Development of a multi-category psychometric function to model categorical loudness measurements

Andrea C. Trevino,[a)] Walt Jesteadt, and Stephen T. Neely
Boys Town National Research Hospital, 555 North 30th Street, Omaha, Nebraska 68131, USA

(Received 9 September 2015; revised 21 July 2016; accepted 5 September 2016; published online 13 October 2016)

A multi-category psychometric function (MCPF) is introduced for modeling the stimulus-level dependence of perceptual categorical probability distributions. The MCPF is described in the context of individual-listener categorical loudness scaling (CLS) data. During a CLS task, listeners select the loudness category that best corresponds to their perception of the presented stimulus. In this study, CLS MCPF results are reported for 37 listeners (15 normal hearing, 22 with hearing loss). Individual-listener MCPFs were parameterized, and a principal component analysis (PCA) was used to identify sources of inter-subject variability and reduce the dimensionality of the data. A representative "catalog" of potential listener MCPFs was created from the PCA results. A method is introduced for using the MCPF catalog and maximum-likelihood estimation, together, to derive CLS functions for additional participants; this technique improved the accuracy of the CLS results and provided a MCPF model for each listener. Such a technique is particularly beneficial when a relatively low number of measurements are available (e.g., International Standards Organization adaptive-level CLS testing). In general, the MCPF is a flexible tool that can characterize any type of ordinal, level-dependent categorical data. For CLS, the MCPF quantifies the suprathreshold variability across listeners and provides a model for probability-based analyses and methods.
 [http://dx.doi.org/10.1121/1.4964106]

[EAS] Pages: 2571–2583

I. INTRODUCTION

Loudness is the perceived intensity of a sound (Fletcher and Munson, 1933). When quantified as a psychophysical variable, loudness has a nonlinear relation to the physical intensity of the sound that is altered when cochlear damage exists (Allen, 2008). There are a number of techniques for measuring loudness (e.g., Marks and Florentine, 2011); in this work we focus on a method called *categorical loudness scaling* (CLS). The CLS task uses category labels that are ecologically valid (e.g., "loud," "soft"), can be administered relatively quickly (<5 min/frequency), and requires little training on the part of the tester or listener. Although CLS has a well-studied relationship with hearing loss (e.g., Allen *et al.*, 1990; Al-Salim *et al.*, 2010; Brand and Hohmann, 2002; Heeren *et al.*, 2013; Oetting *et al.*, 2014), its inter-subject suprathreshold variability is not yet fully understood (Rasetshwane *et al.*, 2015). In general, audiometric-threshold-based models cannot capture this suprathreshold variability. Limited predictability at suprathreshold levels constrains the accuracy of a brief CLS test, reducing the effectiveness of such measurements. We introduce a more comprehensive model for CLS data, the multi-category psychometric function (MCPF), which parameterizes inter-subject variability and can be used to compute accurate CLS results.

The CLS task is conducted by presenting stimuli, typically bandpass noise or tones, to a listener at different levels. After each stimulus presentation, the listener selects the loudness category (e.g., "soft," "medium") that best describes their perception. The set of experimental stimulus levels can either be selected prior to testing, based on the listener's dynamic range (i.e., fixed-level stimuli), or the stimuli can be selected adaptively over the course of testing (i.e., adaptive-level stimuli, Brand and Hohmann, 2002). In a clinical setting, the time that can be devoted to a listening task is limited; in such cases, adaptive stimulus level selection may be employed to evenly sample across the range of a person's loudness perception with a lower number of stimulus presentations. Although presenting a lower number of stimuli provides a practical solution, the results of a brief test are more susceptible to error due to the statistical nature of listener responses. After a CLS task is completed, the listener's CLS function is estimated. The estimation procedure can include outlier removal, regression based on commonly observed functional forms, and the use of median values for each loudness-response category. These methods of computing the results model the random fluctuations in the response data as measurement noise. The proposed technique models the statistics of listener responses, quantifying the statistical distributions of loudness categories and inter-subject variability of suprathreshold perception. By using a probabilistic model, fluctuations in listener responses are interpreted as expected results from a statistical process. This probabilistic model can be utilized for listener simulations or when incorporating probability-based techniques (e.g., estimation, information, or detection theory) into analysis and measurements.

At a single stimulus level, a listener's CLS responses form a probability distribution across loudness categories. A conventional psychometric function represents the boundary between two response types (e.g., "yes"/"no," "heard"/"not

[a)]Electronic mail: andreatrv@gmail.com

heard"), as a function of stimulus level. The multi-category psychometric function (MCPF) is formed by plotting the loudness category probability distributions as a function of stimulus level. In this study, MCPFs were constructed based on data from listeners with normal hearing (NH) and hearing loss (HL). We used a parametric model to fit the MCPFs and a principal component analysis (PCA) to generalize the data. The results of the PCA are used to construct a representative set of MCPFs for a wide range of potential listeners, which we call an MCPF catalog. The MCPF catalog may have a variety of uses. As an illustrative example, we use it to compute a maximum-likelihood (ML) estimate of a new listener's MCPF and CLS function.

The proposed ML estimation approach was inspired by the work of Green (1993), who developed a technique for using estimation theory to determine the psychometric function of a "yes-no" task. His approach, when applied to measurements of audiometric threshold, increased the accuracy of the measurements (Florentine *et al.*, 2000; Wright *et al.*, 1997; Amitay *et al.*, 2006). Similarly, the aim of our MCPF-based ML approach is to improve the accuracy of the estimated CLS function. The International Standards Organization (ISO) recommendations (Kinkel, 2007) for adaptive CLS testing are designed to constrain the test time to a duration that is more clinically acceptable. Once the test is completed, median-based or curve-fitting techniques are used to compute the CLS function from the data. Because of the relatively low number of experimental trials, the natural variability of responses can result in a less accurate CLS function. The maximum-likelihood technique uses the probabilistic models that we have developed to find the MCPF, and from this the CLS function, that best fits the statistics of the listener's responses.

In this paper, we (1) introduce the concept of the MCPF representation using data from listeners with NH and HL, (2) describe a parameterization method for the representation, (3) use PCA to create a representative catalog, and (4) demonstrate how the catalog can be paired with a ML procedure to estimate an individual's MCPF.

II. METHODS

A. Participants

Sixteen NH listeners and 25 listeners with sensorineural HL participated in development of the MCPF catalog. Five additional NH listeners were recruited to test the ML estimation approach. The experimental protocol was approved by the institutional review board of the Boys Town National Research Hospital. Consent was obtained prior to testing and all listeners were compensated for their participation.

Audiometric thresholds of each participant were measured prior to CLS testing. Only one ear was tested per participant; if both ears met the criteria for the study, then the better ear was chosen as the test ear. If thresholds were matched across both ears, a test ear was selected randomly. NH participants had audiometric thresholds ≤ 10 dB HL; participants with HL had audiometric thresholds between 15 and 70 dB HL (thresholds measured at 0.5, 1, 2, 4, and 8 kHz). The inclusion of both listeners with NH and HL is motivated by the need for normative data, along with data that can model typical patients in an audiology clinic. Exclusion criteria were (1) a high false-alarm rate below threshold or (2) only using labeled buttons to respond during the task. On the basis of these criteria, one NH and three HL listeners were excluded from the analysis.

Five additional NH listeners, whose data were not used in the construction of the catalog, were recruited to test the effects of using the ML estimation approach. These listeners completed an additional CLS task adhering to an adaptive-level procedure, described in Sec. II C 2, before completing the fixed-level stimuli task to determine the targets for the estimation.

B. Stimuli

Pure-tone stimuli (1000-ms duration, 25-ms onset/set-off[1] cosine ramps) were sampled at 44 100 Hz. Tones were selected as stimuli due to their established role in loudness experiments; however, the methods described in this paper are not constrained by the type of stimuli. The same methodology could be applied to, for example, bandpass noise or warble-tone stimuli. CLS was measured at 1 and 4 kHz; 1 kHz was selected because it is the standard frequency used in the definition of loudness level, while 4 kHz was selected to capture the effects of high-frequency hearing loss. Stimuli were generated and presented using MATLAB.

C. Procedure

1. Fixed-level procedure

The CLS experimental methods followed the recommendations of the ISO standard (Kinkel, 2007). Stimuli were presented monaurally over headphones (Professional HD 25-II, Sennheiser, Wedemark, Germany) in a double-walled sound booth, with the participant seated at a desk in front of a monitor. Eleven loudness categories were graphically displayed on a computer monitor as horizontal bars that increased in length from bottom to top (see Fig. 1). Every other bar between the "can't hear" and "too loud" categories had a descriptive label (i.e., "very soft," "soft," "medium," etc.). Each bar was a clickable button in the interface. Participants were instructed to select, using a mouse, the bar that corresponded to their loudness perception. Additionally, participants were instructed to use the bars with and without descriptive labels equally, and that the sounds presented may or may not span the full range of "can't hear" to "too loud." Participants were encouraged to take breaks to reduce fatigue.

