JASA ARTICLE

The peak height insertion gain (PHIG) method for quantifying acoustic feedback in hearing aids

Joshua M. Alexander,[1,a)] Stephanie Schmig,[1] Randall P. Wagner,[2] and Steve Armstrong[3]
[1]*Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, Indiana 47907, USA*
[2]*Physical Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA*
[3]*SoundsGood Labs, Burlington, Ontario, Canada*

ABSTRACT:
Hearing aids are commonly fit with ear canals partially or fully open—a condition that increases the risk of acoustic feedback. Feedback limits the audiometric fitting range of devices by limiting usable gain. To guide clinical decision making and device selection, we developed the Peak Height Insertion Gain (PHIG) method to detect feedback spikes in the short-term insertion gain derived from audio recordings. Using a manikin, 145 audio recordings of a speech signal were obtained from seven hearing aids. Each hearing aid was programmed for a moderate high-frequency hearing loss with systematic variations in frequency response, gain, and feedback suppression; this created audio recordings that varied the presence and strength of feedback. Using subjective ratings from 13 expert judges, the presence of feedback was determined and then classified according to its temporal and tonal qualities. These classifications were used to optimize parameters for two versions of the PHIG method based on global and local analyses. When specificity was fixed at 0.95, the sensitivity of the global analysis was 0.86 and increased to 0.95 when combined with the local analysis. Without compromising performance, a clinically expedient version of the PHIG method can be obtained using only a single measurement.
https://doi.org/10.1121/10.0005987

(Received 22 January 2021; revised 12 July 2021; accepted 5 August 2021; published online 7 September 2021)
[Editor: James F. Lynch] Pages: 1635–1651

I. INTRODUCTION

Acoustic feedback occurs when the amplified sound from a hearing aid returns to the microphone and is subsequently re-amplified. If this process continues, the output of the amplifier continues to increase until the saturation level of the hearing aid is reached. High-intensity acoustic feedback is often described as a squealing, whistling, screeching, or howling sound that may cause a patient to experience discomfort and/or embarrassment. For these reasons, acoustic feedback can be a limiting factor that influences the style, technology level, and/or model of hearing aid that a clinician recommends for a patient. Therefore, it is important to develop a quick and simple method for detecting feedback that can be used in a clinical setting as well as in a formal, standardized technical specification. The purpose of this paper is to describe and evaluate a non-intrusive, objective measure for detecting the presence of acoustic feedback in signals amplified by a hearing aid.

A. Acoustic feedback characteristics

Acoustic features that indicate the presence of feedback are a by-product of the conditions that create it. As described by MacIntosh and Cornelisse (2018), the conditions for feedback depend on how the amplitudes and phases of the frequency components in the input signal from the sound source and the output signal from the hearing aid interact at the location of the microphone. The amplitude of the output signal when it returns to the microphone is determined by the gain of the feedforward path and the attenuation of the feedback path. The gain of the feedforward path includes the gain provided by the hearing aid and the gain provided by the ear canal resonance characteristics; these two combine to form the real-ear aided gain. The attenuation of the feedback path is determined by how much the physical fit of the hearing aid blocks the acoustic path from the ear canal to the hearing aid microphone. When the output signal returns to the microphone, its phase is determined by the time it takes for the input signal to be digitally processed and amplified by the hearing aid combined with the propagation time along the feedback path. Frequencies with periods that are integer multiples of this delay time will add in phase while intermediate frequencies will add out of phase. As the gain of the input signal through the feedforward path from the microphone to the ear canal begins to approximate the attenuation of the output signal through the feedback path, this alternating pattern of constructive and destructive interference can create a rippled frequency response in the amplified signal.

The amount of interaction between the input and output signals influences the acoustic characteristics; hence, the perceptual attributes, of the feedback. Based on the amount

[a)]Electronic mail: alexan14@Purdue.edu, ORCID: 0000-0002-4292-7109.

of interaction between these signals, MacIntosh and Cornelisse (2018) have identified three categories of feedback. Two of these categories have been historically labeled "sub-oscillatory" and "oscillatory" (e.g., Cox, 1982; Agnew, 1996). These terms refer to the cyclical re-amplification of the feedback signal when the amplitude of the in-phase output signal exceeds the amplitude of the input signal. In other words, oscillatory feedback occurs when the real-ear aided gain exceeds the attenuation of the feedback path. "Audible feedback" and "tonal feedback" are synonymous with oscillatory feedback because the hearing aid output is dominated by feedback at one or more discrete frequencies. This type of feedback is often observable as narrow peaks or spikes in the output signal spectrum; hence, these frequencies are sometimes referred to as "feedback spikes".

MacIntosh and Cornelisse (2018) also identify an "intermittent" feedback category, which occurs when the real-ear aided gain is very close to the attenuation of the feedback path. In this scenario, there are brief moments when oscillatory feedback is heard as "chirps" or "squeaks" due to changes in the feedback path, such as reflective objects that re-direct the sound escaping from the ear canal back to the hearing aid microphone. Feedback can be problematic even if it is not audible. Sub-oscillatory feedback occurs when the real-ear aided gain is a few dB less than the attenuation of the feedback path. In this scenario, the amplitudes of the in-phase and out-of-phase frequency components of the output signal at the microphone are high enough to interact with the input signal, resulting in a rippled frequency response. Sub-oscillatory feedback may be subtle and affect only the timbre and perceived sound quality, or it might cause a noticeable "ringing" or "echo" in the output that can degrade speech intelligibility (Agnew, 1996).

In summary, acoustic feedback occurs only at frequencies in the output signal that are in phase with the direct sound source and only when the real-ear aided gain of these frequencies exceeds the attenuation of the propagation path from the ear canal to the microphone. Consequently, various digital feedback management techniques have been developed over the years that attempt to eliminate feedback by altering the gain, frequency, and/or phase of the output in an attempt to decorrelate it with the input. While a means to characterize the efficacy of these techniques is the subject of the method discussed here, an in-depth discussion of the techniques is outside the scope of this paper (see Chung, 2004 for a review).

B. Non-intrusive, objective methods of feedback detection

Clinically, audiologists must make many decisions when selecting a hearing aid for their patients, such as brand, technology level, style, receiver power, and many advanced processing features. With open fittings, the ability to provide the prescribed gain without feedback is an especially important consideration. Every manufacturer has feedback management algorithms that attempt to prevent or minimize feedback; however, objective specifications about their effectiveness are lacking. Often, the only information provided to audiologists consists of shaded audiograms for each product that show the fitting range, which may be used to determine the appropriate product and/or receiver strength for the patient. But, again, there is little to no information given as to how these fitting ranges are derived.

Audiometric limits of the fitting range for a particular hearing aid can be inferred by quantifying the maximum stable gain before feedback. Several studies have used human listeners to subjectively determine when a hearing aid goes into feedback as gain is increased (e.g., Ricketts *et al.*, 2008; Marcum and Ricketts, 2011). However, to develop a standardized procedure for quantifying maximum stable gain using a manikin, it is necessary to have an objective method for determining feedback using realistic signals, hearing aid settings, and coupling to the ear. Furthermore, it is desirable if the objective method can be applied *in situ* in the clinic with existing probe microphone equipment because, as noted by Ricketts *et al.* (2008), there are likely to be differences—not only between manikin and human ears, but also between individuals due to differences in: (1) how much sound leaks out of the ear (e.g., due to the coupling of the eartip or earmold in the ear canal); (2) the level of the signal in the ear canal (e.g., due to ear canal impedance and/or geometry); (3) the level of the leaked output signal to the microphone (e.g., due to reflective surfaces). For the objective method to be clinically feasible, it needs to be non-intrusive, which means that it only needs the hearing aid output signal to determine whether the patient is at risk of experiencing oscillatory or sub-oscillatory feedback. Additionally, a feedback-free reference signal, which can be readily determined for a known system, is not available for comparison in this situation. This reference would indicate what the output signal should be in the absence of feedback after the input is amplified by the same gain function and acted on by the nonlinear processing in the hearing aid.

Investigators using objective methods of feedback detection who do not have access to the input signal at the microphone's location must treat the hearing aid as an unknown system (Spriet *et al.*, 2010; Madhu *et al.*, 2011). To simplify the treatment of the hearing aid as an unknown system, users of previous methods for detecting the presence of feedback in the hearing aid output have had to make compromises, they: (1) used stationary input signals; (2) made assumptions about the hearing aid processing; (3) programmed the hearing aid in a way that does not represent how a patient would typically use it; and/or (4) developed a procedure for estimating a feedback-free reference signal from a measurement implemented as a type of requisite reference gain. Table I provides a comparison of the key features of the different methods that have been proposed over the years. For reference, the PHIG method, which is the focus of this paper, is provided in the last row.

The method of feedback detection proposed by Merks and colleagues (Merks *et al.*, 2006; Merks, 2010) compares three measurements of the impulse response for a white noise stimulus at different gain settings. The first measurement is

JASA
https://doi.org/10.1121/10.0005987

TABLE I. A summary of different methods that have been proposed for objectively detecting acoustic feedback in hearing aids. (See the text for citations of the authors.) The last row refers to the method that is being proposed in this paper.

Authors	Method	Input signals	Requisite hearing aid settings	Requisite reference gain	Perceptual basis of method	Notes
Merks *et al.* (2006); Merks (2010)	Impulse response	White noise	No nonlinear elements, including gain	Lower gain setting that is feedback-free	Informal testing by authors	
Shin *et al.* (2007)	Transfer Function Variation Criterion (TVC)	White noise	No nonlinear elements, including gain	Lower gain setting that is feedback-free	Informal listening by authors to set TVC for feedback categories	Performed using a 2 cc coupler in a test-box
Freed and Soli (2006)	Power Concentration Ratio (PCR)	White noise	Disable noise reduction	None	Informal listening by authors to set PCR threshold	
Spriet *et al.* (2008; 2009; 2010); Madhu *et al.* (2011)	Modified TVC and PCR, others	Any	None	Same gain setting but with an occluded fitting	None	Requires estimate of environment / internal noise
Proposed method	Peak Height Insertion Gain (PHIG)	Any	None	None	13 expert judges	A single open-ear response is used for insertion gain

obtained with the hearing aid in the ear but muted or turned off; that is, the real-ear occluded response (REOR). The other two measurements are of the real-ear aided responses (REARs) obtained with the hearing aid gain set low enough so that the output is free from feedback and with the gain set high enough to possibly induce feedback. The impulse response from the REOR is then subtracted from the impulse responses from the two REARs to create two modified impulse responses. The modified impulse response of the low-gain, feedback-free measurement is scaled by the gain difference between the two REARs to create a reference signal. The reference signal is the response that would be expected if the same high-gain settings were applied to the input signal but without feedback. Next, the reference signal is subtracted from the modified impulse response of the high-gain measurement, which is then used to estimate the time segments of the high-gain output signal associated with the acoustic leakage of the fitting, the hearing aid amplification, and the feedback. These time segments are converted back into the frequency domain to compute a gain margin. The caveats with this method are that the input signal needs to be stationary and spectrally flat and the hearing aid processing must be entirely linear, including the gain; otherwise, the reference signal will not be an accurate estimate of the feedback-free, high-gain output.

