Behavior Research Methods

Introduction and application of the multiscale coefficient of variation analysis

Journal:	*Behavior Research Methods*
Manuscript ID	BR-SC-16-011.R1
Manuscript Type:	SCiP Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Abney, Drew; University of California - Merced, Cognitive & Information Sciences Kello, Chris; University of California, Merced, Cognitive and Information Sciences Balasubramaniam, Ramesh; University of California, Merced, Cognitive and Information Sciences

Introduction and application of the multiscale coefficient of variation

Drew H. Abney[1], Christopher T. Kello[1], & Ramesh Balasubramaniam[1]

[1]Cognitive and Information Sciences, University of California, Merced, USA

DRAFT, Please do not cite or distribute

Address correspondence to:
Drew H. Abney
Cognitive and Information Sciences
University of California, Merced
School of Social Sciences, Humanities and Arts
5200 North Lake Rd.
Merced, CA 95343
drewabney@gmail.com
(815) 382-0273

Abstract

Quantifying how patterns of behavior relate across multiple levels of measurement typically requires long time series for reliable parameter estimation. We describe a novel analysis that estimates patterns of variability across multiple scales of analysis suitable for time series of short duration. The multiscale coefficient of variation (MSCV) measures the distance between local coefficient of variation estimates within particular time windows and the overall coefficient of variation across all time samples. We first describe the MSCV analysis and provide an example analytical protocol with corresponding MATLAB implementation and code. Next, we present a simulation study testing the new analysis using time series generated by ARFIMA models that span white noise, short-term and long-term correlations. The MSCV analysis was observed to be sensitive to specific parameters of ARFIMA models varying in the type of temporal structure and time series length. We then apply the MSCV analysis to short time series of speech phrases and musical themes to show commonalities in multiscale structure. The simulation and application studies provide evidence that the MSCV analysis can discriminate between time series varying in multiscale structure and length.

Introduction and application of the multiscale coefficient of variation analysis

Temporal variability is ubiquitous across the behavioral and cognitive sciences. However, measures of temporal variability tend to focus on particular timescales in data, rather than relating variations across timescales. If they do relate across timescales, such as measures of long-range correlations, the methods tend to require very long time series (e.g. more than 1000 points). In this paper, we introduce a new analysis – the Multiscale Coefficient of Variation (MSCV) – to estimate temporal variability across multiple timescales even for short time series.

In Gaussian statistics, variance of the mean or its square root, standard deviation, is the standard measure of variability. Other types of variability include local variability, global variability, and serial correlations (Torre & Balasubramaniam, 2011). Local variability is the difference between adjacent values in a time series (Low, Grabe, & Nolan, 2000; Madison et al., 2009; Torre & Balasubramaniam, 2011). Global variability is the dispersion of a probability distribution typically quantified by the coefficient of variation (σ/μ)[1]. Serial correlation reflects how the values in a time series are related as a function of their distance from each other in time, and in particular whether nearby values tend to be more similar (persistent, positively correlations) or dissimilar (anti-persistent, negatively correlations) than chance (Bassingthwaighte, Liebovitch, & West, 1994; Newell & Slifkin, 1998). Local variability, global variability, and serial correlations are known to be non-independent of one another in certain conditions (Torre, Balasubramaniam & Delignières, 2010; Gilden, 2001; Marmelat, Torre, & Delignières, 2012).

[1] Note that some authors discuss serial correlations like long-range memory as 'global variance'. We use global variance here as a term to distinguish between coefficient variation and serial correlations.

Serial correlations can be found in most natural time series. For example, most biological and behavioral systems exhibit *long-range correlations* (Goldberger et al., 2002; Hausdorff, Peng, Ladin, Wei, & Goldberger, 1995; Ramos-Fernandez et al., 2004; Sims, 2008; West, 2006). In cognitive science, long-range correlations are found in memory processes (Maylor et al., 2001; Rhodes & Turvey, 2007), language structures (Zipf, 1949), and many other types of cognitive phenomena (see Kello et al., 2010, for review).

Long-range correlations can be expressed as scaling laws, i.e. nonlinear functions whereby one variable is relate to another raised to a power, $f(x) \sim x^{\alpha}$. The exponent, α, can be determined by plotting the variables on logged axes and estimating the slope (α) using a regression line. For temporal-based power laws, the variable of interest, $f(T)$, is often some type of variability estimate (e.g., root mean squared error, coefficient of variability, etc.) measured as a function of timescale T. The accuracy with which scaling laws can be estimated from data depends on the length (Delignières et al., 2006) and sample rate (Wijnants, Cox, Hasselman, Bosman, & Van Orden, 2013) of measurement series. Delignières et al. found increased biases and variability of spectral exponent estimation for time series shorter than 1024 data points. For some types of time series, 256 data points were acceptable. However, time series shorter than 1024 points are typically considered to be too short in length for reliable parameter estimation. This restriction is problematic for many behavioral experiments and other measurement conditions in which it is prohibitively difficult to collect more than a few dozen repeated measurements.

The goal of the current paper is to introduce MSCV analysis as a way to measure patterns of variability across multiple timescales for time series far shorter than 1024 data points. The problem we are working to solve is the estimation of patterns of variability across multiple timescales for extremely short time series. The goal is not to estimate scaling laws from data, but rather, to estimate how variability changes across a restricted range of timescales. In the following section, we provide an introduction and description of the MSCV analysis. Then we present a simulation study testing the new analysis using time series generated by ARFIMA models that span white noise, short-term and long-term correlations. In the simulation study, we systematically varied the length of the time series to investigate the sensitivity of the MSCV analysis to signal type and time series length. We will then apply the analysis to short time series of speech phrases and musical themes to show and compare their multiscale structures.

Multiscale coefficient of variation (MSCV)

MSCV analysis was developed to measure the degree to which the coefficient of variation of measured events spans multiple temporal scales. For a time series of event durations (e.g., reaction times, utterance durations, movement distances), the MSCV measures the distance between local coefficient of variation estimates within particular time windows and the overall coefficient of variation across all time samples. It should be reiterated that MSCV values cannot be used to estimate scaling laws. The MSCV analysis simply measures the patterns of variability across multiple timescales. Also, coefficient of variation is only meaningful as a ratio unit, so the user should be aware of what scale of measurement they are using during application.

The sizes of time windows *T* can be set by hand, or similar to scaling law analyses, varied as a power of 2 between a minimum of 2 and maximum of L/2-1, where L is the number of measurements in the time series. The time series is tiled with non-overlapping windows of size *T*, and the coefficient of variation is computed within each window. For window size *T*, coefficients of variation across windows are averaged,

$$MSCV(T) = \langle \frac{\sigma(T)}{\mu(T)} \rangle.$$

The MSCV function can be plotted with window sizes *T* on the x-axis and corresponding MSCV values on the y-axis. The MSCV function can be quantified using a number of measures, such as the range, sum, and normalized sum of MSCV values, and the slope of the function in logarithmic coordinates (see Figure 1).

