Cochlea-scaled spectral entropy predicts rate-invariant intelligibility of temporally distorted sentences[a]

Christian E. Stilp[b]
University of Wisconsin, 1202 West Johnson Street, Madison, Wisconsin 53706

Michael Kiefte
Dalhousie University, 5599 Fenwick Street, Halifax, Nova Scotia B3H 1R2, Canada

Joshua M. Alexander
Purdue University, West Lafayette, Indiana 47907

Keith R. Kluender
University of Wisconsin, 1202 West Johnson Street, Madison, Wisconsin 53706

(Received 3 December 2009; revised 28 July 2010; accepted 29 July 2010)

Some evidence, mostly drawn from experiments using only a single moderate rate of speech, suggests that low-frequency amplitude modulations may be particularly important for intelligibility. Here, two experiments investigated intelligibility of temporally distorted sentences across a wide range of simulated speaking rates, and two metrics were used to predict results. Sentence intelligibility was assessed when successive segments of fixed duration were temporally reversed (exp. 1), and when sentences were processed through four third-octave-band filters, the outputs of which were desynchronized (exp. 2). For both experiments, intelligibility decreased with increasing distortion. However, in exp. 2, intelligibility recovered modestly with longer desynchronization. Across conditions, performances measured as a function of proportion of utterance distorted converged to a common function. Estimates of intelligibility derived from modulation transfer functions predict a substantial proportion of the variance in listeners' responses in exp. 1, but fail to predict performance in exp. 2. By contrast, a metric of potential information, quantified as relative dissimilarity (change) between successive cochlear-scaled spectra, is introduced. This metric reliably predicts listeners' intelligibility across the full range of speaking rates in both experiments. Results support an information-theoretic approach to speech perception and the significance of spectral change rather than physical units of time.

© 2010 Acoustical Society of America. [DOI: 10.1121/1.3483719]

PACS number(s): 43.71.Es, 43.71.Rt [RSN] Pages: 2112–2126

I. INTRODUCTION

Speech signals are extremely complex with many acoustic attributes arrayed in frequency and time. However, this complexity is not without structure. Many acoustic attributes are redundant, and there are systematic relationships among different attributes. Owing to these regularities, perception of speech is remarkably resilient even in the face of extreme signal degradation (e.g., Assmann and Summerfield, 2004; Kluender and Alexander, 2008; Kluender and Kiefte, 2006). For example, real listening environments include noise from competing sources, and surfaces in the environment often introduce reverberation that corrupts temporal aspects of speech. Nevertheless, perception often overcomes such challenges, as listeners are able to rely on attributes that are more accessible in particular listening environments. The present effort is directed to better understand the information used by listeners in conditions that are not optimal due to different types of signal degradation.

Houtgast and Steeneken (1973, 1985) introduced the Modulation Transfer Function (MTF) as one way to assess effects of room acoustics on speech intelligibility. They filtered sentences into octave bands, then determined the amplitudes of temporal modulations within bands as a function of modulation frequency. Across variations in talkers, texts, and modes of speaking, amplitude envelope modulations from octave-filtered medium-rate speech consistently peaked around 3 Hz, with significant modulation present between 1 and 8 Hz. These low-frequency temporal modulations appear to contribute significantly to perception of speech, as listeners are able to understand speech with relatively little temporal fine structure when only low-frequency modulations are available (Dudley, 1939; Shannon *et al.*, 1995). Furthermore, distortion of these modulations results in a decrease in intelligibility (e.g., Drullman *et al.*, 1994a).

Intelligibility data from diverse experiments on perception of temporally distorted speech, as well as analyses of undistorted speech (Houtgast and Steeneken, 1985), have led

[a] Portions of this work were presented at the 153rd [J. Acoust. Soc. Am. **121**, 3185 (2007)] and 154th Meetings of the Acoustical Society of America [J. Acoust. Soc. Am. **122**, 2970 (2007)].

[b] Author to whom correspondence should be addressed. Electronic mail: cestilp@wisc.edu

to suggestions that low-frequency modulation envelopes (specifically, 3–8 Hz) may be critical to speech intelligibility (e.g., Arai and Greenberg, 1998; Greenberg, 1999; Greenberg and Arai, 2001, 2004; Greenberg *et al.*, 1998; Saberi and Perrott, 1999). Across multiple forms of temporal distortion (Arai and Greenberg, 1998; Flanagan, 1951; Greenberg and Arai, 2001; Greenberg *et al.*, 1998; Saberi and Perrott, 1999; Silipo *et al.*, 1999), duration or periodicity of temporal distortion at which intelligibility is minimized corresponds well to peak modulation frequencies in the MTF.

Saberi and Perrott (1999) examined intelligibility of sentences in which successive time segments were locally time-reversed. Performance was near-perfect up to 50-ms segment reversals. Intelligibility was still 50% when 130-ms segments were reversed and did not reach floor performance until segment durations were 200 ms. In similar experiments, Greenberg and Arai (2001) reported that intelligibility decreased much more quickly with increasing distortion (i.e., increased reversed-segment duration). Performance approached 50% with only 60-ms distortion, and approached floor performance with local reversals of 100 ms. At least some of the differences in performance between the two studies may be attributed to Greenberg and Arai's (2001) use of the relatively difficult, multi-talker TIMIT sentence corpus, and their use of explicit measures of intelligibility (Greenberg and Arai, 2001) versus subjective measures (Saberi and Perrott, 1999).

Much earlier, Flanagan (1951) divided speech into twenty frequency bands (spanning 200–6100 Hz, as defined by equal contributions to the Articulation Index—AI; French and Steinberg, 1947). He presented sentences in which onsets of half of these bands were delayed by a fixed duration, and listeners were instructed to report a target word embedded in the carrier sentence, "You will write (target) now." Flanagan reported maximal decrement to intelligibility when the middle half of frequency bands (center frequencies ranging from 920 to 2660 Hz) was delayed 240 ms relative to the remaining lower- and higher-frequency bands, and he concluded that desynchronization delays of approximately this duration were most likely to lead to interference and impairment of intelligibility.

Greenberg and colleagues (Greenberg *et al.*, 1998; Silipo *et al.*, 1999) investigated perception of temporally distorted speech after processing sentences with four one-third-octave filters separated by one-octave stop bands. Despite removing most of the energy in the spectrum, intelligibility was remarkably good (almost 90%). Sentences were then desynchronized such that one pair of bands—either the outermost (335- and 5400-Hz center frequency; CF) or the two center bands (850- and 2135-Hz CFs)—started at a fixed delay relative to the other two. Intelligibility decreased monotonically with increasing delays up to approximately 250 ms, beyond which performance improved somewhat, replicating results reported by Flanagan (1951).

Arai and Greenberg (1998) measured intelligibility of sentences for which spectra were filtered into 19 adjacent, non-overlapping, quarter-octave channels. Channels were randomly shifted in time according to a uniform distribution ranging from zero to a maximum delay, which varied between 60 and 240 ms. The result of this across-channel asynchrony was to "jitter" spectral information relative to the original sentence, similar to temporospectral effects of reverberation that alter relative phase of spectral components, thereby flattening the overall amplitude envelope. Speech intelligibility gradually decreased monotonically with increasing temporal asynchrony but was otherwise remarkably resistant to this distortion (but see Fu and Galvin, 2001; Healy and Bacon, 2007; Remez *et al.*, 2008).[1]

There is a notable congruence between these patterns of performance and the MTF, as the temporal distortion at performance minima (e.g., 100, 200, 240, and 250 ms) correspond reasonably well to peak modulations in the MTF for medium-rate speech (10, 5, 4.2, and 4 Hz, respectively). Multiple researchers (Arai and Greenberg, 1998; Greenberg, 1999; Greenberg and Arai, 2001, 2004; Greenberg *et al.*, 1998; Saberi and Perrott, 1999) have suggested that preservation of peak modulations may be critical to perception of connected speech. This conclusion may be qualified, however, by reports that sentences with compromised MTFs can remain highly intelligible (Drullman *et al.*, 1994a, 1994b), and sentences that have been distorted in ways that maintain a normal MTF can have quite poor intelligibility (Longworth-Reed *et al.*, 2009).

Drullman *et al.* (1994a, 1994b) investigated speech reception following smearing of temporal envelopes. The speech signal was divided into a series of frequency bands (widths of $\frac{1}{4}$, $\frac{1}{2}$, or 1 octave), and amplitude envelopes for each band were low- or high-pass filtered by a wide range of cutoff frequencies, effectively smoothing the MTF to varying degrees (see Drullman *et al.*, 1994b). Substantial improvement in speech reception was observed when low-pass cutoff frequency was increased from 4 to 8 Hz. In addition, envelope information below 4 Hz could be eliminated without significantly affecting performance. Across these experiments, the crossover modulation frequency for low- and high-pass filtered waveform envelopes was around 8 to 10 Hz, nearly independent of frequency channel bandwidths. Perception of distorted sentences was remarkably good despite degraded MTFs.

Conversely, an intact MTF does not always assure accurate speech perception. Longworth-Reed *et al.* (2009) investigated perception of undistorted sentences that were presented in a virtual room whose acoustics changed. In normal listening circumstances, direct-path energy precedes reverberant energy at the listener's ear. Sentence intelligibility was measured when the impulse response of the virtual room was temporally reversed, thus maintaining the MTF but altering room acoustics such that reverberations and acoustic reflections preceded direct-path energy. The simulated spatial location of the speech source was in the center of the virtual room, directly in front of the listener. Despite the MTF being identical in these two conditions,[2] intelligibility was robust in the control condition ("time-forwards" room) but severely compromised in the latter condition ("time-reversed" room). Together, the results of Longworth-Reed *et al.* (2009) and Drullman and colleagues (Drullman *et al.*, 1994a, 1994b) serve to qualify strong conclusions concerning the relationship between reliable speech perception and the MTF.

Experiments reported in the present manuscript serve two goals. The first is to extend previous research by investigating perception of temporally distorted sentences across a wide range of simulated speaking rates. Rate manipulation serves three purposes. First, varying rate generates sentences for which peak modulation frequencies lie at the fringes or altogether outside of the 3–8 Hz range that has been characterized as being critical for speech intelligibility (e.g., Arai and Greenberg, 1998; Greenberg, 1999; Greenberg and Arai, 2001, 2004; Greenberg *et al.*, 1998; Saberi and Perrott, 1999). As a result, presenting sentences synthesized at slow (2.5 syllables per second), medium (5.0), and fast (10.0) rates permit examination of perception across a wider range of modulation frequencies than has been previously investigated. Second, manipulation of rate affords exploration of sentence intelligibility under more realistic circumstances where speaking rate varies. Almost all studies that investigated intelligibility of distorted speech, including those noted above, used sentences that were only modestly representative of conversational speech—relatively clearly spoken at a single measured tempo. Third, measuring intelligibility across different rates allows for evaluation of whether perception is best measured in absolute units, such as milliseconds (or hertz in the case of modulation rates), or in relative units such as duration of distortion relative to the duration of the entire sentence.

Here, we report perceptual effects of two types of acoustic distortion across wide variation in simulated speaking rate with accompanying changes in modulation rates. Investigations of sentence intelligibility will inform about the perceptual significance of the MTF when successive segments of fixed duration are temporally reversed, thus modifying modulation characteristics (exp. 1), and when sentences are processed through four third-octave-band filters, the outputs of which were desynchronized by fixed delays, which has no effect on modulation characteristics (exp. 2). Relationships between modulations, intelligibility, speaking rate, and distortion in both absolute (milliseconds) and relative (proportional) measures are explored. For example, a relative measure of time is proportion of total utterance duration.

