



Objective Assessment System for Hearing Prediction Based on Stimulus-Frequency Otoacoustic Emissions

Trends in Hearing
Volume 25: 1–19
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/23312165211059628
journals.sagepub.com/home/tia


Qin Gong^{1,2} , Yin Liu¹, Runyi Xu¹, Dong Liang¹,
Zewen Peng¹ and Honghao Yang¹

Abstract

Stimulus-frequency otoacoustic emissions (SFOAEs) can be useful tools for assessing cochlear function noninvasively. However, there is a lack of reports describing their utility in predicting hearing capabilities. Data for model training were collected from 245 and 839 ears with normal hearing and sensorineural hearing loss, respectively. Based on SFOAEs, this study developed an objective assessment system consisting of three mutually independent modules, with the routine test module and the fast test module used for threshold prediction and the hearing screening module for identifying hearing loss. Results evaluated via cross-validation show that the routine test module and the fast test module predict hearing thresholds with similar performance from 0.5 to 8 kHz, with mean absolute errors of 7.06–11.61 dB for the routine module and of 7.40–12.60 dB for the fast module. However, the fast module involves less test time than is needed in the routine module. The hearing screening module identifies hearing status with a large area under the receiver operating characteristic curve (0.912–0.985), high accuracy (88.4–95.9%), and low false negative rate (2.9–7.0%) at 0.5–8 kHz. The three modules are further validated on unknown data, and the results are similar to those obtained through cross-validation, indicating these modules can be well generalized to new data. Both the routine module and fast module are potential tools for predicting hearing thresholds. However, their prediction performance in ears with hearing loss requires further improvement to facilitate their clinical utility. The hearing screening module shows promise as a clinical tool for identifying hearing loss.

Keywords

stimulus-frequency otoacoustic emissions, objective assessment system, threshold prediction, hearing screening

Introduction

Audiometric thresholds are the current gold standard for quantitatively evaluating the degree of hearing loss. Pure tone audiometry (PTA) requires subjective responses from individuals and is susceptible to confounding factors such as attention and mental state, making it challenging to apply to certain populations who cannot provide reliable responses to sounds. Electrophysiological measures, such as auditory brainstem response (ABR), can objectively estimate hearing thresholds (Gorga et al., 2006). However, accurate estimates of hearing thresholds from ABR generally rely on skilled human interpretation of recorded responses, making ABR testing expensive and time-consuming (Mertes & Goodman, 2013). It is therefore worthwhile to investigate the accuracy of predicting hearing thresholds from other potentially available objective measures.

Generated as a by-product of the normal function of outer hair cells (OHCs) within the cochlea (Brownell, 1990; Kemp,

1978; Shera & Guinan, 1999), otoacoustic emissions (OAEs) can be useful tools for the non-invasive assessment of cochlear function (for review see (Robinette & Glatcke, 2007)). When cochlear damage that affects the OHCs exists, elevated hearing thresholds, as well as reduced or absent OAEs, are measured. These observations lead to the application of OAE measurements in identifying hearing loss. OAEs are appealing as they can obtain information about the health and integrity of the cochlea and sensory cells. Moreover, OAE testing is performed by placing a

¹Department of Biomedical Engineering, Tsinghua University, Beijing, China
²School of Medicine, Shanghai University, Shanghai, China

Corresponding author:

Qin Gong, Office C212, School of Medicine, Tsinghua University, Beijing 100084, China.
Email: gongqin@mail.tsinghua.edu.cn



small probe in the ear canal, which is non-invasive, affordable and easy to perform (Mertes & Goodman, 2013). Distortion-product OAEs (DPOAEs) and transient-evoked OAEs (TEOAEs) have been routinely measured in clinical settings and are widely used in universal newborn hearing screening and differential diagnostics. A large number of studies indicate that DPOAEs and TEOAEs can identify hearing status (normal hearing vs. hearing loss) (Go et al., 2019; Gorga et al., 1993a; b; Hurley & Musiek, 1994; Mertes & Goodman, 2013; Prieve et al., 1993; Stover et al., 1996). In addition to this simple dichotomous decision, other studies (Gorga et al., 2003; Johnson et al., 2007) proposed an approach that permits individual threshold prediction from DPOAE input/output (I/O) functions. However, large standard errors between the predicted and measured hearing thresholds are still present in these studies. Improved quantitative predictions of hearing thresholds from OAEs are of continued interest in clinical applications.

Stimulus-frequency OAEs (SFOAEs) are sound signals evoked by tonal probes and originate at the same place as the probe in the cochlea (Kemp & Chum, 1980). They are thought to provide most place-specific responses among OAEs (Charaziak et al., 2013; Shera & Guinan, 1999). Moreover, SFOAEs predominantly arise as reflections from a localized region near the peak of the traveling wave (Zweig & Shera, 1995), which are easier to interpret than DPOAEs at the cubic difference frequency, $f_{DP} = 2f_1 - f_2$, whose generation involves the mixing of linear coherent reflection and nonlinear distortion mechanisms (note that source-separated DPOAEs become more place-specific and easy to interpret, but OAE unmixing is beyond the scope of this study). For these reasons, SFOAEs were chosen to investigate their potential as an audiometric prediction tool. Although SFOAEs have been widely investigated as non-invasive probes of cochlear function in humans (Abdala et al., 2019; Charaziak et al., 2013; Kalluri & Abdala, 2015; Keefe et al., 2008; Lineton & Lutman, 2003; Schairer et al., 2006; Shera et al., 2002; Shera & Guinan, 2003), their clinical utility remains limited owing to the complex measurement paradigms (Kalluri & Shera, 2013) and no clinical instruments used to record SFOAEs, as well as a dearth of data relating SFOAEs to hearing thresholds and status. SFOAE measurements have been demonstrated to have potential as a place-specific tool for identifying hearing status at octave frequencies from 0.5–8 kHz (Ellison & Keefe, 2005; Go et al., 2019). Additionally, SFOAEs are significantly correlated with pure-tone thresholds (Ellison & Keefe, 2005). In terms of microstructures, strikingly similar patterns are observed in behavioral hearing thresholds and the amplitudes and delays of SFOAEs. The periodicity and magnitude of this common microstructure have been found to be related to the SFOAE phase-gradient delay and amplitude, respectively (Dewey & Dhar, 2017a). Despite sparse investigations into the relationships between SFOAEs and pure-tone thresholds, these

findings open the possibility that SFOAEs may be a useful audiometric prediction tool.

Machine learning approaches such as support vector machine (SVM), k-nearest neighbor (KNN), back-propagation neural network (BPNN), decision tree, and random forest excel at developing models from large, complex, and information-rich data sets, and are highly effective in solving many complex nonlinear problems. They can automatically learn rules from the input data and then predict the unknown data. Until recently, machine learning techniques have been widely applied to predict sudden sensorineural hearing loss (Bing et al., 2018), noise-induced hearing impairment (Zhao et al., 2019), and sensorineural hearing loss (SNHL) in different inner ear pathologies (Shew et al., 2019). They have been demonstrated to be powerful tools for predicting various types of hearing loss. A preliminary study from our laboratory used BPNN to investigate the ability of SFOAEs to predict hearing thresholds and status for the first time (Gong et al., 2020). The outcomes for different machine learning algorithms were further compared to maximize the potential of SFOAEs in predicting hearing capabilities (Liu et al., 2020), and we found that BPNN, KNN, and SVM algorithms performed well in such hearing prediction tasks. On the basis of these prior studies, BPNN, KNN, and SVM algorithms were selected as the candidates in this study, and we aimed to directly develop an objective system for the prediction of hearing capabilities and validate it on new unknown data.

In this study, a hardware platform consisting of the main control computer, external sound card, power amplifier, miniature speaker and miniature microphone was constructed. Based on this platform, a system for SFOAE recording was developed. Then, we trained machine learning-based models using a large data set of SFOAEs and behavioral thresholds measured in the same ears, and developed an assessment system that allowed the prediction of hearing thresholds and screening for SNHL at several conventional PTA test frequencies (0.5, 1, 2, 4, and 8 kHz). The SFOAE-based assessment system contained three mutually independent test modules, two of which played the same role, that is, providing a quantitative estimate of hearing threshold, but differed from each other in time efficiency, while the other was designed to make dichotomous decisions in identifying the presence or absence of hearing loss. Finally, all of modules were validated on an extra unknown data set.

Materials and Methods

Instrumental Design

Figure 1 shows a diagram of the system connection (Figure 1A) and the actual hardware (Figure 1B). The system hardware consists of the main control computer, external sound card, power amplifier, miniature speaker

and miniature microphone. A computer-generated digital signal was converted to an analog electrical signal using a 24-bit sound card (Fireface UC, RME, Haimhausen, Germany) with a sampling rate of 48 kHz. This signal was amplified by a custom power amplifier designed and developed by ourselves, whose output was transduced to an acoustic signal through the loudspeakers (ER-2, Etymotic Research, Elk Grove Village, IL, USA) and presented to the ear via tubes. A probe containing miniature loudspeakers and a microphone was inserted into the ear. In the signal acquisition pathway, acoustic signals recorded in the ear canal were transduced into an electrical signal by a low-noise miniature microphone (ER-10B+, Etymotic Research, Elk Grove Village, IL, USA) with an amplification of 20 dB, which was then converted to a digital signal via the sound card, and sent back to the computer. The assessment system was developed using C sharp programming language (Microsoft Inc., Redmond, WA, USA) and MATLAB (MathWorks Inc., Natick, MA, USA). Calibration was conducted with reference to the sound pressure level (SPL) at half octave frequencies from 0.125 to 8 kHz. The probe was inserted into a Brüel & Kjær coupler (IEC 711 standard, Type 4157, Nærum, Denmark) and the SPL was measured by the coupler microphone. With this method, we assumed that the voltage that produces the desired SPL in the coupler would produce the same SPL in the ear canal. The probe microphone was calibrated using the coupler microphone as the reference. Following the calibration procedure, multiple probe tones with different SPLs were presented to the earphone to ensure that the SPLs presented equaled the SPLs recorded by the coupler microphone and the total ear canal SPLs at the microphone of ER-10B+ (error $\leq \pm 1$ dB) at any given frequency. The system delay was measured by calculating the phase difference between a 50-ms pure tone presented and recorded at the ER-10B+ microphone.

SFOAE Recording

Stimuli. SFOAEs were recorded based on the two-tone suppression method. The arrangement of probe and suppressor tones for the acquisition of a single SFOAE is illustrated in Figure 2. The stimuli consisted of six intervals (except for the last 5 ms), with intervals M and N added to the traditional four-interval paradigm to eliminate the effects of system and SFOAE delays. There was one interval of $2T_d$ followed by five intervals of T_w (50 ms) in duration. T_d is the pre-measured system delay with a duration of 14.5 ms. The probe and suppressor tones were delivered by two separate speakers. The probe was a continuous pure tone and had the same polarity at intervals A, B, C, D, and N. The suppressor was a tone burst, with the rise and decay time windowed by a 5-ms cosine window. Each of the two stimuli within each interval was a sinusoidal tone with an integral number of periods. Between the rise and decay time of the suppressor tone, the plateau intensity was kept constant. The suppressor at interval D was inverted relative to interval C.