The CLS task was divided into two phases. During the first phase, the listener's dynamic range was determined using an adaptive technique (see Appendix B for implementation details). This initial phase also served as a familiarization phase; therefore, individual response data from this phase, beyond the dynamic range determination, were not used in the analysis. After the completion of phase 1, the data were reviewed for rough monotonicity by the tester before proceeding to the next phase.

The main experimental stimuli were presented in the second phase. For each frequency, a fixed set of stimuli was

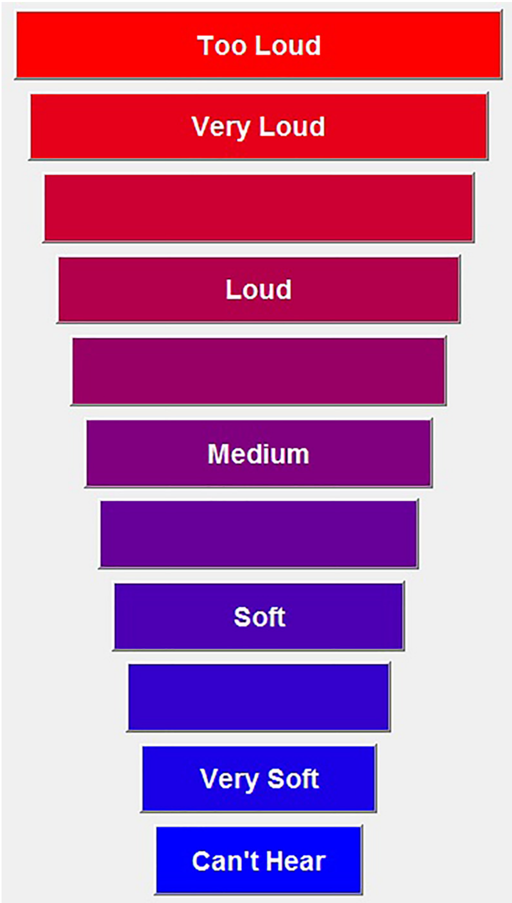

FIG. 1. (Color online) Scale used for loudness judgments. The assigned categorical units (CUs) used in the analysis range from 0 – "can't hear" to 50 – "too loud," in 5 CU increments.

presented to the listener, which was composed of 20 repetitions at each level, with the presentation levels spanning the listener's full dynamic range in 5 dB steps. Stimulus order was pseudorandomized such that there were no consecutive identical stimuli and level differences between consecutive stimuli did not exceed 45 dB (see Appendix A for additional details).

2. Adaptive-level procedure

A CLS test using an adaptive stimulus-level selection procedure was used to test an additional five NH participants, solely for verification of the ML-estimation-based methods. The data from these listeners were excluded from other analyses and from the construction of the representative catalog of MCPFs.

The adaptive stimulus-level procedure conformed to the ISO standard and the stimuli were the same as described in Sec. II B. For each listener, this task was completed after the determination of their dynamic range and before the fixed-level CLS task. The adaptive technique calculated ten levels that evenly spanned a listener's dynamic range and presented pseudorandomized stimuli at these levels three times; the ten levels were re-calculated with each repetition. For comparison, the fixed-level procedure was composed of 440 experimental trials for a listener with a 0–105 dB sound pressure level (SPL) dynamic range, while the ISO adaptive-level procedure required only 30 trials.

D. Model and analysis

1. Multi-category psychometric function

A conventional psychometric function describes the probability of a particular response, in a two-alternative paradigm, as a function of an experimental variable. A multi-category psychometric function (MCPF) generalizes the concept of a psychometric function to more than two possible responses and represents the probability distribution across multiple response categories as a function of an experimental variable (e.g., Torgerson, 1958). We introduce the MCPF within the context of CLS data, to interpret how loudness-category probabilities change with stimulus level. In a MCPF, a family of curves demarcates the probabilities of multiple categories as a function of stimulus level. The vertical distance between curves represents the probability of each categorical response; thus, at each fixed level, the ascending curves plot the cumulative probabilities of the categorical responses, from lowest loudness category to highest. As with any cumulative probability distribution, the total sum of the vertical distances between MCPF curves is always equal to one.

As an example, we demonstrate the construction of a hypothetical CLS MCPF for a NH listener. In this example, the dynamic range spans from 0 to 105 dB SPL. For reference, the 11-category CLS scale used for listener responses is shown in Fig. 1. Figure 2(a) shows the listener's probability distribution across categories at 60 dB SPL. Note that the "soft" category (15 CU) was selected most often by this listener and that the loudness judgments are not constrained to one category, but, instead, form a unimodal probability distribution. Figure 2(b) shows the cumulative probability density function for the data in 2(a). The MCPF [Fig. 2(c)] is constructed by plotting the location of the category boundaries in the cumulative distribution [marked by circles in 2(b)] as a function of stimulus level. In Fig. 2(c), the probabilities of the top three categorical responses at 60 dB SPL are marked with double-headed arrows; the vertical distances between the category-boundary curves at 60 dB SPL match the probabilities shown in Figs. 2(a) and 2(b).

2. Parameterization

The MCPF consists of a family of curves that demarcate the probabilities of each loudness category (see Fig. 2); these curves approximate logistic functions that can be described by a relatively small set of parameters. Fitting parametric models to listener data, organized as MCPFs, allows one to identify trends across listeners and to generalize the individual data from a smaller listener population to a wider variety of hearing configurations.

The four-parameter logistic function, $f(x|\boldsymbol{\theta})$, was selected to represent each of the empirical curves of the MCPF. The four parameters of this function correspond to a curve's minimum asymptote, slope, inflection point, and maximum asymptote, as illustrated in Fig. 3. The four parameters constitute the vector $\boldsymbol{\theta}$.

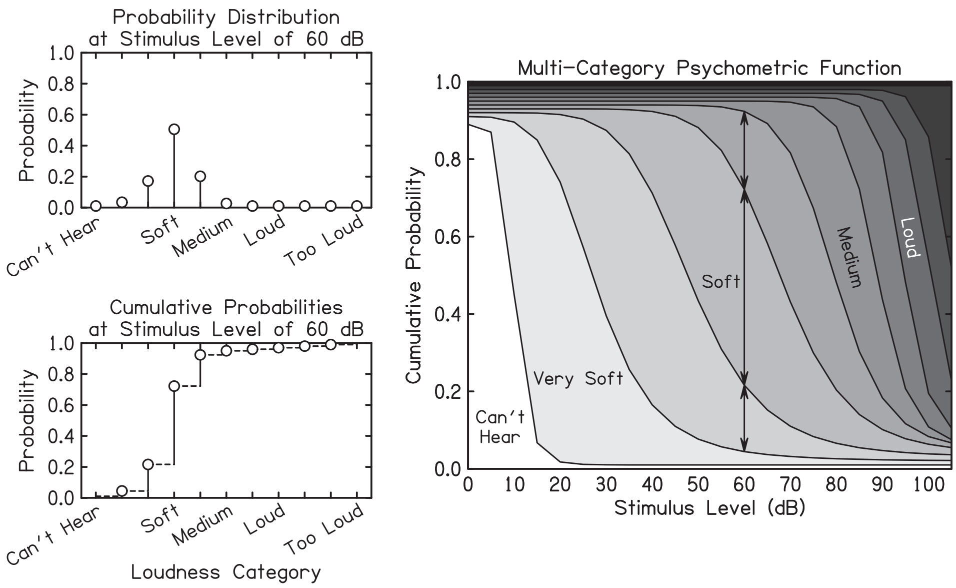

FIG. 2. Example MCPF, constructed from categorical probabilities. (top-left) Probability distribution across loudness categories at 60 dB SPL. (bottom-left) Cumulative probability function for loudness categories at 60 dB SPL; same data as (top-left). (right) MCPF from 0 to 105 dB SPL. Lines represent the category-boundary curves. The vertical distance between curves represents the probability of each category. The 3 highest category probabilities at the 60 dB SPL stimulus level are marked by arrows. Maximum categories are not labeled.

$$f(x|\boldsymbol{\theta}) = D + \frac{A - D}{1 + (x/C)^B},$$
$$x = \text{stimulus level}, \quad \boldsymbol{\theta} = [A \quad B \quad C \quad D]^T,$$
$$A = \text{minimum asymptote}, \quad B = \text{slope},$$
$$C = \text{inflection point}, \quad D = \text{maximum asymptote}. \tag{1}$$

For a MCPF of CLS data, each curve marks the transition between loudness categories along the cumulative probability distribution [Fig. 2(b)]. A single curve's minimum asymptote represents the cumulative likelihood of selecting any category below the curve, independent of stimulus level. The maximum asymptote represents 1 minus the cumulative likelihood of selecting a category above the curve, independent of stimulus level. The distance between two inflection points can be thought of, roughly, as the width of a particular loudness category, in dB. The magnitude of the slope is a measure of how well a listener distinguishes categories; a shallower slope indicates a wider, less-defined distribution of probability across categories, while a sharper slope indicates that the levels adjacent to the curve are perceptually unambiguous.

