Published in final edited form as:

Ear Hear. 2015 January ; 36(1): 111–124. doi:10.1097/AUD.0000000000000089.

Reliability and Clinical Test Performance of Cochlear Reflectance

Daniel M. Rasetshwane¹, Sara E. Fultz^{1,2}, Judy G. Kopun¹, Michael P. Gorga¹, and Stephen T. Neely¹

¹Center for Hearing Research, Boys Town National Research Hospital, Omaha, Nebraska 68131

²Department of Special Education and Communication Disorders, University of Nebraska-Lincoln, Lincoln, Nebraska 68583

Abstract

Objective—Cochlear reflectance (CR) is the cochlear contribution to ear-canal reflectance. CR is equivalent to an otoacoustic emission (OAE) deconvolved by forward pressure in the ear-canal. Similar to other OAE measures, CR level is related to cochlear status. When measured using wideband noise stimuli, potential advantages of CR over other types of otoacoustic emissions include (1) the capability to cover a wider frequency range more efficiently by requiring fewer measurements, (2) minimal influence on the recorded emission from the measurement system and middle ear, (3) lack of entrainment of spontaneous OAEs, and (4) easier interpretation due to the existence of an equivalent linear model, which validates the application of linear systems theory. The purposes of this study were to evaluate the reliability, assess the accuracy in a clinical screening paradigm, and determine the relation of CR to audiometric thresholds. Thus, this study represents an initial assessment of the clinical utility of CR.

Design—Data were collected from 32 normal-hearing (NH) and 58 hearing-impaired (HI) participants. A wideband noise stimulus presented at seven stimulus levels (10 to 70 dB SPL, 10-dB steps) was used to elicit the CR. Reliability of CR was assessed using Cronbach's α , standard error of measurement, and absolute differences between CR data from three separate test sessions. Test performance was evaluated using clinical decision theory. The ability of CR to predict audiometric thresholds was evaluated using regression analysis.

Results—CR repeatability across test sessions was similar to that of other clinical measurements. However, both the accuracy with which CR distinguished NH from HI ears and the accuracy with which CR predicted audiometric thresholds were less than reported in previous studies using distortion-product OAE measurements.

Conclusion—CR measurements are repeatable between test sessions, can be used to predict auditory status and are related to audiometric thresholds. However, under current conditions, CR does not perform as well as other OAE measurements. Further developments in CR measurement

Address correspondence to: Daniel M. Rasetshwane, Center for Hearing Research, Boys Town National Research Hospital, 555 N 30th St. Omaha, Nebraska 68131, Daniel.Rasetshwane@boystown.org, Fax: 1-402- 498-6351, Phone: 1-402-498-6372.

Financial Disclosures/Conflicts of Interest:

This research was supported by grants from the National Institute of Health (R01 DC8318, and P30 DC4662).

and analysis methods may improve performance. CR has theoretical advantages for cochlear modeling, which may lead to improved interpretation of cochlear status.

INTRODUCTION

Otoacoustic emissions (OAEs) are acoustic signals that originate within the cochlea as by-products of its normal function and are dependent on the status of outer hair cells (OHCs) (e.g., Brownell, 1990). OAEs are generated within the organ of Corti by either (1) intermodulation due to OHC nonlinearity or (2) wave reflection due to mechanical irregularity. Both of these mechanisms generate retrograde pressure waves that travel toward the base of the cochlea, through the middle ear, and into the ear canal where they can be detected. OAEs can be evoked using several different types of stimuli. Click-evoked OAEs (CEOAEs) are measured using clicks and, thus, provide information for a wide range of frequencies. Tone-burst-evoked OAEs (TBOAEs) are measured using short-duration sinusoids, and thus, cover a limited frequency range around the frequency of the tone-burst stimulus. CEOAEs and TBOAEs are often collectively referred to as transient-evoked OAEs (TEOAE) because both are evoked using short transient stimuli. Stimulus-frequency OAEs (SFOAEs) are evoked using pure tones and cover a narrow frequency range around the frequency of the stimulus. Distortion-product OAEs (DPOAEs) are evoked using a pair of primary tones and are thought to provide information about a restricted range of frequencies, although there is evidence to suggest that the generation sites extend towards the cochlear base (e.g., Martin et al. 2010). Noise with a spectral density that is band-limited to mimic that of clicks that have been used for CEOAEs has also been used to evoke OAEs (Maat et al. 2000). OAEs can also be produced spontaneously (SOAEs), in the absence of a stimulus.

Sensorineural hearing loss caused by damage to the outer hair cells (OHCs) results in a reduction in OAEs (e.g., Brownell 1990). Several studies have demonstrated a relationship between audiometric status and OAEs. This relationship has been observed for DPOAEs (e.g., Gorga et al. 1993; Stover et al. 1996; Boege and Janssen 2002; Johnson et al. 2010; Kirby et al. 2011) CEOAEs; (e.g., Gorga et al. 1993; Prieve et al. 1993; Hussain et al. 1998; Goodman et al. 2009; Mertes and Goodman, 2013), SFOAEs (e.g., Ellison and Keefe 2005), and TBOAEs (e.g., McPherson et al. 2006; Jedrzejczak et al. 2012). As a consequence of the relationship between auditory status and OAEs, and because of their non-invasive nature, OAEs are used clinically, including in newborn hearing screening. The two most commonly used OAE types are DPOAEs and CEOAEs because of their relationship to auditory status and ease of measurement (at least with currently implemented instrumentation).

Cochlear reflectance (CR) is an alternative measure of cochlear response (Allen et al. 1997; Rasetshwane and Neely 2012). Specifically, CR is the cochlear contribution to total ear-canal reflectance (ECR)¹. CR can be evoked using any type of stimulus that has previously been used to evoke OAEs, including tones. The current study used wideband noise and chirp stimuli to evoke CR. CR has several potential advantages over other types of OAEs. CR has

¹Ear-canal reflectance is the complex ratio of reflected acoustic pressure to forward pressure measured in the ear canal (e.g., Voss and Allen 1994). Ear-canal reflectance has been proposed as a diagnostic method to differentiate between normal ears and ears with specific middle-ear disorders (e.g., Voss et al. 2012).

the capability to cover a wider frequency range more efficiently with few measurements, compared to other measurement approaches. CEOAEs obtained using the double-evoked stimulus and analysis techniques that were designed to preserve the bandwidth (e.g., Keefe et al. 2011) can also cover a wide frequency range. However, the use of a noise stimulus in CR prevents the entrainment of spontaneous OAEs (SOAEs), which can be problematic in CEOAE measurement. To cover a similar frequency range with TBOAEs, SFOAEs or DPOAEs would require multiple measurements made one frequency at a time. Another advantage of CR over other OAEs is that, for CR, there is minimal influence from the measurement system and middle ear on the recorded emission. The procedure for extracting CR from ECR minimizes these influences by characterizing and removing them. Lastly, use of a wideband noise stimulus in the measurement of CR allows for the invocation of de Boer's (1997) nonlinear equivalence (EQ-NL) theorem, which, in turn, theoretically allows for the interpretation of the data using concepts and principles applicable to linear systems such as transfer functions, Fourier analysis, and time–frequency analysis. According to the EQ-NL theorem, for a given class of nonlinear system, (of which the cochlea could be an example) there is an equivalent linear system that has the same response as the nonlinear system for a wideband noise stimulus at a specific level. It is worth noting that the EQ-NL theorem may not apply when its admissibility conditions are not satisfied, as in the case where cochlear amplification involves time-delayed feedback, as suggested by Zweig (1991). The ability to simplify cochlear reflectance using the EQ-NL theorem may be especially useful for testing the validity of cochlear models. Recent modeling results demonstrating the similarity of CR measurements to simulations based on linear models of cochlear mechanics suggest that CR may be interpreted as arising mainly due to linear coherent reflection along the cochlear partition (Rasetshwane et al., 2014).