SFOAE Detection. The pressure responses recorded at intervals A to D were stored in four separate buffers, A to D, respectively. Except for the background noise, the recorded responses in the ear canal consisted of the probe stimulus, R_p , suppressor stimulus, R_s , the SFOAE evoked by the probe, SFE , the SFOAE evoked by the suppressor, SFE_s , and the remaining SFOAE caused by the probe after suppression, SFE' . Both buffers A and B contained R_p and SFE . Buffer C contained R_p , R_s , SFE_s and SFE' , whilst buffer D contained R_p , $-R_s$, $-SFE_s$ and SFE' . The suppressed SFOAE was the subtraction of the pressure responses at intervals (A + B) and (C + D) (see Equation 1) in such a way as to cancel out the R_p , R_s , and SFE_s . If SFOAE could be suppressed completely, whereby SFE' equaled zero, then this only left a residual that equaled the SFOAE in response to the probe. Thus, the SFOAE time waveform was $[(A + B) - (C + D)]/2$.

$$\begin{aligned} \text{residual} &= (A + B) - (C + D) \\ &= (R_p + SFE) + (R_p + SFE) - (R_p + R_s + SFE' \\ &\quad + SFE_s) - (R_p - R_s + SFE' - SFE_s) = 2SFE - 2SFE' \end{aligned} \quad (1)$$

After each acquisition process, a zero-phase shift high-pass filter with 0.5-kHz cut-off frequency for 1–4 kHz measurements and 0.35-kHz cut-off frequency for 0.5 kHz measurement was used to reduce low-frequency background noise. The power spectrum was obtained using Welch's overlapped segment averaging estimator, with each analyzed residual (~ 2410 points) or noise signal windowed by 10 consecutive overlapping 42.7-ms Kaiser windows (2048 points). The number of discrete Fourier transform (DFT) points equaled the sampling rate, i.e., 48000 points. The estimation of the power spectrum was carried out in MATLAB (Ver. MATLAB 2020a, Mathworks) using function "Pw Welch". The extracted power spectra of the SFOAE residual and noise floor are shown in Figure 3. The difference between intervals A and B (A-B) only contained background noise. The trials in which the root-mean-square amplitude of A-B exceeded a set fixed threshold value (0.0008) were rejected in real time to reduce transient artifacts. Additional trial was conducted when the current trial was abandoned, until the pre-defined number of trials were completed. In the absence of SFOAEs, the stimulus pressure response recorded in the ear canal (hereafter referred to as L_{pr}) was calculated as $(C + D)/2$. In the present study, SFOAE transfer function (T_{sf}) magnitude (in dB SPL) was computed as the normalized SFOAE magnitude by subtracting the ear canal sound pressure level of stimulus L_{pr} from the measured SFOAE amplitude.

SFOAE-Based Assessment System for Threshold Prediction and Hearing Screening

Figure 4A shows the framework of the SFOAE-based assessment system consisting of three test modules. The routine test

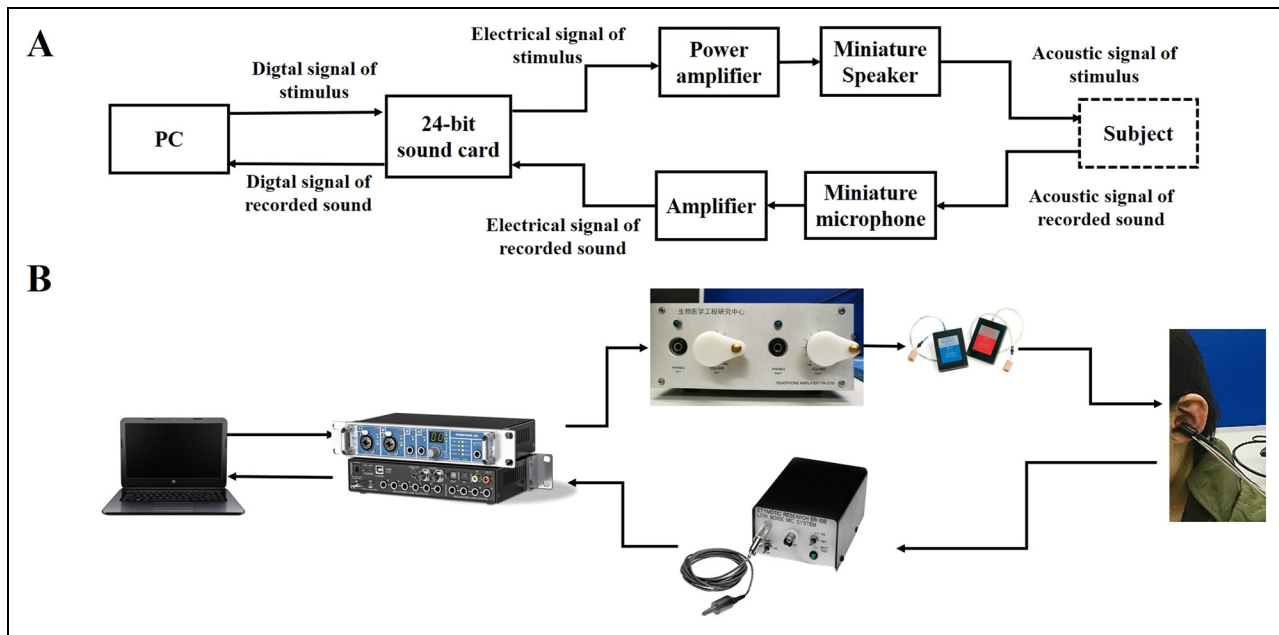


Figure 1. A, schematic connection diagram of the stimulus-frequency otoacoustic emission (SFOAE)-based assessment system. B, The hardware connections.

module provides an objective estimate of hearing threshold from the entire SFOAE I/O function via a machine learning-based regression model, Predictor 1. The fast test module omits SFOAE measurements at low probe levels and terminates when a signal-to-noise ratio (SNR)-based criterion is met, thus completing threshold determination in a relatively short period. It implements threshold prediction by running one of the two regression models, Predictor 2 or Predictor 3, according to whether the SNR-based criterion is met. The purpose of the hearing screening module is to identify hearing status (i.e., to discriminate between normal and impaired ears) from SFOAEs measured at three fixed probe levels (40, 50, and 60 dB SPL) using a trained classifier.

Methods for Model Training. The predictors and classifiers involved in each module are developed based on machine

learning algorithms. The steps of model training are shown in Figure 4B. Data used for model training were collected first. Then feature extraction, described in more detail in an upcoming section, was performed over the large data set, aiming to capture useful information regarding the pure-tone thresholds. Machine learning models must be configured prior to training. These critical configuration variables are called hyperparameters, which typically have a significant impact on the performance of machine learning algorithms. For each candidate machine learning algorithm, hyperparameters are learned through leave-one-out cross-validation (LOOCV) (for the regression models) or k-fold cross-validation (for the classification models), and we selected the optimal combination of model algorithms and hyperparameters for each model. Finally, the trained models were obtained by training on all the collected data for the determined algorithms and hyperparameters.

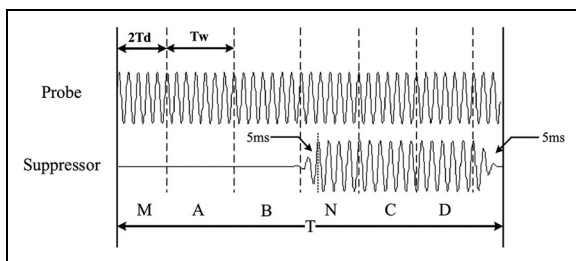


Figure 2. Stimulus synthesis for the acquisition of a single stimulus-frequency otoacoustic emission (SFOAE). The stimuli comprise six intervals. The first interval of $2T_d$ in duration is followed by five intervals of T_w in duration (Gong et al., 2014).

1) Data Collection. Subjects: Data were collected from 245 ears of 131 subjects (66 females) with normal hearing (NH) ($HL \leq 25$ dB HL at octave frequencies from 0.25 to 8 kHz) and 839 ears of 594 subjects (279 females) with SNHL, whose air-conduction (AC) thresholds were > 25 dB HL and ≤ 75 dB HL for at least one octave frequency between 0.5 and 8 kHz. The age for subjects with NH ranged from 18 to 42 years (mean = 23.7, standard deviation [SD] = 4.1) while that for subjects with SNHL was between 12 and 80 years (mean = 47.6, SD = 14.3). All subjects had normal middle ear function, as determined by ≤ 10 dB air-bone gaps and normal 226-Hz tympanometry (defined as peak pressure [PP] between -83 and 0 daPa,

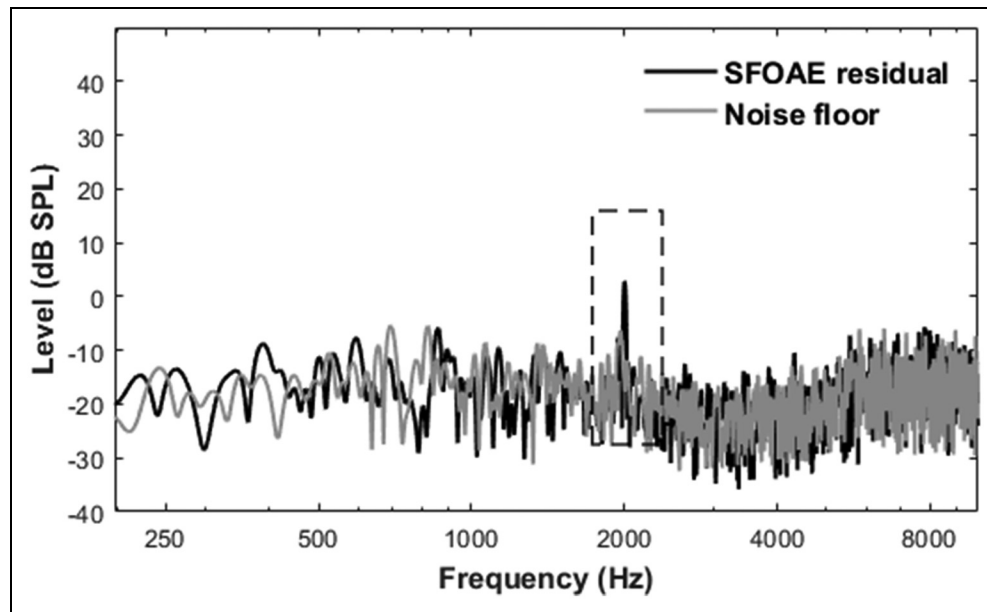


Figure 3. An example of the extracted power spectrum of stimulus-frequency otoacoustic emission (SFOAE) residual (black) and noise floor (gray). The probe frequency is 2 kHz. The black dashed box indicates the presence of the evoked SFOAE.

peak-compensated admittance between 0.3 and 1.4 mmhos, and equivalent ear canal volume (ECV) between 0.6 and 1.5 mL). Table 1 lists the number of ears involved in the tests for each frequency. The classification of NH vs. SNHL was made on a frequency-by-frequency basis for the five test frequencies: 0.5, 1, 2, 4 and 8 kHz; thus, an ear would be classified as NH at some frequencies and SNHL at others. All experiments were carried out in a sound-attenuating chamber. During the OAE test, all subjects were instructed to sit comfortably on a recliner, to sleep or watch silent films with subtitles and to avoid gnashing, chewing, and swallowing to reduce transient noise. The participants were informed of the experimental procedures and objectives, and provided written informed consent. They were given appropriate compensation. All procedures were approved by the institutional review board at Tsinghua University. All data collection was completed by a research assistant within six months.

Procedures: Prior to the test, an external auditory canal examination was performed and cerumen (if present) was removed from the ear canal. Pure-tone AC and bone-conduction thresholds at octave frequencies from 0.25–8 kHz were measured in 5-dB steps using a clinical audiometer (Asteria, Madsen Inc., Denmark). SFOAEs were measured at a fixed probe frequency (F_p) and suppressor frequency ($F_s = F_p - 47$ Hz), while the probe level L_p was increased in 5-dB increments from 5 to 70 dB SPL at 0.5, 1, 2, and 8 kHz and from 5 to 60 dB SPL at 4 kHz. To obtain total suppression, the suppressor level (L_s) equaled 70 dB SPL at the probe levels from 5 to 55 dB SPL, and $L_p + 15$ dB SPL at or above the probe levels of 60 dB SPL. We employed more

averages at lower probe stimulus levels because SFOAEs were difficult to detect under these conditions. The average for each SFOAE response was based on 96 buffers at probe levels of 5–10 dB SPL, 64 buffers at 15–20 dB SPL, and 32 buffers at 25 dB SPL or above. In many previous studies, averaging continued until a target SNR or noise level was achieved, in which the levels of SNR might largely depend on the SFOAE amplitude. In this study however, the number of buffers for averaging remained consistent across subjects to obtain the SNRs under the same condition that might carry valuable information regarding individual hearing thresholds.

2) *Feature Extraction.* Following data collection, feature extraction was performed to capture adequate information regarding hearing thresholds. Table 2 lists the input variables for each predictor and classifier. The input variables for Predictor 1 (Input variables 1) are SFOAE amplitudes, SFOAE SNRs, and T_{sf} magnitudes at all measured probe levels. Slightly differing from Predictor 1, the input variables involved in Predictor 3 (Input variables 3) are SFOAE amplitudes, SFOAE SNRs, and T_{sf} magnitudes at probe levels from 15 to 70 dB SPL (or 60 dB SPL at 4 kHz). Both Predictor 1 and Predictor 3 were trained over all collected SFOAE data, while Predictor 2 was obtained by training data meeting the following SNR-based criterion:

Step 1: The probe level is raised in 5-dB increments from $L_p = L_{min}$ until SFOAE SNR ≥ 9 dB (see the point in the light red shaded area of Figure 5A).