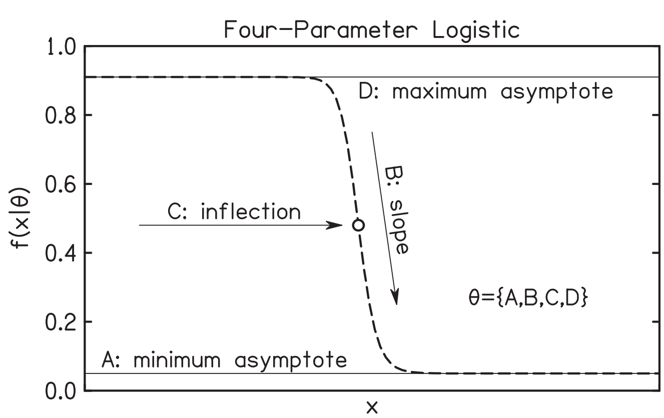

FIG. 3. The 4-parameter logistic function with the roles of each parameter labeled on the plot.

For the 11-category CLS scale, each MCPF, $F(x|\boldsymbol{\Theta})$, consists of 10 category-boundary curves parameterized by a 4×10 matrix $\boldsymbol{\Theta}$. The collection of ten curves, with the ith curve defined by $f(x|\boldsymbol{\theta}_i)$, is denoted as

$$F(x|\boldsymbol{\Theta}) = \{f(x|\boldsymbol{\theta}_i)\}_{i=1}^{10},$$
$$\boldsymbol{\theta}_i = [A_i \quad B_i \quad C_i \quad D_i]^T, \tag{2}$$

where $\boldsymbol{\theta}_i$ is the ith column of $\boldsymbol{\Theta}$.

For N study participants, $\boldsymbol{\Theta}^{(n)}$ is the parameter matrix for the nth participant. Thus, the collection of MCPFs for all study participants is denoted as

$$\{F(x|\boldsymbol{\Theta}^{(n)})\}_{n=1}^{N}.$$

In our implementation, all data were fit using a publically available logistic-regression fitting function L4P (obtained from MATLAB Central File Exchange). A CLS function can be constructed from the MCPF inflection points, which represent the presentation levels that correspond to transitions between loudness categories.

3. Creating a representative catalog

Inspired by the two-category example of Green (1993), we have developed a set of MCPFs designed to represent a typical range of possible listener CLS outcomes. We refer to this set of MCPFs as a "catalog." Each MCPF in the catalog has the same parametric form [Eq. (2)] and is defined by a parameter matrix $\boldsymbol{\Theta}$. The catalog of MCPFs was generalized from the MCPFs of the study participants to capture the naturally occurring range of parameter variability across a large population of listeners.

The catalog was constructed as follows. First, the MCPF from each listener's empirical data was parameterized, as described in Sec. II D 2. Additional transformations of the parameters were implemented to aid in the analysis and generalization, as described by

$$
\begin{aligned}
\widehat{\boldsymbol{\theta}}_i &= \left[\widehat{A_i} \quad \widehat{B_i} \quad \widehat{C_i} \quad \widehat{D_i}\right]^T, \quad i = 1, \ldots, 10, \\
\widehat{A_i} &= (0.01)i, \qquad i = 1, \ldots, 10, \\
\widehat{B_i} &= \frac{1}{B_i}, \qquad i = 1, \ldots, 10, \\
\widehat{C_1} &= C_1, \\
\widehat{C_i} &= C_i - C_{i-1}, \quad i = 2, \ldots, 10, \\
\widehat{D_i} &= 1 - (0.01)(11 - i), \quad i = 1, \ldots, 10,
\end{aligned} \tag{3}
$$

where $\widehat{\boldsymbol{\theta}}_i$ is the ith column of the 4×10 matrix $\widehat{\boldsymbol{\Theta}}$.

The slope and inflection point parameters were transformed to match units and remove level dependency; the reciprocal of each slope was taken to match the units of the inflection points, and the difference between the inflection points of each curve replaced all but the inflection point of the lowest-level curve, $f(x|\boldsymbol{\theta}_1)$. The difference between inflection points captures the width of a listener's loudness category, in dB units, irrespective of the particular levels at which each category lies. The inflection point of the lowest-level curve is related to a listener's degree of hearing loss. The asymptotes were fixed at intervals of 0.01 from 0 (for the minimum) and 1 (for the maximum). Fixing the asymptotes at these values reflects our empirical observation of the rate of random responses, which were presumably due to inattention induced by test fatigue, and reduces the complexity of the model.

A PCA was performed on the transformed slope and intercept parameters $(\widehat{B}, \widehat{C})$ to reduce the dimensionality of the data. The variance explained and the range of listener weightings for each PCA vector were used to reconstruct parameters $(\tilde{B}, \tilde{C})$ that represent a large range of listener responses. Finally, the MCPF catalog is composed of the set of MCPFs which are defined by these parameter sets.

The data for both NH and HL listeners were combined for the PCA in order to identify generalizable sources of variability. In addition, the data for 1 and 4 kHz stimuli were combined for the PCA, as no significant differences between frequencies were observed in the MCPF parameters after accounting for audiometric threshold. Prior to the PCA, the parameter variances across listeners were normalized based on a sensitivity analysis to account for differences in parameter units. The sensitivity analysis measured the resulting mean squared error from perturbing the parameters of a MCPF, and the normalization adjusted the parameter variances proportionally to the mean squared errors. The PCA of all transformed NH and HL listener parameters resulted in a set of eigenvectors that decompose the sources of variability in the data. These eigenvectors can be weighted and combined to reconstruct transformed parameters; we denote transformed parameters that have been constructed from PCA vectors as $\tilde{B}$, $\tilde{C}$. The first PCA vector captures the largest amount of variability across the parameters, with subsequent PCA vectors representing decreasing amounts of variance in the data.

In order to determine the listener weightings for the PCA vectors, each listener-specific set of parameters was projected onto the PCA vectors as follows. Each listener's transformed slope and intercept parameters were reshaped as a vector for the analysis. Let the matrix of all transformed listener parameters be $\mathbf{X}$, then the set of listener weightings, $\mathbf{W}$, from the projections on the PCA vectors $\mathbf{V}$ can be computed by

$$
\mathbf{XV} = \mathbf{W}. \tag{4}
$$

The standard deviations of the listener weightings, for each PCA vector, were computed, and twice the standard deviation (2σ) was used as the range of vector weights. For normally distributed data, the range defined by $\mu \pm 2\sigma$ theoretically contains 95% of the values. The PCA-vector weightings used to reconstruct the parameters for the catalog were determined by evenly sampling from the range of individual listener projections on the PCA vectors. The number of weightings sampled from each PCA vector was approximately proportional to the percentage of energy that the vector captured. The first two PCA vectors were sufficient for capturing $>90\%$ of the energy. A minimum constraint of 1 sampled weight per PCA vector was added in order to maintain valid MCPF parameter values.

The total number of sampled weightings was determined by iteratively increasing the number of sampled weightings, which were proportionally distributed across the PCA vectors, and monitoring the mean squared error between the resulting catalog of reconstructed MCPFs and the original listener data. The mean squared error saturated at a minimum of 0.03 dB with 66 total sampled weightings. As an additional step, the number of sampled weightings allotted to the third and fourth PCA vectors was raised from 1 to 3, in order to allow for a slight increase in the variability represented by the CLS catalog. Permutations of the sampled weightings for the first four PCA vectors were used to approximate different listener characteristics; the mean value was used as the single sample weighting for the remaining 16 PCA vectors. Weighted combinations of all 20 PCA vectors were used to reconstruct the MCPFs for the catalog. Only combinations that formed valid MCPFs (i.e., no curve crossovers or non-negative slopes) were retained, resulting in the 1460 MCPFs that constitute the catalog.

The permutations of sampled weightings, $\tilde{\mathbf{W}}$, represent a wide range of listener types. The new $\tilde{B}$, $\tilde{C}$ parameters for a MCPF were constructed as the weighted sum of the PCA vectors $\tilde{\mathbf{W}}\mathbf{V}^{\mathbf{T}}$, with different parameter sets for each set of eigenvector weightings. Each matrix of derived parameters, $\boldsymbol{\Theta}$, was computed from the resulting parameters via the transformations described in Eq. (3). The set of all derived MCPF parameters defines the catalog $\{F(x|\boldsymbol{\Theta}^{(m)})\}_{m=1}^{M}$, $M = 1460$.

For each listener, there is a single MCPF in the catalog that best fits their categorical loudness perception. The method for determining which MCPF in the catalog best represents a new listener's data is described in Sec. II D 4.

