Evaluating Auditory Contexts and Their Impacts on Hearing Aid Outcomes with Mobile Phones

Syed Shabih Hasan[1], Octav Chipara[1], Yu-Hsiang Wu[2], and Nazan Aksan[3]
[1]Department of Computer Science/Aging Mind and Brain Initiative
[2]Department of Communication Sciences and Disorders
[3]Department of Neurology
University of Iowa
{*syedshabih-hasan, octav-chipara, yu-hsiang-wu, nazan-aksan*} @uiowa.edu

ABSTRACT

This paper evaluates the relationship between auditory contexts, hearing aid features, and hearing outcomes based on real-world measurements. We use a mobile phone application to concurrently evaluate the auditory contexts and hearing aid outcomes using Ecological Momentary Assessments. The collected dataset includes 3437 surveys collected from nineteen patients over ten months. Our analysis indicates that the most frequent listening activities were conversations (32.7% of the time) and listening to media (30.7% of the time), commonly occurring at home, in predominantly quiet environments. Subjects do not attribute equal importance to hearing well in all auditory contexts: it is more important to hear well in contexts that involve social interactions. We show that hearing aid outcomes measures are moderately correlated. By leveraging on these correlations, we propose a method of combining measurements of hearing aid outcomes into a single score to reduce measurement error. Finally, we show that it is possible to discriminate between poor and good hearing aid outcomes with an accuracy of 78% solely based on auditory contexts and hearing aid features. This shows the central role that auditory contexts play in understanding hearing aid outcomes in situ.

1. INTRODUCTION

A 2008 MarkeTrak survey estimates that 11.3% of Americans (approximately 34.25 million) suffer from hearing loss [12]. Left untreated hearing impairment affects communication and can contribute to depression, anxiety, isolation, paranoia, and, possibly, dementia [1, 14–16]. The primary intervention for sensorineural hearing loss and related psychosocial consequences is hearing aid (HA) amplification. However, in spite of significant advancements in HA technology during the past decade, HA use is not prevalent among people with hearing loss [9, 12] and only half of those using HAs are satisfied with their performance in noise [8]. Moreover, several recent clinical studies indicate that the benefit of HA technology (i.e., HA outcome) measured in the laboratory often does not translate to the real world [2,11,17,18]. Therefore, in order to provide better hearing healthcare, there is a critical need to develop assessment techniques that allow researchers and clinicians to understand the factors that affect HA outcomes in the real world.

Measuring HA outcomes in the real world poses significant challenges as it depends on the patient's *auditory context* which includes characteristics of listening partners, listening activities, and acoustic environment. Audiologists typically use interviews and questionnaires to measure HA outcomes. Unfortunately, the accuracy of data collected using survey methods is negatively affected by memory biases as patients are asked to remember the circumstances in which HAs performed poorly long after they occurred. Survey methods are complemented by laboratory-based assessments such as speech recognition tests. During a speech recognition test, a patient placed in a sound-treated booth is presented segments of speech under different noise conditions. As it is extremely difficult to recreate the real world listening conditions in the sound booth, laboratory-based assessments generally fail to be representative of the listening contexts that patients encounter during their daily life. Accordingly, neither self-reporting nor laboratory-based tests are effective in describing the auditory contexts observed by patients in the real world.

In prior work, we have developed AudioSense [5], a novel system for evaluating HA outcomes in the real world using mobile phones. AudioSense includes a mobile phone application that delivers Ecological Momentary Assessments (EMAs). EMA involves the repeated sampling of a subject's current state and experiences in real-time [13]. This is accomplished by delivering electronic surveys either at randomized intervals or when triggered by patients. Compared to other self-reporting methods, EMA has the advantage of reducing memory bias since patients report on their recent experiences (in the previous 5 - 10 minutes). The delivered surveys capture information both the auditory context and the associated HA outcomes.

In this paper, we make the following contributions: (1) We present an empirical study that assess the auditory contexts and their impact on HA outcomes from data collected using mobile phones. We collected a total of 3437 surveys from nineteen subjects using AudioSense to create a detailed record of the auditory contexts that HA users encounter during their daily lives. The scale of our study significantly exceeds previous data collection efforts that used mobile technology. (2) Using the collected dataset, we characterize the common properties of auditory contexts and the importance subjects associate with hearing well in a given context. (3) Research to date shows that HA outcomes show moderate to strong correlations supporting the presence of a latent factor characterizing overall patient experience and opinion. We propose a technique to combine these measures into a single score in order to reduce the measurement error associated with each independent measure. (4) More importantly, we show that it is possible to discriminate between poor and good HA outcomes with an accuracy of 78% solely based on the auditory contexts and HA features. This highlights the central role that auditory contexts play in understanding HA outcomes in situ.