Step 2: The lowest probe level that yields SFOAE SNR ≥ 9 dB is determined as L_p threshold if at least *two* stimulus

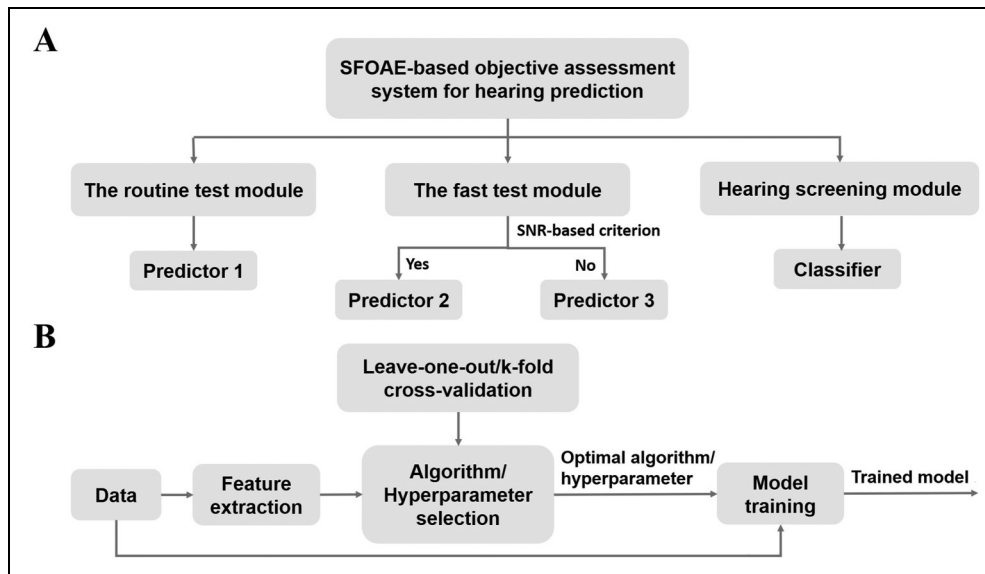


Figure 4. A, The framework of stimulus-frequency otoacoustic emission (SFOAE)-based objective assessment system for hearing prediction, which is composed of the routine test module, the fast test module and hearing screening module. B, The flow chart of a machine-learning model training.

points of the following *three* consecutive stimulus points have $\text{SNR} \geq 6$ dB (see the point in the light purple of Figure 5A). If the number of stimulus points after the candidate, m ($m \geq 1$), is less than *three*, all these points after the candidate are required to meet $\text{SNR} \geq 9$ dB].

Step 3: If the lowest probe level meeting SFOAE $\text{SNR} \geq 9$ dB fails to meet step 2, the latter stimulus points are sequentially checked to find out the new candidate satisfying $\text{SNR} \geq 9$ dB, and the above procedure is repeated until the L_p threshold has been established. If no L_p threshold has been determined until the probe level is raised to the maximum, we assume this ear fails the SNR-based criterion at this frequency (see Figure 5B). The presence of OAE activity is typically defined by $\text{SNR} \geq 6$ dB in clinical applications (Robinette & Glatke, 2007). In many previous studies, a minimum SNR criterion of 6 dB was required for the OAE level to be included in subsequent analyses (e.g., Go et al., 2019; Gorga et al., 2003), while some studies adopted a criterion of $\text{SNR} \geq 9$ dB (Dewey & Dhar, 2017a, 2017b). Based on these commonly used criteria, we redefined the SNR-based criterion as a combination of $\text{SNR} \geq 9$ dB and ≥ 6 dB, with

Table 1. A Summary of the Number of Ears in Each Category for Each Test Frequency.

Category	Frequency (kHz)				
	0.5	1	2	4	8
NH	218	198	206	218	229
SNHL	229	244	239	263	356
Total	447	442	445	481	585

Note: NH = normal hearing; SNHL = sensorineural hearing loss.

the stricter one (i.e., $\text{SNR} \geq 9$ dB) set as the prerequisite for the L_p threshold and the relaxed one (i.e., $\text{SNR} \geq 6$ dB) used as the second condition that allowed a slight reduction in SNR due to the fluctuations of noise floors.

Conceptually similar to the DPOAE thresholds reported previously (Boege & Janssen, 2002; Gorga et al., 2003; Johnson et al., 2007), we considered L_p threshold as the

Table 2. The Input Variables for Each Model.

Module	Model	Input variables
Routine test module	Predictor 1	SFOAE amplitudes; SFOAE SNRs; T_{sf} magnitudes at all measured probe levels
Fast test module	Predictor 2	SFOAE SNRs measured at four probe levels (i.e., at L_p threshold and its three consecutive higher probe levels. If there were less than three probe levels higher than the L_p threshold, the lower probe levels than L_p threshold were used to fill out); L_p threshold; T_{sf} magnitude threshold
	Predictor 3	SFOAE amplitudes, SFOAE SNRs, T_{sf} magnitudes at measured probe levels from 15 dB to 70 (or 60) dB SPL
Hearing screening module	Classifier	SFOAE amplitudes; SFOAE SNRs; T_{sf} magnitudes at 40, 50, 60 dB SPL

Note: SFOAE = stimulus-frequency otoacoustic emission; SNRs = signal-to-noise ratios; T_{sf} = transfer function.

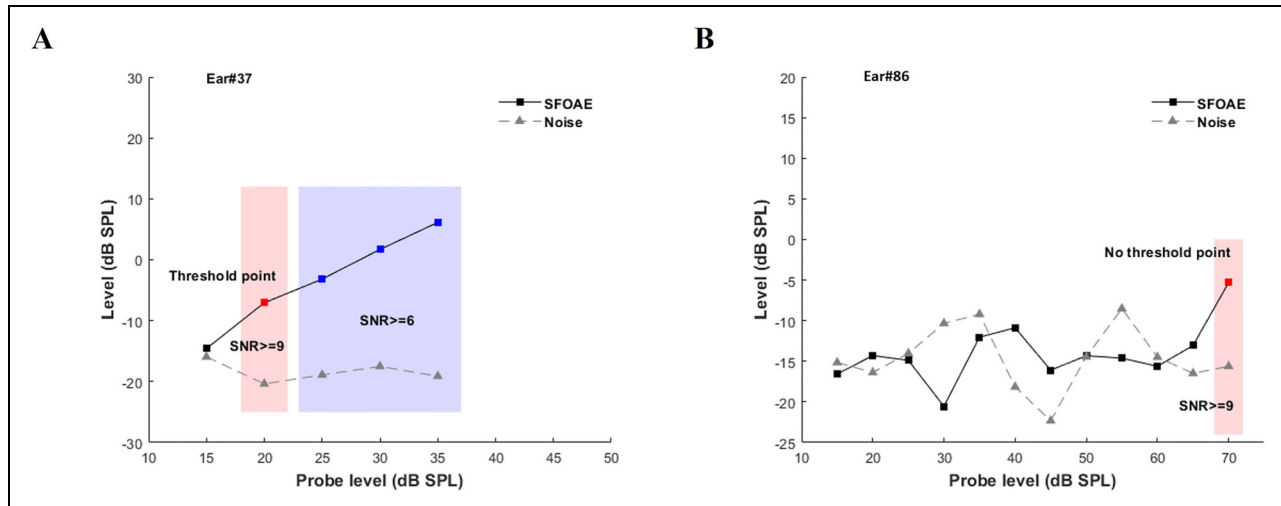


Figure 5. The process of determining L_p threshold (the lowest probe level at which a stimulus-frequency otoacoustic emission (SFOAE) response is detectable) for input/output (I/O) functions. A, a case meeting the signal-to-noise ratio (SNR)-based criterion. The light red shaded area represents the lowest probe level that yields SFOAE SNR ≥ 9 dB (L_p threshold), and the light purple shaded area indicates three points followed by the L_p threshold, in which at least two points are required SNR ≥ 6 dB. B, An example in the absence of L_p threshold given that the lowest probe level meeting SFOAE SNR ≥ 9 dB is at the maximum probe level.

lowest probe level at which a SFOAE response could be detectable. We proposed this SNR-based criterion to determine L_p threshold as much more ears would be excluded when using the inclusion criteria of previous studies (Boege & Janssen, 2002; Gorga et al., 2003). L_p threshold, T_{sf} magnitude threshold (i.e., T_{sf} magnitude at the L_p threshold), and SFOAE SNRs measured at four probe levels (i.e., at L_p threshold and its three consecutive higher probe levels. If there are less than three probe levels higher than the L_p threshold, the lower probe levels than L_p threshold are used to fill out) are collectively taken as the inputs to Predictor 2 (Input variables 2) for threshold prediction. For the classifier contained in the hearing screening module, SFOAE amplitudes, SFOAE SNRs, T_{sf} magnitudes at 40, 50, and 60 dB SPL are used as the input variables. Each predictor or classifier predicts hearing thresholds or status based on SFOAEs measured at a frequency equal to the audiometric frequency. Classification is based on a WHO-defined normal-hearing criterion of ≤ 25 dB HL (i.e., ≤ 25 dB HL = normal).

3) Algorithm/Hyperparameter Selection. A LOOCV (for Predictor 1 in the routine test module and Predictor 2–3 in the fast test module) or k -fold cross-validation (for Classifier in the hearing screening module, $k = 5$) was conducted for model training and validation to select the optimal combination of machine learning algorithms and hyperparameters. In k -fold cross-validation, the dataset is divided into k approximately equal-sized disjoint folds, where a fold is in turn omitted for validating the model trained by other $k-1$ folds. LOOCV is a special case of k -fold cross-validation with k equal to the number of observations in the dataset, n (Cheng et al., 2017). In one of the k runs for LOOCV, each

instance is, in turn, a single-item test set only once to validate the model trained by all other instances. For the three predictors (Predictor 1–3), LOOCV is appealing as the size of the training set is maximized in such a way that the trained models achieve better performance. The test performance was evaluated as the mean accuracy or error of all n observations when individually treated as a single-item test set. It is worth noting that the analysis was based on frequency; therefore, there was a separate model for each frequency.

Two widely used machine learning algorithms, BPNN and KNN, are alternatives to develop the predictors for threshold prediction (Predictor 1–3) and the classifier for hearing screening (Classifier). Another machine learning approach, SVM, is also a candidate for building the classifier in the hearing screening module. Figures 6A–C show the structures of BPNN, KNN, and SVM algorithms, respectively. As shown in Figure 6A, the BPNN model consists of an input layer, a hidden layer and an output layer. The number of nodes in the input layer equals the number of input variables. One node representing the predicted hearing threshold is used in the output layer of the BPNN-based regression models for threshold prediction, while two nodes indicating NH vs. hearing loss are employed in the output layer of the BPNN-based classification models for hearing screening. BPNN training involves forward propagation of the operating signal and back propagation of the error signal. The continuous adjustment of the weights is applied to make the actual output closer to the expected one, until the error is reduced to a set minimum value or the training steps are reached, and then the weights are fixed. In this study, the number of nodes in the hidden layer must be optimized as the hyperparameter for BPNN models. The KNN

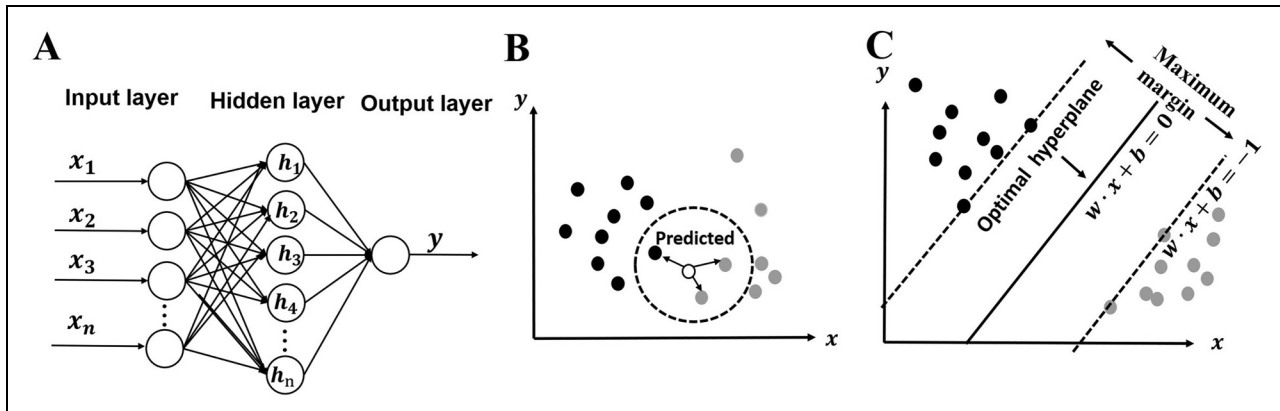


Figure 6. A. The structure of the back-propagation (BP) neural network algorithm. B. The schematic of the k nearest neighbor (KNN) algorithm. C. The schematic of the support vector machine (SVM) algorithm.

classifier uses the labels (i.e., normal or impaired) of the k nearest neighbors of the predicted sample (open circle) vote to determine its category (see Figure 6B). The predicted sample is classified into the category of the majority of the k nearest neighbors. Similarly, the mean hearing threshold of the k nearest neighbors is assigned to the predicted sample as the predicted value of the KNN regression models for threshold prediction. Euclidean distance metric is employed to determine the k nearest neighbors of the predicted sample. The k nearest neighbors play a vital role in prediction accuracy; thus, the hyperparameter of the KNN model to be optimized is the number of neighbors k . As shown in Figure 6C, the basic idea for the SVM-based classifier is to find the optimal hyperplane that has the maximum distance from the closest sample points. These hyperparameters in SVMs must be carefully chosen to obtain good performance. One is the penalty coefficient of the target function, c , which determines the tradeoff between minimizing the training error and minimizing the model complexity, and the other is the coefficient of the kernel function, γ , which implicitly defines the nonlinear mapping from the input space to some high-dimensional feature space (in this study, we focus entirely on the Gaussian kernel).