The method of feedback detection proposed by Shin *et al.* (2007), the Transfer Function Variation Criterion (TVC) is conceptually similar to the method proposed by Merks (2010), except that it operates in the frequency domain. The TVC uses a Wiener filter to estimate the transfer function of the hearing aid by comparing the output signal to a white noise input signal. It is similar to Merks' method because it compares the transfer function of the hearing aid when in a state of possible feedback to the transfer function when it is in a known stable state. Just as with Merks' method, the transfer function in the stable state is scaled by the gain difference between the two hearing aid settings; therefore, an accurate estimate of the feedback-free, high-gain transfer function requires the hearing aid to be completely linear. Furthermore, to control error between measurements, they are conducted with the hearing aid connected to a 2 cm^3 coupler in a test box. A unique aspect of the TVC is that it measures the spectral ripple caused by the alternating pattern of constructive and destructive interference of the input signal by the feedback signal. Therefore, it is reportedly sensitive to the effects of sub-oscillatory feedback. More specifically, the TVC is the maximum peak or valley of the spectral ripple across frequency. TVC values that are greater than or equal to 5 dB and less than 10 dB indicate that the hearing aid is likely in a sub-oscillatory feedback state. TVC values that are greater than 10 dB indicate that the hearing aid is likely in an unstable/oscillatory feedback state.

The method of feedback detection proposed by Freed and Soli (2006), the Power Concentration Ratio (PCR), differs from the other two methods in that it operates only on the output signal from the hearing aid and does not rely on a reference signal for comparison. This feature offers at least three advantages: (1) only one measurement needs to be made; (2) it does not require linear processing in the hearing aid; and (3) it is computationally simple. To compute the PCR, the five frequency bins with the highest power are identified from a Fast Fourier Transform (FFT) on the hearing aid output, then the ratio of their combined power to the total power of the output is calculated. One caveat with this method is that it uses a white noise input signal—which is spectrally flat and stationary—so that spectral peaks in the input are not confounded with the spectral peaks associated with oscillatory feedback. Also, since a white noise test signal is the most favorable input signal for getting adaptive feedback cancellation to converge to an accurate model of the feedback path, this test signal would be one of the least challenging for performing such testing. Finally, in addition to the limited real-world applicability associated with using white noise for an input signal, noise reduction features need to be disabled to prevent them from reducing the gain of the test signal.

Spriet and colleagues (Spriet *et al.*, 2008, 2009, 2010; Madhu *et al.*, 2011) proposed several variants of the TVC and PCR methods to overcome limitations associated with the need to use a white noise input signal and the need to program the hearing aid with completely linear processing. The primary way they could circumvent these restrictions was to develop a new approach for estimating a feedback-free reference signal. Instead of relying on gain reduction to eliminate feedback, they changed from an open fitting for the test signal to an occluded fitting for the reference signal. Then, they applied a filter to the occluded measurement, after accounting for the REOR, to compensate for the difference in the frequency response when going from the open to the occluded fitting. This way, the hearing aid could have identical settings for the test and reference measurements, thereby eliminating the need for linear hearing aid processing or a specific input signal.

Another modification to the TVC and PCR methods introduced by Spriet and colleagues (Spriet *et al.*, 2008, 2009, 2010; Madhu *et al.*, 2011) is that they used a time-segmented analysis; instead of averaging across time, they examined the TVC and PCR in individual time-frequency bins. First, only time-frequency bins that exceed the environmental and internal hearing aid noise by 10 dB are considered for further analysis. Then, the TVC is computed as a function of time by finding the largest difference between transfer functions across frequency at each time segment. The threshold for instability is increased from 5 dB (Shin *et al.*, 2007) to 20 dB, presumably due to short-time analysis leading to more "noise" in the test signals free from audible feedback, that is, spurious fluctuations that otherwise would be averaged out over time. One limitation with this method is that the exact frequencies responsible for the feedback cannot be identified when TVC is expressed as a function of time. The modified PCR method dovetails from the TVC method. First, only time-frequency bins where the difference between transfer functions exceeds 6 dB are considered for further analysis to help minimize errors associated with non-stationary input signals, such as speech. Then, the PCR of the test signal is computed as a function of time after taking into account the PCR of the reference signal, which could contain spectral peaks that originate from the input signal and/or hearing aid processing. One major disadvantage of the methods proposed by Spriet and colleagues (Spriet *et al.*, 2008, 2009, 2010; Madhu *et al.*, 2011) is the complexity due to the need to make additional measurements for the reference signal and environmental/internal noise and the need to derive a compensation filter.

From the literature review and summary in Table I, it is clear that previous methods are limited because they have: (1) restrictions on the type of input signal that can be used for testing; (2) restrictions on the signal processing in the hearing aid; and/or (3) complexity due to additional measurements and computational modeling. Despite these limitations, the primary concern for implementing them in clinical practice or a standardized technical specification is that none of them have been established using formal listening studies done with a group of trained, independent, expert judges/listeners. That is, criteria have been established informally for measurement values that specify when a hearing aid is in oscillatory feedback and some for when a hearing aid is in sub-oscillatory feedback. Furthermore, there is no indication that the parameters used for signal analysis—such as the sampling rate, the FFT length, and the resultant frequency bin width—were based on an optimization procedure.

On the other hand, the Peak Height Insertion Gain (PHIG) method presented herein for detecting feedback: (1) is computationally simple, (2) has been developed using an optimization procedure, and (3) can be applied using realistic signals and realistic use conditions. The PHIG method is designed to find peaks in the output signal after accounting for the frequency-shaping of a speech signal provided by the open ear. It is desirable to remove the open-ear filter properties from the analysis because peaks in the real-ear aided gain (REAG) can create or enhance peaks in a feedback-free signal, and valleys in the REAG can attenuate peaks in a feedback signal. This would confound the optimization procedure, which segregates feedback-free signals from feedback signals using the relative height of the peaks in the output. Therefore, the open-ear filter properties are removed by computing the real-ear insertion gain, which is the difference between the real-ear aided response and the real-ear unaided response for the same source signal.

Instead of comparing the measured gain to a feedback-free reference gain to identify feedback in the frequency-shaped output, the PHIG method relies on the spectral ripple generated by the feedback signal in the insertion gain spectrum. The PHIG method does this by comparing the insertion gain of each frequency component (in time) to the insertion gain of its neighboring frequency components ("sidebands"). In this way, the PHIG method is also sensitive to sub-oscillatory feedback. The sensitivity of the PHIG method was established using a group of independent judges who rated the presence of feedback in 145 recordings generated by seven different hearing aids in a plurality of conditions. The judges also rated the temporal characteristics of the perceived feedback to create different categories of feedback that ranged in severity. The sensitivity of the PHIG method across the different categories of feedback was then used to optimize the parameters used for signal analysis.

II. METHODS[1]

A. Initial device programming

The makes and models of the receiver-in-the-canal (RIC) hearing aids used in this study are shown in Table II.[2] For initial programming, the hearing aids' digital feedback suppression (DFS) and other advanced features (e.g., digital noise reduction, automatic adaptation, directionality) were disabled. The gain of each hearing aid was programmed so that the output levels closely matched the prescriptive targets of the

TABLE II. Hearing aid makes and models used in this study.

Make	Model
Oticon (Smørum, Denmark)	OPN 1
Phonak (Stäfa, Switzerland)	Audeo B90
ReSound (Ballerup, Denmark)	LiNX2 9
Signia (Lynge, Denmark)	Pure Binax 7b×
Starkey (Eden Prarie, MN)	Halo 2 i2400
Unitron (Kitchener, ON, Canada)	Flex:Trial (set to TPro)
Widex (Lynge, Denmark)	Beyond 440

Desired Sensation Level (DSL) v5.0-adult method (Scollie *et al.*, 2005) for speech presented at SPLs of 55, 65, and 75 dB and for tones presented at an SPL of 90 dB. The DSL prescriptive method was selected over other prescriptive methods because it usually prescribes more high-frequency gain for a moderate high-frequency hearing loss (Johnson and Dillon, 2011), thereby increasing the likelihood of feedback. DSL prescriptive targets were generated for the N3 standard audiogram (Bisgaard and Ruff, 2017), which is given in Table III and represents a moderate high-frequency hearing loss commonly treated with a RIC hearing aid and a fully or partially open ear canal. Each hearing aid was programmed in the test box of an Etymonic Design Inc. Audioscan Verifit 1 (Dorchester, ON, Canada). Once a hearing aid was programmed to prescriptive targets, its settings were saved for later retrieval. These settings constitute the "baseline program."

B. Audio recordings

After a hearing aid was programmed in the test box, the DFS was turned back (the other advanced features were left off) and it was mounted with a medium-sized open dome on the right ear of a Knowles Electronics Manikin for Acoustics Research (KEMAR). The KEMAR was equipped with a Knowles Electronics DB-100 Zwislocki Ear Simulator (Itasca, IL) connected to the microphone and amplifier of an Etymōtic Research ER-11 Microphone System (Elk Grove Village, IL) for KEMAR. The KEMAR was centered in a double-walled sound booth measuring 2.13 m by 2.44 m and was positioned 1 m away at 0° azimuth from a Hafler M5 Reference Monitor (Rockford Corp., Tempe, AZ) sound field loudspeaker. The loudspeaker was used to present a standardized speech stimulus at an SPL of 55 dB when measured directly above the KEMAR pinna—the position of the hearing aid microphones. This level is within the range of speech spoken with a soft vocal effort (Scollie *et al.*, 2005); it was chosen because hearing aids with nonlinear amplification apply the most gain for low input levels, thereby increasing the likelihood of feedback.

TABLE III. Audiometric thresholds of the N3 audiogram used for generating prescriptive targets. The threshold at 8000 Hz was extrapolated to extend the slope between the 4000 Hz and 6000 Hz thresholds.

Frequency (Hz)	250	375	500	750	1000	1500	2000	3000	4000	6000	8000
Threshold (dB HL)	35	35	35	35	40	45	50	55	60	65	70

The speech stimulus consisted of two concatenations of Etymonic Design Inc.'s (Dorchester, ON, Canada) 15 s "ear passage" spoken by a male talker, which were preceded and followed by 0.5 s of silence. The first half of the stimulus was used to engage the DFS algorithms and give them time to stabilize; therefore, only the second half of each recording was used to obtain the subjective ratings and optimize the PHIG method's parameters. Audacity open-source software v2.2.2 (Audacity Team, 2018) was used to simultaneously present and record the speech stimulus at its native sampling and bit rates (32 kHz, 24 bits). Using circumaural headphones, the experimenter monitored the output of the ER-11 system from outside the booth by splitting the signal and routing it to an amplifier.