4. Maximum-likelihood estimation

The catalog described in Sec. II D 3 can be used to calculate a ML estimate of a new listener's MCPF. This may be

TABLE I. NH and HL listener characteristics. N is the number of participants. Age, audiometric threshold, and loudness discomfort level are reported. Standard deviations are shown in parentheses. NA: tallies listeners that did not report a LDL from 0 to 105 dB SPL.

	N	Age (years)	Threshold, 1 kHz	Threshold, 4 kHz	Loudness discomfort level, 1 kHz (dB SPL)	Loudness discomfort level, 4 kHz (dB SPL)
NH	15	38 ($\pm$10)	2 dB HL ($\pm$5)	5 dB HL ($\pm$5)	102 ($\pm$4) [6 NA]	102 ($\pm$4) [9 NA]
HL	22	53 ($\pm$16)	31 dB HL ($\pm$18)	51 dB HL ($\pm$12)	102 ($\pm$6) [3 NA]	100 ($\pm$4) [11 NA]

particularly useful when an accurate estimate of a listener's loudness perception is needed, but the number of experimental observations is relatively low, as is the case in most clinically aimed adaptive-level CLS methods.

Given a listener's CLS data, the ML estimate of the MCPF parameters is computed as follows. We denote a listener's CLS data as $[(x_1, c_1), \ldots, (x_T, c_T)]$, where T is the total number of experimental observations, x is the stimulus level, and c is an index of the listener's categorical response. For the 11-category CLS task, the minimum value for c is 1 and the maximum is 11. The likelihood, $\mathcal{L}(\cdot)$, of these observations is computed for each catalog function, with each catalog function being defined by its parameter matrix $\boldsymbol{\Theta}$

$$\mathcal{L}(\boldsymbol{\Theta}; (x_1, c_1), \ \ldots, (x_T, c_T)) = \prod_{t=1}^{T} (f(x_t|\boldsymbol{\theta}_{c_t}) - f(x_t|\boldsymbol{\theta}_{c_t-1})),$$

with the maximization computed over all potential parameter sets in the catalog: $\boldsymbol{\Theta} \in \{\boldsymbol{\Theta}^{(m)}\}_{m=1}^{M}, M = 1460$. For the computation, $f(x_t|\boldsymbol{\theta}_0) = 0$ and $f(x_t|\boldsymbol{\theta}_{11}) = 1$. To reduce the complexity of the computation, the maximization was computed over the monotonically related, and thus functionally equivalent, log-likelihood.

$$\begin{aligned} &\mathrm{argmax}_{\boldsymbol{\Theta}} \ln \mathcal{L}(\boldsymbol{\Theta}; (x_1, c_1), \quad \ldots, (x_T, c_T)) \\ &\quad = \mathrm{argmax}_{\boldsymbol{\Theta}} \sum_{t=1}^{T} \ln(f(x_t|\boldsymbol{\theta}_{c_t}) - f(x_t|\boldsymbol{\theta}_{c_t-1})). \end{aligned}$$

Once the ML parameter matrix $\boldsymbol{\Theta}$ has been determined, it can be used to approximate the listener's MCPF via Eqs. (1) and (2).

III. RESULTS

A. General measures

CLS data for construction of the MCPF catalog were collected from 15 NH listeners and 22 listeners with HL. Table I describes the audiometric characteristics of these listeners. CLS results were initially summarized as CLS functions, which relate loudness categories to level. Each individual listener's CLS function was determined by calculating the median level for each loudness category. Figure 4 displays the mean CLS functions for the group of NH listeners, for the 1 and 4 kHz stimuli; the upper and lower curves represent $\pm$2 standard deviations, respectively. Mean and standard deviations were computed across level for each categorical loudness dimension. These data suggest that, for NH listeners, the frequency of the stimulus did not affect the general trend of loudness perception. Although mean values from the data are reported here for 0 and 50 CUs (i.e., "can't hear" and "too loud," respectively), one should note that actual mean values for these two extreme categories are undefined, because the upper and lower bounds of presentation levels could, theoretically, have been extended by any amount. In general, the data from 1 and 4 kHz stimuli did not produce significantly different relationships between loudness and level, once the higher hearing-loss thresholds for 4 kHz stimuli were accounted for.

Table I reports the listener loudness-discomfort levels (LDLs), along with the number of listeners who did not report any stimulus presentation as "too loud" from 0 to 105 dB SPL (denoted as "NA"). Further detail on the determination of a listener's dynamic range is included in Appendix B. For 4 kHz, the number of NA LDL listeners is about half of both HL and NH groups. The mean loudness attributed to 105 dB SPL was 50 CU for NH listeners and 40 CU for HL listeners. This is relevant to the interpretation of the mean CLS functions. Although the mean trend of loudness growth shown in Fig. 4 might be interpreted as representative, in fact, many of the CLS functions for individual listeners do not reach the highest CUs, even at 105 dB SPL. This can be further observed in the individual data, discussed in Sec. III B.

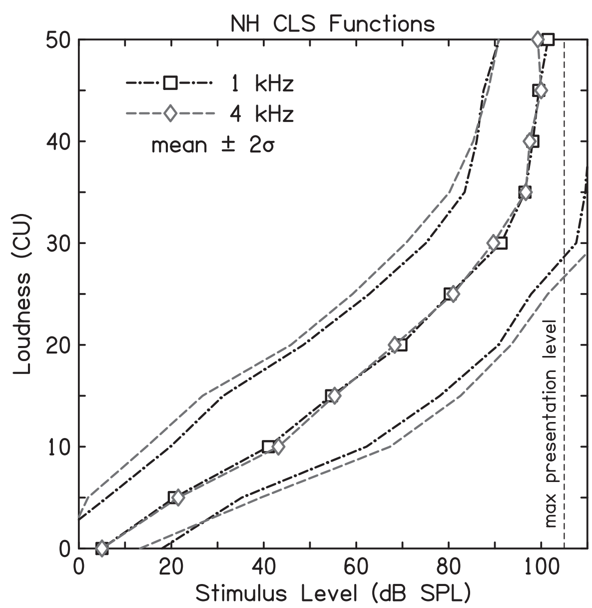

FIG. 4. Summary of CLS function data for NH listeners. Mean and 2 standard deviations are shown for 1 and 4 kHz stimuli.

B. Individual listener MCPFs

Representative CLS functions and corresponding MCPFs from listeners with NH and HL are shown in Figs. 5 and 6, respectively. Each pair of figures shows the data for one stimulus frequency and an individual listener; the CLS function for the listener is shown on the left and the MCPF for the same data is shown on the right. Each MCPF is composed of logistic curves, as described in Eq. (2). The MCPF figures show the category boundary curves for a listener's raw data as solid lines, with dashed-line logistic fits overlaid [logistic fits follow Eq. (1)]. The measured audiometric thresholds are marked on each MCPF with a dashed vertical line.

These NH and HL MCPF examples show some general patterns that apply to both the 1 and 4 kHz stimuli. The leftmost category boundary curve (between 0 and 5 CU), which demarcates the threshold of hearing, drops for most listeners by $\geq 75\%$ within 5 dB; this steep slope indicates that there is low listener uncertainty at the levels that are near threshold. The most across-listener variability in category width was observed for the 5 CU loudness category (i.e., "very soft"), which varied in width across listeners from a maximum of 50 dB (for a NH listener) to a minimum of 3 dB (for a HL listener); this corresponds to variability in the low-level slope of the CLS function. A higher threshold and/or lower LDL produce a horizontally compressed MCPF; this compression narrows the width of all intermediate loudness categories. We see that the steepness of category boundary slopes reflects the uncertainty of a listener's responses (i.e., shallower slopes indicate more decision variance across loudness categories at a fixed level), and that the shallowest slopes are observed across the 5 CU to 25 CU categories. A hearing loss narrows the MCPFs and results in sharper category boundary slopes. Finally, from the different listener examples, we observe that the audiometric threshold is correlated to the inflection point of the leftmost category boundary curve (between 0 and 5 CU). This correlation is further explored in Fig. 7.

Figure 5 shows data for two NH listeners, NH04 at 1 kHz (top) and NH01 at 4 kHz (bottom). Both CLS functions show the characteristic loudness growth that has been documented for NH listeners in the literature, which gradually increases in slope with level. These two listeners have similar leftmost category boundaries, indicating that they have similar detection thresholds. Compared to NH04, listener NH01 has a wider range of levels that correspond to 5 CU ("very soft") and a lower LDL, leaving a smaller dynamic range for the remaining loudness categories. This causes the MCPF to be horizontally compressed, mirroring the listener's sharper growth of loudness in the CLS function. The category boundary curves from 30 to 80 dB SPL for this listener are shallower, indicating probability distributions with more variance across loudness categories at these levels.

