HHS Public Access

Author manuscript

J Am Acad Audiol. Author manuscript; available in PMC 2017 January 25.

Published in final edited form as:

J Am Acad Audiol. 2016 October ; 27(9): 732–749. doi:10.3766/jaaa.15099.

Smartphone-Based System for Learning and Inferring Hearing Aid Settings

Gabriel Aldaz[*], **Sunil Puria**[†], and **Larry J. Leifer**[*]

[*]Department of Mechanical Engineering, Center for Design Research, Stanford University, Stanford, CA

[†]Mechanics and Computation Division, Department of Mechanical Engineering, Stanford University, Stanford, CA

Abstract

Background—Previous research has shown that hearing aid wearers can successfully self-train their instruments' gain-frequency response and compression parameters in everyday situations. Combining hearing aids with a smartphone introduces additional computing power, memory, and a graphical user interface that may enable greater setting personalization. To explore the benefits of self-training with a smartphone-based hearing system, a parameter space was chosen with four possible combinations of microphone mode (omnidirectional and directional) and noise reduction state (active and off). The baseline for comparison was the "untrained system," that is, the manufacturer's algorithm for automatically selecting microphone mode and noise reduction state based on acoustic environment. The "trained system" first learned each individual's preferences, self-entered via a smartphone in real-world situations, to build a trained model. The system then predicted the optimal setting (among available choices) using an inference engine, which considered the trained model and current context (e.g., sound environment, location, and time).

Purpose—To develop a smartphone-based prototype hearing system that can be trained to learn preferred user settings. Determine whether user study participants showed a preference for trained over untrained system settings.

Research Design—An experimental within-participants study. Participants used a prototype hearing system—comprising two hearing aids, Android smartphone, and body-worn gateway device—for ~6 weeks.

Study Sample—Sixteen adults with mild-to-moderate sensorineural hearing loss (HL) (ten males, six females; mean age = 55.5 yr). Fifteen had ≥6 mo of experience wearing hearing aids, and 14 had previous experience using smartphones.

Intervention—Participants were fitted and instructed to perform daily comparisons of settings ("listening evaluations") through a smartphone-based software application called Hearing Aid Learning and Inference Controller (HALIC). In the four-week-long training phase, HALIC

Corresponding author: Gabriel Aldaz, Department of Mechanical Engineering, Center for Design Research, Stanford University, Stanford, CA 94305-2232; zamfir@stanfordalumni.org.

Portions of this article were presented orally at the American Auditory Society conference in Scottsdale, AZ, March 2014, and the International Hearing Aid Research Conference in Lake Tahoe, CA, August 2014.

recorded individual listening preferences along with sensor data from the smartphone—including environmental sound classification, sound level, and location—to build trained models. In the subsequent two-week-long validation phase, participants performed blinded listening evaluations comparing settings predicted by the trained system ("trained settings") to those suggested by the hearing aids' untrained system ("untrained settings").

Data Collection and Analysis—We analyzed data collected on the smartphone and hearing aids during the study. We also obtained audiometric and demographic information.

Results—Overall, the 15 participants with valid data significantly preferred trained settings to untrained settings (paired-samples *t* test). Seven participants had a significant preference for trained settings, while one had a significant preference for untrained settings (binomial test). The remaining seven participants had nonsignificant preferences. Pooling data across participants, the proportion of times that each setting was chosen in a given environmental sound class was on average very similar. However, breaking down the data by participant revealed strong and idiosyncratic individual preferences. Fourteen participants reported positive feelings of clarity, competence, and mastery when training via HALIC.

Conclusions—The obtained data, as well as subjective participant feedback, indicate that smartphones could become viable tools to train hearing aids. Individuals who are tech savvy and have milder HL seem well suited to take advantages of the benefits offered by training with a smartphone.

Keywords

environment classification; hearing aids; smartphone; trainable

INTRODUCTION

In the 1990s, the advent of digital signal processing enabled complex environmentally adaptive technology such as dynamic noise cancellation, adaptive directionality, speech detection algorithms, and multiband compression. Proper adjustment of the large number of amplification parameters available in modern digital hearing aids offers the potential to improve the sound quality experienced by hearing aid users. With this expanded array of amplification parameters comes a corresponding increase in the complexity of the fitting procedure, which may actually hinder hearing-care professionals from prescribing the parameter settings that would be optimal for an individual user. For decades, the fitting procedure has fundamentally consisted of matching a set of measured characteristics (e.g., an audiogram) to a prescriptive fitting rationale. Fitting rationales, derived from a combination of theoretical knowledge and experimental data, are best suited to an idealized "average" person. Keidser and Dillon (2006) reported that only 49% of participants had preferred gain within ±3 dB of the National Acoustic Laboratories-nonlinear 1 (NAL-NL1) prescriptive targets, and 60% had preferred gain within ±3 dB of the National Acoustic Laboratories-nonlinear 2 (NAL-NL2) targets. Furthermore, the clinic where the fitting takes place is usually a low-noise, low-reverberation environment, and the limited range of acoustic stimuli in such a setting is unlikely to lead to an optimal fitting suitable for the real-world environment. Hearing aid users commonly require multiple office visits to fine-tune

the parameters on their devices. Therefore, the audiogram and fitting rationale are a good first step, but only partially address the fact that hearing loss (HL) is an individual experience.

Since individual users will encounter a variety of acoustic scenes during the day, their hearing instruments offer adaptation to changing listening situations. Device adaptation may be implemented either through user control—for example, using multiple programs (Ringdahl et al, 1990)—or automatically (Allegro et al, 2001). Neither strategy has proven completely successful. Nelson et al (2006) found that patients did not regularly and consistently use multiple programs, preferring to spend most of their time in the default program.

Given the limitations of multiple programs, researchers began exploring alternate methods for the customization of hearing aid parameters for the individual user. Dillon et al (2006) described a trainable hearing aid featuring a learning algorithm that promised several potential advantages over traditional hearing aids, including personalization of parameters in actual listening environments, fewer postfitting visits for fine-tuning, and greater user involvement in the fitting process, resulting in increased sense of ownership and engagement. The first commercial hearing aid with some learning capacity, the Siemens Centra, was equipped with algorithms that used information about the user's volume control selection patterns to gradually adjust the default volume setting (Chalupper, 2006; Hayes, 2007). In a real-world experiment using Centras, Chalupper (2006) allowed 19 experienced hearing aid wearers to train the volume setting in different environments. The data showed that participants preferred different amounts of overall gain relative to the prescribed gain, and some individuals preferred different volume settings in different environments. In a follow-up cross over study with Centras, Mueller et al (2008) fitted experienced hearing aid wearers with volume control start-up either >6 dB or <6 dB NAL-NL1 prescriptive targets. Participants trained the overall gain, in real-world situations, to their preferred level. For most participants, the trained preferred gain differed significantly from the NAL-NL1 targets, and the starting level influenced the final preferred gain. In a related experiment conducted in a laboratory environment, Dreschler et al (2008) found that patient-based hearing aid adjustments were a reliable method of individual fine-tuning and provided systematic and reproducible gain-frequency response preferences in different acoustic environments. Nonetheless, the researchers also observed that the initial settings influenced the final preferred settings.

While earlier studies by Chalupper (2006) and Mueller et al (2008) trained only overall gain, researchers began applying wide dynamic range compression learning, in which gain was trained as a function of the level of the input signal. Zakis (2003) implemented a digital research hearing aid called Stereo Hearing Aid Research Processor (SHARP). The SHARP, comprising a body-worn processor unit (90 × 60 × 20 mm) connected via cables to two behind-the-ear hearing aids, featured three user controls on the processor unit: a rotary control, a voting button, and a power/mode switch. Zakis et al (2007) used the SHARP to investigate whether hearing aid users in everyday listening situations preferred trained amplification parameters to a set of untrained parameters prescribed in a clinic. The SHARP's three user controls acted as a program selection switch, a voting button (used to

indicate preferred settings), and a rotary gain-frequency-response adjustment control simultaneously for both ears. In Trial I (one to four weeks), 18 participants were fitted with untrained settings based on the NAL-NL1 prescription, and then attempted to train the amplification settings to their preference in everyday situations (with a minimum of 300 votes). In Trial II (one week), 13 participants blindly compared their trained settings (with noise suppression enabled) to untrained settings (without noise suppression) and voted for their preferred settings in everyday situations. Trial III (one week), with eight participants, was a repeat of Trial II, but with noise suppression disabled for both the trained and untrained settings. Most notably, results for Trial II showed that nine (69%) participants voted for the trained settings significantly more often than the untrained settings in real-life environments, three (23%) participants had nonsignificant preferences, and one (8%) participant had a significant preference for the untrained settings. Results for Trials II and III were not significantly different for seven of the eight participants who participated in both trials. Data logged during the trials showed that the percentage of votes for the trained settings was moderately but not significantly correlated with the hours of aid use during the training period. One limitation of this study was that participants did not train different settings in different environmental sound classes (ESCs), although the SHARP prototype had the capability to do so.

Building on these encouraging results, more sophisticated algorithms have been introduced. Siemens SoundLearning was capable of training the gain-frequency response and compression characteristics by relating preferred gain settings to the input level in four frequency bands (Chalupper et al, 2009). SoundLearning 2.0 added learning in three different ESCs—Speech, Noise, and Music—as well as mixtures of these environments (Powers and Chalupper, 2010). Another commercially available algorithm trained the comfort–clarity balance (Unitron, 2011). Keidser and Alamudi (2013) reported on a study investigating the efficacy and reliability of training commercially available learning hearing aids in everyday environments. The test devices included a modified SoundLearning algorithm with the capability to discriminate between six ESCs (Quiet [QUI], Noise [NOI], Music [MUS], Speech in Quiet [SiQ], Speech in Noise [SiN], and Car Noise) and learn gain-frequency response shape and compression parameters. Among the 18 participants with valid data, 8 (44.4%) preferred the trained response, 8 (44.4%) showed no preference, and 2 (11.1%) preferred the untrained response.

Although trainable hearing aids have been commercially available for years, several important shortcomings inherent in present hearing aids hinder their potential for learning user preferences: (a) a hearing aid has limited sensor inputs, relying entirely on two onboard microphones to collect information about incoming sounds; (b) a hearing aid has a restricted user interface, even if a remote control is available; (c) compared to other computing devices, a hearing aid has reduced processing power, which prevents the implementation of more advanced machine learning algorithms.

To some extent, these limitations may be overcome by regarding the smartphone as part of an intelligent hearing system. Smartphones provide a powerful mobile computing platform, with sensing, processing, communication, and memory capabilities that can add to the processing capabilities of the hearing device and provide more audiological benefits than

previously possible. Smartphone adoption is exploding worldwide. Pew Research estimated that in the United States, smartphone ownership by those aged >65 jumped from 11% to 18% between 2011 and 2013 (Smith, 2012; 2013). Nielsen (2013) estimated that smartphone ownership among Americans aged >55 doubled between 2012 and 2013. Furthermore, the United States lags behind countries such as South Korea, the United Kingdom, and Germany in ownership and access to smartphones (Barker et al, 2013). Recent advances in wireless technology, such as the introduction of Bluetooth 4.0 (Low Energy), promise direct, two-way communication between smartphone and hearing aids, thus eliminating the need for a body-worn gateway device. A number of other smartphone software applications related to hearing have appeared on the market. Most relevant is Ear Machine (Chicago, IL), which allows users to select desired settings by adjusting two controllers, labeled "loudness" and "tone" (Nelson et al, 2014). Researchers have even explored the possibility of using smartphone-based applications as alternatives to traditional hearing aids as a temporary or starter solution for people with a HL (Amlani et al, 2013).