For each of the KNN, BPNN, and SVM algorithms, we tuned these hyperparameters by minimizing the estimated generalization error such as the k -fold cross-validation error or the leave-one-out error. The procedure for tuning hyperparameters for all models was implemented based on the Scikit-learn library (Pedregosa et al., 2011), a well-developed machine-learning library. The optimal combination of hyperparameters and algorithm was finally determined as those for which the model produced the lowest error and did not overfit (we ensured that during cross-validation, the difference in mean absolute error (MAE) between the training and test sets for Predictor 1–3 was limited to < 0.1 dB, or the difference in accuracy between the training and test sets for the classifier was limited to $< 0.5\%$). Table 3 lists the optimal algorithm and hyperparameters for each model.

4) *Model Training.* After acquiring the optimal combination of algorithm and hyperparameters, the final models (i.e., the trained Predictor 1 in the routine test module, Predictor 2 and Predictor 3 in the fast test module, and the classifier in the hearing screening module) were trained over all training data with the determined model algorithm and hyperparameters. It is noteworthy that the training data for Predictor 2 were limited to those frequencies that met the SNR-based criterion.

5) *Model Validation.* To further validate the test performance of the proposed three modules, we directly computed it on an unknown data set containing 44 ears of 23 subjects with NH (age: 23.7 ± 2.88 years) and 85 ears of 57 subjects with SNHL (age: 49.4 ± 14.8 years). Table 4 lists the number of ears in the unknown data set for each test frequency from 0.5–8 kHz. All ears had normal middle function defined by ≤ 10 dB air-bone gaps and normal 226-Hz tympanometry.

Indices for Performance Evaluation on the System. Model performance was evaluated via cross-validation and by testing on the unknown dataset. Test performance computed on all test samples in all k runs of LOOCV or k -fold cross-validation (Given that all cases can be in turn a single-item test set without repeating in each run), referred to as “cross-validation performance” in this study, was predominantly discussed. The three modules were further validated by computing the test performance on all samples in the unknown data set. MAE, defined as the mean of the absolute differences between the predicted and measured hearing thresholds, was computed to quantify the prediction performance for both the routine test module and the fast test module. In addition, the percentage of cases that were predicted within ± 10 dB of the measured hearing thresholds (from now on referred to as 10-dB accuracy) was another indicator for assessing the performance in threshold prediction. Classification accuracy (i.e., the percentage of ears that were correctly classified) was used to evaluate the performance in

Table 3. The Combination of Algorithm and Hyperparameters for Each Predictor or Classifier at all Test Frequencies.

Module	Model	Frequency (kHz)	Model algorithm	Hyperparameter		
				k (KNN-based model)	The number of nodes in the hidden layers (BPNN-based model)	penalty coefficient c /coefficient of kernel function γ (SVM-based model)
Routine test module	Predictor 1	0.5	BPNN	-	200	-
		1	KNN	8	-	-
		2	KNN	8	-	-
		4	BPNN	-	200	-
		8	KNN	8	-	-
Fast test module	Predictor 2	0.5	KNN	9	-	-
		1	KNN	8	-	-
		2	BPNN	-	200	-
		4	KNN	10	-	-
		8	KNN	9	-	-
	Predictor 3	0.5	BPNN	-	200	-
		1	KNN	8	-	-
		2	KNN	8	-	-
		4	KNN	8	-	-
		8	BPNN	-	200	-
Hearing screening module	Classifier	0.5	SVM	-	-	1000/0.00003
		1	SVM	-	-	1000/0.00006
		2	SVM	-	-	1000/0.00003
		4	SVM	-	-	1000/0.00003
		8	SVM	-	-	1000/0.00003

Note: BPNN = back-propagation neural network; KNN = k -nearest neighbor; SVM = support vector machine.

hearing screening. Another two indicators of clinical interest, the false negative rate (i.e., the percentage of ears with hearing loss that went undetected) and the false positive rate (i.e., the percentage of ears with normal hearing that were incorrectly identified as hearing loss), were calculated as well.

The Test Procedure for Each Module

The Routine Test Module. The flow diagram and test interface for the routine test module are shown in Figures 7A-B respectively. After the operator enters the subject information (e.g., name, gender, age and test ear, see Figure 7B), probe frequency (F_p) and probe level (L_p), SFOAE I/O functions are measured when clicking the “Start” button, with L_p increased in 5-dB steps from 5 to a maximum of 70 dB

Table 4. The Number of Ears in Each Category Contained in the Unknown Data set for Each Test Frequency.

Category	Frequency (kHz)				
	0.5	1	2	4	8
NH	42	43	41	39	38
SNHL	33	33	41	43	49
Total	75	76	82	82	87

Note: NH = normal hearing; SNHL = sensorineural hearing loss.

SPL (or 60 dB SPL for 4 kHz). Then, upon clicking the “Predict” button, input variables are extracted for input to the trained Predictor 1 for threshold prediction.

The Fast Test Module. Figures 8A and 8B show the flow diagram and test interface of the fast test module, respectively. The minimum probe level L_{min} can be customized from 5 to 25 dB SPL ($5 \leq L_{min} \leq 25$) to shorten the test time. The system starts to record SFOAEs from L_{min} (e.g., $L_{min} = 15$ dB here) in 5-dB increments immediately after clicking the “Start” button in Figure 8B, while determining L_p threshold according to the above SNR-based inclusion criterion (see details in “Feature extraction” section). The test stops once the L_p threshold has been established (see Figure 5A, in the presence of L_p threshold) and then the system automatically extracts the input variables for Predictor 2 according to the left branch in Figure 8A. The test procedure is discontinued if no L_p threshold has been determined when the probe level reaches the maximum attainable value. In the absence of L_p threshold (see Figure 5B), input variables for Predictor 3 are extracted and then input to the trained Predictor 3 to obtain the estimate of hearing threshold.

Hearing Screening Module. Figure 9 shows the flow diagram (A) and test interface (B) of the hearing screening module. After clicking the “Start” button, SFOAEs are measured at three specific probe levels (40, 50, 60 dB SPL) (see

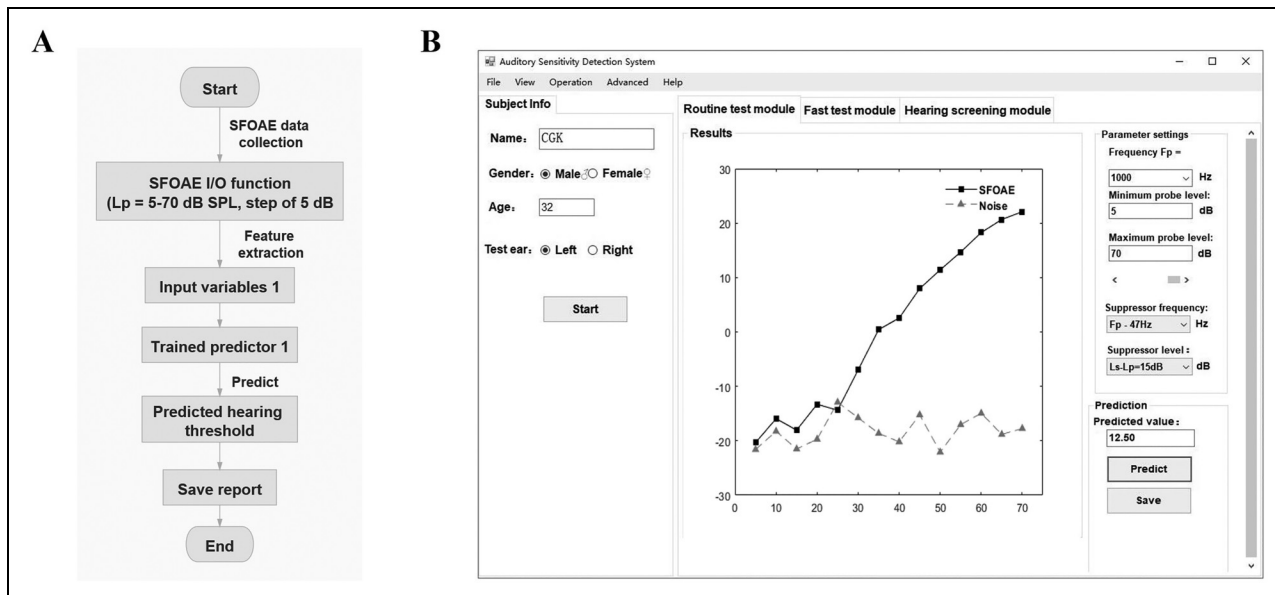


Figure 7. The routine test module. A, Flow diagram B, Test interface and an example of test results.

Figure 9B). After clicking the “Predict” button, the system automatically extracts SFOAE levels, SFOAE SNRs and T_{sf} magnitudes at all measured probe levels as the input variables, which are then input to the trained classifier to be classified as NH or hearing loss.

RESULTS

Cross-Validation Prediction Performance

The Routine Test Module and the Fast Test Module. The MAEs of the routine test module at 0.5–8 kHz computed over the

predictions of the test samples are shown in Table 5. The MAEs ranged from 7.06 (1 kHz) to 11.61 dB (8 kHz). Overall, the MAEs were lower in the normal-hearing group than in the subjects with hearing loss (i.e., HL >25 dB) at all test frequencies. Also shown in Table 5 is the percentage of ears that were predicted within ± 10 dB of the measured hearing thresholds (10-dB accuracy), between 62.05% (8 kHz) and 83.71% (1 kHz). A larger percentage of ears were estimated within ± 10 dB of the measured hearing thresholds in subjects with NH than in those with hearing loss. Across all ears, the routine test module resulted in better performance at 1–4 kHz than at lower and higher frequencies.

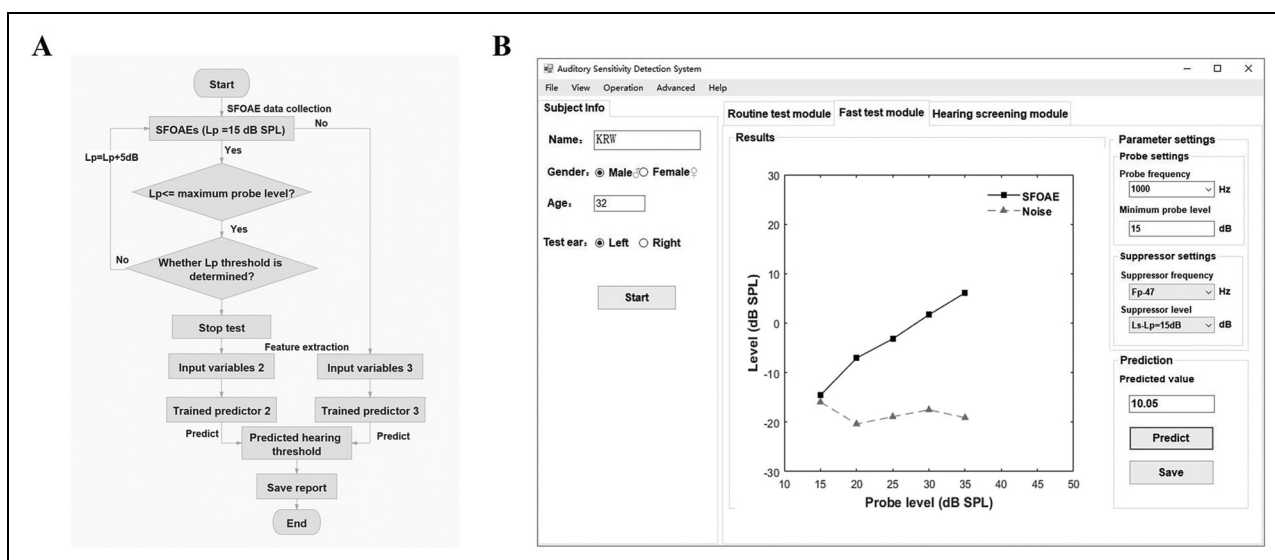


Figure 8. The fast test module. A, Flow diagram. B, Test interface and an example of test results.