Figure 6 shows representative data for listeners with varying degrees of HL at 1 kHz (top two) and 4 kHz (bottom two). Listener HL17 has roughly even widths of each loudness category ($\approx$10 dB SPL), which is reflected in a CLS function that does not sharply steepen within the experimental presentation levels; at 105 dB SPL, only 1 out of 20 responses for this listener corresponded to 40 CU. Listener HL08 has a nearly identical CLS function to HL17, but the MCPF differs. This listener has a wider 5 CU ("very soft") range, slightly shallower category-boundary curves, and a lower LDL. The wider 5 CU range and lower LDL, as was the case with NH listeners, results in a MCPF in which all intermediate categories are narrowed. Data for HL26 and HL11 at 4 kHz are shown in the bottom two rows of Fig. 6. The data from HL26 are an example of an elevated LDL (no responses above 30 CU at 105 dB SPL). The data from HL11 are an example of a listener with hearing loss and an

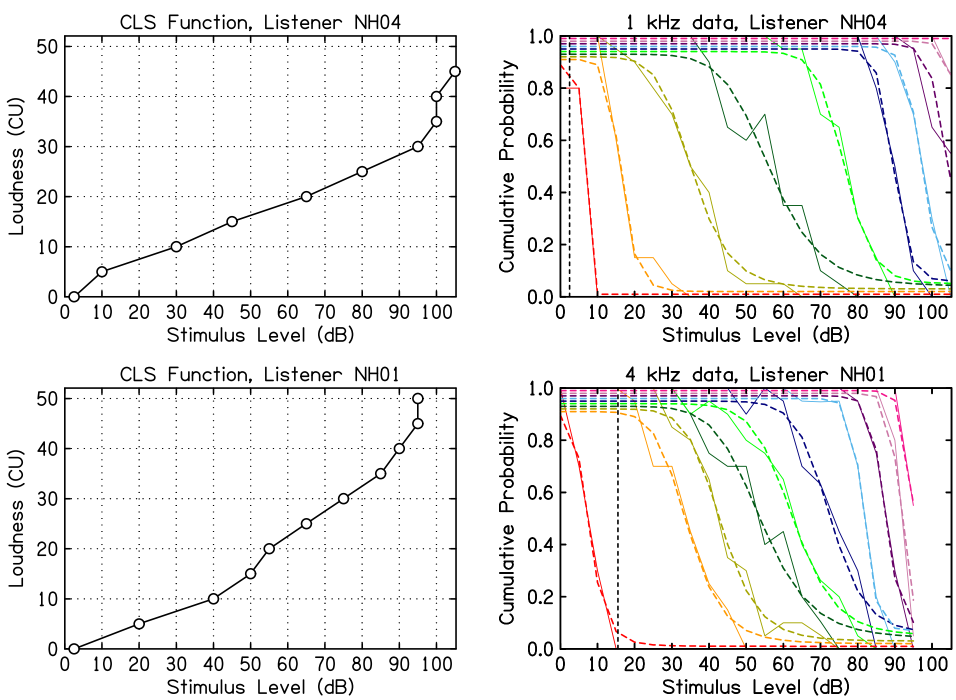

FIG. 5. (Color online) The CLS function and corresponding MCPF for 2 NH listeners: (top) NH04, 1 kHz stimuli, (bottom) NH01, 4 kHz stimuli. The CLS function plots the median level for each CU. The MCPFs show the raw data as solid lines and the parameterized logistic fits as dashed lines. Each listener's audiometric threshold is marked with a vertical line. The maximum presentation levels are constrained by the listener's LDL.

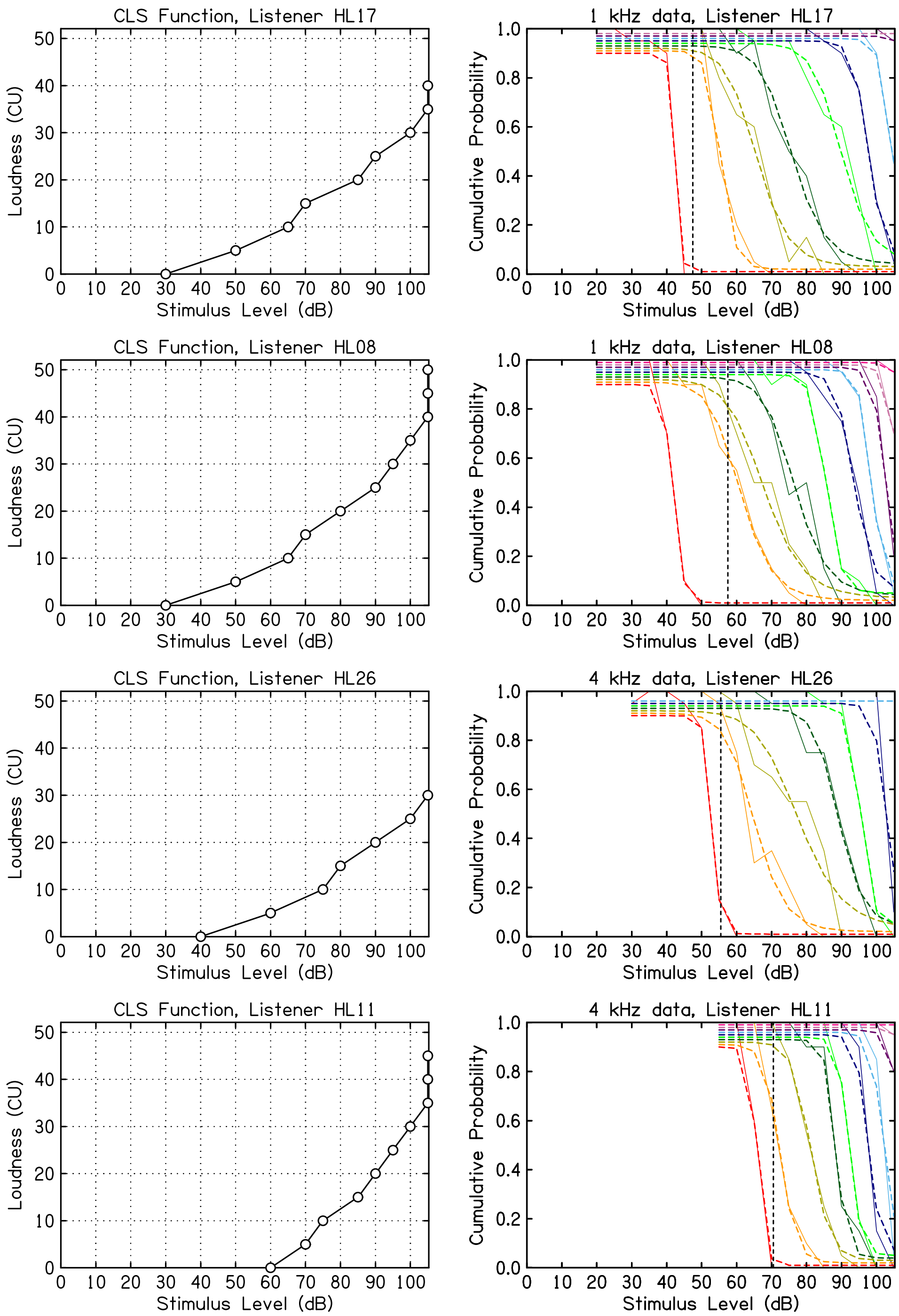

FIG. 6. (Color online) The CLS function and corresponding MCPF for four listeners with HL: (first row) HL17, 1 kHz stimuli, (second row) HL08, 1 kHz stimuli, (third row) HL26, 4 kHz stimuli, (fourth row) HL11, 4 kHz stimuli. The CLS function plots the median level for each CU. The MCPFs show the raw data as solid lines and the parameterized logistic fits as dashed lines. Each listener's audiometric threshold is marked with a vertical line.

LDL that is within the ranges of a NH listener. Despite the compressed dynamic range of HL11, this listener has well-defined perceptual boundaries between categories, indicated by the sharp slopes of the category-boundary curves.

The MCPF boundary curve between 0 and 5 CU describes a listener's sound detection as a function of level. The inflection point of this curve, represented by C_1 in the parameterization of Eq. (2), approximates a listener's threshold for sound detection. This value, which we call the *CLS threshold*, is compared in Fig. 7 to the *audiometric threshold* that was measured for each listener, using standard clinical protocols. This plot also provides additional detail about the distribution of thresholds that were represented by this experimental population. For reference, superimposed lines represent (1) equivalence between CLS and audiometric thresholds (dashed diagonal line) and (2) a linear regression of CLS threshold onto audiometric threshold (solid line). On average, the CLS threshold is approximately 4.4 dB below the audiometric threshold. The largest absolute difference between the audiometric threshold and the CLS threshold is 20 dB.

