5-2018

Development of a Model to Predict the Likelihood of Complaints due to Assorted Tone-in-Noise Combinations

Joonhee Lee
University of Nebraska - Lincoln

Lily M. Wang
University of Nebraska - Lincoln, lwang4@unl.edu

Lee, Joonhee and Wang, Lily M., "Development of a Model to Predict the Likelihood of Complaints due to Assorted Tone-in-Noise Combinations" (2018). *Architectural Engineering -- Faculty Publications*. 119.
https://digitalcommons.unl.edu/archengfacpub/119

Development of a model to predict the likelihood of complaints due to assorted tone-in-noise combinations

Joonhee Lee[a] and Lily M. Wang
Durham School of Architectural Engineering and Construction, University of Nebraska-Lincoln,
1110 South 67th Street, Omaha, Nebraska 68182, USA

(Received 21 September 2017; revised 6 March 2018; accepted 16 April 2018; published online 4 May 2018)

This paper develops a model to predict if listeners would be likely to complain due to annoyance when exposed to a certain noise signal with a prominent tone, such as those commonly produced by heating, ventilation, and air-conditioning systems. Twenty participants completed digit span tasks while exposed in a controlled lab to noise signals with differing levels of tones, ranging from 125 to 1000 Hz, and overall loudness. After completing the digit span tasks under each noise signal, from which task accuracy and speed of completion were captured, subjects were asked to rate level of annoyance and indicate the likelihood of complaining about the noise. Results show that greater tonality in noise has statistically significant effects on task performance by increasing the time it takes for participants to complete the digit span task; no statistically significant effects were found on task accuracy. A logistic regression model was developed to relate the subjective annoyance responses to two noise metrics, the stationary Loudness and Tonal Audibility, selected for the model due to high correlations with annoyance responses. The percentage of complaints model showed better performance and reliability over the percentage of highly annoyed or annoyed.

https://doi.org/10.1121/1.5036731

[SF] Pages: 2697–2707

I. INTRODUCTION

An assortment of building mechanical equipment types generate prominent tones via rotating parts like fans and pumps. These tonal noises can cause an unpleasant evaluation of spaces and potentially increased complaints by building occupants. As mechanical equipment in buildings becomes more energy efficient, the tones produced by such equipment can become more prominent. The heating, ventilation, and air-conditioning industry is consequently interested in developing guidelines for what degree of tonality in different noise signal levels is acceptable to building occupants. Current indoor noise evaluation methods such as Noise Criteria (NC) (Beranek, 1957) and Room Criteria (RC) (Blazier, 1981) do not directly account for tonal characteristics of noises (Ryherd and Wang, 2010). Standards exist that describe metrics to quantify tonality (ISO, 2007; ANSI, 2010), and a few assessment methods consider tonality by adding a penalty to the overall level (Kryter and Pearsons, 1965; AHRI, 2012), which are then often checked against a community noise guideline. These methods apply the same penalty for a given tonality, no matter what the signal's overall level is (45 dBA or 55 dBA, for example). Evidence-based guidelines that, instead, specify maximum acceptable tonality for specific background noise levels in buildings do not currently exist, but manufacturers of equipment that impact building occupants would benefit from such knowledge, particularly in equipment design phases. Thus, this study aims to develop a model to predict how tonal components at frequencies ranging from 125 Hz to 1 kHz in noise affect annoyance responses and subsequent likelihood to complain, with the goal of suggesting maximum allowable levels of tones for a given overall noise level. This study also discusses captured performance outcomes related to task accuracy and completion time; these are analyzed to investigate how tonality, loudness, and the interaction of the two can affect human performance.

Noise-induced annoyance has been considered to be a key factor in environmental noise assessment. However, there is a degree of uncertainty around the term "annoyance" for acoustic researchers, primarily because the aims of noise annoyance studies vary according to the background contexts. According to a definition provided by ISO/TS 15666 (2003), noise-induced annoyance is "one person's individual adverse reaction to noise in various ways including dissatisfaction, bother, annoyance and disturbance." This ISO standard aims to provide specifications for annoyance questionnaires mainly about community noise. The World Health Organization (2011) approaches noise annoyance as having an adverse effect on health. In the WHO report, noise annoyance is defined as "a variety of negative responses, such as anger, disappointment, dissatisfaction, withdrawal, helplessness, depression, anxiety, distraction, agitation or exhaustion." Consequently, noise annoyance can cause psychosocial symptoms like tiredness, stomach discomfort, and stress. Guski *et al.* (1999) approached noise-induced annoyance as a multi-dimensional concept related to behavioral effects such as disturbance and interference and evaluative aspects like nuisance and unpleasantness.

Because the term annoyance embodies broad perceptual concepts, a variety of specific definitions for annoyance have

---

[a] Current address: Department of Building, Civil and Environmental Engineering, Concordia University, 1515 Rue Sainte-Catherine O, Montréal, H3G 2W1, Canada. Electronic mail: Joonhee.Lee@concordia.ca

been suggested by some previous studies (Guski *et al.*, 1999; Perdersen, 2007). Although differences of opinion still exist, there appears to be general agreement that the annoyance caused by a noise is influenced by the noise signal characteristics, the context of the measurement, and personal attributes. In this paper, annoyance is measured in a laboratory setting, in a specific context when participants are carrying out cognitively demanding tasks. No personal attributes are considered except noise sensitivity.

There is a consensus among noise researchers that impulsivity and tonality are two main noise signal characteristics for assessing annoyance besides loudness (Sailer and Hassenzahl, 2000; Marquis-Favre *et al.*, 2005; Pedersen, 2007). A substantial amount of literature has also been published on the annoyance of tones in noise. Hellman (1985) showed that annoyance and noisiness perceptions were highly related to the tonal components in noises. The author also discussed that the number of tones and frequency difference are factors that impact annoyance. Landström *et al.* (1995) investigated the noise annoyance in working spaces exposed to background noises with different spectral shapes. They found that the relation between individual annoyance ratings and sound levels was weak due to the absence of any metric for tonal components in the noise. Miedema and Vos (1998) also suggested extra correction factors for impulsive or tonal components when predicting total annoyance for transportation and industrial noises. Ryherd and Wang (2008) investigated ventilation-type mechanical noises and showed that current indoor noise criteria were not accurately reflecting subjective annoyance because tonal characteristics of noises are not directly included for assessment. More and Davies (2010) examined the effects of tones in aircraft noise on human annoyance and found that subjective loudness and tonality both influenced overall annoyance ratings. Previous work by the authors additionally confirmed that both loudness and tonality impact annoyance, although the influence on task performance using Sudoku puzzles was not statistically significant (Francis *et al.*, 2014; Lee *et al.*, 2017); results from testing two tonal frequencies of 125 and 500 Hz were used to develop a model that predicts annoyance of broadband noises with tones. The current paper presents a more extensive investigation that includes a wider number of tonal signals (four tonal frequencies, from 125 Hz to 1 kHz, at five different tonalities above two different ambient noise levels) and links results not only to annoyance responses but also to likelihood to complain. The result aims to provide a reliable model that can guide manufacturers and others to deal with tonal noise spectra in buildings on generally acceptable limits.

A. Tonal noise metrics

A number of metrics that have been developed to quantify the perception of tonality in noise are utilized in this paper. The term "tonality" for these metrics and, thus in this paper, refers to the perceived magnitude of tonal component (Hansen *et al.*, 2011). Tone-to-Noise Ratio (TNR) and Prominence Ratio (PR) are two widely used metrics for tonality perception that are standardized in ANSI/ASA S12.10/Part 1 (2010) Annex D. Frequency spectra from Fast Fourier Transform (FFT) analysis without any frequency weighting filters are used to calculate the TNR and PR. Caution needs to be taken during the FFT analysis to ensure that the frequency resolution is less than 1% of the tone frequency of interest for PR and 0.25% for TNR. As the name implies, TNR is the decibel level difference of the tone and masking noise within the critical bandwidth centered on the tone frequency. The PR is defined as the exceedance level of the critical band centered on the tone to the average level of the two adjacent critical bands. Both metrics may be used to analyze each tone in a noise signal independently unless multiple tones are in the same critical band. According to the ANSI standard, TNR may be more appropriate for multiple tones in adjacent critical bands, whereas the PR is more accurate for multiple tones within the same critical band.