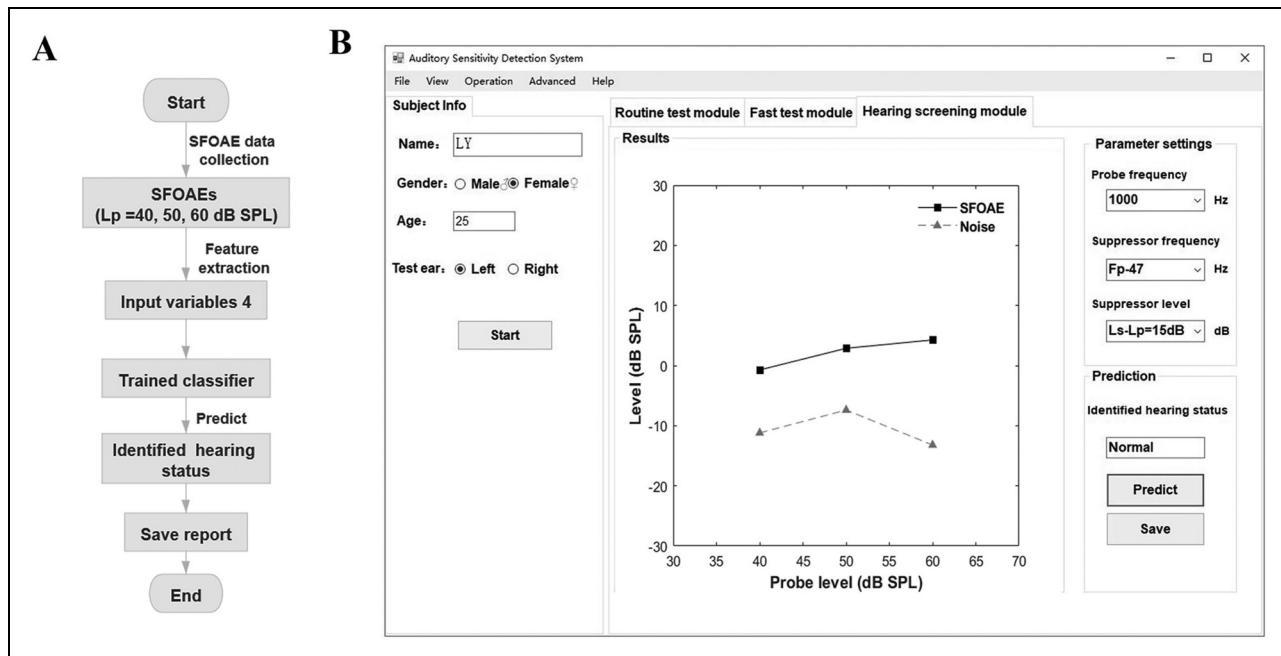


Figure 9. Hearing screening module. A, Flow diagram. B, Test interface and an example of test results.

Table 6 indicates the percentage of ears that met the SNR-based inclusion criterion for determining L_p threshold (“SFOAE evoked rate”) for the fast test module at 0.5–8 kHz, along with the 10-dB accuracy (between 57.26% and 81.22%) and MAEs, ranging from 7.40 (1 kHz) to 12.60 dB (8 kHz). At all test frequencies, over 93% of normal-hearing ears met the SNR-based inclusion criterion, meaning that hearing thresholds were predicted without measuring the entire SFOAE I/O function. However, only 34.06–53.28% of ears with SNHL met the SNR-based inclusion criterion; thus, more time was needed to measure SFOAEs in response to higher probe levels. Despite this, in the SNHL group, the 10-dB accuracy was larger than the SFOAE evoked rate, indicating that in many impaired ears hearing thresholds could be predicted even in the absence of L_p threshold. In agreement with the predictions of the routine test module, the prediction performance in regions of NH was superior to that in regions of hearing loss at all test frequencies (but less evident at 8 kHz), and the prediction performance was poorer at 0.5 and 8 kHz relative to 1–4 kHz.

Histograms of prediction error (i.e., the difference between the predicted and measured hearing thresholds) for the routine test module (black bars) and the fast test module (white bars) are shown in Figure 10A. Each panel from left to right shows data for different frequencies from 0.5 to 8 kHz. Both modules yielded predictions with similar error distributions. A majority of predictions based on the routine test module and the fast test module presented low errors (≤ 10 dB); however, in less than 4% of cases predictions had large errors (≥ 20 dB) at 1–4 kHz, and in 8–15% of cases predictions had errors ≥ 20 dB at 0.5 and 8

kHz. For further observation of error distribution, Figure 10B plots the cumulative percentage of ears in which predictions were within accuracy bands of ± 0 dB (exact), ± 5 dB, ± 10 dB, ± 15 dB, and ≥ 20 dB of the measured hearing thresholds at 0.5–8 kHz for the routine test module (left panel) and fast test module (right panel). Similar trends were observed for both test modules. Overall, there was agreement within 15 dB between the predicted and measured hearing thresholds in approximately 90% of cases at 1–4 kHz. Both modules predicted hearing thresholds to be within 10 dB in a much greater percentage of cases at 1–4 kHz than at 0.5 and 8 kHz.

Overall performance was improved notably when tests were restricted to hearing levels ≤ 60 dB HL, which is typically used in previous studies (Go et al., 2019; Gorga et al., 2003; Mertes & Goodman, 2013). As shown in Table 7, these restrictions resulted in MAE decreases of 0.66–1.34 dB for the routine test module and 0.6–1.25 dB for the fast test module. Indeed, the generation of SFOAEs relies on the normal function of OHCs. It is not surprising that there is an absence of OAEs in the cochlea in ears with $HL > 60$ dB due to a severe or complete loss of OHCs, and naturally no detectable SFOAEs is present in the ear canal (we ensured that all ears had normal middle ear function), which might account for the large prediction errors at high hearing levels.

The Hearing Screening Module. The prediction performance of the hearing screening module for frequencies 0.5–8 kHz is shown in Table 8. Five-fold cross-validation was conducted on performance evaluation so that the test accuracy

Table 5. Cross-Validation Performance of the Routine Test Module in Predicting Hearing Thresholds.

Frequency (kHz)	NH (≤ 25 dB HL)		SNHL (>25 dB HL)		Overall	
	10-dB Accuracy (%)	MAE (dB)	10-dB Accuracy (%)	MAE (dB)	10-dB Accuracy (%)	MAE (dB)
0.5	84.86	7.05	62.88	11.17	73.60	9.16
1	93.43	4.53	75.82	9.11	83.71	7.06
2	92.23	5.75	71.13	9.63	80.90	7.83
4	89.45	5.88	78.33	8.79	83.37	7.47
8	69.00	11.33	57.58	11.79	62.05	11.61

Note: NH = normal hearing; SNHL = sensorineural hearing loss; MAE = mean absolute error; 10-dB accuracy = the percentage of ears that were predicted within ± 10 dB of the measured hearing thresholds.

was computed as the average of five individual runs. For ease of comparison with other studies, the performance of the hearing screening module was also quantified using the area under the receiver operating characteristic curve (AUC). Over 90.82% of ears were identified correctly at 0.5–4 kHz, with the test accuracy ranging from 90.82% (0.5 kHz) to 95.93% (1 kHz). The test accuracy was lowest at 8 kHz, with 88.38% of cases correctly classified. The hearing screening module yielded large AUCs at all test frequencies, with over 0.96 at 0.5–4 kHz and 0.91 at 8 kHz. Also provided in Table 8 are the false negative rate and false positive rate. Overall, the hearing screening module resulted in a low false negative rate at 0.5–8 kHz (2.87–7.02%) but a slightly higher false positive rate at 0.5 and 8 kHz. About 0.95 min was needed in the hearing screening module for identifying hearing status.

Performance Evaluated on an Unknown Data set

Table 9 shows the test performance evaluated on this unknown data set for each module, with the same indicators as the cross-validation. Since the cross-validation performance showed that the hearing thresholds of ears with severe hearing loss almost could not be accurately predicted (Table 7), the hearing thresholds were restricted to ≤ 60 dB HL in the unknown data set. The MAEs for the routine test

module ranged from 6.2 (1 kHz) to 10.2 (8 kHz), with the percentage of ears that were predicted within ± 10 dB of the measured hearing thresholds (10-dB accuracy) ranging between 69.0% and 88.2%. Compared to the routine test module, the fast test module resulted in slightly larger MAEs (0.5–8 kHz: 6.1–11.1 dB) and lower 10-dB accuracy (0.5–8 kHz: 62.1–89.5%) except for 1 kHz. As shown in Table 9, the hearing screening module correctly identified the hearing status of over 92% of ears at 0.5–4 kHz, and that of 88.5% of ears at 8 kHz. Moreover, the hearing screening module resulted in low false negative and false positive rates.

Performance Comparison

Comparisons Between the Routine Test Module and the Fast Test Module. The cross-validation performance of the routine test module and the fast test module were compared in terms of MAE and mean test time. The MAEs and mean test time (minutes) for both hearing categories (normal: HL ≤ 25 dB; SNHL: HL > 25 dB) and across all ears are shown in Figure 11. The gray bars represent the results for the routine test module and the open bars represent the results for the fast test module. Each panel from top to bottom represents data for a separate frequency from 0.5 to 8 kHz. When using the routine test module, approximately 6.2 min was

Table 6. Summary of the Threshold Prediction Results for the Fast Test Module: SFOAE Evoke Rate (%) and Cross-Validation Performance in Predicting Hearing Thresholds.

Frequency (kHz)	NH (≤ 25 dB HL)			SNHL (>25 dB HL)			Overall		
	SFOAE evoked rate (%)	10-dB Accuracy (%)	MAE (dB)	SFOAE evoked rate (%)	10-dB Accuracy (%)	MAE (dB)	SFOAE evoked rate (%)	10-dB Accuracy (%)	MAE (dB)
0.5	95.87	83.03	7.29	34.06	59.39	11.77	64.21	70.92	9.59
1	99.49	93.94	4.73	53.28	70.9	9.57	73.98	81.22	7.40
2	98.54	90.78	5.61	49.37	72.8	10.01	72.13	81.12	7.97
4	98.17	88.53	6.11	45.63	74.52	9.26	69.44	80.87	7.83
8	93.01	60.26	12.29	35.96	55.34	12.79	58.29	57.26	12.60

Note: NH = normal hearing; SNHL = sensorineural hearing loss; MAE = mean absolute error; SFOAE evoked rate = the percentage of ears that met the SNR-based criterion; 10-dB accuracy = the percentage of ears that were predicted within ± 10 dB of the measured hearing thresholds.