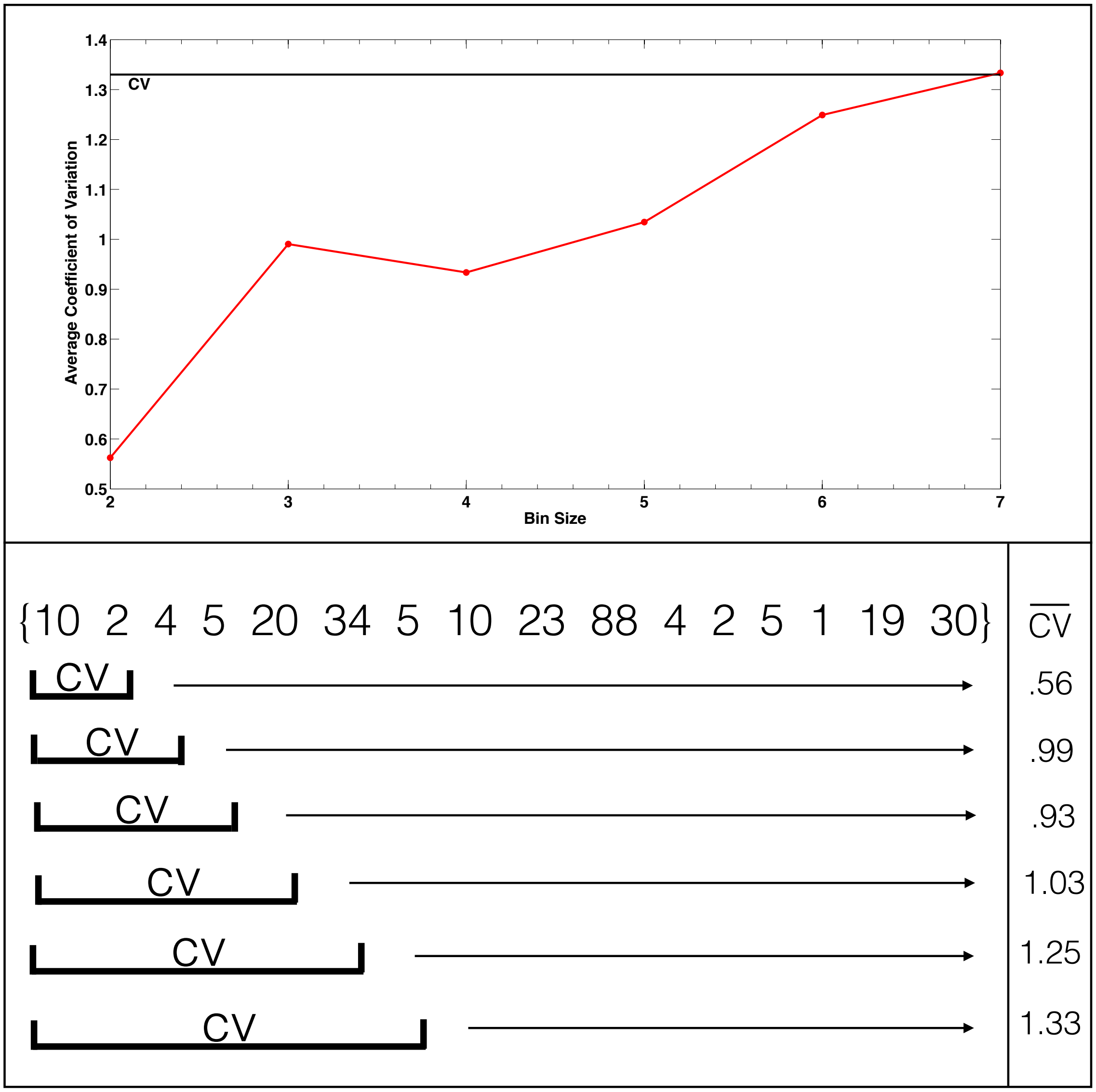

Figure 1. Schematic depiction of MSCV analysis. (top panel) MSCV profile of example time series. (bottom panel) Basic graphical description of MSCV analysis. For each bin size, coefficient of variation is computed across a sliding, nonoverlapping window and averaged. For each bin size, average coefficient of variation is computed (see bottom-right). $MSCV_{range} = .77$, $MSCV_{sum} = 6.10$, $MSCV_{norm} = .76$, $MSCV_{slope} = .61$.

Computing the range and sum of MSCV values is straightforward, but the normalized MSCV, $MSCV_{norm}$, requires some explanation. To compute $MSCV_{norm}$, the MSCV is divided by the global coefficient of variation for the entire time series, and normalized relative to the amount of window sizes N_T,

$$\frac{\frac{\sum_{i=2}^{T} MSCV(T)}{CV}}{N_T}.$$

The $MSCV_{norm}$ value is not bounded by a specific range of values, but it typically ranges between 0 and 1. By normalizing the MSCV by the global coefficient of variation, $MSCV_{norm}$ provides an estimate of the amount of variability across bins that are less than the global coefficient of variation. Time series with random structure will have most window sizes approximate the global coefficient of variation and will therefore have an $MSCV_{norm}$ estimate approximate 1.0. Time series that have more multiscale structure – variability spanning multiple window sizes – will have an MSCVnorm less than one[2].

The MSCV analysis was recently applied to an investigation of how music affects postural sway (Ross, Warlaumont, Abney, Rigoli, Balasubramaniam, 2016). The radial sway of center of pressure measurements of postural sway and musical durations (intervals of onset and offset of sound; Coath et al., 2007; Coath et al., 2009) were subjected to the MSCV analysis. Ross et al. were interested in the multiscale properties of postural sway and musical durations for the purposes of assessing how the multiscale structure of postural sway couples to the multiscale structure of musical durations. Ross et al. observed that $MSCV_{norm}$ estimates of radial sway and musical durations were more similar for nonmusicians relative to musicians, suggesting that nonmusicians couple to the multiscale structure of music more so than musicians. Additional results suggested that the multiscale coupling occurred more for musical durations corresponding to low groove music (Janata et al., 2012).

ARFIMA simulations

[2] For MATLAB scripts go to https://github.com/drewabney/MSCV.git.

The ARFIMA (Auto-Regressive, Fractionally Integrated, Moving Average) modeling method was used to simulate time series with various degrees of short-range and long-range serial correlations. ARFIMA models are extensions to the classical ARMA (Auto-Regressive, Fractionally Integrated, Moving Average) models. ARMA models (p,q) include two components: a pth-order AR process and a qth-order MA process. ARFIMA models (p,d,q) include a dth- order fractional differencing (FI) process (Granger & Joyeux, 1980). We are using ARFIMA models to test the MSCV analysis because ARFIMA modeling has been previously used to estimate and identify long-range dependence and fractal exponents in cognitive and behavioral phenomena (Torre, Delignières, Lemoine, 2007; Torre, Varlet, & Marmelat, 2013; Wagenmakers, Farrell, & Ratcliff, 2004; Farrell, Wagenmakers, & Ratcliff, 2005).

We created three types of time series of durations that are known to vary in statistical structure: persistent long-range correlations (LRC), positive short-range correlations (SRC), and random white noise (WN). For each condition, a pool of 50 series (length=2048) was generated using the ARFIMA modeling method (using the *fracdiff* package in R). All conditions had a mean of 800 and a coefficient of variation of ~6%. The auto-regressive (AR) parameter for LRC, SRC, and WN conditions were 0, 0.6, and 0, respectively. The fractional integration (FI) parameter for LRC, SRC, and WN conditions were 0.45, 0, 0, respectively. The moving average (MA) parameter was set to 0 for all conditions. The specific ARFIMA parameters generated three conditions that varied in memory decay as quantified by the auto-correction function: LRC series exhibited power-law decay over lags suggestive of long-term statistical memory, SRC series exhibited an exponential decay over lags suggestive of short-term statistical memory, and the WN series exhibited no positive or negative autocorrelations across lags. The ARFIMA parameters chosen for the simulation study were shown to affect sensorimotor synchronization performance across auditory stimulus conditions that varied in white noise, short-range, and long-range correlation properties (Torre, Varlet, & Marmelat, 2013).

For each time series, we estimated the $MSCV_{norm}$, for the entire time series (length 2048) and for a random sample of lengths 200, 100, 50, 25, and 10. See Table 1 and Figure 2 for results. Although the MSCV analysis can compute various estimates from the MSCV profile, our aim is to quantify properties of multiscale structure using a single-valued estimate. Therefore, in this simulation study, we chose to only use the $MSCV_{norm}$ estimate.

To test the performance of the $MSCV_{norm}$ estimate against a common multiscale analysis, we also estimated the Hurst exponent using the Anis-Lloyd/Peters corrected rescaled range analysis (Hurst-AL) (see Weron, 2002) for the simulated time series. The rescaled range analysis was first introduced by Mandelbrot and Wallis (1969) and extends Hurst's (1951) calculation of a self-similarity parameter, H. The R/S analysis consists of estimating the range (R) and standard deviation (S) of a subset of a time series. For example, a subset of a time series with a minimum value of 3 and maximum value of 9 will have a range of 6. If the standard deviation of the subset was S=2, then the rescaled range for this particular subset is R/S=3. If we increase the number (n) of observations in the subset, the linear relationship (H) between R/S estimate and n in logarithmic coordinates will approximate H=.5 for a random walk (e.g., white noise) and will be greater than H=.5 for Fractional Brownian motion. The Hurst-AL was used because it was found to improve estimation performance for small time series. To our knowledge, no researcher has used the R/S-AL analysis on extremely small times, e.g., *n*=10.

The results suggest that the $MSCV_{norm}$ estimates are sensitive to signal type. The first observation from the $MSCV_{norm}$ estimates is that WN signals approximate a $MSCV_{norm}$ value approximating 1.0. The second observation is that the $MSCV_{norm}$ decreases from 1.0 as a function of increased multiscale structure, from WN to SRC to LRC. Hereafter, we use the term multiscale structure to refer to the variation can be different or heterogeneous, across timescales. Decreased or low multiscale structure means that variation is similar or homogeneous across timescales. Considering the known statistical dependencies of the three signal types generated from the ARFIMA models, these observations provide two intuitions about the $MSCV_{norm}$ measure. The two intuitions depend on whether the user is interpreting the $MSCV_{norm}$ measure as an absolute or relative measure.

If considering the $MSCV_{norm}$ as an absolute measure, the lower and upper bounds [0.0,1.0] suggest that increasing $MSCV_{norm}$ estimates approaching 1.0 correspond to signals with more homogeneity of variation across bins. Conversely, estimates decreasing from 1.0 suggest more heterogeneity of variation across bins and therefore, a signal that is *more* multiscale.

If considering the $MSCV_{norm}$ as a relative measure, the *directionality* of the $MSCV_{norm}$ estimates between two or more experimental conditions or partitions becomes informative. For example, if a user observed that $MSCV_{norm}$ estimates for Condition A were lower relative to $MSCV_{norm}$ estimates for Condition B, the user could interpret the signals from Condition A to have more heterogeneity across bins and therefore, is consider more *multiscale*, relative to the signals in Condition B.

Another important observation is that the $MSCV_{norm}$ estimates were sensitive to signal type for all time series lengths. However, the $MSCV_{norm}$ estimates failed to discriminate between LRC and SRC signals for the n=25 simulations. At n=10, the LRC and SRC switch orders but both still discriminate between the WN time series. These results suggest that the MSCV analysis is sensitive to different types of time series of extremely short lengths. For extremely short time series, the MSCV analysis can discriminate between time series exhibiting white noise (close to randomness) and time series with specific temporal correlations.