2. RELATED WORK

Several recent clinical studies indicate that the benefit of HA technology (i.e., HA outcome) measured in the lab does not translate to the real world [2,11,17,18]. As a result, there is an increased interest in measuring the prevalence of auditory contexts and HA outcomes in the real world [4,6,17,18]. Ecological Momentary Assessment (EMA) [13] is an established alternative to retrospective self-reporting methods that reduces the problem of memory-bias by collecting data *in the moment*. Additionally, EMA techniques can be implemented using computer technology for scalability. Two computer-based EMA studies have been performed in the field of audiology to date: Henry et al. [6] evaluated the impact of tinnitus on daily lives of people and Galvez et al. [4] assessed patient satisfaction with hearing aids. As part of the ongoing study described in this paper, we have already collected over 3400 surveys, exceeding the scale of the previous computer-based EMA studies in Audiology.

Computer scientists have developed a number of EMA systems [3, 7, 10]. These systems provide a framework that allows for real-time collection of survey and sensor data. However, most often these systems are not deployed as part of clinical or field studies. AudioSense provides similar capabilities to existing EMA systems but emphasizes the collection of data relevant to audiologists such as audio, GPS, and survey data on mobile phones. AudioSense may also replace noise dosimeters (as those used in [19]), which have a larger form-factor to less obtrusive measurement of noise levels in the real world. AudioSense provides the audiologists with a web portal for tracking patient compliance in real-time. The main contribution of this paper is the empirical analysis of the collected data. For the first time, we show that it is feasible to predict HA outcomes based on the characteristics of auditory contexts and HA features. In a broader context, our work contributes to the growing body of literature establishing computer-based EMA as a reliable method for assessing HA technology.

3. AUDIOLOGY APPLICATION

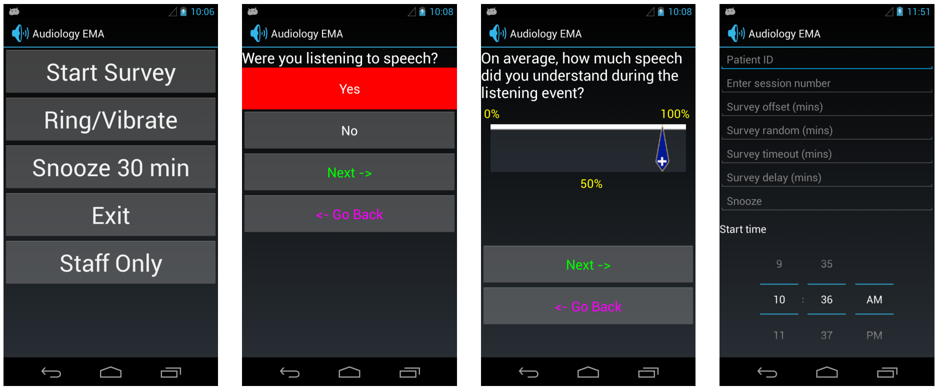
AudioSense integrates mobile phones and web technology to assess hearing aid (HA) outcomes in the real world. We use EMA methods to characterize auditory contexts (e.g., listening activity, room size as compared with an average sized room, and location of speakers relative to the subject) and HA outcomes (e.g., listening effort, speech understanding) associated with these contexts. AudioSense uses a client-server model. The mobile phones are carried by subjects and are used to deliver surveys and collect sensor measurements such as audio signals and GPS location (sensor data is not analyzed in this paper).Mobile phones act as clients in our architecture. The server backend includes two components: a web server and a database. The web server stores the data uploaded by clients in a database. The web server provides a standard web portal interface to visualize the collected data and monitor subject compliance with data collection regiment.

Mobile phones and the web server communicate using HTTP over Wi-Fi or a cellular network. As subjects in our studies are mobile and may live in rural area, wireless connectivity may be intermittent. AudioSense is designed to tolerate intermittent network connectivity by having mobile phones cache the collected data. Periodically, the mobile phone attempts to establish a connection to the web server and, when successful, it uploads the collected data. Note that the storage space available on modern mobile phones is sufficient to store all the data that we collect even in a multi-week deployment. The data is uploaded to the server primarily to track subject compliance.