The second and primary goal of these experiments is to compare the efficacy with which two metrics predict intelligibility of temporally distorted sentences across a wide range of speaking rates. The first metric considered is the Speech Transmission Index (STI), a predictor of speech intelligibility derived directly from the MTF (Steeneken and Houtgast, 1980; Houtgast and Steeneken, 1985). In contrast to spectrally-based models such as the AI (French and Steinberg, 1947), STI predicts intelligibility as a function of changes in amplitude modulations of speech across frequencies. STI expands upon the AI by predicting speech intelligibility amidst nonlinear (e.g., peak clipping) and temporal (e.g., reverberation, echoes) distortions. It bears note that models like STI and AI are designed to predict speech intelligibility. They are not models of perception that can provide pattern recognition or error prediction. The efficacy of STI to predict intelligibility of temporally distorted sentences is inconsistent, as results of Drullman *et al.* (1994a, 1994b) and Longworth-Reed *et al.* (2009) suggest that listener performance relies upon at least some factors that are not captured by the MTF (and consequently, the STI).

The second metric tested is an alternative index of speech intelligibility, cochlea-scaled entropy, that is grounded in psychoacoustic (Kluender *et al.*, 2003) and information-theoretic considerations (Stilp and Kluender, 2010). In accordance with Shannon information theory, there is no new information when events either do not change or are predictable (Shannon, 1948). When there is more entropy (change, or unpredictability), there is more potential information. We introduce an estimate of potential information available to listeners across time.

Stilp and Kluender (2010) measured cochlea-scaled spectral change (CSE) in sentences, then replaced portions rated as having high, medium, or low CSE with noise. As sensorineural systems are tuned nearly exclusively to change, it was predicted that intelligibility would maintain if regions of relatively little spectral change were replaced by noise, but would decrease as more of the sentence's kinematic spectral structure was replaced with static noise. Stilp and Kluender reported a remarkably robust relationship between cochlea-scaled spectral change and intelligibility. This metric is described and employed in the present studies to predict intelligibility of variable-rate, temporally distorted sentences.

CSE and the MTF are correlated in natural connected speech, largely due to their shared relationship to syllable structure, but each measure can be distinguished by varying conditions of speaking rate and distortion. Both metrics are expected to reliably predict listener performance in Experiment 1, where local segment reversals affect both modulation characteristics and spectral change over time. However, the metrics are dissociable in Experiment 2, where desynchronization of bands of speech affects spectral differences over time but does not affect within-band modulation characteristics. For each experiment, comparisons will permit evaluation of the relative efficacy of each measure as a predictor of sentence intelligibility.

II. EXPERIMENT 1

A. Method

1. Listeners

Forty-five native English speakers were recruited from the Department of Psychology at the University of Wisconsin—Madison. None reported hearing impairment and all received course credit as compensation for participation.

2. Stimuli

Synthetic speech was used in the present experiments to systematically vary speaking rate. All stimuli were generated using the AT&T Natural Voices™ Text-To-Speech Synthesizer[3] (Beutnagel *et al.*, 1997) using a male talker with an American English accent ("Mike"). This synthesizer was selected after first attempting to create naturally produced sentences at well-controlled rates. For the range of rates desired, it proved infeasible to create sentences with widely varying rates absent imprecise articulation (fastest rate) or brief pauses (slowest rate). This synthesizer provided good naturalness and direct control over speaking rate independent of vocal pitch.

Experimental materials were sentences selected from the Hearing In Noise Test (HINT) corpus (Nilsson *et al.*, 1994). HINT sentences were chosen because they are semantically more predictable (easier) and have been normed for roughly equal intelligibility when presented in noise. All sentences selected for the present experiments contained seven syllables and ranged from 4 to 7 words in length. All seven-syllable sentences from the HINT lists were synthesized, and sentences containing distorted or ambiguous pronunciation were discarded. In all, 115 sentences were used at each of the three different simulated speaking rates—slow, medium, and fast (2.5, 5.0, and 10.0 syllables/s, respectively)—at a sampling rate of 16 kHz. Rate was adjusted for each sentence using XML-style tags for the "prosody rate" command. Slow sentences had mean duration of 2.6 s (SE=0.05; range 1.5–3.7); medium sentences (i.e., no rate manipulation) had mean duration of 1.4 s (SE=0.02; range 1.0–1.8); and, fast sentences had mean duration of 0.8 s (SE=0.01; range 0.6–1.1). This medium-rate seven-syllable subset of HINT sentences was modestly briefer than average sentence duration in the HINT inventory (mean 1.7 s; SE=0.01; range 1.2–2.5). Visual inspection of sentence waveforms and spectrograms confirmed near uniform linear scaling of time across different rates. From these 115 sentences, 87 sentences were randomly selected for use in Experiment 1.

Sentences were subdivided into equal-duration segments (20, 40, 80, and 160 ms) at the nearest zero crossings in the time waveform. Each segment was time-reversed while maintaining its relative position within the sentence using PRAAT (Boersma and Weenink, 2007). Undistorted sentences were also included as a control condition. From the 87 selected sentences, 12 were used in practice trials, and were presented at only one rate and one segment duration. The remaining 75 sentences, synthesized at each of the three rates, were processed at each of the five frame-reversal durations, generating 1,125 stimuli. Following synthesis, sentences were set to constant overall RMS amplitude and upsampled to 44.1 kHz for presentation.

3. Procedure

All sentences were presented monaurally at 72 dBA via circumaural headphones (Beyer-Dynamic DT-150). Listeners participated individually in a double-wall soundproof booth. Following acquisition of informed consent, listeners read a set of instructions on a computer screen explaining the nature of the experiment. In addition, listeners were told to expect that some sentences would be difficult to understand and that guessing was encouraged because every word of their responses would be scored.

The experiment was conducted in two parts over the course of approximately 30 min. Each sentence was played once, after which the computer prompted listeners to type in any words they understood. Listeners first completed the 12 practice sentences, each hearing the same practice stimuli in the same order. Practice trials were ordered beginning with slow sentences and no distortion, followed by progressively increasing rate and segment duration (i.e., progressively increasing predicted difficulty).

Following practice, each listener heard 75 experimental sentences, one per trial, without hearing any sentence more than once. Although listeners heard each experimental sentence in the same order, the order of speech rates and segment durations was pseudo-randomized. Within each of five 15-trial blocks, listeners heard one sentence in each of the 15 conditions (3 rates by 5 segment reversal durations) in random order. Each unique experimental stimulus was presented exactly once after every 15 participants, and each sentence for a given rate and segment duration was heard three times across all 45 listeners.

Intelligibility for each condition was scored as the proportion of words in each sentence correctly identified. Three raters, blind to experimental conditions, independently scored the typed responses after all data were collected. Responses were scored according to guidelines listed in the Appendix. The same three raters scored responses to all experiments, and inter-rater reliability (measured by intraclass correlation) is provided for each data set.

B. Results

1. Listener performance

Results for Experiment 1 are presented in Fig. 1 (inter-rater reliability: $r > 0.99$). Intelligibility is plotted as a function of reversed-segment duration in Fig. 1(a). Data were analyzed using a repeated-measures, 3 (rate) by 5 (segment reversal duration) analysis of variance (ANOVA). Listener performance changed considerably as a function of speaking rate ($F_{2,88}=1071.85$, $p<0.001$, $\eta_p^2=0.96$). Tukey's Honestly Significant Difference (HSD) *post-hoc* tests indicate intelligibility of each speaking rate was different from the others, with intelligibility related inversely to rate ($\alpha=0.01$). Intelligibility also varied as a function of segment-reversal duration ($F_{4,176}=2064.15$, $p<0.001$, $\eta_p^2=0.98$). Tukey HSD tests indicate that intelligibility decreased with each increase in reversal duration ($\alpha=0.01$).

The interaction between rate and segment duration was also significant ($F_{8,352}=185.12$, $p<0.001$). It appeared that the interaction and main effect for speaking rate were related to the proportion of sentence content distorted. Further, the interaction arises because there is no difference in intelligibility across speaking rates at the shortest and longest reversal durations due to ceiling and floor effects, respectively. To normalize the range of speaking rates, segment durations were transformed to proportion of total utterance (i.e., segment duration divided by mean sentence duration). Analysis using transformed (relative) segment duration revealed that intelligibility for all three rates collapses onto a single curve [Fig. 1(b)]. A comparable ANOVA cannot be conducted for performance as a function of proportions of the sentence distorted because there are data for all three rates at only two of the six proportions of reversed intervals. Nevertheless, several qualitative observations can be made. It appears that reduction in intelligibility across dramatic changes in simu-

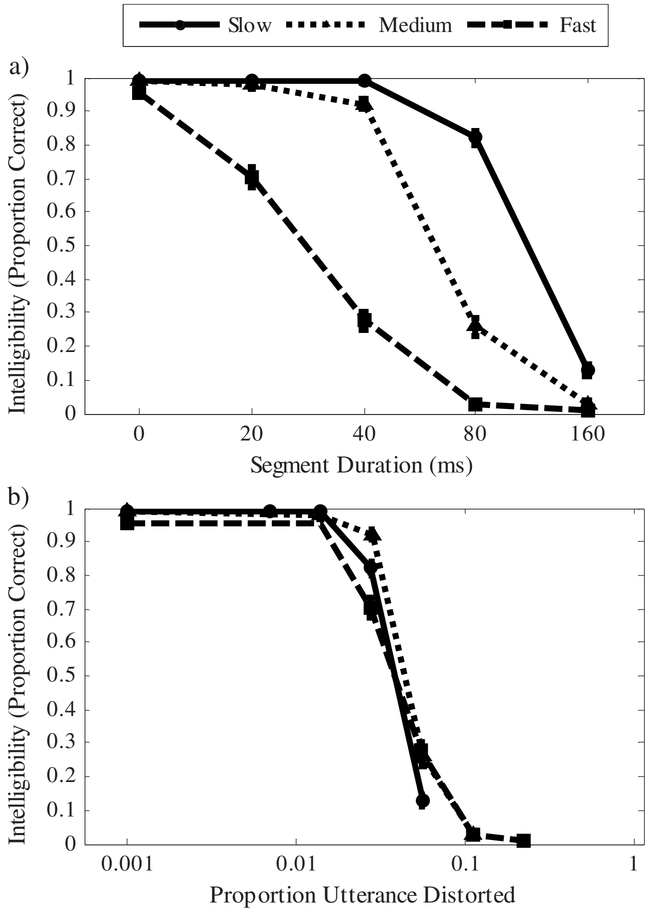

FIG. 1. Results from Experiment 1. (a) Intelligibility is plotted as a function of segment duration. Solid lines with circles denote results for slow-rate sentences, dotted lines with triangles for medium-rate sentences, and dashed lines with squares for fast-rate sentences. Intelligibility decreased with increasing reversal duration (distortion). Error bars depict standard error of the mean. (b) When the same intelligibility results from Experiment 1 are plotted as a function of proportion of utterance distorted, data collapse to a common function. Across speaking rates, the relative minimum in intelligibility occurs at roughly the duration of one syllable (0.14, or one-seventh of the seven-syllable sentence).

lated speaking rate is more directly related to proportion of the utterance distorted rather than absolute duration or modulation frequency.