ISO 1996-2:2007 (2007) Annex C introduces the tonality metric of Tonal Audibility. This metric is calculated based on the steady-state A-weighted frequency spectrum of a noise recording. In the standard, tones are technically defined as local maxima with a 3-dB bandwidth smaller than 10% of the bandwidth of the critical band. Like PR and TNR, the maximum Tonal Audibility value represents total tonal components in a spectrum if multiple tones exist. There are two main differences of Tonal Audibility as compared to PR and TNR. One major difference is that Tonal Audibility includes a frequency correction term in the calculation so that the prominence criteria of tones is constant across frequencies. The other difference is that it uses a linear regression line instead of actual noise components when calculating masking sound levels within the critical bands.

There are a number of other tonality metrics that have been developed by researchers, but these have not been standardized yet in the noise community. Aures (1985) developed a tonality metric that includes the frequency, bandwidth, and levels of all tonal components rather than of a single tone. For a product noise like heating, ventilation, and air-conditioning (HVAC) noise, Hastings *et al.* (2003) proposed modifications of Aures' metric with roll-off rates and bandwidth of narrow-band noises for better correlation of tonality and the metric. Susini *et al.* (2004) investigated the sound quality of indoor air-conditioning units, and found that one of the dominant perceptual structures that determines the sound quality is highly correlated with Noise-to-Harmonic Ratio, the ratio of the broadband noise part and harmonic parts, by resynthesizing the noise with digital signal processing techniques. Spectral Contrast, developed by Berglund *et al.* (2002), is a tonality metric that counts the number of local maxima of Zwicker's (1961) specific loudness critical-band spectra. In their subjective test, Spectral Contrast has the highest correlation with perceptual results.

A variety of noise metrics related to the sound energy or loudness have also been used to assess noise-induced annoyance depending on the context of the studies. The A-weighted equivalent sound level is the most common noise metric for environmental noise assessment because it is easy and convenient to measure. Other widely used loudness metrics are Day-Night average sound level (Kryter, 1982; Miedema and Vos, 1998) for community noises, statistical noise levels for time-fluctuating noises (Tang, 1997), and loudness from the standards ISO532-1:2017 (2017) and ISO532-2:2017 (2017), based on loudness models from Zwicker (1961) and Glasberg and Moore (2006), respectively.

There are a few noise metrics that consider both loudness and tonality to quantify the overall rating of tonal noises. These add penalty values based on the tonality to base signal levels, essentially setting the tonality to be equivalent to some increase in level. Kryter (1960) developed a noise metric named Perceived Noise Level (PNL) for aircraft noise. The metric is based on equal "noisiness" contours from subjective equal annoyance. However, the study by Little (1961) found a weak relation between PNL and annoyance for noises with tones. PNL was revised with a tone-correction factor and named tone-corrected PNL to predict tonal noises (Kryter and Pearsons, 1965). The tone-correction factor varies from 0 to 6.7 dB according to the frequency of tones and level differences between one-third octave band values. They compared subjective noisiness of five levels of tones in octave band noises to the octave band noises without tones at five different frequencies. The Joint Nordic Method (Pedersen *et al.*, 2000) is also standardized in ISO 1996-2:2007 (2007). Penalty k values derived from Tonal Audibility are added to A-weighted sound pressure levels. The penalty values vary from 0 to 6 dB based on subjective tests using artificial and real recordings from industry and wind turbine noises. More recently, the Sound Quality Indicator has been developed for the Air-Conditioning, Heating, and Refrigeration Institute to rate the sound quality of building mechanical product noise (AHRI, 2012), but is not widely in use at this time. The metric is based on the PNL procedure and Zwicker's loudness (Zwicker, 1961). The calculation procedure begins with one-third octave band data, and then the rating is adjusted when a one-third octave band value exceeds the average of the two adjacent bands values by at least 1.5 dB.

Although there has been considerable research on detecting and quantifying tones, many former studies focused mainly on the relation between annoyance and tones for aircraft noises, resulting in development of environmental noise assessment metrics (e.g., tone-corrected PNL and Joint Nordic Method) that assign a tonality penalty to the overall signal level quantity. For HVAC equipment in buildings that produce strong tonal components, however, a suggestion is made herein to approach the problem from another angle: given the tonal noise spectra and location of a piece of mechanical equipment in a building, can one predict if the resulting transmitted spectra at an occupied space will lead to annoyance, or furthermore complaints? Rather than applying penalty values of up to commonly 6 dB based on tonality to meet specified indoor noise criteria, this paper's goal is to suggest an alternative method wherein a tonal signal is run through a model developed to determine the likelihood that the signal results in annoyance and complaints. A subsequent guideline for HVAC tonal noise, then, could be tied to an acceptable annoyance rating or likelihood of complaints.

B. Noise exposure models

One of the main aims of environmental noise studies is to propose acceptable noise levels based on human responses. Noise exposure models that relate noise levels and annoyance have been used in the noise community to suggest maximum allowable noise levels. Generally, the percentage proportion of highly annoyed (%HA) or annoyed (%A) persons is predicted by the model with related noise levels such as A-weighted sound pressure level. Assorted previous studies have used numerical scales for the annoyance survey (Miedema and Vos, 1998; Miedema and Oudshoorn, 2002; Pedersen *et al.*, 2009). In such cases, the commonly-used categorization method is that ratings above 72 out of 100 indicate being highly annoyed, while ratings above 50 out of 100 indicate being annoyed (Pedersen, 2007), although no standardized break point exists. These noise exposure models have typically been developed with a logistic regression model or a quadratic ordinary least squares regression model (Miedema and Vos, 1998). The logistic regression model is a regression with a categorical outcome variable, as expressed by

$$P(Y) = \frac{1}{1 + e^{-(C_0 + C_1X_1 + \cdots + C_nX_n)}}, \qquad (1)$$

where $P(Y)$ is the possibility of outcome Y occurring, C_0, $C_1,\ldots, C_n$ are coefficients of the model and $X_1,\ldots, X_n$ are prediction variables, which are typically noise levels for noise studies. For the logistic regression model, maximum-likelihood estimation is used to estimate the coefficients of the model (Field, 2013).

Dose-response models have been developed using large data sets, sometimes from compiling results across independent noise exposure model studies for a number of noise source types: wind turbines (Pedersen *et al.*, 2009; Janssen *et al.*, 2011), aircraft, road traffic, and railway noise (Schultz, 1978; Fidell *et al.*, 1991; Miedema and Vos, 1998). While dose-response models can be based on huge data sets from field measurements, there still remains some uncertainty and a wide confidence interval in many dose-response relationships (Schomer, 2001, 2005). The study presented in this paper is a one-noise exposure model study that cannot be extended to a broadly tested dose-response model yet, but the objective is to propose a model that uses annoyance ratings and likelihood to complain as the outcome variable, against a number of the noise metrics described in Sec. III C for broader applicability when analyzing acoustic conditions in buildings.

II. METHODS

A. Participants

Twenty listeners (9 females, 11 males) were paid to participate in the subjective test. The participants were recruited with fliers distributed throughout the University of Nebraska at the Omaha campus. The average age across all participants was 24.9 years with a standard deviation (SD) of 4.9. Most participants were university students or staff members. All listeners first participated in a 30-min orientation session including a hearing sensitivity test to confirm that they had hearing thresholds below 25 dB hearing level (HL) from 125 Hz to 8 kHz for both ears.

Noise sensitivity was also measured by using the reduced version of the Noise-Sensitivity-Questionnaire (NoiSeQ), which consists of 13 question items (Schutte *et al.*, 2007); the scores confirmed that the sample is a good approximation of the general population by meeting the normality assumption with a non-significant Shapiro-Wilke test result ($p = 0.82$). Figure 1 shows a histogram of the NoiSeQ values for the 20 participants. The average of NoiSeQ score was 2.75 with a SD of 0.44.