Building on previous research, this work makes two primary contributions. First, we describe the development of a smartphone-based software henceforth referred to as the Hearing Aid Learning and Inference Controller (HALIC). Second, we report on the results of a user study addressing the research question, "Do listeners show a preference for microphone mode and noise reduction settings predicted by their trained system over those predicted by the manufacturer's untrained system?"

The present version of HALIC (Figure 1) takes advantages of a smartphone's built-in sensors, user interface, and computing power. HALIC leverages smartphone sensors to gather information about the user's context. In the field of computer science, context awareness (Schilit et al, 1994) takes into account factors that change on a timescale measured in minutes, hours, or days—such as the user's sound environment, location, or day of week. Thus, context awareness provides an additional layer of information to augment a hearing instrument's real-time processing. HALIC also includes an intuitive user interface, permitting individual users to express their preferences in particular listening situations. Lastly, HALIC employs the smartphone's computing power to run computation-intensive digital signal processing and machine learning algorithms that select the best hearing aid settings (among those in the parameter space) over time.

To answer our research question, we devised a user study in which participants provided subjective feedback in everyday listening environments. Since it was not desirable or realistic for participants to continually communicate their preferred hearing aid settings, a more reasonable approach was to have participants label their preferred hearing aid settings at specific times ("our goal was 8–12 times per day"). We called the mechanism for having users input their preferences via the smartphone app the "listening evaluation." Our objective was to collect as much data from each listening evaluation as possible, while limiting the average duration of an evaluation to 2 min.

In the training phase of the user study, lasting ~4 weeks, HALIC built "trained models" based on individual listening evaluation preferences. During the next two weeks, the validation phase, participants performed blinded listening evaluations comparing settings

predicted by the trained system ("trained settings') to those suggested by the hearing aids' untrained system ("untrained settings"). We demonstrated the benefits of such a hearing system regarding the selection of multiple settings (microphone directionality and noise reduction) intended to enhance speech, suppress transient noise, or perform other useful processing operations. The baseline for comparison was the "untrained system," which is the manufacturer's algorithm for automatically selecting microphone mode and noise reduction state based on acoustic environment.

MATERIALS AND METHODS

The hearing system's off-the-shelf hardware comprises a pair of modern Oticon (Smørum, Denmark) Intiga 10 hearing instruments, an Android-based Google (Mountain View, CA) Nexus Galaxy smartphone, and an Oticon Streamer, a body-worn gateway device with a built-in microphone that wirelessly links the hearing aids to a mobile phone. The gateway serves two primary purposes: it streams sound from a mobile phone or a digital music player to the hearing aids, and it serves as a hearing aid remote control, allowing change of volume and program. Two-way communication between smartphone and gateway takes place via Bluetooth, whereas one-way communication between gateway and hearing aids uses low-power near-field magnetic induction. (Recent technological advances are enabling direct communication between the hearing aids and the smartphone, without an intermediate gateway. The LiNX by GN ReSound [Ballerup, Denmark] is an example.) Due to the increased power requirements of all-day use, both the smartphone and the Streamer were fitted with extralarge (double the normal capacity) batteries.

Sound Environment Classification

Hearing instrument users often prefer different instrument settings in specific acoustic environments (Elberling, 1999). Gatehouse et al (1999) coined the term "auditory ecology" to encompass the listening environments in which people are required to function, the tasks to be undertaken in these environments, and the importance of those tasks in everyday life. Thus, achieving accurate ESC categorization was a key objective for this context-aware hearing system. The authors chose the following six ESCs: QUI, MUS, NOI, SiQ, SiN, and Party (PTY, defined as speech in loud noise, possibly with music present as well). The intention was to compare the performance of HALIC-classified and hearing aid–classified ESCs to the gold-standard ESC.

The gold-standard ESC was obtained as follows. After receiving participant consent, we configured HALIC to record the audio during every listening evaluation (up to 80 sec and saved the resulting sound clip on the smartphone hard drive. We determined the gold-standard ESC by asking two independent raters, both graduate students, to classify the sound clips post hoc. For each listening evaluation, the rater could assign either one or two valid ESCs, the latter being applicable if the environment changed for a significant part of the clip.

HALIC-Classified Environmental Sounds—The environmental sound classification in HALIC was designed for real-time, real-world input and output. However, before HALIC could perform classifications in the real world, it needed training data. The authors generated a training set, the Sound Database, consisting of 30-sec-long sound clips. We

made recordings at restaurants, bars, and other public locations in Copenhagen, Denmark, and Palo Alto, California, making sure to obtain training data for all six ESCs. Since HALIC had access to two microphones with different characteristics (an omnidirectional Galaxy Nexus mic and a directional Streamer mic), 161 clips were recorded using the Galaxy Nexusmic as input and 222 clips using the Streamer microphone as input. Due to the uncontrolled nature of the recordings (one microphone, unknown distance to speech source, etc.), we could not measure objective speech-to-noise ratio criteria for the sound clips. Hence, the discrimination problem faced by HALIC was to classify an incoming signal into one of the six available listening environments based on the training provided by the Sound Database.

Next, the system digitized and stored a specified interval of the signal, regardless of its source, into a buffer. We partitioned the buffer into individual frames. From each frame, the system extracted a set of statistics, in this case, acoustic properties of interest. We based a number of the statistics on the time-domain signal, whereas we derived others from the signal's frequency spectrum. We calculated all statistics on a frame-wise basis (2,048 samples/frame, 50% overlap). Due to hardware limitations, sampling rate from the Streamer microphone was 8 kHz, 16 bit. To ensure that HALIC could easily handle audio from both sources, the smartphone's sampling rate was kept at 8 kHz, 16 bit as well. The features ultimately extracted as inputs for machine learning algorithms were based on an entire buffer (233 frames, spanning 30 sec). The mean (first central moment) and the variance (second central moment) of the statistics were calculated, as the variance may be more discriminative than the mean of the statistic itself (Scheirer and Slaney, 1997).

The computed central-moment values constituted a set of features for the buffer. To obtain a classification, we applied machine-learning algorithms to buffer-level statistics for each feature. Specifically, HALIC used a simple energy threshold to detect QUI, analyzed the persistence of harmonic peaks over successive power spectra to detect MUS, and implemented penalized logistic regression to create separate detectors for NOI, PTY, SiN, and SiQ. We trained Streamer and smartphone mic classifiers separately. To avoid numerical problems with the logistic regression, we scaled each feature as appropriate.

HALIC constantly monitored the smartphone's proximity and light sensors to determine whether the smartphone was concealed (e.g., in a pocket or purse). If the smartphone was out in the open, HALIC used its microphone for sound input and its associated classifier; otherwise, it switched to the Streamer microphone for sound input and used its corresponding classifier.

Hearing Aid–Classified Environmental Sounds—Evaluating the performance of the hearing aids' built-in environmental sound classifier was an important objective of the study. We flashed the test hearing aids with a special version of firmware with 30,000 bytes of blank EEPROM, which permitted storage of hearing aid–classified ESCs during listening evaluations. The hearing aid's built-in classifier attempted to discriminate between the QUI, NOI, SiN, and SiQ subset.

Location

Hearing aid setting preferences may also be dependent on location. The GN ReSound Smart app, for example, allows users to "geotag" their favorite places and associate particular hearing aid settings with each location (Christensen, 2014).

HALIC-Classified Location—If given permission by the user, a smartphone can determine its location using built-in hardware. Global Positioning System (GPS) is satellite based, accurate, and has worldwide coverage, making it an obvious choice for location estimation. However, since GPS requires an unobstructed view of its orbiting satellites, it works poorly inside of buildings, where many people spend the majority of their time (Cheng et al, 2005). It is also possible to obtain location information from cellular towers and wireless access points such as wireless local area network (WiFi) routers (Bahl and Padmanabhan, 2000). WiFi router signal strength is detectable only within a few dozen meters and each router has a unique media access control address whose coordinates Google makes available through a database lookup. If the smartphone detects signal strength from multiple routers, trilateration is possible, resulting in greater accuracy. The same procedure is also applicable to cellular towers.

We segregated location types into two broad categories: Places—stationary coordinates such as home or a restaurant—and Routes, which had moving coordinates. For Routes, the actual location was not as important as the measured speed, which differentiated between types of movement (e.g., walking, running, and sitting in a car). The default category was Place, as HALIC could only calculate Routes when GPS updates were available and the GPS accuracy had stabilized. Unlike environmental sound classification, which occurred continuously, location classification only took place when the user performed a listening evaluation. This approach significantly reduced the battery drain associated with obtaining a GPS fix, with the trade-off that a user had to wait 10 sec (HALIC displayed a countdown timer) before starting the evaluation, thus giving the GPS time to find stable coordinates.

User-Classified Location—During each listening evaluation, HALIC attempted to determine the current coordinates and whether these were moving or stationary. It compared the result with the stored entries in the user-specific location database. If HALIC recognized the coordinates, it suggested the most likely result to the user. Otherwise, HALIC prompted the user to tag the location by first answering the question, "Where are you now?" with "At a place" or "On route." Submenus provided options for 52 Place types and 10 Route types.

Other Sensors

To make the model more complete, every time a listening evaluation took place HALIC also measured sound level (based on the input signal obtained for environment classification, averaging time of 30 sec, hour of day, and day of week. After the study, a sensitivity analysis suggested how the output—in this case, the setting suggested by the model—depended on the input sensors. A standard method of sensitivity analysis is to observe the output when varying one input and holding the others constant, a one-at-a-time technique. For this simple analysis, we held out (dropped) one input sensor at a time and observed the change in output.

Settings

The bulk of previous research in trainable hearing aids has concentrated on learning overall gain, gain-frequency response, and compression based on volume control, sound level, and ESC. Instead, we tested the present context-aware hearing system concept on microphone mode and digital noise reduction. The microphone could operate in omnidirectional or directional modes, while noise reduction state could be active or off. Thus, the four possible settings were omnidirectional microphone mode with noise reduction off (OMNI/NR_OFF), directional microphone mode with NR_OFF (DIR/NR_OFF), OMNI with noise reduction on (OMNI/NR_ON), and DIR with NR_ON (DIR/NR_ON). The directional algorithm used was Oticon's default, implementing four separate adaptive polar patterns, each corresponding to a different frequency band (Flynn, 2004). The noise reduction used was a modified version of Oticon's SuperSilencer fast-attack-and-release algorithm (Schafer et al, 2013), since the results of an earlier pilot study indicated that participants showed no preference for the manufacturer's standard noise reduction algorithm.

Listening Evaluations

Although the smartphone-based HALIC user interface works as a remote control—allowing the user to change programs and adjust volume on the hearing instruments— its primary purpose was to allow users to input their preferences by performing listening evaluations, a form of ecological momentary assessment (EMA). In contrast to traditional retrospective self-report methods, EMA prompts participants to provide an immediate and systematic record of both the context and content of their daily lives at randomized points in time. By sampling an experience the moment it occurs, EMA avoids the memory biases associated with the use of daily diaries or weekly questionnaires. EMA has seen application in the field of audiology. Henry et al (2012) used personal digital assistants and EMA to evaluate the impact of chronic tinnitus on the daily activities of patients while Galvez et al (2012) obtained real-time responses from hearing aid users describing their experiences with challenging hearing situations. Naturally, the proliferation of mobile phones has made EMA more popular. For example, AudioSense (Hasan et al, 2013) is a system for evaluating the performance of hearing aids in the real world combining EMA with the collection of sensor data via smartphone.