Table 1. Results from the ANOVAs and planned comparison for $MSCV_{norm}$ and Hurst-AL estimates across time series lengths.

Time series length	*F(2,47)*	$p_{LRC\ vs.\ SRC}$	$p_{LRC\ vs.\ WN}$	$p_{SRC\ vs.\ WN}$
$MSCV_{norm}$				
2048	217.92^{***}	<.001	<.001	<.001
1024	127.00^{***}	<.001	<.001	<.001
200	117.75^{***}	<.001	<.001	<.001
100	59.59^{***}	.001	<.001	<.001
50	57.46^{***}	>.05	<.001	<.001
25	23.45^{***}	>.05	<.001	<.001
10	12.02^{***}	.02	.02	<.001
Hurst-AL				
2048	2895.90^{***}	<.001	<.001	<.001
1024	1544.00^{***}	<.001	<.001	<.001
200	428.80^{***}	.001	<.001	<.001
100	239.30^{***}	>.05	<.001	<.001
50	107.05^{***}	.01	<.001	<.001
25	36.33^{***}	.02	<.001	<.001
10	4.15^{*}	.01	>.05	.02

Note. *p≤.05, **p≤.01, ***p≤.001. $p_{LRC\ vs.\ SRC}$, $p_{LRC\ vs.\ WN}$, $p_{SRC\ vs.\ WN}$ columns display results from planned comparisons between the three signal types.

The Hurst-AL measure performed similarly to the MSCV analysis at longer time series (*n*=2048, *n*=1024, *n*=200). However, at *n*=100, SRC and LRC time series are statistically indistinguishable, and at *n*=50, the SRC and LRC estimates flip. At *n*=10, the Hurst-AL fails to discriminate between the LRC and WN time series estimates.

Overall, both analyses perform equally well for longer time series. For smaller time series, both analyses also display flipped estimates around *n*=100 (Hurst-AL) and *n*=50 ($MSCV_{norm}$). For extremely short time series, the MSCV analysis – despite a flipping of the SRC and LRC estimates – are able to discriminate between estimates from SRC and LRC time series and estimates from WN time series. The Hurst-AL estimates at *n*=10, showed that WN and LRC estimates were indistinguishable. Considering the results from this simulation study, we would advocate users to employ either analysis for substantially long time series. However, if users desire to estimate properties of multiscale variability for extremely short time series, we advocate utilizing the MSCV analysis.

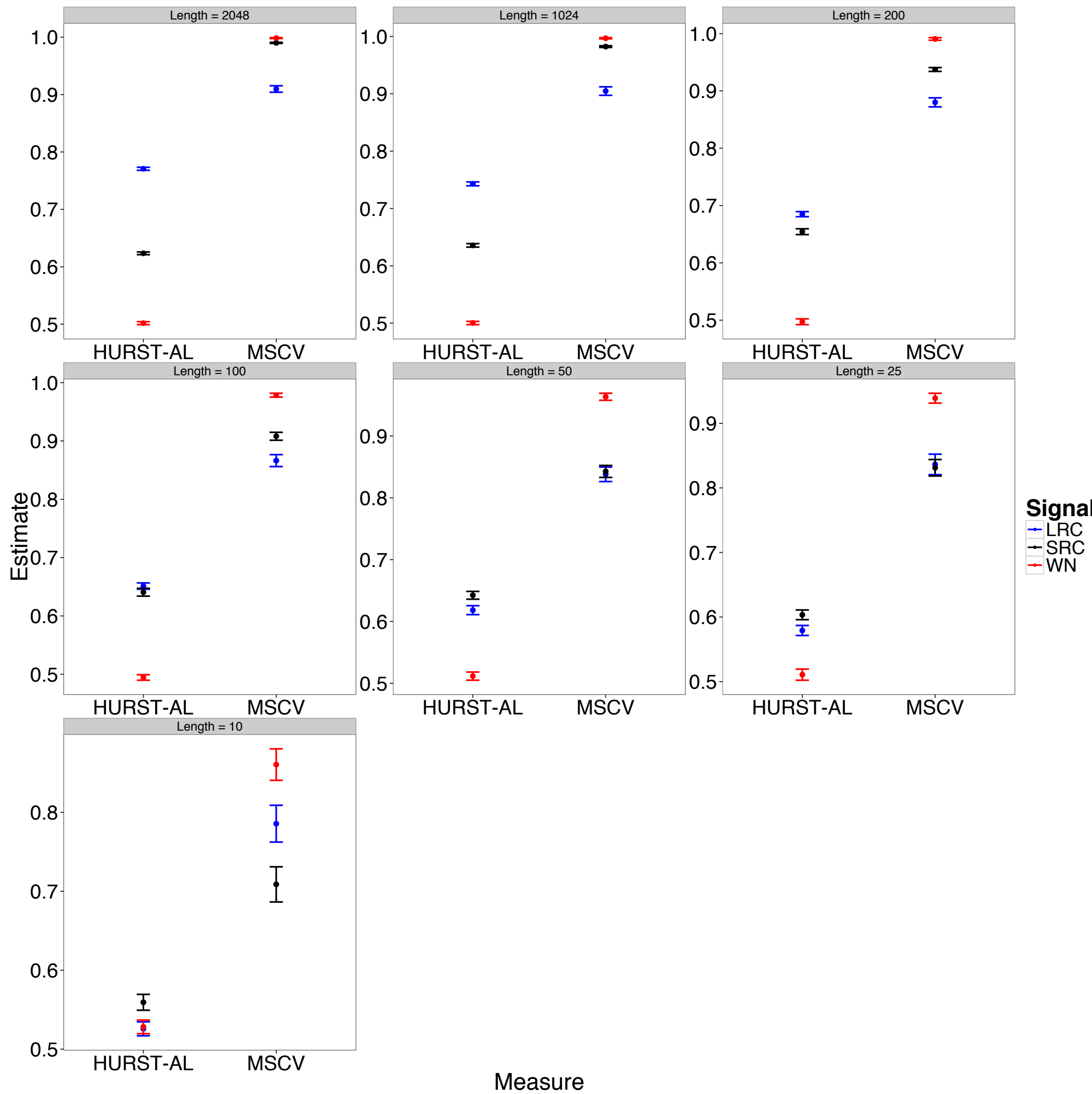

Figure 2. Results for Hurst-AL and MSCV$_{norm}$ estimates as a function of signal type and time series length. Error bars represent standard error of the means.

Overall, the results from the ARFIMA simulations suggest that the MSCV$_{norm}$ provides an intuitive estimate about the multiscale properties of a signal. In the next section, we report an application of the MSCV analysis. We chose our corpora due to the extreme length limitations of the duration series. As previously discussed, a main feature of the MSCV analysis is that it can assess the multiscale structure of extremely short time/event series. Hurst-AL estimates provide information about how the normalized range of values scale across multiple time scales. $MSCV_{norm}$ estimates provide information about how the coefficient of variation at specific time scales relate to the global coefficient of variation. We have demonstrated that for extremely short time series, assessing the coefficient of variation normalized at various time scales (normalized for global coefficient of variation) is more sensitive than assessing the rescale range of values across multiple time scales.

An empirical comparison of multiscale structure in language and music

We now provide an application of the MSCV to a novel comparison of the relationship between speech and music. The study of the relationship between speech and music is generally influenced by a common intuition that both are universal among human cultures (see Patel, 2010). Studying the commonalities between speech and music has lead to rich empirical research programs. Here we focus on the potential common patterns of multiscale structure across speech and music.

For the study of speech and music, one focus has been on prosodic properties like melody and rhythm (Hannon, 2009; Huron & Ollen, 2003; Jusczyk & Krumhansl, 1993; Lerdahl & Jackendoff, 1983; London, 2011; Patel & Daniele, 2003; Patel, Iversen, & Rosenberg, 2006; Ramus, Nespor, & Mehler, 1999). This work was influenced by a hypothesized typology of an isochronous rhythmic organization: stress-timed and syllable-timed languages (Abercrombie, 1967; Pike, 1945). Stress-timed languages were purported to have equal intervals between stresses, and syllable-timed languages were purported to have equal intervals between syllable onsets. Although empirical research does not support this 'isochrony' hypothesis, researchers have started focusing on durational patterns of vocalic and intervocalic intervals.

To measure durational variability, researchers have utilized the normalized pairwise variability index (nPVI), which provides a 'local' measure of the variability of durational patterns:

$$\mathrm{n}PVI = \frac{100}{m-1} \times \sum_{k-1}^{m-1} \left| \frac{d_k - d_{k+1}}{\frac{d_k + d_{k+1}}{2}} \right| ,$$

where m is the number of intervals in a time series and d_k is the duration of the kth interval in the time series. The nPVI is a dimensionless quantity that provides a measure of variability of durational differences for pairs of intervals (i.e., bin size of 2) relative to the average duration of the pair.

Grabe and Low (2002) observed that nPVI measurements of vocalic intervals were greater in stress-timed languages such as British English than in syllable-timed languages such as French. This finding points to earlier work (see Nespor, 1990) suggesting that stress-timed languages are known to exhibit more vowel reduction than syllable-timed languages. Ramus et al., (1999) observed more variability in consonantal durations for stress-timed languages and proposed that stress-timed languages are purported to have more complex syllable structure relative to syllable-timed languages.