B. Equipment and stimuli

The subjective tests were carried out in an indoor acoustic testing chamber at the University of Nebraska (Fig. 2). The 27.8 m^3 room is acoustically isolated from nearby spaces with a room-in-room design and a floating floor. Two side walls are slightly slanted to reduce flutter echoes and minimize effects of room modes. Materials in the room include carpet on the floor, gypsum board walls with additional absorptive panels, acoustic bass traps, and acoustical ceiling tiles. The mid-frequency reverberation time, averaged across the 500–2000 Hz octave bands, is 0.22 s, and the ambient background noise level is 32 dBA (re 20 μPa). Signals were generated through an Armstrong i-ceiling speaker panel (A-50 speaker, D2001 digital processor and D4001 amplifier) and a subwoofer (JBL E250P) in a corner. The i-ceiling speaker panel looks identical to the other ceiling tiles so that participants could not visually identify the location of this sound source. The speaker system was connected to a test computer via a Presonus AudioBox 44VSL mixer. Participants sat roughly in the middle of the chamber during the test.

The test stimuli totaled 45 signals. The 40 test signals were artificially synthesized broadband signals with tonal components. The other five test signals were audio recordings with actual HVAC noises. The five test signals were measured tonal noises from a heat pump, fume hood, and a screw compressor; refer to Ryherd and Wang (2010) for the details of the recorded signals (Signal $T2$ to $T6$). These five signals were only used for prediction model development in Sec. III C. For the artificially synthesized signals, two different broadband sound spectra were used, complying with the room criteria RC-30 and RC-38 neutral contours. Neutral spectra were selected to eliminate perceptual impacts caused by spectral elements other than by the tones.

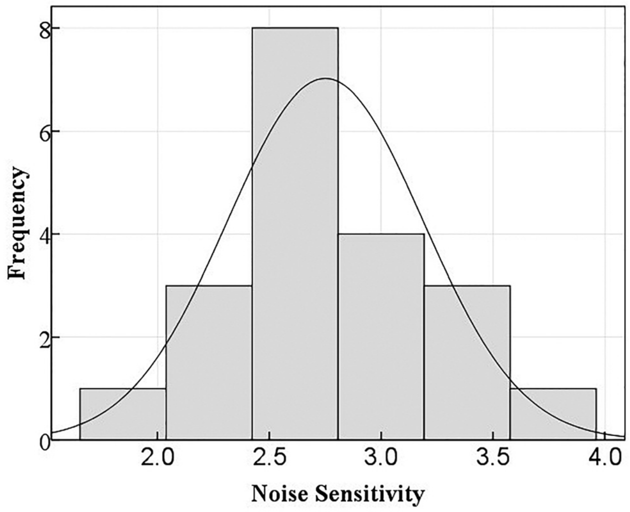

FIG. 1. Histogram showing distribution of test participants' noise sensitivity scores as measured by the NoiSeQ reduced survey. The solid curve shows a normal distribution curve.

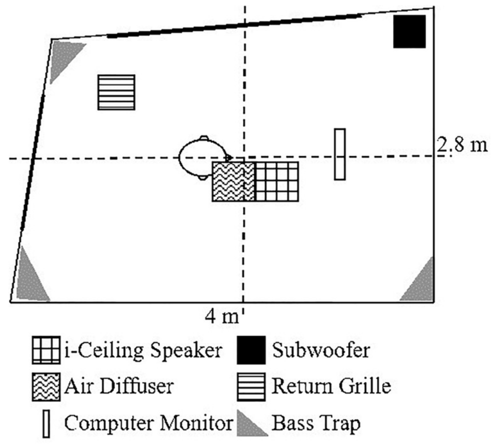

FIG. 2. Schematic plan of the testing chamber.

Measurements of the test signals were averaged over a minute at the listener's ear position (3′7″) in the test chamber using a B&K 4189-A microphone through the B&K PULSE system. The overall sound pressure levels of the two broadband signals at the listener position were 57 dB (re 20 μPa) and 63 dB (re 20 μPa), respectively. Five levels of tones at one of four specific tonal frequencies (125 Hz, 250 Hz, 500 Hz, and 1 kHz) were added separately to the broadband signals. These tonal frequencies were selected as many types of HVAC equipment exhibit tones in this frequency range. Tones at frequencies lower than 125 Hz are also common in modern HVAC equipment, but the sound system used in this study was not able to reproduce tones near 63 Hz at sufficient levels.

The Tonal Audibility metric was used as a reference metric for test signals because it is a frequency-independent metric, which means the signals can be easily compared across frequencies for tonal magnitude. Additionally, in the authors' previous studies (Lee *et al.*, 2017), Tonal Audibility was found to be a better indicator than others with regards to correlation to annoyance. Tonal Audibility values in the current study were between 5 and 19 dB. These tone levels fall within a range that encompasses from below to above the prominence thresholds listed in ANSI S12.10-2010 (2010) for PR and subjectively range from just audible to very prominent for each tonal frequency. The overall A-weighted sound pressure level of the 40 tonal signals ranged from 37.5 dBA (re 20 μPa) to 49.6 dBA (re 20 μPa), as summarized in Table I. The broadband signals and tones were generated using the program Test Tone Generator program (Esser Audio) and digitally synthesized using the program Audacity 2.1.1. Figure 3 shows one-third octave band spectra of the test stimuli for the lowest, middle, and highest tonal levels. The

TABLE I. Tonality and sound levels of artificial noise stimuli used in the subjective test. The same levels of tones are added to both the RC-30 and RC-38 neutral spectra broadband noises.

	Tonal Audibility (dB)				
Frequency (Hz)	Tone level 1	Tone level 2	Tone level 3	Tone level 4	Tone level 5
125	5.4	7.2	9.4	13.2	19.4
250	5.7	7.7	9.7	12.7	19.5
500	5.2	7.5	9.9	12.1	19.0
1000	5.1	7.8	10.1	13.5	19.2
$L_{a,eq}$ (dBA) for RC-30 / RC-38 based noise signals					
125	38.4 / 46.2	38.9 / 46.7	39.7 / 47.5	42.1 / 49.6	46.9 / 54.1
250	38.1 / 46.1	38.7 / 46.5	39.4 / 47.0	40.3 / 48.0	42.1 / 49.6
500	37.6 / 45.6	38.1 / 45.9	38.8 / 46.6	40.0 / 47.7	41.6 / 49.2
1000	37.5 / 45.5	37.9 / 45.9	38.5 / 46.5	39.5 / 47.5	41.0 / 49.0

Tonal Audibility and A-weighted sound pressure level values for the additional five signal recordings were 7.4, 12.5, 4.2, 10.3, and 3 dB, and 44.7, 45.7, 44.4, 45.0, and 44.7 dBA (re 20 μPa), respectively. All test stimuli were steady when played with no obvious fluctuations in time.

