Perceiving and remembering speech depend on nonlinearity in movements producing and exploring speech

Lauren Bloomfield[1], Elizabeth Lane[1], Madhur Mangalam[2] and Damian G. Kelty-Stephen[1]

[1]Department of Psychology, Grinnell College, Grinnell, IA 50112, USA

[2]Department of Physical Therapy, Movement and Rehabilitation Sciences, Northeastern University, Boston, MA 02115, USA

Author for correspondence:

Damian G. Kelty-Stephen

e-mail: keltysda@grinnell.edu

Ethics statement. All participants provided informed consent approved by the Institutional Review Board (IRB) at Grinnell College, Grinnell, IA.

Data accessibility. All data analyzed in the present study are available as supplementary material.

Author contributions: L.B., E.L., and D.G.K-S. conceived and designed research; L.B., E.L., and D.G.K-S. performed experiments; L.B., E.L., M.M., and D.G.K-S. curated and analyzed data; L.B., E.L., M.M., and D.G.K-S. interpreted results of experiments; M.M. and D.G.K-S. prepared figures; L.B., E.L., M.M., and D.G.K-S. drafted manuscript; L.B., E.L., M.M., and D.G.K-S. edited and revised manuscript; L.B., E.L., M.M., and D.G.K-S. approved final version of manuscript.

Competing interests. The authors have no competing interests to declare.

Abstract

Speech perception and memory for speech require active engagement. Gestural theories have emphasized mainly the effect of speaker's movements on speech perception. They fail to address the effects of listener movement, focusing on communication as a boundary condition constraining movement among interlocutors. The present work attempts to break new ground by using multifractal geometry of physical action as a common currency for supporting both sides of the speaker-listener dyads. Participants self-paced their listening to a narrative, after which they completed a test of memory querying their narrative comprehension and the ability to recognize words from the story. The multifractal evidence of nonlinear interactions across timescales predicted the fluency of speech perception. Self-pacing movements that enabled listeners to control the presentation of speech sounds constituted a rich exploratory process. The multifractal nonlinearity of this exploration supported several aspects of memory for the perceived spoken language. These findings extend the role of multifractal geometry in the speaker’s actions to the narrative case of speech perception. In addition to posing novel basic research questions, these findings make a compelling case for imbuing multifractal structure in text-to-speech synthesizers for better perception and memory of speech.

Keywords: acoustic; auditory perception; complexity; learning disability; multifractality; nonlinear

1. Introduction

Listening to speech entails active engagement. Hidden within the seemingly passive act of listening, beyond vibrations from the eardrums, the listener consults the lexicon, a trove of possible words and their corresponding meaning [1]. Search of the lexicon becomes more overt, for example, through our eye movements [2]. The listener must explore the characteristics of the sounds themselves. With the increased use of text-to-speech synthesizers in support of students with learning disabilities, much of the human variability of sound characteristics are lost [3]. Listeners can potentially face an uncertain terrain when exploring the speech around them. The present study investigates how active exploration—yielding physical movements—supports perception and cognition of spoken words.

Speech perception critically depends on the articulatory gestures shaping those acoustic features [4–6]. Human motor processes reshape phonemes in a context- and sequence-dependent way (e.g., in coarticulation) that text-to-speech synthesis has struggled to emulate [7]. Speech sounds reflect movements beyond articulators to distal parts of the body [8]—potentially explaining their utility for diagnosing Parkinson's disease [9]. However, movement is now being recognized as an aspect of speech that otherwise passive listeners can use [10]. Gestural theories have been at most ambivalent about the role of listener's actions in perception and cognition of spoken words (e.g., compare [11] with [12,13]), most recently focusing on communication as a boundary condition that constrains movement among interlocutors [14–16].

The geometry of physical action is crucial for both sides of the speaker-listener dyads. For instance, passive though it seems, auditory perception has strong roots in a listener's bodywide—capability for physical action [17–25], and spoken-language perception depends on action [8,26–28]. Indeed, one early glimpse of this role for listener activity during speech perception is the evidence that spatial organization of speech in the space around the listener's body facilitates spoken-word perception under difficult listening circumstances, for example, with noise [29] or reverberation [30]. We can elicit overt exploratory behaviors, such as key pressing, during speech perception through the "self-paced listening" paradigm requiring participants to make key presses to signal readiness for consecutive words in a sequence [31]. Self-pacing affords the reader more control in exploring linguistic stimuli, with progress or pause indicating the waxing and waning of focus and processing [32].

We aim to align the movement supporting speech production with the movement involved in self-paced listening. The advantages of self-pacing in allocating attention and memory for the text [33] suggest that self-pacing movements could be crucial to sustaining memory of the ongoing speech stimuli [34]. The self-paced listening paradigm allows testing a general hypothesis that perception and cognition of spoken words might depend on the movements of speaker-listener dyads unfolding during conversation. We propose that a single geometric formalism—namely, multifractal geometry—might reveal a causal relationship between 1) motor processes underlying speech stimuli and 2) active exploration of those stimuli by the listeners. Multifractal structure might predict how well individual listeners process the spoken-word stimulus (figure 1*a*, top right) and later exhibit their memory for the spoken words (figure 1*a*, bottom left).

We use multifractality to test the hypotheses in two parts of our study. In Part-1, participants listened to a story, word-by-word, articulated by either a human voice or by a text-to-speech synthesizer. This manipulation allowed not only variations in speech-stream geometry across words but also across movement systems (i.e., natural human speech vs. computer-simulated analogue). In Part-2, they completed a memory test querying their comprehension of the story and memory of words from the story.

1.1. Multifractal geometry helps to formalize the movement-based support for conveying information

Multifractal geometry is a generalization of fractal geometry that encompasses 'multiple' degrees of fractal structure [35–37]. Rare in psycholinguistic research (e.g., [38–40]), variation in fractal structure has begun to demonstrate usefulness in modeling gestural supports for perceiving speech [5,6] as well as modeling how readers explore language in a narrative [41].

1.2. Multifractality offers statistical estimation of nonlinear interactions across scales, indicating how systems blend current, fleeting behaviors into longer-term behaviors

Multifractal analysis examines how variation depends on time in a measured behavior as well as how uniformly or how unevenly variation ranges from a brief glimpse to a longer time-scale of observation. Every observation of behavior depends on the length of observation [42]. Multifractal geometry evaluates variability through a variety of scales within a single measurement. For instance, the analysis segments a time-series measurement, such as audio waveform, into sequential but nonoverlapping bins of the same small size. Each bin represents a short subset of the whole movement at one short time scale. The analysis then estimates variability across those different bins of the same size. Then, multifractal geometry iterates this process for progressively longer timescales.

The average amount of variability in measurement grows as a power law of timescale. Only one power law relates average proportion to timescale for homogeneous time series in which brief and long-term events do not interact. However, for heterogeneous measurements of systems with interactions across scales, a single measurement can contain multiple proportion-to-time power laws. The word 'multifractal' reflects that heterogeneous systems can exhibit multiple power-law relationships each with different fractional exponents. This variety of fractional power-law exponents constitutes a 'multifractal spectrum.'

Two multifractal statistics indicate the strength of relationships between fast and slow events propagating brief events across longer timescales. The first metric, the multifractal spectrum width $\Delta\alpha$, is conceptually analogous to standard deviation but reflects a mixture of linear variability and nonlinear variability. The linear variability concerns histogram width and average drift over time. But the nonlinear variability crucially concerns interactions across scales. Another metric 'multifractal nonlinearity' t_{MF} is a *t*-statistic indicating difference of $\Delta\alpha$ from a sample of expected $\Delta\alpha$ values from strictly linear models of the measurement [36,37]. Bottom panels of figures 2*b* and 2*c* schematize this difference between linear structure and nonlinear interactions across timescales.

Multifractality is the natural, logical outcome when different timescales interact. Both speech production and language perception exemplify just such interactions across timescales. Communicative sounds exhibit a range of fractional power-law exponents [43]. Multifractality nonlinearity t_{MF} may predict expressiveness [44]. Longer-term speech sequences reshape the brief articulation of a single syllable [7]. Speech perception adjusts to compensate for coarticulatory interaction across time [4] and tailors these compensations over longer timescales [45]. Indeed, research has already modeled effects of multifractal structure in supporting speech perception [4–6].

1.3. Multifractal fluctuations in measurements of movement predict how movements support the pick-up, coordination, and later use of information for perceptual judgments

Multifractal geometry has repeatedly captured predictive information about how movement systems loosen their constraints to absorb and then propagate perceptual information. For example, this predictive capacity can be seen in haptic [46,47] and visual perception [48,49], as well as how movement systems integrate information across the body [50,51] and across different modalities [49]. In psycholinguistic domains, greater multifractal nonlinearity in self-paced reading may keep readers'

expectations loose and open-minded enough to navigate unexpected plot twists more fluently, (i.e., with a lower increase in subsequent reading-time), indicating a lighter processing load [41].

Multifractal fluctuations support the sharing of information necessary for dexterous behavior across organisms [52,53], just like they do within organisms. Indeed, attention to a speaker engenders a fractal-like 'complexity matching' supporting the listener's comprehension of linguistic content in the signal [54–56]. There has been growing curiosity in how these findings relate to insights from the multifractal formalism [56,57]. We now test whether such multifractal geometry supports this social-information exchange in speech production and perception.

1.4. Two hypotheses: Multifractality nonlinearity supports the uptake and memory for spoken-language

1.4.1. Hypothesis-1: Greater multifractal nonlinearity of speech sounds predicts more fluent response to individual words, prolonging fluency of perceiving text-to-speech synthesized words comparable to human-voice words

We used a self-paced listening task to experimentally test the effect of variations in multifractality—both from word to word within the same speaker and between two speakers with different styles of producing speech. We enlisted a human speaker and a text-to-speech synthesizer to pronounce each word of a narrative separately but in sequence. Despite being an effective learning tool for students with difficulty reading by sight [58,59], text-to-speech synthesizers can require more time for listeners to process the unfamiliar speech-production style [60–62]. Because of the potential role of multifractal nonlinearity in supporting expressivity, we predicted that multifractal nonlinearity in speech would speed the word-by-word response in self-paced listening (Hypothesis-1a). Likewise, we predicted that multifractal nonlinearity would support the fluency of text-to-speech synthesized speech and defer any gradual slowing of the self-pacing in response to text-to-speech speech relative to human speech (Hypothesis-1b). We explored any effects of multifractal-spectrum width without anticipating any single direction.