Because CR measurement techniques are relatively new, the relationship between CR and auditory status has not been explored. The study of this relationship should provide useful information regarding the clinical utility of CR. Given the advantages of CR measurements, it is of value to evaluate these measurements in normal-hearing and hearing-impaired participants, which is the purpose of this study. The specific goals of this study were to (1) examine the reliability of noise-evoked CR (i.e., repeatability of measurements across sessions), (2) evaluate the clinical utility of CR in terms of identification of auditory status (i.e., normal versus impaired), which we refer to as test performance, (3) evaluate the accuracy of audiometric threshold predictions from CR, and (4) determine optimal stimulus conditions for measuring CR. The similarity in magnitude of CR across test sessions was used to assess reliability. Clinical decision theory (e.g., Swets 1988; Fawcett 2006) was used in the assessment of test performance. Clinical decision theory has been used previously to test the accuracy of dichotomous decisions regarding auditory status based on OAE measurements (e.g., Prieve et al. 1993; Gorga et al. 1993; Stover et al. 1996; Boege and Janssen 2002; Ellison and Keefe 2005; Goodman et al. 2009; Johnson et al. 2010). Univariate analysis, in which CR data at a particular frequency are used to determine test performance at a matching audiometric frequency, as well as multivariate analysis, in which CR data at all frequencies are used to determine test performance at a single audiometric frequency, were performed. Our use of multivariate analyses parallels previous efforts in which these techniques were applied to CEOAEs (Hussain et al. 1998; Mertes and

Goodman, 2013), DPOAEs (Dorn et al. 1999; Gorga et al. 2005; Kirby et al. 2011), and SFOAEs (Ellison and Keefe 2005). A multivariate regression analysis was also used to predict audiometric thresholds from CR. Both multivariate test performance and multivariate threshold prediction analyses were applied to two independent sets of data. The first set of data was used to train multivariate models that related CR to auditory status. The models were subsequently applied to the second set of data without modification to predict auditory status.

METHODS

Participants

A total of 90 participants were enrolled in this study. A subset of 58 participants provided the data used for the *training* set. Of these, 22 were normal-hearing (NH) and 36 were hearing-impaired (HI) participants. An additional 32 participants provided the data used for the *validation* set. Of these, 10 were NH and 22 were HI participants. The average age for the 32 NH participants (22 females and 10 males) was 30.0 yrs. (range = 14–62 yrs., standard deviation, SD = 13.9 yrs.) and the average age for the 58 HI participants (29 females and 29 males) was 56.0 yrs. (range = 20–74 yrs., SD = 13.3 yrs.). The training-data set was collected first with no plans for multivariate analyses. However, given the results of univariate analysis, we elected to perform multivariate analyses on the initial set of data (i.e., the training-data set) to determine if these analytical techniques could be used to improve the accuracy of predictions. When it was observed that multivariate analyses improved performance, we elected to collect an additional set of data (i.e., the validation-data set) as a way to determine if the multivariate solutions were robust. The validation-data set was collected after conclusion of collection of the training-data set and was used to assess the generalizability of these models, as described later. A larger group was recruited for the training-data set than the validation-data set so as to reduce the effects of multicollinearity of the data on the multivariate regression model². Recruiting efforts were focused on having a greater number of HI than NH participants so as to increase variability in the data³. Additionally, we focused on recruiting HI participants with mild-to-moderate hearing losses so as to increase the likelihood of a CR response above the noise floor⁴. Focusing on HI participants with mild-to-moderate hearing loss also provides a stringent evaluation of test performance, as the data from these participants and those with NH are more likely to overlap, compared to the data from NH participants and participants with severe-to-profound HI. Audiometric thresholds were measured at ten octave and inter-octave frequencies (0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 6 and 8 kHz). Thresholds were measured in 5-dB steps, following standard clinical procedure. Participants were classified as NH when their thresholds were ≤ 20 dB HL at all frequencies and as HI when their thresholds were > 20 dB HL for at least one frequency above 1 kHz. The range of thresholds was –5 to 90 dB HL,

²Multicollinearity is a condition where two or more predictors of a multiple regression model are highly correlated. This can reduce the generalizability of a model to independent data.

³The set of data from normal-hearing participants is more homogenous and thus has lower variability compared to that of HI participants because of their limited range of audiometric thresholds. Therefore including more HI participants in the data increases variability and has the potential to produce more robust multivariate models.

⁴Since CR is dependent on OHC viability, it is unlikely that measurable CR will be present in an individual with hearing loss of 60 dB or greater because such a hearing loss is associated with complete damage or dysfunction of OHC.

although the majority had thresholds ≤ 60 dB HL. All participants had normal middle-ear function based on normal otoscopic inspection, a normal 226-Hz tympanogram (peak-compensated static acoustic admittance between 0.3 and 2.5 mmhos and peak tympanometric pressure between -100 and $+50$ daPa), and air-bone gaps ≤ 10 dB. All participants were recruited from a database of potential research participants that is maintained by Boys Town National Research Hospital (BTNRH). Participants were paid for their participation. Data collection was conducted under a protocol that had been approved by the BTNRH Institutional Review Board and informed consent was obtained prior to testing each participant.

All CR measurements were made monaurally. If both ears met the inclusion criteria in NH participants, the ear with better hearing was selected for testing. If both ears had similar audiometric thresholds, the test ear was selected randomly. If both ears met the inclusion criteria in HI participants, the ear with audiometric thresholds in the mild-to-moderate range above 1 kHz was tested to increase the likelihood of a CR response above the noise floor. If both ears had thresholds in this range, the test ear was selected randomly. Only one ear of each participant was tested. In all, data were collected from 16 left and 16 right ears of NH participants and 31 left and 27 right ears of HI participants. The counts for each audiometric frequency and HL combination are shown in Table 1.

Equipment

The sound-delivery system consisted of two modified tweeters (TW010F1, Audax, France) acoustically attached by plastic tubes to an ER-10B+ probe microphone (Etymotic Research, Elk Grove Village, IL, USA). Each stimulus condition was repeated on each tweeter (which we will refer to as the sound source), providing two independent measurements. Amplifiers placed between the soundcard and the tweeters provided power gain and reduced the electrical load on the soundcard output. (The modified tweeters and amplifiers were designed at Northwestern University by J. H. Siegel and S. Dhar). This sound-delivery system is suited for measuring CR at high frequencies because it has a wider bandwidth, lower distortion, and lower crosstalk compared to other emission probes (e.g., Neely and Siegel, 2011). Further description of the measurement system, including a schematic, can be found in Rasetshwane and Neely (2011). The measurement system was calibrated daily prior to data collection to determine the Thévenin-equivalent source impedance and pressure (Allen 1986; Keefe et al. 1992; Rasetshwane and Neely 2011). Thévenin-equivalent source parameters are required in the transformation of ear-canal pressure to CR. Stimulus delivery and data collection were monitored using locally developed software (EMAV; Neely and Liu, 1994). Artifact rejection based on the root-mean-square amplitude of the recordings of time-domain pressure was utilized. In this procedure, if two consecutive sweeps differed by a criterion amount, both sweeps were rejected, a procedure that is effective for rejecting transient artifacts.

Measurement of Cochlear Reflectance

Noise-evoked CR measurements were made following procedures described by Rasetshwane and Neely (2012). The description of the measurements presented here will

focus on stimulus conditions and CR properties. For further theoretical and measurement details, please see Rasetshwane and Neely (2012).

Stimuli were a wideband noise (WBN) and a wideband linear swept-frequency chirp (LSC), both digitally generated at a sampling rate of 48 kHz. The duration of each stimulus/response buffer was 171 ms. Data were collected in response to the LSC stimulus at 60 dB SPL and in response to the WBN stimulus presented at seven levels from 10 to 70 dB SPL in 10-dB steps. These stimulus levels were determined using a sound level meter (System 824, Larson Davis, Provo, UT, USA) with C weighting. C weighting was used, as opposed to flat weighting, because it provided moderate attenuation (up to about 10 dB) for sounds below 0.1 kHz, which includes power line noise from mains electricity. As will be shown later, reliable measurements at frequencies below 1 kHz were problematical and often uninterpretable. Thus, this filtering during calibration had little influence on the main findings in the present study. Following Rasetshwane and Neely (2012), longer averaging times were used at the lower stimulus levels to improve the signal-to-noise ratio (SNR) of the measured response. The stimulus conditions are summarized in Table 2. In a single visit, data collection using the complete set of stimulus conditions was repeated three times, except for 10 participants where data collection was repeated only two times because of their limited availability. The measurement probe was resealed at the beginning of each set of measurements. This approach was followed for both the training- and validation-data sets. Data-collection time (excluding time for consenting and audiometric, tympanometric, and otoscopic assessments) was, on average, 35 minutes per session.