2. MTF/STI analysis

Intensity envelope modulations of experimental sentences were calculated as a function of modulation frequency following the methods of Houtgast and Steeneken (1985). The full corpus of experimental sentences was concatenated for each rate (fast, medium, slow) and octave-band filtered with center frequencies between 0.125 and 4 kHz via IIR filters of order $20\times f_s/\mathrm{BW}$ where f_s is the native sampling rate (i.e., 16 kHz) and BW is the bandwidth of the octave filter in hertz. Filter outputs were squared, downsampled to 100 Hz, and processed by a bank of third-octave IIR filters from 0.25 to 25.4 Hz of order $20\times f_s/\mathrm{BW}$ where f_s is 100 Hz. Outputs of these filters were normalized by the mean of the resampled, squared waveform of the corresponding octave-filter outputs, the result being used to calculate modulation amplitudes which are plotted in Fig. 2. Notably, these measures are corrupted by the segment-reversal distortion in two ways. First, natural modulation characteristics of speech are perturbed by the distortion. Second, time-reversing successive portions of the waveform introduces modulations

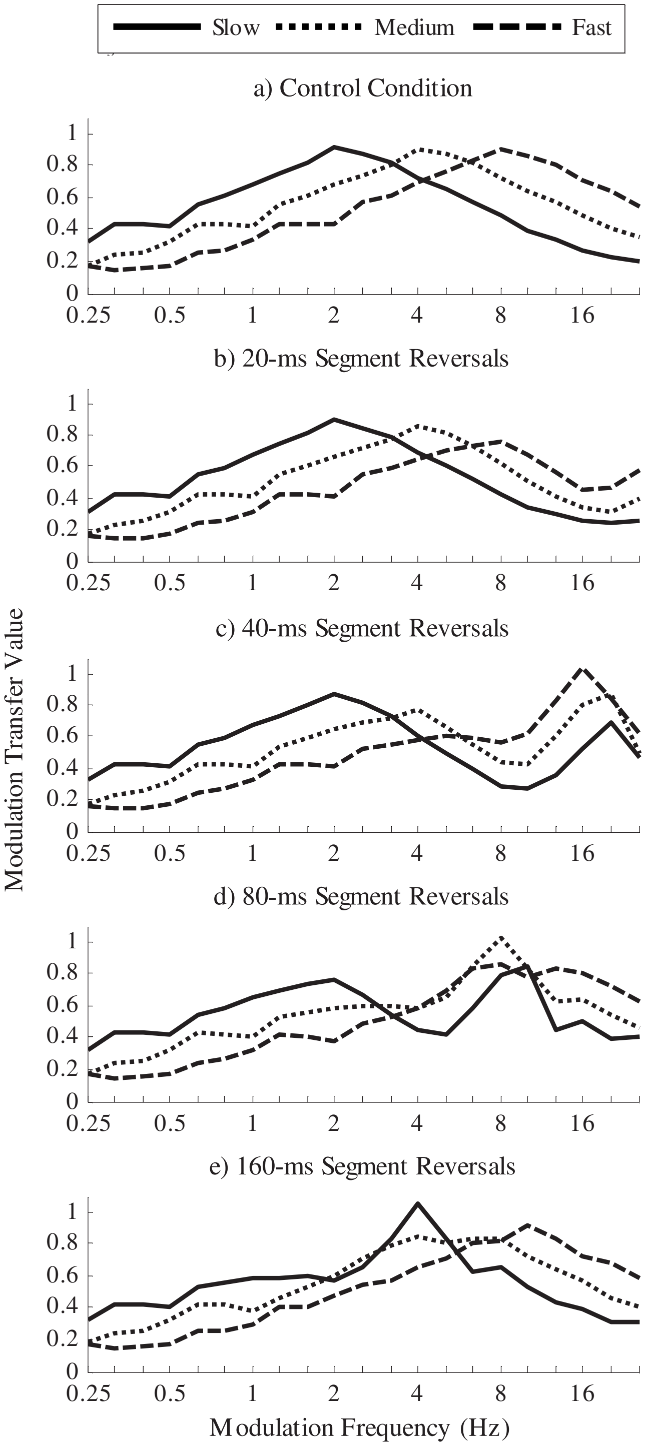

FIG. 2. Modulation transfer functions (MTFs) of experimental sentences in Experiment 1. Solid lines depict slow-rate sentences, dotted lines depict medium-rate sentences, and dashed lines depict fast-rate sentences. Modulation transfer value is plotted on the ordinate and modulation frequency on the abscissa of each graph. MTF for undistorted sentences (a) and following 20-ms (b), 40-ms (c), 80-ms (d), and 160-ms reversals (e) are shown. Additional peaks in MTFs [partially visible in uptick at high-modulation-frequencies in (b)] correspond to modulations introduced by time-reversing segments of the corresponding modulation rate, or the reciprocal of twice the segment duration. Additional peaks for 160-ms reversals (corresponding to 3.125 Hz) overlap with normal modulations at 2-, 4-, and 8-Hz as seen in (a).

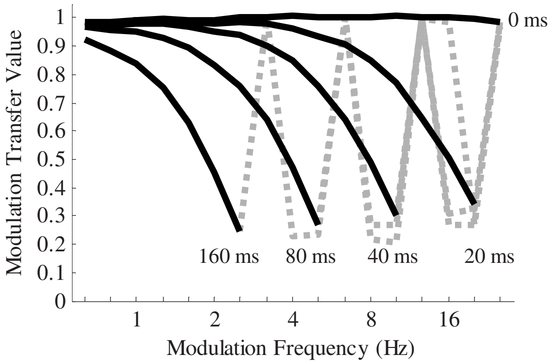

FIG. 3. Modulation transfer functions displaying systematic changes in modulation characteristics following time-reversal of successive fixed-duration segments. Modulation transfer values are plotted on the ordinate, and modulation frequencies from 0.63 to 25.4 Hz in equal-logarithmic steps are plotted on the abscissa. Modulation transfer values were derived from temporally distorted amplitude-modulated noise carriers to directly examine changes in the modulation response of the system (i.e., the input-output function of this means of distortion). Black lines depict modulation transfer values for noise carriers following time-reversal of segments at different durations (labeled at the end of each black line). Dashed gray lines depict aliasing at modulation transfer values at or exceeding the Nyquist frequency of segment-reversal durations. Consequently, derivation of the theoretical STI incorporates meaningful modulations (shown in black) and not those introduced by such distortion (shown in gray).

that are artifacts relative to modulations of unaltered speech (visible as secondary peaks in MTFs in Fig. 2). These peaks in modulation intensities correspond roughly to one-half the periodicity introduced by segment reversals (25, 12.5, 6.25, and 3.125 Hz, corresponding to 20-, 40-, 80-, and 160-ms segment reversals, respectively).

Steeneken and Houtgast (1980) and Houtgast and Steeneken (1985) introduced the speech transmission index (STI) as a predictor of listener performance based upon modulation transfer functions (MTF). The STI provides a scalar value that summarizes reductions of modulation intensities across frequency bands for a given system. One can compute the STI based on the distortion of specific modulation bands; however, calculation of the STI for speech materials used here is corrupted by artifact modulations due to segment reversals in the same way that MTFs are. However, STIs can be estimated from MTFs of amplitude-modulated noise carriers, with time-reversed segments of 20, 40, 80, or 160 ms, in order to assess how modulations within frequency bands are compromised due to additional modulations corresponding to frequency of segment reversals. This permits direct measurement of changes in the modulation response of the system (i.e., changes following this form of temporal distortion) as opposed to only inferring changes in modulation profiles of sentences which have their own modulation characteristics (Fig. 2). Modulation transfer values of noise carriers at each segment-reversal duration are plotted in Fig. 3. A systematic pattern is evident: modulation transfer values remain equal to 1 before declining sharply as modulation frequencies approach those corresponding to the Nyquist frequency of segment-reversal durations, the reciprocal of twice the segment duration (e.g., 25 Hz for a segment duration of 20 ms). Beyond this point, indices are aliased considerably.

Given this observation, theoretical STIs (i.e., based on analyses of system-level responses) can be derived from the methods of Houtgast and Steeneken (1985). The STI is calculated from 13 1/3-oct bands with modulation frequencies ranging from 0.63 to 12.5 Hz. These modulation frequencies can be converted to exponents of rational fractions (i.e., center frequencies of $2^{-2/3}$ to $2^{11/3}$ in exponential steps of 1/3). All are given equal weight according to standard STI calculation. As seen in Fig. 3, reversing segments of a given duration effectively distorts all information above a corresponding modulation rate. Thus, it is assumed that all modulations above this cutoff modulation frequency are maximally distorted resulting in modulation indices of 0 while all other indices are 1 (perfect modulation transfer.) Indices are then averaged across modulation frequencies with no weighting. If bands are treated as continuous, then indices are 1 from $\log_2(2^{-2/3})$ to $-\log_2(2\ell)$, and are 0 from $-\log_2(2\ell)$ to $\log_2(2^{11/3})$ where 2ℓ (measured in seconds) is the shortest period duration distorted by segment-reversal according to the Nyquist theorem, and $-\log_2(2\ell)=\log_2(f)$ where $f=1/2\ell$ is the modulation frequency corresponding to the shortest period affected by this distortion. The average of modulation frequencies with nonzero indices (i.e., equal to 1) is then calculated by Eq. (1):

$$\mathrm{STI} = 3 \times [-\log_2(2\ell) - (-2/3)]/13, \tag{1}$$

where 3 is the number of bands per octave, 13 is the total number of bands, and the difference is the number of octaves over which the score is predicted to be 1. This can be further simplified to

$$\mathrm{STI} = -0.33 \times \ln(\ell) - 0.077. \tag{2}$$

This theoretical STI is consequently linear with the logarithm of segment-reversal duration as defined above, with the only exception being that values cannot exceed 1 consistent with truncation employed by Houtgast and Steeneken (1985). Resulting STI values are subsequently bounded between 0 (no intelligibility) and 1 (perfect intelligibility) as it is assumed that distortion with duration less than 40 ms (corresponding to half the period of the lowest modulation frequency considered) has no effect on STI. Estimates of intelligibility for control sentences (no distortion) were not included in the analysis.[4]

3. CSE analysis

As an alternative to conventional physical units such as milliseconds or hertz, analyses motivated by information-theoretic considerations were conducted to explore factors underlying intelligibility of temporally distorted sentences. Because there is no information in anything that is redundant or predictable, only that which is uncertain or unpredictable conveys information (Shannon, 1948).[5] The complement to predictability is entropy, and here relative entropy of the speech spectrum across time was operationalized as changes in cochlea-scaled spectral Euclidean distance (CSE). Stilp and Kluender (2010) report a robust relationship between sentence intelligibility and the amount of CSE replaced by noise in sentence materials. When equal-duration intervals are replaced with noise, intelligibility is impaired much more when replaced intervals included high versus low CSE. This

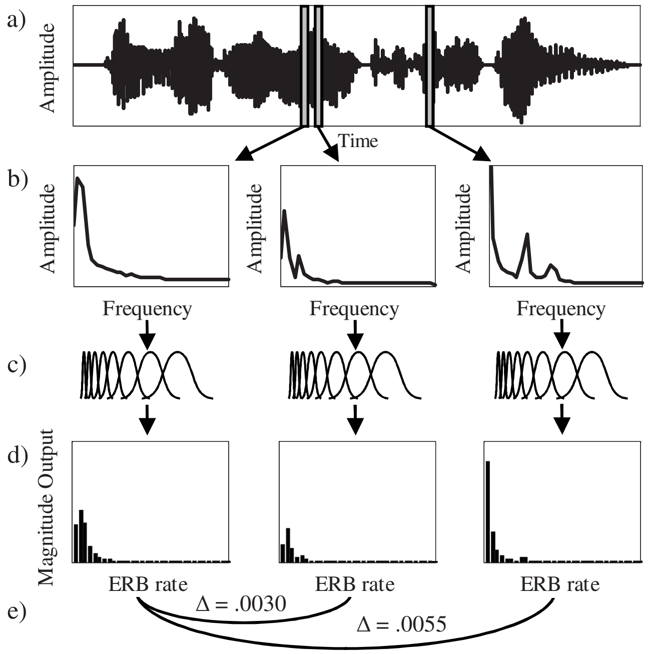

FIG. 4. Process used to measure cochlea-scaled spectral entropy (CSE). (a) Sentences were first RMS-normalized and divided into 16-ms slices, as indicated by the thin rectangular slices. (b) The magnitude spectrum for each slice was captured by a 66-point FFT. (c) ROEX filters were employed to mimic the nonlinear weighting and frequency distribution along the cochlea. The filter bank covered all frequencies up to the Nyquist frequency of 8 kHz. (d) Magnitude spectra processed by 33 ROEX filters, producing magnitude outputs as functions of ERB rate. (e) Euclidean distances were calculated between a given slice of the sentence and all following slices, then averaged for each speaking rate across all sentences.