We designed the listening evaluation to take <2 min to complete. Whenever participants initiated a listening evaluation, they waited for a 10-sec countdown, as necessitated by GPS location. Next, the user compared settings (described below). Then HALIC obtained information from the user about the current ESC and location. Lastly, HALIC presented a thank you screen where users could press a "save evaluation" button to complete the listening evaluation and store the data. At the start of each listening evaluation, HALIC also saved the audio signal, as long as the participant had given consent and the "permission to record audio" box remained checked on the settings screen.

The most important part of every listening evaluation occurred when users entered their setting preferences. Users were unaware that the study consisted of separate training and validation phases. During the training phase, HALIC tried to learn as much as possible about the user by building a preference model, while in the validation phase, this model generated

automatic decisions, via an inference engine, regarding the most appropriate setting for any given classified listening environment.

Training Phase: Listening Evaluations—During the training phase, there were two types of listening evaluations, the A/B Test (Figure 2) and the Self-Adjustment screen, both of which contributed to the trained model. The A/B Test is a standard method of comparing two variants, a paired comparison, applied here to evaluate two hearing aid settings, A and B. Paired comparisons have been shown to be reliable and correlate well with performance measures (Neuman et al, 1987). In the A/B Test, the participant listened to both settings and gave a subjective, relative evaluation ("A is better," "B is better," or "no difference"). The "no difference" response, while not traditionally used in paired-comparison tests, was useful when the participant tried to listen to digital signal processing techniques that were not active or inaudible in that listening environment.

When a participant initiated a training-phase listening evaluation and the current HALIC-classified ESC was MUS, HALIC presented an A/B Test comparing the general program (P1) with the music program (P2). Similarly, when the HALIC-classified ESC was PTY, HALIC presented an A/B Test comparing the general program (P1) with the party program (P3). When the current environment was QUI, NOI, SiN, or SiQ, HALIC randomly selected settings corresponding to A and B. Choosing two of the four available settings (OMNI/NR_OFF, DIR/NR_OFF, OMNI/NR_ON, DIR/NR_ON) gave six possible combinations of A and B. In approximately every seventh A/B Test, HALIC randomly made settings A and B identical, then recorded whether the participant reported "no difference." Thus,

$$\text{Reliability} = \frac{N_{ND,A=B}}{N_{A=B}} \times 100\% \quad (1)$$

where $N_{ND,A=B}$ = number of evaluations answered "no difference" when A and B were identical and $N_{A=B}$ = number of evaluations where A and B were identical.

Although the A/B Test concealed the names of the settings, the Self-Adjustment screen had explicit labels for toggling "noise reduction" and "directional listening," making the settings visible to the users. Details about the two types of listening evaluations, and the characteristics of users who preferred training with each one, appear in Aldaz et al (2014). For the purposes of this article, it is sufficient to note that the outcome of each completed listening evaluation (except those measuring reliability) informed the trained model. At the end of the training phase, therefore, each participant would have created an adaptive self-trained model of his or her listening preferences.

Validation Phase: Listening Evaluations—In the validation phase, HALIC only presented A/B Tests to the users. Without the knowledge of the participants, the meaning of A and B changed, with HALIC randomly assigning A and B to the untrained setting suggested by the hearing aids and the setting suggested by the trained system. For each validation-phase A/B Test i, points awarded were $V_i = +1$ if the user voted that model's

setting, $V_i = -1$ if the user did not vote for the model's setting, and $V_i = 0$ for both models if the user chose "no difference."

After completing L validation-phase listening evaluations, the participant obtained a final trained system score of $S_T = \sum_{i=1}^{L} V_{i,T}$ and untrained system score of $S_U = \sum_{i=1}^{L} V_{i,U}$. For each participant, the preference score between the trained and untrained systems, P, was given by $P = (S_T - S_U)/2$, where the 1/2 factor is needed because of the +1/−1 (difference of 2 per A/B Test) scoring criterion. Variability in number of validation-phase listening evaluations completed by each participant necessitated calculation of the normalized scores, $P_n = \frac{P}{L}$.

The Knowledge-Based Agent

A knowledge-based agent, to use a term from the field of artificial intelligence, incorporates a repository of expert knowledge (the knowledge base) and an automated reasoning component (the inference engine). The knowledge base is simply the data accumulated by HALIC during the training phase. The primary task of the inference engine is to apply some strategy that would best allow it to adapt to varying input conditions and produce the optimal output (among the available choices) as often as possible during the validation phase.

This kind of problem is a good candidate for reinforcement learning, which maps states to actions to maximize a numerical reward signal (Sutton and Barto, 1998). Let S be the set of possible states s based on the outcome of the listening evaluation, so S = {preferred, not preferred}. Outcome states are denoted as s'. The agent has available a five-dimensional evidence vector **e** of smartphone sensor values (Table 1) providing partial knowledge about the state of the world. It does not convey, for example, relevant information about whom the user is conversing with or what sound the user is currently focusing on. Moreover, the sensor readings are noisy and sometimes the classifications are incorrect. Imperfect as they are, these sensor values constitute the evidence for probabilistic inference. Based on this input, the agent can take one of four nondeterministic actions a, each one corresponding to selecting a setting. Thus, set A = {select omni/nr_off, select dir/nr_off, select omni/nr_on, select dir/nr_on}.

The utility function $U(a, s')$ is a mapping from action-state pairs to real numbers, a value that expresses the desirability of that outcome state for that user. A probability is assigned to each possible result for each action, given the proposition that action a is executed in the current state and given the five-dimensional sensor vector **e**. The expected utility EU of taking action a is defined as the average utility value of the outcomes, weighted by the probability that the outcome occurs. The principle of maximum expected utility tells the agent to take action a because it maximizes the expected utility:

$$a = \arg\max_{a \in A} \mathrm{EU}(a|\mathbf{e}) \quad (2)$$

Even though the maximum expected utility principle dictates the right choice of action in a decision problem, it is seldom straightforward to implement (Russell and Norvig, 2009). With some manipulations, we implemented Equation 2; the details are beyond the scope of this article but are treated in Aldaz (2015).

Participants

Sixteen participants (ten males, six females, randomly denoted p1–p16) with symmetric (within 15 dB for any given frequency), sensorineural HL and ranging in age between 22 and 79 yr (mean = 55.5) participated in this study. Participants had a broad range in years of experience using hearing aids (mean = 6.75 yr, standard deviation [SD] = 6.85 yr). Figure 3 shows their mean and SD of audiometric thresholds. The bilateral four-frequency average HL, measured across 0.5, 1, 2, and 4 kHz, ranged from 21 to 63 dB HL (mean = 42 dB HL), indicating that the participants had mild-to-moderate HL. This research was performed under protocols approved by the Western Institutional Review Board (study number: 1138663) and the Human Subjects Research Institutional Review Board panel at Stanford University (protocol ID: 26552).

We recruited participants through four San Francisco Bay area private audiology offices. Each participant visited his or her audiologist at least three times during the course of the study. During the initial visit, we obtained informed consent, including permission to record audio during listening evaluations, and an audiologist took ear impressions. During the second visit, researchers administered a questionnaire to learn about each participant's hearing history, experience with hearing instruments, and expectations of the study. An audiologist fitted all participants bilaterally with the test hearing instruments and custom ear molds vented as appropriate for their loss to avoid occlusion and feedback issues. The audiologist prescribed test instruments through the Genie fitting software, determining gain settings using Oticon's proprietary Voice Aligned Compression (VAC) fitting rationale. Similar to NAL-NL2, VAC applies curvilinear compression to maximize speech intelligibility, emphasize audibility of soft sounds, and keep loud sounds comfortable (Schum, 2011). Unlike NAL-NL2, VAC is not a validated prescriptive method; therefore, it was not possible to verify targets with probe-mic measurements (Sanders et al, 2015).

All four clinics followed their own standard practices during the fitting process. In two of the four offices, the audiologist made real-ear measurements to check the VAC fitting against NAL-NL2 targets. In all offices, the audiologist manually adjusted the VAC fitting based on clinical experience and user feedback. The hearing aids had four user-selectable programs, accessible via the gateway or the smartphone: P1, general; P2, music; P3, party; P4, noise reduction.

Of the 21 persons who signed the consent form, 5 dropped out immediately after the second visit. The dropouts cited one or more of the following reasons for their withdrawal: lack of time to go through study (2), insecurity about proper usage of the technology (2), discomfort wearing ear molds (2), and ear infection (1).

As HALIC relies on Internet connectivity for location-based services, users were required to have a Subscriber Identification Module (SIM) card with a data plan installed in their

Galaxy Nexus phones. For the four participants who transferred their personal SIM cards to the Nexus, we transferred information from the participant's mobile phone (contact numbers, calendar, etc.) and relevant applications downloaded. Ten participants chose to carry both their personal mobile phone and the Nexus with an unlimited voice and provided data SIM card. The remaining two participants began with a Galaxy Nexus but experienced difficulties and opted to carry both phones. Participants received instructions to keep their smartphones with the WiFi radio on and grant location access via GPS satellites as well as WiFi and mobile networks.

In a session lasting ~30 min, a researcher then instructed each individual on how to use the system. The researcher verified that participants understood and were able to use the system by having them sit in a sound booth and perform practice listening evaluations in four different ESCs (MUS, NOI, SiQ, and SiN). As necessary, the researcher's voice served as the speech inputs, while nonspeech sounds were generated in the sound booth by playing prerecorded audio files. All participants departed the clinic wearing the system, with instructions to use it as often as possible during the test period. Every participant contacted us at least once during the test period. At the conclusion of the test period, they returned to the audiology office for a debriefing session. We conducted a postinterview to ask participants open-ended questions about hearing situations that they experienced and to solicit specific feedback about the context-aware hearing system. Participants performed four more listening evaluations in the sound booth.

RESULTS

Each participant's raw data stored on the hearing aids and smartphone were downloaded to a computer. The 16 participants completed 3,578 listening evaluations, an average of >5.5 per person per day for an average 41 days. We analyzed hearing aid data to extract the hearing aid–classified ESC, to verify that settings were as expected during each listening evaluation, and to monitor hearing aid usage. On average, participants used their devices 4–16 hr per day (mean = 10 hr) during the study. Smartphone data contained a wealth of information associated with each listening evaluation, including HALIC-classified ESCs, user-classified ESCs, HALIC-classified locations, user-classified locations, and user setting preferences.

We excluded preference data from p16 after discovering during the postinterview that this participant reported inconsistent preferences. Participant p16 stated, "Program A is best for speech comprehension and not bad for music. Program B is best for music listening—sounds are clearer and expanded and subtly muffled for speech." This participant had not understood that A and B were randomized settings and, having never chosen "no difference," obtained a reliability score of 0%.

We collected the audio recorded during each listening evaluation. The recordings were incomplete, as p10 did not give consent to audio recording and two others (p7 and p11) disabled audio recording during parts of the study. As described earlier, objective speech-to-noise ratio classification criteria for the remaining 3,033 audio clips could not be determined. Instead, we established the gold-standard ESC by having two trained, independent raters listen to and manually classify each sound clip. For each clip, the rater

awarded two possible classifications from the choices QUI, MUS, NOI, PTY, SiN, and SiQ. For example, if the clip contained only traffic noise, the rater would code the clip as NOI, NOI. If the clip contained speech throughout, with intermittent background noise, the rater would code the clip as SiQ, SiN. The two raters, with a Cohen's kappa of 0.775 based on their initial classifications, agreed on a final classification. (Cohen's kappa is a statistic that measures interrater agreement for categorical items, where a value of 1 means perfect agreement.) The duration of each clip was typically 80 sec and the ESC sometimes changed during this time. In 1,558 out of 3,033 (51.4%) sound clips, the raters agreed on two valid ESCs.