1.4.2. Hypothesis-2: Greater multifractal-spectrum width and greater nonlinearity of the self-paced key press interval time series would predict stronger performance on post-narrative test of comprehension and recognition memory

We predicted that key presses for self-pacing of speech stimuli would indicate participants' purposeful exploration of speech. We predicted that stronger evidence of multifractal nonlinearity across key-press sequences would predict better performance on a post-narrative. We predicted that stronger multifractal nonlinearity would entail better memory performance due to richer interaction across timescales (e.g., capacity for brief events to propagate over long periods of time). We explored possible effects of multifractal-spectrum width without anticipating any single direction.

2. Materials and methods

2.1. Participants

Nine healthy men and eleven healthy women (*mean*±*s.d.* age, 20.10±1.29 years) participated in the present study after providing informed consent approved by the Institutional Review Board (IRB) at Grinnell College, Grinnell, IA. Three participants reported having a learning disability.

2.2. Experimental task and procedure

An adult woman and Acapela's U.S. English text-to-speech female voice 'Sharon' (Acapela Inc., Mons, Belgium) using iPad app 'Voice Dream' both produced speech recordings of 2,027 words in sequence from *The Atlantic* article "Torching the Modern-Day Library of Alexandria." Human speech

and text-to-speech both produced words interspersed with pauses to allow parsing. Multiple standard psycholinguistic features of these words were assessed: word frequency, 'logWF,' in SUBTLEX [63]; and number of orthographic and phonological neighbors—'OrthN' and 'PhonN,' respectively—in CLEARPOND [64].

E-Prime software (Psychology Software Tools Inc., Pittsburgh, PA) presented individual audio recordings of each word in original sequence through headphones. Participants sat at an E-Prime-ready computer and were instructed: "Listen to the audio stimuli and press the spacebar after you feel as though you have understood the word you just heard. Try to pay attention to the passage because comprehension and word-memory questions will be asked at the end of the experiment. However, if you miss a word, do not worry and continue to move on because you cannot go back." Word recordings played only once immediately as the spacebar was released. Participants could press the spacebar again, either while or after the recording played, to hear the next word. E-Prime recorded the response time (RT) in milliseconds (up to one decimal place) from each word onset to subsequent button press. After the completion of the self-paced listening task, participants completed a 20-item pen-and-paper test. The first ten items prompted participants to indicate whether each of the ten sentences was 'true' or 'false' based on their comprehension of the narrative. The later ten items prompted participants to demonstrate word-recognition memory by indicating whether each of 10 words appeared in the text.

2.3. Multifractal metrics

We computed multifractal metrics for the absolute value of each spoken word's audio waveform (whether from human speech or text-to-speech) and each participant's RT sequence.

2.3.1. Multifractal spectra width

Chhabra and Jensen's direct method [65] samples non-negative series $u(t)$ at progressively larger scales such that proportion of signal $P_i(L)$ falling within the ith bin of scale L is

$$P_i(L)=\frac{\sum_{k=(i-1)L+1}^{iL} u(k)}{\sum u(t)} \qquad (1)$$

As L increases, $P_i(L)$ represents progressively larger proportion of $u(t)$,

$$P(L)\propto L^{\alpha} \qquad (2)$$

suggesting growth of proportion according to one 'singularity' strength α. $P(L)$ exhibits multifractal dynamics when it grows heterogeneously across timescales L according to multiple potential fractional singularity strengths, such that

$$P_i(L)\propto L^{\alpha_i} \qquad (3)$$

whereby each ith bin may show a distinct relationship of $P(L)$ with L. The spectrum of singularities is itself the multifractal spectrum, and its width $\Delta\alpha$ ($\alpha_{max}-\alpha_{min}$) indicates the heterogeneity of these relationships between proportion and timescale [35,66].

This method estimates $P(L)$ for N_L nonoverlapping bins of L-sizes and accentuates higher or lower $P(L)$ by using parameter $q>1$ and $q<1$, respectively, as follows

$$\mu_i(q,L)=\frac{[P_i(L)]^q}{\sum_{i=1}^{N_L}[P_i(L)]^q} \quad . \qquad (4)$$

See *figure* 2.

$\alpha(q)$ is the singularity for $\mu(q)$ -weighted $P(L)$ estimated by

$$\alpha(q) = -\lim_{L\to\infty} \frac{1}{\ln L} \sum_{i=1}^{N} \mu_i(q,L) \ln P_i(L)$$

$$= \lim_{L\to 0} \frac{1}{\ln L} \sum_{i=1}^{N} \mu_i(q,L) \ln P_i(L) \quad . \qquad (5)$$

Estimates $\alpha(q)$ belong to the multifractal spectrum if Shannon entropy of $\mu(q,l)$ scales with L according to a dimension $f(q)$, where

$$f(q) = -\lim_{L\to\infty} \frac{1}{\ln L} \sum_{i=1}^{N} \mu_i(q,L) \ln \mu_i(q,L)$$

$$= \lim_{L\to 0} \frac{1}{\ln L} \sum_{i=1}^{N} \mu_i(q,L) \ln \mu_i(q,L) \quad . \qquad (6)$$

For q generating scaling relationships (Eqs. 5 and 6) with correlation coefficient, $r > 0.95$, the parametric curve $(\alpha(q), f(q))$ or $(\alpha, f(\alpha))$ constitutes the multifractal spectrum with width $\Delta\alpha = \alpha_{max} - \alpha_{min}$ (figure 2, bottom right).

2.3.2. Multifractal nonlinearity

To identify whether nonzero $\Delta\alpha$ reflected nonlinear temporal correlations [67,68], $\Delta\alpha$ of each original series, was compared to $\Delta\alpha$ of 32 surrogate series obtained using Iterated Amplitude Adjusted Fourier Transformation (IAAFT) [36,37]. IAAFT generates surrogates that randomize phase ordering of the series' spectral amplitudes while preserving only linear temporal correlations. If $\Delta\alpha$ of the original series exceeds a 95% confidence interval (CI) from $\Delta\alpha$ of IAAFT series (i.e., $p < 0.05$), then the original series is considered to have nonzero nonlinear correlations, easily quantifiable using the one-sample *t*-statistic (henceforth, t_{MF}) comparing $\Delta\alpha$ of the original series to that of the surrogates (figure 3).

2.4. Statistical analysis

To test Hypothesis-1, a Poisson regression of self-paced word RT using glmer() in the package *lme4* [69] for *R* [70] included effects of logarithmic sound length (lnSL), word frequency (logWF), number of phonological neighbors (PhonN), number of orthographic neighbors (OrthoN), text-to-speech (TTS = 1 and 0 for synthesizer and human speech, respectively), and t_{MF}. It also included two time predictors spanning trials across the experiment: word# (word number incrementing by 1 with each word) and, because RT shows a general power-law decrease across progressive words for self-paced processing of narrative text [71], 1/word#. The model included the interaction effects of TTS×word#, as well the interactions 1/word#×TTS×logWF, 1/word#×TTS×PhonN, 1/word#×TTS×OrthoN, 1/word#×TTS×t_{MF}, in addition to all other lower-order interactions. Further interactions either were nonsignificant or led to poor model convergence.

To test Hypothesis-2, a Poisson regression of cumulative correct answers in the post-narrative test included effects of test type (WRM = word recognition memory as opposed to comprehension), condition (TTS = having heard the narrative in text-to-speech rather than human speech), orthogonal cubic polynomials of questions 1 through 10 (calculated using the *R*-function poly() [70]) and various metrics on individual participants' self-paced RT series during the narrative including linear features (mean and median) and multifractal metrics ($\Delta\alpha$ and t_{MF}). Modeling tested typeWRM×conditionTTS×RT-$\Delta\alpha$, typeWRM×conditionTTS×RT-t_{MF}, conditionTTS×log(RT-mean), conditionTTS×log(RT-median), as well as question×typeWRM and question×$\Delta\alpha$, and all lower-order

interactions. This model also included a dichotomous predictor 'LD-status' to control for any difference participants self-reporting a learning disability might show. Alternate modeling using principal components to minimize correlation amongst predictors (i.e., between RT-mean and RT-median) showed comparable directions and patterns of significance in effects.

Both models used 1,000 resamplings to generate bootstrapped standard errors, *p*-values, and power estimates. All significant predictors showed at least 80% statistical power (i.e., > 800 resamplings yielded nonzero coefficients in the same direction as the original data) [72].

3. Results

Figure 4 shows illustrative output from multifractal analysis of both speech and RT sequences.

3.1. Model of word-by-word RT

Table 1 lists this model's coefficients.

3.1.1. Testing Hypothesis-1a: Multifractal nonlinearity sped the response in self-paced listening

Speech stimuli with stronger evidence of multifractal nonlinearity evoked faster responses (t_{MF}: $B = -2.77\times10^{-5}$, $p < 0.01$), although the difference was small, ranging from 0.90% to 2.74% as multifractal nonlinearity varied from 1st to 3rd quartiles (table 1). We found comparable effects of word frequency (logWF: $B = -3.22\times10^{-2}$, $p < 0.0001$) and phonological neighborhood size (PhonN: $B = -3.97\times10^{-4}$, $p < 0.05$) in speech perception. RT increased with logarithmically-scaled sound length (lnSL: $B = 3.44\times10^{-1}$, $p < 0.0001$): it takes longer to listen to longer sounds.

3.1.2. Testing Hypothesis-1b: Multifractal nonlinearity deferred the slower response to text-to-speech synthesis—later than other psycholinguistic features did

Text-to-speech posed additional challenges to speech perception. Text-to-speech evoked slower self-paced responses (TTS: $B = -7.16\times10^{-2}$, $p < 0.0001$; TTS×word#: $B = 1.42\times10^{-4}$, $p < 0.0001$) with RTs dwindling with power-law-like decay (e.g., [71]; 1/word#: $B = 1.02\times10^{0}$, $p < 0.001$) leading to a gradual divergence of 95% CI between text-to-speech and human speech (figure 5*a*). This decay was stronger for words spoken with greater multifractal nonlinearity (1/word#×TTS×t_{MF}, $B = 1.81\times10^{-3}$, $p < 0.01$). Hence, greater multifractal nonlinearity led the self-paced listening to resist the progressive linear increase in RT with word# in text-to-speech. Consequently, the expected slower rate of self-paced listening in the Text-to-speech condition only appeared gradually, with 95% CI for the predicted RT diverging on word #1140 for low t_{MF} (1st quartile; thin gray vertical line figure 5*a*) and on word #1280 for high t_{MF} (3rd quartile; thick gray vertical line in figure 5*a*). Hence, varying from low to high multifractal nonlinearity kept the pace of perceptual processing of human speech and text-to-speech speech comparable for 140 words, or 12.28% longer.

