Auditory "bubbles": Efficient classification of the spectrotemporal modulations essential for speech intelligibility

Jonathan H. Venezia,[a] Gregory Hickok, and Virginia M. Richards
Department of Cognitive Sciences, University of California, Irvine, 3151 Social Science Plaza, Irvine, California 92697-5100, USA

(Received 29 October 2015; revised 7 July 2016; accepted 12 July 2016; published online 16 August 2016)

Speech intelligibility depends on the integrity of spectrotemporal patterns in the signal. The current study is concerned with the speech modulation power spectrum (MPS), which is a two-dimensional representation of energy at different combinations of temporal and spectral (i.e., spectrotemporal) modulation rates. A psychophysical procedure was developed to identify the regions of the MPS that contribute to successful reception of auditory sentences. The procedure, based on the two-dimensional image classification technique known as "bubbles" (Gosselin and Schyns (2001). Vision Res. **41**, 2261–2271), involves filtering (i.e., degrading) the speech signal by removing parts of the MPS at random, and relating filter patterns to observer performance (keywords identified) over a number of trials. The result is a classification image (CImg) or "perceptual map" that emphasizes regions of the MPS essential for speech intelligibility. This procedure was tested using normal-rate and 2×-time-compressed sentences. The results indicated: (a) CImgs could be reliably estimated in individual listeners in relatively few trials, (b) CImgs tracked changes in spectrotemporal modulation energy induced by time compression, though not completely, indicating that "perceptual maps" deviated from physical stimulus energy, and (c) the bubbles method captured variance in intelligibility not reflected in a common modulation-based intelligibility metric (spectrotemporal modulation index or STMI). © *2016 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4960544]

[MAH] Pages: 1072–1088

I. INTRODUCTION

Classically, speech perception researchers used speech synthesis to isolate the acoustic cues—discrete spectrotemporal events visible on the speech spectrogram—that underlie perception of individual speech sounds (Cooper *et al.*, 1952; Delattre *et al.*, 1955; Liberman, 1957; Heinz and Stevens, 1961; Blumstein and Stevens, 1980). Recent psychophysical advances allow fairly robust specification of such cues using natural speech tokens (Li *et al.*, 2010; Kapoor and Allen, 2012; Li *et al.*, 2012). However, a great deal of modern speech research suggests that information critical for intelligibility is transmitted via temporal and spectral modulations—global patterns of change across the temporal and spectral axes of the spectrogram (Houtgast *et al.*, 1980; Ter Keurs *et al.*, 1992; Baer and Moore, 1993; Ter Keurs *et al.*, 1993; Baer and Moore, 1994; Drullman *et al.*, 1994a,b; Shannon *et al.*, 1995; Zeng *et al.*, 2004; Henry *et al.*, 2005; Gilbert and Lorenzi, 2006; Litvak *et al.*, 2007; Jørgensen and Dau, 2011).

The present study is concerned with the modulation power spectrum (MPS) (Grace *et al.*, 2003), which considers jointly the temporal and spectral modulations that compose the speech signal. The MPS is obtained by the 2D Fourier transform of the spectrogram. As such, the MPS describes the spectrogram as a weighted sum of 2D-sinusoidal components known as drifting spectrotemporal ripples. A ripple contains energy at a unique combination of temporal (Hz) and spectral (cyc/octave or cyc/kHz) modulation rate, and therefore the MPS shows the distribution of energy across these joint spectrotemporal modulations.

The MPS provides a useful characterization of speech for several reasons. First, the MPS efficiently summarizes the subspace of modulations occupied by speech and accurately describes cases in which the joint distribution of spectral and temporal modulation energy differs from what would be expected considering the spectral and temporal envelopes individually (Singh and Theunissen, 2003; Theunissen and Elie, 2014). Second, the receptive fields of auditory-cortical neurons and neural ensembles can be described in terms of tuning to drifting spectrotemporal ripples in this space (Kowalski *et al.*, 1996; Depireux *et al.*, 2001; Shamma, 2001; Langers *et al.*, 2003; Schönwiesner and Zatorre, 2009) and it has been shown that such tuning facilitates discrimination of natural sounds including speech (Woolley *et al.*, 2005). Third, when speech is degraded, the spectrotemporal modulation profile (i.e., the MPS) is also disrupted and the extent of this disruption can be quantified and used to accurately predict reductions in intelligibility (Chi *et al.*, 1999; Elhilali *et al.*, 2003; Chi *et al.*, 2005; Zilany and Bruce, 2007). Finally, spectrotemporal modulation sensitivity in hearing-impaired listeners predicts speech intelligibility over and above traditional audibility measures (Bernstein *et al.*, 2013; Mehraei *et al.*, 2014).

Given the apparent significance of spectrotemporal modulations in transmitting intelligible speech information,

[a]Electronic mail: jvenezia@uci.edu

 0001-4966/2016/140(2)/1072/17/$30.00

one might ask which particular regions of the MPS carry the bulk of such information. In fact, selective degradation (filtering) of the spectral and temporal speech envelopes, considered in isolation, suggests that most intelligible information is conveyed by spectral modulations up to about 1.5 cycles/octave and temporal modulations up to about 16 Hz (Ter Keurs *et al.*, 1992, 1993; Drullman *et al.*, 1994a,b). Elliott and Theunissen (2009) recently developed a technique that allows selective filtering of joint spectrotemporal modulations. The technique involves nulling particular modulation components on the MPS. The filtered signal is reconstructed by moving back through the spectrographic representation to obtain a filtered speech waveform via iterative spectrogram inversion. Using MPS filtering, Elliott and Theunissen asked normal hearing listeners to comprehend sentences that were low-pass or notch-filtered in the MPS domain. The results indicated a so-called modulation transfer function (MTF) for speech intelligibility in which the crucial modulation information for speech intelligibility fell into a "core" region of the MPS comprising spectral modulations up to ~4 cyc/kHz and temporal modulations up to ~8 Hz, consistent with data based on degradation of spectral and temporal envelopes in isolation (Ter Keurs *et al.*, 1992, 1993; Drullman *et al.*, 1994a,b). The MTF within this critical region for intelligibility was low-pass (<1 cycles/kHz) in the spectral modulation domain and bandpass (1 to 7 Hz) in the temporal modulation domain.

The ability to selectively filter joint spectrotemporal modulations from the speech signal constitutes a significant advance in signal processing, and this advance has improved the characterization of intelligible speech in the MPS domain. However, the paradigm employed by Elliott and Theunissen required testing of many participants (>35) in various filter conditions to obtain the complete speech MTF. As a result, the technique was time consuming and the obtained MTF was rather coarse. Moreover, noise was added to the stimuli, which may have altered listeners' neural representations of spectrotemporal modulations (Chi *et al.*, 1999; Elhilali *et al.*, 2003; Chi *et al.*, 2005). Here, we develop a technique—based on the filtering algorithm of Elliott and Theunissen—that allows efficient classification of the spectrotemporal modulations essential for speech intelligibility *in individual listeners* and without additive noise.

The MPS, like the spectrogram, is just a 2D image in which temporal modulation rates are plotted along the *x* axis, spectral modulation rates are plotted along the *y* axis, and intensity gives the modulation power at a given [*x,y*] coordinate (see Sec. II A 2 b). For this reason, we designed a classification procedure that parallels the 2D-image classification technique from vision research known as "bubbles" (Gosselin and Schyns, 2001). In the bubbles procedure, a visual image (e.g., a face) is overlaid with an opaque mask that is "pierced" with randomly placed, transparent Gaussian apertures (bubbles). The bubbles allow participants to sample (see) different parts of the image on different trials. Participants are asked to perform a task (e.g., gender identification) based on the limited information available to the visual system. After a pre-determined number of trials, bubbles masker patterns across trials can be reverse correlated with participant behavior to derive a classification image (CImg). Essentially, if one imagines overlaying (i.e., summing) all the maskers from correctly identified trials on top of one another, crucial visual information is revealed by virtue of clear elements in the otherwise opaque "summed masker." Visual information that contributes, but may not be crucial, is revealed as a quasi-opaque element in the sum. One can then sum the maskers from incorrect trials and compare them to the correct-trial sum (i.e., take a difference) to derive a CImg. This procedure is essentially a multiple regression that uses opacity in the bubbles masker to predict participant behavior (Chauvin *et al.*, 2005).

We extend the bubbles technique to the auditory speech domain by applying a bubble mask to the MPS. Thus, a sentence's MPS is filtered by multiplication with a randomly generated bubbles masker that allows some spectrotemporal modulations to pass (clear pixels in the visual example) while other modulations are removed. Following Elliott and Theunissen (2009), a filtered spectrogram is then obtained from the inverse transformation of the MPS, and a filtered waveform is recovered by spectrogram inversion. The resulting sentence is degraded in terms of intelligibility such that the extent of degradation depends on the particular pattern of bubbles (i.e., on the particular spectrotemporal modulations preserved in the signal). In the current study, we asked participants to identify keywords from filtered sentences so that we could relate patterns of spectrotemporal modulations to keyword identification performance—i.e., so we could classify the regions of the MPS essential for transmitting intelligible speech. Our study builds on previous work using regression and/or reverse correlation techniques in auditory psychophysics and auditory neurophysiology including some recent applications in the modulation domain (Ahumada and Lovell, 1971; Huang and Richards, 2008; Kumar *et al.*, 2008; Shub and Richards, 2009; Pasley *et al.*, 2012; Santoro *et al.*, 2014; Theunissen and Elie, 2014).

The bubbles technique succeeded in producing reliable CImgs for intelligibility of speech in individual, normal-hearing listeners. Consistent with previous work, the resulting high-resolution CImgs revealed a low-pass form in the spectral modulation domain (3 dB-down cutoff = 1.5 cyc/kHz) and a band-pass form in the temporal modulation domain (peak = 3.7 Hz).[1] We validated CImgs by using individual participant classifications to accurately predict performance on independent (held-out) data from the same participant. To test whether the classification procedure would be sensitive to an expected change in behavior based on a physical change in the stimulus, in a second experiment we obtained CImgs using 2×-time-compressed sentences. The time compression procedure effectively doubled the temporal modulation rates present in the stimuli. This caused the classified region of the MPS to shift up by ~1/3 octave in the temporal modulation domain, confirming the ability of the bubbles procedure to detect changes related to time compression. However, the shift did not strictly follow the change in modulation energy induced by time compression, indicating some limitation either on participants' strategies or on the neural representation of the speech envelope. In a follow-up experiment, we demonstrated that high temporal modulation rates (>10 Hz)

influence intelligibility even for uncompressed speech, suggesting that observers are capable of using information encoded on a more rapid timescale.

II. APPLICATION OF BUBBLES IN THE MPS DOMAIN WITH NORMAL-RATE AND 2×-TIME-COMPRESSED SENTENCES

A. Methods

1. Participants

A total of 27 (16 female) participants were recruited for two main experiments and a follow-up experiment. Twenty-two of the participants were right-handed (self-report) and 20 participants were native English speakers. All participants were classified as fluent (native-like) English speakers based on conversations with the experimenters, who themselves were native English speakers. The participants were between 18 and 45 years of age (M = 21.6, SD = 5.6) and all but one had audiometric thresholds of 20 dB hearing level (HL) or better at frequencies from 250 to 4000 Hz in both ears. One participant had a threshold of 25 dB HL at 500 Hz in the right ear. For each participant we tested the better ear based on pure tone thresholds from 1000 to 4000 Hz. If both ears had equal average thresholds from 1000 to 4000 Hz, the ear with lower average thresholds across the entire audiometric range (250–8000 Hz) was tested.

Twelve participants completed an experiment using normal-rate speech stimuli and a second, unique group of 10 participants completed an experiment using time-compressed speech (see Sec. II A 2). Two participants were dropped from normal-rate group—one failed to complete the entire experiment and one failed to meet the established performance criterion (see Sec. II A 3). Thus, a total of ten participants from each group were included in the final analysis. Five additional participants participated in a follow-up experiment (see Sec. II B 5).