The client-side of AudioSense runs on Android OS. Android OS is available on numerous mobile phones and tablet computers. AudioSense could be deployed on any Android device. The backend server was portable and could be deployed on Mac OS, Linux, and Windows. The web portal was implemented using the Django web framework. SQLite was used to store data and manage metadata associated with the collected sensor readings and surveys.

The mobile application manages the delivery of surveys. A survey is modeled as a set of questions. To keep track of the subjects' choices at run-time, we associate with each question a variable to which we assign a value based on the response of the subject to that question. A subject has the option to navigate through the survey both forwards and backwards. Answers may be revised as necessary. AudioSense supports adaptive surveys by dynamically determining the next question that will be presented to the user based on his/her previous answers. While the EMA component had an extensible architecture, we currently support two types of questions: multiple-choice questions and scale rating. Multiple-choice questions were rendered as a sequence of buttons whose text could be specified by the programmer (see Figure 1(b)). The subject was allowed to select a single option out of those presented. Scale rating questions were rendered using seekbars and the programmer could provide labels to be rendered for the middle and ends of the bar (see Figure 1(c)).

The delivery of electronic surveys is either alarm triggered or subject-initiated. Alarm-triggered surveys are delivered using randomized schedules. After an alarm is delivered, the time to deliver the next survey is determined by adding a constant time offset T_{offset} to a random number picked uniformly from the time interval $[0, T_{rand}]$. The time to deliver the first survey is determined based on the time when the application is started the first time. The surveys in our field study are delivered on average every 1.5 hours and consecu-



(a) Home screen (b) Multiple choice (c) Scale rating (d) Settings screen

Figure 1: AudioSense mobile application

Variable		Statistics
Gender	Male	35%
	Female	65%
Age(years)	Median: 70.5, Range: 65 – 87	
Hearing loss onset(years)	Median:12, Range: 1– 54	
Employment	Full-time	1
	Part-time	1
	Retired	18
Duration of HA use (years)	Median: 8.5, Range : 0 - 40	

Table 1: Demographic information of subjects

Condition	HA use	DM/DMR usage
0	Unaided	–
1	Entry level	Off
2	Entry level	On
3	Premium	Off
4	Premium	On
5	Reliability measure	
99	Training	

Table 2: Study sessions

tive surveys were separated by at least 1 hour (i.e., $T_{offset} =$ 1 hr and $T_{rand} = 1$ hr). Moreover, in order to minimize the interruption burden to subjects, clinicians could select the time interval during a day when surveys could be delivered. To further mitigate the effects of the survey appearing at an undesired time during the aforementioned interval, a *Snooze* button was provided to delay the alarm by 30 minutes. An alarm outside the delivery interval is postponed until the next day.

Appropriate user interface design can have a significant impact on the compliance of subjects with the data collection protocols. This is particularly problematic given that subjects with hearing loss also tend to be older. Accordingly, they not only suffer from hearing loss but may also have impairments associated with vision and/or fine motor control. These considerations influenced our user interface design choices. We opt for large font sizes and an easy-to-distinguish color scheme. Similarly, we use large buttons and enlarged the default seekbar provided by Android OS. The most important decisions in the user interface are related to the delivery of alarms – notifications that the user should complete a survey. After several iterations and feedback from subjects, we decided to deliver survey alarms by vibrating the phone, playing loud ringtones, and turn on/off the flash of the camera. In order to diminish the intrusiveness and irritation of the alarm, we allowed the subjects to press the power button to stop the alarm and added a *Snooze* option to postpone survey completion by 30 minutes.

4. FIELD STUDY

Nineteen older adults are recruited. The participants are hearing impaired, native English speakers, and at least 65 years old. The participants have adult-onset, bilateral, symmetric (within 15 dB), sensorineural hearing loss with thresholds averaged across 0.5-4.0 kHz between 25 and 60 dB HL. This represents a mild-to-moderate level of hearing loss. Both new and experienced HA users are included. The participants are recruited in two ways: (1) The Department of Communication Sciences and Disorders maintains a subject pool from which people who matched the inclusion criteria are invited to participate in the study. (2) The remaining study participants are recruited through word of mouth from other study participants or through hearing screenings in the community. The sample population is representative of the patients commonly seen in audiology clinics. The detailed demographics are included in Table 1.