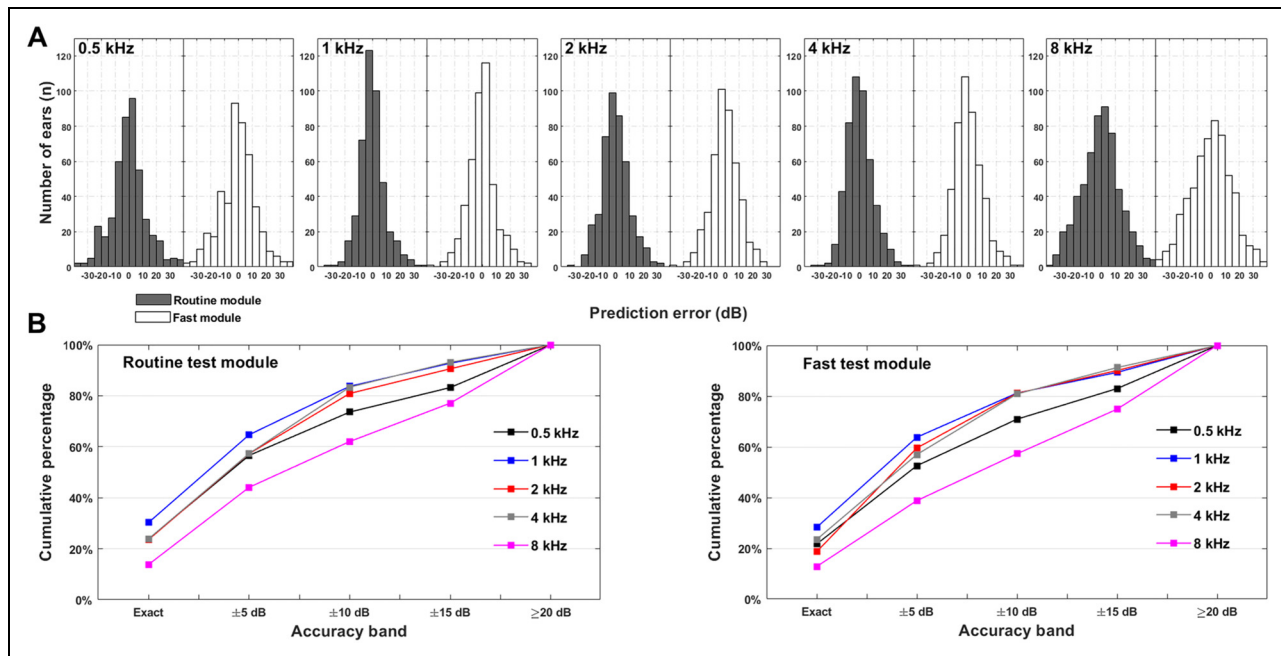


Figure 10. A, histogram of the prediction error (i.e., the difference between the predicted and measured hearing thresholds) for the routine test module (black bars) and fast test module (white bars). B, Cumulative percentage of ears that are predicted within an accuracy band (± 0 dB (exact), ± 5 dB, ± 10 dB, ± 15 dB, ≥ 20 dB) of the measured hearing thresholds at 0.5–8 kHz for the routine test module (left panel) and the fast test module (right panel).

needed for a single frequency (to measure an entire SFOAE I/O function), except for approximately 5.4 min at 4 kHz. The fast test module had shorter test time than the routine test module. Across all ears, the mean test time of the fast module for a single frequency ranged from 3.2 (4 kHz) to 4.0 (8 kHz) min. The mean test time for the fast module was shorter in the normal-hearing group than that in the SNHL group at all test frequencies, which was reduced by 1.25–2.45 min in the normal group relative to the SNHL group from 0.5 to 8 kHz. On average, both modules predicted hearing thresholds within 5 to 10 dB of the measured thresholds at 0.5–4 kHz when hearing levels were within the normal range except for the normal-hearing group at 1 kHz, where the MAE was less than 5 dB. At 8 kHz for the normal group and at 0.5 and 8 kHz for the SNHL group, the MAEs for both modules were within 10 to 15 dB of the measured hearing thresholds.

Comparison Between the Cross-Validation Performance and Performance Evaluated on the Unknown Data set. As shown in Figure 12, the cross-validation performance was compared to the test performance evaluated on the unknown data set for the routine test module (left panel), the fast test module (middle panel) and the hearing screening module (right panel). In the left and middle panels, the gray bars indicate the MAEs for the cross-validation performance, and the open bars represent the MAEs computed on the unknown data set. The open squares indicate the 10-dB accuracy for

the cross-validation performance and the open circles indicate the 10-dB accuracy computed on the unknown data set. Regardless of the routine test module and the fast test module, the MAEs computed on the unknown data set were similar to those evaluated on all test samples of cross-validation (when hearing levels were restricted to ≤ 60 dB). The 10-dB accuracy calculated on the unknown data set was also very close to that evaluated via cross-validation, except at 0.5 and 4 kHz for the fast test module. For the hearing screening module, the test accuracy for the unknown data set was very close to that evaluated via cross-validation.

A repeated-measures (rm) analysis of variance (ANOVA) on MAEs was conducted separately for each frequency, with performance type (i.e., cross-validation performance vs. test performance evaluated on the unknown data set) set to the between-subjects factor, and module type (i.e., the routine test module vs. the fast test module) set to the within-subjects factor. No significant interaction between the performance type and the module type was found at all test frequencies (0.5 kHz: $F_{1, 475} = 0.128$, $p = 0.721$; 1 kHz: $F_{1, 456} = 0.421$, $p = 0.517$; 2 kHz: $F_{1, 464} = 1.044$, $p = 0.307$; 4 kHz: $F_{1, 498} = 0.030$, $p = 0.863$; 8 kHz: $F_{1, 579} = 0.215$, $p = 0.643$). The simple main effect of performance type showed no statistically significant difference in MAEs between the cross-validation performance and the test performance evaluated on the unknown data set (0.5 kHz: $F_{1, 475} = 0.04$, $p = 0.842$; 1 kHz: $F_{1, 456} = 0.00$, $p =$

Table 7. Comparisons of the Mean Absolute Errors (MAEs, dB) Obtained From all Ears and Those From Ears Restricted to ≤ 60 dB HL. The Difference Represents the Decrease in MAEs (dB) From Ears Restricted to ≤ 60 dB HL Relative to Those From all Ears.

Frequency (kHz)	Routine test module			Fast test module		
	Overall	Restricted to ≤ 60 dB HL	Difference	Overall	Restricted to ≤ 60 dB HL	Difference
0.5	9.16	7.82	1.34	9.59	8.45	1.14
1	7.06	6.08	0.98	7.40	6.22	1.18
2	7.83	6.92	0.91	7.97	6.88	1.09
4	7.47	6.81	0.66	7.83	7.23	0.6
8	11.61	10.79	0.82	12.60	11.35	1.25

Note: MAE = mean absolute error.

0.998; 2 kHz: $F_{1, 464} = 0.120, p = 0.729$; 4 kHz: $F_{1, 498} = 0.044, p = 0.835$; 8 kHz: $F_{1, 579} = 0.185, p = 0.668$), indicating a good generalization ability of both the routine test module and the fast test module. No significant effect of module type on MAEs (0.5 kHz: $F_{1, 475} = 3.544, p = 0.060$; 1 kHz: $F_{1, 456} = 0.008, p = 0.930$; 2 kHz: $F_{1, 464} = 0.683, p = 0.409$; 4 kHz: $F_{1, 498} = 2.599, p = 0.108$; 8 kHz: $F_{1, 579} = 3.571, p = 0.059$) indicated that the routine test module and the fast test module resulted in similar performance regardless of performance type. For ears with NH and SNHL, independent samples t -tests were conducted separately for each frequency. Results showed the hearing levels of ears with NH were predicted within comparable absolute errors when using these two modules (0.5 kHz: $p = 0.7616$; 1 kHz: $p = 0.7588$; 2 kHz: $p = 0.8180$; 4 kHz: $p = 0.7107$; 8 kHz: $p = 0.3426$). In addition, there was no statistically significant difference in MAEs between two modules for ears with SNHL (0.5 kHz: $p = 0.4332$; 1 kHz: $p = 0.4806$; 2 kHz: $p = 0.5661$; $p = 0.3773$; $p = 0.1514$). To summarize, no difference in MAEs was apparent for most comparisons between the routine test module and the fast test module. To make test time comparison, a rm ANOVA was conducted on the test time. Results revealed a significant main effect of module type (0.5 kHz: $F_{1, 475} = 1176.12, p < 0.001$; 1 kHz: $F_{1, 456} = 1316.34, p < 0.001$; 2 kHz: $F_{1, 464} = 1261.00, p < 0.001$; 4 kHz: $F_{1, 498} = 1498.78, p < 0.001$; 8 kHz: $F_{1, 579} = 1230.39, p < 0.001$) with no significant performance type \times module type interaction (0.5 kHz: $F_{1, 475} = 0.081, p = 0.776$; 1 kHz: $F_{1, 456} = 0.836, p = 0.361$; 2 kHz: $F_{1, 464} = 0.126, p = 0.723$; 4 kHz: $F_{1, 498} = 0.777,$

$p = 0.378$; 8 kHz: $F_{1, 579} = 1.419, p = 0.234$), which suggested that the test time involved in the fast test module was significantly reduced relative to the routine test module.

Discussion

In this study, a routine test module and a fast test module based on SFOAEs were developed, offering a potential audiometric tool in a frequency-specific manner. Performance in threshold prediction was good in normal-hearing ears; however, worse in ears with hearing loss, which deserves continued efforts to be improved for clinical utility. Standard errors were also calculated to facilitate comparisons between the performance of the present SFOAEs and those of tests based on DPOAEs (Gorga et al., 2003; Johnson et al., 2007), as listed in Table 10. Both the routine test module and the fast test module in the present SFOAE study could make predictions for all ears tested. However, the DPOAE studies of Gorga et al., and Johnson et al., excluded a large percentage of ears originally tested due to a failure to meet their inclusion criteria. Also, the standard errors for both modules were typically equal to or lower than those observed in the DPOAE studies except for 0.5 kHz (For that frequency, the result for the DPOAE study of Gorga et al., may be unreliable given the paucity of data [of 158 ears tested, only 27 ears were predicted]). For these reasons, both the routine test module and the fast test module resulted in better performance in threshold prediction than the previous DPOAE studies (Gorga et al., 2003; Johnson et al., 2007). However, it should be clear that the superior performance in threshold prediction observed in both the routine test module and the fast test module compared to the previous DPOAE studies is a consequence of the machine learning algorithms and multiple combined variables rather than the stimulus type itself.

Regardless of hearing loss, both the routine test module and the fast test module predicted hearing thresholds with comparable performance (no statistically significant difference in MAEs between both modules); however, the test time involved in the fast module was significantly shorter than that in the routine module. Therefore, we suggest that

Table 8. The Cross-Validation Prediction Performance for the Hearing Screening Module at 0.5–8 kHz.

Frequency (kHz)	0.5	1	2	4	8
Test accuracy (%)	90.82	95.93	94.61	94.59	88.38
AUC	95.75	98.52	97.75	98.09	91.16
False Negative rate (%)	5.24	2.87	4.18	4.18	7.02
False Positive rate (%)	13.30	5.56	6.80	6.88	18.78

Note: AUC = area under the receiver operating characteristic curve.

Table 9. Test Performance for the Routine Test Module, Fast Test Module, and Hearing Screening Module That was Evaluated on an Unknown Data set.

Frequency (kHz)	Routine test module			Fast test module			Hearing screening module		
	10-dB Accuracy (%)	MAE (dB)	Mean Test Time (minutes)	10-dB Accuracy (%)	MAE (dB)	Mean Test Time (minutes)	Test accuracy (%)	False negative rate (%)	False positive rate (%)
0.5	82.7	7.50	6.2	82.7	8.43	3.8	92.0	5.3	2.7
1	88.2	6.24	6.2	89.5	6.06	3.1	94.4	2.8	2.8
2	82.9	6.45	6.2	82.9	6.89	3.4	95.1	3.7	1.2
4	80.5	6.90	5.4	76.8	7.42	3.2	93.9	3.7	2.4
8	69.0	10.19	6.2	62.1	11.11	4.1	88.5	1.2	10.3

Note: 10-dB accuracy = percentage of ears that were predicted within ± 10 dB of the measured hearing thresholds; MAE = mean absolute error.

the fast module is better suited than the routine module for clinical utility. In contrast to the single-predictor (Predictor 1) routine test module based on the entire SFOAE I/O function, the fast test module shortened the test time by training two other predictors (Predictor 2 and Predictor 3). Conceptually similar to the OAE thresholds previously reported (Boege & Janssen, 2002; Gorga et al., 2003; Johnson et al., 2007), the L_p threshold containing useful information regarding hearing threshold was determined for the fast routine module based on the SNR-based inclusion criterion. Once the L_p threshold was found, the fast test module stopped the SFOAE test and omitted the SFOAE measurements at higher probe levels. This stopping rule was necessary to avoid prolonged test times, as relatively sufficient information regarding hearing thresholds was captured. The reduced test time for the fast test module relative to the routine test module might depend on the degree of hearing loss. Compared to ears with NH, the mean test time for ears with hearing loss tended to be longer when using the fast test module, for which greater probe levels were needed to yield SFOAE response (i.e., meeting the SNR-based inclusion criterion).