2. Stimuli

a. Uncompressed and 2×-time-compressed sentence stimuli. Auditory sentence stimuli were drawn from the Institute of Electrical and Electronics Engineers (IEEE) sentence corpus (IEEE, 1969). Sentences were spoken by an adult female with an American accent. For each of the 720 sentences in the corpus, the nouns, verbs, pronouns, adverbs, and adjectives were marked as keywords. The "uncompressed" speech stimulus set was composed of 452 sentences containing five keywords each. To generate time-compressed stimuli the uncompressed recordings were compressed by a factor of 2 (i.e., [1/2] duration) using the pitch-synchronous overlap and add (PSOLA) technique (Moulines and Charpentier, 1990) as implemented in PRAAT software (Boersma and Weenik, 2010). The PSOLA procedure reduces the duration of the signal while maintaining the original fundamental frequency contour and largely preserving the original spectrum (including formant spacing). Panel (A) of Fig. 1 plots a one-third octave analysis of the long term average spectrum (LTAS) of the uncompressed (UC) sentence stimuli and the LTAS of the time-compressed stimuli (2×). The LTAS for UC and 2× stimuli

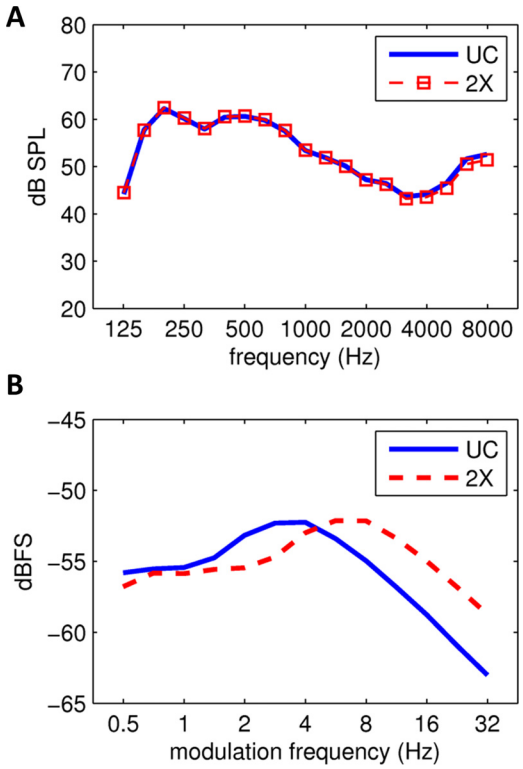

FIG. 1. (Color online) (A) One-third octave analysis of the long-term average spectra of uncompressed (UC) and time-compressed (2×) sentences. (B) One-third octave analysis of the temporal modulation spectrum of uncompressed (UC) and time-compressed (2×) sentences.

are nearly identical, confirming the success of the PSOLA procedure in preserving the spectral content of the 2× stimuli. Panel (B) of Fig. 1 plots a one-third octave analysis of the temporal modulation spectra for UC and 2× stimuli. Temporal modulation power in the 2× stimuli is translated upward in frequency by one octave because the 2× stimuli are half the duration of the UC stimuli.

b. Modulation power spectra of UC and 2× stimuli. To analyze the stimuli in terms of their joint spectrotemporal modulations, we calculated the MPS for both the UC and the 2× stimuli. The MPS is a measure of phase-invariant modulation power obtained by the two-dimensional Fourier transform (2D-FFT) of the spectrogram. By this procedure the spectrogram [Fig. 2(A)] is described as the weighted sum of 2D-sinusoidal components where each component corresponds to a unique broadband ripple sound (analogous to 2D-sinusoidal gratings in vision) characterized by amplitude modulations in time (x axis of the spectrographic representation) and frequency (y axis of the spectrographic representation). Different combinations of spectral and temporal modulation rate yield ripples of unique period and orientation [Fig. 2(B)]. Each pixel (2D-FFT "bin") in the MPS [Fig. 2(C)] reflects the modulation power of a particular ripple component (i.e., at a particular combination of spectral and temporal modulation rate). Note, the MPS has four symmetrical quadrants (upper left = lower right, upper right = lower left), so it suffices to focus on the upper quadrants. By convention, the x axis (Hz) has both positive and negative values, where negative values correspond to upward drifting

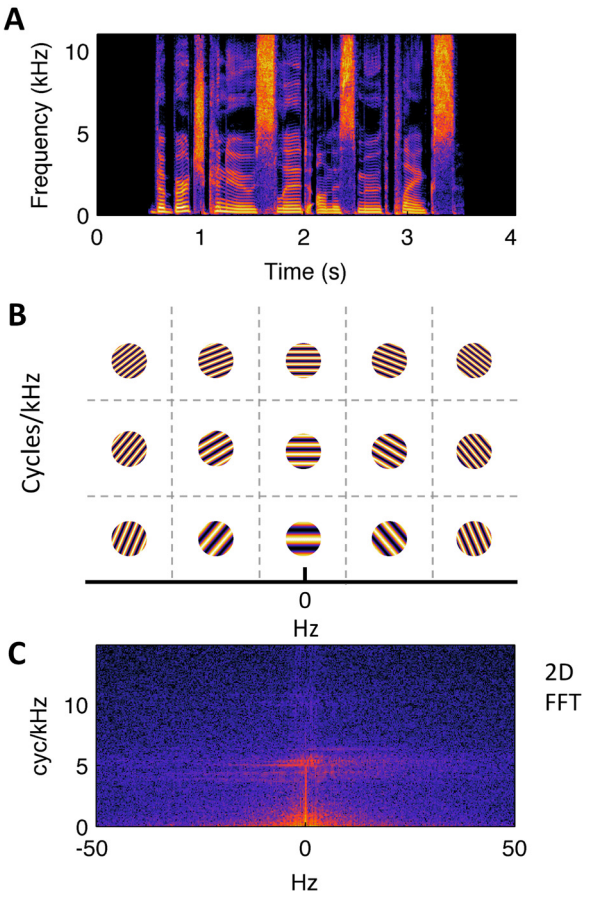

FIG. 2. (Color online) (A) Spectrogram of an example sentence ("the birch canoe slid on the smooth planks"). (B) Schematic of the 2D Fourier transform operation, which decomposes the spectrogram into spectrotemporal ripple components (light/dark gratings) at different temporal (x axis, Hz) and spectral (y axis, cyc/kHz) modulation rates. Each ripple component is a 2D sinusoid in time-frequency space (i.e., a unique combination of spectral and temporal modulation rate). (C) The sentence modulation power spectrum (outcome of the 2D Fourier transform). A pixel in this representation corresponds to a single square as depicted in (B). The color scale (dB, arbitrary ref.) indicates the relative amount of energy across joint spectrotemporal modulations (light = more energy, dark = less energy).

ripples and positive values correspond to downward drifting ripples.

The procedures outlined by Theunissen and colleagues were used to calculate the MPS (Singh and Theunissen, 2003; Elliott and Theunissen, 2009). For each sentence, a time-frequency representation of the stimulus was estimated as the log-power (dB) of a spectrogram obtained with Gaussian windows (4.75 ms-33.5 Hz time-frequency scale). The time-frequency scale of the spectrogram determines the upper bounds of the temporal and spectral modulations represented in the MPS (Singh and Theunissen, 2003; Elliott and Theunissen, 2009). Thus, we chose the parameters of the Gaussian window to ensure that most of the modulation energy in speech would be contained within the MPS boundaries (Elliott and Theunissen, 2009). The MPS was then obtained as the modulus of the 2D-FFT of the spectrogram, converted to power on a dB scale. The boundaries of the MPS at the time-frequency scale of 4.75 ms-33.5 Hz were ± 105 Hz and 14.9 cyc/kHz (Singh and Theunissen, 2003), although further processing of the MPS was restricted to ± 50 Hz on the temporal modulation axis. Stimuli were zero-padded prior to processing to ensure an MPS representation with the same dimensions for each sentence (165×429 pixels for the UC stimuli, and 165×265 pixels for the $2\times$ stimuli). Figure 3 shows the average MPS for UC [panel (A)] and $2\times$ [panel (B)] stimuli. Note that most of the modulation energy for both UC and $2\times$ is near the origin (combinations of low spectral and low temporal modulation rates). A second region containing high modulation power appears near 6 cyc/kHz on the y axis. This region corresponds to modulations around the fundamental frequency (a spectral modulation at 6 cyc/kHz has a period of 166.67 Hz along the frequency axis of the spectrogram), and so can be conceived of as the "pitch" region of the MPS. Figure 3(C) shows a difference MPS, $2\times$ minus UC. This was generated by first upsampling the $2\times$ MPS (linear interpolation) to match the

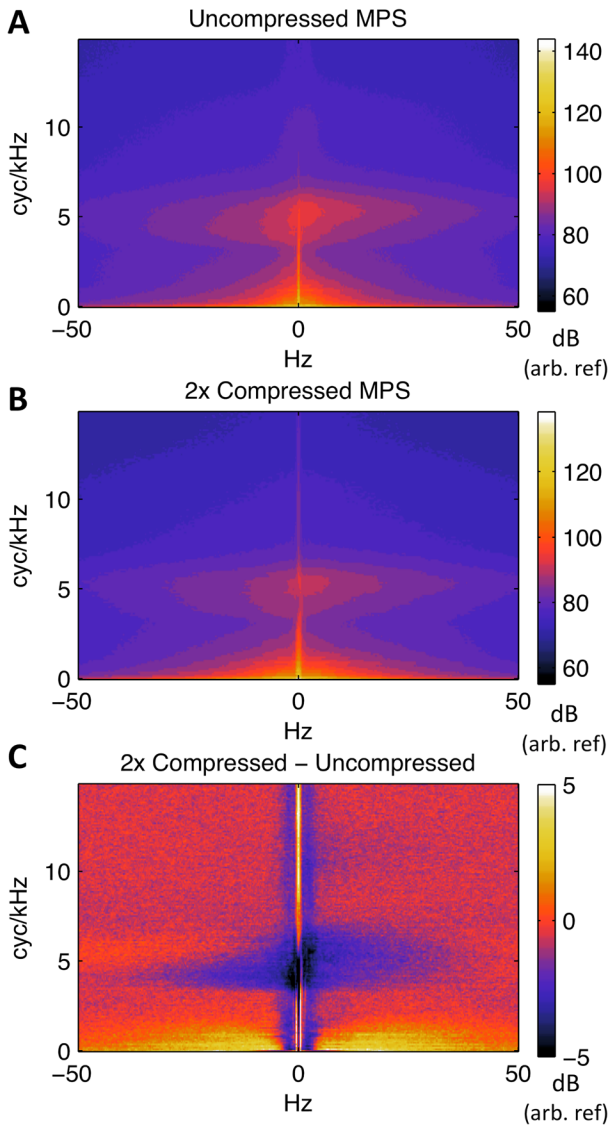

FIG. 3. (Color online) (A) Average modulation power spectrum for uncompressed sentences. (B) Average modulation power spectrum for time-compressed sentences. (C) "Difference" modulation power spectrum formed by subtracting (A) from (B) after equating for total power. Light-colored (yellow online) regions reflect relatively greater modulation energy in time-compressed speech, and dark-colored (purple online) regions reflect relatively greater modulation energy in uncompressed speech. Decibel scales are relative to an arbitrary reference.

dimensions of the UC MPS and scaling to match total power across pixels representing temporal modulation rates above 1 Hz (this was done to avoid the DC and other "noisy" pixels near the origin during scaling). There was a positive difference in energy (2× > UC; bright pixels) at temporal modulation rate above ~4 Hz, but this was restricted to spectral modulation rates below ~2 cyc/kHz. There was a negative difference in energy (UC > 2×; dark pixels) at low temporal modulation rates in the pitch region, indicating that the PSOLA procedure likely flattened some amplitude modulations of the closely spaced harmonics of the fundamental.