In the vein of musical composition, Patel & Daniele (2003) observed that rhythmic patterns in French and British English musical themes had similar rhythmic patterns of the composers' (either French speaking or English speaking) native languages. Using the nPVI to measure local contrast variability, Patel and Daniele found that note durations of British English composers had greater variability relative to note durations of

French composers, which corresponds to what was observed in linguistic nPVI values of speech (Ramus, 2002). Patel and Daniele's results point to a potential common property between speech and music: prosodic patterns via rhythmic durations.

Another potential commonality is that both speech and music are organized across various levels of hierarchical order (Lerdahl & Jackendoff, 1983). In music, meter is the expected pattern of durations, usually denoted by a time signature. Meter is a recurring pattern of durations and displays structure (metric structure) across levels of variation (London, 2000). Using Patel and Daniele's corpus of musical themes, London and Jones (2011) found differences across levels of rhythmic and metrical structure. Recent work in the study of conversational speech has shown that clustering of speech onsets are organized across time scales purported to align with levels of linguistic representation (Abney, Paxton, Dale, & Kello, 2014; Abney, Kello, & Warlaumont, 2015; see also, Luque, Luque, & Lacasa, 2015). In an extension of Patel and colleagues (Patel & Daniele, 2003; Patel, Iversen, & Rosenberg, 2006), we test whether similarities between the multiscale variability of speech and music can be observed across languages that vary on the stress-timed vs. syllable-timed spectrum.

In line with work suggesting that stress-timed languages exhibit more diverse and complex syllable structure (Nespor, 1990; Ramus et al. 1999), we predict that music composed and language produced by native speakers of a stress-timed language (e.g., English) will display more multiscale structure relative to native speakers of syllable-timed languages, e.g., French. To test this prediction, we constructed musical and speech corpora and submitted the musical and speech durations to the MSCV analysis to estimate $MSCV_{norm}$ values. We expect to observe lower $MSCV_{norm}$ for music and speech produced by native speakers of a stress-timed language, which would indicate more multiscale structure. Ramus et al. (1999) observed more consonantal variability for stress-timed language. Therefore, we predict that, controlling for local contrast variability (nPVI), $MSCV_{norm}$ estimates will be lower for stress-timed languages relative to syllable-timed ones.

Musical corpus

Our source of musical material was a subset of the corpus used in Patel and Daniele (2003). Patel and Daniele focused on collecting musical themes written by native-speaking British English and native-speaking French speaking composers who were born in the 1800s and died in the 1900s. The chosen musical themes consisted of at least 12 notes (e.g., eighth, quarter, etc.) with no internal pauses or rests (cf. Patel & Daniele, 2003). Therefore, for each musical theme, we had a time series of note durations. To control for metrical type, in the current analyses we only included musical themes in duple time[3]. Themes with duple time have a binary meter where the meter divides into beats into two subdivisions, e.g., 2/2, 2/4, 6/8. We also excluded musical themes with isochronous durational patterns. A total of 59 English musical themes and 79 French musical themes were included in the current study (see Table 2). For our corpus, the mean duration amount was 20 durations and the minimum duration amount was 12 durations.

To investigate differences in durational variability across musical themes, we estimated CV, nPVI, and $MSCV_{norm}$ for each musical theme. Because we are interested in

[3] In earlier analyses including musical themes exhibiting triple time and other complex metrical structures, we found that the results were neither straightforward nor reliable. Future work with larger corpora should attend to the issues of multiple metrical types.

how the estimates of the MSCV explain variance above and beyond local measures of durational contrast (e.g., nPVI), we also include analyses where nPVI was residualized out of the $MSCV_{norm}$ variable.

The nPVI measures the average degree of durational contrast (or variability) between two successive durations in a time series of discretized events. nPVI can be considered a measure of local variability. The nPVI is a single valued estimate that is computed by (1) estimating absolute difference between two successive intervals durations, (2) normalize by the mean duration of the pair, and (3) multiplied by 100. nPVI estimates closer to 100 are interpreted as having larger durational contrasts relative to lower nPVI estimates. The nPVI has been used in studies of speech and music rhythm (Grabe & Low, 2002; Low, Grabe, & Nolan, 2000; Patel & Daniele, 2003; Ramus, 2002; Ross, Warlaumont, Abney, Rigoli, & Balasubramaniam, 2016).

Table 2. Composers examined in this study.

Composers	# Themes
English	
Bax	6
Delius	10
Elgar	8
Holst	7
Ireland	4
Vaughan Williams	24
French	
Debussy	13
Fauré	6
Honegger	7
Ibert	6
Milhaud	6
Poulenc	5
Ravel	7
Roussel	7
Saint-Saëns	22

English music and French music did not differ in estimates of CV (β=-.05, t[136]=-.30, p=.77) or nPVI values, β=.-21, t(136)=-1.20, p=.23. However, English music did have lower values of $MSCV_{norm}$ relative to French music, β=.61, t(136)=3.29, p=.002. It is important to note that our nPVI results slightly diverge from Patel and Daniele (2003): although we found English music to have higher nPVI estimates relative to French music, this difference was not statistically reliable. One possible explanation for this difference is that we only included musical themes with duple meter, reducing the size of the corpus by almost 25%.

To assess if $MSCV_{norm}$ captured variance not explained by local variability, we residualized out nPVI from $MSCV_{norm}$. After controlling for nPVI, the original pattern of results held, suggesting that English music had lower $MSCV_{norm}$ estimates relative to French music, β=.59, *t*(136)=3.56, *p*<.001 (See Figure 3).

We also submitted the Hurst-AL to the musical corpus. The Hurst-AL analysis yielded estimates for less than 5% of the musical corpus. Given the low percentage of Hurst-AL estimates, we did not proceed to test for differences across the musical corpus. Inspecting the event series in the musical corpus that did and did not yield Hurst-AL estimates provided more insight into the differences between the Hurst-AL analysis and the MSCV analysis. The Hurst-AL analysis could not converge on event series with multiple consecutive identical event durations (e.g., Bax, b508: .5, .5, .5, .5, 1.0, 1.5, .5, .5, .5, .5, .5, 1.0, 1.0…). Because the Hurst-AL estimate relies on a rescaling of ranges for particular window sizes, at small window sizes, the range will be 0. The MSCV analysis relies on coefficient of variation, not a metric of range, and is therefore more flexible for a diverse array of event series types.

A lower $MSCV_{norm}$ estimate suggests that rhythmic durations span more bins of the MSCV profile, which is suggestive of an event series that is more multiscale. Our results suggested that, even when controlling for local variability (nPVI), English music has stronger multiscale properties relative to French music. In other words, there appears to be more heterogeneity of variance across timescales for English music relative to French music.

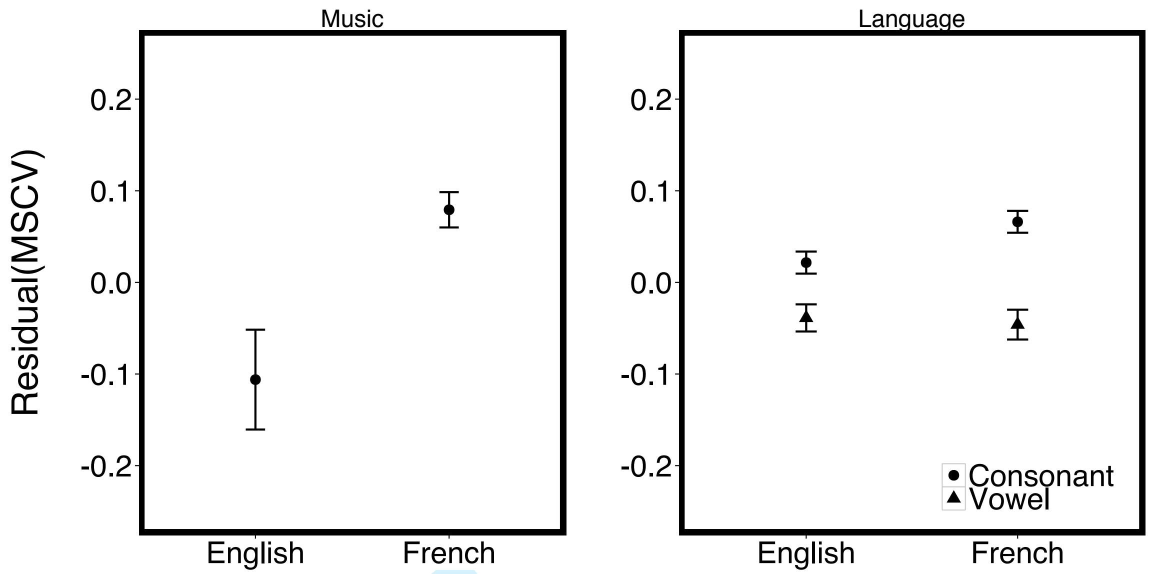

Figure 3. Results of the residual analyses for the musical theme durations (left) and language durations (right). Error bars represent standard error of the means.

Linguistic corpus

The main hypothesis for this application study is that music and spoken language have similar multiscale structure as a function of the composer's native language and the native language of speakers. We can also test whether or not specific units of language – such as vowel durations and consonant durations – display different multiscale structure. Our source of linguistic material was a subset of the BonnTempo Corpus (BTC 1.0; Dellwo et al., 2004). The BTC was originally constructed for the study of rhythmic variability of read speech across languages representing 'stress-timed' (e.g., English and German) and 'syllable-timed' (e.g., French and Italian) rhythmic classes. The text is a passage from a novel 'Selbs Betrug' by Bernhard Schlink.