The measurement of noise-evoked CR involved several steps: (1) measurements of total ear-canal reflectance (ECR) were made using sound sources that had been calibrated as described earlier, (2) next, a subtraction procedure was used to remove ear-canal and middle-ear contributions to total ECR, and (3) finally, time-frequency analysis was used to remove residual ear-canal and middle-ear contributions. Our measurement of ECR followed conventional procedures (e.g., Voss and Allen 1994). The subtraction procedure involves subtracting a response to the LSC from the response to the WBN. The idea is that the two responses include contributions from the ear canal and middle ear that are equal in level, (because of the linearity of these systems) and contributions from the cochlea that differ in level (due to the nonlinearity of the cochlea). The subtraction procedure is similar to the nonlinear residual method that is used to extract TEOAEs (e.g., Kemp et al. 1986). However, in the present study, scaling was not necessary because reflectance is a ratio, and thus is already scaled by the stimulus. Time-frequency analysis was performed using a complex gammatone filterbank with 49 channels (e.g., Patterson and Holdsworth 1996; Hohmann 2002). The center frequencies were logarithmically spaced from $f_c = 0.0625$ to 16 kHz in 1/6-octave steps, and the tuning of each filter was $Q_{ERB} = 3$, where Q_{ERB} is defined as $f_c/ERB(f_c)$ and $ERB(f_c)$ is the equivalent rectangular bandwidth of the filter with center frequency f_c (e.g., Shera et al. 2010).

Filters with the lowest seven center frequencies were removed because they were dominated by noise, leaving a total of 42 filters with center frequencies ranging from 0.14 to 16 kHz. The CR magnitude (CRM) of the output of these filters was calculated following procedures described in Rasetshwane and Neely (2012). CRM is used as the input variable for the

analyses of reliability, test performance and threshold prediction, providing 42 variables per stimulus level per participant.

The steps outlined above for the calculation of CR were carried out independently for measurements from the two sound sources. The mean of the two measurements was used to estimate the CR signal and the mean of the difference between the two measurements was used to estimate the noise. CR SNR was subsequently calculated using the CR signal and noise estimates.

To illustrate the time-frequency analysis of CR and its dependence on level, Fig. 1 shows examples of CR spectrograms at six levels for a representative NH ear. A color bar indicating the relationship between colors in the figure and CR level has been provided. The lines labeled $N_L = 4$ cycles and $N_H = 40$ cycles, together with $t_H = 0.5$ ms and $t_L = 30$ ms, were used as bounds for CR data used in the analyses to follow. The functions N_L and N_H indicate constant numbers of delay cycles or periods, and are defined as $N = tf_c$, where t is time relative to the stimulus onset and f_c is the center frequency (Shera and Guinan, 2003). The figure shows that as stimulus level increases, the energy of the component that is defined as CR (region bounded by N_L , N_H , t_H and t_L) decreases while the energy of the component in the region below N_L remains the same or increases. Although the level of the otoacoustic emission increases with level, CR level decreases because CR level is relative to stimulus level and grows at a slower rate compared to the rate of growth for the stimulus (like other OAEs). High-frequency content temporally precedes low-frequency content, consistent with signals of cochlear origin. We consider the energy below N_L as residual middle-ear and measurement-system activity because it is relatively independent of level. The temporal persistence of CR also depends on level; there is more re-reflected energy closer to $t_L = 30$ ms at the lower levels compared to the higher levels.

Analyses

Within-subject reliability was assessed by comparing CRM from the three measurements. For the reliability calculations, CRM data at each level was reduced from 42 to 10 variables by averaging the CRM data in $\frac{1}{2}$ -octave bands centered around 10 standard audiometric frequencies. Reliability was evaluated using Cronbach's α (Cronbach 1951), standard error of measurement (SEM), and absolute differences (AD). As points of reference, a value of $\alpha = 1$ indicates perfect reliability, while a value of $\alpha > 0.85$ is considered good reliability. SEM is defined as the standard deviation of the measurement errors and is calculated as

$$SEM = SD \sqrt{1 - \alpha} \quad (1)$$

where α is Cronbach's α , and SD is the standard deviation across all participants for a given test condition. A value of $SEM > 2$ dB is considered indicative of significant deviation in the measurements across repetitions because $\pm 1.96 SEM$ is equivalent to the 95% confidence interval of CRM. Cronbach's α and SEM have been used in several studies to assess reliability of OAEs (e.g., Franklin et al. 1992; Wagner et al. 2008; Kumar et al. 2013). The calculations of AD considered all possible session pairs for a particular participant – three pairs for the 80 subjects who participated in all three trials and two for the 10 subject who

participated in only two trials. The calculations of Cronbach's α , SEM and AD were each performed separately for each frequency and stimulus level. Assessment of reliability was performed for all participants for training and validation, and separately for NH and HI participants. Use of Cronbach's α and SEM for assessment of reliability allows for comparison of our results to those of Franklin et al. (1992) and Wagner et al. (2008), while use of absolute and signed differences allows for comparison to results of Wagner et al. and Thorson et al. (2012).

Clinical decision theory (e.g., Swets 1988; Fawcett 2006) was used for the assessment of test performance (identification of normal vs. impaired hearing). A description of CR test performance was obtained by computing hit rates (sensitivity), which is the proportion of HI ears that were correctly identified, and corresponding false-alarm rates (1-specificity), which is the proportion of NH ears incorrectly identified as HI. The assignment of hearing category (NH or HI) was conducted on a frequency-by-frequency basis. The number of ears classified as NH and HI at each frequency are provided in Table 3. Receiver operating characteristic (ROC) curves (plots of hit rate vs. false-alarm rate) were constructed, and the area under each ROC curve (A_{ROC}) was computed. A_{ROC} provides a single estimate of hit rate averaged over all possible false-alarm rates. A value of $A_{ROC} = 0.5$ indicates that hit and false-alarm rates are equal (chance performance), while a value of $A_{ROC} = 1.0$ indicates that the hit rate is 100% for all false-alarm rates, including a false-alarm rate of 0% (perfect test performance). A_{ROC} values were calculated for all stimulus levels and audiometric frequencies.

Two separate test-performance analyses were done. In the first, univariate analysis was performed, where CRM data at a particular frequency were used to determine test performance at the matching audiometric frequency. The second analysis followed a multivariate approach, in which CRM data at all frequencies were used to determine test performance at a single audiometric frequency. It was expected that multivariate analysis would result in better test performance since it uses a larger amount of CRM data to classify participants as NH or HI. Multivariate test-performance analysis was applied to the training- and validation-data sets. The training-data set was used to train the multivariate model and obtain equations relating CRM data to auditory status. This model was subsequently applied to the validation-data set without any modifications to classify ears as NH or HI. A multivariate logistic regression model was used for the analysis of test performance since the outcome of this analysis is a categorical variable (NH or HI). Multivariate analysis was also used for the prediction of audiometric thresholds from CRM; however, in this case we considered audiometric threshold as a continuous variable and used a multiple linear regression model.

Optimal stimulus conditions were determined following test-performance and threshold-prediction analyses as those stimulus conditions that, on average, resulted in the best test performance, (i.e., greatest separation between normal-hearing and hearing-impaired ears) and accounted for the most variance in the prediction of thresholds.

A good multivariate model should be generalizable, i.e., it should perform well when applied to independent sets of data (such as the validation-data set in this study). As a

general rule, the number of participants should be much greater than the number of variables for the development of the multivariate models. A model not meeting this rule can be idiosyncratic, in that the choice of variables and coefficients may be unique to a particular set of data. With only 22 NH and 36 HI participants included in the training set, there was concern that a robust multivariate model might not be produced because data were available from an insufficient number of participants. As a result, we evaluated two methods for reducing the number of variables in the original CRM data. In the first method, we used the mean CRM data at ½-octave bands around standard audiometric frequencies, reducing the number of variables from 42 to 10 per stimulus level per participant. The same approach to data reduction was used for the assessment of reliability. The second method applied principal component analysis (PCA) to reduce the number of variables from 42 to seven per stimulus level per participant. PCA is a procedure for reducing the dimensionality of a large data set consisting of correlated variables to a smaller data set while still retaining most of the variance of the original data set (e.g., Jolliffe 2002). Dimensionality reduction is achieved by transforming to a new set of variables, called the principal components, which are uncorrelated, and which are ordered so that the first component has the highest variance and each succeeding component in turn has the next highest variance possible under the constraint that it be uncorrelated with the preceding components. In the present study, the aim was to retain at least 95% of the variance, was achieved by retaining the first seven principal components.