Validation Phase: Settings Preferences

The validation-phase listening evaluations were A/B Tests randomly comparing trained and untrained settings. Data from p16, who had not understood the experimental procedure, were excluded (so, n = 15). From these participants, we discarded data from incomplete listening evaluations, where the user never pressed the "save evaluation" button. Seven of the 15 participants had a significant preference for the trained settings (binomial test, $p < 0.05$), whereas one participant (p10) had a significant preference for the untrained settings. The remaining seven participants had nonsignificant preferences. Combined, all participants preferred the trained settings 49.2% of the time (binomial test, $p < 0.00001$), untrained settings 27.6% of the time, and indicated no difference in 23.2% of validation-phase listening evaluations. To determine statistical significance for the data as a whole, a paired-samples t test ($\alpha = 0.05$) was used. For the normalized preference scores, P_n, mean = 0.403, SD = 0.616, $t_{(14)} = 2.53$, $p = 0.0238$, indicating that participants' preference for the trained settings was statistically significant at the α level. The effect sizes were $r = 0.560$ and $d = 0.654$. Figure 4 shows the normalized scores for each of the 15 participants.

Reliability (Eq. 1) ranged from 57% to 100% (mean = 87%). According to Pearson's product-moment correlation analysis, degree of preference for trained settings positively correlated with reliability [$r_{(13)} = 0.630$, $p < 0.05$]. The intercept of the trend line in Figure 5 reveals that beyond a reliability of 75% the trained system started being beneficial. As shown in Figure 6, degree of preference for trained settings inversely correlated with dB HL [$r_{(13)} = -0.636$, $p \approx 0.01$]. A weaker inverse correlation also existed between preference for trained settings and years of hearing aid experience [$r_{(13)} = -0.548$, $p < 0.05$].

Participants completed an average of 142 training-phase listening evaluations. The data did not suggest a correlation between preference for trained settings and completed training-phase listening evaluations [$r_{(13)} = 0.099$, $p = 0.726$]. However, among participants who completed more than 150 training-phase listening evaluations, four out of five (80%) had a significant preference for the trained system (the fifth was nonsignificant). Neither participant age [$r_{(13)} = 0.406$, $p = 0.133$] nor average hours of hearing aid use [$r_{(13)} = -0.448$, $p = 0.094$] reached significance in discriminating between the participants' performance.

Training Phase: Settings Preferences

We also analyzed training-phase listening evaluations, discarding data from incomplete listening evaluations or those where the HALIC-classified ESC did not match the gold-standard ESC. Figure 7 depicts the proportion of times that each setting was preferred in each gold-standard ESC, pooling data across participants. (For the A/B Test results, trials in which the user indicated no preference were ignored, and the resulting small discrepancies in the total number of presentations n_i of each setting i corrected by dividing the number of times r_i the setting was chosen by n_i and then dividing each adjusted total $\frac{r_i}{n_i}$ by the grand adjusted total $\sum \frac{r_i}{n_i}$ to obtain corrected proportions.) Preferences were close to uniformly distributed (25%) for each ESC, with DIR/NR_OFF slightly less preferred. Overall, there was minimal variation in setting preference across ESCs. Expressed as percentages of training-phase A/B Tests comparing omnidirectional and directional modes (regardless of noise reduction state), OMNI was favored 33.2% of the time, DIR 27.0%, and no difference in the remaining 39.8%. Expressed as percentages of training-phase A/B Tests comparing noise reduction active and off (regardless of microphone mode), NR_ON was favored 28.9% of the time, NR_OFF 20.2%, and no difference in the remaining 50.9%.

When we broke down the same preference data by participant, a different picture emerged. Figure 8 illustrates individual setting preferences according to gold-standard ESC classifications, where each of the 16 squares represents data from a single participant numbered across rows from p1 in the upper left corner to p16 in the lower right. Three of the squares are entirely blank: p2 (partial training data lost due to technical error), p10 (did not give permission to record audio), and p16 (data not usable). If the participant completed fewer than five valid evaluations in a particular ESC, usually QUI, the column is also blank. Figure 8 shows that individual preferences can be strong and idiosyncratic. One participant (p1) strongly preferred the DIR/NR_ON setting for all four ESCs, while two others (p4 and p14) strongly preferred OMNI/NR_OFF in the ESCs in which they performed evaluations. Two participants, p11 and p15, predominantly preferred OMNI/NR_ON. The χ^2 statistic was used to compare the distribution of individual settings preferences (%) for each of the four ESCs to the overall distribution. None of the participants showed significant ($p < 0.05$) differences from the overall distribution in QUI (n = 6) or SiQ (n = 10). However, the preferences of seven individuals significantly differed from the overall distribution in NOI (n = 13) and eight individuals in SiN (n = 13).

Environmental Sound Classification

Environmental sound classification was the basis for a number of the research findings. The gold-standard ESCs were used to evaluate the relative performance of HALIC's ESC classifier and the hearing aid's built-in classifier. Each classifier's ESC prediction was considered correct if it matched the one or, in 52% of the instances, two gold-standard ESCs for that listening evaluation.

Each column of the confusion matrix (Table 2) gives the HALIC-predicted ESC, while each row lists the gold-standard, or actual, ESC. Correct classifications, therefore, appear along

the main diagonal of the matrix. As expected, HALIC performed best on the QUI class, both in terms of recall and precision. The class with the lowest recall was SiN, as HALIC achieved a true positive rate of 46%. On the other hand, the classifier frequently predicted the PTY class incorrectly (30% precision). Over the six classes, the classifier managed an accuracy of 78% on the full data set.

The hearing aid classifier had only four possible outputs (QUI, NOI, SiN, or SiQ), and this is reflected in Table 3. Over these four classes, the hearing aids achieved 70%. For the purposes of comparison, we deleted the rows and columns corresponding to MUS and PTY from Table 2 (Table 4). Considering only the remaining four classes, HALIC's accuracy was 85%. Although this is a significant improvement, it is important to compare the results between the hearing aids and HALIC with caution. The hearing aids were probably designed with one set of objective classification criteria, and in this study, they were evaluated against a subjective sound environment rating.

Location

In completing 3,758 listening evaluations, participants tagged 3,195 locations as Places and 563 as Routes. The distribution of Places followed a power-law function. Although >50% of all tagged Places fell into the category of Home, the distribution also had a long tail. Of the 52 Place categories available to participants, only 5 (bus station, boat terminal, nightclub, museum, and zoo) were not tagged. Participants also tagged 330 locations as Other, indicating the broad range of Places in which they performed listening evaluations. For instance, participant p7 reported completing listening evaluations in an amusement park and a bowling alley, neither of which appeared in the Places list.

Similarly, the distribution of Routes also followed a power-law function. Car dominated, with 73% of all tagged Routes. Although car and walking together accounted for 90% of all tagged Routes, the long tail was again evident. Of the 10 Route categories available to participants, only two (bicycling and boat) were not tagged, whereas four locations were tagged as Other Routes.

It is important to remember that we obtained data from participants who lived in the United States, and that distributions might look significantly different if a similar user study took place in another part of the world. On average, participants ate out in restaurants regularly, and strongly preferred traveling in cars to public transport. They also traveled frequently, completing listening evaluations throughout California (including Lake Tahoe, Los Angeles, and San Diego) as well as Santa Fe (New Mexico), Houston (Texas), Chicago (Illinois), Ann Arbor (Michigan), Orlando (Florida), upstate New York, and Hawaii.

Sensor Sensitivity

At the conclusion of the training phase, HALIC had built a trained model for each participant based on five sensor classifications—ESC, sound level, location, day of week, and hour of day. Using each participant's trained model, we ran the inference engine for each of the corresponding validation-phase listening evaluations (without reinforcement learning). We took the resulting set of output settings as the ground truth. In subsequent inference engine runs, we removed each one of the five sensors from the model and the

output setting compared against the ground truth (Table 5). The results indicated that the output was most sensitive to ESC, as holding it out of the model resulted in ~20% error in output. The second most sensitive output was sound level, as dropping it resulted in a 15% error in output. Holding out hour of day yielded an 8% error in output, making it the least sensitive sensor.

DISCUSSION

A number of technologies have been developed to increase the benefit provided by hearing instruments to their wearers. The purpose of the present user study was to investigate whether participants could train a prototype hearing system, using a smartphone in the real world, to provide settings preferred to untrained hearing aid settings. The experimental data on 16 participants suggest that the self-trained system became more strongly preferred to the hearing aids' untrained system. Of the 15 participants with valid data, 7 had a significant preference for the trained settings, 1 participant had a significant preference for the untrained settings, and the remaining 7 participants had nonsignificant results.

The finding that about half of the participants who provided valid data reported no preference for either response is in agreement with the Keidser and Alamudi (2013) study. Zakis et al (2007) reported that only about one-quarter of participants had nonsignificant preferences. One possible reason for this discrepancy is that participants in the present study completed an average of 142 training-phase listening evaluations over four weeks, fewer than the 300 adjustments required by Zakis et al. Although no overall significant correlation between number of training-phase listening evaluations and preference for trained settings existed, among participants who completed more than 150 training-phase listening evaluations, 80% (4/5) had a significant preference for trained settings. Zakis et al (2007) and Keidser and Alamudi (2013) also found that training was effective for 75–80% of those participants who had trained the devices sufficiently across several environments.

The participant who demonstrated a significant preference for the untrained settings, p10, had the lowest reliability of any individual with valid data. However, this participant also had the greatest average HL (63 dB HL), and the data suggested a significant inverse correlation between preference for trained settings and dB HL. Keidser and Alamudi also found a tendency for the eight participants who showed no preference to have a higher degree of HL (55.7 dB HL versus 50.8 dB HL; $p = 0.13$) than the eight participants who preferred the trained response. Mueller et al (2008) observed no significant correlation between average dB HL and gain deviation from NAL-NL1 targets.

Combined, participants voted for the trained settings 49.2% of the time. Again, this value is lower than the 68% reported by Zakis et al using the SHARP prototype. One possible explanation is that the SHARP had no other signal processing features—such as directional microphones, feedback reduction, and wind noise reduction—and allowed participants to appreciate fully the difference between trained and untrained compression and noise suppression parameters. In the present study, differences between the trained setting and the untrained setting may not have always been clear; in fact, it was possible that in a given listening evaluation, both the hearing aid and HALIC could present identical settings. For

this reason, a “no difference” button was provided, which was absent in the Zakis et al study, and contributed to the lower percentage of votes for the trained settings.

Directional Microphones and Noise Reduction

Little information exists about training microphone mode and noise reduction preferences. In Trial II of the Zakis et al (2007) study, noise reduction was enabled for the trained settings (with trained noise reduction strength), whereas in Trial III, noise reduction was disabled for both the trained and untrained settings. The outcome for the group data was that the provision of trained noise suppression did not have a significant effect on the preference for the trained settings.

In the present study, the pooled data revealed that preferences were nearly uniformly distributed (25%) for each real-world ESC, with DIR/NR_OFF slightly less preferred in QUI, NOI, SiN, and SiQ (Figure 7). This disfavoring of DIR/NR_OFF may have been due to an increase in audible circuit noise. Initially, venting and phase relationships between the front and rear microphones cause low-frequency roll-off. To compensate for this decrease in audibility while in the directional mode, a process called equalization increases the low-frequency gain so the DIR frequency response nearly matches that of OMNI. Equalization has the undesired side effect of increasing audible circuit noise, which some participants may have found unpleasant.