For each hearing aid, the first set of recordings was made with the baseline program (DFS "off"). If the experimenter detected audible feedback, the hearing aid gain across all frequencies was decreased in the programming software until it could no longer be detected. Then, the mid-frequency gain (1–3 kHz) was increased until feedback was audible. This gain setting was used to make the first recording. To create conditions that varied in the strength and quality of oscillatory or sub-oscillatory feedback, up to three subsequent recordings were made by decreasing the mid-frequency gain by one step in the programming software. If the programming software had an option for the step size of the gain adjustments (e.g., 1, 3, or 5 dB), the smallest step size was used. Although not independently verified, gain adjustments most often corresponded to the programming software label "1 dB". If audible feedback could not be induced, a recording was made with the mid-frequency gain set to maximum. Similarly, if audible feedback could not be eliminated, a recording was made with the mid-frequency gain set to minimum. Starting with the baseline program, this process was repeated a second time for the high-frequency gain (3–8 kHz) and a third time for both mid- and high-frequencies combined. DFS was then activated in the programming software, and the entire process was repeated with each change in the DFS strength. Every manufacturer, except one, allowed DFS strength to be manipulated in the programming software. Recordings were obtained with DFS set to the minimum, medium, and maximum strengths. For each hearing aid, up to 48 recordings were possible: 4 DFS settings (including "off") times three frequency band adjustments (mid, high, mid+high) × 4 gain settings (gain to induce audible feedback, plus 3 steps below this value). Across the seven hearing aids, 336 recordings were possible; however, given the availability of DFS settings in the programming software and given that there were settings for which feedback could not be eliminated or induced (thereby, limiting gain to its minimum or maximum setting, respectively), only 145 recordings were obtained.

C. Subjective ratings

Custom scripts written in MathWorks (Natick, MA) MATLAB were used to prepare and then present the audio recordings for subjective ratings. First, the last 15.5 s of each recording was excised and resampled at 16 kHz since there was no energy above 8 kHz in the speech stimulus. The resampled recordings were then processed by a 32-tap finite impulse response filter to remove the frequency shaping caused by the KEMAR open-ear response—the real-ear unaided gain (REUG) (this is later shown by the dotted line in Fig. 7). This was done to avoid cascading the frequency shaping of the KEMAR ear with the frequency shaping of the judges' ears. Finally, the filtered stimuli were scaled to an SPL of 65 dB for both listening comfort and uniformity.

Using a Lynx Studio Technology, Inc. Lynx TWO-B (Costa Mesa, CA) sound card and Beyerdynamic DT150 (Heilbronn, Germany) circumaural headphone, the stimuli were presented in random order to the right ear of 14 expert judges with clinically normal hearing (≤15 dB HL at each octave frequency from 250 to 8000 Hz). The recordings were tentatively classified into one of seven categories based on the presence of feedback and the temporal and tonal qualities of the feedback. To help train the judges, two audio examples of each of the seven categories, along with their text descriptions, were presented before data collection. Using seven push-buttons on a touchscreen monitor, judges rated the stimuli according to the descriptions provided in Table IV. If a judge determined that a stimulus clearly fell in the Modulated, Weak Continuous Tone, or Strong Continuous Tone category, they could select the appropriate option before the stimulus ended and continue. If they thought a stimulus could fall in the No Feedback, Infrequent, Occasional, or Intermittent category, they were instructed to listen to the entire stimulus before rating it.

Despite the training, the ratings for one judge diverged by almost one category from the modal ratings of the other judges. The modal rating of the remaining 13 judges was used to classify each stimulus based on the categories provided during the experiment. The categories for "Occasional" and "Weak Tonal" were collapsed with other categories due to low counts as shown in Table IV, which resulted in a final total of five feedback categories: None, Infrequent, Intermittent, Modulated, and Tonal.

D. The Peak Height Insertion Gain (PHIG) method

An acoustic signature of feedback is a persistent peak in the insertion gain over a narrow frequency range relative to the gain of neighboring frequencies. Insertion gain accounts for the time-varying acoustics of the signal and the invariant filter effects associated with resonances of the external ear. Therefore, the PHIG method described below can be applied to realistic signals and use conditions, including an open canal configuration. Insertion gain in dB is defined as

$$G = 20\log\left(\frac{p_a}{p_u}\right), \tag{1}$$

where p_a is the sound pressure in the aided ear (or ear simulator) and p_u is the sound pressure in the unaided ear (or ear simulator). To help track feedback spikes over the duration of the stimulus, insertion gain was computed as a function of time and frequency using the discrete Short Time Fourier Transform (STFT) of the stimulus in the aided and unaided conditions.[3] Let $\mathcal{F}(x)$ be the discrete STFT of a time signal x, sampled with frequency f_s, windowed with a Hann window of size N_w, and with time segments overlapping 50%.[4] The result of the STFT is a matrix of complex values representing Fourier amplitudes, with each row representing a single time segment and each column representing the frequency bins of the STFT. Since a spectrogram is a plot of squared magnitudes of the transpose of the STFT, the spectrogram function in MATLAB was used to derive the Insertion Gain Matrix $\boldsymbol{G}$, which can be expressed as

$$\boldsymbol{G} = 10\log_{10}\left|\mathcal{F}(p_a)\right|^2 - 10\log_{10}\left|\mathcal{F}(p_u)\right|^2. \tag{2}$$

In this logarithmic form, insertion gain is expressed in decibels. Insertion gain values <0 dB were set to 0 because it resulted in slightly better performance metrics (see Sec. III A). Since 8000 Hz was the upper-frequency limit of the original signal, only the frequency bins of $\boldsymbol{G}$ representing center frequencies between 200 Hz and 8000 Hz were selected; all other frequency bins were discarded. The mathematical operations described herein are understood to apply to the elements of the STFT matrix, not the matrix itself. Therefore, $\boldsymbol{G}$ was computed by subtracting the squared magnitude of the unaided (open) ear signal response from the corresponding element in the matrix that represented the squared magnitude of the aided ear signal response.[5] The first column in Fig. 1 displays $\boldsymbol{G}$ for stimuli from each of the five feedback categories.

TABLE IV. Descriptions given to the judges for rating the stimuli using one of seven preliminary categories. The column labeled "Count" shows the number of stimuli that fell into each category. Due to low counts for the "occasional" and "weak continuous tone" categories, these stimuli were collapsed with the stimuli from an adjacent category to form the final category classifications shown in the last column.

Description	Preliminary Category	Count	Final Category
No feedback is audible throughout the entire duration	No Feedback	49	None
Feedback overtones or tones are brief and occur one to three times	Infrequent	16	Infrequent
Feedback overtones or tones are brief and occur four to six times	Occasional	6	
Feedback overtones or tones are present during the speech and essentially absent during the pauses	Intermittent	33	Intermittent
Feedback overtones or tones are present throughout, but the loudness varies as the level of speech changes	Modulated	18	Modulated
Feedback is tonal and present throughout, but the speech is generally louder	Weak Continuous Tone	4	Tonal
Feedback is tonal and dominates the signal	Strong Continuous Tone	19	

https://doi.org/10.1121/10.0005987

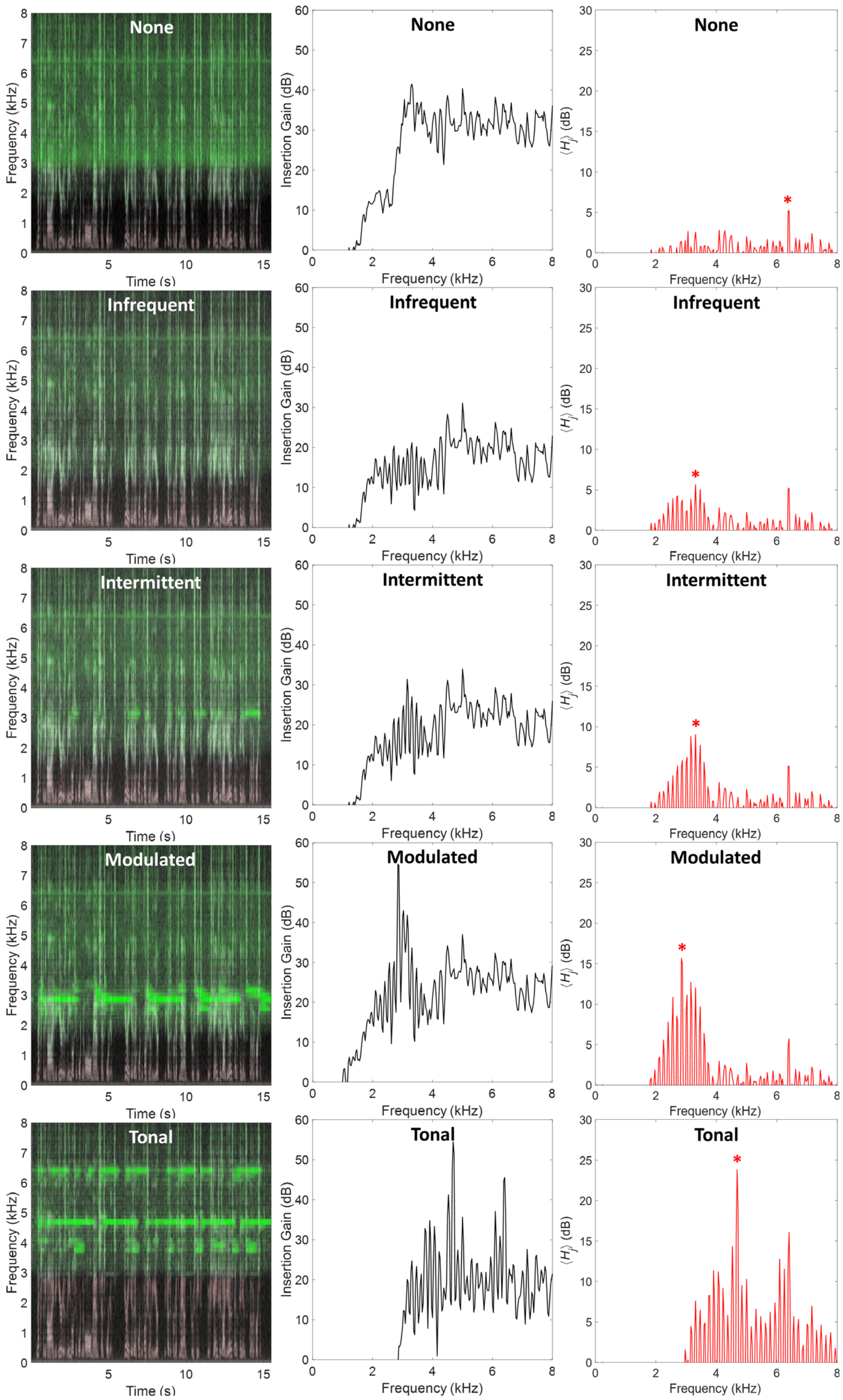

FIG. 1. (Color online) Shown in each row are representative stimuli from each of the five feedback categories. The examples are from the same hearing aid with different settings. (row 1): "None" with DFS set to maximum, (row 2): "Infrequent" with DFS set to medium and gain set to 2 steps down from audible oscillatory feedback, (row 3): "Intermittent" with DFS set to medium and gain set to 1 step down from audible oscillatory feedback, (row 4): "Modulated" with DFS set to minimum, (row 5): "Tonal" with DFS deactivated. The first column shows the insertion gain as a function of time and frequency (note that this is the transpose of $\boldsymbol{G}$). Positive insertion gain is shown in green and negative insertion gain in red. The second column shows the mean insertion gain across time as a function of frequency. The third column shows the mean of H_{ij} across time as a function of frequency $\langle H_j \rangle$. The maximum value of $\langle H_j \rangle$ is the Peak Height Insertion Gain P and is used to determine the presence of feedback. P is denoted by the asterisk on each plot in the last column.