RESULTS

Reliability of Cochlear Reflectance

Figure 2 shows mean values (across participants) of Cronbach's α (top row), SEM (middle row) and absolute differences (bottom row) as functions of frequency, with stimulus level as the parameter. Analyses based on data from all participants are shown in the left column, and separate analyses for NH and HI participants are shown in the middle and right columns, respectively. Values of mean and SD (across level and frequency) for each parameter are provided as insets in the figure panels. Recall that good reliability is considered as $\alpha > 0.85$ and $SEM < 2$.

Considering the results for all participants first, all three measures of reliability depended on frequency with poorer reliability at the extreme low (0.5 kHz) and high (6 kHz) frequencies, compared to the results observed for the mid frequencies. The three measures also depended on stimulus level with levels of 20–40 dB SPL demonstrating good reliability and levels of 60 and 70 dB SPL having poor reliability. The range of Cronbach's α was 0.59–0.94 with mean and SD (across frequency and stimulus level) of 0.86 and 0.07. The range of SEM was 1.4–3.8 dB with mean and SD of 2.3 and 0.5 dB. AD ranged from 1.4 to 6.0 dB with mean and SD of 3.5 and 1.0 dB.

Reliability also depended on hearing status, with CR data from NH participants demonstrating better reliability than CR data from HI participants. The value of Cronbach's α for NH participants (mean = 0.88, SD: 0.09, range = 0.59–0.98) was greater than those for HI participants (mean = 0.80, SD: 0.07, range = 0.53–0.92). Additionally, the values of SEM and AD for NH participants (SEM , mean= 1.8, SD = 0.7, range = 0.8–3.4 dB; AD , mean =

2.8, SD = 1.2, range = 1.0–5.8 dB) were smaller than those for HI participants (*SEM*, mean = 2.6, SD = 0.5, range = 1.1–3.8 dB; *AD*, mean = 3.9, SD = 1.0, range = 1.3–6.1 dB).

To further evaluate the reliability of CRM, a distribution was constructed of the mean (across level and frequency) of the signed difference between CRM measurements. All session pairs for a given participant were considered in the analysis, resulting in a total of 250 session pairs from 90 participants. Recall that 10 of the total 90 participants completed only two test sessions, which reduced the total number of session pairs from $3 \times 90 = 270$ to 250. Figure 3 shows this distribution in the form of a histogram. The signed differences in CRM between the measurements were roughly normally distributed with a mean of 0.0 dB, and a SD of 2.9 dB. The range was –16.4 dB to 9.0 dB; however, 90% of the signed differences were in the range –4.6 to 4.4 dB.

In summary, the reliability of CRM measurements depended on both frequency and level, with good reliability at low stimulus levels and mid frequencies. Reliability also depended on hearing status with NH participants having more reliable CR data than HI participants. Furthermore, the signed differences between the CRM measurements were small, within 4.6 dB for 90% of the measurements.

Cochlear Reflectance Univariate Test Performance

Before we present the results for CR test performance, we will demonstrate the relationship between CR and auditory status. Figure 4 shows mean CRM as a function of frequency for NH participants (thick solid line) and HI participants (thick dashed line) for the training-data set. The relationship between CR and auditory status for the validation-data set was essentially the same and thus it is not shown. Results for levels of 20 to 70 dB SPL are shown in separate panels as indicated in the figure. Estimates of the noise magnitude are included for NH participants (thin dash-dot line) and HI participants (thin dashed line). For each group, CRM was largest at low levels and decreased as stimulus level increased (similar to the trends observed in Fig. 1 for one NH participant). As expected, for a given stimulus level, CRM for NH participants was greater than CRM for HI participants, except at low (< 1 kHz) and high (> 8 kHz) frequencies. However, the noise magnitude for HI participants was larger than that for NH participants at these frequencies. Note also that the noise magnitude was approximately equal to the CRM at low frequencies for both groups. Thus, the larger CRM in HI participants at low frequencies are not meaningful because they cannot be distinguished from the noise.

To further illustrate the relationship between CR and auditory status, Fig. 5 plots audiometric thresholds against CRM for three frequencies (2, 3 and 4 kHz) where CRM was larger than the noise floor (i.e. positive SNR). Separate plots are shown for stimulus levels of 20, 40 and 60 dB SPL as shown on the right hand side of the figure. Values of R^2 are included in the figure panels. The relationship between thresholds and CRM was such that participants with better thresholds, in general, produced larger CRM. CRM explained as much as 52% of the variance in the thresholds. However, at 60 dB SPL for 4 kHz, the variance explained was as low as 0%. This poor relationship was due to the fact that average CRM for HI participants was higher than that for NH participants (see Fig. 4). The reasons for this observation are not yet known. A poor relationship was also observed at 10 dB SPL

across most frequencies (not plotted in Fig. 5). This happened because average CRM for HI participants was indistinguishable from the noise level (see Fig. 4).

Figure 6 shows A_{ROC} for training (top), validation (middle) and the difference between the two sets of A_{ROC} ($A_{ROC} = A_{ROC}$ for validation minus A_{ROC} for training; bottom) as function of frequency. The parameter in all panels is stimulus level. Test performance depends on frequency with the best A_{ROC} values occurring at mid-frequencies (1.5 to 4 kHz) for both training- and validation-data sets. Below and above this frequency range, values of A_{ROC} decrease. Test performance also depends on stimulus level with the best performance occurring, in general, for stimulus levels of 20–40 dB SPL. However, the dependence of test performance on level is complicated by its co-dependence on frequency, i.e. there is no single level that produces the best performance across all frequencies. A level of 40 dB SPL produces the largest A_{ROC} values at lower frequencies (< 1.5 kHz), 30 dB SPL at mid frequencies (1.5 to 3 kHz) and 20 dB SPL at higher frequencies (3 to 6 kHz). This pattern was observed for both training- and validation-data sets. Overall, the best test performance was observed for levels of 20 and 30 dB SPL at 3 kHz for the training data where $A_{ROC} = 0.93$, and for a level of 30 dB SPL at 1.5 kHz for the validation-data set where $A_{ROC} = 0.98$.

The decreased performance at high stimulus levels is due to an increase in the false-negative rate that occurs because more HI participants (especially those with mild-to-moderate HL) produced a CR response and were classified as NH. Values of $A_{ROC} < 0.5$ were observed for low frequencies suggesting that the distributions of CRM for NH and HI participants reversed positions (i.e., participants with HI produced larger CR than participants with NH). For reasons not yet understood, HI participants produced higher noise levels at low frequencies (compared to NH participants), but test-performance analyses did not have a SNR inclusion criterion. Thus, for analysis purposes, HI participants appeared to produce larger CRM, but with CRM and noise indistinguishable, test-performance assessments are meaningless at low frequencies. As a consequence of these observations, only the data above 1 kHz are interpretable. The differences in test performance, A_{ROC} , were small, with an average (across level and frequency) of +5.22% (SD = 15.78%), indicating that the trends observed in test performance, as well as the two data sets, were consistent (see bottom panel of Fig. 6).

Cochlear Reflectance Multivariate Test Performance

Figure 7 shows A_{ROC} for training (top), validation (middle) and A_{ROC} (bottom) as function of frequency, following the conventions used in Fig 6. However, in this case, multivariate analyses were applied to the data, as described above. The parameter in all panels is stimulus level. The left column shows results when the mean CRM data in the $\frac{1}{2}$ -octave bands around each of 10 audiometric frequencies were used to predict auditory status at a single frequency as indicated on the x axis. The right column shows results when auditory status was predicted from a multivariate analysis after reducing the amount of CRM data using PCA. As expected, values of A_{ROC} for multivariate test performance were greater than those for univariate test performance. Note that the ordinate range for Fig. 6 is 0.2–1.0, while that for Fig. 7 is 0.5–1.0. Note also that performance was above the chance level, even for the low frequencies for which univariate analyses resulted in findings that were uninterpretable.

