Probing temporal modulation detection in white noise using intrinsic envelope fluctuations: A reverse-correlation study

Léo Varnet, Christian Lorenzi

Probing temporal modulation detection in white noise using intrinsic envelope fluctuations: A reverse-correlation study

Léo Varnet and Christian Lorenzi

Citation: The Journal of the Acoustical Society of America **151**, 1353 (2022); doi: 10.1121/10.0009629
View online: https://doi.org/10.1121/10.0009629
View Table of Contents: https://asa.scitation.org/toc/jas/151/2
Published by the Acoustical Society of America

JASA ARTICLE

Probing temporal modulation detection in white noise using intrinsic envelope fluctuations: A reverse-correlation study

Léo Varnet[a] and Christian Lorenzi[b]
Laboratoire des Systèmes Perceptifs, Département d'Études Cognitives, École Normale Supérieure, Université Paris Sciences & Lettres, Centre National de la Recherche Scientifique, 75005 Paris, France

ABSTRACT:
Part of the detrimental effect caused by a stationary noise on sound perception results from the masking of relevant amplitude modulations (AM) in the signal by random intrinsic envelope fluctuations arising from the filtering of noise by cochlear channels. This study capitalizes on this phenomenon to probe AM detection strategies for human listeners using a reverse correlation analysis. Eight normal-hearing listeners were asked to detect the presence of a 4-Hz sinusoidal AM target applied to a 1-kHz tone carrier using a yes-no task with 3000 trials/participant. All stimuli were embedded in a white-noise masker. A reverse-correlation analysis was then carried on the data to compute "psychophysical kernels" showing which aspects of the stimulus' temporal envelope influenced the listener's responses. These results were compared to data simulated with different implementations of a modulation-filterbank model. Psychophysical kernels revealed that human listeners were able to track the position of AM peaks in the target, similar to the models. However, they also showed a marked temporal decay and a consistent phase shift compared to the ideal template. In light of the simulated data, this was interpreted as an evidence for the presence of phase uncertainty in the processing of intrinsic envelope fluctuations. © *2022 Acoustical Society of America*.
https://doi.org/10.1121/10.0009629
(Received 6 August 2021; revised 3 February 2022; accepted 3 February 2022; published online 25 February 2022)
[Editor: Joshua G Bernstein] Pages: 1353–1366

I. INTRODUCTION

Band-limited noises, in particular bandpass-filtered Gaussian white noise, show random intrinsic envelope fluctuations[1] from one realization to the next—in other words, their temporal envelopes are not perfectly flat but exhibit a certain amount of variability, both across time and across realizations, as illustrated in Fig. 1. For this reason, the notion of "steady" noise has sometimes been considered as misleading by researchers (Stone *et al.*, 2011; Stone *et al.*, 2012) as steadiness must be understood here as a statistical property, i.e., stationarity of the stochastic process, rather than as an acoustical property. These fluctuations in intensity over time have measurable perceptual consequences: in amplitude-modulation (AM) detection tasks conducted on human participants, for example, the use of a steady noise as a carrier can result in modulation masking effects, because listeners have to separate the target modulation from nonrelevant intrinsic envelope fluctuations (Dau *et al.*, 1997; Dau *et al.*, 1999).

As a first approximation, the average envelope power spectrum of a bandpass Gaussian noise falls linearly as a function of modulation rate, with a maximum rate corresponding to the width of the band [see, e.g., Rice (1944), p. 148; Lawson and Uhlenbeck (1950), p. 62; and Hartmann (2004), p. 535]. This downward-sloping spectrum partly determines the effects of band limited noise on AM detection thresholds for human listeners. For instance, Dau *et al.* (1997) and Dau *et al.* (1999) reported that the bandwidth of the noise carrier had a major influence on the shape of temporal modulation transfer functions (TMTFs), with narrowband carriers yielding bandpass TMTFs while wideband carriers yield lowpass TMTFs. This is because intrinsic envelope fluctuations cover a smaller range of modulation frequencies for narrow-band than for wideband carriers.

In the case of wideband noise carriers such as Gaussian white noise, however, modulation masking properties are not primarily determined by the bandwidth of noise, but by the spectral resolution of the ear (Dau *et al.*, 1997). Indeed, in a normally functioning cochlea, broadband signals are decomposed by the filtering on the basilar membrane into a series of narrow-band signals. Therefore, the envelope power spectrum of the noise in each cochlear channel is directly related to the bandwidth of this channel, expressed on the equivalent rectangular bandwidth (ERB) scale. For example, the filtering of the white-noise stimulus represented in Fig. 1(A) into a 1-ERB-wide bandpass filter centered around 1 kHz gives rise to low-frequency intrinsic envelope fluctuations as shown in Fig. 1(B). Therefore, as described above, these fluctuations can elicit modulation masking in AM detection tasks, with magnitude and tuning directly depending on the width of cochlear filters in the region of the cochlea spanned by the target.

Several methods have been designed to produce random stimuli with a minimal amount of intrinsic envelope

[a]Electronic mail: leo.varnet@ens.psl.eu ORCID: 0000-0002-9702-2649.
[b]ORCID: 0000-0001-7240-1653.

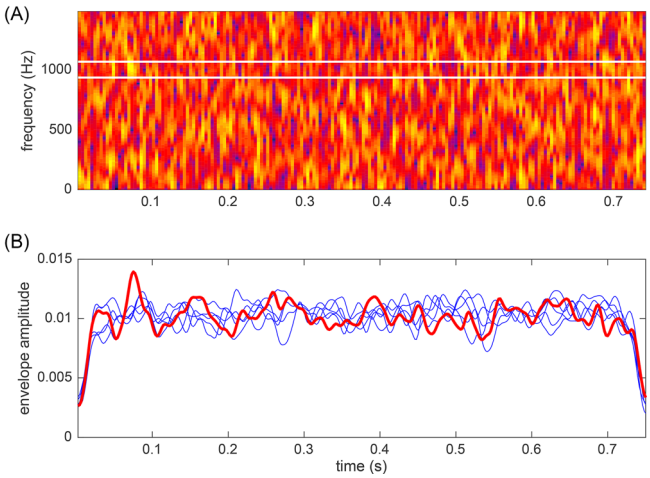

FIG. 1. (Color online) Illustration of intrinsic envelope fluctuations in a band limited noise. The noise stimuli and the procedure for deriving the envelopes are the same as described in Secs. II B and II D. (A) Spectrogram of a 750-ms-long white noise sample. White lines delimit a 1-ERB-wide-band around 1 kHz. (B) Temporal envelopes for 6 realizations of a white noise, filtered with a 1-ERB-wide Butterworth filter around 1 kHz. The red line corresponds to the realization displayed in (A).

fluctuations, such as low-noise noise (Kohlrausch *et al.*, 1997) or pulse-spreading harmonic complexes (Hilkhuysen and Macherey, 2014). At slow modulation rates, AM detection thresholds are reduced by 20 dB when using a low-noise noise narrowband carrier compared to a Gaussian narrowband carrier (Dau *et al.*, 1999). A smaller but still significant gain of ∼8 dB was found by Hilkhuysen and Macherey (2014) using slightly different experimental conditions. This dramatic improvement provides a first estimate of the strong deleterious impact of these intrinsic envelope fluctuations on modulation detection.

In this study, we explored the effects of intrinsic envelope fluctuations on AM detection using a psychophysical reverse-correlation (revcorr) method (Ahumada and Lovell, 1971; Murray, 2011). This "microscopic psychophysics" approach [or "molecular" in the terminology of Green (1964)] aims at revealing the relationship between trial-by-trial random fluctuations in the stimuli and the corresponding behavioral responses of the listener engaged in a simple psychophysical task (for example, a yes/no detection task). Within the auditory domain, it has been applied to a large variety of tasks from loudness perception (Ponsot *et al.*, 2016) to speech comprehension (Varnet *et al.*, 2015; Venezia *et al.*, 2016). At the core of the method is the notion of a kernel. A psychophysical kernel is typically computed as the correlation between the vector of random fluctuation presented in each trial and the corresponding response of the participant. The resulting vector of correlation weights in the stimulus space reveals which characteristics of the masker interfere with the ongoing task. It describes the tuning properties of the overall processing chain engaged by the listener when processing the stimuli, thus allowing for a finer-grained and data-driven characterization of the perception mechanisms than traditional detection-threshold estimation paradigms.

When applied to tone-in-noise detection tasks, the revcorr approach has revealed that listeners are very sensitive to intrinsic envelope fluctuations elicited by the masker in the spectrotemporal region of the target (Ahumada *et al.*, 1975; Gilkey and Robinson, 1986; Schönfelder and Wichmann, 2013; Shub and Richards, 2009). Three studies have confirmed this view with AM detection tasks. The two most recent attempts have relied on a revcorr approach deployed in the modulation domain, that is, based on random fluctuations imposed on the stimulus envelope.

Joosten *et al.* (2016) used stimuli composed of nine white-noise segments, each 33-ms long. Random variations in amplitude were independently applied to these segments and the listeners were instructed to detect the presence of an additional level increment (or decrement) in the central segment. The temporal kernels were obtained by taking the difference between the average of the noise vectors that yielded a positive response of the participant ("I detected the target level increment/decrement") and the average of the noise vectors that yielded a negative response ("I did not detect the target"). This temporal kernel revealed that listeners were particularly sensitive to the presence of noise in the central segment, corresponding to the location of the target, but also in adjacent segments, to a lesser extent. Based on this approach, they were able to obtain an estimate of the selectivity of the auditory system in the modulation domain consistent with previous work based on more traditional masking paradigms.