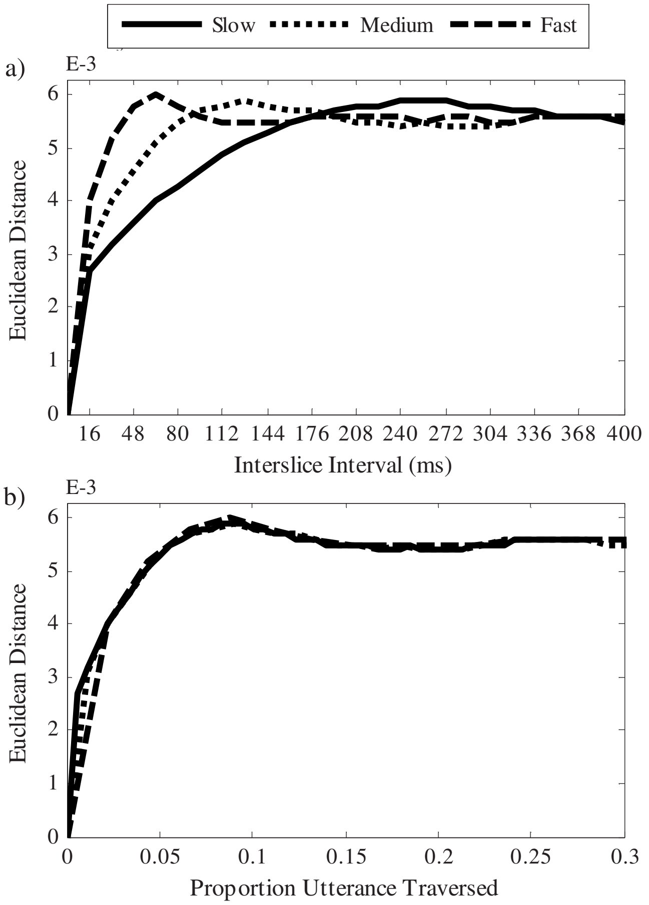

FIG. 5. Measures of cochlea-scaled entropy (CSE) in full-spectrum sentences in Experiment 1. (a) CSE, as Euclidean distance between successive inter-slice intervals. CSE functions peak at different intervals corresponding to speech rate (fast: 64 ms; medium: 128 ms; slow: 256 ms). Following this relative maximum in spectral distance, functions regress to the mean Euclidean distance between any two slices of speech as spoken by the same talker. (b) When inter-slice interval is transformed into proportion of utterance traversed, CSE measures collapse to a common function that peaks at a constant proportion of utterance traversed (approximately 0.1, or slightly less than the duration of one syllable).

strictly psychoacoustic measure is agnostic with respect to the meaning of the message being transmitted, as no knowledge about language or even speech is incorporated into calculations.

CSE analyses were conducted on sentences for each speaking rate by the process depicted in Fig. 4. Sentences were RMS-normalized and divided into 16-ms slices [Fig. 4(a)]. In addition to exceeding duration of a glottal pulse, 16-ms slice duration was chosen for the simple convenience of 256-sample windows at 16 kHz sampling frequency. Analyses were repeated using a wide range of slice durations, some of which yielded slightly but not significantly better results in predicting listener performance. Rather than select slice duration *a posteriori*, 16-ms slice duration was maintained. Prior to filtering with a bank of 33 ROEX filters, magnitude spectra from 66-point FFTs were calculated to maintain consistent density of samples along the frequency axis [Fig. 4(b)].[6] ROEX filters (Patterson *et al.*, 1982) were constructed to mimic the frequency distribution along the cochlea [Fig. 4(c)]. Filters were symmetric in log frequency with parameters defining filter tail shape fixed at four times the center frequency divided by the equivalent rectangular bandwidth (ERB; Glasberg and Moore, 1990). Filters were equally spaced 1 ERB apart with center frequencies ranging from 26 to 7743 Hz [Fig. 4(d)]. Euclidean distances (square root of the sum of squared differences) were calculated between the ERB-scaled spectrum of a given slice and each following slice [Fig. 4(e)], then averaged across all sentences. Absolute values of Euclidean distances appear relatively modest because there are many frequency bands within which either there is little or no energy, or level of energy does not change appreciably. Spectral distance functions measured for control sentences are plotted in Fig. 5. Figure 5(a) depicts spectral dissimilarity as a function of time, while Fig. 5(b) depicts the same measures as a function of utterance proportion. Like listener performance data, CSE measures collapse to a common function of proportion utterance.

4. Predicting listener performance

The STI has historically been used with medium-rate speech. The STI as implemented here is derived from the system modulation response to reversing speech segments, predicting intelligibility strictly as a function of upper modulation cutoff frequency. Therefore, four STI estimates (ℓ =0.02, 0.04, 0.08, 0.16 s corresponding to 20, 40, 80, and 160-ms segment-reversal durations, respectively) were each used three times for comparison with listener data at all three rates. STI estimates and sentence intelligibility across all

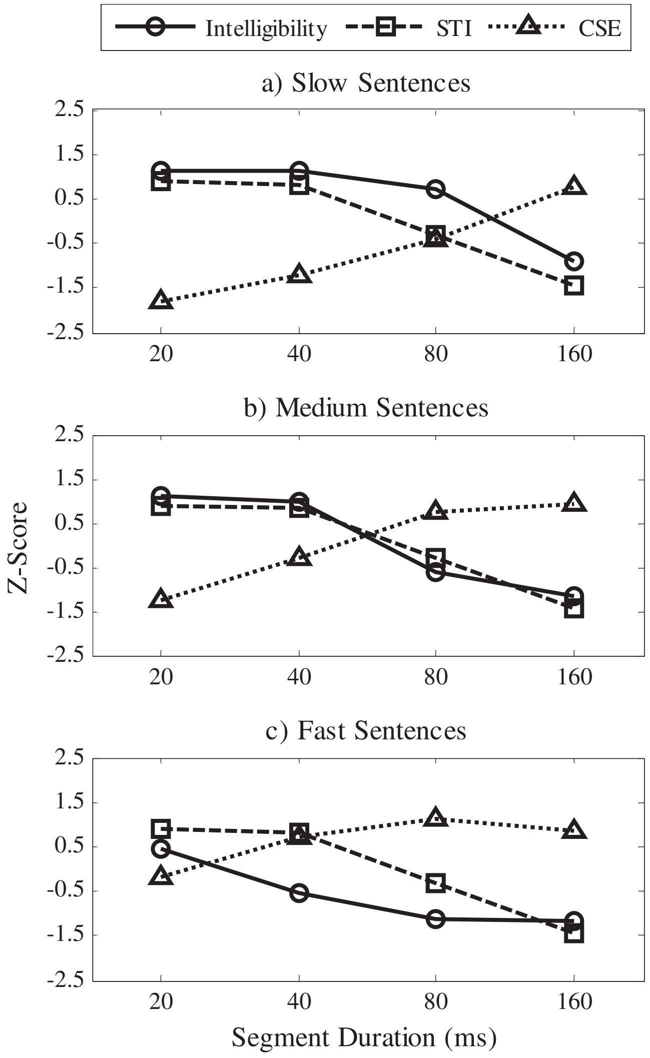

FIG. 6. Intelligibility data in Experiment 1 and corresponding predictions made by STI and CSE measures for slow (top), medium (middle), and fast (bottom) sentences. Proportion sentence correct, STI estimates, and Euclidean distances respectively are all transformed into *z*-scores and plotted to share a common ordinate. Intelligibility data is depicted in solid lines with circles; STI estimates in dashed lines with squares; and CSE measures in dotted lines with triangles. Values on the abscissa correspond to listener performance in that particular experimental condition (intelligibility), modulation frequency cutoff values of 25, 12.5, 6.25 Hz, and 3.125, respectively (STI), and interslice interval (CSE). CSE and STI measures are both correlated with listener performance across all three speaking rates.

conditions were each standardized by *z*-score transformation and are plotted in Fig. 6 with *z*-score on the ordinate and segment duration on the abscissa. Results for each speaking rate are presented in separate graphs for interpretability. Shared variance between intelligibility and STI measures was assessed through Pearson correlation, revealing a good relationship between preserved amplitude modulations and sentence intelligibility ($r^2=0.62$, $p<0.01$).

For CSE, spectral distance functions were linearly interpolated between 16-ms measurement intervals where necessary to derive distances that matched experimental segment durations (20, 40, 80, 160 ms). Similar to the analysis with STI measures, spectral distances at 0-ms lag (i.e., zero distance) were excluded. Remaining spectral distances were *z*-transformed and are plotted with intelligibility and STI as *z*-scores in Fig. 6. This simple, limited measure of information conveyed by spectral change accounts for a substantial amount of listener performance across all rate conditions ($r^2=0.89$; $p<0.001$). To assess relative predictive value of the STI versus CSE for these data, correlation coefficients were derived between each predictor (STI, CSE) and each listener's data. When r^2s were contrasted in a matched-pair *t*-test, CSE (mean=0.84, SE=0.01) accounted for significantly greater proportions of performance than did STI (mean=0.59, SE=0.01) ($t_{44}=14.98$, $p<0.001$).

Some relationship between STI estimates and CSE measures is predictable. Both measures are conceptually related as both are closely associated with syllable structure, albeit in somewhat different ways. Dominant low-frequency amplitude modulations correspond to the relative slow open-close nature of syllables given changes from relatively constricted consonants to relatively open, higher-amplitude vowels followed by consonantal constrictions. Greenberg (1999) and Greenberg and Arai (2004) have emphasized this close relationship between the MTF and the syllable structure of speech.

CSE is also dependent upon syllable structure; however, this is because physical acoustic properties of speech across the full spectrum have a local dependence (similarity) owing to coarticulation. Owing to mass and inertia of articulators (as well as planning), articulatory movements are compromises between where articulators have been and where they are headed. Because the acoustic signal directly reflects these articulatory facts, the frequency spectrum assimilates in the same fashion that speech articulation assimilates. Consequently, cochlea-scaled spectra are more similar (less Euclidean distance) close in time and more distinct at longer intervals, and these time frames are proportional to syllable duration.