These observations are consistent with previous OAE studies that showed better test performance for multivariate analyses compared to univariate analyses (e.g., Vinck et al. 1998; Dorn et al. 1999; Hussain et al. 1998; Gorga et al. 2005; Kirby et al. 2011). The range of values of A_{ROC} for the training-data set was 0.78 to 0.99 when 1/2-octave band method was used for data reduction and was 0.73 to 0.97 when PCA was used for data reduction. Thus, for the training-data set, using data in 1/2-octave bands resulted in better classification of auditory status, compared to using PCA. As was observed in the case of univariate test performance, there was no single stimulus level that resulted in the best performance across all frequencies when multivariate analyses were used. In general, levels of 20–50 dB SPL resulted in the best performance. These levels are lower than the levels that resulted in the best performance for DPOAE measurements (Stover et al. 1996; Johnson et al. 2010; Reuven et al. 2013). These observations hold for both methods of data reduction and for both sets of data. Test performance was better for high frequencies (> 2 kHz) compared to low frequencies (< 2 kHz), especially for the validation data. This trend has been observed for other OAE measurements as well. A_{ROC} values were generally poorer for the validation-data set than the training-data set, for both methods of data reduction. For the validation set, test performance using PCA for data reduction was similar to that using 1/2-octave band analysis. Because test performance for PCA was lower than that for 1/2-octave bands analysis for the training-data set, using PCA resulted in smaller A_{ROC} . Average A_{ROC} (across frequency and stimulus level) was -14.49% ($SD = 8.58\%$) for the analysis using 1/2-octave bands and -7.73% ($SD = 8.98\%$) for the analysis using PCA.

Audiometric Threshold Prediction from Cochlear Reflectance

Standard error (SE)⁵ and the coefficient of determination, R^2 , between measured and predicted audiometric thresholds were used to assess the accuracy with which audiometric thresholds can be predicted from CRM data using multivariate linear-regression analysis. R^2 provides a quantification of the variance that is accounted for by the predictions.

Figure 8 shows SE (units of dB) for training set (top row), validation set (middle row) and the difference between the two sets of SE ($SE = SE$ for validation minus SE for training; bottom row) as function of frequency. The arrangement in Fig. 8 is the same as that in Fig. 7. As in the case of test performance, there is no single stimulus level that results in the lowest SE values across all frequencies. Levels of 20–50 dB SPL result in the lowest SE for both types of analysis and for both data sets. In general, SE values increased with frequency, with the lowest SE values occurring at 0.25 kHz. This was observed for both types of analysis and for both data sets. The low SE values at 0.25 kHz, however, are meaningless because CRM and noise were indistinguishable at low frequencies, thus making the data uninterpretable. SE values for the training-data set are, in general, (but not always) lower than SE values for the validation-data set, as expected. There was less spread in SE across level for the training-data set than for the validation-data set. For the training set, SE values for the analysis using 1/2-octave bands were similar to those for the analysis using PCA.

⁵ SE was calculated as $SE = \sqrt{\frac{1}{N} \sum e^2}$, where e is the difference between measured and predicted thresholds and N is the number of observations.

However, for the validation set, *SE* values for the analysis using PCA were smaller than those for the analysis using ½-octave bands. That is, using PCA resulted in smaller *SE*. Average *SE* (across frequency and stimulus level) was +3.3 dB (SD = 2.0 dB) for the analysis using ½-octave bands and +2.1 dB (SD = 1.8 dB) for the analysis using PCA.

Figure 9 plots the coefficient of determination, R^2 for training (top row), validation (middle row) and the difference between the two sets of R^2 (i.e. $R^2 = R^2$ for validation minus R^2 for training; bottom row) as function of frequency. The arrangement of Fig. 9 follows that of Figs. 7 and 8. Again, levels of 20–50 dB SPL result in the best threshold prediction (highest R^2 values) for both types of analysis and for both data sets, and there is no single stimulus level that results in the highest R^2 values across all frequencies. Threshold prediction was poorest at 10 dB SPL. In general, R^2 values increased with frequency, with the highest R^2 values occurring at frequencies of 3 to 6 kHz. This observation is unlike that observed in Fig. 8 where the best threshold prediction (low *SE* values) was at low frequencies. The reasons for this contradiction are not yet known. R^2 values for the training-data set are generally higher than similar values for the validation-data set, as expected. There was less spread in R^2 across level for the training-data set than for the validation-data set. As was observed in the analysis using *SE* as a metric for evaluating the accuracy of threshold prediction from CRM data, the average R^2 (across frequency and stimulus level) was smaller for the prediction using PCA for data reduction (−31.90% with SD of 23.67%) compared to the prediction using CRM data in ½-octave bands (−43.55% with SD of 23.28%). Changes in R^2 between training and validation were greater than the corresponding changes in A_{ROC} (−7.73% and −14.49% for analyses using PCA and ½-octave bands, respectively). In summary, threshold prediction from CR data depended on both frequency and stimulus level. This was the case for both the training- and validation-data sets. For the training-data set, the variance accounted for was as high as 70% (at 4 kHz).

DISCUSSION

Overall noise-evoked CR measurements were reliable. When considering data for all participants, the mean value of Cronbach's α indicated good reliability. However the mean value of *SEM* was greater than 2, indicating that the differences between measurements from different test sessions were significant. When data for NH and HI participants were considered separately, both α and *SEM* showed that reliability was good for NH participants, but not as good for HI participants. HI subjects produce no response (or responses smaller than those produced by NH subjects), therefore the HI data were less reliable because those data were more dominated by noise. A similar finding, better reliability for NH participants than HI participants, was also reported by Marshall and Heller (1996) in their analyses of TEOAEs.

The reliability depended on both the stimulus level and frequency, with good reliability at levels of 20–40 dB and mid frequencies (1 and 4 kHz). The reason for the poor reliability at high levels is that CRM was low at high stimulus levels. Recall that CRM is high at low levels (see Fig. 4) because CRM is equivalent to an OAE normalized by the forward-pressure component of the stimulus. Therefore, while OAEs grow as level increases, CRM decreases because the growth of the OAE is nonlinear and compressive due to the nature of

the cochlear amplifier. Reliability is poor at the frequency extremes because the noise level is the same size as the CRM level at these frequencies (see Fig. 4). At 10 dB SPL, reliability was poor because the noise level was high.

Cronbach's α , SEM and AD have been used previously to evaluate the reliability of OAE measurements (e.g., Franklin et al. 1992; Wagner et al. 2008). Comparison to studies that made measurements at multiple levels and frequencies are of interest. Franklin et al. evaluated the reliability of DPOAEs measured at six primary frequencies and 12 levels in 12 NH participants. For their measurements made across days, they reported values of Cronbach's α of 0.25–0.98 (mean = 0.87, SD = 0.15) and SEM of 0.8–5.8 (mean = 1.9, SD = 1.0) dB. Wagner et al. (2008) also evaluated the reliability of DPOAEs measured at six frequencies and seven levels in 20 NH participants. For their measurements with probe reseat, they reported values of Cronbach's α of 0.74–0.98 (mean = 0.89, SD = 0.06), SEM of 0.9–3.0 (mean = 1.9, SD = 0.6) dB and AD of 0–22.1 (mean = 2.3, SD = 2.5) dB. Thorson et al. (2012) used AD to evaluate the reliability of DPOAEs measured at three primary frequencies and 17 levels in a group of 16 NH and 58 HI participants. Their mean values of AD were 3.5, 2.6 and 2.9 dB at these three frequencies, with corresponding SDs of 3.5, 2.6 and 2.8 dB. The values of Cronbach's α and SEM reported by Franklin et al. and Wagner et al. are similar to the values observed in our study for NH participants (see Fig. 2, second column). Values of AD reported by Wagner et al. are similar to our corresponding values for NH participants, and values reported by Thorson et al. are similar to our corresponding values for the combined group (see Fig. 2, first column). In summary, CR reliability was within the range that has been observed for other similar measurements, such as DPOAE measurements. As a result of these findings, further evaluation of CR measurements in relation to diagnosis and quantification of hearing loss was warranted.

For frequencies at and above 1 kHz, univariate analyses of CR measurements were able to distinguish ears with NH from those with HI, although not without error (see Fig. 6). Test performance for CR measurements was best for stimulus levels of 20 to 50 dB SPL, whereas other OAE measurements produced better performance at higher stimulus levels. Despite the differences in test performance between CR and other OAE measurements, there would be value in further exploring the accuracy of CR measurements to identify auditory status, given that these measurements provide the greatest control over hardware influences, provide information for a wide range of frequencies in a single measurement, and lend themselves to modeling that may not be possible with other OAE measurements.