C. Principal component analysis

The results of the PCA were such that the first principal component captured 87% of the variability in the MCPF parameters across listeners, indicating strong correlations across the parameters. The first two components combined

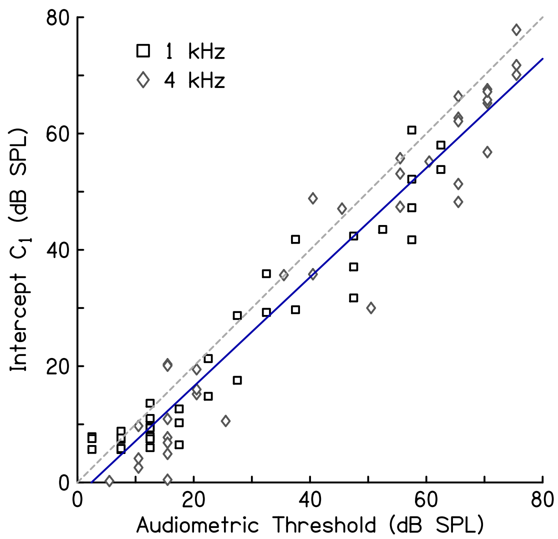

FIG. 7. (Color online) Comparison between the audiometric thresholds and the inflection point of the category-boundary curve between 0 and 5 CU. Individual listener data are shown for all NH and HL participants, for both 1 and 4 kHz stimuli. An equivalence dashed line and a solid best-fit line are shown for reference.

captured more than 90% of the variance. The percentage of variance explained by each principal component is shown in Fig. 8(a).

The individual listener weightings for each principal component are shown in Fig. 8(b) (open circles); these weighting values are calculated from individual listener projections onto the principal components, as described in Eq. (4). In addition to the individual data points, the range represented by two standard deviations about the mean, and the values of the sampled weightings (filled circles) are shown in this figure. The number of sampled weightings per principal component was approximately proportional to the percentage of variance explained.

In order to understand which MCPF parameters contribute the most to inter-subject variability and how the parameters are correlated, we examined the first four principal components in greater detail. The principal component values for each transformed MCPF C parameter are shown in Table II. The C parameters represent the inflection points of the category-boundary curves in the MCPF; the transformed C parameters are calculated as shown in Eq. (3). C_1 is the inflection point of the leftmost category-boundary curve,

TABLE II. Principal component values for the first four PCA vectors (rows). Each column corresponds to one of the transformed C parameters, described in Eq. (3).

	$\widehat{C_1}$	$\widehat{C_2}$	$\widehat{C_3}$	$\widehat{C_4}$	$\widehat{C_5}$	$\widehat{C_6}$	$\widehat{C_7}$	$\widehat{C_8}$	$\widehat{C_9}$	$\widehat{C_{10}}$
PC 1	0.96	−0.18	−0.14	−0.11	−0.09	−0.06	−0.01	−0.01	0	0
PC 2	0.14	0.93	0.11	−0.12	−0.20	−0.17	−0.12	−0.03	−0.02	−0.01
PC 3	0.10	−0.16	0.94	0.01	0.09	−0.27	0.01	0	0.02	−0.01
PC 4	0.20	0.22	0.05	0.78	0.45	0.30	0.05	0.06	−0.03	−0.03

which also represents the listener's hearing loss. The rest of the transformed C parameters are calculated as the difference between inflection points for adjacent category-boundary curves; this roughly represents the width of each particular loudness category, in dB. For example, C_2 is the width of the loudness category between the first and second category-boundary curves, which corresponds to the width of 5 CU or "very soft." From Fig. 8(b), we see that the majority of listener weightings for these four principal components are positive; therefore, mostly positive scalings of the principal components in Table II contribute to listener representations.

The first principal component is dominated by C_1, indicating that the listener's hearing loss is the primary factor that contributes to inter-subject variability. This principal component decreases the values of C_2 to C_8 as the listener's hearing loss (C_1) increases. This inverse relationship between hearing threshold and the width of loudness categories can be observed in the examples of Fig. 6. The second principal component represents the variability across listeners in the width of C_2 (i.e., the "very soft" category); higher-magnitude C_2 values correspond to a decrease in the C_4 to C_{10} values, horizontally compressing the MCPF at higher levels. The third and fourth principal components have the most effect on C_3 and C_4, respectively.

In general, the slope parameters [B in Eqs. (1)–(3)] had relatively low inter-subject variability. The shallowest category-boundary slopes, for both NH and HL listeners, are observed in B_2, B_3, and B_4. The final B parameter values are primarily determined by PCA vectors with lower energy than those which define the C parameter values, indicating that the transformed intercept and slope parameters may not be correlated and verifying our observation that slopes have lower inter-subject variability. The PCA vectors which had the most contribution to the final slope values were determined; while,

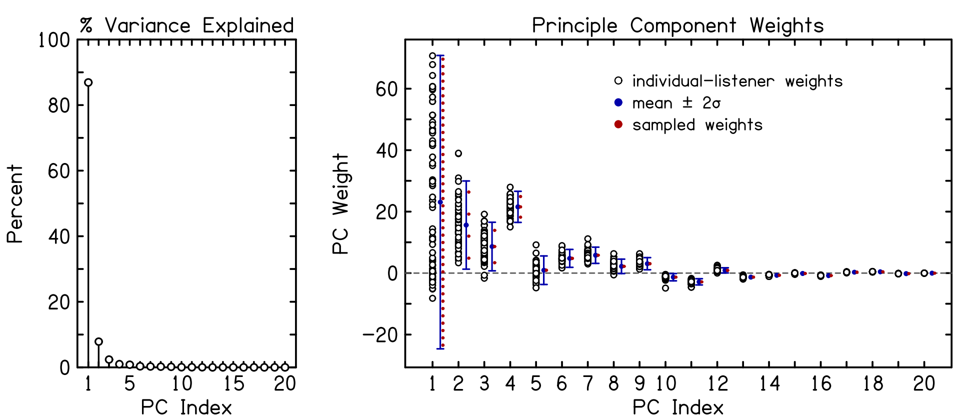

FIG. 8. (Color online) (left) The variance explained by each PCA basis vector. (right) The PCA basis vector weightings for each individual listener's data are displayed as open circles. The range representing twice the standard deviation about the mean of these data are marked with a solid line. The sampled weightings used for reconstruction are shown as solid circles.

in general, different PCA vectors were the primary factor for each B parameter, B_2, B_3, B_4 , and B_5 were mostly determined by PCA vector 11. This indicates a correlation between these particular B parameters.

D. Application to ML estimation

The MCPF may be used in any CLS application that requires a probabilistic model. Here, a ML estimation technique is used in conjunction with the MCPF catalog of potential listener outcomes in order to estimate a new listener's MCPF, as described in the Methods. After an estimate of the MCPF is obtained, a CLS function can be computed from the MCPF curve inflection points. This is done by considering each inflection point to be the level that corresponds to halfway between the CU values of the curve's adjacent loudness categories. For example, the level of the inflection point of the leftmost MCPF curve would define the CLS function at 2.5 CU (halfway between 0 and 5 CU). This ML estimation approach was designed with the aim of improving the accuracy of CLS results.

The ISO adaptive-level technique results in a low number of measurements, typically no more than three total at a particular level. Because of the low number of presentations and the statistical nature of categorical loudness perception, the resulting CLS function from a median-based estimation (as described in Kinkel, 2007) may be inaccurate. In addition, this number of presentations is not sufficient for an accurate MCPF to be fit to the data (fitting described in Sec. II). The ML estimation technique takes advantage of the statistical nature of CLS responses to find the MCPF that would produce a listener's set of responses with the highest statistical likelihood.

Figure 9 compares the results of computing the CLS function using ML estimation to the median-based approach, for two representative-listener examples (out of five total listeners). For the sake of simplicity in this example, we compare the ML approach to an approach that estimates the CLS function as the median level associated with each loudness category. In Fig. 9, both the median and the ML estimate from the same adaptive-level data are superimposed. The fixed-level results are used as the best estimate of the listener's "true" CLS function, and are shown as a solid line, for reference. Although the median-based computation of the CLS function can at times be accurate (see Fig. 9, left), large deviations from the fixed-level results can occur due to the inherent randomness in listener responses (see Fig. 9, right). In both example cases, the CLS function derived from the ML estimate of the listener's MCPF is more accurate. Overall, for the five tested listeners, the average root mean squared error (RMSE) (fixed-level data used as baseline) for the median estimate of the CLS function was 6.7 dB, and the average RMSE for the ML estimate was 4.2 dB. The ML MCPF can be used to estimate a listener's CLS function and additionally provides a probabilistic model of the listener's categorical loudness perception.

Clinical applications of this technique can use the resulting likelihood of the ML result as a goodness-of-fit metric. By reporting a confidence, or goodness-of-fit metric, a researcher or clinician would be able to identify cases that have not yet been represented in the MCPF catalog.