However, on the way towards text-to-speech becoming more cumbersome for perceptual processing, some words were easier to process than others in text-to-speech than human speech—different features led to different divergence of 95% CI. Multifractality prevented this divergence, prolonging the fluency of perceiving text-to-speech. Phonological neighborhood showed the most persistent effect in supporting the greater processing load in text-to-speech (1/word#×TTS×PhonN: $B = 8.21\times10^{-2}$, $p < 0.001$). Response to text-to-speech relative to human speech became much slower first for the words with few phonological neighbors (i.e., 1st quartile of number of words sharing all but one of the same phonemes, low PhonN; thin gray vertical line in figure 5*b*). This shift occurred just about halfway through the text. For roughly 250 words afterward, words with larger phonological neighborhoods (3rd quartile, high PhonN; thick gray vertical line in figure 5*b*) elicited the self-paced response to text-to-speech not significantly different from the response to human speech, but then at the

end of those next 250 words, the benefit of phonological neighborhood size dissipated. Frequency showed a comparable effect (TTS×logWF: $B = 1.65\times10^{-2}$, $p < 0.0001$; 1/word#×logWF: $B = 2.84\times10^{-1}$, $p < 0.0001$). Low-frequency (1st quartile) words began prompting slower responses in text-to-speech than human speech (thick gray vertical line in figure 5*c*). High-frequency words (3rd quartile) began to show an increase in RT in text-to-speech relative to human speech only 40 words later (thick gray vertical line in figure 5*c*). Finally, orthographic neighborhood size exhibited no significant effect ($p > 0.05$), likely reflecting the auditory rather than the visual form of the task.

3.2. Model of cumulative correct responses on the 20-item two-part memory test

Table 2 lists this model's coefficients.

Any benefit of multifractality of response times in self-paced listening for subsequent memory of the narrative was attributable to nonlinearity and strongest in word-recognition memory (figure 6). Multifractality from linear sources (i.e., $\Delta\alpha$) diminished memory performance on average ($B = -2.31\times10^{0}$, $p < 0.001$) especially in word-recognition memory ($B = -4.57\times10^{0}$, $p < 0.0001$; figure 6*b*). Multifractal nonlinearity showed no effect on narrative-comprehension memory ($p > 0.05$; figure 7*a*)but promoted word-recognition memory following self-paced listening to human speech ($B = 2.77\times10^{-2}$, $p < 0.0001$; figure 7*b*). Word-recognition memory became 24% to 147% more accurate from 1st to 3rd quartiles of RT-t_{MF} (table 2). Self-paced listening to text-to-speech made multifractal nonlinearity in RT sequences actually predict weaker memory performance ($B = -1.37\times10^{-2}$, $p < 0.01$), especially in word recognition ($B = -3.06\times10^{-2}$, $p < 0.0001$), and linear sources of multifractality promoted word-recognition memory (typeWRM×conditionTTS×RT-Δα: $B = 2.40\times10^{-2}$, $p < 0.05$; table 2).

Besides the effects on average memory performance, multifractality in response times of self-paced listening predicted the improvement of memory performance as the task continued. Nonetheless, this effect was strictly limited to multifractality from linear sources (i.e., $\Delta\alpha$; poly(Question,3)1×RT-$\Delta\alpha$: $B = 16.47\times10^{0}$, $p < 0.01$; figure 6). Multifractal nonlinearity did not show significant effects (table 2; figure 7). This effect of $\Delta\alpha$ influenced only the linear increase in memory performance with successive questions (poly(Question, 3)1: $B = 8.84\times100$, $p < 0.0001$; figure 6). The quadratic and cubic increases in memory performance with successive questions appeared only in the word-recognition memory test (poly(Question, 2)1×typeWRM: $B = -3.17\times10^{0}$, $p < 0.0001$; poly(Question, 3)1×typeWRM: $B = 1.16\times10^{0}$, $p < 0.05$; figures 6 and 7), indicating that recognizing one word likely cued more accurate recognition of the next word.

Control effects included small effects of central tendency of RTs and positive effects of word-recognition memory, learning disability, and text-to-speech. Logarithmically-scaled mean and median of RT canceled each other out in both human speech (log(RT-mean): $B = 3.35\times10^{0}$, $p < 0.0001$; log(RT-median): $B = -3.89\times10^{0}$, $p < 0.0001$) and text-to-speech (TSS×log(RT-mean): $B = -5.75\times10^{0}$, $p < 0.0001$; TSS×log(RT-median): $B = 5.40\times10^{0}$, $p < 0.0001$). Summing mean and median effects, 3rd quartiles led to less accurate memory performance than 1st quartile by 0.28% and 0.04% in human speech and text-to-speech, respectively (table 2). The model also controlled for positive effects of word-recognition memory test ($B = 2.88\times10^{-1}$, $p < 0.05$), learning-disability status ($B = 3.75\times10^{-1}$, $p < 0.0001$), and text-to-speech ($B = 2.84\times10^{0}$, $p < 0.05$). Word recognition was easier, and participants with learning disabilities may have benefited more than others from the control on speech stimuli from self-paced format (e.g., [33]). Text-to-speech synthesizers might have supported more effective learning with self-pacing, whereas controlling the pacing of human speech might have been distracting.

4. Discussion

We tested two hypotheses about the effects of multifractal nonlinearity t_{MF} in perception and memory for spoken language. Hypothesis-1 was that greater t_{MF} in speech stimuli would make speech perception more fluent, particularly in keeping responses text-to-speech synthesis as fluent as responses to human speech. Hypothesis-2 was that greater t_{MF} in the RT sequence in self-paced listening to a narrative would support better memory of the speech stimuli. Results supported Hypothesis-1 in full but Hypothesis-2 only in the word-recognition part of the memory test and not in the narrative-comprehension portion of the memory test. These findings carry implications not only for basic research into the movement basis for speech perception and memory for speech but also for the development of more effective text-to-speech synthesis for students with learning disabilities.

4.1. Multifractal nonlinearity promotes quicker response to speech stimuli and processing of text-to-speech synthesis later than traditional psycholinguistic features do

Multifractal nonlinearity of speech was associated with more fluent speech perception across a narrative. Whether the speaker was human or machine, the inevitable interactions across timescales as movement systems produce speech sounds (human speech and text-to-speech) promoted spoken-word perception in narrative sequence. Not only articulatory gestures matter to perception [10,11,14–16], the perceptual relevance of gestures goes deeper than linguistic distinctions (e.g., fricatives or stops and voice-onset times) and is rooted in more generic nonlinear dynamics with no currently recognized linguistic meaning [5,6]. The absence of any effects of multifractal spectrum width suggests that this aspect of speech perception depends on nonlinear and not linear relationships across timescales.

Multifractality manifested its role in supporting the processing of synthesized speech later than frequency and phonological neighborhoods, but longer than frequency supported. This result likely reflects that multifractal nonlinearity is a deeper attribute of acoustic waveform. Frequency and phonological neighborhood may support the fluent perception of less intelligible speech early on because these are high-level features anchored in lexicality (i.e., word status). Meanwhile, multifractal nonlinearity is not specific to words and inheres in the heterogeneous movements producing speech sounds. The mind only delves into this deeper movement foundation when linguistic features can no longer support intelligibility.

4.2. Multifractal nonlinearity in self-paced listening to human speech promoted word-recognition memory, and multifractality from linear sources predicted diminished memory on average but increased memory performance as testing continued

Participant's ability to remember the speech stimuli exhibited a more complex relationship between movement systems across the speaker-listener dyad. The two movement systems generating speech (i.e., natural human speech vs. computer-simulated analogue) accrued distinct aspects of the multifractal structure of self-pacing responses: the more intelligible human speech accrued the nonlinear aspects of multifractal self-pacing behaviors (i.e., t_{MF}), and the less intelligible text-to-speech accrued linear aspects (i.e., $\Delta\alpha$). Self-paced listening to speech presented word by word may have emphasized individual words at the expense of attention to the narrative's discourse-level aspects.

5. Conclusions

The present work is the latest step in the new enterprise of modeling multifractal effects supporting spoken language use. It addresses current calls in the literature for a straightforward multifractal extension of the notion that a speaker-listener dyad depends on 'complexity matching' between the respective movement systems [54–57]. It rests on a long history of investigating the multifractal support for coordinating perceptual information within and across organisms [46–51,73].

We have threaded a multifractal relationship from the speaker to the listener. First, the multifractal evidence of nonlinear interactions across timescales predicts the fluency of speech perception. Second, self-pacing movements that enabled listeners to control the presentation of speech sounds constituted a rich exploratory process. Multifractal nonlinearity of this exploration supports some but not yet all aspects of memory for perceived spoken language. Nonetheless, the present findings extend the role of multifractal geometry in the speaker's actions to the narrative case of speech perception. To the point, our findings reflect a clearer glimpse into the role of the listener's action in speech perception.

Self-paced listening does not approximate speech perception "in the wild" (e.g., a listener listening to a speaker producing words at the speaker's own preferred pace). However, multifractality by no means characterizes only self-paced language perception. Future work could examine postural or head sway in listeners to determine whether these subtle movements carry multifractal signatures sensitive to speech similar to haptic information (e.g., [46,47,50,51,73]). Even admitting the contrived status of self-paced listening, the present results validate our idea that speech perception and memory reflects an alignment (or not) between the structures of the speaker's and the listener's movements. Speech memory reflected a compromise between the speaker's movement system and the listener's actions' multifractality. The self-pacing paradigm prompting the listener to act to get the next word coaxes some action whose geometry is also multifractal. Different movement systems can accrue specifically different (i.e., linear or nonlinear) aspects of the multifractality in self-paced timing. A more familiar speaker (i.e., human speech) sooner leads to multifractal nonlinearity to support speech memory. In contrast, the less familiar, less intelligible speaker (i.e., text-to-speech [3]) sooner leads to linear sources of multifractality to support speech memory.