The nearly uniform distribution across settings supports previous literature investigating user preference for microphone mode and noise reduction in the real world. In a study by Walden et al (2004), 17 hearing-impaired participants used pen and pencil to record their preferred microphone mode while describing the current listening situation in terms of five variables: background noise (“present” or “absent”), signal location (“front” or “other”), signal distance (“near” or “far”), reverberation (“low” or “high”), and noise location (“front” or “other”). Results suggested that knowing background noise presence, signal location, and signal distance were sufficient to make a reasonable guess of preferred mode. Participants showed a preference for omnidirectional mode in relatively quiet listening environments and in background noise with a relatively distant signal source. Only when background noise was present but the signal source was in “front” and relatively “near” the listener did participants clearly prefer the directional mode. Overall, OMNI was favored 37% of the time, DIR 33%, and there was no preference for either microphone mode in the remaining 30%. These values are similar to this study’s results:OMNI (33%), DIR(27%), and no difference (40%).

Meanwhile, manufacturer-specific noise reduction algorithms vary in the number of channels used, the time constants, and the magnitude of gain reduction as a function of frequency and sound level (Dreschler et al, 2001), making it difficult to compare results across studies. Boymans and Dreschler (2000) measured the effects of both a directionality system—the twin-microphone system (TMS)—and a noise reduction algorithm—speech-sensitive processing (SSP). The study combined laboratory experiments and field trials. Alongside objective data, the researchers collected subjective data via paired comparisons, with a test after week 4 and a retest after week 12. The four settings were TMS_OFF/SSP_OFF, TMS_ON/SSP_OFF, TMS_OFF/SSP_ON, and TMS_ON/SSP_ON. The only

ESC tested was SiN, specifically speech in cocktail party noise background and speech in car noise. Compared to the baseline TMS_OFF/SSP_OFF, results of the paired comparisons showed ~60% higher preference for the directional mode (TMS_ON/SSP_OFF), and 10–20% higher preference for noise reduction (TMS_OFF/SSP_ON). There was not a significant difference between TMS_ON/SSP_OFF and TMS_ON/SSP_ON. Crucially, the paired comparisons took place in the laboratory, with noise from three sides (90°, 180°, and 270°) and speech from 0° azimuth. A number of studies (Walden et al, 2000; Surr et al, 2002; Cord et al, 2004) found no relationship between laboratory performance of speech perception in noise, which showed highly significant directional advantages, and field ratings of speech understanding in noise, which yielded minimal directional benefit. Furthermore, research findings investigating changes in speech intelligibility due to implementation of digital noise reduction algorithms have been mixed, tending only to indicate that such algorithms can work to increase listener comfort and not showing a clear preference for noise reduction (Bentler, 2005).

Although the data in the present study suggest a relatively similar distribution of preferences for the four settings and environments across participants, individual results varied significantly. Walden et al (2004) observed as much, noting that one participant preferred OMNI in 90% of listening situations, whereas another preferred DIR in 73% of listening situations. These observations suggest that personalized settings could provide a benefit for some individuals not obtainable with an untrained algorithm.

Strengths and Limitations

Including a smartphone as part of a hearing system demonstrated several advantages. Three smartphone-related factors may have contributed to the participants' statistically significant preference for trained settings. The first consideration is the use of built-in sensors to create context awareness. In decreasing order of importance, these were ESC, sound level, location, day of week, and hour of day. The second factor is the graphical user interface, giving participants the ability to train the system via listening evaluations. The third contributor is smartphone memory and processing power. The use of machine learning and digital signal processing algorithms on the smartphone improved environmental sound classification performance; in QUI, NOI, SiN, and SiQ, the HALIC classifier achieved 84% accuracy, compared to 70% accuracy for the hearing aid classifier. In addition, the knowledge-based engine implemented reinforcement learning, which could adjust to user preferences even in the validation phase.

All 16 participants (those who remained after five dropped out during the initial phase) completed the study, although two participants required extra time to complete the requisite number of evaluations. Participants averaged 5.5 listening evaluations per day for ~6 weeks, a remarkable achievement considering that they did not receive any reminders or prompts from the research team. Every participant encountered technical difficulties and received either in-person or telephone support at some point during the study. During the postinterview, 14 of 16 participants reported positive feelings of clarity, competence, and mastery of listening evaluations. The sustained commitment of the participants, resulting in their successful completion of the study, demonstrated their extreme motivation when

engaged in an issue as fundamental to well-being as hearing. The participants' positive attitude and involvement supports the claim by Dillon et al (2006) that trainable hearing aids offer the advantage of increased sense of ownership and engagement.

The process of selecting participants for the study indicated that smartphones are not an appropriate part of a hearing system for a segment of the hearing-impaired population. The average age of the study participants was 55.5 yr, compared to the average age of a hearing aid user, ~70 yr (Kochkin, 2005). A major exclusion criterion was lack of technology use (persons who did not own computers or cell phones, or were never online), which is more prevalent among the 80+ age group. Of the 21 individuals who agreed to participate in the study, 5 dropped out just after getting started, with 2 of them citing insecurity about proper usage of the technology. It is possible that these people could have completed the study had they received more instruction, not only on the HALIC system but also on the function of the Streamer and the smartphone.

The present study had a number of notable limitations. First, the fitting rationale was manufacturer specific, and therefore could not be verified with probe-mic measurements. Given that starting gain has been shown to be associated with final hearing aid gain settings (Mueller et al, 2008), over- or under amplification may have significantly affected user preferences for trained and untrained hearing aid settings. Second, for the 16 individuals who did complete the study, we were not able to control for acclimatization to the hearing aids. It is possible that preferences for trained or untrained settings would change over time as participants acclimated to the new amplification. User test data showed a weak inverse correlation between preference for trained settings and years of hearing aid experience. Lastly, we acknowledge the limitations of training hearing aids using a smartphone. Participants did not train the system uniformly across ESCs and locations due to varying levels of commitment as well as social norms regarding use of mobile phones.

Future Work

This article has presented the design and evaluation of a novel trainable hearing system comprising two hearing aids and a smartphone, wirelessly connected through a body-worn gateway, which was required at the start of the study but is already becoming obsolete. The space of settings chosen was limited to microphone mode and digital noise reduction, and further restricted to on/off functionality. We hope to extend the prototype system to include other parameters as well. The question of which gain-frequency response to implement is almost as old as electronic hearing aids themselves, and has been addressed in past trainable hearing aids. A smartphone-based system could provide a novel method for the selection of gain-frequency responses and compression parameters, as is evident by the development of recent smartphone-based fitting algorithms (Nelson et al, 2014). Although participants felt confident using the A/B Test that they were giving "good data," a limitation of paired comparisons is that they require many iterations to train if the setting space is large.

Due to institutional review board considerations, the prototype hearing system only controlled the hearing aid settings during listening evaluations, as this minimized the chances for the system to malfunction or cause harm. Consequently, participants reported frustration with selecting a preferred setting in a listening evaluation and then having the

hearing instruments revert back to an undesired setting immediately afterward. Participant p6 expressed disappointment at being a "passive recorder" and lamented that the system did not become "a dynamic participant in my hearing." Clearly, the capability to remain in the preferred setting after the conclusion of the listening evaluation would provide an improved user experience. As the hearing system learned, the user's need to perform listening evaluations should decrease over time, giving the user a greater sense of involvement.

A second possible enhancement of the system would be to allow gradual changes. For example, if a participant responded well to a mild noise reduction algorithm, HALIC could present the wearer with increasingly more aggressive responses to find the optimum setting. This gradual change method is readily applicable to frequency-gain adjustments as well.

Although the results of this study are encouraging, future research on the reliability of training and the effect of training on additional objective and subjective measures of benefit is needed to evaluate the overall value of training with a smartphone-based hearing system.

Acknowledgments

This research was funded by the Oticon Foundation, Smørum, Denmark.

The authors wish to thank the many engineers, audiologists, and researchers of William Demant Holding—both at Oticon A/S (Smørum, Denmark) and the Eriksholm Research Centre (Snekkersten, Denmark)—without whom this research would not have been possible. In particular, the authors thank advisers Christian Binau Nielsen and Søren Kamaric Riis for their guidance. Ariane Laplante-Lévesque and Graham Naylor provided extremely helpful input. Kasper Nyborg Nielsen, Piotr Sapieżyński, and Andreas Borup Svendsen spent countless hours helping to code and debug the Android software. The authors also thank the students who helped set up and run the user study (Dafna Szafer and Tyler Haydell), the two independent raters, and the audiologists in California who assisted with participant recruitment, fitting, and troubleshooting: Dr. Darcy Benson, Dr. Ramsay Poindexter, and Dr. Teresa Testa (California Hearing, San Mateo), Dr. Jane Baxter, Dr. Shu-En Lim, and Dr. Peg Lisi (Pacific Hearing, Menlo Park), Dr. Debbie Clark, Dr. Erin Harrigan, and Dr. Brook Raguskus (Pacific Hearing, Los Altos), and Dr. Bill Diles (Kenwood Hearing, Petaluma). Lastly, but perhaps most importantly, the authors thank the user study participants for their time and patience.

Abbreviations

DIR	directional microphone mode
EMA	ecological momentary assessment
ESC	environmental sound class
GPS	Global Positioning System
HALIC	Hearing Aid Learning and Inference Controller
HL	hearing loss
MUS	Music environmental sound class
NAL-NL1	National Acoustic Laboratories-nonlinear 1
NAL-NL2	National Acoustic Laboratories-nonlinear 2
NOI	Noise environmental sound class

NR_OFF	noise reduction off
NR_ON	noise reduction on
OMNI	omnidirectional microphone mode
PTY	Party environmental sound class
QUI	Quiet environmental sound class
SD	standard deviation
SHARP	Stereo Hearing Aid Research Processor
SIM	Subscriber Identification Module
SiN	Speech in Noise environmental sound class
SiQ	Speech in Quiet environmental sound class
SSP	speech-sensitive processing
TMS	twin-microphone system
VAC	Oticon Voice Aligned Compression rationale
WiFi	wireless local area network

REFERENCES

Aldaz, G., Haydell, T., Szafer, D., Steinert, M., Leifer, L. User experience in training a personalized hearing system. In: Markus, A., editor. Design, User Experience, and Usability. User Experience Design for Everyday Life Applications and Services. New York, NY: Springer; 2014. p. 3-14.

Aldaz, G. Smartphone-based system for learning and inferring hearing aid settings. Ph.D. thesis. Stanford, CA: Stanford University; 2015.

Allegro, S., Büchler, MC., Launer, S. Automatic Sound Classification Inspired by Auditory Scene Analysis. Aalborg, Denmark: Eurospeech; 2001.

Amlani AM, Taylor B, Levy C, Robbins R. Utility of smartphone-based hearing aid applications as a substitute to traditional hearing aids. Hear Rev. 2013; 20:16–18.

Bahl, P., Padmanabhan, VN. Proc IEEE Infocom 00. Los Alamitos, CA: IEEE CS Press; 2000. RADAR: An In-Building RF-Based User Location and Tracking System; p. 775-784.

Barker, J., Asmundson, P., Lee, P. 2013 Global Mobile Survey: Divergence Deepens. London: Deloitte; 2013.