Multivariate analysis improved test performance, with values of A_{ROC} for both $\frac{1}{2}$ -octave band and PCA methods of data reduction being greater than values of A_{ROC} for univariate analysis for the training-data set (compare Fig. 6 to Fig. 7). This was an expected observation since multivariate analysis incorporates more information than univariate analysis to classify ears as NH or HI. These findings are also consistent with previous OAE studies that showed better test performance for multivariate analyses compared to univariate analyses (e.g., Dorn et al. 1999; Hussain et al. 1998; Vinck et al. 1998; Gorga et al. 2005). It should be noted, however, that the amount of improvement that is obtained with multivariate analysis depends on the combination of variables used (e.g., emission amplitude and SNR, results from multiple levels, results from multiple frequencies). A_{ROC} values for the

validation-data set were less than those for the training-data set. This observation is not surprising because the multivariate model equations used in the test-performance evaluation were based on the training-data set. Although using PCA for data reduction resulted in lower A_{ROC} for the training set, this method resulted in more generalizable test performance (lower A_{ROC}) than the $\frac{1}{2}$ -octave bands method. Figure 10 compares test performance of CR to that of SFOAE (Ellison and Keefe, 2005), CEOAE (Mertes and Goodman, 2013) and DPOAE (Kirby et al. 2011). The best CR test performances for univariate analysis (training-data set at 30 dB SPL) and multivariate analysis (training data-set at 50 dB SPL) are reproduced from Figs. 6 and 7, respectively. Comparisons are made to our training data because the other studies only included one set of data, which is equivalent to our training data, and these studies did not validate their results using independent data sets. Univariate analysis was used in the SFOAE study of Ellison and Keefe and in the CEOAE study of Mertes and Goodman. Ellison and Keefe evaluated multivariate analysis and obtained results similar to those of their univariate analysis, so we will only compare our data to their univariate results. Multivariate analysis was used in the DPOAE study of Kirby et al. (2011). Univariate CR test performance was similar to performance for the other OAEs that utilized univariate analysis, but only over the frequency range of 1.5–3 kHz. Outside this range univariate CR test performance was poor. As mentioned previously, multivariate analysis improved CR test performance and made it better than test performances for SFOAE and CEOAE. Multivariate CR test performance was also similar to multivariate DPOAE test performance, except for frequencies ≤ 1 kHz and ≥ 6 , where DPOAEs performed better.

CR test performance, like reliability, was poor at low and high frequencies. This is a consequence of the low CR SNR at these frequencies, as was seen in Fig. 4. The SNR was further explored, as shown in Fig. 11, in which CR SNR is plotted as a function of frequency for each stimulus level for the training-data set. Separate plots are shown for participants with NH (solid lines) and for participants with HL (dashed lines). The SNR is low (less than 3 dB) at low frequencies (<0.5 kHz for most levels), and at high frequencies (> 8 kHz). The poor SNR at low frequencies has been observed in measurements of other types of OAEs and it is mainly due to breathing noise and movement by the test participants (e.g., Moleti and Sisto 2008; Gorga et al. 2011; Rasetshwane et al. 2013). The reasons for the poor SNR at frequencies above 8 kHz is not obvious but may relate to as yet undetermined characteristics of the sound-delivery system.

Another observation from Fig. 11 is that the SNR for HI participants was poorer than that for NH participants at all frequencies, except between 4 and 8 kHz, where there was a rise in the SNR for HI participants. The reasons for the latter observation are not yet known. A possible reason for the poorer SNR of HI participants is that this group consisted mostly of older participants (average age of 57 years vs. average age of 30 for the NH group) who might have produced higher noise due to louder breathing. In combination with the reduced CRM in HI ears, the higher noise level probably accounts for the previously described observation that CR reliability was less in HI ears compared to NH ears.

DPOAE and TEOAE test performance has benefited from many years of research, whereas this study represents the initial assessment of CR test performance. Several studies have investigated factors that influence OAE test performance, dating back at least 18 years (e.g.,

Prieve et al. 1993; Whitehead et al. 1995; Stover et al. 1996; Fitzgerald and Prieve 2005; Sanford et al. 2009; Johnson et al. 2010; Kirby et al. 2011). Kirby et al. (2011) provided a summary of factors influencing DPOAE test performance, including (1) stimulus level, (2) relationship among primary levels and frequencies, (3) relative contributions of distortion and coherent-reflection source generators, (4) the particular distortion product that is measured, (5) calibration, (6) middle-ear transfer, and (7) data-analysis approach. While some of the factors that are relevant to measurement of CR were incorporated in the present study, our CR study has not benefited from as thorough an exploration of these factors which are known to influence test-performance for other OAE measurements.

The relationship between CRM and audiometric thresholds is such that ears with better thresholds produce larger CRM (see Fig. 5). Results of threshold prediction (Fig. 8 and 9) showed that (1) no single stimulus level results in the best accuracy (lowest *SE* or highest correlation coefficient) across all frequencies, (2) the accuracy was best for levels from 20 to 40 dB SPL, (3) the validation-data set resulted in lower prediction accuracy and more variability compared to the training-data set, and (4) data reduction using PCA resulted in more generalizable test performance, compared to ½-octave bands method. These results are, in general, consistent with the results of the multivariate test-performance analyses. Audiometric thresholds were predicted with less accuracy by CR than by other types of OAEs. Values of *SE* between CR-predicted and measured thresholds were in general higher than values reported by Rogers et al. (2010) for DPOAEs and Vinck et al. (1998) for TBOAEs and CEOAEs. Values of R^2 were less than those reported by Rogers et al. (2010) for DPOAEs, Vinck et al. (1998) for TBOAEs and CEOAEs, and Mertes and Goodman (2013) for CEOAEs. It is worth noting that at the present moment, even measurements of standard OAEs (i.e., DPOAEs and TEOAEs) are not used in the clinic to predict audiometric thresholds. This suggests that further efforts are needed, for both OAEs and CR, before these measurements can be applied in the clinic for this purpose.

Overall, levels of 20 to 50 dB SPL produced the best results for reliability, test performance and threshold prediction. However, the level that produced the best results varied with frequency, i.e. there was no single level that produced the best α , *SEM*, *AD*, A_{ROC} , *SE* or R^2 values across all frequencies. Both multivariate CR test performance and threshold prediction are not as accurate as results obtained with DPOAEs, a class of OAEs for which test performance and threshold prediction have been studied extensively. However, with further development, CR could be proven to have clinical utility. Potential strategies include a combination of (1) a better training model, (2) larger data set, (3) further refinements to the methods for data collection and (4) further refinements to the analysis methods to extract CR from ear-canal reflectance. Refinement of methods for data collection can include increased averaging time or artifact rejection procedures to improve SNR. Refinements to CR analysis procedures may include adjustments to the definitions of the boundary used to define the region associated with CR in the time-frequency domain (see Fig. 1).

CONCLUSION

In summary, CR measurements evoked using wideband noise stimuli are repeatable between test sessions. The relationship between CRM and audiometric threshold is such that, in

general, NH ears have higher CRM than HI ears. The ability of CR to distinguish between NH and HI ears and to predict audiometric thresholds is less accurate than other classes of OAE measurements. Overall, the best accuracy was obtained at levels of 20–50 dB SPL and at mid frequencies. Future efforts should focus on these stimulus conditions. With refinement to the methods for data collection and analysis, CR may eventually be proven to have clinical utility.

Acknowledgments

This research was supported by grants R01 DC8318 and P30 DC4662 from the NIH.