Future work could investigate aspects of human or text-to-speech voices influencing relative linear or nonlinear effects of multifractality on perception and subsequent memory for speech. For instance, multifractality could enhance text-to-speech synthesis as a learning tool. The present evidence that multifractality of speech movements is essential for speech perception is silent on whether multifractality of speech movements have consequences for memory. Unambiguous comparison of multifractality across dyads was precluded by the radically different data formats between speaker and listener recordings (multiple 44.1 kHz digital recordings vs. single sequences of 2,026 inter-key press intervals). Whether this alignment rests firmly on multifractality could be tested in a 'shadowing' paradigm (i.e., listeners repeating immediately each spoken word (e.g., [74]).

The observed multifractal effects stand alongside and not in tension with traditional psycholinguistic features. The invocation of multifractal geometry situates the present work theorizing on the multi-scaled mechanics of physiological systems [24,75], going deeper into the physiological weeds than prior calls to implicate embodiment of perception. We feel that it elaborates the bodily 'meshing' of perceptual and cognitive processes [34]. The peaceful coexistence of significant multifractal with significant psycholinguistic predictors likely signals that the multifractal approach is an option rather than a necessity. Nonetheless, we hope that multifractal elaborations of bodily 'mesh' supporting perception and cognition is a rich potential avenue of research.

References

1. Aitchison J. 2012 *Words in the Mind: An Introduction to the Mental Lexicon*. Oxford, UK: Wiley-Blackwell.

2. Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, Sedivy JC. 1995 Integration of visual and linguistic information in spoken language comprehension. *Science* **268**, 1632–1634. (doi:10.1126/science.7777863)

3. Paul D, Shifas MP V, Pantazis Y, Stylianou Y. 2020 Enhancing speech intelligibility in text-to-speech synthesis using speaking style conversion. *arXiv Prepr. arXiv2008.05809*

4. Viswanathan N, Magnuson JS, Fowler CA. 2014 Information for coarticulation: Static signal properties or formant dynamics? *J. Exp. Psychol. Hum. Percept. Perform.* **40**, 1228–1236. (doi:10.1037/a0036214)

5. Hasselman F. 2015 Classifying acoustic signals into phoneme categories: Average and dyslexic readers make use of complex dynamical patterns and multifractal scaling properties of the speech signal. *PeerJ* **3**, e837. (doi:10.7717/peerj.837)

6. Ward RM, Kelty-Stephen DG. 2018 Bringing the nonlinearity of the movement system to gestural theories of language use: Multifractal structure of spoken english supports the compensation for coarticulation in human Speech Perception. *Front. Physiol.* **9**, 1152. (doi:10.3389/fphys.2018.01152)

7. Hill DR, Taube-Schock CR, Manzara L. 2017 Low-level articulatory synthesis: A working text-to-speech solution and a linguistic tool. *Can. J. Linguist.* **62**, 371–410. (doi:10.1017/cnj.2017.15)

8. Pouw W, Paxton A, Harrison SJ, Dixon JA. 2020 Acoustic information about upper limb movement in voicing. *Proc. Natl. Acad. Sci.* **117**, 11364–11367. (doi:10.1073/pnas.2004163117)

9. Tsanas A, Little MA, McSharry PE, Ramig LO. 2011 Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. *J. R. Soc. Interface* **8**, 842–855. (doi:10.1098/rsif.2010.0456)

10. Galantucci B, Fowler CA, Turvey MT. 2006 The motor theory of speech perception reviewed. *Psychon. Bull. Rev.* **13**, 361–377. (doi:10.3758/BF03193857)

11. Fowler CA. 1996 Listeners do hear sounds, not tongues. *J. Acoust. Soc. Am.* **99**, 1730–1741. (doi:10.1121/1.415237)

12. Fadiga L, Craighero L, Buccino G, Rizzolatti G. 2002 Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *Eur. J. Neurosci.* **15**, 399–402. (doi:10.1046/j.0953-816x.2001.01874.x)

13. Fowler CA. 2008 The FLMP STMPed. *Psychon. Bull. Rev.* **15**, 458–462. (doi:10.3758/PBR.15.2.458)

14. Shockley K, Baker AA, Richardson MJ, Fowler CA. 2007 Articulatory constraints on interpersonal postural coordination. *J. Exp. Psychol. Hum. Percept. Perform.* **33**, 201–208. (doi:10.1037/0096-1523.33.1.201)

15. Shockley K, Santana M-V, Fowler CA. 2003 Mutual interpersonal postural constraints are involved in cooperative conversation. *J. Exp. Psychol. Hum. Percept. Perform.* **29**, 326–332. (doi:10.1037/0096-1523.29.2.326)

16. Fowler CA, Richardson MJ, Marsh KL, Shockley KD. 2008 Language use, coordination, and the emergence of cooperative action. In *Coordination: Neural, Behavioral and Social Dynamics* (eds A Fuchs, VK Jirsa), pp. 261–279. Berlin, Heidelberg: Springer. (doi:10.1007/978-3-540-74479-5_13)

17. Repp BH, Knoblich G. 2007 Action can affect auditory perception. *Psychol. Sci.* **18**, 6–7. (doi:10.1111/j.1467-9280.2007.01839.x)

18. Stoffregen TA, Pittenger JB. 1995 Human echolocation as a basic form of perception and action. *Ecol. Psychol.* **7**, 181–216. (doi:10.1207/s15326969eco0703_2)

19. Stoffregen TA, Villard S, Kim C, Ito K, Bardy BG. 2009 Coupling of head and body movement with motion of the audible environment. *J. Exp. Psychol. Hum. Percept. Perform.* **35**, 1221–1231. (doi:10.1037/a0014251)

20. Stoffregen TA, Ito K, Hove P, Yank JR, Bardy BG. 2010 The postural responses to a moving environment of adults who are blind. *J. Vis. Impair. Blind.* **104**, 73–83. (doi:10.1177/0145482X1010400203)

21. Wallmeier L, Wiegrebe L. 2020 Self-motion facilitates echo-acoustic orientation in humans. *R. Soc. Open Sci.* **1**, 140185. (doi:10.1098/rsos.140185)

22. Pawlukowska W, Siwczak K, Szylińska A, Turoń A, Rotter I. 2017 The impact of muscle tone on the development of auditory perception – research orientation. *Fam. Med. Prim. Care Rev.* **19**, 382–386. (doi:10.5114/fmpcr.2017.70812)

23. Myers JC, Mock JR, Golob EJ. 2020 Sensorimotor integration can enhance auditory perception. *Sci. Rep.* **10**, 1496. (doi:10.1038/s41598-020-58447-z)

24. Ingber DE. 2006 Cellular mechanotransduction: Putting all the pieces together again. *FASEB J.* **20**, 811–827. (doi:10.1096/fj.05-5424rev)

25. Boulenger V, Nazir TA. 2010 Interwoven functionality of the brain's action and language systems. *Ment. Lex.* **5**, 231–254. (doi:10.1075/ml.5.2.05bou)

26. McNeill D. 2016 *Why We Gesture: The Surprising Role of Hand Movements in Communication*. Cambridge, UK: Cambridge University Press.

27. Morett LM, Roche JM, Fraundorf SH, McPartland JC. 2020 Contrast is in the eye of the beholder: Infelicitous beat gesture increases cognitive load during online spoken discourse comprehension. *Cogn. Sci.* **44**, e12912. (doi:10.1111/cogs.12912)

28. Pouw W, Harrison SJ, Esteve-Gibert N, Dixon JA. 2020 Energy flows in gesture-speech physics: The respiratory-vocal system and its coupling with hand gestures. *J. Acoust. Soc. Am.* **148**, 1231–1247. (doi:10.1121/10.0001730)

29. Viswanathan N, Kokkinakis K, Williams BT. 2018 Listeners experience linguistic masking release in noise-vocoded speech-in-speech recognition. *J. Speech, Lang. Hear. Res.* **61**, 428–435. (doi:10.1044/2017_JSLHR-H-17-0215)

30. Viswanathan N, Kokkinakis K. 2019 Listening benefits in speech-in-speech recognition are altered under reverberant conditions. *J. Acoust. Soc. Am.* **145**, EL348–EL353. (doi:10.1121/1.5100898)

31. Ferreira F, Henderson JM, Anes MD, Weeks PA, McFarlane DK. 1996 Effects of lexical frequency and syntactic complexity in spoken-language comprehension: Evidence from the auditory moving-window technique. *J. Exp. Psychol. Learn. Mem. Cogn.* **22**, 324–335. (doi:10.1037/0278-7393.22.2.324)

32. Unsworth N, Robison MK. 2016 Pupillary correlates of lapses of sustained attention. *Cogn. Affect. Behav. Neurosci.* **16**, 601–615. (doi:10.3758/s13415-016-0417-4)

33. Piquado T, Benichov JI, Brownell H, Wingfield A. 2012 The hidden effect of hearing acuity on speech recall, and compensatory effects of self-paced listening. *Int. J. Audiol.* **51**, 576–583. (doi:10.3109/14992027.2012.684403)

34. Glenberg AM. 1997 What memory is for. *Behav. Brain Sci.* **20**, 1–19. (doi:10.1017/S0140525X97000010)

35. Mandelbrot BB. 1999 *Multifractals and 1/f Noise*. New York, NY: Springer-Verlag.

36. Ihlen E. 2012 Introduction to multifractal detrended fluctuation analysis in Matlab. *Front. Physiol.* **3**, 141. (doi:10.3389/fphys.2012.00)

37. Mandic DP, Chen M, Gautama T, Van Hulle MM, Constantinides A. 2008 On the characterization of the deterministic/stochastic and linear/nonlinear nature of time series. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **464**, 1141–1160. (doi:10.1098/rspa.2007.0154)

38. Kello CT, Anderson GG, Holden JG, Van Orden GC. 2008 The pervasiveness of 1/*f* scaling in speech reflects the metastable basis of cognition. *Cogn. Sci.* **32**, 1217–1231. (doi:https://doi.org/10.1080/03640210801944898)