IV. DISCUSSION

The relation between perceptual judgments and physical intensity is a classical problem in psychophysics (Thurstone, 1927; Braida and Durlach, 1972). Many studies have examined the relationship between categorical loudness perception and stimulus level (e.g., Allen *et al.*, 1990; Brand and Hohmann, 2002). This study introduces the MCPF representation to CLS data. The MCPF quantifies a novel dimension of CLS data: the probability distributions across categories. The CLS MCPF represents how the probability distributions of categorical loudness perception vary with stimulus level. This representation can be used to model suprathreshold variability across listeners and to guide CLS algorithms.

CLS data were collected for listeners with NH and HL. The distribution of audiometric thresholds is shown in Fig. 7. Listeners with similar audiometric thresholds displayed a wide range of suprathreshold variability in their loudness perception (see Fig. 4). This suprathreshold variability could be due to individual differences in peripheral loss (e.g., deafferentation) or central adaptations (Kujawa and Liberman, 2009; Niemeyer, 1971). Several listeners did not report loudness discomfort within the stimulus range of 0–105 dB SPL, while others reached levels of discomfort by 90 dB SPL (Table I and Fig. 4). These observations show that, although all listeners have a general trend of steepening CLS function-slopes at high levels, one cannot assume that high-level loudness perception is the same across a group of listeners, even when their thresholds are similar. Data were collected for both 1 and 4 kHz stimuli; in general, subjects had higher audiometric thresholds at 4 kHz, but this did not significantly change the relationships among MCPF parameters that were observed. This suggests that the observations of this study can be generalized across frequency. The MCPF

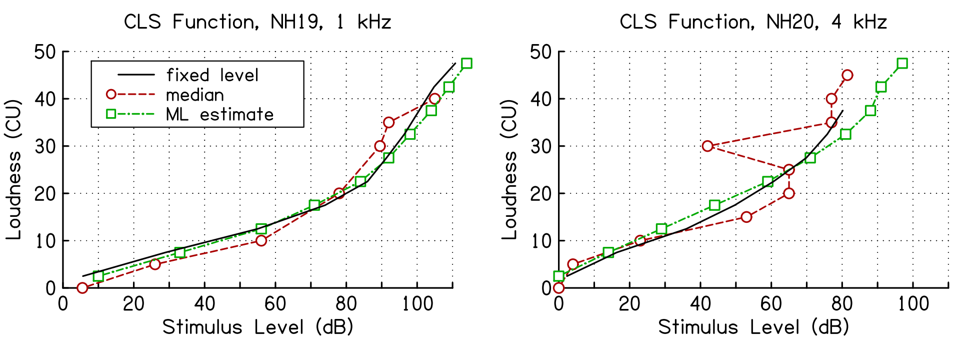

FIG. 9. (Color online) Comparison of NH CLS results for two representative NH listeners. The CLS function based on the fixed-level experiment is shown as a solid line, representing the "best" estimate of the listener's loudness perception. The ISO-adaptive median is shown with circle markers. The ML-based estimate is shown with square markers.

methods are not restricted to any particular type of stimulus, so the techniques described in this paper can also be generalized to different stimulus types (e.g., warble tones or half-octave-band noise).

Individual MCPFs, along with the corresponding CLS functions, are examined in Figs. 5 and 6. A number of observations can be made from these MCPFs. The largest variability across listeners was observed for the hearing threshold and the horizontal width of the "very soft" loudness category. These observations were verified by the PCA analysis. In addition, we observe that increased hearing loss, reduced loudness discomfort level, or wide low-level loudness categories all result in a horizontally compressed MCPF. This compression affects many loudness categories and is indicated by a steepening of the CLS function. The listener's variance across loudness categories, which is quantified by the MCPF category-boundary slopes, could not be inferred from the CLS function. The leftmost category-boundary curve provides a measure of the threshold of hearing by demarcating the boundary between the "can't hear" and "very soft" loudness categories. The smooth transition of this category boundary curve for each listener indicates that loudness is not discontinuous at threshold, which is consistent with current loudness models (Moore *et al.*, 1997). The inflection point of this curve, which is the CLS threshold, is compared to the measured audiometric threshold in Fig. 7. On average, the audiometric threshold was higher than the CLS threshold, with the largest observed difference being 20 dB. This demonstrates that listeners detect (with finite probability) the presence of the stimulus at and below audiometric threshold. This observation can guide future CLS task implementations as it demonstrates that the audiometric threshold should not be assumed to represent the lowest point of a listener's dynamic range; a more conservative estimate, that would provide some CLS measurements at 0 CU for most listeners, would be to initially use the audiometric threshold minus 20 dB when determining the lowest point of the dynamic range.

This paper describes how we fit a logistic parametric model to empirical MCPFs. This approach captured the patterns observed in the measured data with four parameters per category-boundary curve. The resulting parameters and a PCA were used to study inter-subject variability. A general catalog of potential listener MCPFs was developed, based on the PCA results. The catalog of listener models can be applied in a variety of ways. A common usage for such probabilistic listener models is to simulate listener behavior when developing experiments or listening devices. Probabilistic models also allow one to apply concepts from detection, information, and estimation theory to the analysis of results and the methodology of the experiment.

As an example of such an application, we show how to combine ML estimation with the catalog of listener MCPFs. The ML estimate is able to more-accurately predict a listener's underlying CLS function than a median-based estimate, without pre-processing to remove outliers or requiring that the CLS trend conforms to a single functional form. Although the MCPFs conform to a logistical functional form, the CLS function is only constrained to having a monotonic non-decreasing relationship. This ML application not only estimates the CLS function, but also gives additional insight into the listener's categorical perception by providing a MCPF model, without exceeding clinically acceptable test times. Some listener types may not be included in the 1460 MCPFs of the current catalog; one way to account for this would be to present the researcher or clinician with the resulting likelihood value for the ML result as a measure of confidence. Subsequent work on this ML application should examine the range of goodness-of-fits given a large clinical population; this would identify the frequency of CLS responses that are not represented by the catalog, and would produce the range of likelihoods that correspond to reasonable agreement between the ML CLS function and a fixed-level CLS function. In future work, the results of this ML application of the MCPF catalog could be compared to existing model-function fitting approaches for CLS function estimation.

The results of all CLS approaches are limited by the attention of the listener to the task. In addition, due to the inter-subject variability at the upper and lower bounds of the dynamic range, the initial determination of each listener's dynamic range that controls the limits of the adaptive or fixed-level CLS task must be implemented with redundancies (see Appendix B for further details).

The MATLAB code for the CLS tasks and the resulting MCPF catalog are publically available (Boys Town National Research Hospital, 2015). MCPF data gives insight into the internal noise that is inherent in the categorical decision process. In future work, the MCPF catalog can be used to model the behavior and outcomes of typical listeners. One such use would be to predict listener outcomes with different hearing aid algorithms. In addition, the model of listener statistics can be used in future simulations when designing new CLS data collection algorithms. As this work continues, the database of listener MCPFs will be extended to include less-common listener types, such as listeners with hyperacusis. Our future work will focus on using the MCPF catalog to design an adaptive-level tracking algorithm for CLS which uses concepts from information theory to maximize the information provided by each stimulus presentation.

V. CONCLUSIONS

The MCPF is a probabilistic representation that can be used to study any type of ordinal categorical data. The application of MCPF to the modeling of CLS data is novel and provides insight into inter-subject suprathreshold variability. A MCPF catalog, such as the one described in this paper, could serve as a basis for predicting loudness perception or developing novel experimental techniques.

ACKNOWLEDGMENTS

This research was supported by grants from the National Institutes of Health: T32 DC000013, R01 DC011806 (W.J.), R01 DC008318 (S.T.N.), and P30 DC004662. We would like to thank Katyarina Brunette, Sara Walker, Katie Thrailkill, Brenda Ohlrich, Judy Kopun, Mary Thomas, Marta Delgado, and Colleen Gibilisco for their help in recruiting subjects and data collection.

APPENDIX A: SEQUENTIAL EFFECTS

The fixed-level CLS data from Jesteadt and Joshi (2013) was used in order to investigate sequential effects in CLS testing. Sequential effects occur when the nature of the previously presented stimuli affect the perception of the current stimulus. Some listeners were found to have a higher susceptibility to sequential effects than others. Based on the maximum observed sequential effect across stimulus levels (0–105 dB SPL) in the Jesteadt and Joshi data, we limited the jumps in intensity between adjacent stimuli to no more than 45 dB to prevent sequential effects for all listeners.

For verification, we analyzed possible sequential effects in our data. A regression between previous stimulus level and categorical response was calculated for each stimulus level and listener. On the basis of the regression, the levels at which sequential effects may be affecting a listener's categorical response were identified. We found that 8% of the data points may have been affected by sequential effects, and that in the majority of cases these points corresponded to categorical responses that were adjacent (i.e., ±5 CU) to the median response category. We concluded that sequential effects cannot be entirely eliminated by limiting the level differences between stimuli, but that the resulting sequential effects are relatively small. In order to appropriately model the sequential effects that can result from standard CLS testing, we did not attempt to remove sequential effects from the model or analyses. In interpreting the variability across categorical responses for a fixed stimulus level, it is important to keep in mind that approximately 8% of the variability in the distribution may be due to such sequential effects.