More recently, Ponsot *et al.* (2021) explored the perception of spectrotemporal modulation (STM) applied to a noise carrier. The task consisted of the detection of an elementary STM (temporal modulation rate = 7.1 Hz, spectral modulation rate = 1 cycle/oct) embedded in a noise composed of other STMs with various orientations. As in Joosten *et al.* (2016), the authors relied on a reverse-correlation approach to relate the content of the noise in the modulation domain, i.e., the amplitude of each individual STM component in the masker, to the particular response of the listener. Based on the obtained kernel, they were able to assess the selectivity of the participants' listening strategy in the spectrotemporal orientation space, compared to the optimal strategy which would consist of taking into account only the amplitude of the target STM. They complemented the analysis with simulation data and showed that the particular shape of the human perceptual kernel could be accounted for by a modulation filterbank (MFB) model of the human auditory system (Dau *et al.*, 1997; King *et al.*, 2019; Vecchi *et al.*, 2021).

The two above-described studies of modulation perception mechanisms using revcorr both rely on the introduction of random perturbations in the modulation domain. Theoretically, however, the use of a modulated noise is not required because, as discussed above, a steady noise also induces random perturbations in the modulation domain due to intrinsic envelope fluctuations, which interfere with the mechanisms responsible for AM perception. As of today, only one revcorr study by Ardoint *et al.* (2007) explored modulation perception using a steady-noise masker. The

https://doi.org/10.1121/10.0009629

experiment consisted of a 4-Hz AM detection task in a Gaussian white noise presented at a fixed signal-to-noise ratio (SNR) of +6 dB. Once the data were collected, the authors passed each realization of the masker through a bank of cochlear filters, and computed the temporal kernel by averaging the resulting filtered noise envelopes conditional on the corresponding behavioral response of the participant. In 4 listeners out of 10, the psychophysical kernels showed a prominent 4-Hz component—indicating that this subgroup of participants was confused by the presence of a 4-Hz component in the intrinsic envelope fluctuations of the masker similar to the target. On the contrary, the remaining 6 individuals did not show any clear pattern in their perceptual kernel. The authors interpreted these mixed results as reflecting poorer envelope encoding (higher envelope-phase uncertainty) in these listeners.

In this context, the purpose of the present study was twofold:

(1) To explore the effects of intrinsic envelope fluctuations on AM detection using a reverse-correlation experiment similar to that of Ardoint *et al.* (2007). For this purpose, we collected a large set of behavioral data for an AM-detection task in white noise. Each participant was then characterized by calculation of temporal kernels, representing the relationship between the intrinsic envelope fluctuation pattern in each trial and the corresponding decision of the listener. We decided to use a lower SNR than in Ardoint *et al.* (2007) (−10 dB instead of +6 dB), as we suspected that this would yield more robust and more consistent kernels.
(2) As a second step, human perceptual kernels were compared to those obtained by a computational model of the human auditory system, an approach comparable to that used by Ponsot *et al.* (2021). While the structure of the model front-end is generally agreed upon, little is known about the cues and statistics upon which the auditory system bases its decision (Ewert and Dau, 2000; King *et al.*, 2019; Osses Vecchi and Kohlrausch, 2021; Strickland and Viemeister, 1996). Here, we capitalized on the perceptual effect of intrinsic envelope fluctuations to compare internal representations and decision strategies in real and in artificial listeners. This complementary modeling approach will therefore help us to interpret the kernels measured in humans. In particular, we will compare three simulated decision strategies differing with respect to their use of envelope phase information.

II. MATERIALS AND METHODS

A. Listeners

Nine listeners took part in the experiment. Their pure-tone detection thresholds were tested at 250, 500, 750, 1000, 1500, 2000, 3000, 4000, 6000, and 8000 Hz using a Madsen Itera II (Otometrics, Taastrup, DK) audiometer. Eight listeners had audiometric thresholds at or below 20 dB hearing level for both ears at all frequencies tested up to 4000 Hz. The last participant (S4) had only one normal-hearing ear. Their audiograms are shown in supplementary Fig. 1.[2]

The data from one participant was removed from further analysis as she was not able to perform the experiment: she obtained very low performance level compared to the group (see supplementary[2] Fig. 22). The remaining participants were 8 listeners (4 females) aged between 26 and 35 years.

All listeners provided written consent and received financial compensation for their participation. The study received the approval of the local ethical committee of University Paris Descartes (IRB 00012020–12).

B. Stimuli

All stimuli were 750-ms long. They consisted of AM tones (target) and unmodulated tones (nontarget) embedded in a white noise masker with a 24 kHz bandwidth. AM tones were generated using the following formula:

$$AM(t) = [1 + m\cos(2\pi f_m t + \pi)]\sin(2\pi f_c t), \quad (1)$$

where m is the AM depth, f_m (= 4 Hz) is the AM rate, f_c (= 1 kHz) is the carrier frequency, and t is the time-sample vector. Note that the modulation phase was fixed to π so that the stimulus endpoints correspond to minima of the envelope, with maxima at 125, 375, and 625 ms. Modulated and unmodulated tones were added to a white noise at −10 dB SNR, corresponding to a local SNR of +33.8 dB at 1 kHz (expressed as the signal level relative to the dB/Hz spectrum level of the noise). This SNR was chosen to strike a balance between preserving audibility of the carrier (i.e., not making SNR too low) and making it more likely that noise fluctuations could actually interfere with the modulations in the stimulus (i.e., not making SNR too high). The resulting stimuli were gated on and off with 75-ms raised-cosine ramps and equalized in root mean square (rms) amplitude.

The overall sound level was fixed to 65 dB sound pressure level (SPL) throughout the experiment. Stimuli were generated digitally at a sample rate of 48 kHz using MATLAB R2019 and sent to a Fireface RME audio interface (24-bit resolution). They were presented diotically using Sennheiser HD 650 headphones within an IAC sound-proof booth via a wall patch (except for participant S4: stimuli presented only to the best ear).

C. Procedure

Participants completed a set of 3000 trials each, consisting of 1500 AM tones and 1500 unmodulated tones, embedded in a white noise and presented in random order. For each trial, they were asked to listen carefully and to indicate, by a button press, whether the tone was modulated in amplitude or not. The experiment was divided into 10 blocks of 300 trials, separated with breaks. Participants were explicitly informed that modulated and unmodulated tones were equally likely, and their responses were checked for the

presence of a bias (see Sec. III). No feedback was provided. Before the beginning of each block, subjects were allowed to perform a short practice session similar to the test phase except that the correct answers were displayed after each trial and that an additional button permitted replaying the stimulus.

Within each block, the AM depth was adjusted on a dB scale (20logm) using a transformed 2-down 1-up staircase procedure (starting value: –1 dB). The initial staircase stepsize was 2 dB; it was decreased by 10% after each step until it reached a minimum value of 0.5 dB, resulting in step sizes of 2, 1.8, 1.62 dB, etc. The purpose of this adaptive procedure was to continuously track the 70.7% correct performance level, despite potential fluctuations in attention or strategy shifts (Levitt, 1971).

D. Analysis

The objective of the present experiment was to carry a reverse correlation analysis on the intrinsic envelope fluctuations of the masker. Envelopes were extracted from the waveform by bandpass filtering between 935 and 1068 Hz (1000 Hz ± 0.5 ERB, Glasberg and Moore, 1990) using a second-order Butterworth filter, in order to account for the spectral resolution of the human ear in the region of the carrier frequency,[3] followed by half-wave rectification and low-pass filtering (cut-off frequency = 30 Hz, first-order Butterworth filter). The obtained envelopes were then downsampled to 480 Hz.

In the present experiment, target and non-target trials only differed by the presence of a sinusoidal AM. Only the non task-relevant information in the stimuli (i.e., the tone carrier and the noise, excluding the sinusoidal AM target) was taken into account when extracting the envelopes, in order to ensure that the psychophysical kernel reflects the auditory processing of the stimulus and not the structure of the stimulus itself.

For each participant, psychophysical kernels were derived separately for target-present and target-absent trials. The target-present kernel was obtained as the difference between the average envelope in hit trials (correctly identified AM tone) and the average envelope in miss trials (pure-tone response while an AM tone was played). A target-absent kernel was similarly obtained as the difference between the average envelope in false-alarm trials (AM-tone response while a pure tone was played) and the average envelope in correct-rejection trials (correctly identified pure tone). Each kernel was therefore computed on 1500 trials; however, the number of trials aggregated in the averaged envelope can vary from one participant to the other depending on their hit and false-alarm rates. Finally, a "general kernel" was obtained as the mean of the target-present and target-absent kernels (Ahumada, 1996; Murray, 2011). For each kernel, we derived a single summary metric attempting to capture the temporal decay in the weights, obtained as the ratio between the rms weights in the first and the second half of the kernel.

Each temporal kernel was then transformed to the complex Fourier domain using a zero-padded fast Fourier transform. The magnitude spectrum of the temporal kernel is hereafter called Fourier kernel. Individual complex Fourier kernels were then characterized by the phase and amplitude of their 4-Hz component.

A nonparametric bootstrap procedure was used to assess the reliability of each estimate, i.e., all individual temporal and Fourier kernels as well as metrics derived from these kernels (Ponsot *et al.*, 2021). For each participant, 200 new estimates were computed from 200 random subsets of the data (obtained by sampling with replacement), and their distribution was used as an approximation of the sampling variability of the estimator. The bootstrap distribution was reported in the text as 95% confidence intervals. In addition, 95% confidence intervals under the null hypothesis of no relationship between the noise and the listeners' percept were derived using a randomization procedure based on 200 random permutations of each participant's response vector. Either bootstrap or randomization intervals were used depending on whether the focus was on assessing the robustness of the point estimate (bootstrap) or on testing whether it significantly differed from the chance distribution (randomization). As confidence intervals under the null hypothesis were very similar across participants, only the average intervals across participants are reported.