C. Subjective testing procedure

The subjective test procedures, reviewed and approved by the University of Nebraska – Lincoln Institutional Review Board (IRB No. 20130313196EP), consisted of one orientation session and four main testing sessions. After

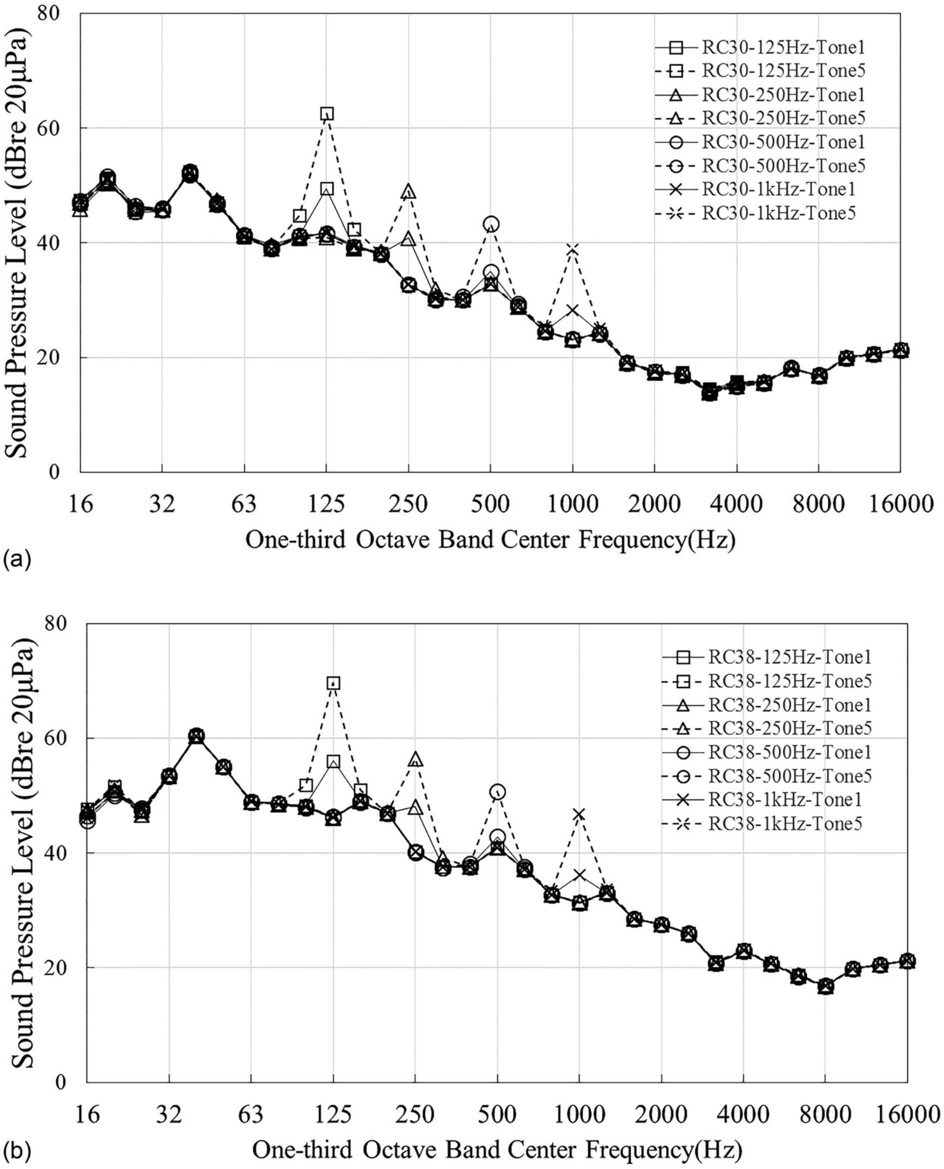

FIG. 3. One-third octave band spectra of (a) RC30 based noise signals and (b) RC38 based noise signals, for the lowest, middle, and highest tonal levels.

signing an informed consent form and completing the hearing threshold screening test during the orientation session, participants were informed about the general purpose of the study with annoyance defined as "individual adverse reaction such as dissatisfaction, distraction, bother and annoyance." Participants were informed that the context of measurement was as if they were working in an office environment. The participants also familiarized themselves with the main task by practicing it for 10 min at the end of the orientation session. They were asked to focus on accurately completing each task as quickly as possible.

The participants were next asked to attend four 30-min sessions, each of which included 24 test trials. Trials using only RC-30 neutral broadband noise were inserted between trials with tonal test signals to eliminate back-to-back comparisons of tonal test signals, so each 30-min session involved 12 tonal test signals and 12 neutral broadband noises only, except for the last session. The last session contained nine trials under the tonal test signals and 11 trials under the neutral noise. Each of the four main test sessions used the same neutral broadband noise condition across the entire 30-min session (RC-30N). The sessions started with the neutral broadband noise and ended with the tonal test signals. The presentation order of tonal test signals was balanced through a Latin-square design across all participants to avoid biasing the results due to signal order. For each trial, participants were asked to perform a digit span task in which they memorized a series of numbers in the reverse order of presentation while exposed to one of the test signals. The digit span task is a measure of short-term working memory commonly used in psychology experiments (Mølhave *et al.*, 1986; Jahanshahi *et al.*, 2008). Digit span tasks ranging from four digits and increasing up to eight digits (two at each digit level) were used over a duration of approximately a minute under one of the test signals. The exact time for the digit span task varied with participants. Conventionally, the digit span task ends when subjects fail to answer two consecutive questions correctly, but in this study, the maximum length was manually set to eight digits, regardless of participants' answers, to fix the duration time of each test stimulus.

A custom-written graphical user interface (GUI) in MATLAB controlled the presentation of all the trials and test signals; the program also measured the accuracy of answers and completion time of responses. At the end of each trial, the participants were asked to fill out a subjective questionnaire with two items indicating how annoyed they were by the test signal, and whether they would complain about the test signal to which they were exposed during the previous digit span tasks. No other information was provided regarding to whom they would complain or whether it was a single or recurring exposure. The annoyance question was answered on an 11-point continuous scale, while the complaint question was a dichotomous choice. Above the questionnaire in the MATLAB GUI was a statement advising participants not to consider their responses from any previous test signals they had heard or sessions they had attended.

III. RESULTS

A. Task performance

While not a primary goal of this research, two outcome variables related to performance were gathered and statistically analyzed to investigate the effects of tonal test signals on task performance: (1) the maximum number of correct digits achieved for a single digit span test trial and (2) the time it took for the participant to complete a single digit span test trial.

One-way repeated measure analysis of variance (ANOVA) was adopted to investigate the two performance measures across test signals. The first outcome variable associated with the maximum number of correct digits achieved for a single test trial did not show any statistically significant differences between test stimuli at all. However, ANOVA results showed that all completion times under tonal noise conditions were significantly longer than completion times under broadband noise conditions with RC-30N only [$F(40\,760) = 3.2$, $p < 0.001$, $\eta_p^2 = 0.14$]. Figure 4 illustrates

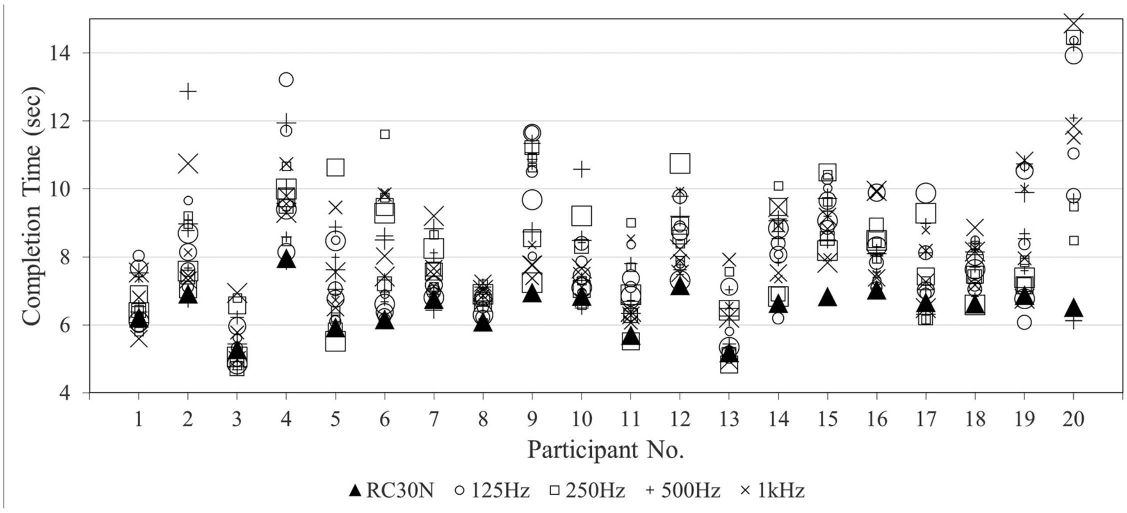

FIG. 4. Measured completion times for the digit span task under assorted tonal noise conditions above the RC-30 background noise across participants. The size of each marker corresponds to the tone level of each frequency, with larger markers indicating higher tone levels.