The Insertion Gain Matrix $\boldsymbol{G}$ is an $m \times n$ matrix. The ith row represents the ith time segment, and the jth column represents the jth frequency bin. Feedback is characterized by a relatively high amount of gain in one or more columns corresponding to the frequency bins. Therefore, the gain of each time-frequency bin was compared to the average gain of the neighboring frequency bins to compute a variable, the local Peak Height Insertion Gain H_{ij}. H_{ij} was used for the detection and classification of feedback and was defined as the difference between the insertion gain of the jth frequency bin and the average insertion gain of the 2 r adjacent frequency bins (sidebands)

$$H_{ij} = G_{ij} - \frac{1}{2r}\sum_{k=1}^{r}(G_{ij-k} + G_{ij+k}). \qquad (3)$$

H_{ij} was defined only for frequency bins where $r < j \leq n - r$. Figure 2 demonstrates how H_{ij} was computed from the insertion gains at a given time segment i.

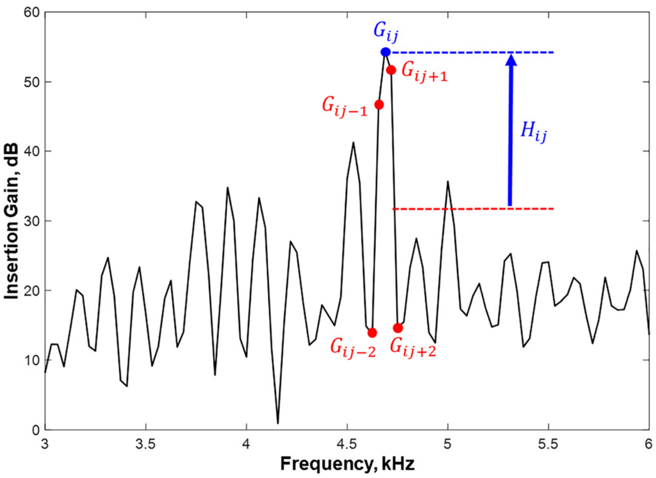

FIG. 2. (Color online) Demonstration of how H_{ij} was computed from the insertion gain at a given time segment i. Shown is the insertion gain between 3 kHz and 6 kHz for a signal with tonal feedback. The insertion gain for the frequency of interest G_{ij} is shown in blue. Show in red are the sidebands, which are comprised of two frequency bins (i.e., sideband range $r = 2$) below (G_{ij-2}, G_{ij-1}) and above (G_{ij+1}, G_{ij+2}) the frequency of interest.

The final step converted the matrix H_{ij} into the Peak Height Insertion Gain P that was used to classify each stimulus as containing feedback or not:

$$P = \max\left(\frac{1}{m}\sum_{i=1}^{m} H_{ij}\right), \tag{4}$$

where P is the grand maximum H_{ij} determined over the full frequency range after averaging across all time frames. The third column in Fig. 1 shows examples of H_{ij} averaged across time as a function of frequency. For comparison, the second column in Fig. 1 shows insertion gain averaged across time as a function of frequency from which clear feedback spikes can be seen for the stimuli categorized as having modulated and tonal feedback. A feedback threshold criterion c was set to equal the 95th percentile of the P for the set of stimuli that the judges categorized as having no feedback ("None" category). This value of c fixed the false-positive rate at $p = 0.05$. A given stimulus was classified as containing feedback if $P > c$. It is important to optimize the procedure with a relatively small false-positive rate so that a manufacturer would not be unfairly "punished" if the metric was used as a standard for benchmarking hearing aid performance. In addition, if the metric was used during a routine hearing aid fitting, a false positive may cause the clinician to unnecessarily reduce gain.

The values obtained for H_{ij} in Eq. (3), and consequently for P in Eq. (4), depended on the duration of the Hann window used to generate $\mathbf{G}$ in Eq. (2). The duration of the window, in turn, depended on its size N_w relative to the sampling frequency f_s of the signal. Since the original recordings (not the filtered and resampled stimuli used for the subjective ratings) were used for the PHIG method, f_s was 32 kHz. In addition, H_{ij} and P depended on the range of the frequency neighborhoods (sidebands) used to compute the relative height of the peak in each bin. The frequency range of each sideband was equal to the frequency resolution of each column in $\mathbf{G}$ times the number of columns for r. Since the frequency resolution depended on N_w, a brute-force approach was used to find the parameters and methods of computation that maximized feedback detection while minimizing false positives. Values of N_w equal to $2^x \mid x = \{8, \ldots, 15\}$ were explored, which corresponded to bin widths ranging from 125 to 1 Hz (all values referring to the frequency resolution of the bins here and throughout are rounded to the nearest Hz). Then, for each N_w, H_{ij}, and P were computed using sideband ranges corresponding to $r = 1$ to 20.

III. RESULTS

Results are presented in three sections. The first section presents analyses of H_{ij} averaged across the duration of the stimulus, known as the global PHIG method. The second section presents analyses of H_{ij} at spectrally and temporally local points of the stimulus, known as the local PHIG method. The third section examines how the global PHIG method is affected by a mismatch between the open ear responses that contribute to the aided and unaided output.

A. Global PHIG method

Every stimulus was processed by each combination of STFT bin width and sideband range. For each combination, values of P for the stimuli were sorted by feedback category. Stimuli in the None category were used to establish the feedback criterion c for each combination, which was then used to determine the hit rates for the remaining categories. The color map in Fig. 3 shows c as a function of the STFT bin width (columns) and sideband range (rows). As shown in the figure, the values of c displayed a decreasing trend as bin width increased from 1 to 31 Hz. This occurred because the feedback signals were concentrated over a very narrow frequency range; so, as the bin widths of the frequency

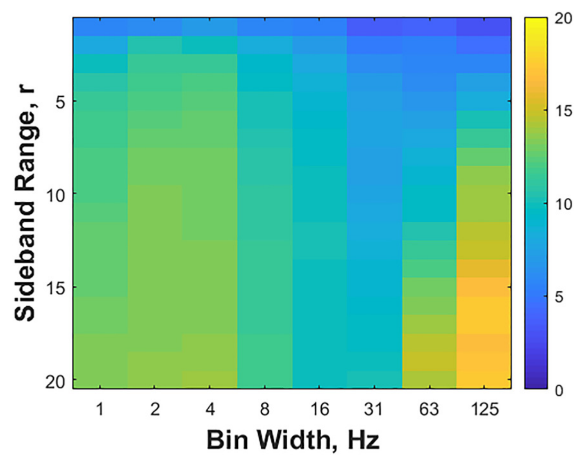

FIG. 3. (Color online) For each bin width and sideband range r (number of bins) combination, the feedback criterion c in dB is shown. Values of c closer to zero are shown in blue and values of c closer to 20 dB are shown in yellow.

analysis windows increased to include more off-peak frequencies, the overall energy in the window with the feedback signal decreased. The values of c increased for the 63 Hz and 125 Hz bin widths because the frequency analysis windows were likely wide enough to capture tonal effects produced by the primary feedback signal at nearby frequencies. The values of c also increased slightly as the sideband range increased because they included more frequencies away from the feedback signal for bin widths <=31 Hz, and more frequencies away from the feedback overtones for bin widths >=63 Hz.

Using the same format as Fig. 3, the color map in Fig. 4 indicates the hit rates for each feedback category. Figure 4(a) shows that the PHIG method was robust at detecting tonal feedback: The hit rate for most combinations of bin width and sideband range was 1.0 except for the combinations where bin width and sideband range were both low and both high. This occurred because the insertion gain for the feedback signal was so high for these cases (see Fig. 1) that normalizing it relative to its neighboring frequencies added no new information for detection. Figure 4(b) indicates that hit rates for modulated feedback generally