E. Modeling: MFB front-end

To better understand the shape of the human kernels, the above-described analysis was also carried out on simulated data. The objective of this modeling effort was not to reproduce details of the human kernels, but rather to identify the mechanisms which may explain patterns observed in the behavioral data.

For this purpose, a widely accepted MFB front-end model of temporal-envelope processing by the human auditory system was used. A block diagram of the MFB front-end is shown in supplementary[2] Fig. 42. This computational model, which corresponds to a simplified version of the model developed by Dau *et al.* (1997), has proven successful in reproducing behavioral data in a wide range of modulation perception tasks (Cabrera *et al.*, 2019; King *et al.*, 2019; Ponsot *et al.*, 2021; Wallaert *et al.*, 2018). For the sake of simplicity, and given that the target signal comprised a sinusoidal modulation frequency applied to a pure-tone carrier, a mono-channel version of the model was considered here, with a single cochlear filter tuned to the carrier frequency (1 kHz) and a single modulation channel tuned to the modulation frequency to be detected (4 Hz). The implementation of the MFB front-end as used in this study is available as the routine king2019 within the AMT toolbox (v1.0) for MATLAB (Majdak *et al.*, 2021; Vecchi *et al.*, 2021).

The model included the following stages, in sequential order:

https://doi.org/10.1121/10.0009629

(1) A linear 1-ERB-wide Gammatone filter centered at 1 kHz, corresponding to a level-independent approximation of the on-frequency cochlear filter. The MFB model uses the all-pole implementation of the Gammatone filters described by Hohmann (2002).
(2) Amplitude compression using a broken-stick input-output function. The function is linear up to a knee-point of 30 dB SPL and compressive above this kneepoint, using a power law with an exponent of 0.3.
(3) Half-wave rectification followed by low-pass filtering (first-order Butterworth filter, 1000-Hz cutoff). This stage performs envelope extraction.
(4) Adaptation through high-pass filtering (first-order Butterworth filter, 3-Hz cutoff).[4] The compression, envelope extraction and adaptation stages form a simple model of peripheral auditory processing.
(5) The resulting envelope is further band-limited by a modulation filter centered around the target (first-order band-pass Butterworth filter centered at 4 Hz). A quality factor of 1 (hence a 4-Hz bandwidth) was chosen for that filter in accordance with Ewert and Dau (2000) [see also Jørgensen and Dau (2011) and Wallaert *et al.* (2017)].
(6) The internal representation of the stimulus is finally down-sampled to 4 kHz. Note that envelope phase is preserved throughout all stages of the model.

F. Modeling: MFB decision device

Relatively little is known about how exactly the auditory system processes continuous, time-varying internal representations into single binary decision statistics [see, for example, Osses Vecchi and Kohlrausch (2021), Table I], in contrast to the well established front-end processing described in Sec. II E. In the following, three alternative decision devices were considered.

The "optimal detector" (O-D) strategy for the task consists of selecting one of the two responses on the basis of the cross correlation between the internal representation of the presented stimulus (IR_{stim}) and those of the unmodulated ($IR_{nontarget}$) and modulated (IR_{target}) tones (Dau *et al.*, 1997), corresponding to the image of the sounds stored in memory—or "template." IR_{target} and $IR_{nontarget}$ were computed in the absence of external noise maskers and normalized to unit energy similar to Dau *et al.* (1996a). In the case of a single-interval paradigm such as the one used here, the model is provided with *a priori* information about the relative probabilities of possible stimulus alternatives (Dau, 1999). This corresponds to setting the decision criterion c_1 in the following decision rule:

$$response = \begin{cases} \text{"AM tone"} & \text{if } (IR_{stim} \star IR_{target})[0] \\ & \quad -(IR_{stim} \star IR_{nontarget})[0] \geq c_1 \\ \text{"pure tone"} & \text{otherwise,} \end{cases} \tag{2}$$

where $\star$ indicates the non-normalized cross correlation operator $(f \star g)[n] = (1/N)\Sigma_{k=1}^{N} f[k]g[k-n]$. In the present study, it was assumed that the model targets an equal rate of "AM" and "pure tone" responses, making it similar to real participants (see Sec. II C). For this purpose, the decision criterion c_1 was determined empirically by simulation so that the proportions of the two alternatives were approximately equal.

A second decision device considered in the present study is the envelope power-spectrum detector (EPS-D), which has also been used in several modeling studies (Ewert and Dau, 2000). Contrary to the optimal detector, this decision device does not involve an explicit template representation of the target and non-target stimuli. Instead, it bases its decision on the envelope power in the modulation channel corresponding to the target. In the present study, the model was restricted to the target channel (see Sec. II E) and therefore the relevant envelope statistics could be simplified to $(1/N)\Sigma_{k=1}^{N} IR_{stim}^2[k] = (IR_{stim} \star IR_{stim})[0]$. As before, the decision criterion c_2 was fixed empirically to yield equal proportions of target and nontarget responses,

$$response = \begin{cases} \text{"AM tone"} & \text{if } (IR_{stim} \star IR_{stim})[0] \geq c_2 \\ \text{"pure tone"} & \text{otherwise.} \end{cases} \tag{3}$$

Note that, compared to the optimal detector, this strategy discards the envelope phase information in the stimulus as only the overall energy in the envelope is taken into account, irrespective of its temporal organization within the observation interval.

The two above decision rules are based on cross-correlations operators: cross correlation with zero lag for the optimal detector, and auto-correlation with zero lag in the case of the envelope power-spectrum detector. A third cross correlation-based decision device was finally considered, as an intermediate option between the two above. Like the "optimal detector" it involves an explicit template representation of the target signal, but also includes some uncertainty about envelope phase making it more similar to the envelope power spectrum. Here, the cross correlation function was not evaluated at zero lag, as before, but at all positive or negative lags n, and the maximum cross correlation was retained,

TABLE I. Summary of real and artificial listeners' performance in the task.

	EPS-D	O-D	O-Du	XC-D	XC-Du	Participants
Percent correct	70.8	71.0	71.3	70.9	70.7	70.72 ± 0.37
Average m (dB)	−11.9	−30.7	−31.0	−21.6	−16.8	−16.8 ± 2.5
Rate of "AM tone" response	0.49	0.49	0.50	0.51	0.49	0.43 ± 0.1

$$response = \begin{cases} \text{“AM tone”} & \text{if } \max_n(IR_{stim} \star IR_{target})[n] \geq c_3 \\ \text{“pure tone”} & \text{otherwise.} \end{cases} \quad (4)$$

This decision device will be referred to as XC-D in this article. In practice, this suboptimal template-matcher scans the internal representation of the stimulus for a pattern resembling the target AM template. A similar “max of cross correlation” decision statistics introducing a certain degree of envelope phase uncertainty into the model was used by Osses Vecchi and Kohlrausch (2021) in the context of a three-interval forced-choice task [see also King *et al.* (2019) and Wallaert *et al.* (2018)].

Crucially, two decision devices (O-D and XC-D) must be provided with templates of the expected sound representation (IR_{target} and/or $IR_{nontarget}$), which will determine the strategy of the artificial listener (i.e., “what to listen to” in the input signal). For this reason templates are usually derived in a suprathreshold condition, where the cues are easily detectable (Dau *et al.*, 1996a). Here, the templates for O-D and XC-D were calculated once at the beginning of the experiment from non-noisy stimuli at the top of the staircase (Dau *et al.*, 1996b; King *et al.*, 2019; Osses Vecchi and Kohlrausch, 2021). Nevertheless, two additional decision device were considered (O-Du and XC-Du) in which IR_{target} was updated at each step of the staircase with the clean target stimulus at the corresponding modulation depth. These last two models provide a coarse simulation of the influence of contextual effects on internal template during the course of the experiment (see the appendix in Derleth and Dau, 2000, for a similar approach).

All artificial listeners were tested on the same set of noise and in the same trial order for better comparability between detectors.

III. RESULTS

A. Performance in the task

Despite the long duration of the experiment (3000 trials/participant, i.e., ≈ 3 h including breaks), participants showed a rather stable behavior, both in terms of performance and bias, indicating no or little learning effect (see Table I and supplementary[2] Fig. 22). They reported no perceived effect of excessive mental fatigue over the course of the experiment, although some participants admitted experiencing occasional and brief attention loss. The average proportion of correct responses [± standard deviation (SD)] for the group was 70.72 ± 0.37%, close to the 70.7% score targeted by the adaptive staircase algorithm. The mean modulation depth across all participants was $m = -16.8 \pm 2.5$ dB. Overall, participants showed only a slight response bias (mean rate of “AM” answer = 0.43 ± 0.1).

A complementary picture of the participants’ behaviour over the experiment can be obtained by plotting their correct response rates as a function of the depth of the modulation to be detected—that is, the current state of the staircase—and trial type (Fig. 2). This figure makes it apparent that performance in target-present trials (hit rate) depended on modulation depth, while performance in target-absent trials (correct rejection rate) did not.

B. Psychophysical kernels

Figure 3 displays the temporal kernels for the eight participants, calculated for all trials (general kernel) or separately in the target-present and in the target-absent conditions, together with the 95% confidence interval under the null hypothesis. As explained in Sec. II, these kernels correspond to the difference in the average envelopes between hits and misses and between false alarms and correct rejections. Also, shown in this figure is the ideal template for the task (dotted line), defined as the weighting of temporal information used by an optimal detector operating directly in the temporal envelope domain. In practice, the ideal template was calculated as the difference between the envelopes of the target and nontarget stimuli.