CSE functions in Fig. 5(a) reveal peaks at 64 (fast sentences), 128 (medium), and 256 ms (slow), or a constant proportion of mean syllable duration across speaking rates [Fig. 5(b)]. These intervals reflect the fact that the acoustic realizations of consonant and vowel sounds are largely conditioned by preceding vowels or consonants until they begin to assimilate more to the next speech sound. For English CVCs, the identity of the second C is largely independent of the first C, and identities of vowels in successive syllables also are largely independent. Consequently, beyond these relative maxima, distances regress toward the mean Euclidean distance of any spectral sample to the long-term spectrum of speech from the same talker. This conceptual relationship between STI and CSE, both relating to syllable structure, is statistically apparent in the correlation between the two measures across all rates ($r^2=0.49$; $p<0.05$).

C. Discussion

Results from Experiment 1 are in good agreement with previous experiments in which speech is distorted in the same manner (Greenberg and Arai, 2001; Saberi and Perrott, 1999). The present experimental design followed that of Greenberg and Arai as opposed to designs that either: employed multiple presentations of the same sentence with different distortion (e.g., Greenberg *et al.*, 1998; Saberi and Perrott, 1999; Silipo *et al.*, 1999); measured intelligibility of target words rather than sentences (e.g., Flanagan, 1951); or, reported subjective intelligibility (e.g., Saberi and Perrott, 1999). Previous studies report intelligibility scores from repeated presentations fell to 50% with 130-ms distortion (Saberi and Perrott, 1999), while single presentations of more difficult sentence materials on each trial resulted in similar performance with only 60-ms distortions (Greenberg and Arai, 2001). Despite the introduction of dramatic variability in speaking rate from trial to trial, intelligibility of temporally distorted sentences is highly consistent and systematic. Previous findings are extended by demonstration that, as temporal distortion increased, rate of performance decrement was significantly less for slower (longer) sentences and significantly greater for faster (shorter) ones [Fig. 1(a)].

STI estimates reveal a consistent relationship between preserved amplitude modulations and sentence intelligibility across speaking rates and segment-reversal durations. However, two points should be made concerning STI measures. Because the STI weights modulation frequencies equally, an increase/decrease in speaking rate which results in an upward/downward shift in the location of the peak in the 3–6 Hz region of the modulation spectrum results in almost no change in the STI calculation. As long as the gross pattern of modulations across frequency bands remains intact, STI estimates remain constant with changing rate of speech. Second, congruence between STI estimates and listener performance observed in Experiment 1 are likely limited by a restriction of range in experimental conditions. Theoretical STI measures are linear with the logarithm of modulation rate. Although listener performance is roughly linear over this same range of modulation rates, presentation of sentences with shorter and longer segment-reversal durations than those presently investigated would produce a sigmoidal STI function, resulting in ceiling and floor performance, respectively [i.e., leftward and rightward extrapolation of Fig. 1(b)]. This possibility is among the reasons for employing a different type of temporal distortion in Experiment 2, through which comparisons of STI and CSE can be made across a wider range of temporal distortions.

While STI and CSE are clearly not independent of one another, they capture different as well as similar aspects of the signal and of listener performance. This can be inferred from the considerable but not complete overlap between measures ($r^2=0.49$). Unlike the STI, measures of potential information (e.g., CSE) change with speaking rate and thus make different predictions about intelligibility data across all three speaking rates. Similar to intelligibility data, measures of CSE over segment-duration intervals reveal increasing spectral dissimilarity with increasing temporal displacement. In addition, while CSE measures are less affected by the constrained range of modulations investigated, they are distinguished by better fits at longer durations of temporal distortion. Figure 5 reveals that CSE asymptotes at longer lags, much in the same way that listener performance remains at floor levels with longer segment reversals.

Another attractive property of CSE is that it requires no explicit rate normalization. CSE measures naturally accommodate variable-rate sentence materials, while STI estimates are indifferent to rate. While there have been substantial efforts to better understand how listeners normalize across speaking rate when reporting perception of individual consonants (e.g., Miller, 1981; Miller and Liberman, 1979), vowels (e.g., Ainsworth, 1972, 1974; Gottfried *et al.*, 1990), or words (e.g., Miller and Dexter, 1988), much less attention has been paid to how distortion affects sentence recognition across variation in rate. To the extent that potential information, not time and frequency per se, accounts for perception, concerns about normalization of time or frequency toward some iconic standard may dissolve. While durations and modulation frequencies may vary, information content remains relatively constant and requires no such normalization.

III. EXPERIMENT 2

Experiment 2 is designed to investigate perceptual resilience to temporal distortion, across a wider range of distortions, for varying rates of speech. In addition to providing this greater range, this experiment permits greater dissociation between STI and CSE as predictors of listener performance. Adopting methods of Flanagan (1951) and parameters of Greenberg *et al.* (1998), the speech spectrum is filtered into four one-third-octave bands, and the onsets of two of these are desynchronized by fixed delays relative to the other two. While this method of temporal distortion greatly distorts CSE measures, MTF (and STI) measures do not change because amplitude modulations within each band are preserved. A control study was also conducted to determine the intelligibility of individual pairs of speech bands.

A. Method

1. Listeners

Eighty-four native English speakers were recruited from the Department of Psychology at the University of Wisconsin—Madison (42 in the control study; 42 in Experiment 2). No listeners participated in more than one experiment. None reported any hearing impairment and all received course credit for participation.

2. Stimuli

All processing was performed in MATLAB. The control study and Experiment 2 used the full corpus of 115 synthesized sentences described in Experiment 1. Following the methods of Greenberg *et al.* (1998), sentences were filtered by third-octave 6$^{\text{th}}$-order Butterworth filters centered at 335, 850, 2135, and 5400 Hz such that each channel was separated by a one-octave stop band.

In the control study, stimuli varied in seven filter conditions: all six pairwise combinations of the four bands as well as the four-band composite. Each of 105 sentences was synthesized at 3 simulated speaking rates and arranged in 7 filter pairings, generating 2205 stimuli. In addition, 10 sentences were presented at one rate and one filter pair each as practice sentences, resulting in 2215 stimuli in all. Following synthesis, sentences were normalized to equal overall RMS amplitude and upsampled to 44.1 kHz for presentation.

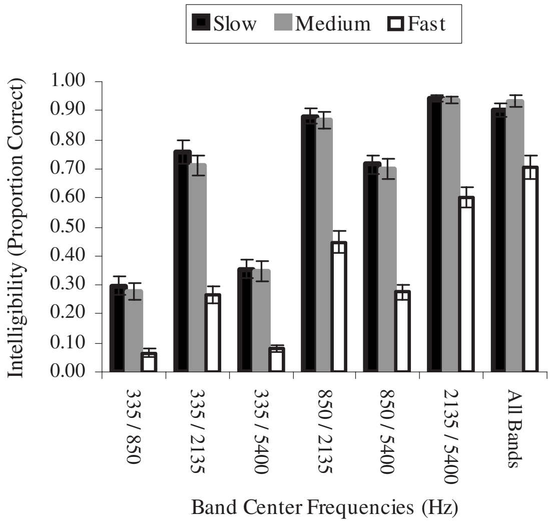

FIG. 7. Results from the control study preceding Experiment 2. Intelligibility (proportion of words correctly identified) is plotted as a function of distortion condition (six two-band conditions, one four-band composite condition). Black bars denote performance for slow-rate sentences, gray bars for medium-rate sentences, and white bars for fast-rate sentences. Error bars depict standard error of the mean. Middle-frequency bands (850-/2135-Hz CF) and low- and high-frequency bands (335-/5400-Hz CF) were paired respectively for use in Experiment 2.

Following Greenberg *et al.* (1998), in Experiment 2 the lowest and highest bands (335- and 5400-Hz CFs) were paired separately from the two middle-frequency bands (850- and 2135-Hz CFs), and were delayed relative to the middle-frequency pair by fixed amounts.[7] Onsets of the 335- and 5400-Hz bands were delayed by multiple durations (0, 25, 50, 100, 200, 400, and 600 ms) relative to onsets of the 850- and 2135-Hz bands. From the 115 sentences, 10 were used as practice trials and were presented at only one speaking rate and one band pair delay. The remaining 105 sentences, generated at each of the three speaking rates, were distorted at each of seven relative delays. Sentences were again normalized to equal overall RMS amplitude and upsampled to 44.1 kHz.

3. Procedure

Procedures for the control study and Experiment 2 are the same as that of Experiment 1 with four exceptions. First, the number of trials changed from 12 practice and 75 experimental to 10 and 105, respectively. A second difference is that each stimulus was presented twice across the group of listeners instead of three times. Third, 42 listeners participated in the control study and Experiment 2, versus 45 in Experiment 1. Finally, the experimental session lasted approximately 40 min.

B. Results

1. Listener performance

a. Control study Results from the control study are plotted in Fig. 7 (inter-rater reliability: $r > 0.99$). Intelligibility is plotted for each distortion condition (six two-band conditions, one four-band composite condition). Data were analyzed in a repeated-measures ANOVA with three levels of rate (slow, medium, and fast) and seven levels of distortion (six band pairs and one four-band composite). Listener performance differed across speaking rate ($F_{2,82}=517.30$, $p < 0.001$, $\eta^2_p=0.93$). Tukey HSD *post-hoc* tests reveal that slow and medium-rate conditions were not significantly different from one another, but performance with fast sentences was significantly poorer ($\alpha=0.05$). Intelligibility also varied as a function of band pairs ($F_{6,246}=289.75$, $p < 0.001$, $\eta^2_p=0.88$). The interaction between rate and band pairs was also significant ($F_{12,492}=9.01$, $p < 0.001$).

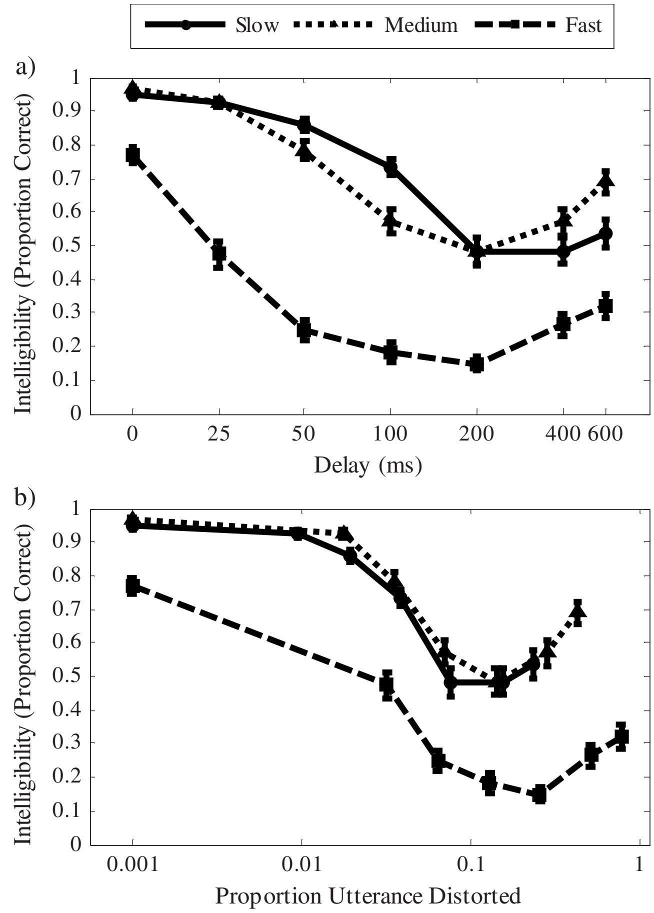

FIG. 8. Results from Experiment 2. (a) Intelligibility is plotted as a function of band pair delay. Solid lines with circles depict results for slow-rate sentences, dotted lines with triangles for medium-rate sentences, and dashed lines with squares for fast-rate sentences. Error bars depict standard error of the mean. Across sentence rates, intelligibility decreases to a relative minimum before improving at longer delays. (b) The same intelligibility results are plotted as a function of proportion of utterance distorted. Across speaking rates, relative minima in intelligibility all occur at roughly equal proportions of the utterance, relating to the duration of one syllable.

b. Experiment Results from Experiment 2 are presented in Fig. 8 (inter-rater reliability: $r > 0.99$). Intelligibility in 0-ms delay conditions of Experiment 2 correspond well to the same condition in the control study (four-band composite; Fig. 7). Results were analyzed in a repeated-measures, 3 (rate) by 7 (delay) ANOVA. Similar to Experiment 1, simulated rate differentially affected intelligibility ($F_{2,82}=580.35$, $p < 0.001$, $\eta^2_p=0.93$). Similar to the control study, Tukey HSD tests reveal that slow and medium-rate sentences resulted in similar intelligibility performance, which was significantly better than performance for fast sentences ($\alpha=0.05$). Intelligibility follows a non-monotonic function of delay ($F_{6,246}=145.92$, $p < 0.001$, $\eta^2_p=0.78$).