In the BonnTempo Corpus, speakers were instructed to first read the passage in their 'normal reading' rate. After the first reading, speakers were instructed to read the passage again at different speech tempi. We only included read speech from native English and French speakers reading the passage at a 'normal reading' rate. The BonnTempo Corpus consists of Praat™ textgrid files (Nijmegen, NL) with human-coded labeling of syllables, consonantal intervals, and vowel intervals. We created custom Praat™ scripts to extract consonantal and vowel intervals from textgrid files. Our linguistic corpus consisted of 49 read phrases in English and 42 read phrases in French. For our corpus, mean duration amount was 27 durations with a minimum duration amount of 13 durations. For each read speech phrase, we created event series, akin to the rhythmic durations in the musical themes for consonantal durations and vowel durations. To investigate differences in durational variability across read speech, we estimated CV, nPVI, and $MSCV_{norm}$ for each phrase and duration type. Similar to the analysis of musical themes, we also included an analysis where nPVI was residualized out of the $MSCV_{norm}$ variable.

CV estimates were higher for English speakers (M=.51, SE=.01), relative to French speakers (M=.44, SE=.02), β=-.72, t(178)=-3.53, p<.001. CV estimates did not vary across consonantal durations (M=.46, SE=.01) and vowel durations (M=.50, SE=.02), β=.02, t(178)=.14, p=.86. We observed a Language X Duration Type interaction, β=.59, t(178)=2.06, p=.04, suggesting that CV estimates for French consonant durations (M=.41, SE=.02) were lower than English consonant durations (M=.51, SE=.02), t=-3.53, p=.003, but estimates for French (M=.49, SE=.02) and English vowel durations (M=.51, SE=.02) were not reliably different, t=-.63, p=.92.

nPVI estimates were higher for English speakers (M=57.06, SE=1.51), relative to French speakers (M=50.82, SE=1.45), β=-.73, t(178)=-3.61, p<.001. nPVI estimates were higher for consonant durations (*M*=55.44, *SE*=1.54) relative to vowel durations (*M*=52.91, *SE*=1.50), β=-.45, *t*(178)=-2.32, *p*=.02. We observed a Language X Duration Type interaction, β=.29, *t*(178)=2.11, *p*=.03, suggesting that nPVI estimates for French consonant durations (*M*=49.70, *SE*=2.10) were lower than English consonant durations (*M*=60.37, *SE*=1.99), *t*=-3.61, *p*=.002, but estimates for French (*M*=51.93, *SE*=2.02) and English vowel durations (*M*=53.76, *SE*=2.18) were not reliably different, *t*=-.62, *p*=.92.

$MSCV_{norm}$ estimates for English speakers (*M*=.92, *SE*=.01) and French speakers (*M*=.93, *SE*=.01) were not reliably different, β=.31, *t*(178)=1.60, *p*=.10. $MSCV_{norm}$ estimates were higher for consonant durations (*M*=.97, *SE*=.01) relative to vowel durations (*M*=.88, *SE*=.01), β=-.65, *t*(178)=-3.54, *p*<.001. We did not observe a Language X Duration Type interaction, β=-.40, *t*(178)=-1.47, *p*=.14.

Finally, to control for local variability estimated by the nPVI, we residualized out the variance explained by the nPVI estimates and constructed a new model for the $MSCV_{norm}$ estimates. Residual $MSCV_{norm}$ estimates for English speakers (*M*=-.009, *SE*=.01) were reliably smaller relative to French speakers (*M*=.01, *SE*=.02), β=.43, *t*(178)=2.26, *p*=.02. Residual $MSCV_{norm}$ estimates for vowel durations (*M*=-.04, *SE*=.01) were reliably smaller relative to consonant durations (*M*=.04, *SE*=.01), β=-.59, *t*(178)=-3.19, *p*=.002. We observed a marginal Language X Duration Type interaction, β=.29, *t*(178)=-.50, *p*=.06. However, planned comparisons suggested that French consonant durations (*M*=.07, *SE*=.01) were not reliably different than English consonant durations (*M*=.02, *SE*=.01), *t*=1.61, *p*=.38, nor were estimates for French (*M*=-.05, *SE*=.02) different from English vowel durations (*M*=-.04, *SE*=.01), *t*=-.47, *p*=.96 (see Figure 3).

We also submitted the Hurst-AL analysis to the language corpus. The Hurst-AL analysis provided estimates for 97.8% (n=178) of the language corpus, and therefore, we tested for differences across language and duration type. Hurst-AL estimates for English speakers (M=.45, SE=.07) and French speakers (M=.44, SE=.08) were not reliably different, β=-.12, t(174)=-.59, p=.55. Hurst-AL estimates were higher for vowel durations (M=.46, SE=.07) relative to consonant durations (M=.43, SE=.08), β=.44, t(174)=2.16, p=.03. This result corroborates with the results from the $MSCV_{norm}$ estimates suggesting that vowel durations have more multiscale structure relative to consonant durations. We did not observe a Language X Duration Type interaction, β=-.19, t(174)=-.64, p=.52. Residual Hurst-AL estimates (controlling for nPVI) did not differ across language, duration type, nor the language X duration type interaction, all ps>.10.

Interim discussion of the application results

The results from the residual $MSCV_{norm}$ estimates for musical themes suggests that the composer's native language has an influence on the multiscale structure of his or her work. Similar to other past studies (Patel & Daniele, 2003; see also London & Jones, 2011), we applied a quantitative measure of a proposed property of speech and music, multiscale variability, to the music of composers from stress-timed (British English) and syllable-timed (French) languages. We found that, controlling for local variability (nPVI values), English classical music had more multiscale variability, as suggested by observing lower $MSCV_{norm}$ estimates, relative to French classical music. We limited our corpus of musical themes to only consist of themes with duple meter. In a re-analysis of Patel and Daniele (2003), London and Jones (2011) investigated two levels of linguistic structure and found that only themes in duple time showed the differing patterns of local variability across British English and French themes.

We observed that English-read speech had more multiscale variability, as suggested by observing lower $MSCV_{norm}$ estimates, relative to French-read speech. We also observed that vowel durations had more multiscale variability relative to consonant durations. Finally, although we observed marginally interaction suggesting that, for consonantal durations, English-read speech has more multiscale variability relative to French-read speech. However, subsequent analyses suggested that this was only a nominal difference. Nevertheless, we can speculate that these results relate to the idea that stress-timed languages have more complex syllables (Dauer, 1983). Ramus et al. (1999) observed that consonantal durations in for stress-timed languages have more variability relative to syllable-timed languages (Ramus et al., 1999). Again, however, these interpretations are speculative considering the lack of a reliable effect in subsequent statistical tests.

Across the results of speech and music, one observation is that the patterns of $MSCV_{norm}$ estimates were most similar across the language of the composer (musical corpus) and speaker (language corpus), suggesting that at least for one stress-timed language, English, there appears to be more multiscale variability. If cultural differences do in fact influence the composition of music, perhaps this pattern suggests that the complexity of syllable structure influences the degree to which a musical theme is composed. This conjecture could be informed by future work with larger and more diverse speech and music corpora.

Discussion and conclusion

Methods for estimating patterns of variation across scales of measurement typically require the user to have substantially large time series. In this paper, we introduced a new analysis that affords researchers the ability to estimate patterns of variability across temporal scales using time series of limited length.

In the simulation study, we observed that the MSCV analysis was sensitive to different types of time series that varied depending on the temporal structure generated from ARFIMA models. From the MSCV profile, the user can choose from a variety of estimates that, in various ways, quantify the pattern of variability across temporal scales. We observed that the $MSCV_{norm}$ estimate generally ranges from 0.0 to 1.0. In the simulation study, ARFIMA models generating white noise time series produced $MSCV_{norm}$ estimates around 1.0. ARFIMA models generating long-range and short-range correlations produced $MSCV_{norm}$ estimates less than 1.0. Notably, long-range correlations are known to display multiscale structure across temporal scales and had the smallest $MSCV_{norm}$ estimates. As previously noted, the MSCV analysis is not meant to assess the fractality of a time or event series. Researchers interested in assessing whether or not a time series is fractal are encouraged to use previously existing methods (Eke, Hermán, Kocsis, & Kozak, 2002; Goldberger et al., 2002; Hausdorff, Peng, Ladin, Wei, & Goldberger, 1995; Holden, 2005).

In the simulation study we also compared the MSCV analysis with common multiscale analysis, the rescaled range analysis (Hurst-AL). We found that both the MSCV and rescaled range analyses performed equally well for longer time series. However, the MSCV outperforms the rescaled range analysis for extremely short time series and for event series with diverse properties, e.g, consecutive identical durations in the musical corpus.