c. Experimental stimuli (bubbles-filtered sentences). To create "bubbles" versions of the UC and 2× stimuli, we implemented a filtering technique that removed portions of the MPS from each stimulus at randomly selected locations (Fig. 4). For each sentence, the MPS was obtained as described above with the exception that the modulation spectrum was not converted to log power (dB). Rather, subsequent operations were performed on the magnitudes of the 2D-FFT, which we will continue to refer to as the MPS for convenience. Given the MPS, a multiplicative mask (i.e., filter) with randomly placed apertures (bubbles) was applied. Specifically, a binary bubbles mask of the same dimensions as a single quadrant of the MPS (165 × 215 for UC, 165 × 133 for 2×) was created. Beginning with an all-zero image, a number of randomly selected pixels were set to value 1. The particular number of pixels set to 1 determined the number of bubbles in the filter. The image was then smoothed with a symmetric Gaussian filter with sigma = 7 pixels for UC. For 2×, an asymmetric Gaussian filter was used with sigma = 7 pixels in the spectral modulation dimension and sigma = 4.33 pixels in the temporal modulation dimension. This ensured that the sigma of the Gaussian filter in the temporal modulation dimension, expressed in Hz rather than pixels, was matched across the UC and 2× stimuli. Following smoothing, all values above 0.1 were set to 1 and all other values were set to zero, creating a binary mask with circular apertures (i.e., bubbles). Note, when the number of bubbles was large the apertures "bumped into" each other and created a variety of shapes as in Fig. 4. The binary mask was smoothed again with a Gaussian filter (sigma = 1 pixel) to prevent abrupt changes in spectrotemporal modulation energy across neighboring pixels, which could potentially lead to spurious noise (i.e., splatter) on filtered spectrograms (see below). This final mask was multiplied symmetrically to all quadrants of the MPS, removing modulation power at some locations but not at others. Application of the same mask in each quadrant ensured that modulation power at upward- and downward-drifting ripple components was filtered equally, a simplification that increased the statistical power of the data analysis [see Sec. II A 4]. The resultant filtered MPS was then converted back to a spectrogram representation (inverse 2D-FFT), and a filtered stimulus waveform was obtained by iterative spectrogram inversion (Griffin and Lim, 1984). Twenty iterations were used in the inversion procedure. To quantify the percent error introduced by spectrogram inversion, we compared the desired spectrogram (prior to inversion) with the actual spectrogram of the final filtered waveform by squaring the differences between them, dividing by the power of the desired spectrogram, and summing over time and frequency (Elliott and Theunissen, 2009). The mean error of the spectrogram inversion procedure was 4.43% (SD = 0.60) for the UC stimuli and 3.61% (SD = 0.66) for the 2× stimuli. Error tended to increase as the number of bubbles decreased (UC: mean at 20 bubbles = 5.53%, mean at 100 bubbles = 3.64%; 2×: mean at 150 bubbles = 2.75%, mean at 30 bubbles = 4.98%). The MATLAB code for modulation filtering, spectrogram inversion, and error calculation was obtained from Elliot and Theunissen (2009) who themselves modified code from Slaney (1998).

For each of the 452 sentence stimuli, filtered versions were created using random, independent bubbles masks. Further, separate sets of filtered stimuli were created using different numbers of bubbles in the masks (UC: 20–100 in steps of five; 2×: 30–150 in steps of five). All stimuli were generated and stored prior to running the experiments. The

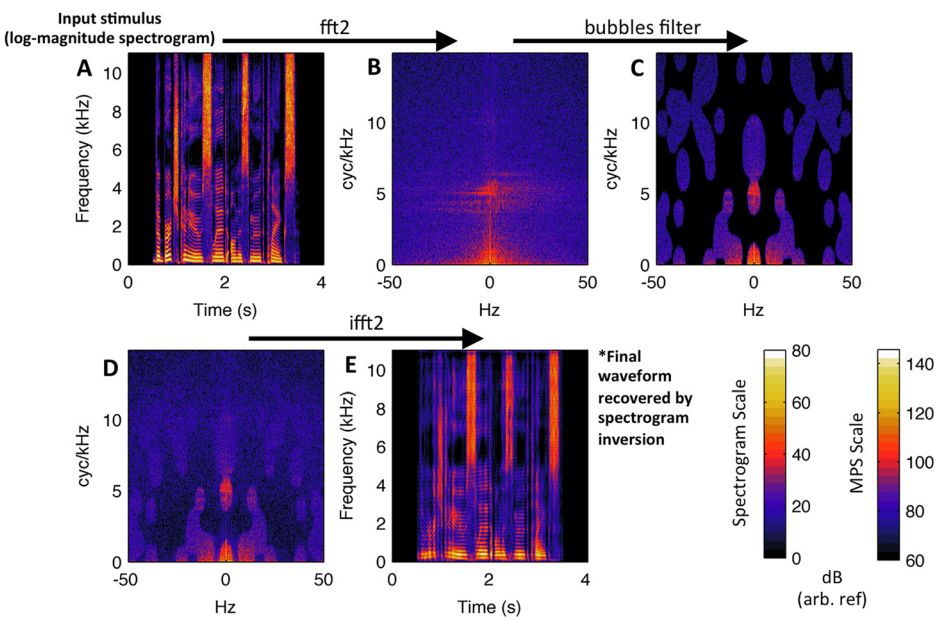

FIG. 4. (Color online) Schematic of the bubbles filtering procedure. The MPS is obtained by 2D Fourier transform of the spectrogram, and random components of the MPS are nulled according to the bubbles filter pattern (top row). A filtered speech spectrogram is obtained from the modified MPS by inverse 2D Fourier transform (bottom row). Finally, a filtered waveform is obtained by iterative spectrogram inversion (not pictured). (A) Stimulus spectrogram. (B) Unfiltered modulation power spectrum. (C) Application of bubbles. (D) Filtered modulation power spectrum. (E) Filtered spectrogram. Decibel scales are relative to an arbitrary reference.

unique bubbles mask for each stimulus was also stored for later analysis. Examples of bubbles stimuli are provided in supplementary[2] audio 1–3.

During the experiments, the filtered speech stimuli were presented at an overall level of 70 dB sound pressure level (SPL). The target speech signals were generated digitally at a sampling rate of 22 050 Hz and presented monaurally via a 24-bit soundcard (Envy24 PCI audio controller, VIA technologies, Inc.), passed through a programmable attenuator and headphone buffer (PA4 and HB6, Tucker-Davis Technologies, Inc.) and presented to the listener though either a Sennheiser HD410 SL or a Sennheiser HD600 headset.

3. Procedures

The UC and 2× experiments were conducted in two 1–1.5 h sessions for each listener. During the first session, listeners typically completed nine blocks of 25 trials of sentence recognition with the bubbles-filtered speech stimuli. During the second session, listeners typically completed an additional eight blocks of 25 trials and one block of 27 trials (18 total blocks). On each trial, the target sentence was drawn at random, without replacement, from the 452 potential sentences. Thus, no listener heard the same sentence twice and each listener heard all 452 items in the stimulus set. Data from the first block of each session were discarded (see adaptive procedure below) leaving 402 total trials from 16 blocks for subsequent analysis.

Listeners were seated in a double-walled sound-attenuated booth and were asked to follow the on-screen instructions displayed on a PC monitor. A mouse and a keyboard were provided to the listener. On each trial, after the presentation of the stimulus, the listener was asked to type the sentence s/he heard into an edit-box on the screen. After the listener confirmed her/his response, the typed text remained on the screen but could not be edited. At this point, five boxes appeared on the screen, each of which was labeled using a true keyword from the target sentence presented on that trial. The listener was instructed to select the keywords that were correctly identified in their typed response by clicking the corresponding buttons. By instruction, errors in tense or obvious typos were counted as correct. The next trial began after the self-scoring process was complete.

An up-down adaptive tracking procedure (Levitt, 1971) was implemented in which the number of bubbles applied to the filtered stimulus varied from trial to trial. When listeners correctly identified more than half of the keywords presented in the trial, the number of bubbles was decreased by 5. Otherwise, the number of bubbles was increased by 5. A new track was started at the beginning of each testing session. The particular bubbles mask associated with each trial was stored for later analysis along with the participant's response (correct or incorrect for each keyword). An increased number of bubbles generally allowed more spectrotemporal-modulation information through to the listener. Thus, the task was easier with a large number of bubbles and harder with a small number of bubbles. The tracking procedure converged on a performance level in which listeners correctly identified three or more keywords in 50% of trials.

After data collection was completed, verification of participant self-scoring was carried out. Scores were recalculated by an automated algorithm that checked stored participant response strings against the actual keywords from each trial. Any items for which the automated score disagreed from a participant's self-score were checked manually by an experimenter for minor spelling errors or errant keystrokes that "fooled" the automated algorithm. A final percent agreement (self-scores vs experimenter-verified scores) was tabulated for each participant. Percent agreement exceeded 96.5 for every participant (mean = 99.1).

Prior to testing, listeners in the 2× experiment were given two training sessions (25 trials each) with unfiltered time-compressed speech. The training was carried out in the same listening environment as the test sessions. Fifty additional, unfiltered 2× stimuli were generated by randomly selecting items from the IEEE corpus with greater or less than five keywords (i.e., from the remaining 268 items after excluding the 452 experimental items) and submitting the items to the PSOLA time-compression procedure. The 2× participants were asked to listen to each item and repeat out loud what they heard. After pressing the "enter" key on the keyboard the actual text of the sentence was displayed on the computer screen to provide the participant with feedback. Oral responses were recorded on the internal computer microphone and stored but were not analyzed. The goal of the practice session was to allow listeners to acclimate to 2× speech.

4. Analysis

a. Construction of classification images. The main purpose of the UC and 2× experiments was to create classification images (CImgs) in the MPS domain—i.e., "heat maps" showing precisely which spectrotemporal modulations contributed maximally to speech intelligibility. To create CImgs for each participant, the pattern of responses (number of keywords correct) across trials was related to the pattern of bubbles in the MPS domain across trials. This procedure is equivalent to a multiple linear regression with behavioral responses serving as the criterion variable and across-trial bubble-mask values at each pixel serving as the predictor variables (Chauvin *et al.*, 2005).

Classification images were estimated as follows. First, responses on each trial were coded so that each correctly identified keyword was assigned the value "1" and the remaining keywords were assigned the value "0." Each trial was assigned an overall score equal to the sum across all five keywords (i.e., a number ranging from 0 to 5 depending on the number of correctly identified keywords). Scores were then converted to "deviation scores" by subtracting the mean score across trials. A CImg was produced by summing the bubble-mask values (β) over all trials (t) at each pixel location (x, y), weighting each trial by its deviation score (δ)

$$CImg_{x,y} = \sum_{t=1}^{ntrials} \delta_t \beta_t \quad \text{for } x = 1, 2, \ldots, 215\,(133); \quad y = 1, 2, \ldots, 165. \tag{1}$$

The result was a map of coefficients that constituted a "raw" CImg with dimensions equal to one quadrant of the MPS. Deviation scores were chosen as weights due to the following desirable properties: they take on large positive values for trials in which many keywords are correctly identified; they take on large negative values for trials in which few or no keywords are correctly identified; and they sum exactly to zero over all trials, which yields an unbiased classification.

We created multiple raw CImgs for each participant using a hold-one-block-out procedure (Kohavi, 1995). This was done to avoid violations of independence when the data used to construct the CImgs were also used to validate the regression model (see Sec. II A 4 c). Overall individual-participant CImgs were taken to be the average of the raw, hold-one-out CImgs. These overall CImgs were smoothed with a Gaussian filter (sigma = 5 pixels) and converted to z-scores. To create a group-level CImg, the raw CImgs from each participant were summed across all participants in a group (UC, 2×), smoothed with a Gaussian filter (sigma = 5 pixels) and converted to z-scores. Null-distribution CImgs were created for each participant and both groups by repeating the analysis steps listed above 1000 times with the order of responses (δ_t) randomly shuffled across trials. Thus, participant and group-level null distributions were distributions of z-scores. Significant pixels were those for which the z-score in the true CImg fell outside the 2.5–97.5 percentiles of the null distribution (i.e., two-sided $p < 0.05$). A difference CImg was formed by subtracting the group-level CImg for UC from the group-level CImg for 2×. Prior to subtraction, the 2× CImg was upsampled (linear interpolation) to match the dimensions of the UC CImg. Both the UC and 2× CImgs were then scaled to have a maximum value of 1 and a minimum value of 0, such that the difference CImg could range from −1 to 1.

b. Spectral and temporal modulation transfer functions. While individual pixels in the CImg represent the relative contribution to intelligibility of joint spectrotemporal modulations, individual spectral and temporal MTFs can be estimated by collapsing across one or the other dimension of the CImg. We estimated the spectral modulation transfer function (SMTF) by obtaining the maximum z-score across each row (i.e., across every temporal modulation rate in Hz for a given spectral modulation rate in cyc/kHz). Similarly, we estimated the temporal modulation transfer function (TMTF) by obtaining the maximum z-score across each column (i.e., across every spectral modulation rate in cyc/kHz for a given temporal modulation rate in Hz). The SMTF and TMTF were estimated from each participant's CImg and both group-level CImgs. The shape of the SMTF was consistently low-pass and the shape of the TMTF was consistently band-pass. Accordingly, we measured for each participant and each group the 3-dB down cutoff of the SMTF (cyc/kHz), the peak of the TMTF (Hz), and the 3-dB down point at the right of the TMTF peak (Hz). The MTF parameters extracted from individual-participant CImgs were compared between groups using an independent-samples t-test with Bonferroni correction to account for multiple comparisons (corrected threshold $p < 0.0167$).

c. Model validation. As stated above, a bubbles CImg is essentially a regression model that predicts responses from the pattern of a bubbles filter on a given trial. In this case, the responses are the number of correct keywords for each sentence (0 to 5 correct). It is of interest to generate predictions for trials not used to construct the model (the CImgs), which we achieved using a hold-one-out cross-validation procedure (Kohavi, 1995). For each participant, a CImg was constructed using a subset of the data for which trials from one block were excluded. For each of the excluded trials, a decision variable was formed by multiplying the trial-specific bubbles masker and the estimated CImg, and summing the result across all pixels. High values of the decision variable indicated that the CImg "hot spot" was largely revealed by the bubbles masker on a given trial. Keep in mind that the mask itself, not the stimuli filtered by the mask, were used to form the decision variable. This process was repeated so that each of the 16 total blocks was excluded once, yielding an observed decision variable (i.e., a prediction) for all 402 trials.