The hearing screening module of the SFOAE-based system identified hearing status with great accuracy at all test frequencies in less than one minute for a single frequency, which was useful for objective SNHL screening. We compared the performance of the present SFOAE-based study to the best results of DPOAEs (Gorga et al., 2000), in which the AUCs were separately approximated as 0.96, 0.975, 0.98, 0.975, 0.98 from 0.5–8 kHz. The performance of our models in predicting hearing status was generally similar to that of DPOAE-based tests at 0.5–4 kHz; however, slightly poorer than DPOAEs at 8 kHz (Gorga et al., 2000). Compared to the SFOAE study of Ellison and Keefe (2005), in which the best AUC obtained at 0.5–8kHz was between 0.83 and 0.93, the performance in identifying hearing loss for the hearing screening module was improved remarkably. Given its good performance, we suggest that the hearing screening module is promising for clinical applications.

It is possible that the prediction errors partly result from the measurement errors from the coupler calibration, especially for higher frequencies. For stimulus calibration, coupler calibration may be preferable to in-the-ear SPL calibration (Neely & Gorga, 1998), but it is also problematic as the pressure measured in the coupler can differ greatly from that presented to the eardrum if the coupler impedance is much different from the ear-canal impedance. For microphone calibration, there is a concern that such coupler calibration would not mitigate standing waves in the ear canal for higher frequencies, where ear canal resonances and probe fitting would result in changes in relative stimulus intensity at the cochlea between subjects. In future studies, the system is expected to be calibrated with more advanced calibration techniques such as the forward pressure level/emitted pressure level approach (Maxim et al., 2019) and “in-situ” calibration (Chen et al., 2014). Another possible factor that affects the accuracy of hearing prediction is the use of high stimulus levels for the probe and suppressor given that they could elicit efferent reflex feedback and potentially even middle ear muscle reflex (Guinan et al., 2003). While this risk for SFOAEs is lower than that for DPOAEs, it should be noted that the recorded SFOAEs at higher stimulus levels involve a modulation by these two reflexes.

One preliminary study from our laboratory (Gong et al., 2020) have investigated the ability of SFOAEs to predict capabilities using BPNN algorithm, and another (Liu et al., 2020) further maximized the test performance by comparing multiple machine learning algorithms. The current study advanced beyond our previous work in three ways. First, the present study proposed a fast test module that reduced a large amount of test time without a significant decrease in threshold prediction performance. Similarly, the developed hearing screening module could identify hearing status with comparable performance to the studies of Gong et al. (2020) and Liu et al. (2020); however, its test duration was reduced by approximately 80%. Second, more data were collected in this study, particularly at 4 and 8 kHz. Moreover, the hearing thresholds

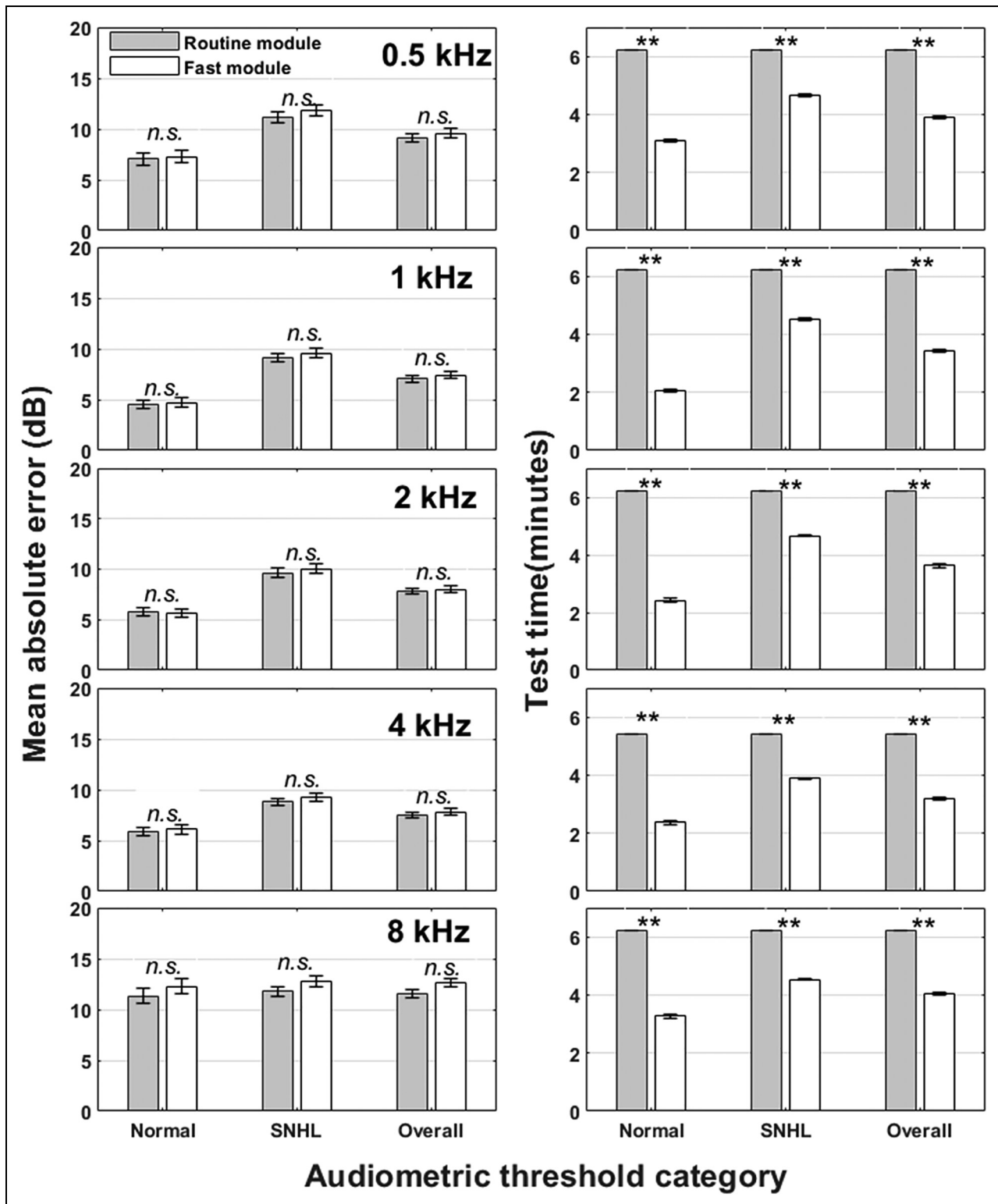


Figure 11. For two hearing categories (normal and sensorineural hearing loss [SNHL]) and across all ears (overall), the mean absolute error (MAE) (left panel) and mean test time (right panel) for the routine test module are compared to those for the fast test module. The gray bars indicate the results for the routine test module, and the open bars indicate the results for the fast test module. *n.s.* indicates no significance. ** $p < 0.01$. The error bars represent the standard errors of the mean.

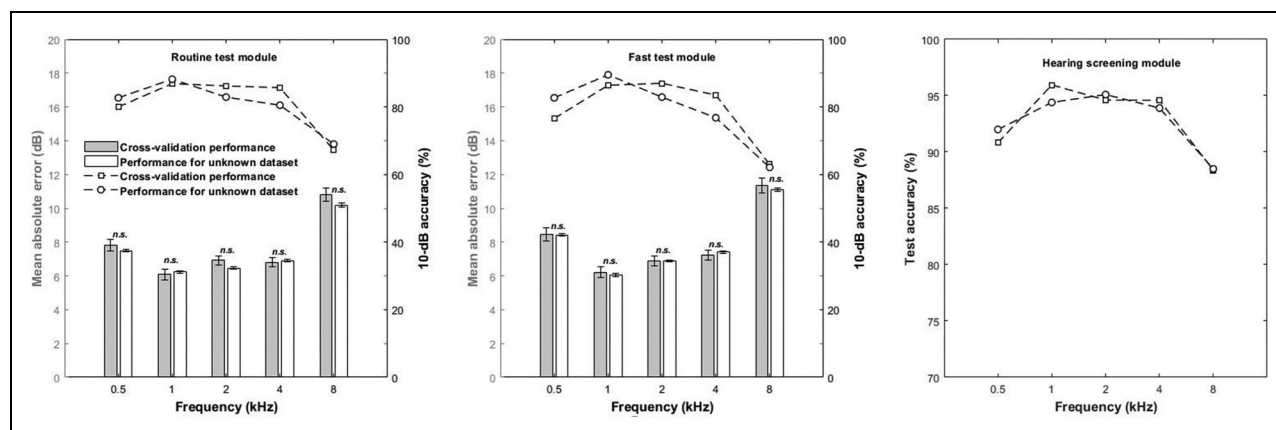


Figure 12. The cross-validation performance (when hearing levels were restricted to ≤ 60 dB HL) is compared to the test performance evaluated on the unknown data set for the routine test module (left), fast test module (middle) and hearing screening module (right). In the left and middle panels, the gray bars indicate the mean absolute errors (MAEs) for the cross-validation performance, and the open bars indicate the MAEs calculated on the unknown data set. *n.s.* indicates no significance. The error bars represent the standard errors of the mean. The open squares indicate the 10-dB accuracy for the cross-validation performance, and the open circles indicate that computed on the unknown data set.

of all ears involved in the SFOAE tests could be predicted via the present routine test module and the fast test module, while the BPNN predictors in Gong et al. (2020) were built based solely on those data meeting the inclusion criterion. Given that many ears failing the inclusion criterion were excluded from our previous study (Gong et al., 2020), it is not surprising that larger MAEs were observed in the current study than in the previous study. Finally, we generated machine learning models using optimized hyperparameters and algorithms that can be directly used for unknown data. Furthermore, the proposed modules in this study have been validated on a new unknown data set, while the test performance in the two preliminary studies from our laboratory was just assessed via cross-validation due to it only being used to investigate or maximize the ability of SFOAEs in predicting hearing capabilities.

One of the limitations is the clinical utility of the routine test module and the fast test module, which may be limited by a large amount of required measurement time, even for the hearing screening module. A significant reduction in the time effort is needed to improve the clinical utility of SFOAEs, for example, a fast approach to measure SFOAE I/O functions could be the use of chirp stimuli. Second, the large variability of the predicted thresholds for both the routine test module and the fast test module requires further reduction. In particular, hearing thresholds of ≥ 60 dB HL could not be predicted in the present assessment system. Indeed, physiological data have suggested a lack of OHC functioning when hearing thresholds exceeded a “rule of thumb” 50 dB (Stebbins et al., 1979); therefore, OAEs failed to account for $HL > 60$ dB (or even down to 50 dB). For example, several previous OAE-related studies did not

Table 10. Performance Comparison in Threshold Prediction Between Both Modules in the Present SFOAE Study and the DPOAE Studies of Gorga et al. (2003) and Johnson et al. (2007). *n/N* indicates the ratio of the number of ears predicted (*n*) and tested (*N*). Dashes indicated that predictions were not reported at that frequency.

Study	<i>n/N</i>					Standard error (dB)				
	Frequency (kHz)					Frequency (kHz)				
	0.5	1	2	4	8	0.5	1	2	4	8
The routine module in the present SFOAE study	447/447	442/442	445/445	481/481	585/585	12.5	9.9	10.4	9.8	15.1
The fast module in the present SFOAE study	447/447	442/442	445/445	481/481	585/585	13.0	10.6	10.6	10.3	16.3
DPOAEs in Gorga et al. (2003)	27/158	88/268	110/273	149/272	81/270	9.0	11.6	10.6	11.2	19.2
DPOAEs in Johnson et al. (2007)	-	-	117/205	164/205	-	-	-	9.9	10.3	-

Note: The standard error of the estimate in a regression is the standard deviation of the residuals of the regression, which is calculated as $\sqrt{\sum (Y - Y')^2 / (n - 2)}$ equivalent to the previous studies used (*Y* is the actual hearing threshold, *Y'* is the predicted hearing threshold, and *n* is the number of pairs of hearing thresholds).

SFOAE = stimulus-frequency otoacoustic emission; DPOAEs = distortion-product otoacoustic emissions.

attempt to predict hearing thresholds of > 60 dB HL (Boege & Janssen, 2002; Go et al., 2019; Gorga et al., 2003; Johnson et al., 2007; Mertes & Goodman, 2013).

In conclusion, the present SFOAE-based assessment system consisting of the routine test module, the fast test module and the hearing screening module provides a potential tool for objectively assessing hearing loss. The fast test module can predict hearing thresholds in an ear within ± 10 dB with an accuracy of 57.3–81.2% quantitatively in a relatively short time. The hearing screening module can identify hearing status with high accuracy and a low false negative rate.