APPENDIX B: DETERMINING THE DYNAMIC RANGE LIMITS

Much of the inter-subject variability is observed at the lower end of listeners' dynamic range. Loudness perception at the upper end of listeners' dynamic range is a significant factor in determining the comfort of a hearing device. For these reasons, correct determination of both ends of an individual's dynamic range is critical when assessing loudness perception.

The ISO standard method for determining the limits of a listener's dynamic range uses two interleaved tracks of stimuli, one ascending and one descending in level. The tracks begin at 65 dB SPL and spread out along the dynamic range until one response of "can't hear" and one response of "too loud" are given by the listener, or until the bounds of allowable stimulus presentations have been reached.

In our pilot data, we found that listeners are prone to responding "can't hear" at stimulus levels above threshold, especially during the listener's initial determination of the dynamic range. In addition, we observed that the first "too loud" that they respond with is not necessarily the upper limit of their dynamic range, as they are still becoming acclimated to loudness testing. We developed a more redundant algorithm for determining the dynamic range, compliant with the ISO standard; this algorithm adds at most 30 s to the loudness testing. The lower and upper bounds of our presentation levels were 0 and 105 dB SPL, respectively.

Using interleaved stimuli tracks that are simultaneously ascending and descending has the potential to produce sequential effects in the listener, once the level difference between adjacent stimuli is greater than 45 dB. Because of this, our technique first determines the lower bound of the dynamic range, followed by the upper bound. The tradeoffs between sequential effects and possible biases due to not using interleaved stimuli require further study as these methods are evaluated for inclusion in clinical testing.

To determine the lower bound of for CLS testing, our algorithm initially set the estimated lower bound as the audiometric threshold minus 30 dB. A stimulus at 65 dB SPL was played. A fixed set of stimuli at 50, 35, 20, 5, 15, 0, 25, and 45 dB SPL (excluding any stimuli more than 30 dB below threshold) were then tested, in that order. Using a fixed set of stimuli prevents random "can't hear" responses from prematurely ending the test. On the basis of the cumulative responses over level, the lowest level corresponding to a 0 CU response was set as the lower bound. The algorithm then tested stimuli at levels −10, 0, +10, +20 dB relative to the lower bound. The lower bound was re-calculated and this final level minus 10 dB was used as the lower bound of presentation levels for the listener. This ensured that the CLS task would have data for the "can't hear" category. If the estimated lower bound was higher than the audiometric threshold, then threshold minus 5 dB was used as the lower bound for CLS testing.

To determine the upper bound of a listener's dynamic range, we simply added redundancy to the existing ISO technique. The stimuli begin at 65 dB SPL, increasing by 10 dB below 90 dB SPL, and increasing by 5 dB above 90 dB SPL. Once that a listener responds with "too loud" or the presentation level reaches 105 dB, the stimulus level resets to 65 dB SPL. The upper bound determination ends when the listener responds "too loud" twice or the 105 dB SPL presentation level is reached three times. The lowest level which the listener indicated was "too loud" with the majority of their responses was set as the loudness discomfort level.

APPENDIX C: REDUCING FATIGUE ERRORS

The monotonous nature of the task can lead to loudness responses that do not reflect perception due to inattention or fatigue. The most common anomalous responses that we observed in pilot data were loudness judgments of "can't hear" for stimuli that were >10 dB above the listener's audiometric threshold. The "can't hear" response above threshold could be observed for stimuli as high as 90–105 dB SPL, and was more common for listeners with HL. Besides encouraging breaks during testing, a "repeat" button was added to the graphical user interface that would repeat the stimulus one time, if needed, to increase the likelihood that responses would reflect perception, not fatigue.

APPENDIX D: BIAS FOR LABELS

The CLS task interface, shown in Fig. 1, has bars with meaningful labels and unlabeled intermediate bars. A percentage of listeners (≈10%) showed some bias toward the labeled bars. Because of this, when instructing the listener on the CLS task, the tester should clearly explain why some buttons are unlabeled and explicitly tell the listener to use both types of buttons during the task.

[1]The term setoff is introduced in this paper as an alternative to offset to avoid confusion with the more common meaning of offset as a vertical displacement.

Allen, J. B. (**2008**). "Nonlinear cochlear signal processing and masking in speech perception," in *Springer Handbook of Speech Processing* (Berlin, Springer), pp. 27–60.

Allen, J. B., Hall, J. L., and Jeng, P. S. (**1990**). "Loudness growth in 1/2-octave bands (LGOB)—A procedure for the assessment of loudness," J. Acoust. Soc. Am. **88**(2), 745–753.

Al-Salim, S. C., Kopun, J. G., Neely, S. T., Jesteadt, W., Stiegemann, B., and Gorga, M. P. (**2010**). "Reliability of categorical loudness scaling and its relation to threshold," Ear Hear. **31**(4), 567–578.

Amitay, S., Irwin, A., Hawkey, D. J. C., Cowan, J. A., and Moore, D. R. (**2006**). "A comparison of adaptive procedures for rapid and reliable threshold assessment and training in naive listeners," J. Acoust. Soc. Am. **119**(3), 1616–1625.

Boys Town National Research Hospital (**2015**). "Categorical loudness scaling," http://audres.org/cel/cls/ (Last viewed September 26, 2016).

Braida, L., and Durlach, N. I. (**1972**). "Intensity perception. II. resolution in one-interval paradigms," J. Acoust. Soc. Am. **51**(2B), 483–502.

Brand, T., and Hohmann, V. (**2002**). "An adaptive procedure for categorical loudness scaling," J. Acoust. Soc. Am. **112**(4), 1597–1604.

Fletcher, H., and Munson, W. A. (**1933**). "Loudness, its definition, measurement and calculation," Bell Syst. Tech. J. **12**(4), 377–430.

Florentine, M., Buus, S., and Geng, W. (**2000**). "Toward a clinical procedure for narrowband gap detection: I. A psychophysical procedure," Audiology **39**(3), 161–167.

Green, D. M. (**1993**). "A maximum-likelihood method for estimating thresholds in a yes–no task," J. Acoust. Soc. Am. **93**(4), 2096–2105.

Heeren, W., Hohmann, V., Appell, J. E., and Verhey, J. L. (**2013**). "Relation between loudness in categorical units and loudness in phons and sones," J. Acoust. Soc. Am. **133**(4), EL314–EL319.

Jesteadt, W., and Joshi, S. N. (**2013**). "Reliability of procedures used for scaling loudness," Proc. Meet. Acoust. **19**, 050023.

Kinkel, M. (**2007**). "The new ISO 16832 'Acoustics–loudness scaling by means of categories'," in *8th EFAS Congress/10th Congress of the German Society of Audiology*, pp. 1–4.

Kujawa, S. G., and Liberman, M. C. (**2009**). "Adding insult to injury: Cochlear nerve degeneration after 'temporary' noise-induced hearing loss," J. Neurosci. **29**(45), 14077–14085.

Marks, L. E., and Florentine, M. (**2011**). "Measurement of loudness, Part I: Methods, problems, and pitfalls," in *Loudness*, edited by M. Florentine, A. N. Popper, and R. R. Fay (Springer, New York), pp. 17–56.

Moore, B. C. J., Glasberg, B. R., and Baer, T. (**1997**). "A model for the prediction of thresholds, loudness, and partial loudness," J. Audio Eng. Soc. **45**(4), 224–240.

Niemeyer, W. (**1971**). "Relations between the discomfort level and the reflex threshold of the middle ear muscles," Int. J. Audiol. **10**(3), 172–176.

Oetting, D., Brand, T., and Ewert, S. D. (**2014**). "Optimized loudness-function estimation for categorical loudness scaling data," Hear. Res. **316**, 16–27.

Rasetshwane, D. M., Trevino, A. C., Gombert, J. N., Liebig-Trehearn, L., Kopun, J. G., Jesteadt, W., Neely, S. T., and Gorga, M. P. (**2015**). "Categorical loudness scaling and equal-loudness contours in listeners with normal hearing and hearing loss," J. Acoust. Soc. Am. **137**(4), 1899–1913.

Thurstone, L. L. (**1927**). "A law of comparative judgment," Psychol. Rev. **34**(4), 273–286.

Torgerson, W. (**1958**). *Theory and Methods of Scaling* (Wiley, New York), 460 pp.

Wright, B. A., Lombardino, L. J., King, W. M., Puranik, C. S., Leonard, L. M., and Merzenich, M. M. (**1997**). "Deficits in auditory temporal and spectral resolution in language-impaired children," Nature **387**, 176–178.