It is clear from this figure that all kernels showed a very similar pattern across participants, with significant positive weights in the regions corresponding to the first two peaks of the target modulation (i.e., around 125 and 375 ms) and significant negative weights in the region of the first trough

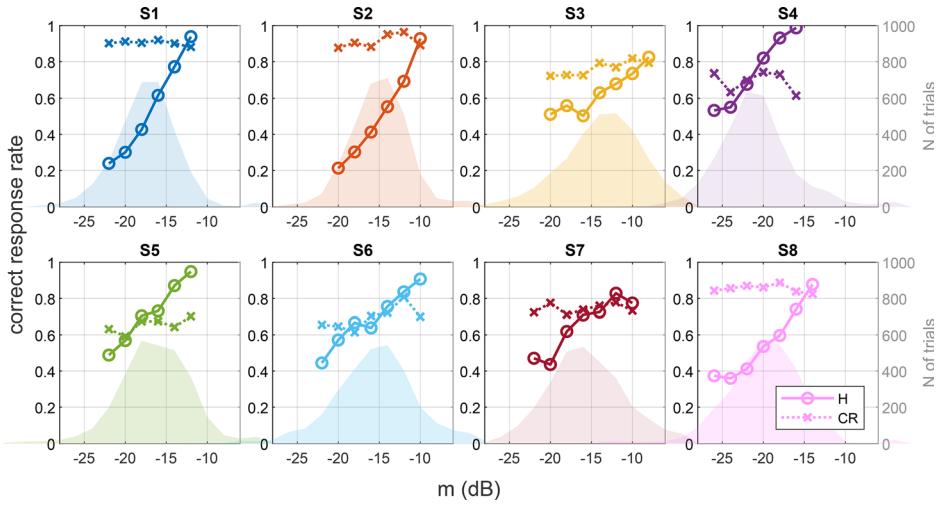

FIG. 2. (Color online) Proportion of hits (solid line, circles) and correct rejections (dotted line, crosses) for each participant included in the study as a function of the depth of the modulation to be detected. The distribution of trials is shown as a shaded histogram. Correct response rates are shown only for modulation depth with a sufficient number of trials (N ≥ 150).

https://doi.org/10.1121/10.0009629

(250 ms). For example, within a 50-ms segment around the first modulation trough, the estimated kernels from all eight participants departed from the 95% confidence interval of no effect. Importantly, this temporal structure was not an artefactual distortion induced by the target signal, as it was overall preserved in the target-absent condition where noise is the only source of trial-by-trial variability. Notably, during the first period of the target modulation, the fluctuations of the general temporal kernels were so large that, in all participants, they reached positive and negative amplitudes that were very unlikely to occur by chance only—as confirmed by the randomization 95% confidence intervals. This is a strong evidence that intrinsic envelope fluctuations in the white noise masker interfered with the decision of the participant in a systematic way. In other words, it is possible to relate the exact response of the observer to the specific configuration of the noise intrinsic envelope fluctuation on a trial-by-trial basis. In particular, all participants were significantly more likely to respond "AM tone" (respectively "pure tone") when there was a large concentration of noise envelope energy around $t = 125$ ms (respectively, $t = 250$ ms), whether the target was presented or not. Note, however that there are some slight differences between target-absent and target-present kernels, which will be discussed later in this section.

The alternation of positive and negative weights observed in Fig. 3 results in a strong spectral peak when the kernels are represented in the Fourier domain (Fig. 4). The 4-Hz component was significantly larger than chance in each individual kernel, although the peak frequency was slightly lower than 4 Hz in most participants (see below). Some, but not all, Fourier kernels showed significant weights corresponding to the harmonics of the target modulation frequency, which could suggest an effect of nonlinearities in envelope processing (Ewert *et al.*, 2002; Lorenzi *et al.*, 2001). As these weights did not appear for all participants tested, we will not investigate this possibility further here. As can be seen in Fig. 3, temporal kernels deviated from the ideal template in several other ways, including a temporal decay (the first period of the modulation was weighted more heavily than the last one) and a phase shift, particularly visible in the target-absent kernels.

A temporal decay ratio was obtained as the ratio of the rms weights in the first vs the second half of the kernel for each individual temporal kernel. Individual complex Fourier kernels were characterized by their peak frequency and the phase of their 4-Hz component. The same metrics were also extracted from the bootstrapped kernels, allowing for the derivation of confidence intervals for each estimate. These results are shown in Fig. 5 and supplementary[2] Fig. 32.

Although there was a certain amount of interindividual variability, temporal decay ratios confirmed that, in all participants, the impact of noise on the decision was significantly stronger in the first than in the second half of the stimulus (mean temporal decay ratio $= 1.65 \pm 0.17$). The peak frequency of all general Fourier kernels was lower than 4 Hz (mean $= 3.38 \pm 0.65$ Hz), although the difference was not significant at the individual level in four participants. Finally, a clear pattern emerged from the 4-Hz-component phase estimates: the 4-Hz component was almost aligned with the modulation to be detected in target-present kernels (empty circles, mean phase $= 3.37 \pm 0.19$ radians) but it was significantly shifted towards a higher phase in target-absent kernels for all participants except S2 (filled circles, mean phase $= 4.32 \pm 0.64$ radians). In fact, S2 (orange symbols) showed a slightly higher response bias

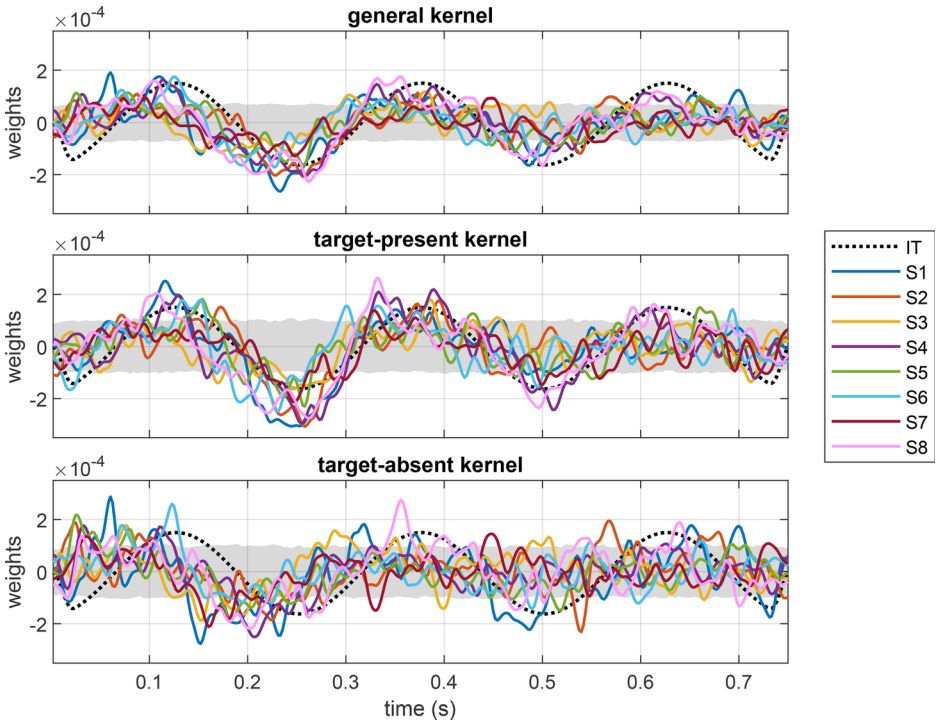

FIG. 3. (Color online) Temporal kernels calculated separately for the eight participants, for the whole experiment (top) and in the target-present (middle) and target-absent (bottom) conditions. The dotted line corresponds to the ideal template for this task (see description in text) and the shaded region to the average 95% confidence interval obtained by randomization (i.e., the weight amplitudes that would be measured under the hypothesis of no effect of the noise, see Sec. II D). The amplitude of the ideal template is arbitrary.

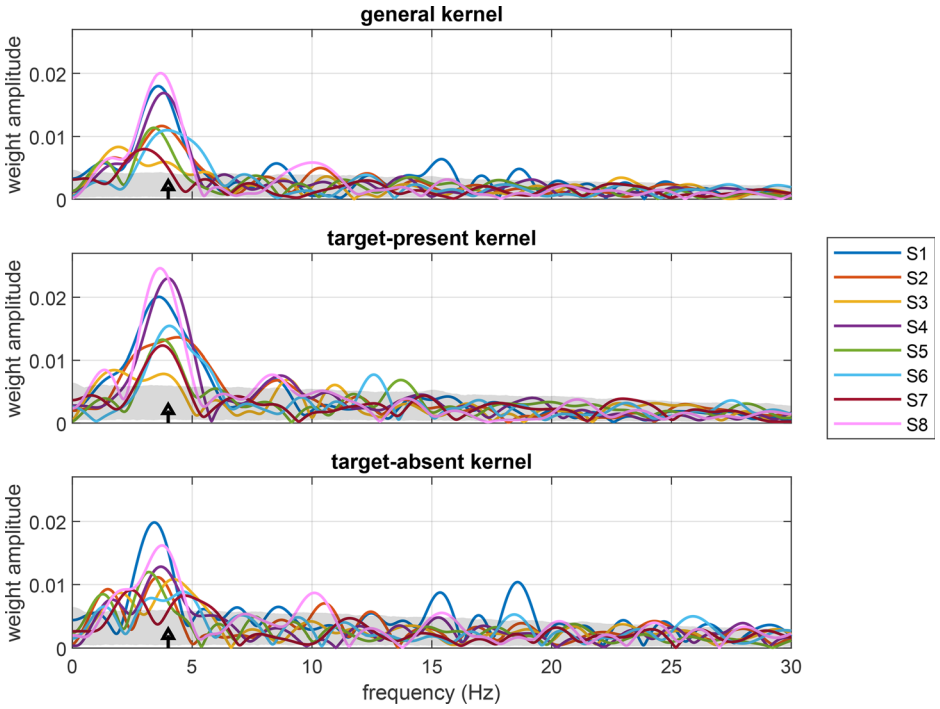

FIG. 4. (Color online) Fourier kernels, obtained by Fourier transform of the temporal kernels from Fig. 3 (same color code). The arrows indicate the target modulation frequency (4 Hz).

than the rest of the group and it is therefore likely that the phase estimates were not sufficiently reliable for this participant because of the lesser number of trials in one response category. At the group level, this phase-shift is clearly visible as a difference of $\sim$ 40 ms between the onsets of the condition-specific kernels in Fig. 3.