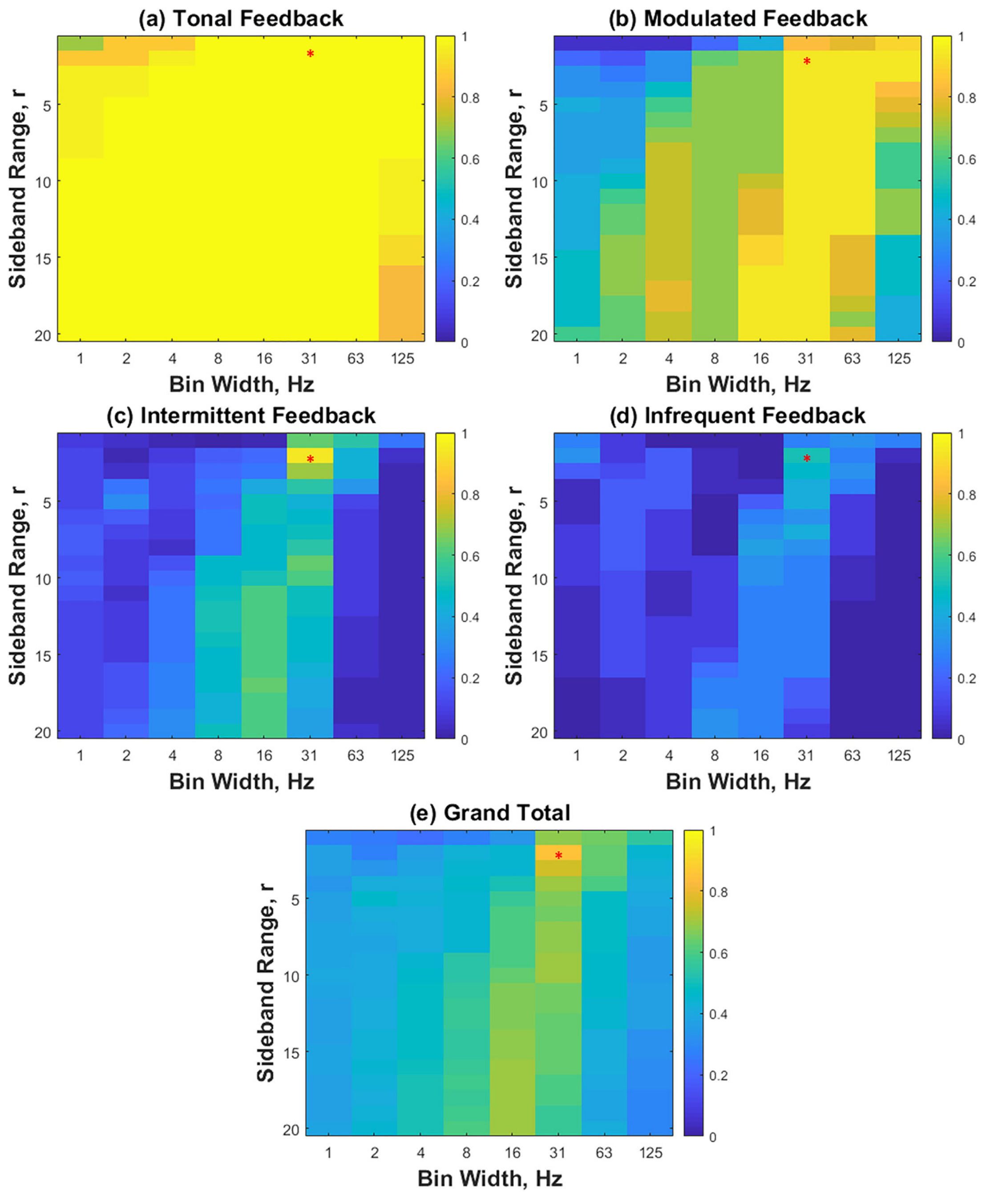

FIG. 4. (Color online) In each panel, the hit rates by bin width and sideband range are shown for the different feedback categories. Hit rates closer to zero are shown in blue and hit rates closer to one are shown in yellow. The red asterisk in each panel corresponds to the combination of bin width and sideband range r that generated the highest grand total hit rate (bottom-center panel).

increased as the bin width increased up to 31 Hz and 63 Hz, where hit rates were equal to 1.0 across a wide range of sideband ranges. Unlike tonal and modulated feedback, Figure 4(c) shows that hit rates for intermittent feedback never reached 1.0. Hit rates for intermittent feedback were highest for bin widths between 8 Hz and 31 Hz, and the maximum hit rate (0.94) was achieved only with a bin width of 31 Hz and a sideband range of $r = 2$. Not surprisingly, hit rates for infrequent feedback were lowest among the feedback categories as indicated by Fig. 4(d). Hit rates for infrequent feedback were generally highest for the 31 Hz bin width, where the maximum hit rate (0.50) was achieved with a sideband range of $r = 2$. This is the same combination that achieved the maximum hit rate for intermittent feedback. Given that the hit rates for tonal and modulated feedback were less dependent on precise combinations of bin width and sideband range, the "Grand Total" hit rates shown in Fig. 4(e) (the weighted average hit rates) were influenced more by the pattern of results for intermittent and infrequent feedback. Therefore, the maximum grand total hit rate (0.86) also occurred with a bin width of 31 Hz and a sideband range $r = 2$. This combination represents the optimized parameters and is denoted by the red asterisks in each panel shown in Fig. 4.

Using the optimized parameters (bin width = 31 Hz and sideband range $r = 2$), the distributions of P for each feedback category were computed. The distributions of P are shown in Fig. 5. The solid line corresponds to the 95th percentile of P for stimuli in the None category. Stimuli with P exceeding this line ($c = 5.2$ dB) were identified as containing feedback. For reference, the dashed line corresponds to the 100th percentile of P for the stimuli in the None category ($c = 7.4$ dB). From the figure, it can be seen that the distribution of P for stimuli with infrequent feedback encompassed the range of P for stimuli with no feedback. The distribution of P progressively increased for stimuli with intermittent, modulated, and tonal feedback. Table V provides the descriptive statistics of P for each category.

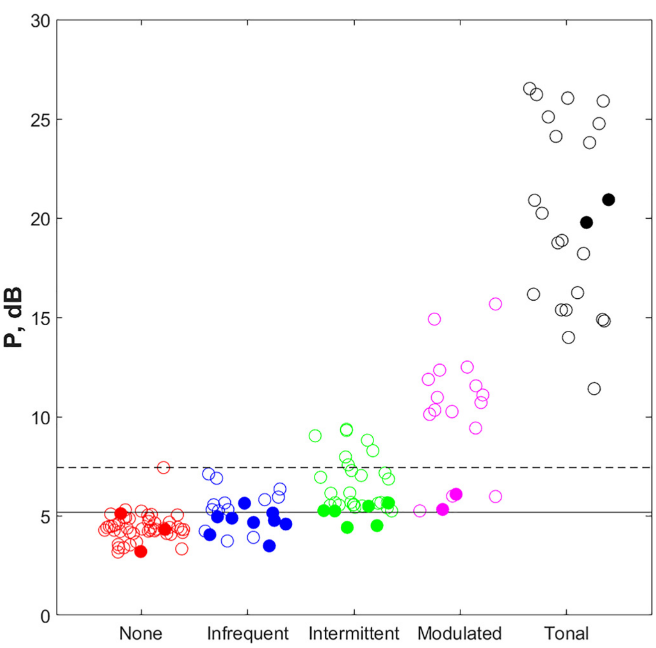

FIG. 5. (Color online) For each feedback category, the distribution of P is shown by the colored data points. The solid line corresponds to the 95th percentile of P for stimuli with no feedback ("None"). Stimuli with P exceeding this line are identified as containing feedback. For reference, the dashed line corresponds to the 100th percentile of the stimuli in the "None" category. The data points above the solid line in the None category correspond to false positives. The cases identified as containing feedback based on the local PHIG method (see section titled "Local PHIG Method") are shown by the filled data points.

B. Local PHIG method

As indicated by Figs. 4 and 5, the PHIG method for detecting and identifying specific feedback categories was poorer at detecting less persistent—and possibly sub-oscillatory—feedback (infrequent and intermittent) compared to persistent oscillatory feedback (modulated and tonal). This outcome is not surprising because the global PHIG method averages the local peak height insertion gain, H_{ij}, across the duration of the stimulus so that the average gain of the frequency component containing a brief feedback signal is relatively low. Therefore, while the presence of a very brief feedback signal in a stimulus might be detectable to a human listener, the global PHIG method is less likely to separate it from other stimuli containing no feedback. The local PHIG method was specifically developed to detect the stimuli missed by the global PHIG method by focusing on the spectrally and temporally local values of H_{ij}.

The basic concept behind the local PHIG method was to find the individual time-frequency tiles with oscillatory or sub-oscillatory feedback that caused the judges to classify a given stimulus as containing feedback. In addition to bin width and sideband range, the local PHIG method required an additional parameter in the optimization procedure. For each bin width and sideband range combination, a local threshold k had to be determined for deciding when a time-frequency tile should be classified as containing feedback: $H_{m,n} > k$. Based on a preliminary analysis, a range of $k = 10.5$ dB to 50 dB in 0.5 dB steps was explored for each bin width and sideband range combination. For each of the 12 800 permutations of parameter values (8 bin widths $\times$ 20 sideband ranges $\times$ 80 threshold values), the proportion of time-frequency tiles identified as containing feedback was computed as the decision variable p. Just as with the global

TABLE V. Descriptive statistics of P in dB, for the different feedback categories.

Feedback Category	N	Mean	SD	Min	Max
None	49	4.4	0.7	3.2	7.4
Infrequent	22	5.2	0.9	3.7	7.1
Intermittent	33	6.4	1.4	3.5	9.4
Modulated	18	10.0	3.1	5.2	15.7
Tonal	23	19.9	4.6	11.4	26.5

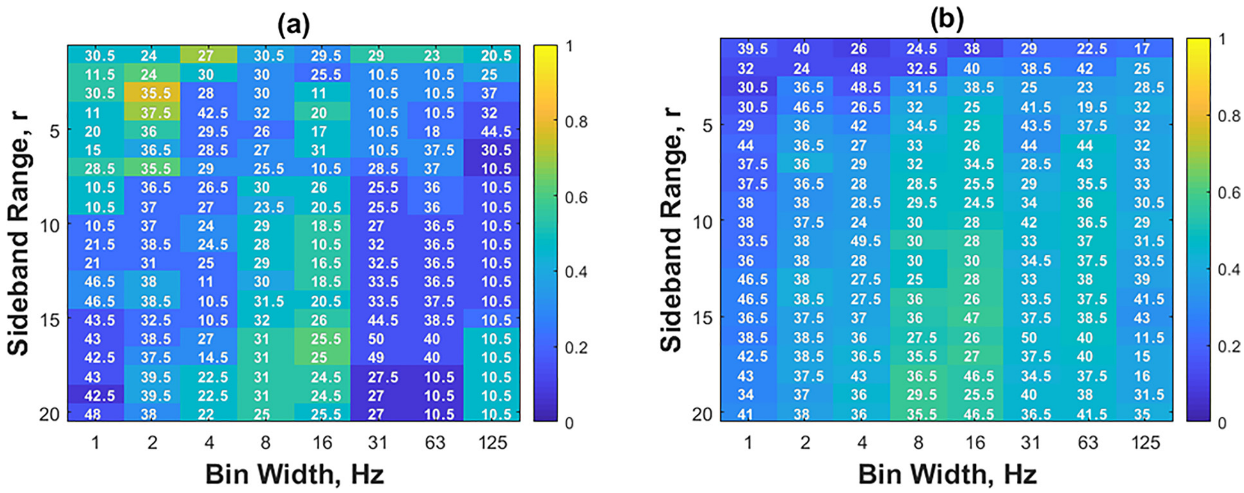

FIG. 6. (Color online) Using the same color map as Fig. 3, the grand total hit rates by bin width and sideband range as shown for the local PHIG method: (a) corresponds to analyses conducted only on the subset of stimuli missed by the global PHIG method, (b) corresponds to analyses conducted on the full set of stimuli. The numbers in each cell correspond to the optimized values for the local threshold k. In cases where multiple values of k corresponded to the maximum hit rate for a given combination of bin width and sideband range, the lowest value of k is shown.

PHIG method, the criteria c for calculating hit rates were set to equal the 95th percentile of the p values obtained from the None category.