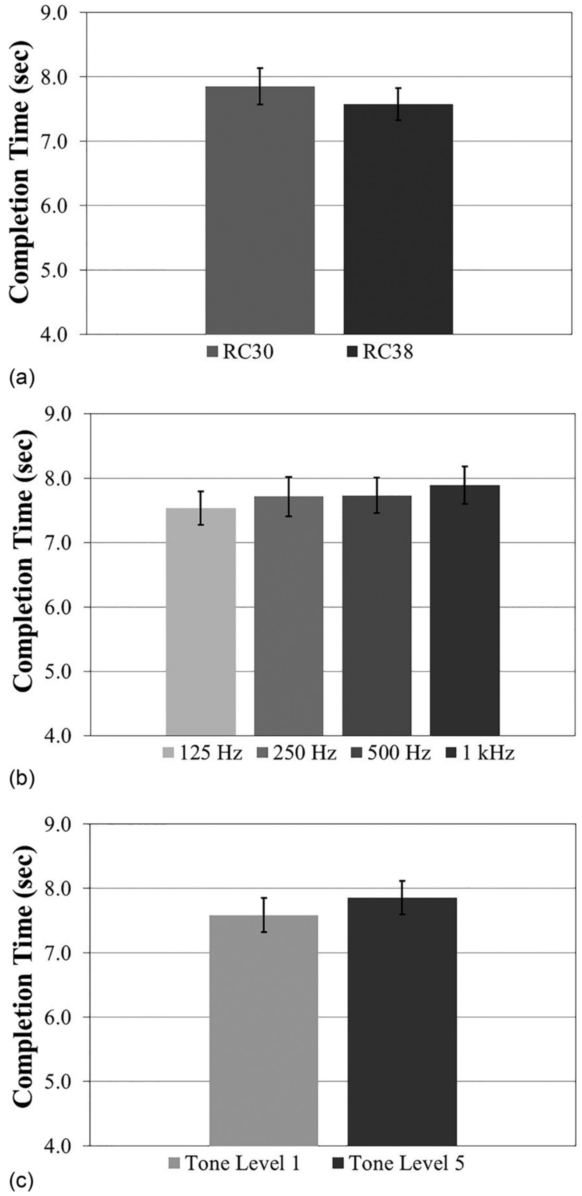

FIG. 5. Effect of (a) background noise level, (b) tone frequency, and (c) tone level on completion time. Only tone level was found to be statistically significant.

each individual completion time for each tonal noise condition over the RC-30 background level. Responses from signal conditions over the RC-38 noise level showed similar trends.

A three-way repeated-measures ANOVA was conducted to investigate the relationships between background noise level, tone frequency, and tone level on completion time across the 40 tonal signals. Figure 5 compares the completion times between the two different background noise levels, four different tone frequencies, and the least and most prominent tone levels tested.

Statistical analyses indicated that the effects of background noise level and tone frequency on completion time were not statistically significant, even though a trend of longer completion times with higher frequency tones was observed. The only significant factor found was tone level $[F(1,19) = 12.2,\ p = 0.002,\ \eta_p^2 = 0.4]$, with higher tone strength resulting in longer completion times. This result is noteworthy: higher tone strength can affect performance on a digit span task in terms of increasing the total amount of time taken to complete the task, but not its accuracy. This is in line with some other tonal noise studies that investigated performance and did not find significant results in terms of accuracy (Ryherd and Wang, 2010; Lee *et al.*, 2017); none of those studies looked at completion times, though. If subjects can maintain task accuracy under more strongly tonal conditions but at the expense of taking more time to do so, the result is still a significant loss in productivity.

B. Relationship between noise metrics and annoyance

To compare the relation between noise metrics and annoyance, Spearman's nonparametric correlation coefficients were calculated because the annoyance responses did not meet the normality assumption. Among the noise metrics previously introduced in Sec. I A, the following were chosen and calculated for each test stimuli: PR, TNR, and Tonal Audibility (ΔL_{ta}) for tonality metrics; un-weighted sound pressure level ($\mathrm{SPL_z}$), A-weighted sound pressure level ($\mathrm{SPL_a}$), ISO532-2:2017 (2017) Loudness (referred to as M-G Loudness in this paper), and ISO532-1:2017 (2017) Loudness (referred to as Zwicker Loudness in this paper) for loudness metrics; and tone-corrected Perceived Nose Level (PNLT), Joint Nordic Method (dBA+k), and Sound Quality Indicator (SQI) for combined metrics.

Four subjects' responses were excluded for all subsequent analyses since they submitted the same minimum annoyance rating across all signals; these four are shown as subjects 17 through 20 in Fig. 4. The results are analyzed in three groups separately: first with all signals included and then with each base background noise level (RC-30 or RC-38) separately. Table II presents the correlation coefficients for each analysis.

M-G Loudness shows the highest correlation coefficients with annoyance ratings across all signals. When separating signals into the two background noise levels, though, tonality metrics show on par or slightly higher correlation with annoyance than loudness metrics. Among tonality

TABLE II. Nonparametric Spearman correlation coefficients between noise metrics and annoyance (two-tailed, $^{**}p < 0.01$, $^{*}p < 0.05$).

	Tonality Metrics			Loudness Metrics				Combined Metrics		
	PR	TNR	ΔL_{ta}	$\mathrm{SPL_z}$ (dB)	$\mathrm{SPL_a}$ (dBA)	M-G Loudness	Zwicker Loudness	PNLT	dBA+k	SQI
All	0.105**	0.119**	0.157**	0.485**	0.539**	0.570**	0.557**	0.530**	0.532**	0.536**
RC-30N	0.169**	0.212**	0.246**	0.050	0.220**	0.246**	0.214**	0.207**	0.241**	0.215**
RC-38N	0.129*	0.179**	0.184**	0.062	0.124*	0.178**	0.149**	0.138*	0.133*	0.111*

TABLE III. Coefficients of the logistic regression model predicting whether a participant would (a) complain (95% BCa bootstrap confidence intervals based on 1000 samples), (b) be annoyed, or (c) be highly annoyed. CI = confidence interval.

% Complaint

		95% CI for Odds Ratio		
	b	Lower	Odds	Upper
Constant	−21.11 (−25.43, −17.75)			
M-G Loudness (phon)*	0.30 (0.25,0.36)	1.28	1.35	1.42
ΔL_{ta} (dB)**	0.07 (0.03,0.12)	1.03	1.07	1.12

Note. $R^2 = 0.25$, $\chi^2(2) = 209.72$, $p < 0.001$. * $p < 0.001$. ** $p < 0.001$.

% Annoyed

		95% CI for Odds Ratio		
	b	Lower	Odds	Upper
Constant	−21.29 (−26.01, −18.04)			
M-G Loudness (phon)*	0.30 (0.25, 0.37)	1.27	1.35	1.44
ΔL_{ta} (dB)**	0.02 (−0.02, 0.07)	0.98	1.02	1.06

Note. $R^2 = 0.19$, $\chi^2(2) = 144.74$, $p < 0.001$. * $p < 0.001$. ** $p = 0.29$.

% Highly Annoyed

		95% CI for Odds Ratio		
	b	Lower	Odds	Upper
Constant	−23.84 (−50.82, −15.65)			
M-G Loudness (phon)*	0.29 (0.16, 0.69)	1.14	1.34	1.56
ΔL_{ta} (dB)**	0.12 (0.02 0.22)	1.03	1.13	1.23

Note. $R^2 = 0.06$, $\chi^2(2) = 44.62$, $p < 0.001$. * $p < 0.001$. ** $p = 0.56$.

metrics, Tonal Audibility demonstrates slightly better correlation than TNR and PR for all analyses. The results indicate that loudness is the most important feature of noise to predict annoyance, but also tonality of noise should be included for the annoyance model, especially when background noise levels are kept constant. Combined metrics such as the dBA+k, PNLT, and SQI did not show better performance than loudness metrics, even though they were significantly related to annoyance ratings. The results show that imposing set penalty values to loudness levels may not be the most effective way to quantify overall annoyance of the noise.

