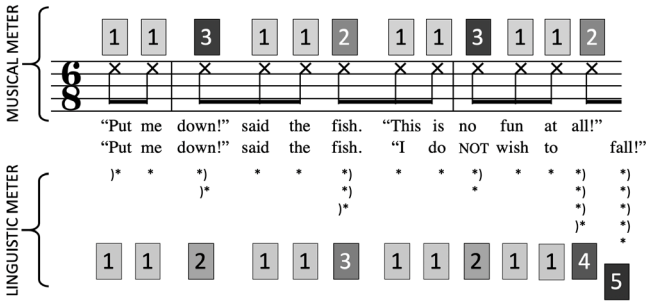

Figure 1: One stanza of The Cat in the Hat *annotated with the musical metric model (top) and the linguistic metric model (bottom).*

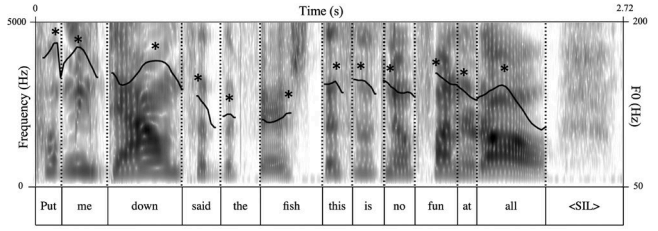

Figure 2: Word F0 measurement. An excerpt from one production is plotted as a time-frequency spectrogram. Identified word and silence boundaries are indicated by dashed vertical lines. The F0 contour generated for this excerpt is plotted in black over the spectrogram, with the parabolically-interpolated maximum F0 for each word indicated with an asterisk.

Analysis

Using linear mixed-effects regression, we fit a model of maximum F0 using the control factors. The data were fit on a word-by-word basis. The fully-saturated model included all control fixed effects, a random effect of speaker, and random slopes over speaker for each fixed effect. This model did not converge, so we iteratively removed random slopes accounting for the least variance and refit the model until it converged. Fixed effects were then individually removed and the simpler model was compared to the more complex model using a likelihood ratio test (Baayen et al., 2008). Factors accounting for significantly more variance in the more complex model remained in the final control model.

We then fit an experimental model by adding fixed effects corresponding to the predictions of musical meter and linguistic meter. We added musical metric strength as both a fixed effect and random slope over

participant, coded using simple contrast coding with metric strength level 2 (intermediate) as the reference level. We added linguistic metric strength as both a fixed effect and random slope over participant, coded using backward difference coding where each higher level is contrasted with the level below.

Results

Normalized maximum F0 is shown for words at each level of the musical metric hierarchy in Figure 3, and at each level of the linguistic metric hierarchy in Figure 4. The final model parameters appear in Table 1. Results demonstrate the speakers consistently signal two levels of musical metric strength with pitch: words aligned with metric strength level 3 (beat 1 in a 6/8 measure structure) are produced with higher pitch than words aligned with the other levels. Although words aligned with metric strength level 2 were produced with numerically higher F0 than level 1, this contrast did not reach significance in the model when including control parameters. Results also demonstrate that speakers use pitch to signal the end of large linguistic metric constituents, as evidenced by significant pitch decreases from metric strength level 3 to level 4, and from level 4 to level 5. Control parameters indicate that speakers use higher pitch to cue open class words (as opposed to closed class words), lower frequency words, words written in SMALL CAPS, and the first mention of words (as opposed to repeated mentions).

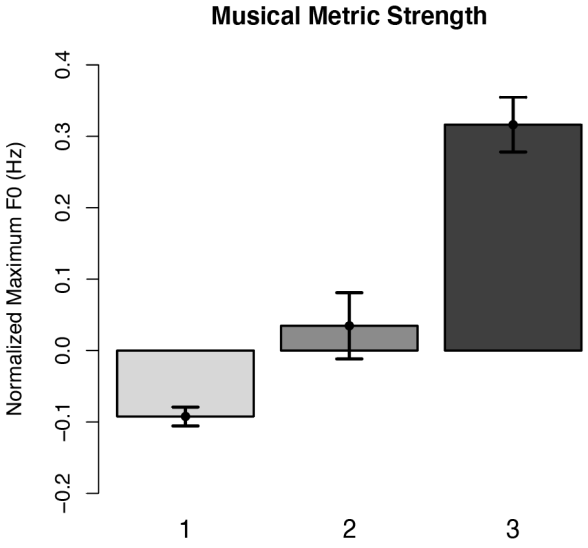

Figure 3. Values of normalized maximum F0 calculated across three levels of musical metric strength. Error bars indicate standard error.