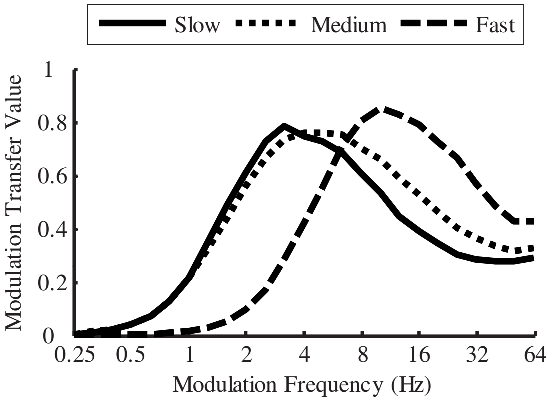

FIG. 9. MTFs of experimental sentences in Experiment 2. Solid lines depict slow-rate sentences, dotted lines medium-rate sentences, and dashed lines fast-rate sentences. Modulation transfer value is plotted on the ordinate and modulation frequency on the abscissa. Because band pair delay has no effect on modulation transfer values, this single set of MTFs represents sentence modulations in all durations of band pair delay.

Tukey HSD tests reveal that for shorter delays, performance decreased with increasing delay of the low- and high-frequency bands (0-ms-delay intelligibility > 25 ms, α = 0.05; 25 ms > 50 ms > 100 ms, α = 0.01; 100 ms > 200 ms, α = 0.05). At longer delays, performance recovered slightly (600 ms > 200 ms, α = 0.01), replicating the same observation in prior studies examining intelligibility of desynchronized speech bands (Flanagan, 1951; Silipo *et al.*, 1999). The interaction between rate and distortion was also significant ($F_{12,492} = 17.72$, $p < 0.001$).

Delay duration was transformed into proportion of utterance distorted [delay duration divided by mean sentence duration; Fig. 8(b)]. Rescaled results exhibit a high degree of overlap. However, fast sentences filtered into four bands were considerably more difficult across all delay conditions (see also four-band composite results in Fig. 7). Across speaking rates, performance minima are roughly equivalent at approximately average syllable duration (i.e., 0.14 or one-seventh of total seven-syllable utterance).[8] Beyond this point, intelligibility improved with increasing delays. With greater delays, performance approached intelligibility of mid-frequency bands presented in isolation in the control experiment (Fig. 7).

2. MTF/STI analysis

MTF and STI analyses in Experiment 2 differed significantly from those reported in Experiment 1. MTFs of experimental sentences were derived following the same procedure outlined in Experiment 1. However, MTF calculations are insensitive to band delay. Therefore, all experimental sentences are represented by only three MTF functions, one per speaking rate, depicted in Fig. 9. This is in contrast to Experiment 1, where each reversal-duration condition altered the MTF (see Figs. 2 and 3). In addition, bandpass filtering sentences in Experiment 2 seemingly high-pass filtered the MTF of slow sentences, as modulation indices 4 Hz and below overlap extensively with those in medium-rate sentences; otherwise, Fig. 9 resembles Fig. 2(a) (MTFs of full-spectrum, undistorted sentences used in Experiment 1).

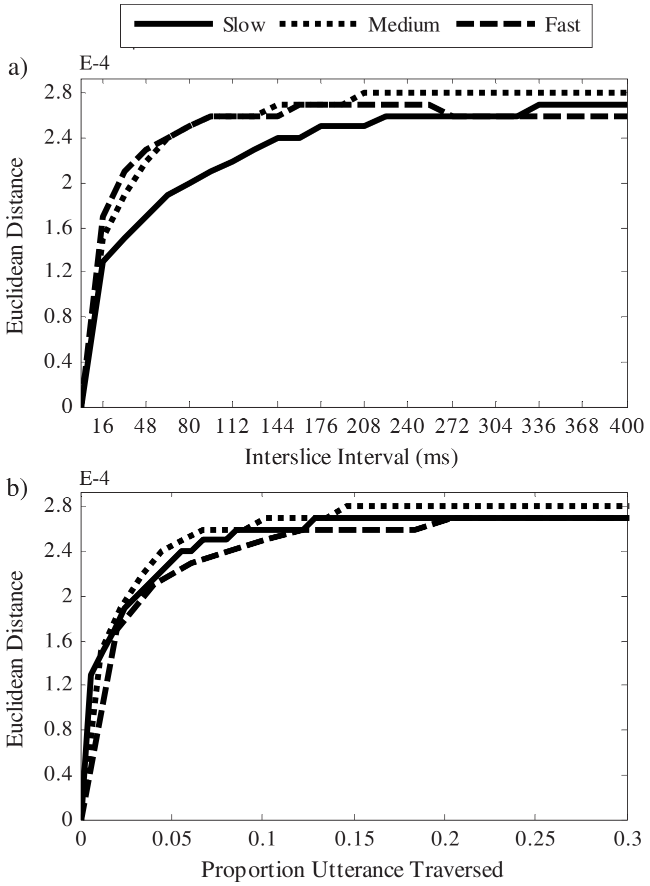

FIG. 10. Measures of CSE in filtered-spectrum sentences in Experiment 2. As in Fig. 5, CSE is shown as functions of inter-slice interval (a) and proportion of utterance traversed (b). Due to far less sampling in the frequency domain following filtering, CSE functions are less well-defined than those observed for full-spectrum sentences in Fig. 5. CSE functions lack the clear relative maximum and regression to mean Euclidean distance, instead asymptoting at the mean Euclidean distance between any two slices of speech as spoken by the same talker. Distance functions across this wide range of speaking rates converge to a common function when rescaled onto proportion of utterance traversed (b).

Because band delay has no effect on modulations as measured by the MTF, STI cannot predict changes in performance at all. Thus, Experiment 2 presents another example, in addition to the results of Longworth-Reed *et al.* (2009), where listener intelligibility is severely compromised despite unperturbed MTF.

3. CSE analysis

Analyses of CSE were conducted on the expanded corpus of four-band, filtered sentences. The only difference between this analysis and that described in Experiment 1 is sampling in the frequency domain. In Experiment 1, CSE analyses included 33 real points of the 66-point FFT, spanning frequencies up to 8 kHz. Here, 25 of the 33 ERB-spaced samples in the frequency domain fall within stop bands following filtering. As a result, CSE analyses use the same 66-point FFTs to maintain spectral resolution, but Euclidean distances were calculated between only outputs of the eight ERB filters (two per passband). Mean results across all control sentences are plotted in Fig. 10. Similar to Fig. 5(a), CSE increases with increasing duration between samples [Fig.

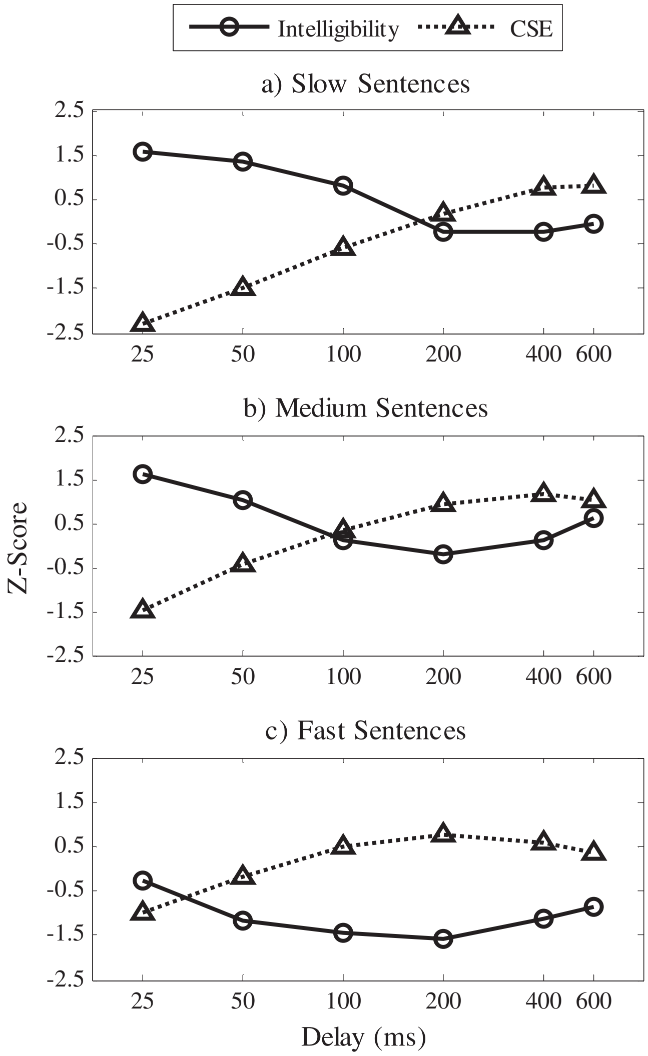

FIG. 11. Intelligibility data from Experiment 2 and corresponding predictions made by CSE measures for slow (top), medium (middle), and fast (bottom) sentences. As the MTF (and subsequent STI) is insensitive to band-pair delay, it produces the same predictions for all delay conditions and is not shown here. All labeling matches that used in Fig. 6. Delay conditions of 25, 50, 100, 200, 400, and 600 ms are plotted on the abscissa of each graph. Similar to Experiment 1, CSE measures are significantly correlated with listener performance across all three speaking rates.

10(a)]. When transformed into proportions of utterance, measures collapse to a common function as also observed in Fig. 5(b).

CSE measures were z-transformed, as were intelligibility results from Experiment 2, and are plotted in Fig. 11. Similar to Experiment 1, CSE is significantly inversely related to listener intelligibility (r^2=0.40; $p<0.01$). CSE measures across such a small number of filter outputs are expected to be poorer predictors of listener performance, especially to the extent that there is no measure of listeners' ability to 'fill in' gaps via expectations from experience with full-spectrum speech. In addition, the overall correlation is undermined by performance data for fast sentences, for which intelligibility is lower at every level of distortion. Separate r-squared values for performance at six band delays at each rate are somewhat higher (fast r^2=0.75, $p<0.05$; medium r^2=0.75, $p<0.05$; slow r^2=0.92, $p<0.01$). Even sparse CSE measures significantly contribute to predicting listener performance.