In the application study, we applied the MSCV analysis to a comparison between speech and music. Previous research had shown that the rhythmic properties of music and speech, as quantified by the nPVI, vary as a function of whether the composer's native language was stress-timed (e.g., English) or syllable-timed, e.g., French (Patel & Daniele, 2003). In our application study, we investigated whether multiscale properties of speech and music, as quantified by the $MSCV_{norm}$ estimate, differed as a function of stress- and syllable-timed languages, too. We observed that $MSCV_{norm}$ estimates for note durations of music and read speech differed across English and French. Specifically, English music and speech has lower $MSCV_{norm}$ estimates relative to French music and speech. These results suggest that stress-timed languages have stronger multiscale properties relative syllable-timed languages. Conversely, these results suggest that the variability of syllable-timed languages do not span as many levels of temporal structure. These results provide evidence for a common property linking music and speech: multiscale structure. The application study provided a good example of how the MSCV analysis can differentiate between time series of short durations.

In both the simulation study and the application study, we used the default binning parameters for each time series, [2,(L/2)-1]. However, the MATLAB scripts can be adjusted to define any range of bins as long as the minimum bin is a whole number greater than 1. The MSCV analysis can be applied to a wide range of datasets with duration- or interval-level data points. Coefficient of variation is a dimensionless number because it is independent of the unit of measurement specific to a dataset. The analysis has already been applied to measurements of postural sway (Ross et al. 2016), musical durations estimated from an auditory saliency model (Ross et al. in press; see also Coath et al., 2007, 2009), and durations from musical themes and spoken language (current study).

The results from the simulation and application studies suggest that the MSCV analysis can discriminate between time series that vary in multiscale structure. Importantly, the results from the simulation study suggest that even short time series (e.g., lengths of 50 or 25 data points), can vary in multiscale structure and can be differentiated using the MSCV analysis. It should be noted that for extremely short time series (e.g., 25 data points), the MSCV analysis failed to discriminate between time series of specific temporal correlations, e.g., LRC vs. SRC. Nevertheless, even for extremely short time series, the MSCV analysis, and specifically the $MSCV_{norm}$ estimate, was sensitive to whether a time series had heterogeneous structure (e.g., LRC and SRC) or homogeneous structure across timescales, e.g., white noise. Future research should try the MSCV analysis on a wide corpus of short and long sequences of behavioral data such as speech, human motor performance, and reaction times, and continuous measurements of neural data such as spike trains and time-varying EEG signals.

Acknowledgements

We thank Justin London for providing the musical theme corpus and Volker Dellwo for providing the BonnTempo corpus.

References

Abercrombie, D. (1967). Elements of general phonetics (p. 97). Chicago, IL: Aldine.

Abney, D. H., Paxton, A., Dale, R., & Kello, C. T. (2014). Complexity matching in dyadic conversation. *Journal of Experimental Psychology: General*, *143*(6), 2304.

Abney, D. H., Kello, C. T., & Warlaumont, A. S. (2015). Production and convergence of multiscale clustering in speech. *Ecological Psychology*, *27*(3), 222-235.

Bassingthwaighte, J. B., Liebovitch, L. S., & Bruce, J. (1994). West. *Fractal Physiology (Methods in Physiology 2 An American Physiological Society Book). An American Physiological Society Book, illustrated edition edition*.

Coath, M., Denham, S. L., Smith, L., Honing, H., Hazan, A., Holonwicz, P., & Purwins, H. (2007). An auditory model for the detection of perceptual onsets and beat tracking in singing. *Neural Information Processing Systems, Workshop on Music Processing in the Brain.*

Coath, M., Denham, S. L., Smith, L., Honing, H., Hazan, A., Holonowicz, P., & Purwins, H. (2009). Model cortical responses for the detection of perceptual onsets and beat tracking in singing. *Connection Science*, *21*(2-3), 193–205.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. Journal of Phonetics, 11, 51–62.

Delignieres, D., Ramdani, S., Lemoine, L., Torre, K., Fortes, M., & Ninot, G. (2006). Fractal analyses for ‘short’time series: a re-assessment of classical methods. *Journal of Mathematical Psychology*, *50*(6), 525-544.

Dellwo, V., Steiner, I., Aschenberner, B., Dankovičová, J., and Wagner, P. (2004). The BonnTempo-Corpus and BonnTempo-Tools: A database for the study of speech rhythm and rate. Proceedings of the 8th ICSLP, Jeju Island, Korea.

Eke, A., Herman, P., Kocsis, L., & Kozak, L. R. (2002). Fractal characterization of complexity in temporal physiological signals. *Physiological measurement*, *23*(1), R1.

Gilden, D. L. (2001). Cognitive emissions of 1/f noise. *Psychological review*, *108*(1), 33.

Goldberger, A. L., Amaral, L. A., Hausdorff, J. M., Ivanov, P. C., Peng, C. K., & Stanley, H. E. (2002). Fractal dynamics in physiology: alterations with disease and aging. *Proceedings of the National Academy of Sciences*, *99*(suppl 1), 2466-2472.

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven & N. Warner, Laboratory phonology (pp. 515–546). 7. Berlin: Mouton de Gruyter.

Hannon, E. E. (2009). Perceiving speech rhythm in music: Listeners classify instrumental songs according to language of origin. *Cognition*, *111*(3), 403-409.

Hausdorff, J. M., Peng, C. K., Ladin, Z. V. I., Wei, J. Y., & Goldberger, A. L. (1995). Is walking a random walk? Evidence for long-range correlations in stride interval of human gait. *Journal of Applied Physiology*, *78*(1), 349-358.

Holden, J. G., Van Orden, G. C., & Turvey, M. T. (2009). Dispersion of response times reveals cognitive dynamics. *Psychological review*, *116*(2), 318.

Huron, D., & Ollen, J. (2003). Agogic Contrast in French and English Themes: Further Support for Patel and Daniele (2003). *Music Perception*, *21*(2), 267.

Janata, P., Tomic, S. T., & Haberman, J. M. (2012). Sensorimotor Coupling in Music and the Psychology of the Groove. *Journal of Experimental Psychology: General, 141*, 54-75. doi: 10.1037/a0024208

Jusczyk, P. W., & Krumhansl, C. L. (1993). Pitch and rhythmic patterns affecting infants' sensitivity to musical phrase structure. *Journal of experimental psychology: Human perception and performance*, *19*(3), 627.

Kello, C. T., Beltz, B. C., Holden, J. G., & Van Orden, G. C. (2007). The emergent coordination of cognitive function. *Journal of Experimental Psychology: General*, *136*(4), 551.

Kello, C. T., Brown, G. D., Ferrer-i-Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., & Van Orden, G. C. (2010). Scaling laws in cognitive sciences. *Trends in cognitive sciences*, *14*(5), 223-232.

Lerdahl, F., & Jackendoff, R. (1983). An overview of hierarchical structure in music. *Music Perception*, 229-252.

London, J. (2000). Hierarchical representations of complex meters. In *Proceedings of the 6th International Conference on Music, Perception and Cognition*.

London, J., & Jones, K. (2011). Rhythmic refinements to the nPVI measure: A reanalysis of Patel & Daniele (2003a). *Music Perception*, *29*(1), 115-120.

Low, L. E., Grabe, E., & Nolan, F. (2000). Q uantitative Characterizations of Speech Rhythm: Syllable-Timing in Singapore English. *Language and speech*, *43*(4), 377-401.

Luque, J., Luque, B., & Lacasa, L. (2015). Scaling and universality in the human voice. *Journal of The Royal Society Interface*, *12*(105), 20141344.

Madison, G., Forsman, L., Blom, Ö., Karabanov, A., & Ullén, F. (2009). Correlations between intelligence and components of serial timing variability. *Intelligence*, *37*(1), 68-75.

Mandelbrot, B. B., & Wallis, J. R. (1969). Some long-run properties of geophysical records. *Water resources research*, *5*(2), 321-340.

Marmelat, V., Torre, K., & Delignières, D. (2012). Relative roughness: an index for testing the suitability of the monofractal model. *Frontiers in physiology*, *3*.

Maylor, E. A., Chater, N., & Brown, G. D. (2001). Scale invariance in the retrieval of retrospective and prospective memories. *Psychonomic Bulletin & Review*, *8*(1), 162-167.

Nespor, M. (1990). On the rhythm parameter in phonology. In I. Rocca, Logical issues in language acquisition (pp. 157–175). Dordrecht: Foris.

Patel, A. D. (2007). *Music, language, and the brain*. Oxford university press.

Patel, A. D., & Daniele, J. R. (2003). An empirical comparison of rhythm in language and music. *Cognition*, *87*(1), B35-B45.

Patel, A. D., Iversen, J. R., & Rosenberg, J. C. (2006). Comparing the rhythm and melody of speech and music: The case of British English and French. *The Journal of the Acoustical Society of America*, *119*(5), 3034-3047.

Peters, E. E. (1994). *Fractal market analysis: applying chaos theory to investment and economics* (Vol. 24). John Wiley & Sons.

Pike, K. N. (1945). The intonation of American English (pp. 34–35). Ann Arbor, MI: University of Michigan Press.

Ramos-Fernández, G., Mateos, J. L., Miramontes, O., Cocho, G., Larralde, H., & Ayala-Orozco, B. (2004). Lévy walk patterns in the foraging movements of spider monkeys (Ateles geoffroyi). *Behavioral Ecology and Sociobiology*, *55*(3), 223-230.

Ramus, F. (2002). Acoustic correlates of linguistic rhythm: perspectives. Proceedings of Speech Prosody 2002, Aix-en-Provence (pp. 115–120). Aix-en-Provence: Laboratoire Parole et Langage.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. Cognition, 73, 265–292.