Bentler RA. Effectiveness of directional microphones and noise reduction schemes in hearing aids: a systematic review of the evidence. J Am Acad Audiol. 2005; 16(7):473–484. [PubMed: 16295234]

Boymans M, Dreschler WA. Field trials using a digital hearing aid with active noise reduction and dual-microphone directionality. Audiology. 2000; 39(5):260–268. [PubMed: 11093610]

Chalupper J. Changing how gain is selected: the benefits of combining data logging and a learning VC. Hear Rev. 2006; 13:46–55.

Chalupper J, Junius D, Powers T. Algorithm lets users train aid to optimize compression, frequency shape, and gain. Hear J. 2009; 62:26–33.

Cheng, Y-C., Chawathe, Y., Krumm, J. Proc 3rd Int Conf Mobile Systems, Applications, and Services. New York, NY: ACM Press; 2005. Accuracy Characterization for Metropolitan-scale Wi-Fi Localization; p. 233-245.

Christensen L. Introducing ReSound LiNX, Made for iPhone Hearing Aid. Audiol Online. 2014 [Accessed October 25, 2014] http://www.audiologyonline.com/interviews/introducing-resound-linx-made-for-12569.

Cord MT, Surr RK, Walden BE, Dyrlund O. Relationship between laboratory measures of directional advantage and everyday success with directional microphone hearing aids. J Am Acad Audiol. 2004; 15(5):353–364. [PubMed: 15506497]

Dillon H, Zakis J, McDermott H, Keidser G, Dreschler WA, Convery E. The trainable hearing aid: what will it do for clients and clinicians? Hear J. 2006; 59:30–36.

Dreschler WA, Verschuure H, Ludvigsen C, Westermann S. ICRA noises: artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. International Collegium for Rehabilitative Audiology. Audiology. 2001; 40(3):148–157. [PubMed: 11465297]

Dreschler WA, Keidser G, Convery E, Dillon H. Client-based adjustments of hearing aid gain: the effect of different control configurations. Ear Hear. 2008; 29(2):214–227. [PubMed: 18490863]

Elberling C. Loudness scaling revisited. J Am Acad Audiol. 1999; 10(5):248–260. [PubMed: 10331617]

Flynn MC. Maximizing the voice-to-noise ratio (VNR) via voice priority Processing. Hear Rev. 2004; 11:54–59.

Galvez G, Turbin MB, Thielman EJ, Istvan JA, Andrews JA, Henry JA. Feasibility of ecological momentary assessment of hearing difficulties encountered by hearing aid users. Ear Hear. 2012; 33(4):497–507. [PubMed: 22531573]

Gatehouse, S., Elberling, C., Naylor, G. Aspects of Auditory Ecology and Psychoacoustic Function as Determinants of Benefits from and Candidature for Non-Linear Processing in Hearing Aids. Proc 18th Danavox Symposium: Auditory Models and Non-Linear Hearing Instruments; Kolding, Denmark. 1999. p. 221-233.

Hasan, SS., Lai, F., Chipara, O., Wu, Y-H. Proc IEEE 26th International Symposium on Computer-Based Medical Systems. Los Alamitos, CA: IEEE CS Press; 2013. AudioSense: Enabling Real-Time Evaluation of Hearing Aid Technology In-Situ; p. 167-172.

Hayes D. Empowering the hearing aid wearer through logging plus learning. Hear J. 2007; 60:20–25.

Henry JA, Galvez G, Turbin MB, Thielman EJ, McMillan GP, Istvan JA. Pilot study to evaluate ecological momentary assessment of tinnitus. Ear Hear. 2012; 33(2):179–290. [PubMed: 21960147]

Keidser, G., Dillon, H. What's new in prescriptive fittings Down Under?. In: Palmer, CV., Seewald, R., editors. Hearing Care for Adults 2006. Stafa, Switzerland: Phonak AG; 2006. p. 133-142.

Keidser G, Alamudi K. Real-life efficacy and reliability of training a hearing aid. Ear Hear. 2013; 34(5):619–629. [PubMed: 23575461]

Kochkin S. MarkeTrak VII: Customer satisfaction with hearing instruments in the digital age. Hear J. 2005; 58:30–43.

Mueller HG, Hornsby BW, Weber JE. Using trainable hearing aids to examine real-world preferred gain. J Am Acad Audiol. 2008; 19(10):758–773. [PubMed: 19358456]

Nelson, J., Kiessling, J., Dyrlund, O., Groth, J. Integrating Hearing Instrument Datalogging into the Clinic. Minneapolis, MN: Am Acad Audiol.; 2006.

Nelson, P., Van Tasell, D., Sabin, A., Gregan, M., Hinz, J., Brandewie, E., Svec, A. User Self-Adjustment of a Simulated Hearing Aid in Laboratory versus Real-World Noise. Scottsdale, AZ: Am Audiol Soc.; 2014.

Neuman AC, Levitt H, Mills R, Schwander T. An evaluation of three adaptive hearing aid selection strategies. J Acoust Soc Am. 1987; 82(6):1967–1976. [PubMed: 3429734]

Nielsen. [Accessed October 3, 2014] Mobile Majority: U.S. Smartphone Ownership Tops 60%. 2013. http://www.nielsen.com/us/en/insights/news/2013/mobile-majority–u-s-smartphone-ownership-tops-60-.html

Powers, T., Chalupper, J. Soundlearning 2.0: approaching optimum gain, compression and frequency shape in different listening situations. Erlangen, Germany: Siemens Publication; 2010. p. 1-11.

Ringdahl A, Eriksson-Mangold M, Israelsson B, Lindkvist A, Mangold S. Clinical trials with a programmable hearing aid set for various listening environments. Br J Audiol. 1990; 24(4):235–242. [PubMed: 2224290]

Russell, S., Norvig, P. Artificial Intelligence: A Modern Approach. 3rd. Upper Saddle River, NJ: Prentice Hall Press; 2009.

Sanders J, Stoody T, Weber J, Mueller HG. Manufacturers' NAL-NL2 Fittings Fail Real-ear Verification. Hear Rev. 2015; 21:24.

Schafer EC, Sanders K, Bryant D, Keeney K, Baldus N. Effects of voice priority in FM systems for children with hearing aids. J Educ Audiol. 2013; 19:12–24.

Scheirer, E., Slaney, M. Proc IEEE Int Acoustics, Speech, and Signal Processing. Washington, DC: IEEE CS Press; 1997. Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator; p. 1331-1334.

Schilit, B., Adams, N., Want, R. Proc IEEE 1st Int Workshop Mobile Computing Systems Applications. Los Alamitos, CA: IEEE CS Press; 1994. Context-Aware Computing Applications; p. 85-90.

Schum, DJ. The audiology of Oticon Intiga. White paper, Smørum. Denmark: Oticon A/S; 2011.

Smith, A. 46% of American Adults Are Smartphone Owners. Washington, DC: Pew Research Center; 2012.

Smith, A. Smartphone Ownership—2013 Update. Washington, DC: Pew Research Center; 2013.

Surr RK, Walden BE, Cord MT, Olson L. Influence of environmental factors on hearing aid microphone preference. J Am Acad Audiol. 2002; 13(6):308–322. [PubMed: 12141388]

Sutton, R., Barto, A. Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press; 1998.

Unitron. Next. Everything you need to succeed. Publication 08-020 028-5274-02, 6. Plymouth, MN: Unitron Hearing; 2011.

Walden BE, Surr RK, Cord MT, Edwards B, Olson L. Comparison of benefits provided by different hearing aid technologies. J Am Acad Audiol. 2000; 11(10):540–560. [PubMed: 11198072]

Walden BE, Surr RK, Cord MT, Edwards B, Dyrlund O. Predicting hearing aid microphone preference in everyday listening. J Am Acad Audiol. 2004; 15(5):365–396. [PubMed: 15506498]

Zakis, JA. A trainable hearing aid. Ph.D. thesis. Australia: University of Melbourne; 2003.

Zakis JA, Dillon H, McDermott HJ. The design and evaluation of a hearing aid with trainable amplification parameters. Ear Hear. 2007; 28(6):812–830. [PubMed: 17982368]

Figure 1.

Principal components of HALIC: context (ESC, location, time, and movement) is gathered via the smartphone's built-in sensors, user input is captured via listening evaluations, and the knowledge-based agent is in charge of learning and inference. Accel = accelerometer (detects movement); GSM = Global System for Mobile Communications (this cellular tower has a uniquely identifiable ID, making it useful in location estimation).

Figure 2.
Smartphone screen shots of the A/B Test sequence: Listening to setting A (left), listening to setting B (center), and choosing the best setting (right).

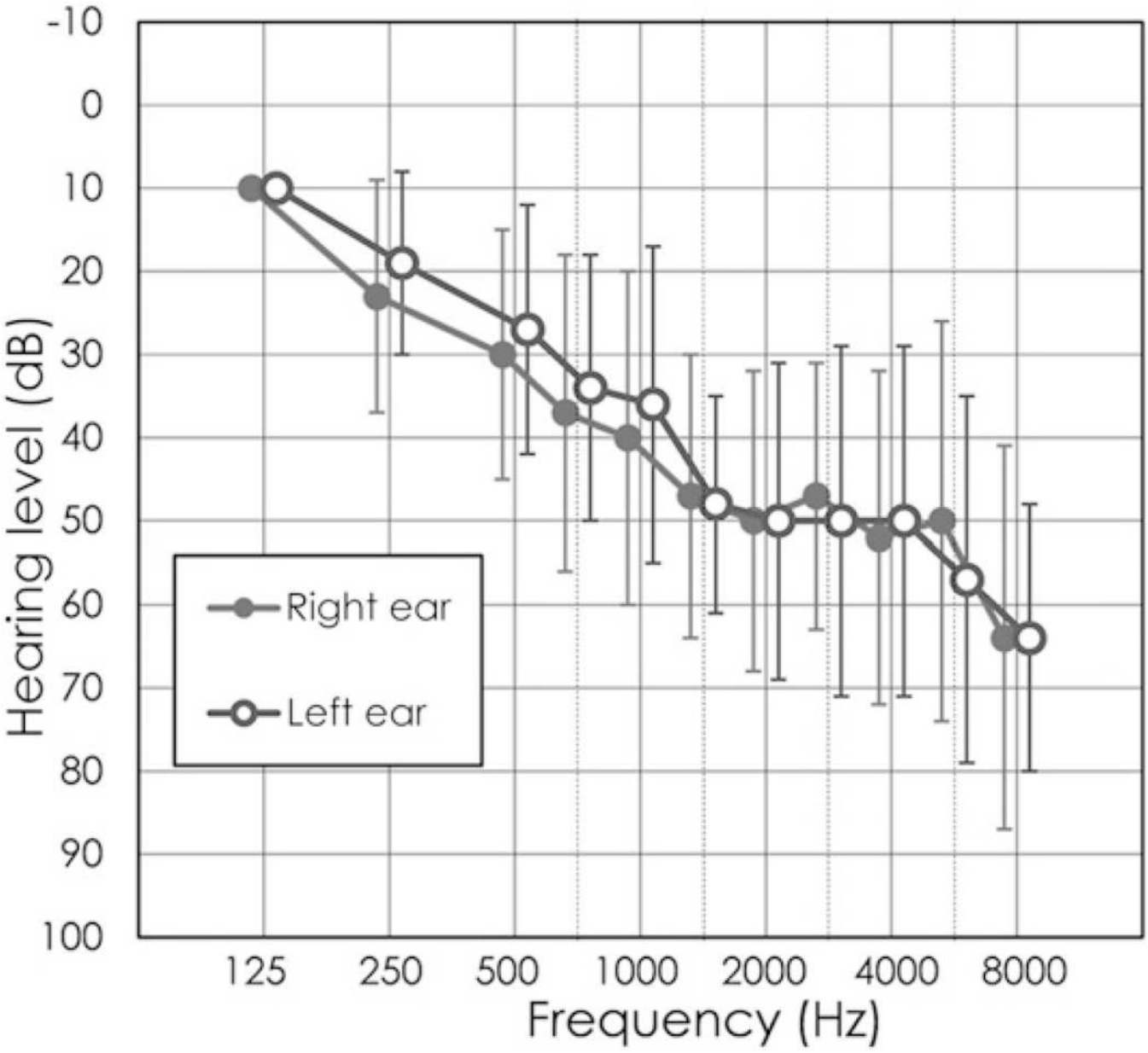

Figure 3.
Mean left- and right-ear audiometric thresholds (± SD) of the 16 participants.