Each subject is enrolled in six sessions, each session lasting for a week. The sessions differ in the types of HA devices used and what features are enabled (see Table 2). This is a single-blind study: participants are not aware of what type or features of the HA are active in a given session (but the research team is). To understand the impact of HA technology, we select the following hearing aids: (1) a low-cost, entry-level model (Phonak Bolero Q50) with a low-end adaptive directional microphone (DM) and digital noise reduction (DNR) and (2) a premium level hearing aid (Phonak Bolero Q90) with advanced DM and DNR features. The devices are used with both the DM/DNR features enabled and disabled.

HA outcomes depend on both HA capabilities and the auditory contexts in which HAs are used. AudioSense is used to simultaneously characterize the auditory context and measure the HA outcomes associated with that context. The impact of HA features is evaluated by comparing the results obtained in different sessions. The surveys evaluate the auditory contexts and HA outcomes across multiple dimensions (see Table 3). We leverage on AudioSense's capability to dy-

Context	Variable	Question
Activity context	Activity type	What were you listening to?
	Location	Where were you?
Acoustic context	Noise level	How noisy was it?
	Noise location	Where was the noise coming from?
	Talker location	Where was the talker?
	Room size	How larger was the room?
	Carpeting	Was there carpeting?
Social context	Visual cues	Could you see the talker's face?
	Familiarity	Are you familiar with the talker(s)?
Perception	Speech perception (SP)	How much speech did you understand?
	Listening effort (LE)	How much effort was required to listen?
	HA satisfaction (ST)	How satisfied were your with the hearing aid?
	Sound localization (LCL)	Could you tell where sounds were coming from?
	Loudness (LD2)	Were you satisfied with the loudness?
	Activity participation (AP)	How your hearing affected what you wanted to do?
Importance	Importance	How important was it to hear well?

Table 3: Measured variables

namically determine the next question in the survey based on prior answers in order to reduce the number of questions asked. A typical survey includes a median of 22 questions (range: 12 – 26 questions).

The auditory contexts are evaluated across three dimensions: (1) The activity context captures the type of listening activities (e.g., conversing vs. music listening) and the location of these activities (indoor vs. outdoor). (2) The acoustic context includes elements that affect noise level, location, and degree of reverberation (as determined by room size and presence of carpeting). (3) The social context characterizes the interactions between speakers including visual cues and familiarity. Empirical evidence exists in audiology literature to support that each of these factors may have an impact on HA outcomes. However, as discussed in related work, most of these experiments were not performed using computerized EMA. The HA outcomes are evaluated across multiple dimensions including: listening effort, speech understanding, satisfaction with HAs, the ability to localize sounds, level of loudness, and impact on activity participation.

The first patient was enrolled in the study in February 2013 and the trial is ongoing. By the end of the trial, we will collect data from 50 subjects. The results presented in this paper are based on 3437 surveys collected from the 19 subjects. To the best of our knowledge this is the largest audiology dataset collected using mobile phones. This showcases the feasibility of mobile phones as a data collection platform in field and clinical studies.

5. RESULTS

In this section, we characterize the interplay between auditory contexts, HA features, and HA outcomes based on *real world data*. Specifically, our analysis focuses on the following questions:

- What are the typical auditory contexts subjects encounter in the real world and what is the relative importance they assigned to hearing well in that context?
- Are the HA outcome measures correlated and, if they are, can they be combined into a single HA outcome score?
- Can the HA outcomes be predicted based on auditory contexts and HA features?

Answering these questions will provide a sound basis for understanding some of the factors that affect HA outcomes. This information is valuable to both audiologists that are interested in measurements of HA outcomes in the real world and to computer scientists that are interested in improving EMA systems.

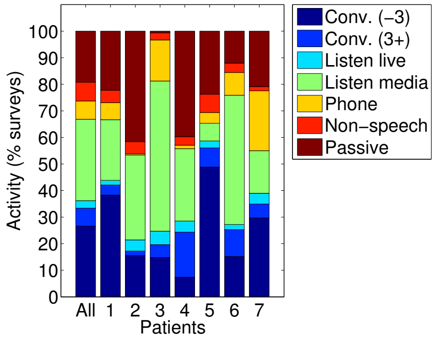
5.1 Properties of Auditory Contexts

We analyzed the distribution of auditory contexts both per subject and over the entire sample, as we are interested in characterizing both the average likelihood of a context and its variation between subjects. The prevalence of a context per subject is the fraction of surveys that the subject indicated to be in that context. The prevalence of a context over the entire sample was computed by averaging the context prevalence over all subjects. Due to space limitations, we focus on listening activities, their locations, and noise level as they have a significant impact on HA outcomes. The subjects rated the importance of hearing well in a given context on a 1 – 100 scale. The analysis presented in this section uses data from all sessions as the HA features have no bearing on context prevalence.