Rhodes, T., & Turvey, M. T. (2007). Human memory retrieval as Lévy foraging. *Physica A: Statistical Mechanics and its Applications*, *385*(1), 255-260.

Ross, J.M., Warlaumont, A.S., Abney, D.H., Rigoli, L.M., Balasubramaniam, R. (2016). Influence of Musical Groove on Postural Sway. *Journal of Experimental Psychology: Human Perception and Performance.*

Sims, D. W., Southall, E. J., Humphries, N. E., Hays, G. C., Bradshaw, C. J., Pitchford, J. W., ... & Morritt, D. (2008). Scaling laws of marine predator search behaviour. *Nature*, *451*(7182), 1098-1102.

Slifkin, A. B., & Newell, K. M. (1998). Is variability in human performance a reflection of system noise?. *Current directions in psychological science*, 170-177.

Torre, K., Delignieres, D., & Lemoine, L. (2007). Detection of long-range dependence and estimation of fractal exponents through ARFIMA modelling. *British Journal of Mathematical and Statistical Psychology*, *60*(1), 85-106.

Torre, K., Balasubramaniam, R. & Delignières, D. (2010). Oscillating in synchrony with a metronome: serial dependence, limit cycle dynamics, and modeling. *Motor Control* 13: 323-343.

Torre, K., & Balasubramaniam, R. (2011). Disentangling stability, variability and adaptability in human performance: Focus on the interplay between local variance and serial correlation. *Journal of experimental psychology: human perception and performance*, *37*(2), 539.

Torre, K., Varlet, M., & Marmelat, V. (2013). Predicting the biological variability of environmental rhythms: Weak or strong anticipation for sensorimotor synchronization?. *Brain and cognition*, *83*(3), 342-350.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General*, *132*(3), 331.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2005). Human cognition and 1/f scaling. *Journal of Experimental Psychology: General*, *134*(1), 117.

Wagenmakers, E. J., Farrell, S., & Ratcliff, R. (2004). Estimation and interpretation of 1/fα noise in human cognition. *Psychonomic bulletin & review*, *11*(4), 579-615.

West, B. J. (2006). *Where medicine went wrong: Rediscovering the path to complexity* (pp. 119-72). London: World Scientific.

Weron, R. (2002). Estimating long-range dependence: finite sample properties and confidence intervals. *Physica A: Statistical Mechanics and its Applications*, *312*(1), 285-299.

Wijnants, M. L., Cox, R. F. A., Hasselman, F., Bosman, A. M. T., & Van Orden, G. (2012). Does sample rate introduce an artifact in spectral analysis of continuous processes?. *Frontiers in physiology*, *3*.

Zipf, G. K. (1949). Human behavior and the principle of least effort.

Dear Dr. Mike Jones,

Thank you for the opportunity to revise BR-SC-16-011, titled "Introduction and application of the multiscale coefficient of variation analysis". We thank you and all three reviewers for the thoughtful comments on the revised manuscript. We believe that these revisions have both improved and clarified the manuscript. We hope that the revised manuscript is suitable for publication in *Behavior Research Methods*.

The most major change is adding an additional analysis to the simulation study. This was requested by Reviewer 1 in order to compare our new method with existing multiscale analyses. We believe this additional analysis provides unique insights into the positives and negatives of our method and existing methods. Overall, both analyses perform equally well for long time series. However, the two analyses diverge at shorter and shorter time series lengths in interesting ways. We highlight these differences within the revision.

We believe that the current manuscript is much more clear and direct and importantly, accessible to the *Behavior Research Methods* readership.

We now provide a detailed summary of each reviewer's comments and our subsequent actions.

Revisions based on comments of the Reviewer 1:

R1, Comment 1:
Broadly, I'd like to see a clearer statement of what problem the authors are trying to solve with MSCV, and why MSCV solves it better than canonical approaches to multi-scale structure (e.g. DFA, multi scale entropy, wavelet methods). If the problem is just that these analyses don't work on small samples, then a much more thorough statistical treatment is needed to show that MSCV avoids their pitfalls (see below). If the problem is that these analyses don't capture the aspects of multi-scale structure that the authors are interested in, then the relevant aspects need to be stated precisely and a more direct comparison of the different techniques is needed.

If it's the former, then the provided ARFIMA simulations are not sufficient support the claim: the authors only used three condition with hand-picked parameters. Presumably, there's some relationship between the size of the effect and the sample size required to detect it. The AR parameter in the SRC condition was set to 0.6, which is a fairly strong time dependence — it's not surprising that it was distinguishable with a relatively small sample size. But how big a sample size would you need if the dependence is weaker? Similarly, the FI parameter is near the upper boundary of what can be set without the process becoming non-stationary. It would be more compelling to show a full parameter sweep and calculate statistical power at different combinations of sample size & effect size.

We have now provided a comparison analysis of the simulations using a type of rescaled *range analysis, the R/S-AL, which has been known to provide better estimation performance for short time series. Although the R/S-AL (and the Hurst-AL parameter estimate) has been used for time series as short as n=256 (see Weron, 2002), to our knowledge, we are not aware of any other work using the R/S-AL on time series of extremely short lengths, e.g., n=25 or n=10. We also added an additional time series length to the simulation, n=10, to test the lower limits of what is considered an extremely short time series. Overall, we find both estimate techniques perform equally well for longer time series. However, both techniques start breaking down at shorter time series (n=100 for Hurst-AL and n=50 for MSCVnorm). Crucially, we find that for extremely short time series, the MSCVnorm is able to distinguish between WN time series and LRC/SRC time series, whereas the Hurst-AL is able to distinguish between WN time series and SRC time series, but not between WN and LRC time series. We conclude that both techniques have limitations, but for extremely short time series, the MSCV analysis outperforms the R/S-AL analysis.*

We also submitted the Hurst-AL to the music and language corpus within the application study section of the revision. Notably, for the musical theme corpus, the Hurst-AL only provided estimates for less than 5% of the themes. Upon further inspection of the corpus, we found that the Hurst-AL has difficulty providing parameter estimates for event durations with consecutive identical durations (e.g., .5, .5, .5, .5, 1, 1, 1, 1, .25, .25, …). This difficulty is likely due to the Hurst-AL relying on standardized range values in its computation. We find this to be an interesting property that further distinguishes between the MSCV analysis and the Hurst-AL estimates.

Also, we have now made it more explicit that we implemented previously-used ARFIMA parameters from a study distinguishing performance of sensorimotor synchronization performance across various conditions that varied in random, short-range, and long-range properties (Torre, Varlet, & Marmelat, 2013). Our choice to use the ARFIMA properties from Torre et al. was to use properties in our simulations that are known to impact human performance. Torre et al., found that auditory sensorimotor synchronization differed as a function of the ARFIMA properties. This was not made explicit in the previous manuscript, and so we understand if this free parameter in our simulations was considered ill-defined.

R1, Comment 2:

Since you're estimating MCSV_norm from (possibly small) empirical time series, another statistical problem is deriving a sampling distribution to use for hypothesis testing. This seems a bit tricky, since the estimated MSCV for a particular window size is the mean of a sequence of noisy CV estimates, and the MSCV_norm further averages across window sizes. Analytic proofs are clearly outside the scope of the paper, but you could at least show the results of a simulation where you sample time series of a certain length with a known underlying MSCV_norm and show that your estimation procedure is unbiased and that the error scales with sample size in a reasonable way. This would also help make the point about MSCV working for small sample sizes (DFA, for example, is known to be biased for finite sample sizes; Bryce & Sprague, 2012).

This is an important comment and we think our new simulations provide insights into the biases R1 is mentioning. First, it is important to state that there are biases for both analytic techniques that vary in degree and function for different time series lengths. We make these biases explicit in the revision and find these biases to be important for readers interested in this topic of research. Overall, these results show how impactful time series length is for estimation. However, because we are interested in communicating the potential value of this new method, it is important to point out again that the MSCV analysis outperforms the R/S-AL analysis at extremely short time series, as long is performance is considered as the successful discrimination of SRC and LRC time series with WN time series.

R1, Comment 3:

1. It's worth noting that the coefficient of variation is only meaningful for measurements on ratio scales. All the suggested applications use event durations, which have a fixed zero point, but it's an important disclaimer for the reader nonetheless.

We have now added this disclaimer.

R1, Comment 3:

2. Captions for Figure 2 and Figure 3 should explain how error bars were calculated (e.g. bootstrapped 95% confidence intervals? SE of mean?)

We have now added more information about the error bars for Figs. 2 and 3.

R1, Comment 4:

3. In the final section, you refer to nPVI several times before defining or explaining it.

We have now provided additional information about the nPVI.

R1, Comment 5:

4. At several points in the "Linguistic Corpus" section, you run two separate t-tests and use the difference in significance to claim a difference in the conditions. However, a difference in significance doesn't imply a significant difference: you have to test the interaction (Nieuwenhuis, Forstmann, & Wagenmakers, 2011).

We thank R1 for this comment. We have now updated our entire results section and included more comprehensive regression models. In doing so, we did not find a reliable interaction between Language and Duration Type (p=.06) as per subsequent post-hoc tests. Although we did not find a reliable interaction, we did find other interesting results, which we discuss in the Results and Discussion sections.

R1, Comment 6:

5. Are there different multi-scale variance structures that would end up having the same MSCV_norm value? It seems like collapsing everything down to a scalar value could be "too lossy" in the sense of making meaningful differences look the same at the level of the DV.