Acknowledgments

We thank Professor Mario Ruggero (Northwestern University) for help with text editing and Fei Ji (The General Hospital of the People's Liberation Army) for help with data collection.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Natural Science Foundation of China (grant number 61871252) and the Foundation of Jiangsu Province Science and Technology (grant number BE2020635).

ORCID iD

Qin Gong  <https://orcid.org/0000-0002-5664-8477>

Supplemental Material

Supplemental material for this article is available online.

References

- Abdala C., Luo P., & Guardia Y. (2019). Swept-Tone stimulus-frequency otoacoustic emissions in human newborns. *Trends in Hearing*, 23, 1–15. <https://doi.org/10.1177/2331216519889226>.
- Bing D., Ying J., Miao J., Lan L., Wang D., Zhao L., & Wang Q. (2018). Predicting the hearing outcome in sudden sensorineural hearing loss via machine learning models. *Clinical Otolaryngology*, 43(3), 868–874. <https://doi.org/10.1111/coa.13068>.
- Boege P., & Janssen T. (2002). Pure-tone threshold estimation from extrapolated distortion product otoacoustic emission I/O-functions in normal and cochlear hearing loss ears. *The Journal of the Acoustical Society of America*, 111(4), 1810–1818. <https://doi.org/10.1121/1.1460923>.
- Brownell W. E. (1990). Outer hair cell electromotility and otoacoustic emissions. *Ear and Hearing*, 11(2), 82–92. <https://doi.org/10.1097/00003446-199004000-00003>.
- Charaziak K. K., Souza P., & Siegel J. H. (2013). Stimulus-Frequency otoacoustic emission suppression tuning in humans: Comparison to behavioral tuning. *Journal of the Association for Research in Otolaryngology*, 14(6), 843–862. <https://doi.org/10.1007/s10162-013-0412-1>.
- Chen S., Zhang H., Wang L., & Li G. (2014). An in-situ calibration method and the effects on stimulus frequency otoacoustic emissions. [journal article; research support, Non-U.S. Gov't]. *Biomedical Engineering Online*, 13, 95. <https://doi.org/10.1186/1475-925X-13-95>.
- Cheng H., Garrick D. J., & Fernando R. L. (2017). Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction. *Journal of Animal Science and Biotechnology*, 8(3), 733–737. <https://doi.org/10.1186/s40104-017-0164-6>.
- Dewey J. B., & Dhar S. (2017a). A common microstructure in behavioral hearing thresholds and stimulus-frequency otoacoustic emissions. *The Journal of the Acoustical Society of America*, 142(5), 3069–3083. <https://doi.org/10.1121/1.5009562>.
- Dewey J. B., & Dhar S. (2017b). Profiles of stimulus-frequency otoacoustic emissions from 0.5 to 20 kHz in humans. *Journal of the Association for Research in Otolaryngology*, 18(1), 89–110. <https://doi.org/10.1007/s10162-016-0588-2>.
- Ellison J. C., & Keefe D. H. (2005). Audiometric predictions using SFOAE and middle-ear measurements. *Ear and Hearing*, 26(5), 487–503. <https://doi.org/10.1097/01.aud.0000179692.81851.3b>.
- Go N. A., Stamper G. C., & Johnson T. A. (2019). Cochlear mechanisms and otoacoustic emission test performance. *Ear and Hearing*, 40(2), 401–417. <https://doi.org/10.1097/AUD.0000000000000625>.
- Gong Q., Liu Y., & Peng Z. (2020). Estimating hearing thresholds from stimulus-frequency otoacoustic emissions. *Trends in Hearing*, 24, 1–15. <https://doi.org/10.1177/2331216520960053>.
- Gong Q., Wang Y., & Xian M. (2014). An objective assessment method for frequency selectivity of the human auditory system. *Biomedical Engineering Online*, 13(1), 171. <https://doi.org/10.1186/1475-925X-13-171>.
- Gorga M. P., Johnson T. A., Kaminski J. K., Beauchaine K. L., Garner C. A., & Neely S. T. (2006). Using a combination of click- and toneburst-evoked auditory brainstem response measurements to estimate pure-tone thresholds. *Ear and Hearing*, 27(1), 60–74. <https://doi.org/10.1097/01.aud.0000194511.14740.9c>.
- Gorga M. P., Neely S. T., Bergman B., Beauchaine K. L., Kaminski J. R., Peters J., & Jesteadt W. (1993a). A comparison of transient-evoked and distortion product otoacoustic emissions in normal-hearing and hearing-impaired subjects. *The Journal of the Acoustical Society of America*, 94(5), 2639–2648. <https://doi.org/10.1121/1.407348>.
- Gorga M. P., Neely S. T., Bergman B., Beauchaine K. L., Kaminski J. R., Peters J., & Jesteadt W. (1993b). Otoacoustic emissions from normal-hearing and hearing-impaired subjects: Distortion product responses. *The Journal of the Acoustical Society of America*, 93(41), 2050–2060. <https://doi.org/10.1121/1.406691>.
- Gorga M. P., Neely S. T., Dorn P. A., & Hoover B. M. (2003). Further efforts to predict pure-tone thresholds from distortion product otoacoustic emission input/output functions. *The Journal of the Acoustical Society of America*, 113(6), 3275. <https://doi.org/10.1121/1.1570433>.
- Gorga M. P., Nelson K., Davis T., Dorn P. A., & Neely S. T. (2000). Distortion product otoacoustic emission test performance when both 2f1–f2 and 2f2–f1 are used to predict auditory status. *The Journal of the Acoustical Society of America*, 107(4), 2128–2135. <https://doi.org/10.1121/1.428494>.

- Guinan J. J., Backus B. C., Lilaonitkul W., & Aharonson V. (2003). Medial olivocochlear efferent reflex in humans: Otoacoustic emission (OAE) measurement issues and the advantages of stimulus frequency OAEs. *Journal of the Association for Research in Otolaryngology*, 4(4), 521–540. <https://doi.org/10.1007/s10162-002-3037-3>.
- Hurley R. M., & Musiek F. E. (1994). Effectiveness of transient-evoked otoacoustic emissions (TEOAEs) in predicting hearing level. *Journal of the American Academy of Audiology*, 5(3), 195–203.
- Johnson T. A., Neely S. T., Kopun J. G., Dierking D. M., Tan H., Converse C., & Gorga M. P. (2007). Distortion product otoacoustic emissions: Cochlear-source contributions and clinical test performance. *The Journal of the Acoustical Society of America*, 122(6), 3539–3553. <https://doi.org/10.1121/1.2799474>.
- Kalluri R., & Abdala C. (2015). Stimulus-frequency otoacoustic emissions in human newborns. *The Journal of the Acoustical Society of America*, 137(1), L78–L84. <https://doi.org/10.1121/1.4903915>.
- Kalluri R., & Shera C. A. (2013). Measuring stimulus-frequency otoacoustic emissions using swept tones. *The Journal of the Acoustical Society of America*, 134(1), 356–368. <https://doi.org/10.1121/1.4807505>.
- Keefe D. H., Ellison J. C., Fitzpatrick D. F., & Gorga M. P. (2008). Two-tone suppression of stimulus frequency otoacoustic emissions. *The Journal of the Acoustical Society of America*, 123(3), 1479–1494. <https://doi.org/10.1121/1.2828209>.
- Kemp D. T. (1978). Stimulated acoustic emissions from within the human auditory system. *The Journal of the Acoustical Society of America*, 64(5), 1386–1391. <https://doi.org/10.1121/1.382104>.
- Kemp D. T., & Chum R. (1980). Properties of the generator of stimulated acoustic emissions. *Hearing Research*, 2(3–4), 213–232. [https://doi.org/10.1016/0378-5955\(80\)90059-3](https://doi.org/10.1016/0378-5955(80)90059-3).
- Lineton B., & Lutman M. E. (2003). The effect of suppression on the periodicity of stimulus frequency otoacoustic emissions: Experimental data. *The Journal of the Acoustical Society of America*, 114(2), 871–882. <https://doi.org/10.1121/1.1582437>.
- Liu Y., Xu R., & Gong Q. (2020). Maximising the ability of stimulus-frequency otoacoustic emissions to predict hearing status and thresholds using machine-learning models. [journal article]. *International Journal of Audiology*, 60(4), 263–273. <https://doi.org/10.1080/14992027.2020.1821252>.
- Maxim T., Shera C. A., Charaziak K. K., & Abdala C. (2019). Effects of forward- and emitted-pressure calibrations on the variability of otoacoustic emission measurements across repeated probe fits. *Ear and Hearing*, 40(6), 1345–1358. <https://doi.org/10.1097/AUD.0000000000000714>.
- Mertes I. B., & Goodman S. S. (2013). Short-latency transient-evoked otoacoustic emissions as predictors of hearing status and thresholds. *The Journal of the Acoustical Society of America*, 134(3), 2127–2135. <https://doi.org/10.1121/1.4817831>.
- Neely S. T., & Gorga M. P. (1998). Comparison between intensity and pressure as measures of sound level in the ear canal. [comparative study; journal article; research support, U.S. Gov't, P.H.S.]. *The Journal of the Acoustical Society of America*, 104(5), 2925–2934. <https://doi.org/10.1121/1.423876>.
- Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., & Duchesnay E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Prieve B. A., Gorga M. P., Schmidt A., Neely S., Peters J., Schultes L., & Jesteadt W. (1993). Analysis of transient-evoked otoacoustic emissions in normal-hearing and hearing-impaired ears. *The Journal of the Acoustical Society of America*, 93(6), 3308–3319. <https://doi.org/10.1121/1.405715>.
- Robinette M. S., & Glatke T. J. (2007). *Otoacoustic emissions: Clinical applications* (3rd editioned.). Thieme Medical Publishers.
- Schairer K. S., Ellison J. C., Fitzpatrick D., & Keefe D. H. (2006). Use of stimulus-frequency otoacoustic emission latency and level to investigate cochlear mechanics in human ears. *The Journal of the Acoustical Society of America*, 120(2), 901–914. <https://doi.org/10.1121/1.2214147>.
- Shera C. A., & Guinan J. J. (1999). Evoked otoacoustic emissions arise by two fundamentally different mechanisms: A taxonomy for mammalian OAEs. *The Journal of the Acoustical Society of America*, 105(2), 782–798. <https://doi.org/10.1121/1.426948>.
- Shera C. A., & Guinan J. J. (2003). Stimulus-frequency-emission group delay: A test of coherent reflection filtering and a window on cochlear tuning. *The Journal of the Acoustical Society of America*, 113(5), 2762–2772. <https://doi.org/10.1121/1.1557211>.
- Shera C. A., Guinan J. J., & Oxenham A. J. (2002). Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. [journal article; research support, U.S. Gov't, P.H.S.]. *Proceedings of the National Academy of Sciences of the United States of America*, 99(5), 3318–3323. <https://doi.org/10.1073/pnas.032675099>.
- Shew M., New J., Wichova H., Koestler D. C., & Staecker H. (2019). Using machine learning to predict sensorineural hearing loss based on perilymph micro RNA expression profile. *Scientific Reports*, 9, 1. <https://doi.org/10.1038/s41598-019-40192-7>.
- Stebbins W. C., Hawkins J. J. E., Johnson L. G., & Moody D. B. (1979). Hearing thresholds with outer and inner hair cell loss. *American Journal of Otolaryngology*, 1(1), 15–27. [https://doi.org/10.1016/S0196-0709\(79\)80004-6](https://doi.org/10.1016/S0196-0709(79)80004-6).
- Stover L., Gorga M. P., Neely S. T., & Montoya D. (1996). Toward optimizing the clinical utility of distortion product otoacoustic emission measurements. *The Journal of the Acoustical Society of America*, 100(2), 956–967. <https://doi.org/10.1121/1.416207>.
- Zhao Y., Li J., Zhang M., Lu Y., Xie H., Tian Y., & Qiu W. (2019). Machine learning models for the hearing impairment prediction in workers exposed to Complex industrial noise: A pilot study. *Ear and Hearing*, 40(3), 690–699. <https://doi.org/10.1097/AUD.0000000000000649>.
- Zweig G., & Shera C. A. (1995). The origin of periodicity in the spectrum of evoked otoacoustic emissions. *The Journal of the Acoustical Society of America*, 98(4), 2018–2047. <https://doi.org/10.1121/1.413320>.