Once decision variables were formed for each trial, they were transformed to have the same properties as responses, as follows. First, the true number of responses of each type (0–5 keywords correctly identified) were counted, $n_0, n_1, \ldots, n_5$. The decision variables were then sorted from low to high, and the top n_5 values were assigned a predicted response of 5, the next n_4 values a predicted response of 4, etc. The actual participant responses from each trial (number of keywords correct from 0 to 5) were sorted in the same order to facilitate comparison of predicted and true responses. Model performance was quantified using two methods, each of which provided a different way of summarizing the relationship between predicted and true responses. First, the percent agreement between predicted and true responses was estimated for each of the six possible response types. Second, Kendall's tau (τ), a measure of association between two ordinal variables that takes values from −1 to 1, was computed between the full-length predicted and true response vectors. Group-level model performance was calculated by taking the mean of each performance measure (percent agreement, τ) across all participants in a group (UC, 2×).

Chance performance was determined using permutation testing in which the hold-one-out cross-validation procedure was repeated 1000 times for each participant using a randomly shuffled order of the true responses in each iteration. Group-level performance measures were calculated for each iteration resulting in 1000 draws from a group-level null distribution. Model performance was taken to be significantly above chance when a group-level measure calculated with unshuffled responses exceeded the value in the 97.5 percentile of the group-level null distribution for that measure (two-sided $p < 0.05$).

B. Results

1. Behavioral data

On average, the UC group correctly identified more than half of keywords on 50.5% of trials (SD = 0.7%) and the 2×

group correctly identified more than half of keywords on 50.7% of trials (SD = 0.9%). The mean number of bubbles for the UC group was 53.7 (SD = 8.5), and the mean number of bubbles for the 2× group was 90.7 (SD = 11.0). These data indicate that the up-down tracking procedure generally converged on the target performance criterion. The distribution of responses (0–5 keywords correctly identified) is shown in Fig. 5 for each group. Both groups identified either zero or five keywords correctly in a large proportion of trials, with the remainder of responses distributed rather evenly across 1–4 keywords correct. The UC group on average had a larger number of trials with zero or five keywords correct.

2. Classification images

Representative CImgs (unthresholded) from four participants in the UC and 2× groups are displayed in Figs. 6 and 7, respectively. These CImgs are plotted in the space of the upper right quadrant of the MPS, with the *x* axis reflecting temporal modulation rate in Hz and the *y* axis reflecting spectral modulation rate in cyc/kHz. The right quadrant of the MPS reflects downward-sweeping spectrotemporal ripple components, while the left quadrant reflects upward-sweeping components. Because the stimuli were filtered symmetrically in the left and right quadrants, the classification results are essentially averaged over both quadrants. Large positive values in the CImg occur at pixels for which modulation energy tended to be removed on trials with poor performance (no or few keywords identified) and preserved on trials with good performance (many or all keywords identified). Thus, pixels with large positive values mark the spectrotemporal modulations that contribute most to speech intelligibility. The opposite interpretation can be given to pixels with large negative values, namely, that modulation energy at such pixels hinders intelligibility, although significant negative-valued pixels were never observed. Also, significant positive values were only observed within a restricted range of spectrotemporal modulations (<20 Hz and <3 cyc/kHz). Therefore, CImgs are plotted over a subspace of the MPS ranging from 0 to 25 Hz on the temporal modulation axis and 0–5 cyc/kHz on the spectral modulation axis.

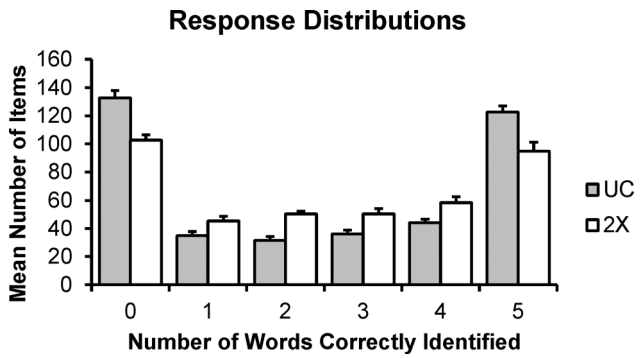

FIG. 5. Distribution of responses in the UC (grey bars) and 2× (white bars) groups. Responses are binned by number of keywords correctly identified (i.e., performance; *x* axis). Height of bars gives the average number of trials for which a given level of performance was achieved. Error bars reflect 1 SEM.

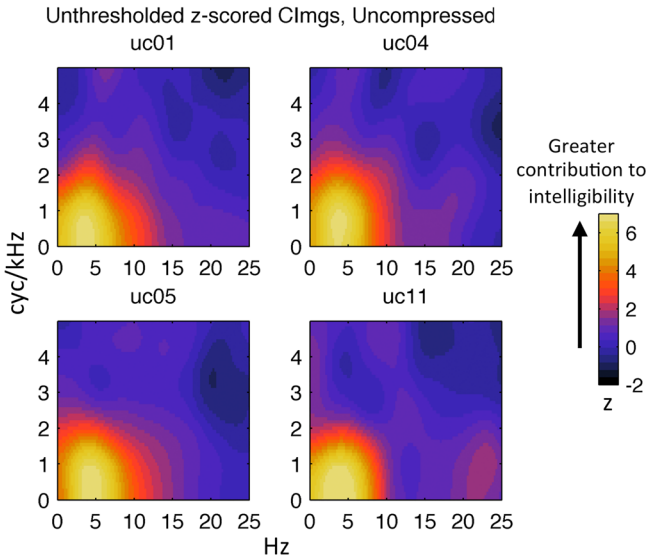

FIG. 6. (Color online) Individual participant CImgs for the UC group. Colormap reflects the normalized magnitude (z-score) of the CImg, where larger z-scores indicate a greater contribution to intelligibility. Temporal modulation rate (Hz) is plotted along the *x* axis and spectral modulation rate (cyc/kHz) is plotted along the *y* axis. Individual CImgs are labeled with participant codes.

A cursory examination of Figs. 6 and 7 reveals some general features of the measured CImgs. The CImgs consistently display a single "hot spot" near the origin corresponding to combinations of low spectral (<3 cyc/kHz) and low temporal (<20 Hz) modulation rates. This region of the MPS has been identified previously as contributing significantly to intelligibility (Elliott and Theunissen, 2009). However, the current classification technique provides a high-resolution depiction of relative contributions to intelligibility within the "hot spot," as evidenced by the consistent "bull's-eye" shape

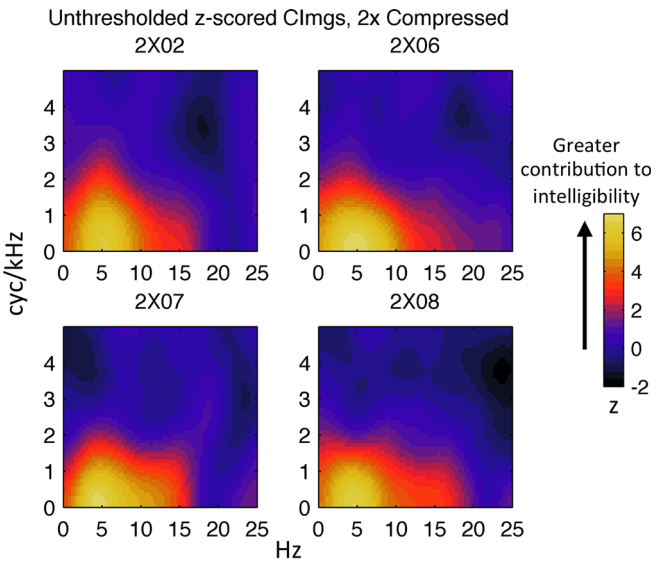

FIG. 7. (Color online) Individual participant CImgs for the 2× group. Colormap reflects the normalized magnitude (z-score) of the CImg, where larger z-scores indicate a greater contribution to intelligibility. Temporal modulation rate (Hz) is plotted along the *x* axis and spectral modulation rate (cyc/kHz) is plotted along the y-axis. Individual CImgs are labeled with participant codes.

where modulations at the center of the bull's-eye contributed relatively more to intelligibility. Additionally, it is clear that there was relatively little variation in the broad pattern of CImgs across individuals, although significant individual variation can be seen at high temporal modulation rates particularly in the 2× group (supplementary[2] Figs. 1 and 2). The group-level CImgs (Fig. 8, A/C) largely mirror the broad patterns observed in individual-participant CImgs from the respective groups.

Group (UC vs 2×) differences are apparent in the individual participant CImgs (Figs. 6 and 7) and the group-level CImgs (Fig. 8, A/C). The "hot spot" for the 2× group is reduced in the spectral modulation domain (y axis) and enlarged in the temporal modulation domain (x axis). This can be seen most clearly by examining the thresholded CImgs for each group (Fig. 8, B/D), which show the colormap only for those pixels that contributed significantly to intelligibility (permutation test, $p < 0.05$). Thresholded individual-participant CImgs also show this pattern (supplementary[2] Figs. 1 and 2). Also clearly apparent in the group-level CImgs is that the "hot spot" is low-pass for spectral modulations (center of the bull's-eye crosses zero cyc/kHz) and band-pass for temporal modulations (center of the bull's eye occurs at ~4 Hz and does not include 0 Hz). The efficacy of the bubbles technique is demonstrated in supplementary[2] audio 4–6, which give examples of UC sentences with the entire hot spot filtered out from the MPS, rendering the filtered sentences far less intelligible.

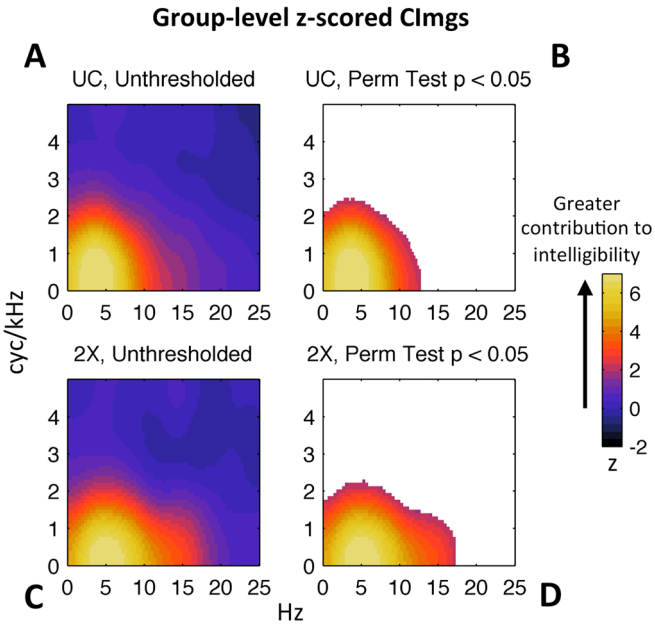

FIG. 8. (Color online) Group-level CImgs. Colormaps reflect the normalized (z-score) magnitude of the group CImg, which is obtained by summing CImgs across participants. In all CImgs, temporal modulation rate (Hz) is plotted along the x axis and spectral modulation rate (cyc/kHz) is plotted along the y-axis. (A) Unthresholded group-level CImg for uncompressed speech. (B) Thresholded group-level CImg for uncompressed speech. Pixels not exceeding threshold have not been assigned a color value (i.e., appear white). The threshold criterion was established on the basis of a null distribution formed by estimating 1000 group-level CImgs with participants' responses shuffled. (C) Unthresholded group-level CImg for time-compressed speech. (D) Thresholded group-level CImg for time-compressed speech.