C. Discussion

Experiment 2 replicates results reported by Flanagan (1951) and Silipo *et al.* (1999) that intelligibility of temporally desynchronized bands in medium-rate speech improves slightly at very long delays. That finding is extended here, as performance improvement appears to occur at delays that exceed mean syllable duration across a wide range of speaking rates. Improvement in listener performance at longer delays may result, at least in part, from ability to integrate information from band pairs with spectra effectively independent of one another on temporal and spectral bases.

Aside from being indifferent to speech rate, STI is insensitive to band-pair delay and is incapable of predicting listener performance in Experiment 2. Across all conditions in both experiments, only CSE corresponds well to listener performance. Changes in intelligibility (Fig. 8), despite stable long-term modulation characteristics (Fig. 9), reveals another instance in which amplitude modulations are poor predictors of listener performance (e.g., Drullman *et al.* 1994a, 1994b; Longworth-Reed *et al.*, 2009). The fact that CSE accounts for over 40% of the variance in intelligibility scores in Experiment 2, and shares over 80% of variance with listener data in Experiment 1, when all frequencies are included, suggests that relative change in spectral composition may be a major contributor to listener performance. This measure of potential information generalizes across speaking rates and means of temporal distortion, capturing patterns of listener performance in ways that the STI or the MTF cannot.

IV. GENERAL DISCUSSION

Perception of speech is remarkably resilient to extreme signal degradation, owing to the substantial experience listeners have with speech and its multiplicity of largely redundant cues (e.g., Assmann and Summerfield, 2004; Kluender and Alexander, 2008; Kluender and Kiefte, 2006). Perceptual resilience to temporal distortion was investigated here in two distinct manners. Successive segments of constant duration were time-reversed (Experiment 1), and onsets of two filtered bands were desynchronized by a constant delay relative to the other two bands (Experiment 2). Consistent with prior findings of Arai and Greenberg (1998), Flanagan (1951), Greenberg and Arai (2001), Saberi and Perrott (1999), and Silipo *et al.* (1999), perception is resilient when faced with modest amounts of temporal distortion.

How much temporal distortion is necessary to significantly impair performance varies substantially and systematically with changing speech rate. The present experiments employed a wide range of speaking rates, from 2.5 (slow) syllables per second, to 5 (medium) to 10 (fast). When intelligibility data are transformed to proportions of utterance distorted [Figs. 1(b) and 8(b)], performance is remarkably consistent across all rates. Aside from the increased difficulty of filtered fast sentences (cf., Figs. 7 and 8 for Experiment 2), performance collapses onto one common function [Experiment 1, Fig. 1(b)] or follows repeated and largely overlapping patterns [Experiment 2, Fig. 8(b)].

Several reports have noted the relationship between intelligibility of temporally distorted speech and measurements of intensity modulation in undistorted speech using the MTF (Houtgast and Steeneken, 1985), and some authors have suggested that low-frequency modulation envelopes (3–8 Hz) corresponding roughly to syllabic structure are critical to speech intelligibility (e.g., Arai and Greenberg, 1998; Greenberg, 1999; Greenberg and Arai, 2004; Greenberg *et al.*, 1998; Saberi and Perrott, 1999). The significance of amplitude modulation information was assessed by calculating MTFs (Houtgast and Steeneken, 1985) for each set of experimental sentences, and correlating derived STI measures with listener intelligibility. Results were mixed. Predictions made by STI estimates in Experiment 1 (Fig. 6) were substantial but difficult to evaluate given the restricted range of time-reversed temporal distortions. With a wider range of temporal distortion using band-pair delays, MTFs (and subsequent STI estimates) yielded predictably poor correspondence to intelligibility because they are insensitive to this type of temporal distortion.

Notably, this inability to predict performance for experimentally distorted speech employed here does not undermine the practical utility of the MTF and its scalar counterpart the STI. The MTF and STI continue to serve as useful predictors of speech perception under adverse conditions, such as room acoustics for which the measures were designed. It is likely that the predictive values of the MTF and STI lie in the way they characterize the relative integrity of syllable structure, particularly by sensitivity to modulations in the region of 3–8 Hz.

Perhaps the most striking result from these experiments is the consistency with which psychoacoustically-inspired measures of potential information (CSE) predict listener performance across speaking rates. More sophisticated measures that capture psychoacoustic properties in addition to spectral change may account for even more of the variance in performance than did the simple measure employed above. For example, measures of periodicity could be incorporated to capture differences between periodic and relatively aperiodic portions of the signal or other changes in fundamental frequency. Further, the relatively simple calculation of CSE is not expected to bear such close correspondence to listener intelligibility across all types of signal distortion. There exist manipulations of the speech signal, such as entire sentence reversal, which will preserve the rate of spectral change over time (thus preserving CSE) but render speech unintelligible. Nevertheless, the present data clearly recommend that emphasis upon physical metrics (e.g., 3–8 Hz modulations) may miss the central fact that perception can be better explained by equally quantifiable measures such as psychoacoustic entropy.

It is no accident that intelligibility coincides so closely with CSE to the extent that relative spectral similarity is a consequence of the structure of speech signals. Connected speech is produced with a series of opening and closing gestures (consonants) surrounding intervals when the vocal tract is relatively open (vowels). For English and most other languages, one consonant or vowel (n) minimally predicts the next consonant or vowel ($n+1$); predictability of the next sound ($n+2$) is even lower (e.g., Shannon, 1948, 1951). Across speaking rates, at approximately two-thirds of mean syllable duration, CSE reaches its relative maximum [peaks in Fig. 5(a); approximate asymptotes in Fig. 10(a)]. Soon after this point of relative spectral independence, performance across experiments reaches its relative minimum, coinciding roughly with mean syllable duration. Simple measures of time would fail to capture this robust relationship. However, viewing intelligibility as a function of proportions of utterance distorted allows time to be collapsed across all three talker rates and for results to be viewed with a single, common metric.

Results from Experiment 2 pose interesting questions about which spectral-temporal details listeners may use to combine information across frequency bands. One hypothesis is that onsets may play an important role in synchronizing temporally displaced acoustic information. The auditory system is known to be selectively tuned to these acoustic events (e.g., Rhode, 1991; Rhode and Greenberg, 1992, 1994) and the degree to which sentences possess sharp changes in onset characteristics has been shown to be positively correlated with sentence intelligibility in virtual rooms with different reverberation characteristics (Longworth-Reed *et al.*, 2009).

Results from Experiment 2 also present an interesting exception to the otherwise close relationship between CSE and intelligibility. Replicating the same effect observed by Flanagan (1951) and Silipo *et al.* (1999), intelligibility increased rather than decreased at especially long periods of desynchronization. CSE measures, on the other hand, asymptote soon after mean syllable duration. What characteristics of the signal explain listeners' improved performance despite such grave temporal distortion? One possibility is that, when cross-channel asynchrony is increased beyond average syllable duration, channels may be perceptually segregated such that listeners might utilize only the two central bands while ignoring the other two. Although this is not the only possible hypothesis, it bears note that intelligibility at long delays approximates performance with mid-frequency bands only (Fig. 7).

These factors suggest that listeners hearing temporally-displaced frequency bands may use information in the speech signal that is independent of CSE. For example, differences in fundamental frequency information (f_0) across channels may help listeners be more resilient to such temporal asynchrony as seen in Experiment 2, and the present measure of changes in spectral shape is relatively insensitive to f_0. Fundamental frequency is known to be a reliable component for separating acoustic events from the environment (Bregman, 1990). Differences in f_0 of vowels heard simultaneously facilitate perception of both vowels, while similar f_0 information conversely impairs perception and limits listeners' ability to report which vowels were played (Assmann and Summerfield, 1989; Broadbent and Ladefoged, 1957; Darwin, 1981). Thus, local f_0 differences between syllables may aid listeners in segregating desynchronized bands. This perceptual segregation may actually aid correct identification in that listeners can focus attention on a single band pair rather than relying upon integration of potentially conflicting desynchronized bands. Finally, perceptual segregation of adjacent bands may be enhanced by their delayed start times (e.g., Bregman, 1990; Darwin, 1981). It is possible that intelligibility may be overestimated because listeners may compensate for constant temporal differences detected between channels that were apparent at the beginning of each sentence.

Results presented here provide evidence that encourage information-theoretic approaches to understanding speech perception, somewhat in contrast to traditional approaches that emphasize physical acoustic dimensions such as time that subsequently may require additional scaling or normalization. Some attention to potential information appears necessary to account for the present data, and it shares a long history of utility for understanding human performance. Perceptual systems do not record absolute levels of stimulus dimensions, whether loudness, pitch, brightness, or color, and this has been demonstrated perceptually in every sensory domain (e.g., Kluender *et al.*, 2003). A host of recent experimental findings demonstrate that change, versus absolute acoustic characteristics, is fundamental to speech perception (Kluender *et al.*, 2003; Alexander and Kluender, 2008; Stilp and Kluender, 2010), and the auditory system calibrates to predictability in the signal in ways that emphasize unpredictable characteristics (information) for perception of speech and nonspeech sounds (e.g., Kiefte and Kluender, 2008; Stilp *et al.*, 2010). Present results illustrate the close inverse relationship between CSE and perceptual resilience of temporally distorted speech. These findings encourage increased investigation of information-theoretic approaches to speech perception and to perception more broadly.

ACKNOWLEDGMENTS

The authors are grateful to Brian C.J. Moore for thoughtful suggestions, to Benjamin Hesprich, Stephanie Jacobs, Amanda Baum, Rebecca Edds, Tricia Nechedom, and Rebecca Strauss for assistance in conducting these studies, and to Elizabeth Shaut, Kyira Hauer, and Erin Nagy for data scoring. The authors also thank Steven Greenberg, Kourosh Saberi, and two anonymous reviewers for comments on a previous draft of this manuscript. This work was supported by Grant DC 009532 from the National Institute of Deafness and Communicative Disorders to the first author, a grant from the Canadian Social Sciences and Humanities Research Council to the second author, and Grant DC 004072 from the National Institute of Deafness and Communicative Disorders to the fourth author.

APPENDIX: SCORING GUIDELINES

Three independent raters were trained to score listener data according to the following guidelines using data from pilot versions of these experiments.

1. One point was awarded for each word correctly identified, up to the maximum number of points (i.e., number of words in the sentence), then was divided by that maximum.
2. Responses of ‘X’ or no response at all earned zero points.
3. No penalty was instituted for guessing or incorrect words.
4. Punctuation, capitalization, and word order were not factored into scoring.
5. Per the HINT scoring guidelines (Nilsson *et al.*, 1994), “a” and “the” were considered interchangeable.
6. Words partially correct were scored depending upon how they sounded, given that the correct meaning was present.
   a. Incorrect regular verb forms: incorrect verb endings were scored as correct so long as the pronunciation of the verb root was unchanged (*e.g.*, ‘help’ is a correct response to ‘helped’, but ‘drink’ is an incorrect response to ‘drank’).
   b. Noun number: incorrect noun number was scored as correct so long as the pronunciation of the noun was unchanged (*e.g.*, ‘coin’ is a correct response to ‘coins’, but ‘man’ is an incorrect response to ‘men’).
   c. Homophone part-words: responses with the same initial sounds but clearly different meaning than the target word was scored as incorrect (*e.g.*, ‘personal’ is an incorrect response to ‘purse’).
7. Typographical errors such as two-letter transpositions were acceptable, so long as the intended response was clear (*e.g.*, ‘mathces’ is a correct response to ‘matches’).