C. Model to predict percentage of complaints and those annoyed by tones in noise

A model has been developed from the gathered likelihood of complaints to determine thresholds of acceptability for tonality, using a multiple logistic regression model. Based on the correlation analysis in Sec. III B, M-G Loudness and Tonal Audibility are chosen as the two prediction variables for the regression model. Three logistic regression models are then constructed: one predicting the percentage of subjects who indicated they would complain to the noise condition under test (%Complaint), one predicting the percentage of subjects considered to be annoyed (%A), and one predicting the percentage of subjects considered to be highly annoyed (%HA). The break-points to convert the continuous scale data to the categorical data were set to 5.0 and 7.2 on an 11-point scale (from 0 to 10) for the percentage of annoyed and highly annoyed persons, respectively, as has been done in previous work (Pedersen, 2007).

Table III presents coefficient values and statistics for all three models. The chi-square (χ^2) value indicates how much each model prediction is improved against the model with no predictor. The R^2 of the logistic regression, which is similar to the R^2 of linear regression, is a measure of how well the prediction model fits the response data. It can be calculated with the chi-square and maximum log-likelihood values. Several methods are proposed to calculate the R^2 for the logistic regression and, in this paper, the R^2 from Cox and Snell (1989) is calculated. The odds ratio is the exponential of the coefficient of the model. The ratio indicates how the "odds" of the outcome occurrence will change with a unit of predictor change. A ratio greater than one indicates a positive relation between the predictor and the odds of the outcome.

All models of %Complaint, %A, and %HA are statistically significant ($p < 0.001$) and the M-G Loudness significantly improved the model fit to the model of %Complaint based on chi-square statistics. However, the Tonal Audibility predictor significantly improved the model fit in the %Complaint model ($p < 0.001$), whereas the Tonal Audibility is not the significant predictor for %A or % HA models ($p = 0.29$ for %A and 0.56 for %HA).

Specifically, the logistic regression equation for %Complaint can be expressed as

$$\%\,Complaint = \frac{1}{1 + e^{-21.11-0.30[Loudness]-0.07[\Delta L_{ta}]}}, \tag{2}$$

where %Complaint is the percentage of complaints expected to be lodged against a particular tonal signal condition. This %Complaint model yields a chi-square ($\chi 2$) of 209.72, which is highly significant ($p < 0.001$). The accuracy of the model's prediction against observed responses is 76.3%. Recall that participants were not explicitly told to whom they would complain or whether it was due to a single or recurring exposure of the test signal; this logistic regression model, then, can represent a worst-case scenario, resulting in a complaint due to a single exposure.

Figure 6 illustrates the logistic regression lines with actual responses. The result shows that the %Complaint model is more similar to the %A model rather than the %HA. The %Complaint model also showed better performance with regards to chi-square statistics and confidence intervals. Current guidelines suggest dividing the continuous scale into certain breakpoints for the %A or %HA logistic regression models. However, the results from this study show that these noise-exposure models show lower chi-square statistics and wider confidence intervals. One reason for this may be that subjects may still feel confused about the meaning of the term "annoyance," even though they are informed about it and provided a definition of annoyance in the orientation session. The question of whether they are going to complain or not may feel easier for the subjects to answer because it is a more behaviorally-based question. Another reason is that setting the breakpoint at 72 (or 50)

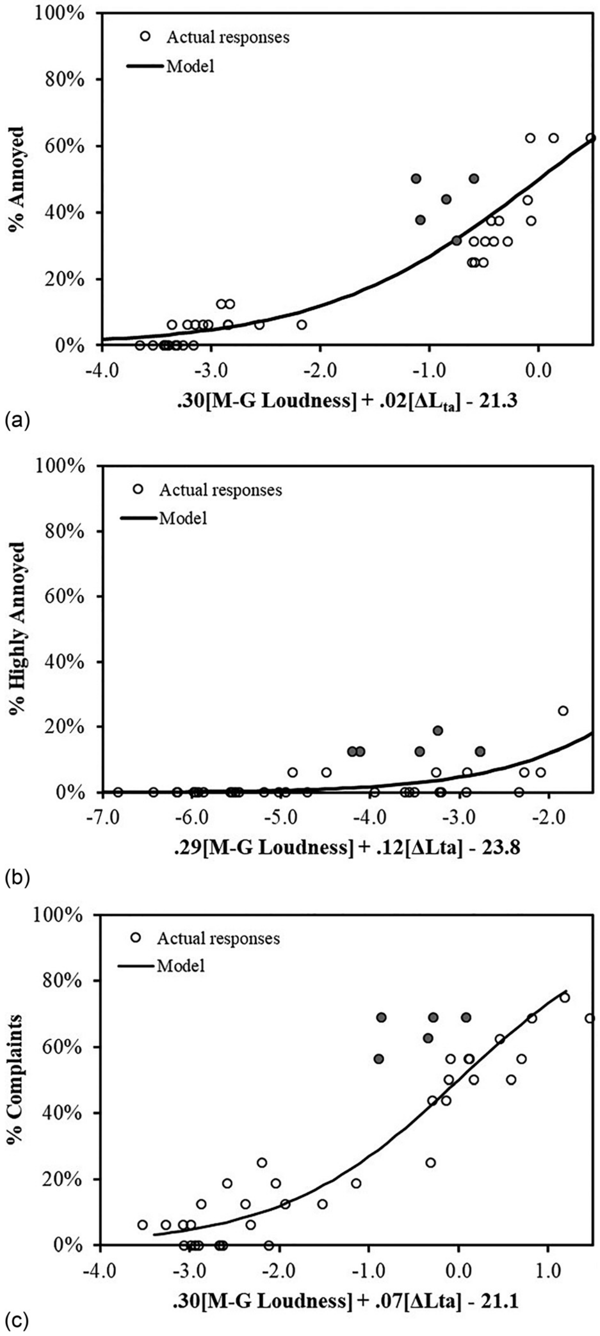

FIG. 6. Logistic regression models (as given in Table III) of percentage of persons (a) annoyed, (b) highly annoyed, or (c) who would submit complaints due to a given noise condition with certain Moore-Glasberg Loudness and Tonal Audibility. The filled markers are response rates for five actual noise recordings.

points and over implies a very distinct difference for responses near the breakpoint; 73 points will be counted as highly annoyed and 71 points will be counted as annoyed, even though the actual responses are close. Thus, the results suggested that a predictive model based on %Complaint, rather than %A or % HA, is recommended over the others. However, the %Complaint model still showed the wide range of confidence intervals for each metric due to the small sample size despite their statistical significance. With utilizing more test signals and participants, a more accurate dose-response model can be proposed. Additionally, the annoyance responses were answered on an 11-point continuous scale and comparison between annoyance responses using a 5-point verbal scale and %Complaint should be investigated in a future study.

Tenfold cross-validation analysis was utilized to estimate the prediction performance of the models to the general population and compare the performance of each model by dividing the data randomly in ten subsets (Friedman *et al.*, 2001). The analysis cannot be performed with the %HA model due to the lack of variability. The average accuracies are 76.5% with a SD of 3% for %Complaint and 78.3% (SD = 3%) for %A. Figure 7 shows receiving operating characteristic (ROC) curves of the two %Complaint and %A models. The axes of the graph are related to type I (false positive) and type II (false negative) errors, respectively. The box plots show variances of the curve values by ten-folding cross validation. Ideally, the curve should approach the top left corner steeply and have small variations for more accurate classification. The area under the curve (AUC) values were 0.82 (SD = 0.03) for %Complaint and 0.78 (SD = 0.04) for %A. The variance of the error and steepness of the curve indicates that %Complaint is a better model than %A in terms of reliability.

To suggest allowable tonality limits, the points at which 30, 40, or 50% of participants would complain can be determined from the logistic regression model to determine maximum Tonal Audibility for a given M-G Loudness in phons (Fig. 8). The authors do not intend to prescribe that these percentages should be of more or less utility than other percentages, but instead present them as a starting point. Fidell *et al.* (2011) have defined community tolerance level (CTL) as being the level at which 50% of a community is highly annoyed, so there is some precedent in using similar percentages. The criteria lines in the figure demonstrate that the thresholds of acceptable tonality decrease as overall background noise level increases. The results indicate that lower levels of tonal components may not be acceptable when the overall background noise is louder.