3. Modulation transfer functions

Spectral and temporal MTFs were calculated by taking the maximum z-score across one dimension of a CImg (e.g., the maximum z-score at 4 Hz across all values of cyc/kHz) across all locations in the other dimension (e.g., repeated for all Hz). Group-level MTFs were formed from the group-level CImgs, and individual-participant MTFs were formed from individual-participant CImgs. The group-level MTFs are shown in Fig. 9 where the maximum height is set to 1. The TMTFs [panel (A)] are band-pass, while the SMTFs [panel (B)] are low-pass. Additionally, the TMTF for the 2× group is translated upward in frequency relative to the TMTF for the UC group (UC peak at 3.7 Hz, 2× peak at 4.9 Hz). The modulation frequencies at the peak and 3-dB down (i.e., a value of 0.707) from the peak were measured from the TMTF of each individual participant. The mean TMTF peak across participants in the UC group was 3.72 Hz (SEM = 0.15) and the mean TMTF peak across participants in the 2× group was 4.85 (SEM = 0.14). The mean 3-dB down point (i.e., temporal modulation rate at 3-dB down to the right of the TMTF peak) was 7.14 Hz (SEM = 0.15) for UC and 9.48 Hz (SEM = 0.23) for 2×. These differences were statistically reliable [peak: $t(18) = 5.7$, $p < 0.001$; 3-dB: $t(18) = 8.7$, $p < 0.001$]. Thus, the effect of time-compression was to shift CImg "hot spot" up in temporal modulation frequency by just over one third of an octave. A shift in this direction was expected given the upward shift in the modulation spectrum induced by time-compression [Fig. 1(B)], although the shift in the CImg did not perfectly follow the shift in stimulus energy (one-third octave versus a full octave).

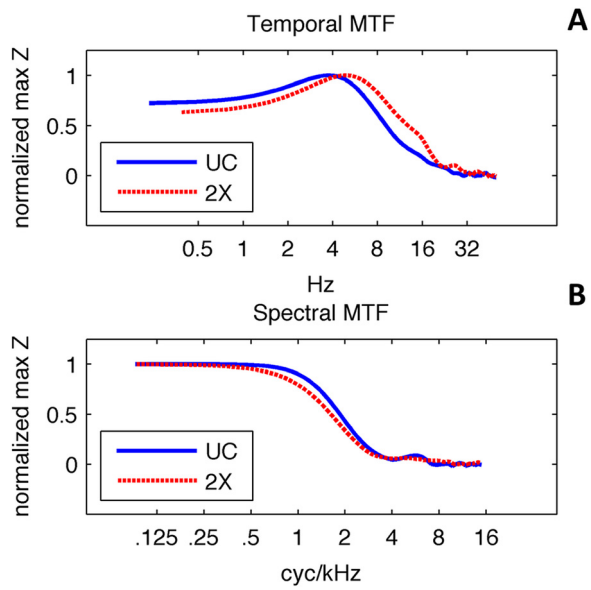

FIG. 9. (Color online) (A) Group-level temporal modulation transfer functions. Plots reflect the maximum z-score value obtained along the spectral modulation dimension of the group CImg, as a function of temporal modulation rate (Hz). Function heights have been normalized to a maximum value of 1. (B) Group-level spectral modulation transfer functions. Plots reflect the maximum z-score value obtained along the temporal modulation dimension of the group-level CImg, as a function of spectral modulation rate (cyc/kHz). Function heights have been normalized to a maximum value of 1.

The SMTF for the 2× group was also shifted relative to the SMTF for the UC group, but the shift was downward in frequency in the spectral modulation dimension [Fig. 9, panel (B)]. As with the TMTFs, SMTFs were estimated from both group-level and individual-participant CImgs. The group-level SMTF cutoff at 3-dB down was 1.45 cyc/kHz for UC and 1.11 cyc/kHz for 2×. The mean cutoff of individual-participant SMTFs for the UC group was 1.45 cyc/kHz (SEM = 0.04) and the mean cutoff for the 2× group was 1.20 cyc/kHz (SEM = 0.04), and the group difference was statistically reliable [$t(18) = 4.14$, $p < 0.001$]. Thus, a second effect of time-compression was to shift the CImg "hot spot" down one quarter of an octave in the spectral modulation dimension. This effect was not expected but is consistent with the effect of time compression on the MPS (Fig. 3). The largest relative increase in modulation power induced by time compression occurred for temporal modulation rates above 4 Hz, but only at very low spectral modulation rates (<2 cyc/kHz). This is apparent in Fig. 3(C), which shows a difference MPS (2× minus UC). To explore the relation between differences in the UC and 2× CImgs and differences in MPS energy for UC and 2× stimuli, a "difference" CImg was generated as follows. Group-level CImgs were scaled to have a maximum value of 1 and a minimum value of 0, and the UC CImg was subtracted from the 2× CImg such that the difference CImg ranged from −1 to 1. Figure 10 compares the difference MPS (panels A/B) and the difference CImg (panels C/D). The difference CImg overlaps strongly with the difference MPS, especially for low temporal and spectral modulation rates (panels B/D; note the change in axes for panels B and D relative to panels A and C), which suggests that CImgs follow stimulus energy to a considerable extent. Note also the similar boundary between bright and dark regions of the difference MPS and the difference CImg (dotted black line, panels B/D).

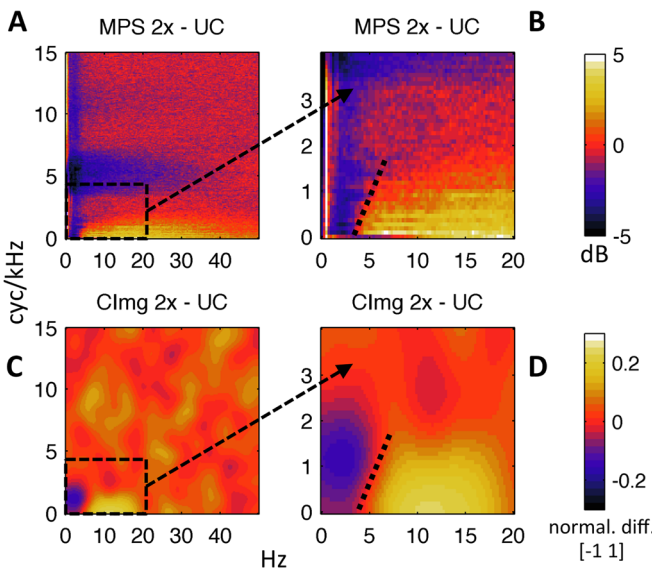

FIG. 10. (Color online) (A) "Difference" modulation power spectrum (2× − UC) reproduced from Fig. 3(C). (B) Zoomed view of (A) focusing on low temporal and spectral modulation rates. Note the scale change on the axes. (C) "Difference" CImg (2× − UC) formed by scaling z-scored group-level CImgs to the range [0 1] and subtracting the scaled UC CImg from the scaled 2× CImg. Values can take the range [−1 1]. Light pixels (yellow online) reflect a relatively greater contribution to intelligibility for 2× vs UC. Dark pixels (purple online) reflect a relatively greater contribution to intelligibility for UC vs 2×. (D) Zoomed view of (C) focusing on low temporal and spectral modulation rates. Note the scale change on the axes. Dotted lines in B/D show the boundary between light (yellow online) and dark (purple online) pixels. For all plots, temporal modulation rate is plotted along the *x* axis (Hz) and spectral modulation rate is plotted along the *y* axis (cyc/kHz).

4. Model validation

The results reported in the preceding sections suggest that (a) the bubbles technique yielded a highly reliable classification of the spectrotemporal modulations that support intelligibility, and (b) the resolution of the technique was sufficient to detect fine-grained changes in the relative importance of different modulations after time compression. Given the fact that bubbles CImgs are essentially a regression model formed under the assumption that trial-to-trial variation in bubbles masks predicts trial-to-trial variation in behavior, we can also ask how well the CImg predicts behavior. This was evaluated using a hold-one-out cross-validation procedure (see Sec. II A 4 c). Separate CImgs were constructed on subsets of the data by holding out data from one block, performing the classification analysis on the remaining data, and repeating this procedure holding out each block. In each iteration, the CImg was used to generate observations of a decision variable, which was a simple multiplication-and-sum of the CImg and the bubbles mask from each trial of the held out block. These observations formed the basis for model predictions.

The average percent agreement between true responses and responses predicted by the model, binned by type of response (0–5 keywords correct), is shown in Fig. 11. For example, the height of the bar for "five keywords correctly identified" shows, for those trials in which the participants actually identified five keywords, the proportion of predicted responses (model guesses) equal to 5. Two features of this plot are noteworthy, and they apply equally to the UC and 2× groups. First, the pattern of model performance across bins mirrors the distribution of participant responses (Fig. 4), a fact that was ensured by the prediction strategy which

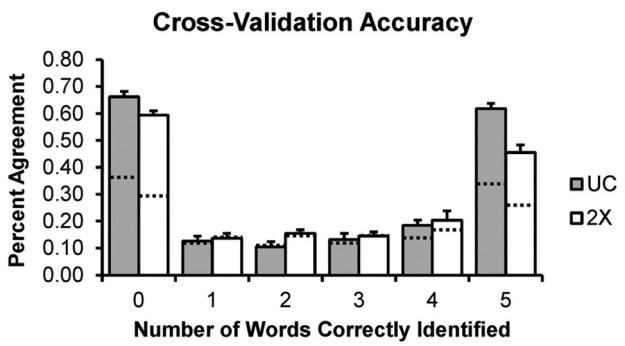

FIG. 11. Performance in hold-one-block-out cross validation of bubbles-based CImgs for UC (grey bars) and 2× (white bars) groups. Height of the bars indicates the across-participant average percent agreement between predicted and true responses. True responses are binned by number of keywords correctly identified (0–5, *x* axis) with percent agreement calculated separately for each bin. Dashed lines show chance performance determined on the basis of a null distribution formed by repeating the cross-validation procedure 1000 times with the order of true responses shuffled.

matched the proportion of model guesses to the true distribution of responses for each participant. Second, model predictions most clearly exceeded chance (dotted lines, determined by permutation testing) for 0 and 5 keywords correct, indicating that these trials were by far the most informative to the classification analysis. The next most informative trials were those with four keywords correct. The model was generally not successful at discriminating between trials with 1–3 keywords correct, indicating that significant variation in participant behavior could not be accounted for. However, although the model did not succeed in generating the correct number of keywords for trials with 1–3 keywords correct, guesses were "close" to the correct category. This is reflected in the strength of association (τ) between predicted responses and true responses, which was large for each group (UC: $M = 0.53$, $SEM = 0.01$; $2\times$: $M = 0.45$, $SEM = 0.01$) and greatly exceeded the critical values ($P < 0.05$) established by permutation testing (UC: $\tau_{crit} = 0.04$; $2\times$: $\tau_{crit} = 0.03$).

5. Follow-up experiment

The results of the UC experiment suggest that temporal modulations above ~10 Hz are not crucial for the intelligibility of speech spoken at a normal rate [Fig. 9(A)]. However, some potentially useful linguistic information is encoded at these rapid timescales (e.g., individual phonemes; Poeppel, 2003). It is possible that the bubbles procedure failed to capture the perceptual significance of temporal modulations above 10 Hz. Specifically, since the bubbles procedure is essentially a multiple regression, it will pick out components of the MPS that capture the majority of the variance in intelligibility, potentially underweighting components that contribute to intelligibility less reliably or to a lesser extent. We conducted a follow-up experiment with five additional participants to investigate this possibility. In the follow-up experiment, the bubbles procedure was repeated using new versions of the experimental UC stimuli in which spectrotemporal modulations that contributed significantly to intelligibility in the original UC experiment [Fig. 8(B)] were attenuated by a fixed amount on each trial equal to 0.15 times the initial magnitude of the MPS. This ensured that variation in the stimulus introduced by the bubbles procedure was limited to pixels outside the original hot spot. The experimental procedures were identical to those described in Sec. II A 3, and the classification analysis was repeated as described in Sec. II A 4 a. The distribution of responses was much flatter in the follow-up experiment (mean N for 0–5 keywords correct = 47.4, 58, 75.4, 78, 69, 74.2), suggesting that trials with 1–4 keywords correct were weighted more heavily than in the original UC experiment. This further suggests that variance in performance was driven by more fine-grained differences in the information conveyed by the stimuli relative to the original UC experiment. The group-level CImg for the follow-up experiment is plotted in Fig. 12 using the full axes of the MPS to allow proper visualization of high spectrotemporal modulation rates. The figure demonstrates that temporal modulation rates above 10 Hz contributed significantly to intelligibility. Specifically, the follow-up hot spot is shifted to the right of the original and centered at ~15 Hz. Thus, when the slow temporal modulation rates in the original hot-spot were prevented from accounting for variance in intelligibility, the contribution of faster temporal modulation rates was statistically revealed.