39. Van Orden GC, Holden JG, Turvey MT. 2003 Self-organization of cognitive performance. *J. Exp. Psychol. Gen.* **132**, 331–350. (doi:10.1037/0096-3445.132.3.331)

40. Gilden DL. 2001 Cognitive emissions of 1/*f* noise. *Psychol. Rev.* **108**, 33–56. (doi:10.1037/0033-295X.108.1.33)

41. Booth CR, Brown HL, Eason EG, Wallot S, Kelty-Stephen DG. 2018 Expectations on hierarchical scales of discourse: Multifractality predicts both short- and long-range effects of violating gender expectations in text reading. *Discourse Process.* **55**, 12–30. (doi:10.1080/0163853X.2016.1197811)

42. Adolph KE, Robinson SR. 2011 Sampling development. *J. Cogn. Dev.* **12**, 411–423. (doi:10.1080/15248372.2011.608190)

43. Kello CT, Bella SD, Médé B, Balasubramaniam R. 2017 Hierarchical temporal structure in music, speech and animal vocalizations: Jazz is like a conversation, humpbacks sing like hermit thrushes. *J. R. Soc. Interface* **14**, 20170231. (doi:10.1098/rsif.2017.0231)

44. Roeske TC, Kelty-Stephen D, Wallot S. 2018 Multifractal analysis reveals music-like dynamic structure in songbird rhythms. *Sci. Rep.* **8**, 4570. (doi:10.1038/s41598-018-22933-2)

45. Viswanathan N, Kelty-Stephen DG. 2018 Comparing speech and nonspeech context effects across timescales in coarticulatory contexts. *Attention, Perception, Psychophys.* **80**, 316–324. (doi:10.3758/s13414-017-1449-8)

46. Teng DW, Eddy CL, Kelty-Stephen DG. 2016 Non-visually-guided distance perception depends on matching torso fluctuations between training and test. *Attention, Perception, Psychophys.* **78**, 2320–2328. (doi:10.3758/s13414-016-1213-5)

47. Doyon JK, Hajnal A, Surber T, Clark JD, Kelty-Stephen DG. 2019 Multifractality of posture modulates multisensory perception of stand-on-ability. *PLoS One* **14**, e0212220. (doi:10.1371/journal.pone.0212220)

48. Jacobson N, Berleman-Paul Q, Mangalam M, Kelty-Stephen DG. 2020 Multifractality in postural sway supports quiet eye training in aiming tasks: A study of golf putting. *bioRxiv* , 258053. (doi:10.1101/2020.08.19.258053)

49. Carver NS, Bojovic D, Kelty-Stephen DG. 2017 Multifractal foundations of visually-guided aiming and adaptation to prismatic perturbation. *Hum. Mov. Sci.* **55**, 61–72. (doi:10.1016/j.humov.2017.07.005)

50. Mangalam M, Carver NS, Kelty-Stephen DG. 2020 Multifractal signatures of perceptual processing on anatomical sleeves of the human body. *J. R. Soc. Interface* **17**, 20200328. (doi:10.1098/rsif.2020.0328)

51. Mangalam M, Kelty-Stephen DG. 2020 Multiplicative-cascade dynamics supports whole-body coordination for perception via effortful touch. *Hum. Mov. Sci.* **70**, 102595. (doi:10.1016/j.humov.2020.102595)

52. Stephen DG, Dixon JA. 2011 Strong anticipation: Multifractal cascade dynamics modulate scaling in synchronization behaviors. *Chaos, Solitons & Fractals* **44**, 160–168. (doi:10.1016/j.chaos.2011.01.005)

53. Carver NS, Kelty-Stephen DG. 2017 Multifractality in individual honeybee behavior hints at colony-specific social cascades: Reanalysis of radio-frequency identification data from five different colonies. *Phys. Rev. E* **95**, 22402. (doi:10.1103/PhysRevE.95.022402)

54. Abney DH, Paxton A, Dale R, Kello CT. 2014 Complexity matching in dyadic conversation. *J. Exp. Psychol. Gen.* **143**, 2304–2315. (doi:10.1037/xge0000021)

55. Ramirez-Aristizabal AG, Médé B, Kello CT. 2018 Complexity matching in speech: Effects of speaking rate and naturalness. *Chaos, Solitons & Fractals* **111**, 175–179. (doi:10.1016/j.chaos.2018.04.021)

56. Abney DH, Kello CT, Warlaumont AS. 2015 Production and convergence of multiscale clustering in speech. *Ecol. Psychol.* **27**, 222–235. (doi:10.1080/10407413.2015.1068653)

57. Schloesser DS, Kello CT, Marmelat V. 2019 Complexity matching and coordination in individual and dyadic performance. *Hum. Mov. Sci.* **66**, 258–272. (doi:10.1016/j.humov.2019.04.010)

58. Wood SG, Moxley JH, Tighe EL, Wagner RK. 2017 Does use of text-to-speech and related read-aloud tools improve reading comprehension for students with reading disabilities? A meta-analysis. *J. Learn. Disabil.* **51**, 73–84. (doi:10.1177/0022219416688170)

59. Košak-Babuder M, Kormos J, Ratajczak M, Pižorn K. 2018 The effect of read-aloud assistance on the text comprehension of dyslexic and non-dyslexic English language learners. *Lang. Test.* **36**, 51–75. (doi:10.1177/0265532218756946)

60. Wester M, Watts O, Henter GE. 2016 Evaluating comprehension of natural and synthetic conversational speech. In *Proc. Speech Prosody*, pp. 736–740.

61. Hux K, Knollman-Porter K, Brown J, Wallace SE. 2017 Comprehension of synthetic speech and digitized natural speech by adults with aphasia. *J. Commun. Disord.* **69**, 15–26. (doi:10.1016/j.jcomdis.2017.06.006)

62. Giannouli V, Banou M. 2020 The intelligibility and comprehension of synthetic *versus* natural speech in dyslexic students. *Disabil. Rehabil. Assist. Technol.* **15**, 898–907. (doi:10.1080/17483107.2019.1629111)

63. Brysbaert M, New B. 2009 Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behav. Res. Methods* **41**, 977–990. (doi:10.3758/BRM.41.4.977)

64. Marian V, Bartolotti J, Chabal S, Shook A. 2012 CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PLoS One* **7**, e43230. (doi:10.1371/journal.pone.0043230)

65. Chhabra A, Jensen R V. 1989 Direct determination of the $f(\alpha)$ singularity spectrum. *Phys. Rev. Lett.* **62**, 1327–1330. (doi:10.1103/PhysRevLett.62.1327)

66. Halsey TC, Jensen MH, Kadanoff LP, Procaccia I, Shraiman BI. 1986 Fractal measures and their singularities: The characterization of strange sets. *Phys. Rev. A* **33**, 1141–1151. (doi:10.1103/PhysRevA.33.1141)

67. Kelty-Stephen DG, Palatinus K, Saltzman E, Dixon JA. 2013 A tutorial on multifractality, cascades, and interactivity for empirical time series in ecological science. *Ecol. Psychol.* **25**, 1–62. (doi:10.1080/10407413.2013.753804)

68. Veneziano D, Moglen GE, Bras RL. 1995 Multifractal analysis: Pitfalls of standard procedures and alternatives. *Phys. Rev. E* **52**, 1387–1398. (doi:10.1103/PhysRevE.52.1387)

69. Bates D, Mächler M, Bolker B, Walker S. 2015 Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **1**, 1–48. (doi:10.18637/jss.v067.i01)

70. Team RC. 2013 R: A language and environment for statistical computing. (doi:URL https://www.R-project.org/.)

71. Wallot S, O'Brien BA, Haussmann A, Kloos H, Lyby MS. 2014 The role of reading time complexity and reading speed in text comprehension. *J. Exp. Psychol. Learn. Mem. Cogn.* **40**, 1745–1765. (doi:10.1037/xlm0000030)

72. Efron B, Tibshirani R. 1986 Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.* **1**, 54–75.

73. Palatinus Z, Kelty-Stephen DG, Kinsella-Shaw J, Carello C, Turvey MT. 2014 Haptic perceptual intent in quiet standing affects multifractal scaling of postural fluctuations. *J. Exp. Psychol. Hum. Percept. Perform.* **40**, 1808–1818. (doi:10.1037/a0037247)

74. Liu H, Bates E, Powell T, Wulfeck B. 1997 Single-word shadowing and the study of lexical access. *Appl. Psycholinguist.* **18**, 157–180. (doi:10.1017/S0142716400009954)

75. Turvey MT, Fonseca ST. 2014 The medium of haptic perception: A tensegrity hypothesis. *J. Mot. Behav.* **46**, 143–187. (doi:10.1080/00222895.2013.798252)

Table 1. Coefficients of the Poisson regression of self-paced word RT[a] (section 3.1).

effect[b]	***mean***	***s.e.m.*** **(bootstrapped)**	***p***[c] **(bootstrapped)**
intercept	5.54×10^{0}	2.53×10^{-1}	$\mathbf{6.22\times10^{-107}}$
lnSL	3.44×10^{-1}	2.42×10^{-2}	$\mathbf{2.83\times10^{-46}}$
logWF	-3.22×10^{-2}	2.60×10^{-3}	$\mathbf{2.49\times10^{-35}}$
PhonN	-3.97×10^{-4}	2.06×10^{-4}	$\mathbf{2.73\times10^{-02}}$
OrthN	-8.26×10^{-5}	4.88×10^{-4}	4.33×10^{-1}
TTS	-7.16×10^{-2}	1.75×10^{-2}	$\mathbf{2.19\times10^{-5}}$
t_{MF}	-2.77×10^{-5}	1.06×10^{-5}	$\mathbf{4.60\times10^{-3}}$
TTS×logWF	1.65×10^{-2}	3.65×10^{-3}	$\mathbf{3.29\times10^{-6}}$
TTS×PhonN	5.66×10^{-5}	3.48×10^{-4}	4.35×10^{-1}
TTS×OrthN	7.42×10^{-4}	7.53×10^{-4}	1.62×10^{-1}
TTS×t_{MF}	3.46×10^{-5}	1.66×10^{-5}	$\mathbf{1.87\times10^{-2}}$
word#	-2.66×10^{-4}	5.44×10^{-6}	0.00×10^{0}
TTS×word	1.42×10^{-4}	8.04×10^{-6}	$\mathbf{2.55\times10^{-70}}$
1/word#	1.02×10^{0}	3.20×10^{-1}	$\mathbf{7.21\times10^{-4}}$
1/word#×logWF	2.84×10^{-1}	2.60×10^{-3}	$\mathbf{2.49\times10^{-35}}$
1/word#×PhonN	1.73×10^{-2}	1.96×10^{-2}	1.89×10^{-1}
1/word#×OrthN	-8.35×10^{-2}	5.83×10^{-2}	7.61×10^{-2}
1/word#×TTS	-4.06×10^{-1}	5.29×10^{-1}	2.21×10^{-1}
1/word#×t_{MF}	-1.43×10^{-4}	5.21×10^{-4}	3.92×10^{-1}
1/word#×TTS×logWF	-8.41×10^{-2}	1.97×10^{-1}	3.35×10^{-1}
1/word#×TTS×PhonN	8.21×10^{-2}	2.59×10^{-2}	7.65×10^{-4}
1/word#×TTS×OrthN	3.03×10^{-2}	8.40×10^{-2}	3.59×10^{-1}
1/word#×TTS×t_{MF}	1.81×10^{-3}	6.47×10^{-4}	$\mathbf{2.55\times10^{-3}}$