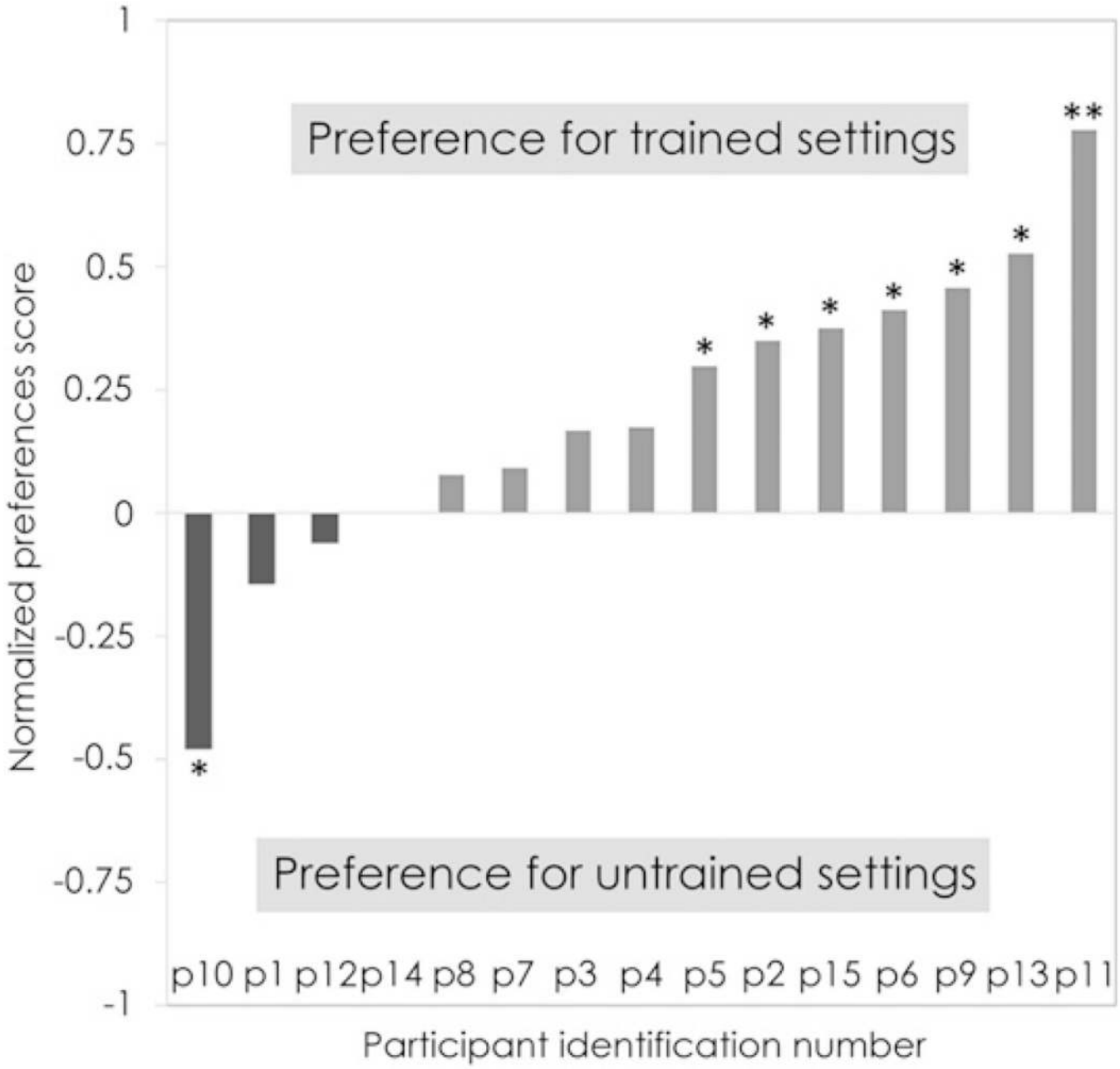

Figure 4.
Validation-phase normalized preference score for trained (positive values) versus untrained (negative values) settings for each participant. *$p < 0.05$; **$p < 0.01$.

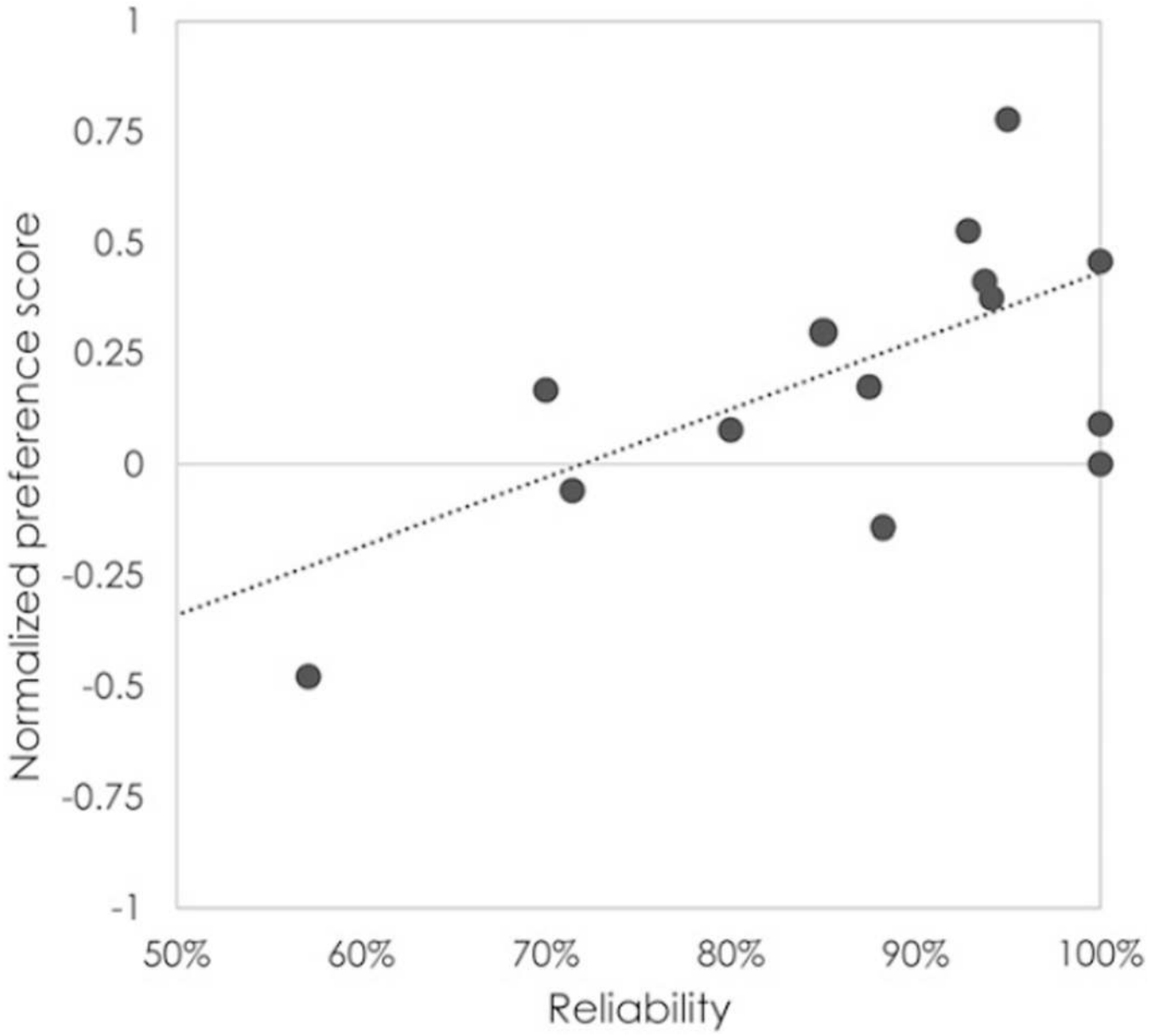

Figure 5.
Validation-phase normalized preference score versus reliability, defined in Equation 1 (minimum 50% reliability).

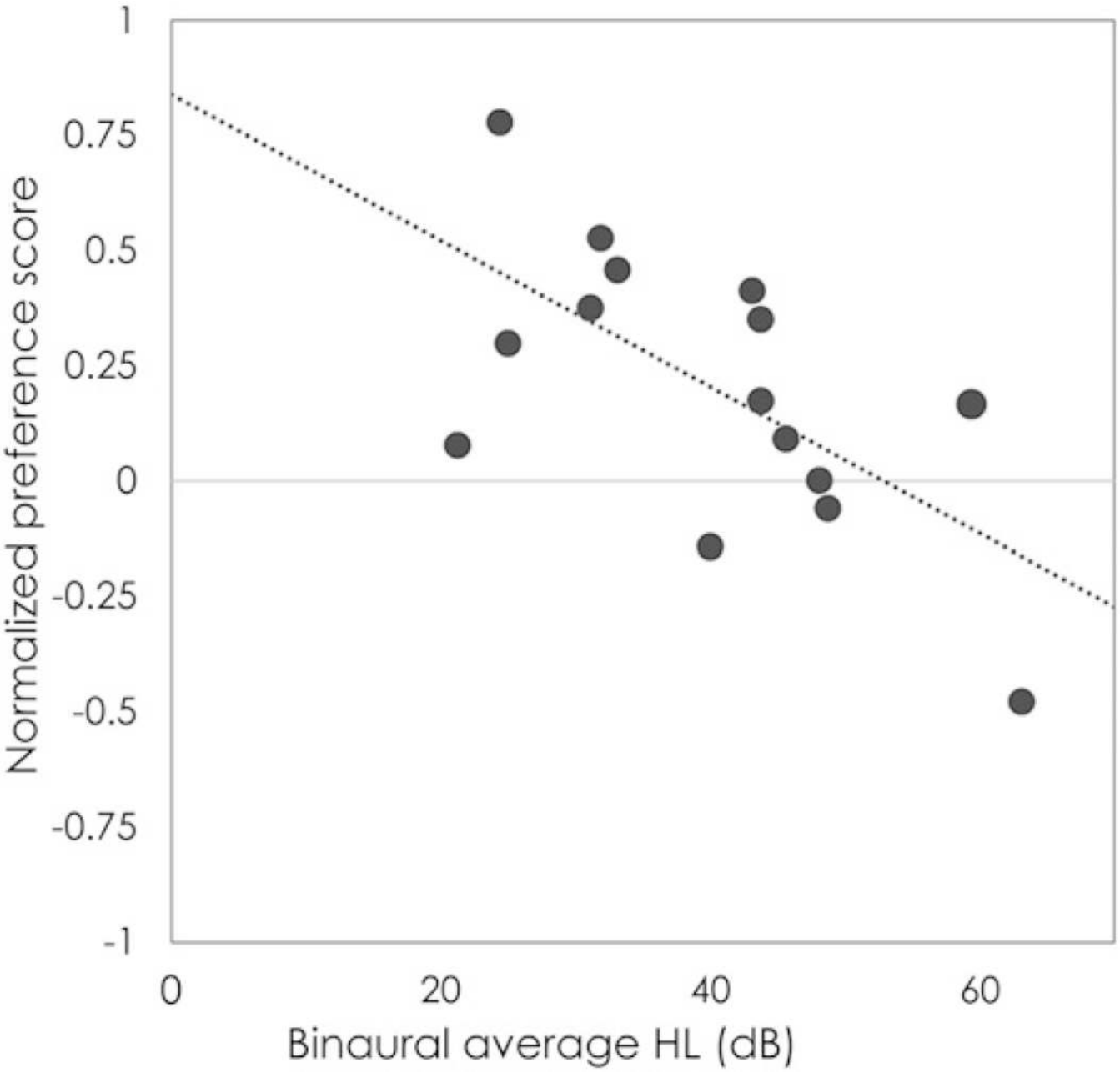

Figure 6.
Validation-phase normalized preference score versus degree of HL.