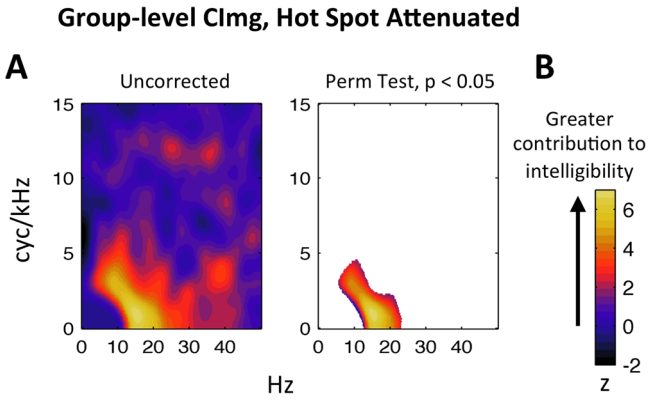

FIG. 12. (Color online) (A) Unthresholded group-level CImg for uncompressed speech with the original bubbles "hot spot" removed by attenuating the magnitude of the modulation power spectrum in that region by a factor of 0.15. This prevented the hot spot region from accounting for variance in performance. (B) Thresholded version of (A). Pixels not exceeding threshold have not been assigned a color value (i.e., appear white). The threshold criterion was established on the basis of a null distribution formed by estimating 1000 group-level CImgs with participants' responses shuffled. Colormap reflects the normalized magnitude (z-score) of the CImg, where larger z-scores indicate a greater contribution to intelligibility. Temporal modulation rate (Hz) is plotted along the *x* axis and spectral modulation rate (cyc/kHz) is plotted along the *y* axis.

III. GENERAL DISCUSSION

The primary goal of the current study was to develop an efficient procedure to classify the spectrotemporal modulations essential for speech intelligibility. Our approach involved adaptation of the "bubbles" technique from vision research (Gosselin and Schyns, 2001) to the auditory domain. We implemented the bubbles procedure by randomly filtering out portions of the MPS (Elliott and Theunissen, 2009), and we examined the effect of this filtering on the perception of uncompressed (UC) and $2\times$-time-compressed ($2\times$) sentences in normal-hearing listeners. Some of the randomly generated filter patterns impaired perception more than others, providing a means of relating trial-by-trial patterns in the bubbles filters to trial-by-trial patterns in behavior (multiple regression). The procedure determined weights on each pixel of the MPS (i.e., on particular spectrotemporal modulations), with larger weights indicating larger contributions to intelligibility. Overall, the procedure succeeded in producing robust, reliable classifications in individual participants (Figs. 6 and 7, supplementary[2] Figs. 1 and 2), and we were able to use individual-participant CImgs to predict sentence intelligibility on independent (held out) data (Fig. 11). In the UC experiment, the bubbles procedure highlighted a particular subregion or "hot spot" of the MPS consisting of low temporal (<10 Hz) and spectral (<2 cyc/kHz) modulation rates [Fig. 8(B)]. The MTF for UC speech was low-pass in the spectral modulation domain (SMTF cutoff = 1.5 cyc/kHz) and band-pass in the temporal modulation

domain (TMTF peak = 3.7 Hz). The location and shape of the MTF accorded well with previous results as described in the introduction (Ter Keurs *et al.*, 1992, 1993; Drullman *et al.*, 1994a,b; Elliott and Theunissen, 2009). However, Elliott and Theunissen (2009) found spectral modulations less than 1 cyc/kHz were most crucial for intelligibility, whereas we observed a cutoff in the SMTF at 1.5 cyc/kHz. This difference might be explained in part by smearing introduced by the bubbles procedure, which essentially applies a 2-D low-pass filter to the true MTF.

We expected the hot spot to be translated upward in the temporal modulation domain for 2× speech due to the doubling of temporal modulation rates induced by time compression [Fig. 1(B)]. In fact, the hotspot was translated upward by one-third octave in the temporal modulation domain [TMTF peak = 5.5 Hz; Fig. 8(C), 8(D), 9(A)] and downward by one-quarter octave in the spectral modulation domain [SMTF cutoff = 0.8 cyc/kHz; Fig. 8(C), 8(D), Fig. 9(B)]. The latter finding was surprising but was reconciled by comparing the MPS for UC speech to the MPS for 2× speech (Fig. 10), which demonstrated that time compression produced an increase in modulation energy only at very low spectral modulation rates. Together, these findings indicate that the results of the bubbles procedure reflected the modulation energy in the stimulus, though not completely. The implications of these findings are discussed further below.

A. Does the bubbles procedure simply track modulation energy?

The results of the bubbles procedure were found to be related to patterns of modulation energy in the stimulus. To some degree this would be expected, i.e., we would not expect the procedure to identify the upper right corner of the MPS where there is a relative absence of modulation energy. As such, it is important to ask whether the bubbles procedure adds information beyond simple examination of the MPS. Indeed, direct comparison of the bubbles CImgs for UC and 2× speech [Figs. 10(C), 10(D)] yields a remarkably similar pattern to direct comparison of the MPS for 2× speech [Figs. 10(A) and 10(B)]. However, a close examination of Fig. 10 shows that the results of the bubbles procedure clearly deviated from the patterns predicted by differences in stimulus energy, particularly at high temporal (>20 Hz) and spectral (>3 cyc/kHz) modulation rates. For instance, the "pitch region" of the MPS, which occurs at combinations of high spectral and low temporal modulation rates (see Sec. II A 2 b), contained significant modulation energy that varied in power between UC and 2× speech (Figs. 3 and 10), yet this region was not identified by the bubbles procedure for either UC or 2× [Figs. 8, 10(C) and 10(D)]. Moreover, the band-pass TMTFs observed for both UC and 2× (Fig. 9) did not obey the 1/f trend visible along the temporal modulation axis of the MPS [Figs. 3(A) and 3(B), *x* axis].

One possibility is that the bubbles procedure tracked modulation energy in a transformed representational space. Specifically, it has been proposed that acoustic signals are modified in the ascending auditory pathway prior to extraction and representation of modulation energy at higher levels (Bacon and Grantham, 1989; Dau *et al.*, 1997; Chi *et al.*, 1999; Joris *et al.*, 2004; McDermott and Simoncelli, 2011). In the context of the current study, this implies that the MPS, which is not faithful to peripheral auditory processing, was not an appropriate representation of modulation energy. This may explain why bubbles CImgs deviated from patterns of modulation energy observed on the MPS. To explore this possibility, we determined the predicted cortical response to bubbles stimuli using a computational model of the auditory system (Chi *et al.*, 1999; Shamma, 2001; Chi *et al.*, 2005). The model, implemented via a freely available MATLAB toolbox (Neural Systems Laboratory, 2001), proceeds through several peripheral stages including cochlear filtering (constant Q filters, 24 per octave, five-octave range), a hair cell stage, and a lateral inhibitory network. Peripheral processing yields an "auditory spectrogram" that is subsequently processed through a cortical stage in which the spectrogram is decomposed by "cortical neurons" with spectrotemporal receptive fields tuned to particular ripple patterns (i.e., by a spectrotemporal modulation filterbank). The spectrotemporal decomposition is performed within peripheral frequency channels and across time, yielding a complex-valued 4-D representation with the following dimensions: time (s), channel (Hz), temporal modulation rate (Hz), and spectral modulation rate (cyc/oct).

For each participant in our study, we obtained the predicted cortical representation (corticogram) of the 402 bubbles sentences heard by that participant, along with a corticogram of the unprocessed version of each sentence. For each corticogram, the magnitude was averaged across time and channel to obtain a real-valued 2-D representation (11 × 22 pixels). This 2-D representation showed "neural" energy across a range of temporal (1–32 Hz, half-octave steps, positive and negative axes) and spectral (0.25–8 cyc/oct, half-octave steps) modulation rates similar to those represented in the MPS. The goal was to evaluate, for each item, the effect of bubbles filtering on the corticogram. This was achieved by directly comparing the 2-D corticogram of each individual bubbles sentence to a clean speech template, which was the average 2-D corticogram of the unprocessed sentences. The difference between these corticograms was quantified via a modified version of the spectrotemporal modulation index (STMI; Elhilali *et al.*, 2003; Grant *et al.*, 2008)

$$\mathrm{STMI} = 1 - \frac{\|T(N-T)\|}{\|T^2\|}, \tag{2}$$

where T is the clean speech template, N is the corticogram of the bubbles stimulus, and $\|\ \|$ is the 2-norm. The STMI takes values from 0 to 1 with values closer to 0 indicating a greater disruption of the "neural" modulation energy pattern.

If the bubbles procedure tracks neural modulation energy rather than stimulus modulation energy, then values of the STMI should accurately predict the effects of bubbles filtering on intelligibility. In fact, when we repeated the model validation procedure described in Sec. II A 4 c using the STMI as the decision variable,[3] agreement was as good as or better than the agreement achieved by a decision variable derived from bubbles CImgs (Table I). Figure 13 plots

TABLE I. Comparison of model validation performance using the bubbles-derived decision variable versus the STMI. Percent agreement between predicted and true responses displayed at left, binned by true number of keywords correct (as in Fig. 11). Kendall's tau computed between predicted and true responses displayed at right. Table entries reflect the group mean with standard error of the mean in parentheses below. UC = uncompressed speech group, 2× = time-compressed speech group, ORIG = model performance determined using the bubbles-based decision variable to generate predicted responses, STMI = model performance determined using the STMI to generate predicted responses.

	Percent agreement						
	Number of keywords correct						
	0	1	2	3	4	5	Kendall's τ
UC-ORIG	66.2 (2.0)	12.5 (1.9)	10.4 (1.9)	13.2 (2.2)	18.3 (2.0)	61.8 (2.0)	0.53 (0.01)
UC-STMI	75.6 (2.2)	15.6 (2.9)	15.3 (2.6)	19.1 (2.3)	18.7 (1.8)	65.3 (2.8)	0.63 (0.02)
2×-ORIG	59.4 (1.6)	13.6 (1.8)	15.4 (1.3)	14.4 (1.5)	20.3 (3.4)	45.6 (2.7)	0.45 (0.01)
2×-STMI	67.0 (1.6)	18.5 (1.9)	14.7 (1.0)	14.3 (1.9)	20.1 (2.6)	44.3 (2.4)	0.48 (0.02)

values of the STMI (right axis, dashed) binned by number of keywords correct and averaged across items and participants for the UC [panel (A)] and 2× [panel (B)] groups. The same relation is plotted for the bubbles-based decision variable (left axis, solid). On average, the STMI strongly overlaps the bubbles-based decision variable. This appears to support the conclusion that the bubbles procedure tracked the neural representation of modulation energy.

To further verify the relationship between the STMI and the bubbles procedure, we re-ran the bubbles classification analysis (Sec. II A 4 a) using responses predicted by the STMI (0–5 keywords correct) as the dependent variable. Surprisingly, CImgs generated on this basis deviated significantly from the original CImgs (supplementary[2] Fig. 3). The

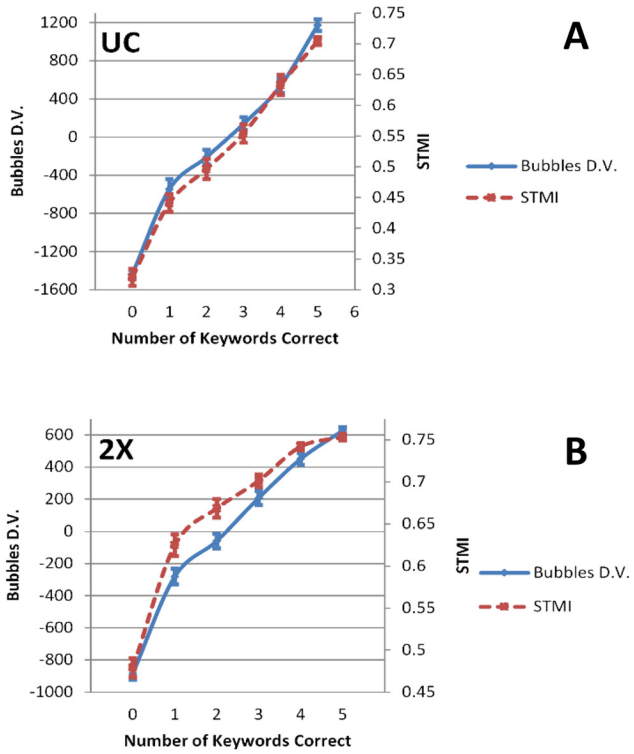

FIG. 13. (Color online) (A) Mean values of the bubbles-based decision variable (solid, left axis) and STMI (dashed, right axis) for the UC group. Values are binned by number of keywords correctly identified (i.e., performance, x axis). (B) Plot organized as in (A), but for the 2× group. For all plots, error bars reflect ±1 SEM.