C. Model performance in the task

Five simple models of the human auditory system, all based on the same model front-end but using different decision devices (O-D, O-Du, EPS-D, XC-D, XC-Du, see Secs. III E and III F), were tested on the same task as real participants. Their performance in the task is presented in Table I. By construction, the response bias of each model was very close to 0.5, as decision criteria (c_1, c_2, and c_3 in the decision rules from Sec. II F) were fixed empirically to yield equal proportions of target and nontarget responses. The performance level was considerably higher for the optimal detector than for the suboptimal models, reflecting the fact that O-D is only limited by the presence of external noise, while the other models included other sources of uncertainty, in particular on the target envelope phase.

Note that, in the context of the present study, performance level should not be used as a criterion for comparing artificial listeners to real listeners. Indeed, the MFB model usually includes an internal source of stochasticity that further limits its performance (Dau *et al.*, 1996b; King *et al.*, 2019). As the focus on the present study was on revealing the systematic part of the decision process using a source of external noise, an additional source of internal noise would obscure the results, and we therefore decided not to include it in the models.

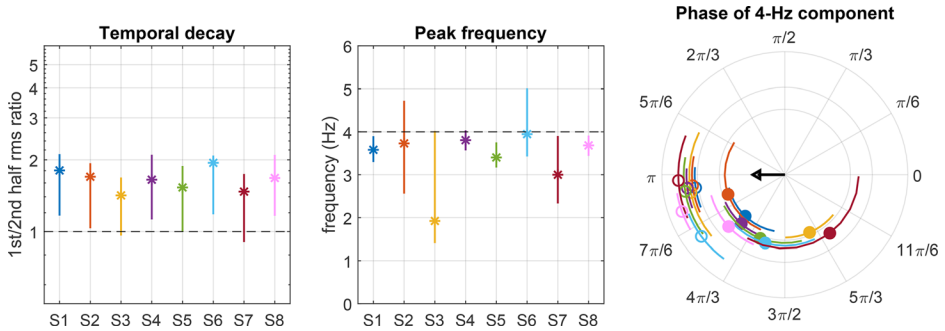

FIG. 5. (Color online) Metrics extracted from the individual kernels: temporal decay ratio of the general temporal kernel, peak frequency in the general Fourier kernel, and phase of the 4-Hz component for the target-present and target-absent Fourier kernels. Error bars show the 95% confidence interval for each estimate. Dotted lines show the target modulation frequency (4 Hz) and the theoretical ratio in the absence of temporal decay; the arrow indicates the phase of the AM target (π). Stars correspond to general kernels, empty circles to target-present kernels, filled circles to target-absent kernels. Each participant is represented with a different colour.

https://doi.org/10.1121/10.0009629

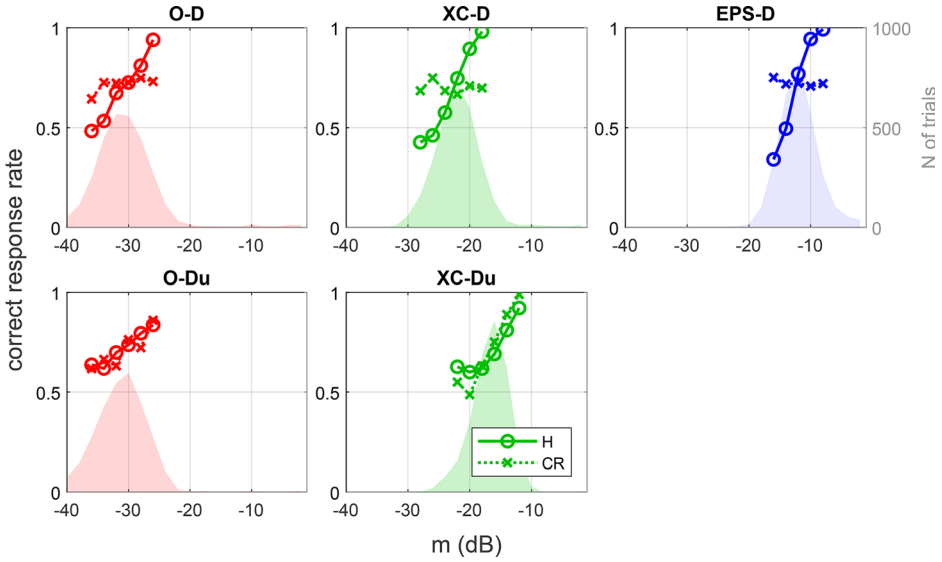

FIG. 6. (Color online) Proportion of hits (solid line, circles) and correct rejections (dotted line, crosses) as a function of current target modulation depth for models O-D, O-Du, EPS-D, XC-D and XC-Du. Same representation as in Fig. 2.

In contrast to the modeled performance levels, the investigation of hit rates and correct rejection rates provides valuable insight into the decision process (Fig. 6). As was the case for real participants (Fig. 2), EPS-D, O-D, and XC-D showed correct rejection rates independent of the current target modulation depth. In contrast, for O-Du and XC-Du, the proportion of correct rejection in target-absent trials increased with the discriminability of the target. In other words, the salience of the target-present stimulus affected the performance of the models in the nearby target-absent trials. This surprising behavior results from the fact that these decision devices relied on a template which is constructed from a "copy" of the current target (see Sec. II F), and that this template was updated from trial to trial.

Given their inconsistency with human data, and the fact that they led to psychophysical kernels very similar to the "non-updated" versions, we did not further evaluate the O-Du and XC-Du decision devices.

D. Model kernels

For each artificial listener, the temporal and Fourier kernels were derived in the same way as for real participants and compared to the average human kernel (Fig. 7). Similar to the kernels derived from the perceptual data, the kernels for O-D and XC-D decision devices showed a prominent temporal modulation, which was expected given that their templates were constructed from an internal representation of the 4-Hz target. Perhaps more surprisingly, the EPS-D

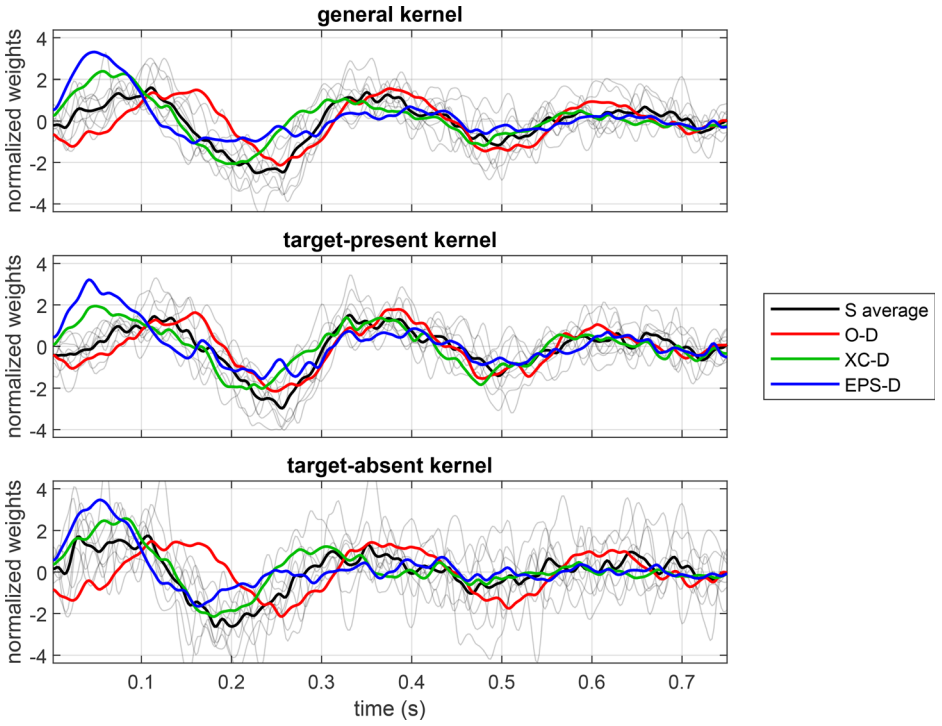

FIG. 7. (Color online) Temporal kernels for the three artificial listeners (same representation as in Fig. 3). The thin gray lines correspond to the individual kernels from Fig. 3 and the black line to their average. To facilitate comparison, weights are shown on a standardized scale.