Using all the stimuli from the None category and the subset of stimuli with feedback that were missed by the global PHIG method (N = 13; intermittent = 2; infrequent = 11), the highest total hit rate among all values of k was determined for each combination of bin width and sideband range. For each combination of bin width and sideband range, the color map in Fig. 6(a) shows the total hit rates for the subset of stimuli and the text shows the optimized values of k. The combination that led to the highest total hit rate (0.77) was bin width = 2 Hz, sideband range $r = 3$, and local threshold $k = 35.5$ dB. For this particular combination, the 95th percentile of the stimuli in the None category (the criterion c) was 44 tiles out of 3991 frequency bins times 59 time segments or $p = 0.022\%$. Table VI shows the hit rates for each feedback category when computed with the global PHIG method and when this method was followed by the local PHIG method. With the addition of the local PHIG method, the grand total hit rate improved from 0.86 to 0.95, and the hit rate for stimuli with intermittent feedback improved to 1.0. Figure 5 shows that the stimuli classified as containing feedback (filled data points) using the optimized parameters for the local PHIG method were almost exclusive of the stimuli classified by the global PHIG method (open data points above the solid line). Interestingly, very few stimuli with persistent feedback (modulated and tonal) were detected with the local PHIG method using this particular set of parameters (see Discussion).

TABLE VI. For each feedback category, the hit rate is shown for the optimized parameters using the global PHIG method and when these results were combined with the optimized parameters using the local PHIG method. The false-positive rate for each method was nominally set at 0.05.

Feedback Category	Global hit rate	Local hit rate
Infrequent	0.50	0.77
Intermittent	0.94	1.00
Modulated	1.00	1.00
Tonal	1.00	1.00
Grand Total	0.86	0.95

For comparison, Fig. 6(b) shows the grand total hit rates and optimized values of k when the full set of stimuli was analyzed by the local PHIG method. This figure indicates that the local PHIG method by itself did not perform as well as the global PHIG method [cf. Fig. 4(e)] and that the optimized parameters depended on whether the full set or a subset of stimuli were being analyzed [cf. Fig. 6(a)].

C. Simulations of mismatched REUGs

Since the REUG is not typically measured in routine clinical practice, we investigated whether a stored, age-corrected, or other standard REUG can be used in place of a patient's actual REUG. More specifically, we evaluated the efficacy of the PHIG method when there is a mismatch between the contribution of the ear canal resonance to the REAR, p_a in Eq. (1), and the contribution of the ear canal resonance to the REUR, p_u in Eq. (1). As shown in Fig. 7, individual ear responses were simulated from the open-ear response measured in KEMAR. First, the KEMAR REUG used in this study (dotted lines) was subtracted from the REAR of each stimulus to create a KEMAR-removed REAR for each one. Then, the KEMAR REUG was circularly shifted by –1000 Hz to +4000 Hz, which moved the resonant peak from 2400 Hz down to 1400 Hz and up to 6400 Hz (thin solid lines). These simulations represent the REUGs of adults with longer ear canals and infants with shorter ear canals, respectively. The thick solid line in each panel in Fig. 7 shows the mismatch between the KEMAR and simulated REUGs. Finally, the simulated REUGs were added back to the KEMAR-removed REAR of each

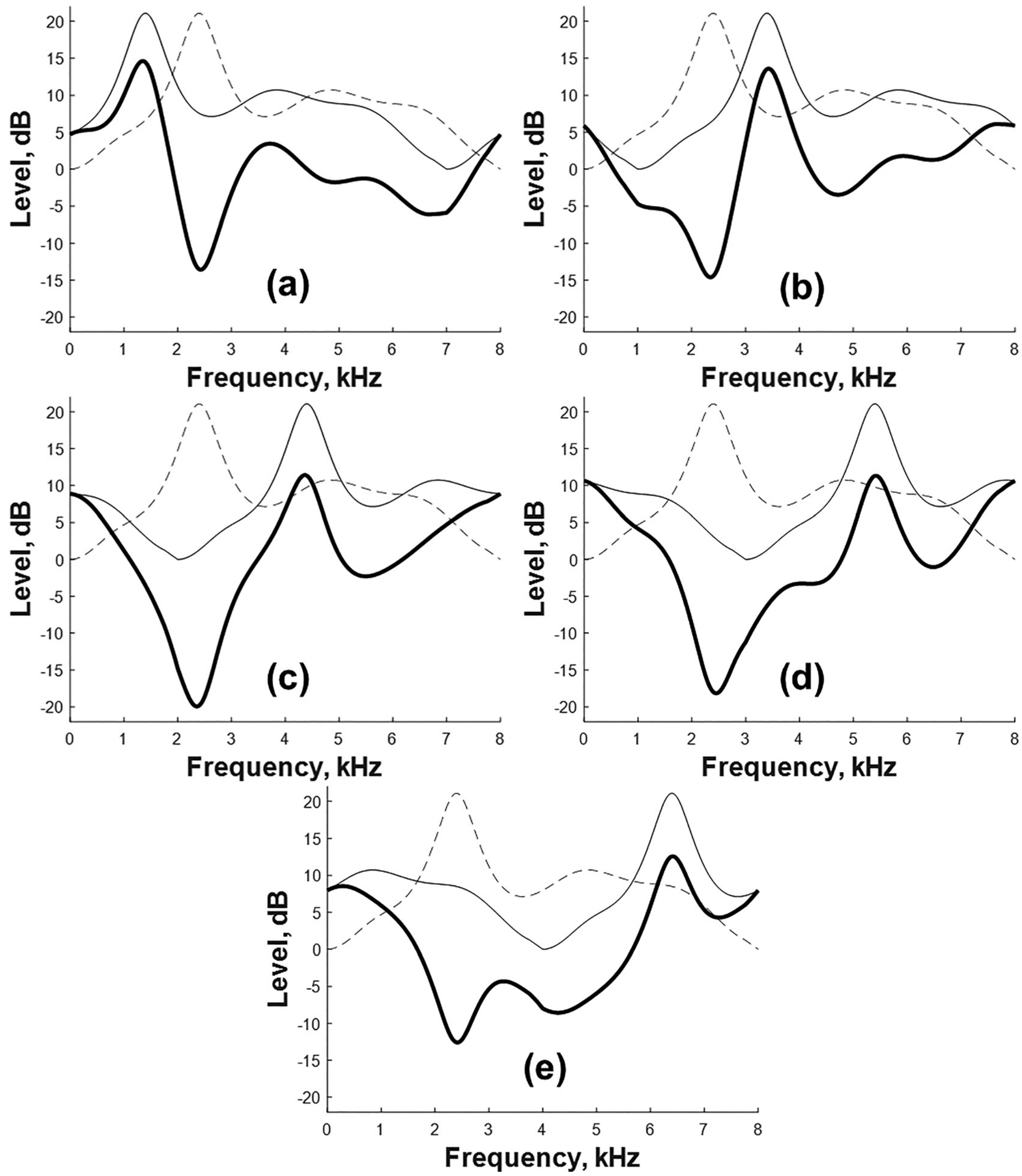

FIG. 7. Simulated open-ear responses (thin solid lines) of the stimuli compared to the KEMAR open-ear response from which the insertion gains were derived (dotted line). The thick solid lines represent the mismatch between the real and simulated responses. These were added to the real-ear aided responses of the stimuli to show the effects of computing $\boldsymbol{G}$ from individual ear responses that differ from a fixed reference response. The panels correspond to different frequency shifts of the KEMAR open-ear response that generated the simulated individual ear responses: (a) –1000 Hz, (b) +1000 Hz, (c) +2000 Hz, (d) +3000 Hz, and (e) +4000 Hz.

stimulus to create a modified REAR for each one. The KEMAR REUR from the original study was subtracted from these modified REARs to compute the Insertion Gain Matrix $\boldsymbol{G}$ for the global PHIG method.

Using the optimized parameters for the global PHIG method (bin width = 31 Hz, sideband range $r = 2$; criterion $c = 5.2$ dB), $\boldsymbol{G}$ was computed for the stimuli with mismatched open-ear responses for the REAR and REUR. On average, across all stimuli and simulations, the absolute difference between the original values for P and the new values for P was 0.05 dB for frequency shifts between –1000 Hz and +3000 Hz and was 0.18 dB for the +4000 Hz frequency shift. As such, the false-positive rate remained unchanged for all frequency shifts, except for the +4000 frequency shift where it increased to 0.16. Due to the slightly higher values of P for the +4000 Hz frequency shift, the hit rate for infrequent feedback increased to 0.55. However, due to the increased false-positive rate, this condition will not be considered further. Table VII shows that the hit rates for the modulated and the tonal feedback categories were unaffected when the simulated resonant peak was shifted from −1000 Hz to +3000 Hz. The hit rates for the infrequent and intermittent feedback categories varied slightly but essentially remained unchanged. These simulations show that a

TABLE VII. For each feedback category, the hit rate for the global PHIG method is shown when the resonance peak of the output response (REAR) was shifted –1000 Hz to +4000 Hz to create a mismatch with the reference open-ear response (REUR). Except for the +4000 Hz condition, the false-positive rate was unaffected by the mismatched responses and remained equal to 0.05.

Feedback Category	Matched	−1000 Hz	+1000 Hz	+2000 Hz	+3000 Hz	+4000 Hz
Infrequent	0.50	0.41	0.50	0.50	0.45	0.55
Intermittent	0.94	0.91	0.94	0.94	0.94	0.94
Modulated	1.00	1.00	1.00	1.00	1.00	1.00
Tonal	1.00	1.00	1.00	1.00	1.00	1.00
Grand Total	0.86	0.83	0.86	0.86	0.85	0.88

generic open-ear response can be used when computing P from a single REAR measurement in the clinic. Infants and young children have shorter ear canals, which creates a high-frequency resonant peak in their open-ear response. Therefore, an age-corrected REUR could be applied to the generic open-ear response from KEMAR to improve the accuracy of the PHIG method. Given the parameters used to compute PHIG in these simulations, the changes to the insertion gains due to a mismatched open-ear response were much more gradual than the narrow frequency range over which H_{ij} was computed. Therefore, the global PHIG method was robust in the face of broad discrepancies between the actual and simulated open-ear responses. For this reason, the local PHIG method, which was optimized using an even smaller bin width, is also expected to be unaffected by a mismatch between open-ear responses.

IV. DISCUSSION

The Peak Height Insertion Gain (PHIG) method is an advancement over previous methods developed for objectively detecting when hearing aid output is contaminated by feedback. As noted in the Introduction and the summary in Table I, compared to other methods, the PHIG method makes fewer restrictions on the nature of the input signal and on how the hearing aid is programmed. The PHIG method is also computationally simple because it requires only a single measurement of the hearing aid output. In addition to these technical details, the stimuli used to optimize the PHIG method have been quantified using a group of independent judges. This is the first method that has been established using a formal subjective listening test and perceptual judgments from individuals other than the study authors. Using the ratings from the judges in our study, the PHIG method has been documented to have high sensitivity and specificity across a variety of feedback categories ranging in severity. Two of the categories, tonal and modulated, have clear attributes of oscillatory feedback as indicated from their descriptions in Table IV. The PHIG method was the most robust for these categories, as indicated by hit rates at or near ceiling across a wide range of parameter combinations in Fig. 4. The other two categories, intermittent and infrequent, most closely correspond to MacIntosh and Cornelisse's (2018) intermittent and sub-oscillatory categories. No attempt was made in the work presented here to identify sub-oscillatory feedback as such, other than to give judges the option for flagging stimuli as having "feedback overtones". While perceptual attributes were not the focus of this study, this is the first work to identify the characteristics of feedback with labels that have been established using a group of judges. A more exhaustive investigation of this topic may reveal more descriptive and reliable labels.