We think that the simulation results suggest that the MSCV analysis and the Hurst-AL estimates are both sensitive to various properties such as the length of the time series. Therefore, it is possible that different types of structure (e.g., LRC, SRC) will have similar estimates. However, this is not a new insight and has been documented in various outlets regarding multiscale and fractal methods.

This is an issue that pervades work in multiscale and fractal estimation and is considered an open research/methodological question. We believe that it is important for the user of any of these methods (including the MSCV) to consider the range of time series lengths used in their sample and work to constrain this range in experimental design and, if necessary, in post-hoc decisions about sample inclusion. These thoughts have now been added into the manuscript as a disclaimer.

Revisions based on comments of the Reviewer 2:

Reviewer: 2

Comments to the Author
The authors' introduce a statistic, the multiscale coefficient of variation (MSCV), that is capable of distinguishing between data with different temporal structure. The authors describe MSCV clearly enough that I would feel comfortable in correctly implementing the algorithm from the details supplied in the submitted manuscript.

The authors demonstrate the statistic's sensitivity by first showing it can distinguish between simulated data with known temporal properties. The authors then demonstrate the practical value of this statistic to behavioral research by demonstrating it can be used to identify commonalities and differences in linguistic and musical data that depend on the native language of the speaker or composer. Specifically, the authors find evidence of more multiscale structure in music and speech when it is produced by native speakers of stress-timed (rather than syllable-timed) languages. This was revealed through an analysis of note durations for music and consonantal durations for language.

The main reason why behavioral researchers might want to the MSCV in analysis is that it has sensitivity even with with very short time series. This is not true of other tools that quantify scaling properties of temporal data. Describing MSCV and demonstrating it has sensitivity that can distinguish time series with different temporal properties is a useful contribution for behavioural scientists working with time series data.

R2, Comment 1:

The theoretical introduction to MSCV and analysis of simulation data were very clear. However, I became confused when the authors began talking about musical themes. My best understanding is that the authors calculated MSCV using note durations of a variety of compositions in duple time. It was unclear to me how one measures note duration. Do you have to get someone to play a piece? Can it be measured from a musical score? It was likewise unclear to me what duple time is and why it would be a good type of music to use for analysis here. I believe more explicit description would help musically unaware readers conceptualize what the authors are actually studying in this section of the paper.

We have now added information about the musical themes included in the sample, including information about duple meter.

R2, Comment 2:

I found the authors' description of their analysis of linguistic data confusing. At first, I got the impression that the authors had participants read passages that were extracted from the BonnTempo Corpus. After re-reading the section a few times, I now understand that the BonnTempo Corpus actually contains data of read passages. This section could be made more clear. Specifically, my source of confusion existed at the transition from paragraph 1 to paragraph 2 on page 17.

We have now provided additional information and detail about the BonnTempo Corpus.

R2, Comment 3:

Page 18 line 8. "French speaker" should be "French speakers".

Thank for you for pointing this out. We have corrected it accordingly.

Revisions based on comments of the Reviewer 3:
Reviewer: 3

Comments to the Author
In this study, the authors present a new analytical method, MSCV, for assessing temporal variability across multiple timescales, with the particular advantage in that it can be used on short timescales. Other approaches, which estimate scaling laws, are problematic in that accuracy is penalized with shorter series. The authors showcase MSCV across two sections: the first is in generating time series (using ARFMA) with known statistical structures (exhibiting long-range correlations, short-term correlations, and white noise) and determining whether MSCV can distinguish each as the lengths of time series are manipulated, and the second section is to apply MSCV to determine differences/similarities in language and speech rhythms across short time series signals.

R3, Comment 1:

My general impression is that this paper is very good, but could be improved upon by developing and better connecting each of the two sections described above. In particular, the last section on language/speech rhythms did not seem well integrated with the overarching goal (stated on page 3; demonstrating temporal structural variability across a range of short and long time series). It felt like an entirely new paper in many ways, with its own set of theoretical questions, debates, and points of emphasis. Although I enjoyed reading about the relationships between music and speech and why the music of speakers of stress-timed or syllable-timed languages might exhibit similar structural properties, I was expecting more on what MSCV might specifically bring to the table. Why is MSCV necessary for these types of data? How would other measures based on scaling law estimates fail to capture the temporal variability?

We thank R3 for this question. In response to R1's comment 1, we have now added a new measure in the simulation study, the Hurst-AL, which is a standard measure used in fractal estimation. Our new simulation results suggest that both analyses – our MSCV and the existing Hurst-AL – do a good job for long time series, but both analyses perform differently for time series decreasing in length.

R3, Comment 2:

One of the strengths of the MSCV is that it is a straightforward measure. I felt like the mechanisms of implementation were well explained. But I was not so clear on the motivating rationale for why each term in the MSCVnorm equation was included, and why it was expected to capture multiscale properties. For example, why normalize over the global coefficient of variation? Is this what is picking up on potential differences at longer timescales? So- just a bit more on these issues would be appreciated.

We have now provided additional information and detail regarding the specific properties of the $MSCV_{norm}$ *estimate.*

R3, Comment 3:

Another issue I'm mostly just curious about is the sensitivity of MSCVnorm. For example, what if you generated a sequence of ARFMA signals that had a finer gradation in their transitions from WN, SRC and LRC values? Would the correlation with MSCVnorm be near 1? You don't have to report this. I'm just curious.

This is a good question. We used ARFIMA models and exact parameters from a recent study (Torre, Varlet, & Marmelat, 2013) to generate our time series for the simulation study. From the simulation study, we know that the White Noise condition has a p(AR) of 0 and a d(FI) of 0. The SRC has a p(AR) of .6 and a d(FI) of 0 and the LRC has a p(AR) of 0 and a d(FI) of .45. For a time series with random structure, an MSCVnorm will approximate 1.0; even short time series (e.g, n=25). Although we do not know exact estimates from more finer gradations of these transitions, due to the current findings, we *expect that the MSCVnorm estimates will vary as a function of the multiple parameters of the an ARFIMA model. Of course, we could explore the entire parameter space of the ARFIMA model, which would be an interesting future project.*

R3, Comment 4:

Also, related to the previous point, is it a fair characterization to say that MSCV is good for just showing relative differences between two signals - that one signal generally has more heterogeneity in variance than another - but if you wanted to know precisely by how much, you would be better off using something that estimates fractal exponents (assuming long enough time series)?

We believe R3's intuitions are correct regarding the end-user applications. Also, we have now made it clear in the revision that the MSCV analysis is not for estimating fractal exponents nor is it a power law analysis.

R3, Comment 5:

The issue with nPVI is a bit confusing. The term is introduced on page 12 without really being defined (it's not until page 15 that a clear definition is given). You also state that nPVI was residualized out of the MSCVnorm variable. But I'm still not entirely clear about the rationale for doing so. On page 14, it's also noted that greater consonational variability necessitates controlling for local contrast variability. But again, why?

We have now provided additional information about the nPVI in the revised manuscript.

R3, Comment 6:

Also, is there something odd going on with the results as shown in Figure 2, the top-left figure? Is it the case that MSCVnorm cannot discriminate between SRC and WN for long time series? Isn't it supposed to?

The estimates for the SRC and WN time series (n=2048) are actually statistically different. We wanted to keep the y-axis consistent across the various facet plots, so the difference is qualitatively less distinct.

R3, Comment 7:

In the music/speech section, there really was no explicit mention of the lengths of the music and speech time series, other than that the corpora were selected because of extreme length limitations. How extreme are we talking about?

We have now added this information into the revision. The minimum duration amounts were 12 and 13 for the music and language corpora, accordingly. The average duration amounts were 20 and 27 durations for the music and language corpora, accordingly.

R3, Comment 8:

OTHER ISSUES:

MSCV is also a rather straightforward measure to implement and given it is being showcased in BRM I was expecting some sort of accompanying tool or software. At the very least, it may be worth making the Matlab code available on Github. Also, this manuscript suggests that MSCV has already been introduced elsewhere (e.g., Ross et al., in press). How novel is the presentation here in BRM - of the method itself - not in how you applied it?

We have provided the Matlab code on the 1st author's Github account. The link is provided in a footnote in the manuscript. The MSCV analysis was created and used for the Ross et al. project. Once we completed that project, we decided this analysis should be made accessible to the public and figured the most responsible/accessible route was through a methodological publication (and code provided on Github). The results in the applications section (music and language corpora) of the current manuscript are completely novel.

R3, Comment 9:

I also think that I'm missing something obvious in understanding the interpretation in the interim discussion. It has to do with the sentences on page 19 (of 28), from about lines 20 to 30. So, for consonational durations, you observe greater multiscale variability. But then you go on to say that this result may be because stress-timed languages have more complex syllables. But shouldn't it be because they have something more complex going on with consonants?

Yes, this was influenced by some of Ramus' previous work. We have now provided more detailed results and interpretations in the revised manuscript.

R3, Comment 10:

Minor issue, on page 7 (bottom), the full name for the ARMA acronym shouldn't have the "Fractionally-Integrated" bit.

In the simulation study, we actually did use ARFIMA models to generate the time series. We used specific parameters from a recent study (Torre, Varlet, & Marmelat, 2013). We have now added more detail to mitigate any confusion regarding this issue.