[a]model in *R* syntax: glmer(I(10*RT) ~ lnSL + I(1/word#)*logWF*condition + I(1/word#)*PhonN*condition + I(1/word#)*OrthN*condition + I(1/word#)*condition*t_{MF} + word#*condition + (1|participant), data = DATA, family = poisson).

[b]abbreviations: lnSL = ln(sound length); logWF = $\log_{10}$(word frequency); PhonN = number of phonological neighbors; OrthoN = number of orthographic neighbors; TTS = text-to-speech; t_{MF} = multifractal nonlinearity; word# = word number in the presented sequence.

[c]boldfaced values indicate statistical significance at the alpha level of 0.05.

Table 2. Coefficients of the Poisson regression of cumulative correct answers in the post-narrative test[a] (section 3.2).

effect[b]	***mean***	***s.e.m.*** **(bootstrapped)**	***p***[c] **(bootstrapped)**
intercept	4.62×10^{0}	1.07×10^{0}	$\mathbf{8.34\times10^{-6}}$
poly(question,3)1	8.84×10^{0}	1.40×10^{0}	$\mathbf{1.41\times10^{-10}}$
poly(question,3)2	-1.25×10^{0}	1.16×10^{0}	1.40×10^{-1}
poly(question,3)3	1.39×10^{-1}	9.22×10^{-1}	4.40×10^{-1}
LD-status	3.75×10^{-1}	8.24×10^{-2}	$\mathbf{2.62\times10^{-6}}$
typeWRM	2.88×10^{-1}	1.29×10^{-1}	$\mathbf{1.27\times10^{-2}}$
conditionTTS	2.84×10^{0}	1.49×10^{0}	$\mathbf{2.80\times10^{-2}}$
log(RT-mean)	3.35×10^{0}	5.08×10^{-1}	$\mathbf{1.99\times10^{-11}}$
log(RT-median)	-3.89×10^{0}	6.08×10^{-1}	$\mathbf{7.70\times10^{-11}}$
RT-$\Delta\alpha$	-2.31×10^{0}	8.85×10^{-1}	$\mathbf{4.49\times10^{-3}}$
RT-t_{MF}	6.49×10^{-3}	4.24×10^{-3}	6.31×10^{-2}
poly(question,3)1×typeWRM	9.45×10^{-1}	9.10×10^{-1}	1.50×10^{-1}
poly(question,3)2×typeWRM	-3.17×10^{0}	8.10×10^{-1}	$\mathbf{4.49\times10^{-5}}$
poly(question,3)3×typeWRM	1.16×10^{0}	6.15×10^{-1}	$\mathbf{2.99\times10^{-2}}$
conditionTTS×log(RT-mean)	-5.75×10^{0}	6.67×10^{-1}	$\mathbf{3.03\times10^{-18}}$
conditionTTS×log(RT-median)	5.40×10^{0}	7.30×10^{-1}	$\mathbf{7.23\times10^{-14}}$
poly(question,3)1×RT-$\Delta\alpha$	16.47×10^{0}	5.95×10^{0}	$\mathbf{2.80\times10^{-3}}$
poly(question,3)2×RT-$\Delta\alpha$	-5.67×10^{0}	4.81×10^{0}	1.19×10^{-1}
poly(question,3)3×RT-$\Delta\alpha$	3.47×10^{0}	3.61×10^{0}	1.68×10^{-1}
typeWRM×conditionTTS	3.73×10^{-2}	2.14×10^{-1}	4.31×10^{-1}
typeWRM×RT-$\Delta\alpha$	-4.57×10^{0}	6.84×10^{-1}	$\mathbf{1.13\times10^{-11}}$
typeWRM×RT-t_{MF}	2.77×10^{-2}	3.05×10^{-3}	$\mathbf{4.05\times10^{-20}}$
conditionTTS×RT-$\Delta\alpha$	1.13×10^{0}	1.17×10^{0}	1.67×10^{-1}
conditionTTS×RT-t_{MF}	-1.37×10^{-2}	5.94×10^{-3}	$\mathbf{1.06\times10^{-2}}$
typeWRM×conditionTTS×RT-$\Delta\alpha$	2.40×10^{0}	1.13×10^{0}	$\mathbf{1.71\times10^{-2}}$
typeWRM×conditionTTS×RT-t_{MF}	-3.06×10^{-2}	4.53×10^{-3}	$\mathbf{7.07\times10^{-12}}$

[a]model in *R* syntax: glmer(cum-corr ~ + poly(question,3) + LD_status + poly(question,3)*type + condition*[log(RT_mean) + log(RT_median)] + poly(question,3)*RT_$\Delta\alpha$ + type*condition*RT_$\Delta\alpha$ + type*condition*RT_t_{MF} + (1|participant), data = DATA, family = poisson).

[b]abbreviations: poly(question,3)i = i[th]-order orthogonal polynomial ($1\leq i\leq 3$); LD-status = 1 or 0 depending on whether participant did or did not self-report learning disability; typeWRM = word-recognition memory subset of memory test; conditionTTS = condition involving self-paced listening

with text-to-speech; RT = response time; RT-$\Delta\alpha$ = multifractal-spectrum width of RT sequence; RT-t_{MF} = multifractal nonlinearity in RT sequence.

[c]boldfaced values indicate statistical significance at the alpha level of 0.05.

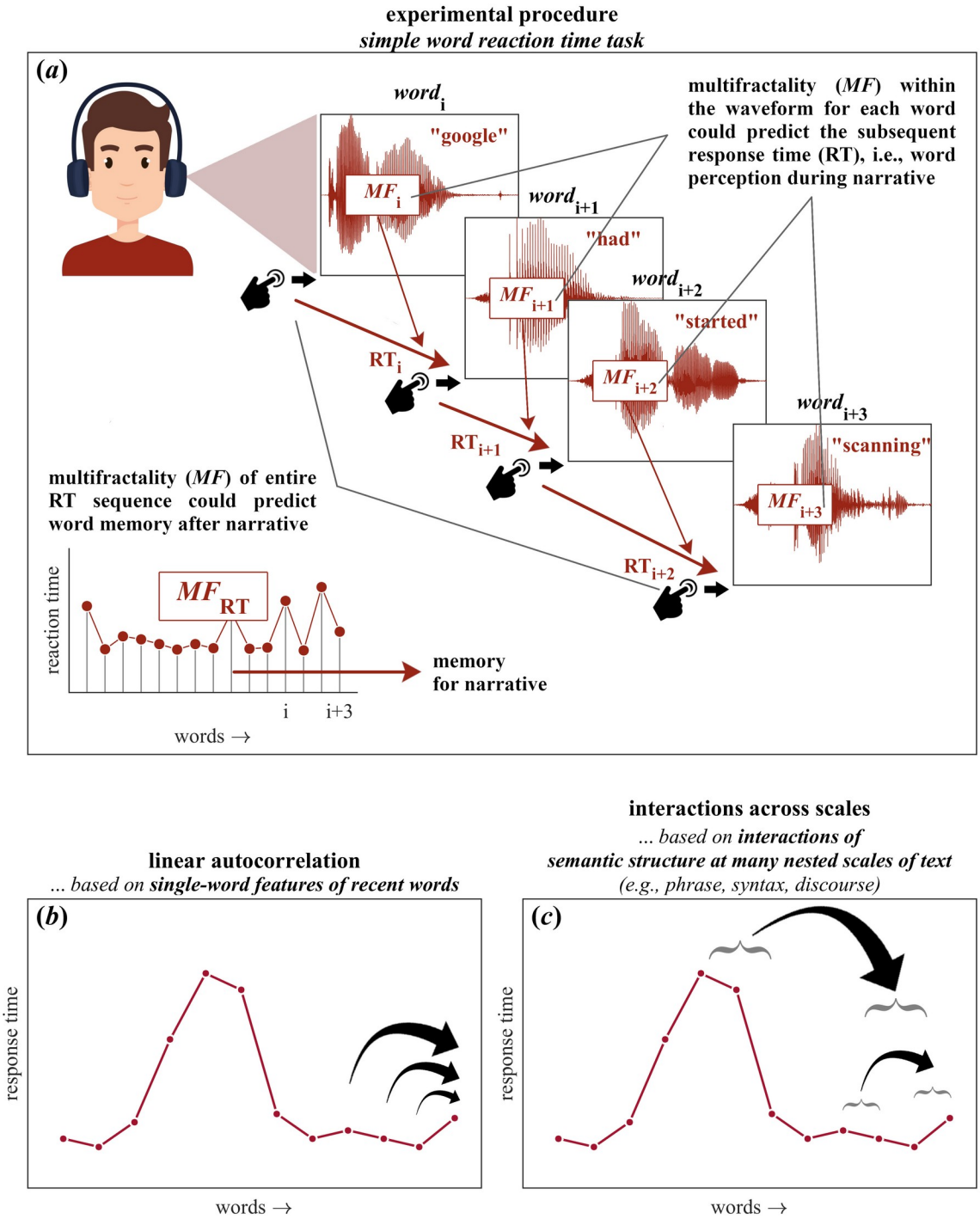

Figure 1. Experimental procedure and two perspectives on individual word's response times (RT). (*a*) Word perception (see Section 2.2 for details). (*b*) The autoregressive perspective that each word's RT entails the cumulative summing of subsequent and independent processes regarding lexical aspects of preceding words. (*c*) The multifractal perspective that each word's RT entails the interactions unfolding among processes across scales.