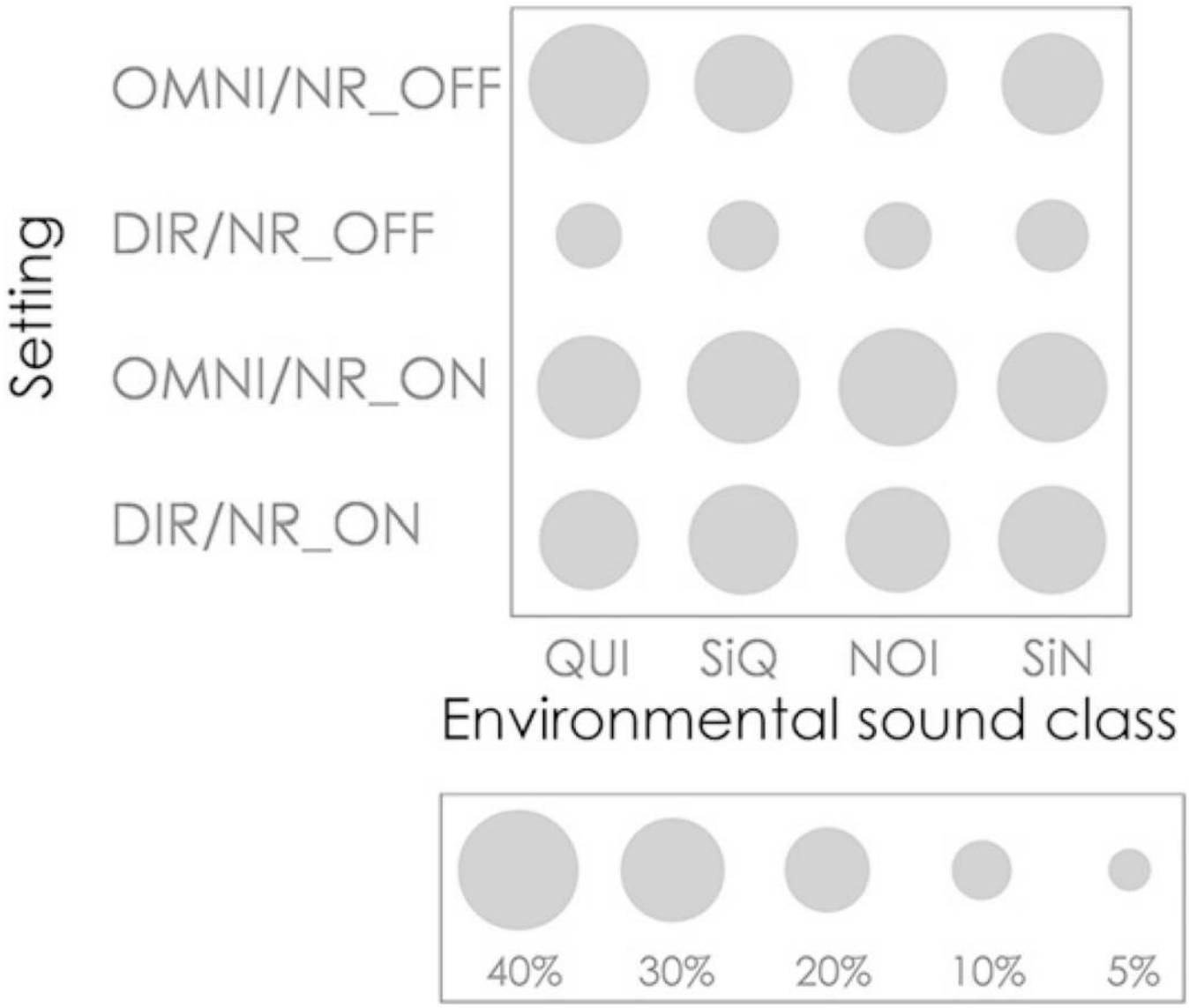

Figure 7.
Pooled training-phase setting preferences according to gold-standard ESC. Area of circles indicates proportions of preference for each ESC.

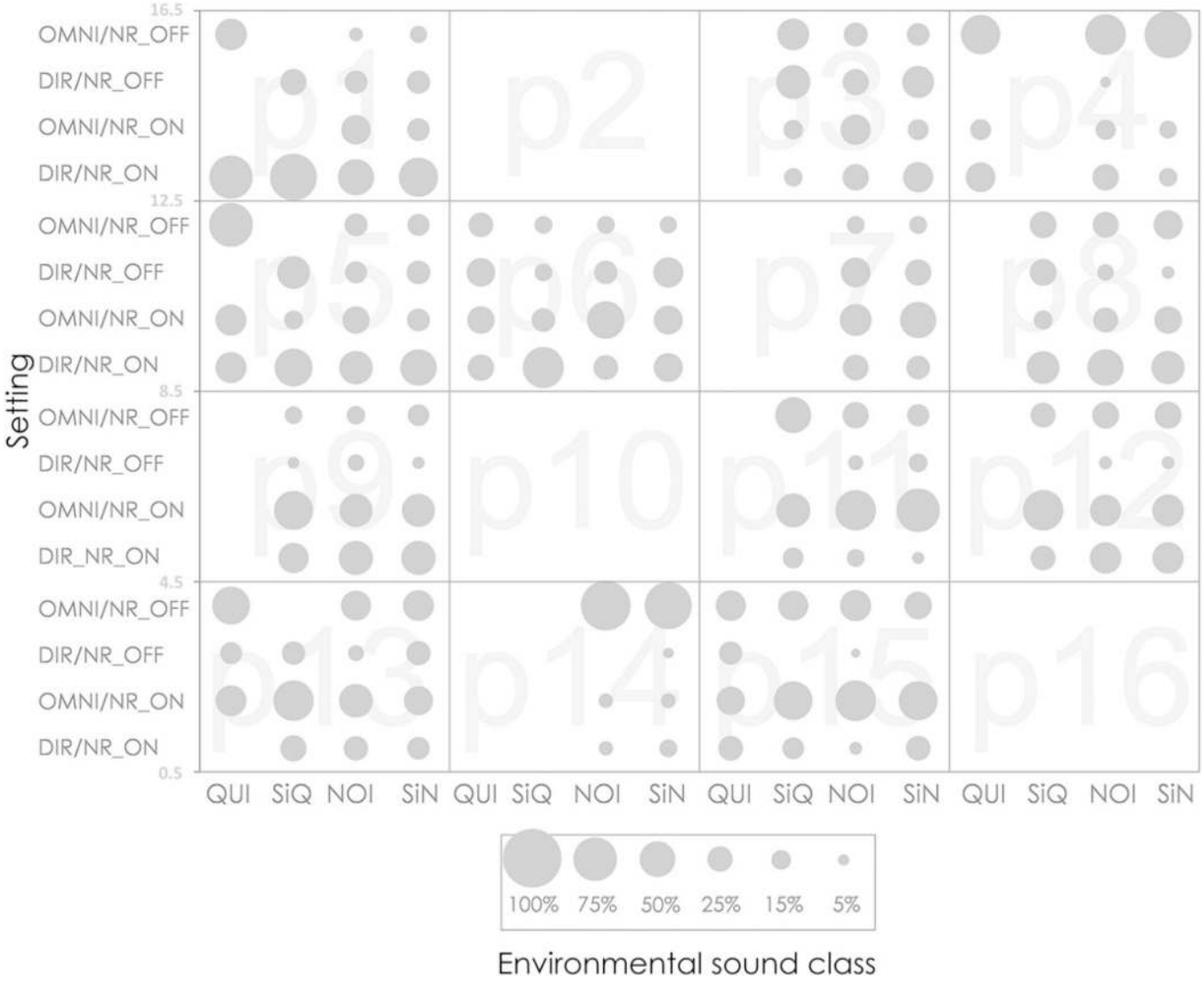

Figure 8.

Individual training-phase setting preferences according to gold-standard ESC. Data from participants, p2, p10, and p16, were excluded for various reasons (see text). Area of circles indicates proportions of preference for each ESC.

Table 1

List of Smartphone Sensors Used in the Prototype Hearing System

Sensor	Number	Values
ESC	6	QUI, MUS, NOI, PTY, SiQ, SiN
Sound level	5	Silent, soft, moderate, loud, very loud
Location	52 Places	Home, work, airport, restaurant, zoo, …
	10 Routes	In car, running, walking, …
Time of day	24	00:00–00:59, 01:00–01:59, 02:00–02:59, …
Day of week	7	Monday, Tuesday, Wednesday, …

Table 2

Confusion Matrix for HALIC Classifier over Six Classes

	Predicted						
	QUI	**SiQ**	**NOI**	**SiN**	**MUS**	**PTY**	**Correct (%)**
Actual							
QUI	**172**	2	7	2	0	0	94
SiQ	1	**162**	12	6	3	3	87
NOI	17	23	**1,277**	62	57	22	88
SiN	7	29	168	**282**	84	49	46
MUS	3	6	46	16	**302**	16	78
PTY	0	0	3	1	0	**38**	90
Correct (%)	86	73	84	76	68	30	**78**

Note: Values in boldface denote correct classifications.

Table 3

Confusion Matrix for Hearing Aid Classifier

	Predicted				
	QUI	**SiQ**	**NOI**	**SiN**	**Correct (%)**
Actual					
QUI	**121**	7	2	28	81
SiQ	2	**138**	13	131	76
NOI	9	131	**742**	253	65
SiN	1	39	39	**743**	72
Correct (%)	91	27	93	71	**70**

Note: Values in boldface denote correct classifications.

Table 4

Confusion Matrix for HALIC Classifier over Four-Class Subset

	Predicted				
	QUI	**SiQ**	**NOI**	**SiN**	**Correct (%)**
Actual					
QUI	**172**	2	7	2	94
SiQ	1	**162**	12	6	90
NOI	17	23	**1,277**	62	93
SiN	7	29	168	**282**	58
Correct (%)	87	75	87	80	**85**

Note: Values in boldface denote correct classifications.

J Am Acad Audiol. Author manuscript; available in PMC 2017 January 25.

Table 5

Effect of Dropping a Sensor on Classification Accuracy

Sensor	Misclassifications	% Error
Drop none (ground truth)	0	0
Drop ESC	212	19.7
Drop sound level	164	15.3
Drop day of week	133	12.4
Drop location	113	10.5
Drop hour of day	87	8.1

Note: n = 1,075 validation-phase listening evaluations.