It is noteworthy, though, that the statistically significant %Complaint model presented in Table III still shows a wide confidence interval range for each metric (e.g., from 1.03 to 1.12 for odd ratio of tonal audibility change and 1.28 to 1.42 for odd ratios of loudness), due to the smaller sample size in this investigation. Future testing utilizing more test signals and participants is recommended so that the prediction models and suggested allowable tonality limits for a given loudness level presented above can be refined further.

IV. CONCLUSIONS

A subjective study on how tonal noise conditions, such as those produced by building mechanical systems, can impact participant performance and annoyance has been presented. Results reveal that while there was no statistically significant effect on accuracy, even the least prominent tonal signals increased the time it took for participants to complete the digit span task compared to test conditions with broadband noise alone. Additionally, the level of tone was found to have a statistically significant effect on the performance

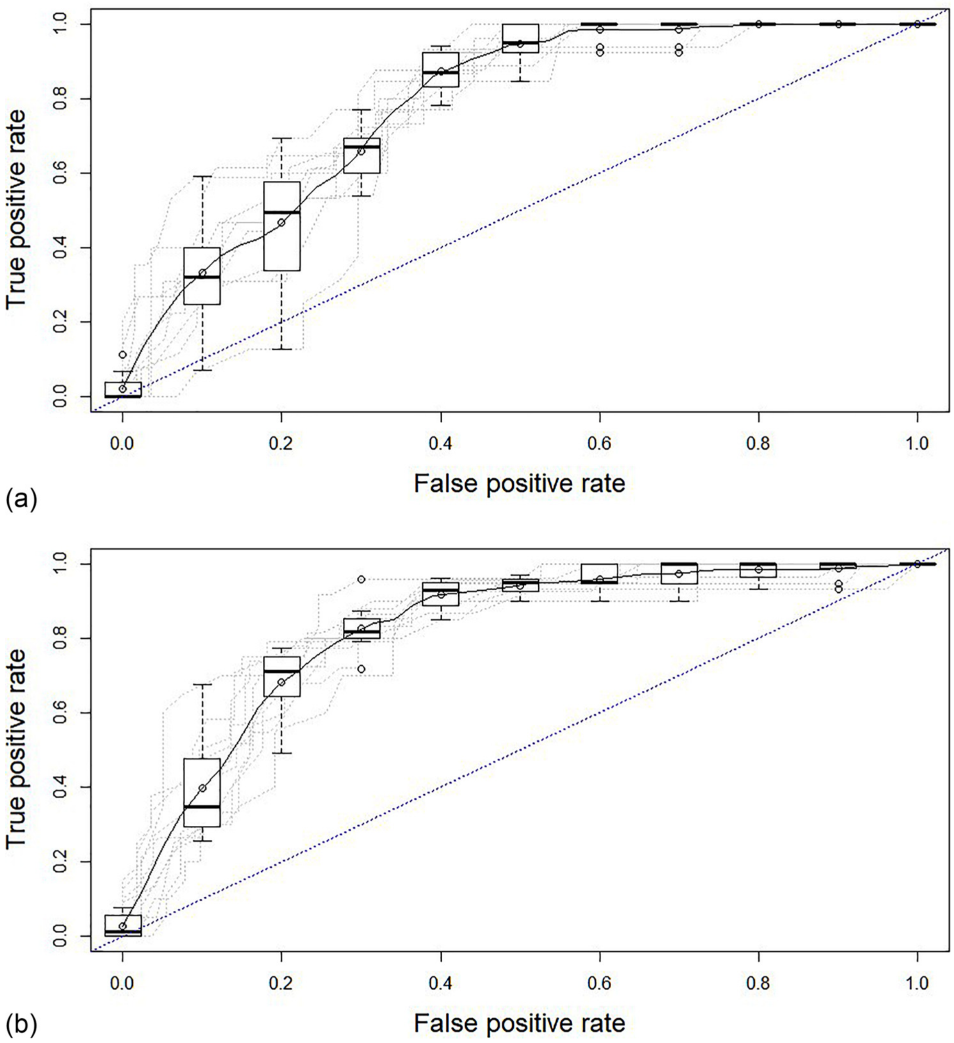

FIG. 7. (Color online) ROC curves with ten-folded subsets of the participant data for (a) %Annoyed and (b) %Complaint due to a given noise condition with certain Moore-Glasberg Loudness and Tonal Audibility. The connected dots and solid lines inside the box plots represent average and median values across ten subsets, respectively, and dots outside the box plots show outliers.

metric of completion time, with higher tonal levels causing subjects to take longer to complete the task. A louder background noise level (RC-38 versus RC-30) and varying tone frequencies (from 125 Hz to 1 kHz) did not. More comprehensive testing is suggested to generalize the findings on how tonal levels can affect human performance or short-term memory.

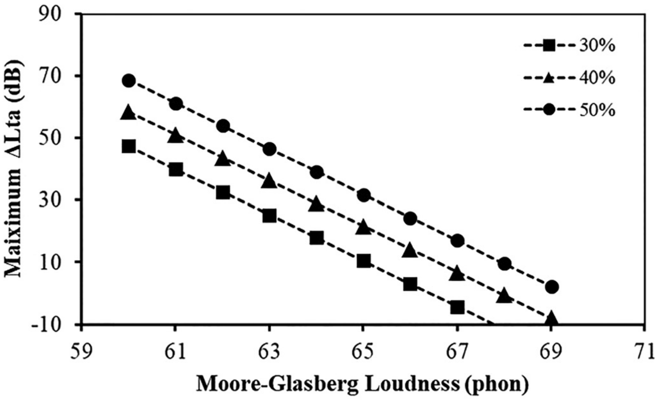

FIG. 8. Maximum allowable Tonal Audibility criteria for a given Moore-Glasberg Loudness (in phons). The results shown correspond to 30%, 40%, or 50% of subjects complaining, but other percentages may also be considered.

Based on the annoyance responses and likelihood to complain, predictive logistic regression models have been developed. The reliability of the models depends on the selected noise metrics, which should correlate strongly to the perception of the noise. Based on correlational analyses, the loudness metric M-G Loudness showed the highest correlation overall to the annoyance responses, while the tonality metric Tonal Audibility also demonstrated significant correlation with the annoyance. Thus, these two noise metrics for loudness and tonality, respectively, were chosen to develop binary logistic regression models related to %Complaint, %A, and %HA responses. The %Complaint model fits the actual responses best and has the least wide confidence interval among the models, suggesting that similar studies in the future should focus on asking about the likelihood of subjects to complain due to a noise condition, rather than asking subjects to rate their annoyance. The %Complaint regression model is subsequently used to suggest maximum allowable tonality limits for a given M-G Loudness in phons. Future work is recommended with an increased number of tonal test signals and participants to validate these findings further. In particular, tones in the low frequency range below 125 Hz were not investigated in this study and are suggested for future research. Manufacturers of HVAC systems are now commonly publishing sound data down to at least the 63 Hz octave band, as HVAC equipment can produce significant tones in that frequency range. At sufficiently high levels, tones at frequencies lower than 125 Hz may impact the occupants' likelihood to complain, and these should be incorporated into future models.

AHRI/ANSI 1140-2012 (**2012**). *Sound Quality Evaluation Procedures for Air-Conditioning and Refrigeration Equipment* (Air-Conditioning, Heating, and Refrigeration Institute, Arlington, VA).

ANSI/ASA S12.10-2010/Part 1 (**2010**). *Acoustics—Measurement of Airborne Noise Emitted by Information Technology and Telecommunications Equipment—Part 1: Determination of Sound Power Level and Emission Sound Pressure Level* (American National Standards Institute, New York).

Aures, W. (**1985**). "Procedure for calculating the sensory euphony of arbitrary sound signals," Acustica **59**, 130–141.

Beranek, L. L. (**1957**). "Revised criteria for noise in buildings," Noise Control **3**, 19–27.