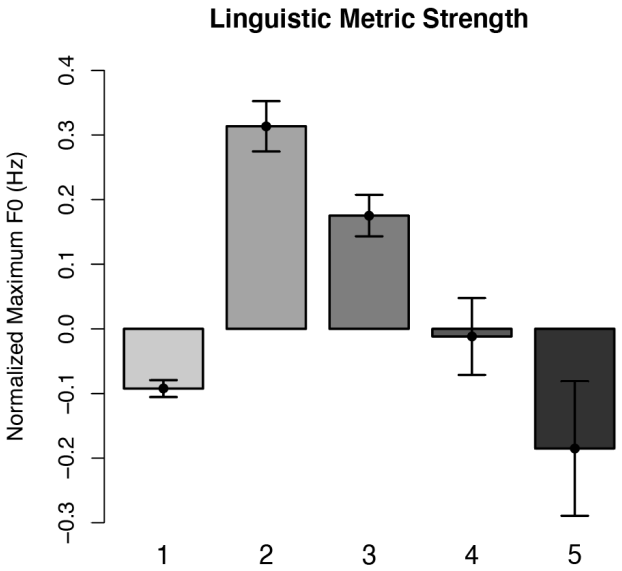

Figure 4. Values of normalized maximum F0 calculated across five levels of linguistic metric strength. Error bars indicate standard error.

Table 1: Fixed effects in the model of normalized maximum F0 by linguistic metric structure and control predictors

	Normalized Maximum F0			
Fixed Effects	*Estimates*	*SE*	*t*	*p*
(Intercept)	0.71	0.06	11.51	**<0.001**
Music Meter 3v2	0.44	0.17	2.60	**0.009**
Music Meter 1v2	0.28	0.24	1.16	0.245
Ling Meter 5v4	-0.15	0.04	-3.79	**<0.001**
Ling Meter 4v3	-0.18	0.04	-4.92	**<0.001**
Ling Meter 3v2	0.25	0.17	1.50	0.134
Ling Meter 2v1	0.12	0.17	0.72	0.472
Word Class	-0.18	0.02	-8.66	**<0.001**
Lexical Frequency	-0.08	0.00	-18.26	**<0.001**
Font Emphasis	0.41	0.05	8.36	**<0.001**
Intra-stanza repetition	-0.10	0.02	-6.76	**<0.001**

Discussion

The current study was designed to investigate the signaling of metric structure through pitch variation in child-directed productions of highly metric poetry. Results demonstrate that speakers use pitch variation to signal metric structure in poetic production in multiple ways, consistent with work in both music performance and prose speech production. Specifically, readers cue metrically strong syllables (i.e., beat 1 in a 6/8 musical meter) with higher pitch than metrically weak syllables (all other beats). In addition, speakers cue three hierarchical levels of metric units by decreasing pitch.

Although it has long been observed that poetry shares features with both music and language (Lerdahl, 2001), there have been few empirical investigations of this claim. The current study provides such empirical support by demonstrating that models of both musical and linguistic metric structure simultaneously account for pitch variation in child-directed productions of poetry. Interestingly, this hybrid realization of musical and linguistic metric structure in pitch differs from the patterns we have previously demonstrated for word intensity and duration in this corpus, which respectively predominantly signal either musical or linguistic metric structure alone.

For musical metric structure, pitch signaled a two-level metric hierarchy with sole accent on the downbeat (strength level 3). This contrasts with our previous investigations of both intensity, which signaled a three-level hierarchy with primary accent on the downbeat and secondary accent on strength level 2 (Fitzroy & Breen, 2020), and word duration, which signaled a three-level hierarchy with primary accent on strength level 2 and secondary accent on the downbeat (Fitzroy & Breen, 2018). The realization of downbeat accent in pitch suggests that, like intensity, musical metric structure is signaled via pitch variation. However, the realization of only two levels of musical metric strength in pitch indicates that this aspect of metric structure is realized with lower fidelity in pitch than in intensity.

For linguistic metric structure, pitch marked differences between the three highest strength levels. This clearly contrasts with our previous investigation of duration, which unambiguously signaled five levels of linguistic metric strength (Breen, 2018), and differs somewhat from our previous findings for intensity, which marked differences between strength levels 2 and 3, and between strength levels 4 and 5 (Fitzroy & Breen, 2018). The similarity of linguistic metric structure realization in pitch and intensity is consistent with prior findings that physical aspects of these prosodic channels lead them to correlate somewhat in speech production (e.g., Gramming et al., 1988). However, the clearer distinction between linguistic metric levels in pitch than in intensity suggests that linguistic metric structure is realized more clearly in pitch. Taken together, our present pitch results and prior investigations of metric realization in word intensity and duration in this same corpus demonstrate that metric structure is realized in poetic production in a manner that reflects both musical and linguistic features of poetry, but that the balance of these features differs across prosodic channels.

The realization of metric structure through pitch and other acoustic cues provides important temporal information to child listeners of poetic texts like *The Cat in the Hat.* Specifically, they can use the metric structure to generate temporal expectations about upcoming information. For example, event-related potential studies demonstrate that listeners use metric structure, cued by in part by pitch variation, to direct their attention to metrically strong events in both music (Fitzroy & Sanders, 2015) and speech (Breen et al., 2014). Child listeners who hear metrically-regular poetry like *The Cat in the Hat* with clear prosodic cues to metric structure can direct their attention to metrically strong moments, facilitating phonological learning.