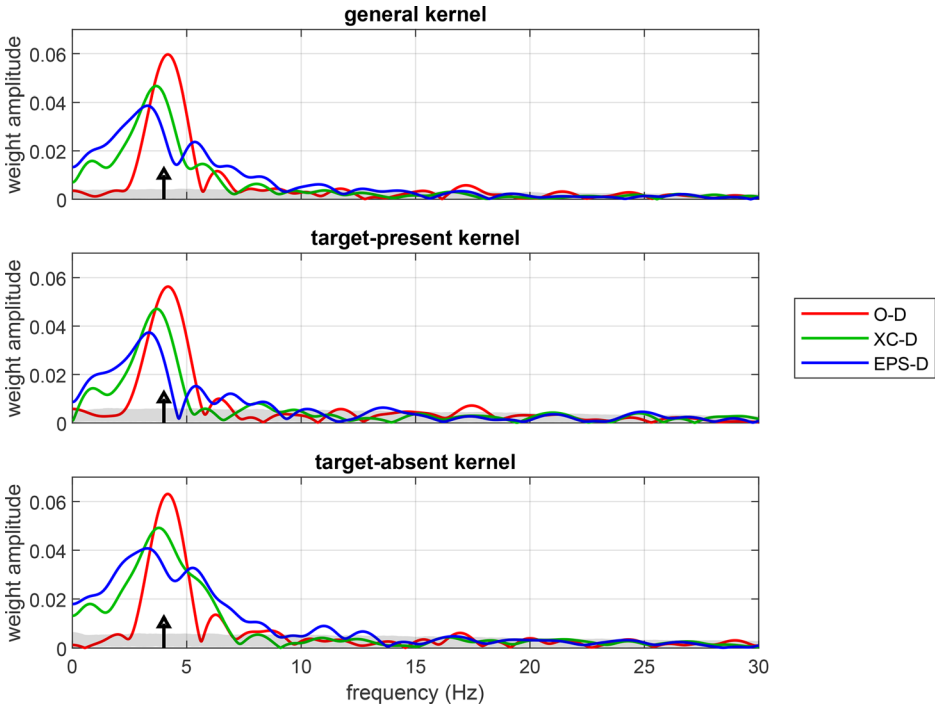

FIG. 8. (Color online) Fourier kernels for the 3 models (same representation as in Fig. 4). The arrows indicate the target modulation frequency (4 Hz).

model, which was not based on a template-matching process, was also associated with a clearly defined temporal pattern. Indeed, as explained in Sec. III B, psychophysical kernels reveal the regions of the stimulus where the presence of noise had a stronger impact on the decision. Although the EPS-D did not include an explicit template, its sensitivity to noise was not the same in every time region of the stimulus, as revealed by the temporal kernel. As seen in Fig. 8, the 4-Hz component of the kernel was highly significant for the three models, i.e., well above the 95% confidence threshold obtained by randomization (corresponding to an upper limit of 0.004 at 4 Hz). The specific shape of the simulated kernels are further discussed below.

As with real participants, three metrics were extracted from the estimated and bootstrapped kernels: temporal decay ratio, frequency of the peak and phase of the 4-Hz component (Fig. 9 and supplementary[2] Fig. 32). Overall, the EPS-D showed the strongest onset effects, resulting in an unrealistic value of the temporal decay ratio (95% confidence interval = [3.25, 4.7], compared to 1.65 for real listeners). The XC-D decision device, for which the kernel presented a moderate onset effect, also overestimated the temporal decay ratio compared to the group of real listeners (95% confidence interval = [2.46, 3.36]). Finally, the O-D model showed no onset effect and a temporal decay ratio close to one (95% confidence interval = [1.19, 1.51]), meaning that it underestimated the true value for most of the participants. A key aspect of the Fourier kernel measured in real listeners is that peak frequencies were lower than 4 Hz. The EPS-D and XC-D best reproduced this characteristic (95% confidence interval = [3.04, 3.42] for EPS-D, [3.57, 3.75] for XC-D), while O-D was associated with a peak frequency slightly but significantly above 4 Hz (95% confidence interval = [4.12, 4.23]). Finally, the phase metrics further highlighted the difference between decision devices with (XC-D and EPS-D) or without (O-D) phase uncertainty. In the latter, the temporal positions of the peaks and troughs of the target were exactly known by the model. Therefore, the resulting kernel was precisely aligned with the target, similar to the phase observed in real target-

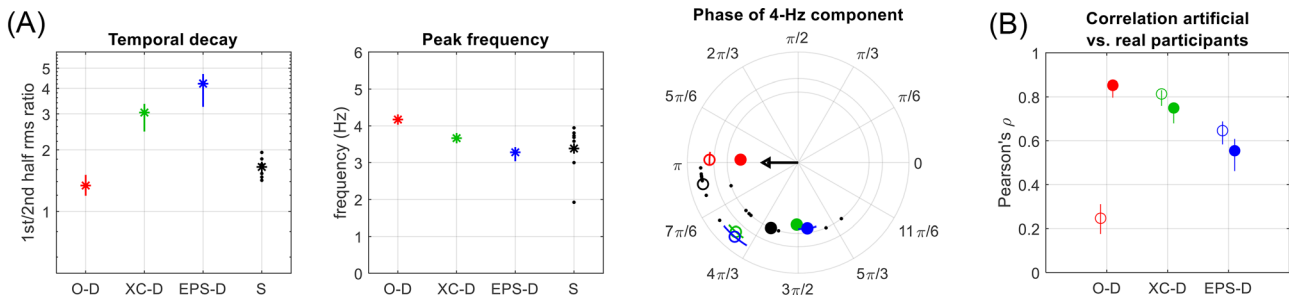

FIG. 9. (Color online) (A) Metrics extracted from the simulated (colored symbols) and real (black dots) kernels. Black symbols correspond to the average across the eight real subjects. Same representation as in Fig. 5. (B) Pearson's correlation coefficient between the temporal kernels of the three artificial listeners and the average human kernels (i.e., black vs colored lines in Fig. 7).

https://doi.org/10.1121/10.0009629

present kernels. On the contrary, for phase-uncertain decision devices, the 4-Hz component in the kernel was shifted towards higher phase, similar to real target-absent kernels. Importantly, only EPS-D and XC-D successfully reproduced the effect of condition on kernel phase by showing a $\sim\pi/6$ phase shift between target-present and target-absent kernels.

To quantify the accuracy of the listening strategies of the three decision devices in the task compared to the strategy of real participants, we measured Pearson's correlation between real and simulated kernels [Fig. 9(B)]. All simulated kernels were significantly correlated to the average human kernel (i.e., $\rho > 0.31$, the 95% significance threshold determined by randomization) except for O-D in target-absent trials. This was expected given the strong phase difference observed in this condition. On the contrary, the target-present condition was well accounted for by the O-D model (95% confidence interval for $\rho_{target-present} = [0.80, 0.87]$). Overall, XC-D was the best at reproducing human-like kernels across conditions (95% confidence interval for $\rho_{target-absent} = [0.76, 0.83]$; $\rho_{target-present} = [0.68, 0.77]$).

IV. DISCUSSION

The most recent studies of modulation perception using a revcorr approach have relied on dimensional noise, that is, random fluctuations deployed in the modulation domain (Joosten *et al.*, 2016; Ponsot *et al.*, 2021). In their article, Joosten *et al.* wrote that, when applying psychophysical reverse correlation to AM processing.

Adding acoustic white-noise is inappropriate, partly because the envelope fluctuations it induces are difficult to control and exercise adequately, and partly because the dimensionality of the space needing characterization is impractically large to measure.

As a matter of fact, at this time the only previous attempt at deriving psychophysical kernels on an AM detection task in white noise yielded rather mixed results with only four participants out of ten showing a distinct pattern in their temporal kernels (Ardoint *et al.*, 2007). In the present study, we showed that it is in fact possible to measure reliable kernels for AM detection using a white-noise masker, provided that the SNR is sufficiently low [–10 dB SNR in the present study vs +6 dB SNR in Ardoint *et al.* (2007)].

More precisely, the clear temporal pattern shown in each individual kernel in Fig. 3 provides strong evidence that random intrinsic envelope fluctuations in the white noise masker interfered with the decision of the participant in a systematic way. This supports the idea that intrinsic envelope fluctuations of notionally steady noise have a measurable effect on auditory perception (Stone *et al.*, 2011; Stone *et al.*, 2012), consistent with the results obtained by Varnet *et al.* (2013) and Varnet *et al.* (2015) on phoneme perception tasks using a similar revcorr approach. By examining how a specific pattern of intrinsic envelope fluctuations can bias perception towards detecting a 4-Hz AM target or not, we can obtain a fine-grained characterization of the mechanisms underlying AM detection using a data-driven approach. Temporal kernels are a useful tool to explore the systematic component of categorization errors. They can be conceived (up to a linear approximation) as the pattern of phase-locked random fluctuation that would most efficiently bias the participant's decision towards detecting the target. Conversely, the negative of the temporal kernel corresponds to the pattern that is most likely to bias the participant's decision towards not detecting the target. Therefore, the revcorr analysis can reveal the patterns of fluctuation critical for the decision if they are phase-locked to the onset of the stimulus.

As expected, general temporal kernels measured in the present AM-detection task revealed that 4-Hz noise intrinsic envelope fluctuations with phase $\approx\pi$ are likely to be confused with the AM target (Fig. 3, upper panel). This is consistent with the conclusions from Ardoint *et al.* (2007), although the present results were more homogeneous, probably because of the use of a lower SNR level. More surprisingly, however, the contour of the average temporal kernel did not exactly match that of the modulation to be detected. This suggests that the most efficient masker in this 4-Hz AM detection task would not be exactly a 4-Hz AM, contrary to the findings of Viemeister (1977) using the so-called "temporal-probe" method based on the masking patterns produced by an AM noise on a click target. More precisely, the kernels in Fig. 3 deviate from the ideal template (thick red line) in several ways. First, they show a clear temporal decay, that is, an increased weighting of the first few hundred milliseconds of a sound compared to later temporal segments. Similar primacy effects have often been reported in revcorr studies of loudness perception (Fischenich *et al.*, 2021; Ponsot *et al.*, 2016), lateralization (Brown and Stecker, 2010; Stecker, 2014), spectral processing (Richards *et al.*, 2021), and sample discrimination tasks (Berg, 1989), although the underlying mechanism is still debated. Other deviations from the ideal template include a peak frequency below 4 Hz, and a phase shift between the target-present and target-absent kernels. However, as for temporal decay, it is not clear which auditory processes may explain these observations. In line with previous studies (Ponsot *et al.*, 2021), we relied on computational auditory modeling tools to identify the types of nonlinearities involved in the decision process which may explain the structure of the observed kernels. In particular, we focused our investigation on late nonlinearities, that is, decision rules—although the MFB front-end also includes other nonlinearities such as amplitude compression.