[1] Asynchrony tolerance has been shown to be related to the amount of spectral information available for across-frequency integration of timing cues. Healy and Bacon (2007) report that intelligibility of pairs of pure tones amplitude-modulated by speech-shaped intensity envelopes is reduced by as little as 12.5 ms of asynchrony, with floor performance reached at 100-ms asynchrony. Remez *et al.* (2008) drastically reduced intelligibility of sine-wave replicas of speech by desynchronizing the second formant analog by only 50 ms. Similarly, Fu and Galvin (2001) report that removal of fine spectral structure renders speech perception considerably more susceptible to effects of cross-channel asynchrony.

[2] This effect is not predictable based on the magnitude component of the MTF, but it bears note that the phase component of the MTF does change (e.g., Greenberg and Arai, 2001).

[3] Online demonstration available at http://www2.research.att.com/~ttsweb/tts/demo.php (Last viewed 8/30/10).

[4] Aside from attempting to calculate the log of 0 (control condition/no segment reversal), several factors preclude assuming perfect intelligibility for control sentences (i.e., setting the STI equal to 1), including but not limited to speaking rate manipulations, sentence predictability, listener attention, and the fact that listener performance is affected by at least some factors that cannot be captured by amplitude modulations (e.g., Drullman *et al.*, 1994a, 1994b; Longworth-Reed *et al.*, 2009).

[5] A conventional unit of measure in information theory, the bit, is often used to estimate potential information amidst a fixed number of possible outcomes with known probabilities. Because current efforts are interested in investigating potential information transmission across the continuously variable, complex spectrum of speech, measurement in bits is not practical, necessary, or perhaps appropriate here (see e.g., Linsker, 1988).

[6] Measures of differences between spectral slices vary insignificantly depending upon whether frequency is scaled in linear Hz (FFT) or in ERB, and amplitude is scaled in linear pressure/volts or in logarithmic dB. All combinations correlate well with performance data in Experiment 1: $\mathrm{FFT}_{\mathrm{V}}$ $r^2=0.87$; $\mathrm{ERB}_{\mathrm{V}}$ $r^2=0.89$ (used here); $\mathrm{FFT}_{\mathrm{dB}}$ $r^2=0.84$; $\mathrm{ERB}_{\mathrm{dB}}$ $r^2=0.89$.

[7] Results from the control study suggest that perhaps pairing bands in an alternating fashion would provide a stronger test of intelligibility, given highly comparable performance for band pairs 335/2135 and 850/5400 as illustrated in Fig. 7. Nevertheless, bands were paired in the described fashion to replicate the methods of Greenberg *et al.* (1998). In addition, intelligibility performance in Greenberg *et al.* (1998) was reported to be equivalent whether onset of middle-frequency bands was delayed relative to onset of low- and high-frequency bands or vice versa.

[8] Relative minima in Fig. 8 also give some suggestion of alignment at 200 ms desynchronization across all speaking rates. It is unlikely that 200 ms is a privileged duration across all speaking rates, given that it is markedly shorter than mean slow-sentence syllable duration and longer than mean fast-sentence syllable duration. Further, post hoc tests indicate that fast

sentence performance at 100 ms and 200 ms desynchronization do not significantly differ from one another, and the same holds for slow sentence performance at 200 and 400 ms desynchronization.

Ainsworth, W. A. (**1972**). "Duration as a cue in the recognition of synthetic vowels," J. Acoust. Soc. Am. **51**, 648–651.

Ainsworth, W. A. (**1974**). "The influence of precursive sequences on the perception of synthesized vowels," Lang Speech **17**, 103–109.

Alexander, J. M., and Kluender, K. R. (**2008**). "Spectral tilt change in stop consonant perception," J. Acoust. Soc. Am. **123**, 386–396.

Arai, T., and Greenberg, S. (**1998**). "Speech intelligibility in the presence of cross-channel spectral asynchrony," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 933–936.

Assmann, P. F., and Summerfield, Q. (**1989**). "Modeling the perception of concurrent vowels: Vowels with the same fundamental frequency," J. Acoust. Soc. Am. **85**, 327–338.

Assmann, P. F., and Summerfield, Q. (**2004**). "The perception of speech under adverse conditions," in *Speech Processing in the Auditory System*, edited by S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay (Springer, New York), Vol. **14**, pp. 231–308.

Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., and Syrdal, A. (**1997**). "AT&T natural voices text-to-speech [computer software]," http://www2.research.att.com/~ttsweb/tts/demo.php (Last viewed 8/30/2010).

Boersma, P., and Weenink, D. (**2007**). "Praat: Doing phonetics by computer (version 9 4.5.12) [computer program]," http://www.praat.org/ (Last viewed 1/31/2007).

Bregman, A. (**1990**). *Auditory Scene Analysis* (MIT, Cambridge, MA), pp. 1–790.

Broadbent, D. E., and Ladefoged, P. (**1957**). "On the fusion of sounds reaching different sense organs," J. Acoust. Soc. Am. **29**, 708–710.

Darwin, C. J. (**1981**). "Perceptual grouping of speech components differing in fundamental frequency and onset-time," Q. J. Exp. Psychol. A **33**, 185–207.

Drullman, R., Festen, J. M., and Plomp, R. (**1994a**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**, 1053–1064.

Drullman, R., Festen, J. M., and Plomp, R. (**1994b**). "Effect of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Am. **95**, 2670–2680.

Dudley, H. (**1939**). "Remaking speech," J. Acoust. Soc. Am. **11**, 169–177.

Flanagan, J. L. (**1951**). "Effect of delay distortion upon the intelligibility and quality of speech," J. Acoust. Soc. Am. **23**, 303–307.

French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.

Fu, Q.-J., and Galvin, J. J., III (**2001**). "Recognition of spectrally asynchronous speech by normal-hearing listeners and Nucleus-22 cochlear implant users," J. Acoust. Soc. Am. **109**, 1166–1172.

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.

Gottfried, T. L., Miller, J. L., and Payton, P. E. (**1990**). "Effect of speaking rate on the perception of vowels," Phonetica **47**, 155–172.

Greenberg, S. (**1999**). "Speaking in shorthand—A syllable-centric perspective for understanding pronunciation variation," Speech Commun. **29**, 159–176.

Greenberg, S., and Arai, T. (**2001**). "The relation between speech intelligibility and the complex modulation spectrum," in 7th International Conference on Speech Communication and Technology, Scandinavia, pp. 473–476.

Greenberg, S., and Arai, T. (**2004**). "What are the essential cues for understanding spoken language?," IEICE Trans. Inf. Syst. **E87-D**, 1059–1070.

Greenberg, S., Arai, T., and Silipo, R. (**1998**). "Speech intelligibility derived from exceedingly sparse spectral information," in Proceedings of the 5th International Conference on Spoken Language Processing, Sydney, Australia, pp. 74–77.

Healy, E. W., and Bacon, S. P. (**2007**). "Effect of spectral frequency range and separation on the perception of asynchronous speech," J. Acoust. Soc. Am. **121**, 1691–1700.

Houtgast, T., and Steeneken, H. J. M. (**1973**). "The modulation transfer function in room acoustics as a predictor of speech intelligibility," J. Acoust. Soc. Am. **54**, 557.

Houtgast, T., and Steeneken, H. J. M. (**1985**). "A review of the MTF-concept in room acoustics," J. Acoust. Soc. Am. **77**, 1069–1077.

Kiefte, M., and Kluender, K. R. (**2008**). "Absorption of reliable spectral characteristics in auditory perception," J. Acoust. Soc. Am. **123**, 366–376.

Kluender, K. R., and Alexander, J. M. (**2008**). "Perception of speech sounds," in *The Senses: A Comprehensive Reference*, edited by A. I. Basbaum, A. Kaneko, G. M. Shepard, andG. Westheimer (Academic, San Diego, CA), Vol. **3**, pp. 829–860.

Kluender, K. R., Coady, J. A., and Kiefte, M. (**2003**). "Sensitivity to change in perception of speech," Speech Commun. **41**, 59–69.

Kluender, K. R., and Kiefte, M. (**2006**). "Speech perception within a biologically-realistic information-theoretic framework," in *Handbook of Psycholinguistics*, edited by M. A. Gernsbacher and M. Traxler (Elsevier, London), pp. 153–199.

Linsker, R. (**1988**). "Self-organization in a perceptual network," Computer **21**, 105–117.

Longworth-Reed, L., Brandewie, E., and Zahorik, P. (**2009**). "Time-forward speech intelligibility in time-reversed rooms," J. Acoust. Soc. Am. **125**, EL13–EL19.

Miller, J. L. (**1981**). "Effects of speaking rate on segmental distinctions," in *Perspectives on the Study of Speech*, edited by P. D. Eimas and J. L. Miller (Erlbaum, Hillsdale, NJ,), pp. 39–74.

Miller, J. L., and Dexter, E. R. (**1988**). "Effects of speaking rate and lexical status on phonetic perception," J. Exp. Psychol. Hum. Percept. Perform. **14**, 369–378.

Miller, J. L., and Liberman, A. M. (**1979**). "Some effects of later-occurring information on the perception of stop-consonant and semivowel," Percept. Psychophys. **25**, 457–465.

Nilsson, M., Soli, S., and Sullivan, J. (**1994**). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.

Patterson, R. D., Nimmo-Smith, I., Weber, D. L., and Milroy, D. (**1982**). "The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold," J. Acoust. Soc. Am. **72**, 1788–1803.

Remez, R. E., Ferro, D. F., Wissig, S. C., and Landau, C. A. (**2008**). "Asynchrony tolerance in the perceptual organization of speech," Psychon. Bull. Rev. **15**, 861–865.

Rhode, W. S. (**1991**). "Physiological-morphological properties of the cochlear nucleus," in *Neurobiology of Hearing: The Central Auditory System*, edited by R. A. Altschuler, B. M. Clopton, B. M. Bobbin, and D. W. Hoffman (Raven, New York), pp. 47–77.

Rhode, W. S., and Greenberg, S. (**1992**). "Physiology of the cochlear nuclei," in *The Mammalian Auditory Pathway: Neurophysiology*, edited by A. N. Popper and R. R. Fay (Springer, New York), pp. 94–152.

Rhode, W. S., and Greenberg, S. (**1994**). "Encoding of amplitude modulations in the cochlear nucleus of the cat," J. Neurophysiol. **71**, 1797–1825.

Saberi, K., and Perrott, D. R. (**1999**). "Cognitive restoration of reversed speech," Nature (London) **398**, 760.

Shannon, C. E. (**1948**). "A mathematical theory of communication," Bell Syst. Tech. J. **27**, 379–423 and 623–656.

Shannon, C. E. (**1951**). "Prediction and entropy of printed English," Bell Syst. Tech. J. **30**, 50–64.

Shannon, R. V., Zeng, F., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

Silipo, R. Greenberg., S., and Arai, T. (**1999**). "Temporal constraints on speech intelligibility as deduced from exceedingly sparse spectral representations," in Proceedings of the 6th European Conference on Speech Communication and Technology, pp. 2687–2690.

Steeneken, H. J. M., and Houtgast, T. (**1980**). "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**, 318–326.

Stilp, C. E., Alexander, J. M., Kiefte, M., and Kluender, K. R. (**2010**). "Auditory color constancy: Calibration to reliable spectral properties across nonspeech context and targets," Atten. Percept. Psychophys. **72**, 470–480.

Stilp, C. E., and Kluender, K. R. (**2010**). "Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility," Proc. Natl. Acad. Sci. U.S.A. **107**, 12387–12392.