References

- Allen, JB. Measurement of eardrum acoustic impedance. In: Allen, JB.; Hall, JL.; Hubbard, A.; Neely, ST.; Tubis, A., editors. *Peripheral Auditory Mechanisms*. Springer-Verlag; New York: 1986.
- Allen, JB. Derecruitment by multiband compression in hearing aids. In: Jesteadt, W., editor. *Modeling Sensorineural hearing loss*. Lawrence Erlbaum Associates; New Jersey: 1997. p. 99-112.
- Boege P, Janssen T. Pure-tone threshold estimation from extrapolated distortion product otoacoustic emissions I/O-functions in normal and cochlear hearing loss ears. *J Acoust Soc Am*. 2002; 111:1810–1818. [PubMed: 12002865]
- Brownell WE. Outer hair cell electromotility and otoacoustic emissions. *Ear Hear*. 1990; 11:82–92. [PubMed: 2187727]
- Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951; 16:297–334.
- de Boer E. Connecting frequency selectivity and nonlinearity for models of the cochlea. *Audit Neurosci*. 1997; 3:377–388.
- Dorn PA, Piskorski P, Gorga MP, et al. Predicting audiometric status from distortion product otoacoustic emissions using multivariate analyses. *Ear Hear*. 1999; 20:149–163. [PubMed: 10229516]
- Ellison JC, Keefe DH. Audiometric predictions using stimulus-frequency otoacoustic emissions and middle ear measurements. *Ear Hear*. 2005; 26:487–503. [PubMed: 16230898]
- Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*. 2006; 27:861–874.
- Fitzgerald TS, Prieve BA. Detection of hearing loss using $2f_2-f_1$ and $2f_1-f_2$ distortion-product otoacoustic emissions. *J Speech Lang Hear Res*. 2005; 48:1165–1186. [PubMed: 16411804]
- Franklin DJ, McCoy MJ, Martin GK, et al. Test/retest reliability of distortion-product and transiently evoked otoacoustic emissions. *Ear Hear*. 1992; 13:417–29. [PubMed: 1487104]
- Gorga MP, Neely ST, Bergman B, et al. A comparison of transient-evoked and distortion product otoacoustic emissions in normal-hearing and hearing-impaired participants. *J Acoust Soc Am*. 1993; 94:2639–2648. [PubMed: 8270740]
- Gorga MP, Dierking DM, Johnson TA, et al. A validation and potential clinical application of multivariate analyses of distortion-product otoacoustic emission data. *Ear Hear*. 2005; 26:593–607. [PubMed: 16377995]
- Gorga MP, Neely ST, Kopun J, et al. Growth of suppression in humans based on distortion-product otoacoustic emission measurements. *J Acoust Soc Am*. 2011; 129:801–816. [PubMed: 21361439]
- Goodman SS, Fitzpatrick DF, Ellison JC, et al. High-frequency click-evoked otoacoustic emissions and behavioral thresholds in humans. *J Acoust Soc Am*. 2009; 125:1014–1032. [PubMed: 19206876]
- Hohmann V. Frequency analysis and synthesis using a gammatone filterbank. *Acustica*. 2002; 88:433–442.
- Hussain DM, Gorga MP, Neely ST, et al. Transient evoked otoacoustic emissions in patients with normal hearing and in patients with hearing loss. *Ear Hear*. 1998; 19:434–449. [PubMed: 9867292]

- Jedrzejczak WW, Kochanek K, Trzaskowski H, et al. Tone-burst and click-evoked otoacoustic emissions in participants with hearing loss above 0.25, 0.5, and 1 kHz. *Ear Hear.* 2012; 33:757–767. [PubMed: 22710662]
- Johnson TA, Neely ST, Kopun JG, et al. Clinical test performance of distortion-product otoacoustic emissions using new stimulus conditions. *Ear Hear.* 2010; 31:74–83. [PubMed: 19701088]
- Jolliffe, IT. *Principal Component Analysis*. 2. Springer-Verlag; New York: 2002.
- Keefe DH, Ling R, Bulen RC. Method to measure acoustic impedance and reflection coefficient. *J Acoust Soc Am.* 1992; 91:470–485. [PubMed: 1737890]
- Keefe DH, Goodman SS, Ellison JC, et al. Detecting high-frequency hearing loss with click-evoked otoacoustic emissions. *J Acoust Soc Am.* 2011; 129:245–261. [PubMed: 21303007]
- Kemp DT, Bray P, Alexander L, Brown AM. Acoustic emission cochleography—practical aspects. *Scand Audiol Suppl.* 1986; 25:71–95. [PubMed: 3472324]
- Kirby BJ, Kopun JG, Tan H, et al. Do “optimal” conditions improve distortion product otoacoustic emission test performance? *Ear Hear.* 2011; 32:230–237. [PubMed: 21057318]
- Kumar UA, Methi R, Avinash MC. Test/retest repeatability of effect contralateral acoustic stimulation on the magnitudes of distortion product ototacoustic emissions. *Laryngoscope.* 2013; 123:463–471. [PubMed: 23161444]
- Maat B, Wit HP, Van Dijk P. Noise-evoked otoacoustic emissions in humans. *J Acoust Soc Am.* 2000; 108:2272–2280. [PubMed: 11108368]
- Marshall L, Heller LM. Reliability of transient-evoked otoacoustic emissions. *Ear Hear.* 1996; 17:237–254. [PubMed: 8807266]
- Martin GK, Stagner BB, Lonsbury-Martin BL. Evidence for basal distortion-product otoacoustic emission components. *J Acoust Soc Am.* 2010; 127:2955–2972. [PubMed: 21117746]
- McPherson B, Li SF, Shi BX, et al. Neonatal hearing screening: Evaluation of tone-burst and click-evoked otoacoustic emission test criteria. *Ear Hear.* 2006; 27:256–262. [PubMed: 16672794]
- Mertes IB, Goodman SS. Short-latency transient-evoked otoacoustic emissions as predictors of hearing status and thresholds. *J Acoust Soc Am.* 2013; 134:2127–2135. [PubMed: 23967943]
- Moleti A, Sisto R. Comparison between otoacoustic and auditory brainstem response latencies supports slow backward propagation of otoacoustic emissions. *J Acoust Soc Am.* 2008; 123:1495–1503. [PubMed: 18345838]
- Neely, ST.; Liu, Z. EMAY: Otoacoustic emission averager, Technical Memo No 17. Boys Town National Research Hospital; Omaha, NE: 1994.
- Neely, ST.; Siegel, JH. Comparison of sound sources for ear canal measurements. poster presented at the 34th Annual Midwinter Research Meeting of the Association for Research in Otolaryngology; Baltimore, MD. 2011.
- Patterson, RD.; Holdsworth, J. A functional model of neural activity patterns and auditory images. In: Ainsworth, AW., editor. *Advances in Speech, Hearing and Language Processing*. JAI Press; London: 1996. p. 547-563.
- Prieve BA, Gorga MP, Schmidt A, et al. Analysis of transient-evoked otoacoustic emissions in normal-hearing and hearing-impaired ears. *J Acoust Soc Am.* 1993; 93:3308–3319. [PubMed: 8326059]
- Rasetshwane DM, Neely ST. Inverse solution of ear-canal area function from reflectance. *J Acoust Soc Am.* 2011; 130:3873–3881. [PubMed: 22225043]
- Rasetshwane DM, Neely ST. Measurements of wide-band cochlear reflectance in humans. *J Assoc Res Otolaryngol.* 2012; 13:591–607. [PubMed: 22688355]
- Rasetshwane DM, Argenyi M, Neely ST, et al. Latency of tone-burst-evoked auditory brain stem responses and otoacoustic emissions: Level, frequency, and rise-time effects. *J Acoust Soc Am.* 2013; 133:2803–2817. [PubMed: 23654387]
- Rasetshwane, DM.; Liu, Y.; Neely, ST. Active contribution to ear-canal reflectance in a model of cochlear mechanics. poster presented at the 37th Annual Midwinter Research Meeting of the Association for Research in Otolaryngology; San Diego, CA. 2014.
- Reuven ML, Neely ST, Kopun JG, et al. Effect of calibration method on distortion-product otoacoustic emission measurements near 4 kHz. *Ear Hear.* 2013 In Press.