Perceptual filters derived from human data were compared with those simulated from a simplified version of the model described by Dau *et al.* (1997) using five different decision rules. Two of them (O-D and XC-D) relied on an internal template derived from the internal representation of the target stimulus. A recurring but untested hypothesis in the auditory modeling literature is that such templates are formed by the listener at the beginning of the experiment (i.e., in a suprathreshold condition) and then remain unchanged even when the target becomes harder to detect

(Dau *et al.*, 1997; Osses Vecchi and Kohlrausch, 2021). Here, we tested this assumption by comparing the fixed-template models O-D and XC-D to a version where the template was updated from trial to trial (O-Du and XC-Du). The comparison of the proportion of hits and correct rejections as a function of current target modulation depth with that of real listeners (Table I) revealed that the updating rule failed to account for an important aspect of the data: the fact that the proportion of correct responses in target-absent trials did not depend on the difficulty in target-present trials. This supports the idea that the internal template is generated once at a level well above threshold, as assumed from the earliest MFB modeling studies [see Dau *et al.* (1997), note 5]. Note that any other updating rule where the template at trial n depends on the discriminability of the target, such as the one described in Derleth and Dau (2000), is likely to yield the same pattern of behavior. Hence, real participants used a decision rule which was either fixed (i.e., calculated only once at the beginning of the experiment) or at least not directly dependent on the difficulty of the task.[5]

Following this macroscopic psychophysical analysis, we compared the behavior of models and humans at a microscopic level, that is, we explored if they were confused by the same features in the noise envelopes at the trial level. As for real participants, kernels were computed as the correlation between the vector of intrinsic envelope fluctuation presented in each trial and the corresponding response of the O-D, XC-D, and EPS-D models. Overall, the three decision devices yielded kernels with a strong 4 Hz component, similar to real participants. However, their exact time courses reflected the type of mechanisms included in their decision devices.

The O-D model, the typical decision device for the model developed by Dau *et al.* (1997), realizes an optimal detector acting on the internal representation of the stimuli. All information available at this processing stage—in particular envelope phase information—is used in an optimal way resulting in O-D kernels that are well defined and similar across conditions and, importantly, in phase with the target modulation (Fig. 7). On the contrary, EPS-D implements a different decision strategy based on envelope spectrum power cues only and therefore discarding envelope phase information (Ewert and Dau, 2000). The most prominent feature in its kernel is a strong weighting of the stimulus onset. Indeed, any instance of the noise containing more energy in the first 100 ms of the stimulus will lead to a strong ringing effect in the modulation filter which will be confused with the presence of a target. In contrast, the presence of additional noise energy in any other temporal region has a smaller effect on the envelope spectrum power and thus on the decision. This explains the particular shape of the EPS-D temporal kernel, with an enhancement of the onset region ($\sim$ 50 ms) which slightly precedes the first peak of the target modulation (125 ms) resulting in a strong phase shift relative to π and a high temporal decay ratio (Fig. 9). Finally, an interesting feature of the EPS-D is the difference between target-present and target-absent kernels (the former showing a 4 Hz component in phase with the target in addition to the onset effect). Indeed, the presence of a modulation in the signal reinforces the effect of intrinsic envelope fluctuations temporally aligned with the peaks of the modulation. Finally, we investigated a third decision device which lies somewhere in between O-D and EPS-D as it only partially preserves information about envelope phase. More precisely, XC-D includes a template encoding a perfect internal representation of the target. However, unlike O-D, the decision is based on a "max of cross correlation" corresponding to an uncertainty on the starting phase of the stimulus. As the template shows a strong 4-Hz component, this decision device actually performs a 4-Hz detection, making it in essence more similar to the EPS-D. Practically, XC-D shows the same pattern of behavior as EPS-D, as revealed by its kernels, although there is a difference in the strength of the onset effect relative to the 4-Hz component.

Measuring Pearson's correlation coefficient between observed and simulated kernels confirmed that XC-D model provided a close match to the human data, although the target-present condition was best accounted for by the O-D model. More specifically, the participants' mean target-absent kernel displayed an enhanced onset, as confirmed by the 4-Hz phase metrics. This reveals that, in target-absent trials, human listeners were likely to confuse the onset of the noise masker with the first peak of the modulation to be sought, similar to phase-discarding models. In target-present trials, however, this phenomenon was considerably reduced, indicating that the presence of a target modulation provided a temporal landmark for the detection, resulting in no phase confusion. A decision device based on a max of cross correlation rule (XC-D model) is able to capture this phase-locking to the target, as indicated by the high correlations in both target-present and target-absent trials.

The present microscopic analysis highlights a weakness of the O-D, the dominant approach for modeling low-rate AM perception tasks. Although macroscopic analyses usually assume that a detection mechanism with no phase uncertainty is sufficient for accounting for human behavior in fixed-phase AM detection experiments, here we show that its description of the processing of (random-phase) noise envelope fluctuation is not accurate. Indeed, only noise instances with intrinsic envelope fluctuations in phase with the target can mislead the O-D (leading to kernels in phase with the target in Fig. 9). In contrast, real participants and phase-uncertain models are also confused, to a certain extent, by intrinsic envelope fluctuations with a slight phase shift relative to the target, as revealed by their kernels. This stresses the importance of using accurate phase-processing mechanisms even in fixed-phase detection tasks. The temporal decay observed in all simulated kernels can be explained in terms of transient response characteristics of the modulation filter, as confirmed by a complementary analysis reported in supplementary saterials.[2] In real participants, damping may result from temporal asymmetry at different levels of the auditory system, including ringing at the output of the cochlear and modulation filters, short term adaptation

https://doi.org/10.1121/10.0009629

or memory effects (Irino and Patterson, 1996; Wallaert *et al.*, 2018). An alternative interpretation could be that the listeners' strategy may be increasingly inconsistent and noisier toward the end of the stimulus, resulting in lower perceptual weights or more inconsistent phase-locking to the envelope.

On a methodological side, the present study aimed at shedding light on the functioning and operability of the auditory revcorr approach using a steady white noise masker. Such maskers are useful when experimenters have no *a priori* information on the location of the cues to be sought—or do not want to rely on this knowledge (Varnet *et al.*, 2013). However, as pointed by Joosten *et al.* (2016), it is likely that the introduction of random perturbations in the modulation domain (i.e., modulated noise) would make the auditory kernel estimation more efficient than relying only on intrinsic envelope fluctuations. Further work will be dedicated to quantitatively assess this hypothesis.

V. CONCLUSION

In summary, the macroscopic and microscopic analysis of the influence of a white noise masker in a 4-Hz AM detection task allowed us to draw important conclusions about the decision rules used by humans:

(1) Random intrinsic envelope fluctuations in the steady noise masker have a *systematic* influence on the decision, namely, they can bias the participant's response towards one alternative or the other depending on their trial-specific configuration.
(2) Overall, the temporal kernels measured on 8 normal-hearing participants suggest that the pattern of intrinsic envelope fluctuation that is most likely to mislead the listeners in the task does not correspond exactly to a 4-Hz AM.
(3) In an experiment with no explicit feedback such as the present one, the proportion of correct responses in target-absent trials does not depend on the current modulation depth, suggesting that the internal representation of the target is not updated from trial to trial to track the difficulty of the task.
(4) The classic optimal detector (O-D) rule is not consistent with human data at a microscopic level because it does not account for the effect of intrinsic envelope fluctuations in the masker that are out of phase with the target AM.
(5) The decision device based on a max of cross-correlation rule (XC-D) implementing an intermediate form of envelope phase uncertainty between O-D and EPS-D gives the best (although not perfect) match with the real listeners' data.

ACKNOWLEDGMENTS

The authors would like to warmly thank Alejandro Osses for his valuable comments on an earlier version of the manuscript, as well as Lucie Pierquin, Charlotte Delagnes, and François Trublereau for their help with designing the experiment. This study was funded by the ANR grant "fastACI" attributed to L.V. (Grant No. ANR-20-CE28-0004). Both authors were also supported by Grant No. ANR-17-EURE-0017.

[1]This phenomenon has been referred to under different names in the psychoacoustics and psycholinguistics literature, including "inherent fluctuations" Dau *et al.* (1997), "intrinsic fluctuations" Dau *et al.* (1999), "spurious modulations" (Drullman, 1995), and "stochastic modulations" (Noordhoek and Drullman, 1997). In the following, we will use the term "intrinsic envelope fluctuations."

[2]See supplementary materials at E-JASMAN-151-07268 for links to all data and scripts used in this study, supplementary figures, and a complementary simulation.

[3]We do not consider the possibility that listeners used "off frequency" cues for AM detection (Moore and Sek, 1994) because SNR in cochlear channels tuned above and below was extremely poor. As a matter of fact, the same revcorr analysis carried on two cochlear channels centered on 1000 Hz + 1 ERB and 1000 Hz − 1 ERB yielded no significant result.

[4]Some additional tests, not reported here, were performed to ensure that the exact value of the adaptation parameter (cut-off frequency) did not critically affect the model's kernel. As a matter of fact, even removing the adaptation module did not substantially alter the results.

[5]As pointed out by an anonymous reviewer, it is possible that this conclusion depends on the experimental design used in this study. In particular, the absence of any explicit feedback during the task provides less opportunity for subjects to update their internal templates. Therefore, adding an explicit feedback at the end of each trial could help listeners improve their listening strategy over the course of the experiment. Consistent with this idea, the revcorr study by Joosten *et al.* (2016), which included an explicit feedback, demonstrated a change of kernels between the first and the second half the experiment.