Another advancement of the PHIG method over previously reported methods is that the signal analysis parameters for its computation have been optimized using performance metrics derived from perceptual judgments. Perhaps the most important parameter when analyzing feedback is the size of the FFT bin width. If the bin width is too wide, then less intense feedback spikes may be missed since the analysis bin will contain energy from frequencies where there is no feedback and/or the energy in the feedback spikes may be canceled out when summed with the spectral valleys that result from the destructive interference by the out-of-phase components in the feedback signal. If the bin width is too narrow, then a method that averages across time may miss a feedback signal that changes frequency as a by-product of the DFS and/or a dynamic feedback path. Most of the methods listed in Table I used a 16 kHz sampling rate and 256-point FFT, which corresponds to a 63 Hz bin width. One exception is Freed and Soli (2006), who used a 24 kHz sampling rate and 1024-point FFT, which corresponds to a 23 Hz bin width. Interestingly, the optimal bin width for the global PHIG method was determined to be 31 Hz (32 kHz sampling frequency, 1024-point FFT), which falls within the range of values that have been used in the past.

The optimal sideband range when computing H_{ij} depends on the likelihood that the sidebands are wide enough to capture some of the low-amplitude portions of the valleys adjacent to the peaks in the rippled spectrum because this will lead to the largest values of H_{ij}. The density of the spectral ripple depends on the throughput delay in the hearing aid, which ranged from 2 to 7 ms for the manufacturers used in this study (Alexander, 2016), plus the travel delay back through the feedback path, which is around 0.17 ms according to MacIntosh and Cornelisse (2018). These values for the combined delay lead to a separation of 70–230 Hz between adjacent spectral peaks and valleys. The optimal sideband range for a 31 Hz bin width was found to be $r = 2$, which corresponds to a frequency range for the sideband of 63 Hz. While this value does not fall within the values expected based on time delays, it should be noted that a wide range of values of r led to hit rates of 1.0 for modulated and tonal feedback. Recall that the optimized parameters for the global PHIG method were determined mainly by the intermittent and infrequent feedback categories, which may have a more complex frequency response given the less frequent and less intense feedback, including sub-oscillatory feedback.

To increase the hit rate for the intermittent and infrequent feedback categories, a secondary method, the local PHIG method, was implemented. Note that the PHIG method is a time-segmented analysis like the modified TVC method proposed by Spriet and colleagues (Spriet *et al.*, 2008, 2009, 2010; Madhu *et al.*, 2011). Whereas Spriet and colleagues computed the maximum TVC as a function of time, the global PHIG method computed P as a function of frequency. Averaging H_{ij} across time leads to a metric (P) that is more sensitive to persistent feedback at localized frequencies; however, it loses temporal specificity that may be needed to detect less frequent feedback. Hence, the local PHIG method did not average across time or frequency. Interestingly, the local PHIG method led to different optimized parameters depending on whether only the misses from the global PHIG method were analyzed [Fig. 6(a)] or if all the stimuli were analyzed [Fig. 6(b)]. Whereas the optimized parameters for the misses (mostly infrequent and intermittent feedback) were spectrally localized (2 Hz, $r = 3$), the optimized parameters for the full pool were broader (8 Hz, $r = 18$ and r = 19). Very few stimuli with persistent feedback (modulated and tonal) were detected with the local PHIG method using the parameters that optimized the detection of the infrequent and intermittent feedback because the sidebands were so narrow ($\pm$6 Hz) that they likely included the feedback spike, thereby resulting in a small value for H_{ij}. Perhaps, the spectrally localized parameters picked up on time segments in the stimuli with infrequent and intermittent feedback where the DFS erred and injected a brief tonal signal, different from oscillatory feedback—a condition known as entrainment. This explanation is plausible considering that the optimized threshold was $k = 35.5$ dB for the local PHIG method compared to $c = 5.2$ dB for the global PHIG method. With this amount of relative insertion gain, a short-duration tone would have certainly been audible to the normal-hearing judges. Another reason for the significant increase in the threshold between the two methods is likely that individual time-frequency bins without feedback or entrainment contain spurious fluctuations that increase the value of H_{ij}. When computed over time as with the global PHIG method, this "noise" is averaged out. This explanation was also offered in the Introduction for why the TVC threshold for instability was 5 dB when averaged over time by Shin *et al.* (2007) compared to 20 dB when analyzed at individual time segments by Spriet and colleagues (Spriet *et al.*, 2008, 2009, 2010; Madhu *et al.*, 2011).

Whereas previous feedback detection metrics, like PCR and TVC, informally established criterion values that specified when a hearing aid was considered to be in oscillatory feedback and in sub-oscillatory feedback, criterion values for the PHIG method have been established using perceptual judgments. As indicated by the descriptive statistics of P for each feedback category in Table V and the data in Fig. 5, P increases with the severity of the category; however, there is a lot of overlap between categories. Interestingly, there is a break in the values for P around 10 dB between the infrequent/intermittent feedback categories and the modulated/tonal feedback categories. As noted earlier, the former categories may be considered to include sub-oscillatory and intermittent feedback, while the latter categories may be considered to be oscillatory feedback only. Considering that a feedback criterion of $c = 5.2$ dB has been established for identifying stimuli as containing feedback, these values are very close to the values informally established by Shin *et al.* (2007) for the TVC method. The TVC method is conceptually similar to the PHIG method in that the TVC corresponds to the maximum peak or valley of the spectral ripple across frequency. Stimuli with TVC $<$ 5 dB were classified as having no feedback, stimuli with TVC $\geq$ 5 dB and $<$10 dB were classified as having sub-oscillatory feedback, and stimuli with TVC $\geq$ 10 dB were classified as having oscillatory feedback. Values for P can also be used informally in a clinic setting to grade the severity of feedback into these broad categories.

As noted in the Introduction, the PHIG method was designed to address two primary needs. The first need is a non-intrusive, objective method that establishes the maximum stable gain before feedback in a formal technical specification of feedback suppression measurements in hearing aids. This information can inform clinicians of the expected audiometric fitting range of a particular hearing aid in actual use conditions. Another need addressed by the PHIG method is an early warning for clinicians that a hearing aid is in sub-oscillatory feedback or on the verge of oscillatory feedback. Given that the PHIG method was established using a speech stimulus, it is compatible with routine clinical practice since speech or a speech-like stimulus is often used *in situ* to fit a hearing aid to prescriptive targets. That is, the PHIG method would not require any additional measurements other than the open-ear response or REUG. However, simulations showed that the sensitivity of the PHIG method should be unaffected when a generic or an age-corrected open-ear response is used to compute P from a single measurement of the REAR in the clinic.

V. LIMITATIONS AND FUTURE DIRECTIONS

The PHIG method was established using a single male talker with an average pitch of 191 Hz, which could slightly bias the metric since different talkers have different short-time spectra that could influence the likelihood of feedback and possibly its detection. Specifically, a talker with a different pitch could influence the optimized parameters for the infrequent and intermittent feedback categories because the maximum sensitivity of the PHIG method for these categories was tied to a specific combination of bin width and sideband range. For example, a talker with a lower pitch could influence how the rippled frequency response of the feedback transfer function interacts with the more closely spaced harmonics of the source signal. In this case, the insertion gain of the sidebands could increase relative to the peak so that a narrower bin width (e.g.,16 Hz instead of 31 Hz) would be necessary to optimize feedback detection. Future work should investigate whether the parameters of the PHIG method need to be re-optimized if a different talker is used for the standard. However, it is not expected that other differences between talkers will influence feedback detection with the PHIG method since the overall structure of speech is similar across talkers. For example, resonances associated with different vocal tract lengths manifest as shifts in the formant frequencies; and it has been shown in Fig. 7 that the PHIG method is robust when broad filter resonances are shifted up and down.

Another potential limitation is that the PHIG method was established using a single audiogram. Like with different talkers, the concern is *not* that some audiograms will increase the likelihood of feedback or change the feedback frequencies. Instead, the concern is that some audiograms will influence the detectability of feedback using the PHIG method by changing the relationship between the insertion gain of a peak and its sidebands. For example, more severe hearing losses require higher compression ratios, which flatten the spectral shape across frequency, especially as the number of independent channels increases and the duration of the compression time constants decreases (e.g., Alexander and Masterson, 2015). However, the optimized parameters of the global PHIG method encompass a 156 Hz range (the 31 Hz wide bin corresponding to the peak and the two bins above and below it), which is much narrower than typical hearing aid channels. Therefore, differences in compression parameters are not expected to influence the sensitivity of the PHIG method.

The PHIG method parameters were optimized with all advanced signal processing features in the hearing aids deactivated in the programming software, except the DFS. (It should be acknowledged that some manufacturers do not allow the clinician to completely turn off certain features). However, during a fitting, hearing aids are programmed with the features set as the patient will use them outside of the clinic. It is possible that certain hearing aid features, such as noise reduction, post-filtering, multiband directionality, and frequency lowering, could influence the sensitivity of the PHIG method. As noted in the previous paragraph, the PHIG method will be influenced to the extent that these signal processing features alter the insertion gain over a very narrow frequency range. While this study has shown that the PHIG method is sensitive to a wide variety of DFS methods and core signal processing schemes, these questions will need to be addressed by future research.

Another area of future research is the validation of the optimized PHIG parameters. To do this, novel stimuli produced from additional recordings would first be classified by the PHIG method and then compared to the ratings of human judges. As noted in the preceding paragraphs, validation could be extended to include different talkers, different audiograms, and activation of various hearing aid features.

Finally, as noted by others (e.g., Freed and Soli, 2006; Spriet *et al.*, 2008, 2009, 2010; Madhu *et al.*, 2011), there are limitations inherent to DFS performance metrics regarding how well they address all aspects of a hearing aid's DFS. Every objective or subjective method that simply quantifies the presence or the amount of feedback in a hearing aid fails to consider other factors that may affect how a patient experiences the effects of the DFS in everyday scenarios. For example, a DFS algorithm might be highly resistant to oscillatory feedback; however, this might come at the expense of gain reduction and poor sound quality resulting from sub-oscillatory feedback or entrainment from tonal, signals such as music. In other words, a determination of how well a hearing aid eliminates feedback by applying a non-intrusive, objective method like the PHIG method is only the first step in the overall evaluation of the DFS performance.