STMI-based CImg "hot spots" were weighted toward the origin, and there was less evidence of a rightward shift for the 2× CImg (i.e., the STMI was less sensitive to the doubling of modulation rates induced by time compression). Thus, while the STMI and bubbles-based decision variable were similar on average—with lower average values of each measure for low intelligibility items and higher average values of each measure for high intelligibility items—the two measures differed in important ways. The average correlation between the measures was 0.66 ± 0.01 SEM for UC and 0.61 ± 0.02 SEM for 2× (all $p < 0.001$), which indicates that although the STMI and bubbles-based decision variable tended to be high on average for high intelligibility items, the items producing the highest STMI values were not necessarily the same items that produced the highest values of the bubbles-based decision variable. In short, the STMI tended to be largest when spectrotemporal modulations near the origin of the MPS were preserved and smallest when these modulations were filtered out. On the other hand, the bubbles procedure primarily emphasized modulations around 4–6 Hz (roughly the syllable rate; Arai and Greenberg, 1997). This suggests the STMI is weighted more toward stimulus *energy*, while the bubbles procedure tracks linguistically relevant *information* in the stimulus. The ability of both measures to predict intelligibility suggests that intelligibility depends on the integrity of modulations carrying both stimulus energy and linguistically relevant information.

B. Perception fails to fully track the changes induced by time compression

One of the key findings of the current study was that, while time compression induced a doubling of the modulation rates present in the stimulus, the hot spot of the 2× CImg was shifted upward by only one-third octave in the temporal modulation domain relative to the UC CImg. Thus, if the syllable rate in UC speech was ~4 Hz, then information at the syllable rate in 2× speech (~8 Hz) contributed relatively less to intelligibility or was represented less efficiently. Perhaps this accounts for the reduction in performance observed in the current study for 2× speech relative to UC speech. Specifically, although 2× speech is typically highly intelligible when presented in favorable listening conditions (Versfeld and Dreschler, 2002), in the current study

performance was generally worse for the 2× group as evidenced by the fact that less degradation of the signal was required to drive performance to the 50% correct threshold for 2× relative to UC (Sec. II B 1; confirmed by the STMI analysis in Sec. III A). Despite tremendous redundancy in the signal, evidenced by the large degree of overlap between UC and 2× CImgs (Fig. 8; see also Sec. III C), linguistically relevant information in a narrow range of spectrotemporal modulations appears to be crucial for intelligibility under challenging listening conditions. In order to optimize intelligibility, this range—characterized in the temporal modulation domain by modulations around the peak of the TMTF—should have doubled for 2× speech relative to UC speech. Why was this not case?

One possibility is that participants failed to fully adapt to 2× speech. This possibility seems somewhat unlikely given previous research on the effects of perceptual learning with time-compressed speech. Specifically, previous research suggests that perceptual learning reaches an asymptote after exposure to approximately 20 time-compressed sentences (Dupoux and Green, 1997; Peelle and Wingfield, 2005; Adank and Janse, 2009). Participants in the 2× group of the current study received 50 trials of exposure to clear (unfiltered) 2× speech prior to starting the experiment. However, this does not completely discount a role for perceptual learning or the lack thereof. All users of spoken language have necessarily been exposed to speech across a range of speaking rates, but centered on normal conversational rates. Thus, we are expertly trained to listen for information at the temporal modulation rates characteristic of normal-rate speech. It may take a tremendous amount of exposure to time-compressed speech, far beyond what is reasonable for a laboratory study, to offset the effects of this natural training. Although the auditory system is flexible enough to perceive 2× speech accurately, it may not be flexible enough to optimally allocate attention to the most informative modulations in the 2× signal. Alternatively, one could say, the auditory system is optimally tuned to the most informative modulations in normal-rate speech, which is far more likely to be encountered in the natural world.

Another possibility is that there are neural limitations on the encoding of fast temporal modulation rates in speech. A landmark study by Ahissar *et al.* (2001) showed that intelligibility of time-compressed speech is correlated with the degree of phase-locking to the speech envelope in neural ensembles of the auditory cortex. Reductions in phase-locking were observed exactly when the signal was compressed beyond 2×. Subsequent work has emphasized the role of endogenous cortical rhythms, especially in the theta (~4–8 Hz) range, in tracking the speech envelope, including further demonstration of the link between such tracking and intelligibility (Giraud *et al.*, 2007; Luo and Poeppel, 2007; Ghitza and Greenberg, 2009; Giraud and Poeppel, 2012; Peelle and Davis, 2012; Peelle *et al.*, 2013). Failure to observe a full octave shift of the bubbles hot spot for 2× speech in the current study may be due to the fact that the envelope of 2× speech could not be reliably tracked by listeners' auditory cortical neurons. In contrast to this conclusion, data from ECoG recordings in human participants demonstrate that envelopes of time-compressed speech can be tracked in the high gamma activity (70–250 Hz) of neuronal populations in core fields of the auditory cortex, even at very high compression rates (e.g., 5×) that render speech completely unintelligible (Nourski *et al.*, 2009). Moreover, high temporal modulation rates can be effectively encoded in the firing *rate* of auditory cortical neurons in the absence of phase locking (Joris *et al.*, 2004; Bendor and Wang, 2008; Pasley *et al.*, 2012). Thus, failure to observe a full octave shift in the 2× experiment cannot be attributed to a complete failure to represent high temporal modulation rates in the auditory cortex. Furthermore, these high temporal modulation rates were shown to contribute significantly to intelligibility in our follow-up experiment (Sec. II B 5), suggesting that such rates were not only represented at a high level of the nervous system (e.g., in cortex) but were also functionally relevant. Nonetheless, it remains possible that theta-rate neural oscillations contribute relatively more to intelligibility than other mechanisms for encoding speech (Giraud *et al.*, 2007; Giraud and Poeppel, 2012). Such relative weighting of encoding mechanisms could be reflected in bubbles CImgs, potentially "locking" CImg hot spots to temporal modulation rates around 4–8 Hz.

C. Bubbles CImgs do not fully characterize intelligible speech

A final important point of discussion concerns the extent to which the bubbles classification procedure can be used to characterize intelligible speech. The model validation results (Sec. II B 4) demonstrate straightforwardly that bubbles CImgs do not fully capture the variation in intelligibility present in the data. Namely, model predicted responses were only partially correlated with true responses ($\tau = 0.53$ for UC and $\tau = 0.45$ for 2×). This means that remaining variation in intelligibility could be captured by one of two factors: (1) the contribution of spectrotemporal modulations outside the bubbles hot spot or (2) the contribution of other cues such as temporal fine structure (Lorenzi *et al.*, 2006; Moore, 2008; Sheft *et al.*, 2008; Hopkins and Moore, 2009; Shamma and Lorenzi, 2013). We will only address (1) at further length.

The bubbles procedure may be insensitive to certain infrequent spectrotemporal features of speech. The bubbles procedure is essentially a multiple regression designed to identify MPS pixels that predict intelligibility. For a pixel to be assigned a large regression weight it must be reliably informative—that is, the information conveyed by that pixel must contribute to intelligibility, and the information must be reliably present across many sentences. Pixels that only occasionally convey useful information are likely to be underweighted. In the current study we observed that CImgs reliably identified low spectral and temporal modulation rates corresponding roughly to sentence components containing vowel formants (see also Elliott and Theunissen, 2009). More finely resolved spectral components such as those composing the pitch contour also contribute to intelligibility (van Santen *et al.*, 2008), but this contribution may be more pronounced

for some sentences than others, which would effectively reduce the weight on the pitch region.

It is important to remember, then, that MPS pixels shown to be maximally informative according to the bubbles procedure do not uniquely convey intelligible speech. This can be confirmed by a simple demo. If the most informative pixels [the maximal region of the bubbles CImg visible as the brightest pixels in Fig. 8(B)] are completely filtered out from the signal, speech remains intelligible (supplementary[2] audio 7–9). We may have expected that removal of maximally informative pixels would have some effect on intelligibility, but the demo clearly demonstrates this is not the case. Rather, there is considerable redundancy in the encoding of intelligible information such that spectrotemporal modulations linked relatively less strongly to performance in the current study [less-bright/orange pixels in Fig. 8(B)]) sufficiently encode a fully intelligible signal. As such, we must take caution in remembering that the bubbles procedure emphasizes the *most* informative regions of the MPS given the set parameters of the experiment. This point was driven home by our follow-up experiment, which showed that regions of the MPS that failed to reach significance in the original UC experiment emerged as significant when the original hot spot was severely attenuated on each trial (i.e., when the spectrotemporal modulations previously labeled as most informative were prevented from contributing to variation in intelligibility).

IV. SUMMARY

We adapted the image classification procedure known as "bubbles" to the auditory domain in order to efficiently classify the spectrotemporal modulations that convey intelligible speech information. This was achieved by randomly filtering out components of the MPS and relating variation in the filter patterns to variation in performance (keyword identification). The procedure identified a particular region of the MPS that contributed significantly to intelligibility. The region was confined to low spectral (<2 cyc/kHz) and temporal (<10 Hz) modulation rates, demonstrating a low-pass shape in the spectral modulation domain and a band-pass shape in the temporal modulation domain. We validated this result by using individual-participant classification images to predict performance on independent datasets. We also found that classification results were highly reliable across participants. Critically, the reliability of the bubbles procedure suggests that a robust classification can be achieved in far fewer trials than the number used here. A rapid version of the bubble procedure would be ideal to detect individual differences in populations where such differences are more likely to be observed, such as in hearing-impaired populations. When the procedure was repeated using 2×-time-compressed speech, the classified region shifted upward by only one-third octave in the temporal modulation domain. This result was shown to have significant implications in terms of understanding the perceptual and/or neural limitations that can ultimately lead to a breakdown in information processing for time-compressed speech. Intelligibility was also shown to depend substantially on both the spectrotemporal modulations that carry significant modulation energy and the spectrotemporal modulations that carry linguistically relevant information. Finally, we showed, as expected, that the bubbles procedure is inherently limited in that the classification results were tied to the stimulus parameters and task associated with the current experimental design.

ACKNOWLEDGMENTS

This work was supported by the National Institute on Deafness and Other Communication Disorders Award R21 DC013406 to V.M.R. and Y.S. During the investigation, J.H.V. was supported by the UC Irvine Center for Hearing Research via National Institute on Deafness and Other Communication Disorders Award No. T32 DC010775. Additionally, the authors thank Steven Thurman for suggesting the follow-up experiment reported in Sec. II B 5.

[1]It should be noted from the outset that bubbles CImgs provide an estimate of the underlying MTF for speech intelligibility. The boundaries of the estimated MTF should serve as an "upper limit" for boundaries of the true MTF. Specifically, bubbles act as a low-pass filter that smears the underlying MTF in proportion to the size of the bubbles. We chose a bubble size (see Sec. II A) that struck a balance between minimizing smearing and minimizing acoustic distortion to the stimulus, which increases with decreasing bubble size. All of the reported results should be interpreted with this in mind.

[2]See supplementary material at http://dx.doi.org/10.1121/1.4960544.

[3]Note the "hold-one-out" component of the analysis was no longer necessary because no behavioral data were used to derive the STMI.

Adank, P., and Janse, E. (**2009**). "Perceptual learning of time-compressed and natural fast speech," J. Acoust. Soc. Am. **126**, 2649–2659.

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., and Merzenich, M. M. (**2001**). "Speech comprehension is correlated with temporal response patterns recorded from auditory cortex," Proc. Natl. Acad. Sci. U.S.A. **98**, 13367–13372.

Ahumada, A., and Lovell, J. (**1971**). "Stimulus features in signal detection," J. Acoust. Soc. Am. **49**, 1751–1756.

Arai, T., and Greenberg, S. (**1997**). "The temporal properties of spoken Japanese are similar to those of English," in *EUROSPEECH*.