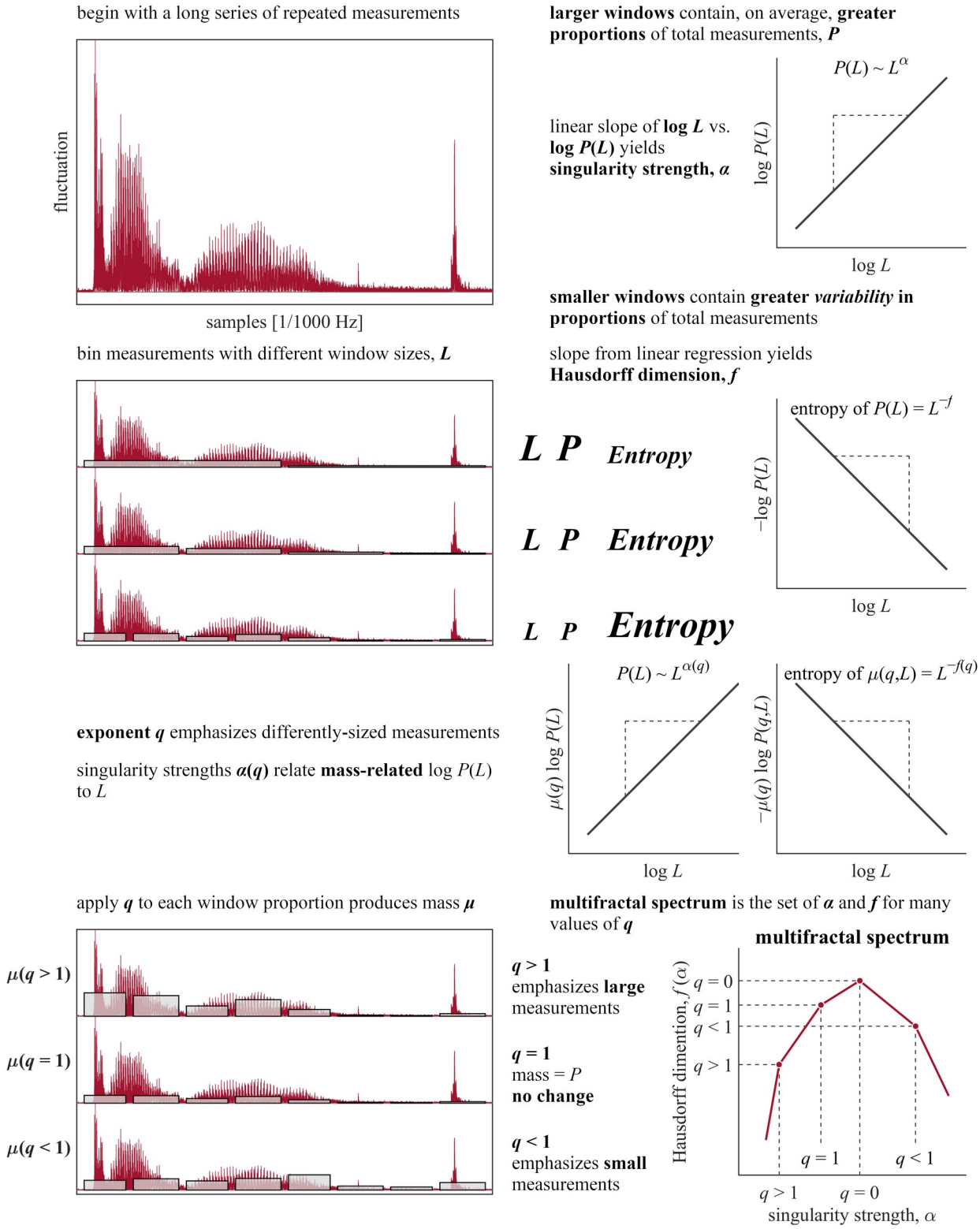

Figure 2. Schematic of Chhabra and Jensen's [65] method used to estimate multifractal spectrum $f(\alpha)$. See Section 2.4.1 for further details.

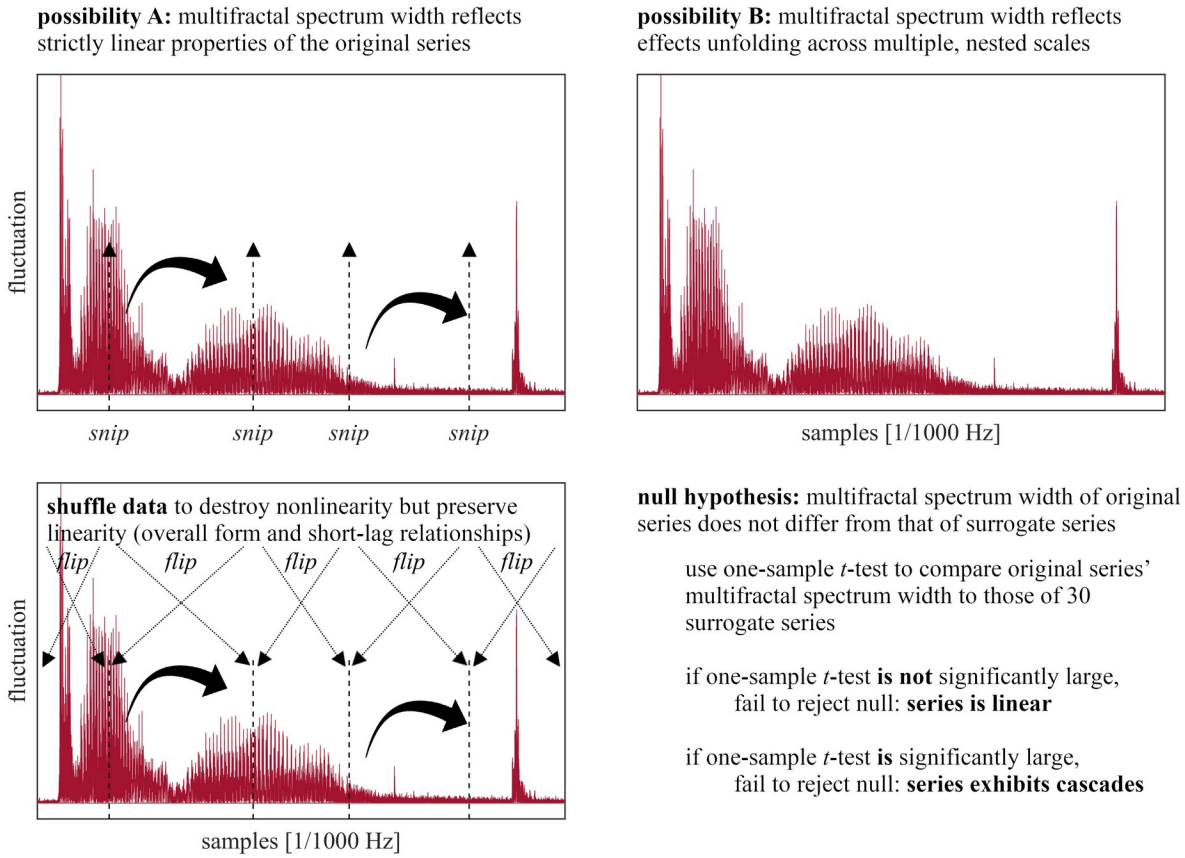

Figure 3. Surrogate analysis used to identify whether a series exhibits multifractality due to nonlinearity (cascades). See Section 2.4.2 for further details.