- Rogers AR, Sienna RB, Kopun JG, et al. Influence of calibration method on distortion-product otoacoustic emission measurements: II. Threshold prediction. *Ear Hear.* 2010; 31:546–554. [PubMed: 20458245]
- Sanford CA, Keefe DH, Liu Y, et al. Sound-conduction effects on distortion-product otoacoustic emission screening outcomes in newborn infants: Test performance of wideband acoustic transfer functions and 1-kHz tympanometry. *Ear Hear.* 2009; 30:635–652. [PubMed: 19701089]
- Shera CA, Guinan JJ. Stimulus-frequency-emission group delay: a test of coherent reflection filtering and a window on cochlear tuning. *J Acoust Soc Am.* 2003; 113:2762–2772. [PubMed: 12765394]
- Shera CA, Guinan JJ, Oxenham AJ. Otoacoustic estimation of cochlear tuning: validation in the chinchilla. *J Assoc Res Otolaryngol.* 2010; 11:343–365. [PubMed: 20440634]
- Stover L, Gorga MP, Neely ST, et al. Toward optimizing the clinical utility of distortion product otoacoustic emissions. *J Acoust Soc Am.* 1996; 100:956–967. [PubMed: 8759949]
- Swets J. Measuring the accuracy of diagnostic systems. *Science.* 1988; 240:1285–1293. [PubMed: 3287615]
- Thorson MJ, Kopun JG, Neely ST, et al. Reliability of distortion-product otoacoustic emissions and their relation to loudness. *J Acoust Soc Am.* 2012; 131:1282–1295. [PubMed: 22352502]
- Vinck BM, Van Cauwenberge PB, Corthals P, et al. Multi-variant analysis of otoacoustic emissions and estimation of hearing thresholds: Transient evoked otoacoustic emissions. *Audiology.* 1998; 37:315–334. [PubMed: 9888189]
- Voss SE, Allen JB. Measurement of acoustic impedance and reflectance in the human ear canal. *J Acoust Soc Am.* 1994; 95:372–384. [PubMed: 8120248]
- Voss SE, Merchant GR, Horton NJ. Effects of middle-ear disorders on power reflectance measured in cadaveric ear canals. *Ear Hear.* 2012; 33:195–208. [PubMed: 22037477]
- Wagner W, Heppelmann G, Vonthein R, et al. Test-retest repeatability of DPOAEs. *Ear Hear.* 2008; 29:378–391. [PubMed: 18382378]
- Whitehead ML, McCoy MJ, Lonsbury-Martin BL, et al. Dependence of distortion product otoacoustic emissions on primary levels in normal and impaired ears: I. Effects of decreasing L_2 below L_1 . *J Acoust Soc Am.* 1995; 97:2346–2358. [PubMed: 7714254]
- Zweig G. Finding the impedance of the organ of Corti. *J Acoust Soc Am.* 1991; 89:1229–1254. [PubMed: 2030212]

Fig. 1.

Illustration of CR in the time-frequency domain for levels of 20–70 dB SPL. The units of the color map are dB SPL. The region of high energy of the gammatone spectrogram enclosed by $N_L = 4$ cycles, $N_H = 40$ cycles, $t_H = 0.5$ ms and $t_L = 30$ ms includes the energy that we associate with CR. The region below N_L includes residual middle-ear and measurement-system activity that was not removed by the subtraction procedure. The activity beyond $t_L = 30$ ms is due to re-reflection of the traveling wave. CR can be observed at the four lowest stimulus levels, and its magnitude decreases with increasing stimulus level. There is no evidence of CR at the highest level. CR persists longer at lower stimulus levels than at higher levels; there is more re-reflected energy close to $t_L = 30$ ms at the lowest level (from Rasetshwane and Neely, 2012).

Fig. 2.

Assessment of reliability using Cronbach's α (top row), standard error of measurement (SEM ; middle row) and absolute differences (AD ; bottom row). The columns from left to right show results for all participants, NH participants and HI participants. Values of mean and SD (across frequency and stimulus level) are provided as insets. Good reliability is considered to be $\alpha > 0.85$ and $SEM < 2$ (both indicated using dashed lines).

Fig. 3.

Histogram of the mean (across frequency and level) signed difference in CRM between test sessions. The signed differences in CRM between test sessions were roughly normally distributed with a mean of 0.0 dB and SD of 2.9 dB. Ninety percent of the data were within the interval -4.6 to 4.4 dB.

Fig. 4.

Cochlear reflectance magnitude (CRM) for NH participants (thick solid line) and HI participants (thick dashed line). Estimates of the noise magnitude for NH participants (thin dash-dot line) and HI participants (thin dashed line) are also shown. CRM for NH participants is greater than CRM for HI participants, except in the low (< 1 kHz) and high (> 8 kHz) frequencies where CRM is indistinguishable from the noise. Also, the noise level for NH participants is lower than that for HI participants.

Fig. 5.

Relationship between audiometric thresholds and CRM. Columns present data at different frequencies, as indicated by labels at the top. Rows present data at different stimulus levels, as indicated by labels on the right. The open circles are the individual data and the lines are the simple regression of thresholds to CRM. Values of R^2 for the simple regression lines are included as insets in the figure panels. The regression lines have negative slope, indicating that participants with better hearing have larger CRM.

Fig. 6.

Univariate test performance. The top panel shows area under the receiver operating characteristic curve (A_{ROC}) for training-data set and the middle panel shows A_{ROC} for validation-data set, both as functions of frequency. The difference in test performance between the two data sets (ΔA_{ROC}) is shown in the bottom panel. Stimulus level is the parameter in the figure panels.

Fig. 7. Multivariate test performance. Top row shows A_{ROC} for the training-data set and middle row shows A_{ROC} for the validation-data set as functions of frequency. The difference in test performance between the two data sets (ΔA_{ROC}) is shown in the bottom row. The left and right columns show results where data reduction used $1/2$ -octave bands analysis and PCA, respectively. Stimulus level is the parameter in the figure panels.

Fig. 8. Standard error (*SE*) for threshold prediction. Arrangement of results is as described in Fig. 7. $\frac{1}{2}$ -octave band analysis results in lower *SE* for the training-data set, but PCA results in more generalizable threshold predictions (lower *SE*).

Fig. 9.

R^2 for threshold prediction. Arrangement of results is as described in Fig. 7. $\frac{1}{2}$ -octave band analysis results in higher correlation for the training-data set, but PCA results in more generalizable threshold predictions (R^2 closer to 0).

Fig. 10.

Comparison of test performance of CR to that of SFOAE (Ellison and Keefe, 2005), CEOAE (Mertes and Goodman, 2013) and DPOAE (Kirby et al. 2011).

Fig. 11.

Average SNR (across participants) of cochlear reflectance as function of frequency. Stimulus level is the parameter in the figure. Separate plots are shown for NH (solid lines) and HI participants (dashed lines). The average SNR for NH participants is higher than that for HI participants. The average SNR for both groups is low at low (<0.5 kHz) and high frequencies (> 8 kHz).

TABLE 1

Distribution of audiometric thresholds by frequency.

	Frequency (kHz)									
	0.25	0.5	0.75	1	1.5	2	3	4	6	8
Threshold (dB HL)	-5	1	0	1	0	1	0	1	2	3
	0	8	7	1	6	4	10	12	7	8
	5	16	12	16	15	14	10	6	12	5
	10	15	17	16	13	10	6	7	11	6
	15	15	15	15	8	7	6	7	4	5
	20	10	8	4	6	5	8	4	2	4
	25	7	5	6	7	9	5	1	4	4
	30	8	9	3	6	8	5	5	2	1
	35	4	4	7	2	4	8	6	4	3
	40	2	5	6	9	5	7	5	4	2
	45	1	2	4	3	5	6	6	4	1
	50	1	1	3	4	3	4	5	3	5
	55	2	1	2	3	6	2	8	8	6
	60	0	1	2	3	1	3	4	9	12
	65	0	0	0	1	3	4	4	9	8
	70	0	3	1	0	1	4	9	7	2
	75	0	0	2	2	0	1	1	2	5
	80	0	0	1	2	4	0	0	0	3
	85	0	0	0	0	0	1	0	0	3
	90	0	0	0	0	0	0	0	0	2

TABLE 2

Stimulus type, level, number of averages per sound source and test time.

Test	Stimulus	Level (dB SPL)	Number of averages	Test time (s)
1	WBN	20	64	192
2	WBN	30	32	96
3	WBN	40	16	48
4	WBN	50	8	24
5	WBN	60	4	12
6	WBN	70	2	6
7	LSC	60	2	6

Number of ears classified as normal hearing (NH) and hearing impaired (HI) at each frequency for training- and validation-data sets. For analyses of test performance, ears were classified as NH or HI on a frequency-by-frequency basis.

TABLE 3

		Frequency									
		0.25	0.5	0.75	1	1.5	2	3	4	6	8
Training	NH	35	34	31	26	25	22	22	23	22	22
	HI	23	24	27	32	33	36	36	35	36	36
Validation	NH	20	17	18	16	11	10	10	9	9	10
	HI	12	15	14	16	21	22	22	23	23	22