Bacon, S. P., and Grantham, D. W. (**1989**). "Modulation masking: Effects of modulation frequency, depth, and phase," J. Acoust. Soc. Am. **85**, 2575–2580.

Baer, T., and Moore, B. C. (**1993**). "Effects of spectral smearing on the intelligibility of sentences in noise," J. Acoust. Soc. Am. **94**, 1229–1241.

Baer, T., and Moore, B. C. (**1994**). "Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech," J. Acoust. Soc. Am. **95**, 2277–2280.

Bendor, D., and Wang, X. (**2008**). "Neural response properties of primary, rostral, and rostrotemporal core fields in the auditory cortex of marmoset monkeys," J. Neurophys. **100**, 888–906.

Bernstein, J. G., Mehraei, G., Shamma, S., Gallun, F. J., Theodoroff, S. M., and Leek, M. R. (**2013**). "Spectrotemporal modulation sensitivity as a predictor of speech intelligibility for hearing-impaired listeners," J. Am. Acad. Audiol. **24**, 293–306.

Blumstein, S. E., and Stevens, K. N. (**1980**). "Perceptual invariance and onset spectra for stop consonants in different vowel environments," J. Acoust. Soc. Am. **67**, 648–662.

Boersma, P., and Weenink, D. (**2010**). PRAAT, http://www.fon.hum.uva.nl/praat/ (Last visited 08/08/2016).

Chauvin, A., Worsley, K. J., Schyns, P. G., Arguin, M., and Gosselin, F. (**2005**). "Accurate statistical tests for smooth classification images," J. Vision **5**, 659–667.

Chi, T., Gao, Y., Guyton, M. C., Ru, P., and Shamma, S. (**1999**). "Spectro-temporal modulation transfer functions and speech intelligibility," J. Acoust. Soc. Am. **106**, 2719–2732.

Chi, T., Ru, P., and Shamma, S. A. (**2005**). "Multiresolution spectrotemporal analysis of complex sounds," J. Acoust. Soc. Am. **118**, 887–906.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., and Gerstman, L. J. (**1952**). "Some experiments on the perception of synthetic speech sounds," J. Acoust. Soc. Am. **24**, 597–606.
Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997**). "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," J. Acoust. Soc. Am. **102**, 2892–2905.
Delattre, P. C., Liberman, A. M., and Cooper, F. S. (**1955**). "Acoustic loci and transitional cues for consonants," J. Acoust. Soc. Am. **27**, 769–773.
Depireux, D. A., Simon, J. Z., Klein, D. J., and Shamma, S. A. (**2001**). "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," J. Neurophysiol. **85**, 1220–1234.
Drullman, R., Festen, J. M., and Plomp, R. (**1994a**). "Effect of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Am. **95**, 2670–2680.
Drullman, R., Festen, J. M., and Plomp, R. (**1994b**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**, 1053–1064.
Dupoux, E., and Green, K. (**1997**). "Perceptual adjustment to highly compressed speech: Effects of talker and rate changes," J. Exp. Psychol. Human Percept. Perform. **23**, 914–927.
Elhilali, M., Chi, T., and Shamma, S. A. (**2003**). "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech Commun. **41**, 331–348.
Elliott, T. M., and Theunissen, F. E. (**2009**). "The modulation transfer function for speech intelligibility," PLoS Comput. Biol. **5**, e1000302.
Ghitza, O., and Greenberg, S. (**2009**). "On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence," Phonetica **66**, 113–126.
Gilbert, G., and Lorenzi, C. (**2006**). "The ability of listeners to use recovered envelope cues from speech fine structure," J. Acoust. Soc. Am. **119**, 2438–2444.
Giraud, A.-L., Kleinschmidt, A., Poeppel, D., Lund, T. E., Frackowiak, R. S., and Laufs, H. (**2007**). "Endogenous cortical rhythms determine cerebral specialization for speech perception and production," Neuron **56**, 1127–1134.
Giraud, A. L., and Poeppel, D. (**2012**). "Cortical oscillations and speech processing: Emerging computational principles and operations," Nat. Neurosci. **15**, 511–517.
Gosselin, F., and Schyns, P. G. (**2001**). "Bubbles: A technique to reveal the use of information in recognition tasks," Vision Res. **41**, 2261–2271.
Grace, J. A., Amin, N., Singh, N. C., and Theunissen, F. E. (**2003**). "Selectivity for conspecific song in the zebra finch auditory forebrain," J. Neurophysiol. **89**, 472–487.
Grant, K. W., Elhilali, M., Shamma, S. A., Walden, B. E., Surr, R. K., Cord, M. T., and Summers, V. (**2008**). "An objective measure for selecting microphone modes in OMNI/DIR hearing aid circuits," Ear Hearing **29**, 199–213.
Griffin, D. W., and Lim, J. S. (**1984**). "Signal estimation from modified short-time Fourier transform," IEEE Trans. Acoust. Speech Sign. Process. **32**, 236–243.
Heinz, J. M., and Stevens, K. N. (**1961**). "On the properties of voiceless fricative consonants," J. Acoust. Soc. Am. **33**, 589–596.
Henry, B. A., Turner, C. W., and Behrens, A. (**2005**). "Spectral peak resolution and speech recognition in quiet: Normal hearing, hearing impaired, and cochlear implant listeners," J. Acoust. Soc. Am. **118**, 1111–1121.
Hopkins, K., and Moore, B. C. (**2009**). "The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise," J. Acoust. Soc. Am. **125**, 442–446.
Houtgast, T., Steeneken, H., and Plomp, R. (**1980**). "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics," Acta Acust. Acust. **46**, 60–72.
Huang, R., and Richards, V. M. (**2008**). "Estimates of internal templates for the detection of sequential tonal patterns," J. Acoust. Soc. Am. **124**, 3831–3840.
IEEE Subcommittee on Subjective Measurements IEEE Recommended Practices for Speech Quality Measurements. (**1969**). *IEEE Transactions on Audio and Electroacoustics*, Vol. 17, pp. 227–246.
Jørgensen, S., and Dau, T. (**2011**). "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," J. Acoust. Soc. Am. **130**, 1475–1487.
Joris, P., Schreiner, C., and Rees, A. (**2004**). "Neural processing of amplitude-modulated sounds," Physiol. Rev. **84**, 541–577.
Kapoor, A., and Allen, J. B. (**2012**). "Perceptual effects of plosive feature modification," J. Acoust. Soc. Am. **131**, 478–491.
Kohavi, R. (**1995**). "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, pp. 1137–1145.
Kowalski, N., Depireux, D. A., and Shamma, S. A. (**1996**). "Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra," J. Neurophysiol. **76**, 3503–3523.
Kumar, S., Forster, H. M., Bailey, P., and Griffiths, T. D. (**2008**). "Mapping unpleasantness of sounds to their auditory representation," J. Acoust. Soc. Am. **124**, 3810–3817.
Langers, D. R., Backes, W. H., and van Dijk, P. (**2003**). "Spectrotemporal features of the auditory cortex: The activation in response to dynamic ripples," NeuroImage **20**, 265–275.
Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.
Li, F., Menon, A., and Allen, J. B. (**2010**). "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," J. Acoust. Soc. Am. **127**, 2599–2610.
Li, F., Trevino, A., Menon, A., and Allen, J. B. (**2012**). "A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise," J. Acoust. Soc. Am. **132**, 2663–2675.
Liberman, A. M. (**1957**). "Some results of research on speech perception," J. Acoust. Soc. Am. **29**, 117–123.
Litvak, L. M., Spahr, A. J., Saoji, A. A., and Fridman, G. Y. (**2007**). "Relationship between perception of spectral ripple and speech recognition in cochlear implant and vocoder listeners," J. Acoust. Soc. Am. **122**, 982–991.
Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. (**2006**). "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," Proc. Natl. Acad. Sci. U.S.A. **103**, 18866–18869.
Luo, H., and Poeppel, D. (**2007**). "Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex," Neuron **54**, 1001–1010.
McDermott, J. H., and Simoncelli, E. P. (**2011**). "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis," Neuron **71**, 926–940.
Mehraei, G., Gallun, F., Leek, M. R., and Bernstein, J. G. (**2014**). "Spectrotemporal modulation sensitivity for hearing-impaired listeners," J. Acoust. Soc. Am. **136**, 301–316.
Moore, B. C. (**2008**). "The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people," J. Assoc. Res. Otolaryngol. **9**, 399–406.
Moulines, E., and Charpentier, F. (**1990**). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun. **9**, 453–467.
Neural Systems Laboratory (2001), http://www.isr.umd.edu/Labs/NSL/Downloads.html (Last visited 08/08/2016).
Nourski, K. V., Reale, R. A., Oya, H., Kawasaki, H., Kovach, C. K., Chen, H., Howard, M. A., and Brugge, J. F. (**2009**). "Temporal envelope of time-compressed speech represented in the human auditory cortex," J. Neurosci. **29**, 15564–15574.
Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T., and Chang, E. F. (**2012**). "Reconstructing speech from human auditory cortex," PLoS-Biol. **10**, 175.
Peelle, J. E., and Davis, M. H. (**2012**). "Neural oscillations carry speech rhythm through to comprehension," Front. Psychol. **3**, 1–17.
Peelle, J. E., Gross, J., and Davis, M. H. (**2013**). "Phase-locked responses to speech in human auditory cortex are enhanced during comprehension," Cerebral Cortex **23**, 1378–1387.
Peelle, J. E., and Wingfield, A. (**2005**). "Dissociations in perceptual learning revealed by adult age differences in adaptation to time-compressed speech," J. Exp. Psychol. Human Percept. Perform. **31**, 1315–1330.
Poeppel, D. (**2003**). "The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time,'" Speech Commun. **41**, 245–255.
Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., and Formisano, E. (**2014**). "Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex," PLoS Comput. Biol. **10**, e1003412.
Schönwiesner, M., and Zatorre, R. J. (**2009**). "Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI," Proc. Natl. Acad. Sci. U.S.A. **106**, 14611–14616.
Shamma, S. (**2001**). "On the role of space and time in auditory processing," Trends Cogn. Sci. **5**, 340–348.

Shamma, S., and Lorenzi, C. (**2013**). "On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system," J. Acoust. Soc. Am. **133**, 2818–2833.
Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Sheft, S., Ardoint, M., and Lorenzi, C. (**2008**). "Speech identification based on temporal fine structure cues," J. Acoust. Soc. Am. **124**, 562–575.
Shub, D. E., and Richards, V. M. (**2009**). "Psychophysical spectro-temporal receptive fields in an auditory task," Hear. Res. **251**, 1–9.
Singh, N. C., and Theunissen, F. E. (**2003**). "Modulation spectra of natural sounds and ethological theories of auditory processing," J. Acoust. Soc. Am. **114**, 3394–3411.
Slaney, M. (**1998**). https://engineering.purdue.edu/∼malcolm/interval/1998-010/ (Last viewed 08/08/2016).
Ter Keurs, M., Festen, J. M., and Plomp, R. (**1992**). "Effect of spectral envelope smearing on speech reception. I," J. Acoust. Soc. Am. **91**, 2872–2880.
Ter Keurs, M., Festen, J. M., and Plomp, R. (**1993**). "Effect of spectral envelope smearing on speech reception. II," J. Acoust. Soc. Am. **93**, 1547–1552.
Theunissen, F. E., and Elie, J. E. (**2014**). "Neural processing of natural sounds," Nat. Rev. Neurosci. **15**, 355–366.
van Santen, J., Mishra, T., and Klabbers, E. (**2008**). "Prosodic processing," in *Springer Handbook of Speech Processing* (Springer, Berlin), pp. 471–488.
Versfeld, N. J., and Dreschler, W. A. (**2002**). "The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners," J. Acoust. Soc. Am. **111**, 401–408.
Woolley, S. M., Fremouw, T. E., Hsu, A., and Theunissen, F. E. (**2005**). "Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds," Nat. Neurosci. **8**, 1371–1379.
Zeng, F.-G., Nie, K., Liu, S., Stickney, G., Del Rio, E., Kong, Y.-Y., and Chen, H. (**2004**). "On the dichotomy in auditory perception between temporal envelope and fine structure cues (L)," J. Acoust. Soc. Am. **116**, 1351–1354.
Zilany, M. S., and Bruce, I. C. (**2007**). "Predictions of speech intelligibility with a model of the normal and impaired auditory-periphery," in *International IEEE/EMBS Conference on Neural Engineering, 2007. CNE'07*, 3rd ed. (IEEE, Piscataway, NJ), pp. 481–485.