Berglund, B., Hassmén, P., and Preis, A. (**2002**). "Annoyance and spectral contrast are cues for similarity and preference of sounds," J. Sound Vib. **250**, 53–64.

Blazier, W. E. (**1981**). "Revised noise criteria for application in the acoustical design and rating of HVAC systems," Noise Control Eng. **16**, 64–73.

Cox, D. R., and Snell, E. J. (**1989**). *Analysis of Binary Data* (CRC Press, Boca Raton, FL), Vol. 32.

Fidell, S., Barber, D. S., and Schultz, T. J. (**1991**). "Updating a dosage–effect relationship for the prevalence of annoyance due to general transportation noise," J. Acoust. Soc. Am. **89**, 221–233.

Fidell, S., Mestre, V., Schomer, P., Berry, B., Gjestland, T., Vallet, M., and Reid, T. (**2011**). "A first-principles model for estimating the prevalence of annoyance with aircraft noise exposure," J. Acoust. Soc. Am. **130**, 791–806.

Field, A. (**2013**). *Discovering Statistics Using IBM SPSS Statistics* (SAGE Publications, London), Chap. 8, pp. 264–315.

Francis, J. M., Lee, J., and Wang, L. M. (**2014**). "SE-14-C043: Determining annoyance thresholds of tones in noise," ASHRAE Tran. **120**, 1–8.

Friedman, J., Hastie, T., and Tibshirani, R. (**2001**). *The Elements of Statistical Learning* (Springer, New York).

Glasberg, B. R., and Moore, B. C. (**2006**). "Prediction of absolute thresholds and equal-loudness contours using a modified loudness model," J. Acoust. Soc. Am. **120**, 585–588.

Guski, R., Felscher-Suhr, U., and Schuemer, R. (**1999**). "The concept of noise annoyance: How international experts see it," J. Sound Vib. **223**, 513–527.

Hansen, H., Jesko, L. V., and Reinhard, W. (**2011**). "The magnitude of tonal content. A review," Acta Acust. united Acust. **97**, 355–363.

Hastings, A., Lee, K. H., Davies, P., and Surprenant, A. M. (**2003**). "Measurement of the attributes of complex tonal components commonly found in product sound," Noise Control Eng. J. **51**, 195–209.

Hellman, R. P. (**1985**). "Perceived magnitude of two-tone-noise complexes: Loudness, annoyance, and noisiness," J. Acoust. Soc. Am. **77**, 1497–1504.

ISO 532-1:2017 (**2017**). *Acoustics—Method for Calculating Loudness—Part 1: Zwicker Method* (International Organization for Standardization, Geneva, Switzerland).

ISO 532-2:2017 (**2017**). *Acoustics—Method for Calculating Loudness—Part 2: Moore-Glasberg Method* (International Organization for Standardization, Geneva, Switzerland).

ISO/TS 15666:2003 (**2003**). *Acoustics—Assessment of Noise Annoyance by Means of Social and Socio-Acoustic Surveys* (International Organization for Standardization, Geneva, Switzerland).

ISO 1996-2:2007 (**2007**). *Acoustics—Description, Measurement and Assessment of Environmental Noise—Part 2: Determination of Environmental Noise Levels* (International Organization for Standardization, Geneva, Switzerland).

Jahanshahi, M., Saleem, T., Ho, A. K., Fuller, R., and Dirnberger, G. (**2008**). "A preliminary investigation of the running digit span as a test of working memory," Behav. Neurol. **20**, 17–25.

Janssen, S. A., Vos, H., Eisses, A. R., and Pedersen, E. (**2011**). "A comparison between exposure-response relationships for wind turbine annoyance and annoyance due to other noise sources," J. Acoust. Soc. Am. **130**, 3746–3753.

Kryter, K. D. (**1960**). "The meaning and measurement of perceived noise level," Noise Control **6**, 12–27.

Kryter, K. D. (**1982**). "Community annoyance from aircraft and ground vehicle noise," J. Acoust. Soc. Am. **72**, 1222–1242.

Kryter, K. D., and Pearsons, K. S. (**1965**). "Judged noisiness of a band of random noise containing an audible pure tone," J. Acoust. Soc. Am. **38**, 106–112.

Landström, U., Åkerlund, E., Kjellberg, A., and Tesarz, M. (**1995**). "Exposure levels, tonal components, and noise annoyance in working environments," Environ. Int. **21**, 265–275.

Lee, J., Francis, J. M., and Wang, L. M. (**2017**). "How tonality and loudness of noise relate to annoyance and task performance," Noise Control Eng. J. **65**, 71–82.

Little, J. W. (**1961**). "Human response to jet engine noises," Noise Control **7**, 11–13.

Marquis-Favre, C., Premat, E., and Aubrée, D. (**2005**). "Noise and its effects—A review on qualitative aspects of sound Part II: Noise and annoyance," Acta Acust. united Acust. **91**, 626–642.

Miedema, H. M., and Vos, H. (**1998**). "Exposure-response relationships for transportation noise," J. Acoust. Soc. Am. **104**, 3432–3445.

Miedema, H. M., and Oudshoorn, C. (**2002**). *Position Paper on Dose Response Relationships Between Transportation Noise and Annoyance. EU's Future Noise Policy, WG2–Dose/Effect* (European Communities, Luxembourg).

Mølhave, L., Bach, B., and Pedersen, O. F. (**1986**). "Human reactions to low concentrations of volatile organic compounds," Environ. Int. **12**, 167–175.

More, S., and Davies, P. (**2010**). "Human responses to the tonalness of aircraft noise," Noise Control Eng. J. **58**, 420–440.

Pedersen, E., van den Berg, F., Bakker, R., and Bouma, J. (**2009**). "Response to noise from modern wind farms in The Netherlands," J. Acoust. Soc. Am. **126**, 634–643.

Pedersen, T. H. (**2007**). *The "Genlyd" Noise Annoyance Model: Dose-Response Relationships Modelled by Logistic Functions* (DELTA Acoustics & Vibration, Hørsholm, Denmark).

Pedersen, T. H., Søndergaard, M., and Andersen, B. (**2000**). *Objective Method for Assessing the Audibility of Tones in Noise Joint Nordic Method—Version* 2 (DELTA Acoustics & Vibration, Hørsholm, Denmark).

Ryherd, E. E., and Wang, L. M. (**2008**). "Implications of human performance and perception under tonal noise conditions on indoor noise criteria," J. Acoust. Soc. Am. **124**, 218–226.

Ryherd, E. E., and Wang, L. M. (**2010**). "AB-10-018: The effects of noise from building mechanical systems with tonal components on human performance and perception (1322-RP)," ASHRAE Tran. **116**, 541–552.

Sailer, U., and Hassenzahl, M. (**2000**). "Assessing noise annoyance: An improvement-oriented approach," Ergonomics **43**, 1920–1938.

Schomer, P. D. (**2001**). *A White Paper: Assessment of Noise Annoyance* (Schomer and Associates, Champaign, IL).

Schomer, P. D. (**2005**). "Criteria for assessment of noise annoyance," Noise Control Eng. J. **53**, 125–137.

Schultz, T. J. (**1978**). "Synthesis of social surveys on noise annoyance," J. Acoust. Soc. Am. **64**, 377–405.

Schutte, M., Marks, A., Wenning, E., and Griefahn, B. (**2007**). "The development of the noise sensitivity questionnaire," Noise Heal. **9**, 15–24.

Susini, P., McAdams, S., Winsberg, S., Perry, I., Vieillard, S., and Rodet, X. (**2004**). "Characterizing the sound quality of air-conditioning noise," Appl. Acoust. **65**, 763–790.

Tang, S. K. (**1997**). "A distribution function applicable to office noise study," J. Sound Vib. **208**, 603–615.

World Health Organization (**2011**). *Burden of Disease from Environmental Noise: Quantification of Healthy Life Years Lost in Europe* (WHO Regional Office for Europe, Copenhagen, Denmark).

Zwicker, E. (**1961**). "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," J. Acoust. Soc. Am. **33**, 248.