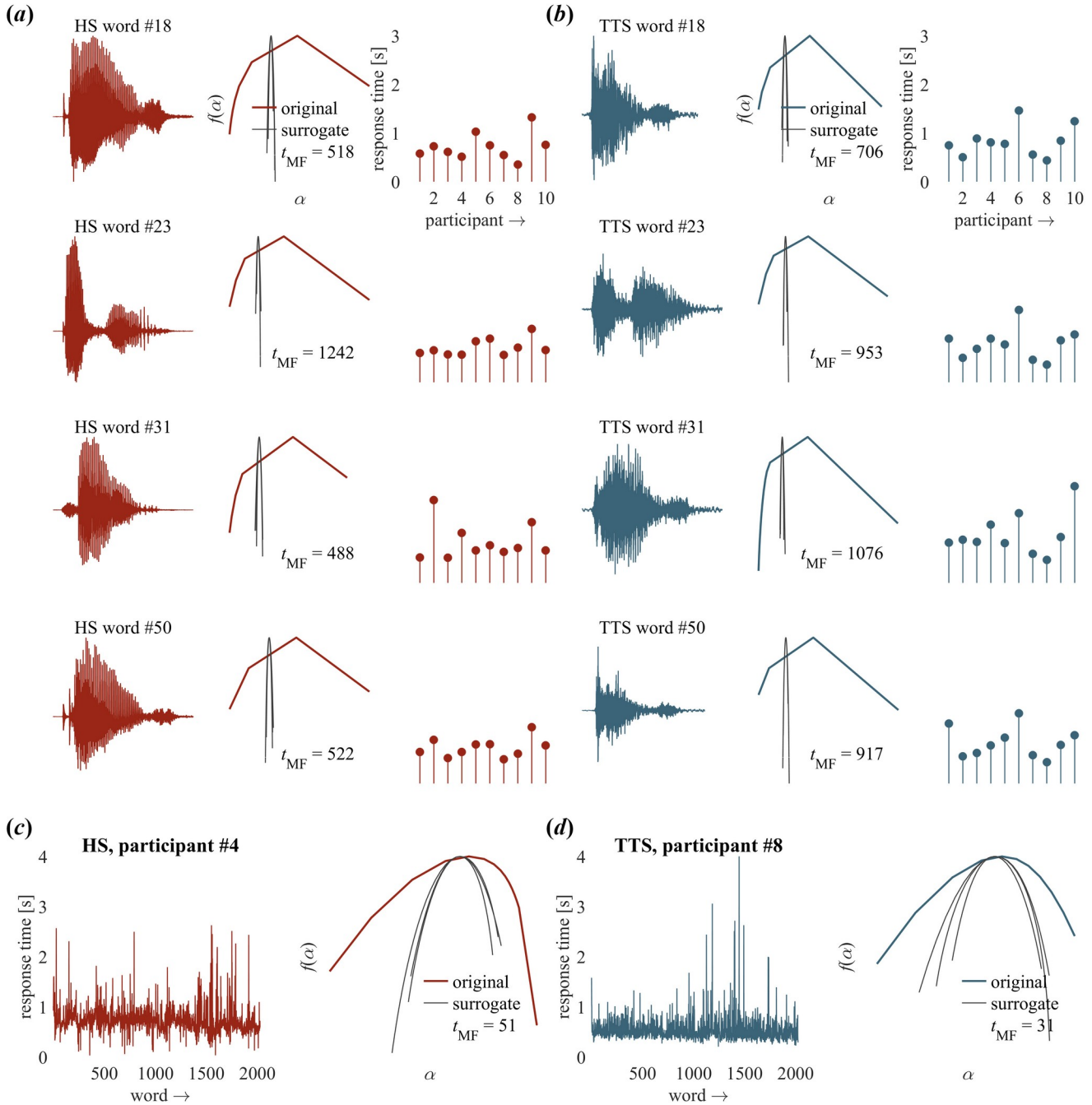

Figure 4. Multifractal spectra of human-speech (HS) and text-to-speech (TTS) words, as well as response time for an individual participant in both conditions. (*a*) Multifractal spectra of representative human-speech words and the respective response times for each participant (n = 10). (*b*) Multifractal spectra of representative text-to-speech words and the respective response times for each participant (n = 10). (*c*) Multifractal spectrum of response time series for a representative participant in human-speech condition. (*d*) Multifractal spectrum of response time series for a representative participant in text-to-speech condition.

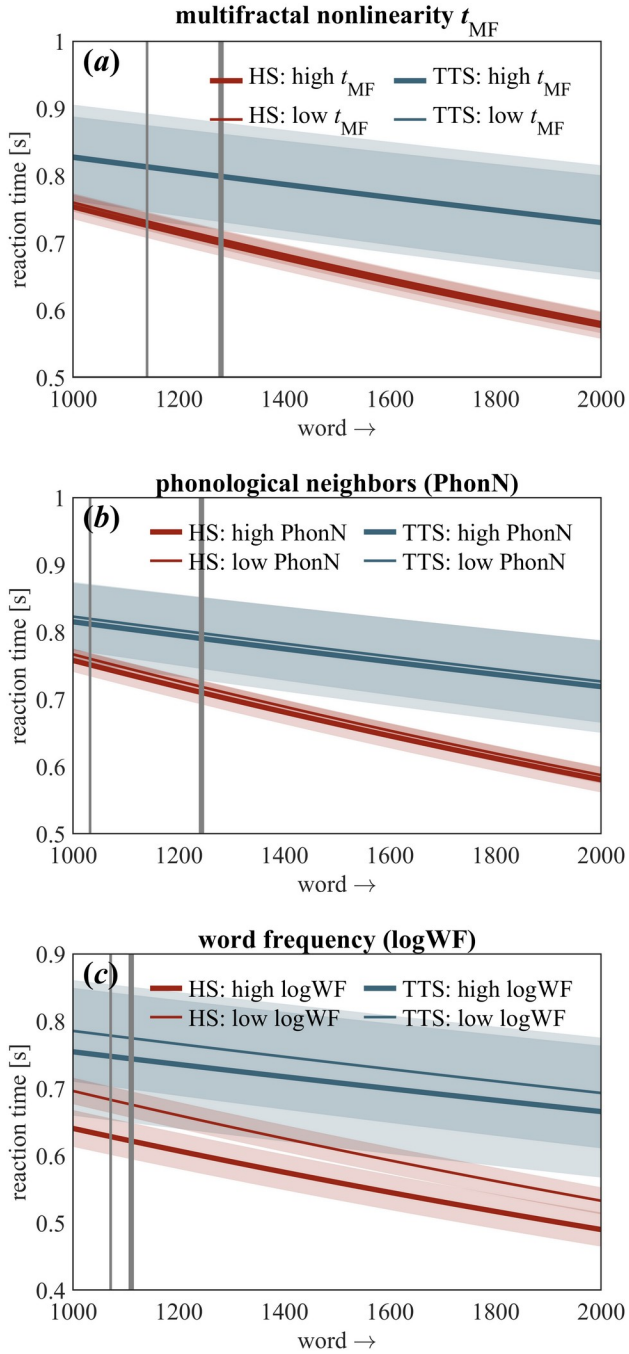

Figure 5. Model predictions for response time (section 3.1.2). (*a*) Effect of multifractal nonlinearity t_{MF}. Thin and thick gray vertical lines indicate differences between human speech (HS) and text-to-speech (TTS) in the low t_{MF} (word #1140) and high t_{MF} (word #1280), respectively. (*b*) Effect of phonological neighbors (PhonN). Thin and thick gray vertical lines indicate differences between human speech and text-to-speech in the low PhonN (word #1032) and high PhonN (word #1243), respectively. (*c*) Effect of word frequency (logWF). Thin and thick gray vertical lines indicate differences between human speech and text-to-speech in the low logWF (word #1071) and high logWF (word #1110), respectively. Shaded bars indicate ±2*s.e.m.*

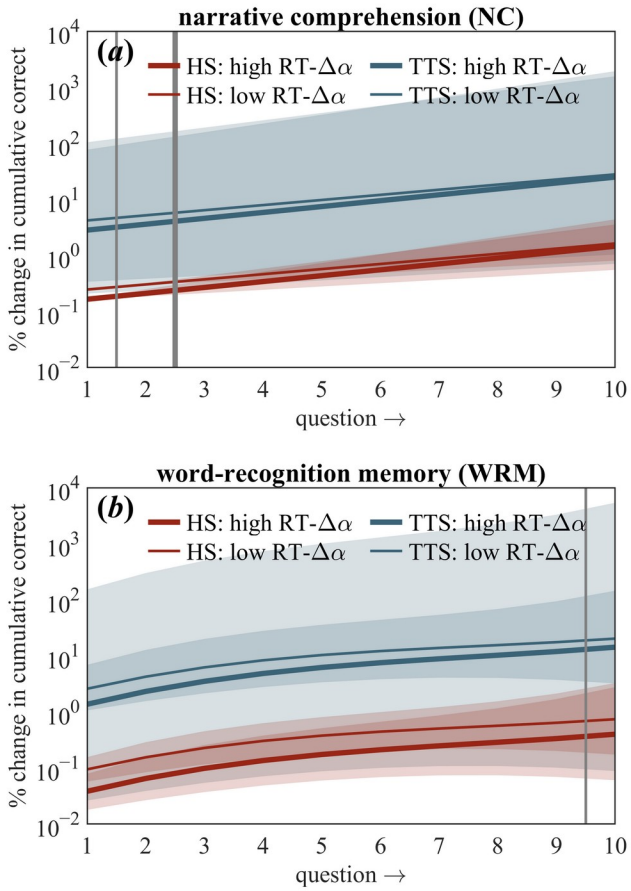

Figure 6. Model predictions for performance in narrative comprehension and word-recognition memory tests (section 3.2). (*a*) Effect of RT-$\Delta\alpha$ on % change in cumulative correct answers in narrative comprehension test. Thin and thick gray vertical lines indicate when low RT-$\Delta\alpha$ and high RT-$\Delta\alpha$, respectively, stops predicting better test performance compared to low RT-$\Delta\alpha$ and high RT-$\Delta\alpha$, respectively, in speech. (*b*) Effect of RT-t_{MF} on % change in cumulative correct answers in word-recognition memory test. Gray vertical line indicates when low RT-$\Delta\alpha$ stops predicting better test performance compared to low RT-$\Delta\alpha$ in speech. Shaded bars indicate ±2*s.e.m.*

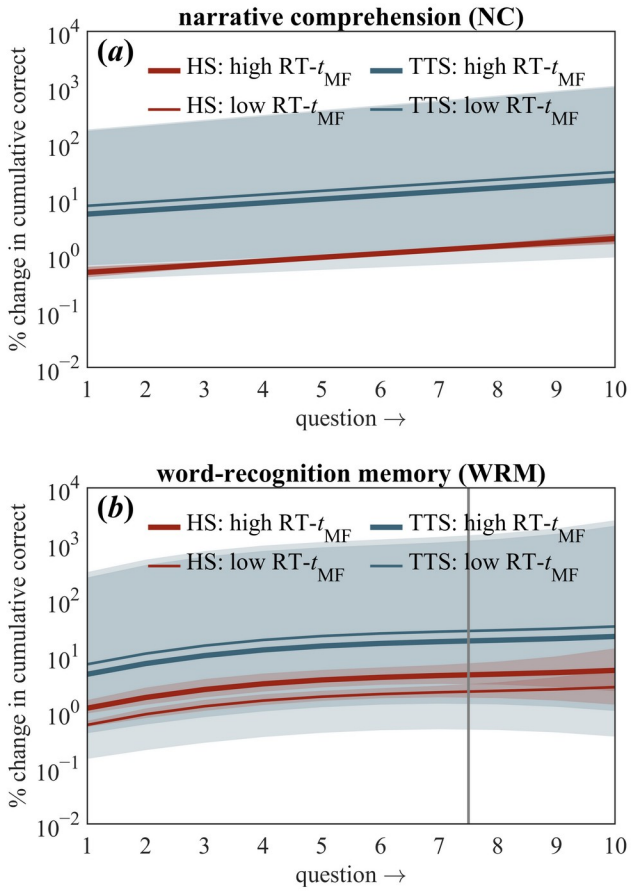

Figure 7. Model predictions for performance in narrative comprehension and word-recognition memory tests (section 3.2). (*a*) Effect of RT-t_{MF} on % change in cumulative correct answers in narrative comprehension test. (*b*) Effect of RT-t_{MF} on % change in cumulative correct answers in word-recognition memory test. Gray vertical line indicates when low RT-t_{MF} stops predicting better test performance compared to low RT-t_{MF} in speech. Shaded bars indicate ±2*s.e.m.*