Conclusion

The current study demonstrates that speakers of child-directed poetic text use pitch variation to consistently signal two levels of musical metric structure and three levels of linguistic metric structure. These results are consistent with previous work in both prose speech production and Western music composition. Moreover, these results complement prior demonstrations that duration and intensity variation signal musical and linguistic metric structure in the same corpus, while providing further evidence that metric structure is realized differently across these prosodic channels.

Acknowledgements

The authors wish to thank Kathryn Guarino, who collected the recordings and provided initial analysis of the data, Sarah Weidman, who provided assistance with analysis and inspired discussion of the corpus, and Michele Cubillo, who implemented the syntactic analysis. This work was supported by the James S. McDonnell Foundation [Understanding Human Cognition Scholar Award to MB].

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Boersma, P., & Weenink, D. (2018). *Praat: Doing phonetics by computer [Computer software]. Version 6.0. 43*.

Breen, M. (2018). Effects of metric hierarchy and rhyme predictability on word duration in The Cat in the Hat. *Cognition*, *174*, 71–81. https://doi.org/10.1016/j.cognition.2018.01.014

Breen, M., Dilley, L. C., McAuley, J. D., & Sanders, L. D. (2014). Auditory evoked potentials reveal early perceptual effects of distal prosody on speech segmentation. *Language, Cognition and Neuroscience*, *29*(9), 1132–1146. https://doi.org/10.1080/23273798.2014.894642

Breen, M., Fedorenko, E., Wagner, M., & Gibson, E. (2010). Acoustic correlates of information structure. *Language and Cognitive Processes*, *25*(7–9), 1044–1098. https://doi.org/10.1080/01690965.2010.504378

Dr. Seuss. (1957). *The cat in the hat* (Vol. 1). Random House Books for Young Readers.

Drake, C., & Palmer, C. (1993). Accent Structures in Music Performance. *Music Perception: An Interdisciplinary Journal*, *10*(3), 343–378. https://doi.org/10.2307/40285574

Fabb, N., & Halle, M. (2008). *Meter in poetry: A new theory*. Cambridge University Press. https://doi.org/10.1017/CBO9780511755040

Fitzroy, A. B., & Breen, M. (2018). *Readers use duration and intensity variation to signal hierarchical metric structure in child-directed poetic speech*. 15th International Conference on Music Perception and Cognition, Montreal, Quebec, Canada.

Fitzroy, A. B., & Breen, M. (2020). Metric Structure and Rhyme Predictability Modulate Speech Intensity During Child-Directed and Read-Alone Productions of Children's Literature. *Language and Speech*, *63*(2), 292–305. https://doi.org/10.1177/0023830919843158

Fitzroy, A. B., & Sanders, L. D. (2015). Musical meter modulates the allocation of attention across time. *Journal of Cognitive Neuroscience*, *27*(12), 2339–2351. https://doi.org/10.1162/jocn_a_00862

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, *1*(2), 126–152. https://doi.org/10.1177/002383095800100207

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, *39*(3), 192–193.

Gramming, P., Sundberg, J., Ternström, S., Leanderson, R., & Perkins, W. H. (1988). Relationship between changes in voice pitch and loudness. *Journal of Voice*, *2*(2), 118–126. https://doi.org/10.1016/S0892-1997(88)80067-5

Hudson Kam, C. L., & Matthewson, L. (2017). Introducing the infant bookreading database (IBDb). *Journal of Child Language*, *44*(6), 1289–1308. https://doi.org/10.1017/S0305000916000490

Huron, D. (2006). *Sweet anticipation*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/6575.001.0001

Huron, D., & Royal, M. (1996). What is melodic accent? Converging evidence from musical practice. *Music Perception*, *13*(4), 489–516. https://doi.org/10.2307/40285700

Lerdahl, F. (2001). The sounds of poetry viewed as music. *Annals of the New York Academy of Sciences*, *930*(1), 337–354. https://doi.org/10.1111/j.1749-6632.2001.tb05743.x

Lerdahl, F., & Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. MIT Press.

Palmer, C. (1996). Anatomy of a performance: Sources of musical expression. *Music Perception*, *13*(3), 433–453. https://doi.org/10.2307/40286178

Shanahan, D., & Huron, D. (2011). *Interval Size and Phrase Position: A Comparison between German and Chinese Folksongs*. https://doi.org/10.18061/1811/52948

Temperley, D. (2003). End-accented phrases: An analytical exploration. *Journal of Music Theory*, *47*(1), 125–154. https://doi.org/10.1215/00222909-47-1-125

Todd, N. (1985). A model of expressive timing in tonal music. *Music Perception*, *3*(1), 33–57. https://doi.org/10.2307/40285321