Ahumada, A. J., Jr. (**1996**). "Perceptual classification images from vernier acuity masked by noise," in *ECVP'96 Abstracts*.
Ahumada, A. J., Jr., and Lovell, J. (**1971**). "Stimulus features in signal detection," J. Acoust. Soc. Am. **49**(6B), 1751–1756.
Ahumada, A. J., Jr., Marken, R., and Sandusky, A. (**1975**). "Time and frequency analyses of auditory signal detection," J. Acoust. Soc. Am. **57**(2), 385–390.
Ardoint, M., Mamassian, P., and Lorenzi, C. (**2007**). "Internal representation of amplitude modulation revealed by reverse correlation," in *Abstract AR0 n 919, 30th ARO Midwinter Meeting*, February 10–15, Denver, CO.
Berg, B. G. (**1989**). "Analysis of weights in multiple observation tasks," J. Acoust. Soc. Am. **86**(5), 1743–1746.
Brown, A. D., and Stecker, G. C. (**2010**). "Temporal weighting of interaural time and level differences in high-rate click trains," J. Acoust. Soc. Am. **128**(1), 332–341.
Cabrera, L., Varnet, L., Buss, E., Rosen, S., and Lorenzi, C. (**2019**). "Development of temporal auditory processing in childhood: Changes in efficiency rather than temporal-modulation selectivity," J. Acoust. Soc. Am. **146**(4), 2415–2429.
Dau, T. (**1999**). *Modeling Auditory Processing of Amplitude Modulation* (BIS Verlag, Oldenburg).
Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997**). "Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration," J. Acoust. Soc. Am. **102**(5), 2906–2919.
Dau, T., Püschel, D., and Kohlrausch, A. (**1996a**). "A quantitative model of the 'effective' signal processing in the auditory system. I. Model structure," J. Acoust. Soc. Am. **99**(6), 3615–3622.
Dau, T., Püschel, D., and Kohlrausch, A. (**1996b**). "A quantitative model of the 'effective' signal processing in the auditory system. II. Simulations and measurements," J. Acoust. Soc. Am. **99**(6), 3623–3631.
Dau, T., Verhey, J., and Kohlrausch, A. (**1999**). "Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers," J. Acoust. Soc. Am. **106**(5), 2752–2760.
Derleth, R. P., and Dau, T. (**2000**). "On the role of envelope fluctuation processing in spectral masking," J. Acoust. Soc. Am. **108**(1), 285–296.

Drullman, R. (**1995**). "Temporal envelope and fine structure cues for speech intelligibility," J. Acoust. Soc. Am. **97**(1), 585–592.
Ewert, S. D., and Dau, T. (**2000**). "Characterizing frequency selectivity for envelope fluctuations," J. Acoust. Soc. Am. **108**(3), 1181–1196.
Ewert, S. D., Verhey, J. L., and Dau, T. (**2002**). "Spectro-temporal processing in the envelope-frequency domain," J. Acoust. Soc. Am. **112**(6), 2921–2931.
Fischenich, A., Hots, J., Verhey, J., and Oberfeld, D. (**2021**). "Temporal loudness weights are frequency specific," Front. Psychol. **12**, 588571.
Gilkey, R. H., and Robinson, D. E. (**1986**). "Models of auditory masking: A molecular psychophysical approach," J. Acoust. Soc. Am. **79**(5), 1499–1510.
Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**(1–2), 103–138.
Green, D. M. (**1964**). "Consistency of auditory detection judgments," Psychol. Rev. **71**(5), 392–407.
Hartmann, W. M. (**2004**). *Signals, Sound, and Sensation* (Springer Science & Business Media, New York).
Hilkhuysen, G., and Macherey, O. (**2014**). "Optimizing pulse-spreading harmonic complexes to minimize intrinsic modulations after auditory filtering," J. Acoust. Soc. Am. **136**(3), 1281–1294.
Hohmann, V. (**2002**). "Frequency analysis and synthesis using a Gammatone filterbank," Acta Acust. Acust. **88**(3), 433–442.
Irino, T., and Patterson, R. D. (**1996**). "Temporal asymmetry in the auditory system," J. Acoust. Soc. Am. **99**(4), 2316–2331.
Joosten, E. R. M., Shamma, S. A., Lorenzi, C., and Neri, P. (**2016**). "Dynamic Reweighting of Auditory Modulation Filters," PLoS Comput. Biol. **12**(7), e1005019.
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**(3), 1475–1487.
King, A., Varnet, L., and Lorenzi, C. (**2019**). "Accounting for masking of frequency modulation by amplitude modulation with the modulation filter-bank concept," J. Acoust. Soc. Am. **145**(4), 2277–2293.
Kohlrausch, A., Fassel, R., Heijden, M. V. D., Kortekaas, R., Par, S. V. D., Oxenham, A. J., and Püschel, D. (**1997**). "Detection of tones in low-noise noise: Further evidence for the role of envelope fluctuations," Acustica **83**(4), 659–669.
Lawson, J. L., and Uhlenbeck, G. E. (**1950**). *Threshold Signals* (Dover Publications, New York).
Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**(2), 467–477.
Lorenzi, C., Simpson, M. I., Millman, R. E., Griffiths, T. D., Woods, W. P., Rees, A., and Green, G. G. (**2001**). "Second-order modulation detection thresholds for pure-tone and narrow-band noise carriers," J. Acoust. Soc. Am. **110**(5 Pt 1), 2470–2478.
Majdak, P., Hollomey, C., and Baumgartner, R. (**2021**). "AMT 1.0: The toolbox for reproducible research in auditory modeling," https://www.amtoolbox.org/ (Last viewed 2/16/2022).
Moore, B. C., and Sek, A. (**1994**). "Effects of carrier frequency and background noise on the detection of mixed modulation," J. Acoust. Soc. Am. **96**, 741–751.
Murray, R. F. (**2011**). "Classification images: A review," J. Vision **11**(5), 1–25.
Noordhoek, I. M., and Drullman, R. (**1997**). "Effect of reducing temporal intensity modulations on sentence intelligibility," J. Acoust. Soc. Am. **101**(1), 498–502.
Osses Vecchi, A., and Kohlrausch, A. (**2021**). "Perceptual similarity between piano notes: Simulations with a template-based perception model," J. Acoust. Soc. Am. **149**(5), 3534–3552.
Ponsot, E., Susini, P., and Oberfeld, D. (**2016**). "Temporal weighting of loudness: Comparison between two different psychophysical tasks," J. Acoust. Soc. Am. **139**, 406–417.
Ponsot, E., Varnet, L., Wallaert, N., Daoud, E., Shamma, S. A., Lorenzi, C., and Neri, P. (**2021**). "Mechanisms of spectrotemporal modulation detection for normal- and hearing-impaired listeners," Trends Hear. **25**, 233121652097802.
Rice, S. O. (**1944**). "Mathematical Analysis of Random Noise," Bell Syst. Tech. J. **23**(3), 282–332.
Richards, V. M., Tisby, M. K., Suzuki-Gill, E. N., and Shen, Y. (**2021**). "Sub-optimal construction of an auditory profile from temporally distributed spectral information," J. Acoust. Soc. Am. **149**(3), 1567–1578.
Schönfelder, V. H., and Wichmann, F. A. (**2013**). "Identification of stimulus cues in narrow-band tone-in-noise detection using sparse observer models," J. Acoust. Soc. Am. **134**(1), 447–463.
Shub, D. E., and Richards, V. M. (**2009**). "Psychophysical spectro-temporal receptive fields in an auditory task," Hear. Res. **251**(1-2), 1–9.
Stecker, G. C. (**2014**). "Temporal weighting functions for interaural time and level differences. IV. Effects of carrier frequency," J. Acoust. Soc. Am. **136**(6), 3221–3232.
Stone, M. A., Füllgrabe, C., Mackinnon, R. C., and Moore, B. C. J. (**2011**). "The importance for speech intelligibility of random fluctuations in 'steady' background noise," J. Acoust. Soc. Am. **130**(5), 2874–2881.
Stone, M. A., Füllgrabe, C., and Moore, B. C. J. (**2012**). "Notionally steady background noise acts primarily as a modulation masker of speech," J. Acoust. Soc. Am. **132**(1), 317–326.
Strickland, E. A., and Viemeister, N. F. (**1996**). "Cues for discrimination of envelopes," J. Acoust. Soc. Am. **99**(6), 3638–3646.
Varnet, L., Knoblauch, K., Meunier, F., and Hoen, M. (**2013**). "Using auditory classification images for the identification of fine acoustic cues used in speech perception," Front. Human Neurosci. **7**, 1–12.
Varnet, L., Knoblauch, K., Serniclaes, W., Meunier, F., and Hoen, M. (**2015**). "A psychophysical imaging method evidencing auditory cue extraction during speech perception: A group analysis of auditory classification images," PLoS One **10**(3), e0118009.
Vecchi, A., Varnet, L., Carney, L. H., Dau, T., Bruce, I. C., Verhulst, S., and Majdak, P. (**2021**). "A comparative study of eight human auditory models of monaural processing," arXiv:2107.01753.
Venezia, J. H., Hickok, G., and Richards, V. M. (**2016**). "Auditory 'bubbles': Efficient classification of the spectrotemporal modulations essential for speech intelligibility," J. Acoust. Soc. Am. **140**(2), 1072–1088.
Viemeister, N. F. (**1977**). "Temporal factors in audition: A systems analysis approach," in *Psychophysics and Physiology of Hearing*, edited by E. F. Evans and J. R. Wilson (Academic, London), pp. 419–427.
Wallaert, N., Moore, B. C. J., Ewert, S. D., and Lorenzi, C. (**2017**). "Sensorineural hearing loss enhances auditory sensitivity and temporal integration for amplitude modulation," J. Acoust. Soc. Am. **141**(2), 971–980.
Wallaert, N., Varnet, L., Moore, B. C. J., and Lorenzi, C. (**2018**). "Sensorineural hearing loss impairs sensitivity but spares temporal integration for detection of frequency modulation," J. Acoust. Soc. Am. **144**(2), 720–733.