ACKNOWLEDGMENTS

Hearing aid measurement work for R.W. was supported through an interagency agreement between NIST and the U.S. Department of Veterans Affairs. The authors would like to thank Daniel Warren, the reviewers at NIST, and two anonymous reviewers for their valuable comments that improved the description of the PHIG method and other key concepts in the paper. The stimuli, data, and code referenced in this paper are publicly available at https://purr.purdue.edu/publications/3684/1, "Dataset of the peak height insertion gain (PHIG) method for quantifying acoustic feedback in hearing aids v.1.0(#3684)."

APPENDIX

Uncertainty analysis

Standard and expanded ($k = 2$) uncertainties for the measured P values discussed herein were determined by applying published guidelines for evaluating uncertainties (Taylor and Kuyatt, 1994). A type A standard uncertainty component was determined for the aided ear measurements by pooling the variances of results obtained for aids programmed to create a representative aided ear stimulus sample for each of the five different categories of feedback (None, Infrequent, Intermittent, Modulated, Tonal). Data were acquired for ten trials of each of the five representative aided ear stimulus samples and used with a single unaided ear sample to provide a total of 50 values for P. These values were used to determine the standard uncertainty for the aided ear measurements by pooling the variances calculated for the five samples. A type A standard uncertainty component was also determined for the unaided ear measurements by pooling the variances of the results obtained using a single trial for each of the five representative aided ear stimulus samples. These five separate samples were used with ten trials of the unaided ear to provide another set of 50 values for

TABLE VIII. Standard and expanded ($k = 2$) uncertainties for the measured values of P.

Source of uncertainty	Type	Standard uncertainty (%)
Repeatability of aided ear measurements	A	2.1
Repeatability of unaided ear measurements	A	2.2
Linearity of measurement system	B	1.7
Estimate of combined standard uncertainty (%)		3.5
Estimate of expanded (k = 2) uncertainty (%)		7.0
Estimate of expanded (k = 2) uncertainty (dB)		0.6

P. These values were used to determine the standard uncertainty for the unaided ear measurements by pooling the variances calculated for the five samples. In addition, a type B standard uncertainty component was determined for the linearity of the measurement system from input–output data acquired using sinusoidal test signals over a 60 dB range at 200, 1000, and 8000 Hz. These measurement data were used to estimate the limits of a rectangular probability distribution chosen to model deviations from linear behavior. The standard uncertainty component for the linearity of the measurement system was determined from this distribution. These three standard uncertainty components were combined using the root-sum-squared method to determine a combined standard uncertainty and an expanded ($k = 2$) uncertainty. Values for all of these uncertainties are provided in Table VIII.

[1]The hearing-aid programming, the acquisition of audio recordings and subjective ratings, and the data analysis (except for the uncertainty analysis) for the results reported herein were completed at the Experimental Amplification Research (EAR) Laboratory of the Purdue University Department of Speech, Language, and Hearing Sciences (IRB Protocol #1406014978). None of the work completed at Purdue was either conducted or supported by NIST. Preliminary development of the PHIG method was done with audio recordings made of hearing aids in place on a KEMAR manikin at NIST, and with computational analysis of this audio measurement data carried out jointly by NIST and SoundsGood Labs (Burlington, ON, Canada) and discussed in meetings of Working Group S3/WG48, Hearing Aids, and its Hearing Aid Digital Feedback Suppression Subgroup. This working group was established by Accredited Standards Committee S3, Bioacoustics, which is administered by the Acoustical Society of America and accredited by the American National Standards Institute.

[2]Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

[3]Note that feedback detection could be considered a subset of the general problem of tone detection in noise. A rich set of literature exists for this problem, especially in psychophysics, for which the Theory of Signal Detection was developed (e.g., Green, 1958; Green *et al.*, 1959; Green and Swets, 1966; Swets *et al.*, 1959). Relevant to our methods, So *et al.* (1999) demonstrate that the discrete Fourier transform with periodogram spectral averaging is an optimal tone detector in white noise because the random fluctuations in amplitude and phase in the noise cancel over time. However, the general problem becomes more complicated when trying to detect a feedback tone against a speech background instead of noise. Therefore, alternative approaches that rely on the side effects caused by feedback may be necessary. The proposed method differs from classic tone detection approaches because it exploits the low amplitude valley in the spectral ripple (sidebands to the peaks) caused by the interaction of the feedback signal with the direct signal. In fact, the depth of the valleys could be used to help detect the presence of feedback: the TVC method of Shin *et al.* (2007) is equal to the maximum absolute value of the peaks and valleys. We explored several variants of our proposed method that explicitly considered the depths of the valleys, including the maximum absolute difference between a frequency component (e.g., a valley) and its sidebands (e.g., a peak), the peak-to-valley ratio (the minimum value in the sidebands), and cepstral analysis to quantify the depth of the modulations (ripples) in the periodogram. None of these variants demonstrated as high a sensitivity as the PHIG method, so they were not considered further.

[4]In the frequency domain, window shapes are characterized by the width of the mainlobe and the relative level of the side lobes (e.g., Harris, 1978; Kaiser and Schafer, 1980). In theory, different window shapes and overlap methods could influence the insertion gain of a peak relative to its sidebands. Therefore, until the effects of different window shapes can be determined, it is recommended that the methodology adopted in the present work (Hann window with 50% overlap) be used when applying the PHIG method.

[5]To minimize the computational complexity of the PHIG method, no attempt was made to compensate for the throughput delay of the hearing aid before subtracting the magnitude spectrograms. Alexander (2016) documented that the throughput delay of hearing aids produced by the manufacturers of the hearing aids used in this study ranged from 2 to 7 ms. Measurements from the most recent products from these manufacturers confirm that these delays have not changed considerably over the years. Because these delays are relatively short compared to the temporal smoothing used in the magnitude spectrograms, we do not expect the performance of the PHIG method to significantly improve if the effects of delay are compensated. In addition, compensating for the hearing aid delay would create delays in the other direction between the part of the aided response that comes from the direct path through the open ear canal.

Agnew, J. (**1996**). "Acoustic feedback and other audible artifacts in hearing aids," Trends Amplif. **1**(2), 45–82.

Alexander, J. M. (**2016**). "Hearing aid delay and current drain in modern devices," https://canadianaudiologist.ca/hearing-aid-delay-feature/ (Last viewed July 1, 2020).

Alexander, J. M., and Masterson, K. (**2015**). "Effects of WDRC release time and number of channels on output SNR and speech recognition," Ear Hear. **36**(2), e35–e49.

Audacity Team (**2018**). See https://www.audacityteam.org/ for Audacity version 2.2.2 (Last viewed August 25, 2021).

Bisgaard, N., and Ruf, S. (**2017**). "Findings from EuroTrak surveys from 2009 to 2015: Hearing loss prevalence, hearing aid adoption, and benefits of hearing aid use," Am. J. Audiol. **26**(3S), 451–461.

Chung, K. (**2004**). "Challenges and recent developments in hearing aids: Part II. Feedback and occlusion effect reduction strategies, laser shell manufacturing processes, and other signal processing technologies," Trends Amplif. **8**(4), 125–146.

Cox, R. (**1982**). "Combined effects of earmold vents and suboscillatory feedback on hearing aid frequency response," Ear Hear. **3**(1), 12–17.

Green, D. M. (**1958**). "Detection of multiple component signals in noise," J. Acoust. Soc. Am. **30**(10), 904–911.

Green, D. M., McKey, M. J., and Licklider, J. C. R. (**1959**). "Detection of a pulsed sinusoid in noise as a function of frequency," J. Acoust. Soc. Am. **31**(11), 1446–1452.

Green, D. M., and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics* (Krieger, Huntington, NY).

Harris, F. (**1978**). "On the use of windows for harmonic analysis with the discrete Fourier transform," Proc. IEEE **66**(1), 51–83.

Johnson, E., and Dillon, H. (**2011**). "A comparison of gain for adults from generic hearing aid prescriptive methods: Impacts on predicted loudness, frequency bandwidth, and speech intelligibility," J. Am. Acad. Audiol. **22**(7), 441–459.

Kaiser, J., and Schafer, R. (**1980**). "On the use of the 10-sinh window for spectrum analysis," IEEE Trans. Acoust. Speech Signal Process. **28**(1), 105–107.

MacIntosh, S., and Cornelisse, L. (**2018**). "Understanding and managing feedback in hearing instruments," Unitron whitepaper 027-6249-02.

Madhu, N., Wouters, J., Spriet, A., Bisitz, T., Hohmann, V., and Moonen, M. (**2011**). "Study on the applicability of instrumental measures for black-box evaluation of static feedback control in hearing aids," J. Acoust. Soc. Am. **130**(2), 933–947.

Marcum, S., and Ricketts, T. A. (**2011**). "Assessment of modern feedback reduction systems," in *Invited Poster at the American Auditory Society Annual Meeting*, Scottsdale, AZ.

Merks, I. (**2010**). "Method and apparatus for measurement of gain margins of a hearing assistance device," U.S. patent no. 7664281 B2.

Merks, I., Banerjee, S., and Trine, T. (**2006**). "Assessing the effectiveness of feedback cancellers in hearing aids," Hear. Rev. **13**(4), 53–57.

Ricketts, T., Johnson, E., and Federman, J. (**2008**). "Individual differences within and across feedback suppression hearing aids," J. Am. Acad. Audiol. **19**(10), 748–757.

Scollie, S., Seewald, R., Cornelisse, L., Moodie, S., Bagatto, M., Laurnagaray, D., Beaulac, S., and Pumford, J. (**2005**). "The desired sensation level multistate input/output algorithm," Trends Amplif. **9**(4), 159–197.

Shin, M., Wang, S., Bentler, R. A., and He, S. (**2007**). "New feedback detection method for performance evaluation of hearing aids," J. Sound Vib. **302**, 350–360.

So, H. C., Chan, Y. T., Ma, Q., and Ching, P. C. (**1999**). "Comparison of various periodograms for sinusoid detection and frequency estimation," IEEE Trans. Aerosp. Electron. Syst. **35**(3), 945–952.

Spriet, A., Eneman, K., Moonen, M., and Wouters, J. (**2008**). "Objective measures for real-time evaluation of adaptive feedback cancellation algorithms in hearing aids," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Lausanne, Switzerland.

Spriet, A., Moonen, M., and Wouters, J. (**2009**). "Objective evaluation of feedback reduction techniques in hearing aids," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Glasgow, UK.

Spriet, A., Moonen, M., and Wouters, J. (**2010**). "Evaluation of feedback reduction techniques in hearing aids based on physical performance measures," J. Acoust. Soc. Am. **128**(3), 1245–1261.

Swets, J. A., Shipley, E. F., McKey, M. J., and Green, D. M. (**1959**). "Multiple observations of signals in noise," J. Acoust. Soc. America **31**(4), 514–521.

Taylor, B. N., and Kuyatt, C. E. (**1994**). "Guidelines for evaluating and expressing the uncertainty of nist measurement results," NIST Technical Note 1297—1994 edition (U.S. Government Printing Office, Washington DC, 1994). https://www.nist.gov/document-18373 (Last viewed July 1, 2020).