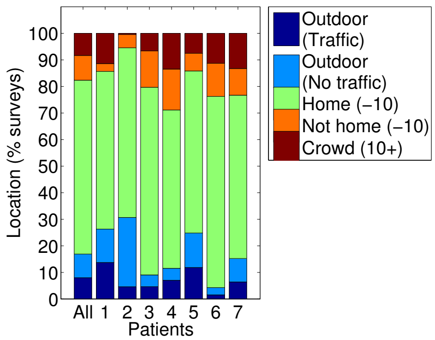
Figure 2(a) plots the activity type for a representative subset of seven patients and for the entire sample (labeled `All` in figures). Subjects spent about 19.2% of the time listening passively. The most common activities are conversations (32.7%) and listening to media (30.7%), accounting for total of 63.4% of the time. The remaining time (17.3%) is spent talking on the phone (6.8%) and listening to live presentations (2.8%) or non-speech sounds (7.1%). Approximately 80% of the conversations involve at most three participants (`Conv. (-3)`), only 20% involving more than three participants (`Conv (3+)`). We observe a significant variability across patients. For example, patient 1 spends 42.1% of his time compared to just 17.2% for patient 2 in conversations. A similar trend may be observed for other activities.

Figure 2(b) shows that subjects spend 16.9% of their time outdoors and 83.1% indoors. About half of the time spent outdoors was spent driving a car (`Outdoor (Traffic)`). Most of the time spent indoors is at home, in the presence of fewer than 10 people (`Home (-10)`). Our subjects spent a significant fraction of time (17.65%) engaging in social activities either outside (`Not home (-10)`) the house or in crowds (`Crowd (10+)`). Similar to the activity type, we observe a significant variation in the distribution of locations across patients.

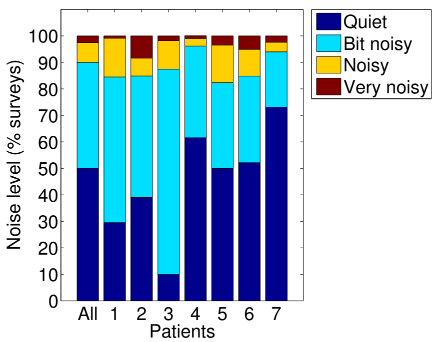
Figure 2(c) plots the noise level reported by subjects. Most of the time subjects report low levels of noise: `Quiet` (50.1%) or `Bit noisy` (39.9%). The low levels of noise can be partly



(a) Activity type



(b) Location



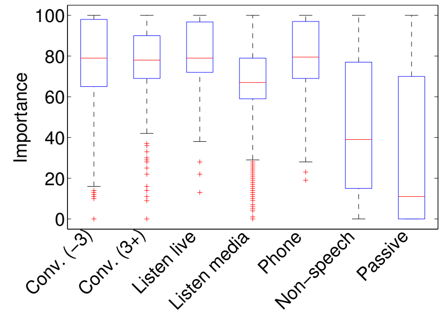
(c) Noise Level

Figure 2: Auditory type, location, and noise level for sample and representative subjects

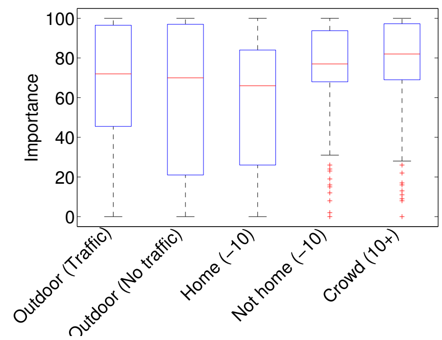
justified by subjects being at home where they can adjust the noisiness of their environment. The propensity of low noise levels is common across all patients.

Result: *Most frequent listening activities were conversations and listening to media, commonly occurring at home, in predominantly quiet environments. Results indicate significant variability between subjects in both listening activities and locations.*

The importance of activity type and location are plotted in Figures 3(a) and 3(b), respectively. The plots show that passive listening or listening to non-speech sounds are associated with low importance ratings. Listening to media is associated with higher importance ratings. In contrast, conversations and listening to live presentations are associated with the highest importance ratings. These insights are corroborated by importance ratings assigned to locations. Most important locations are `Not home` and `Crowd`, where the patient is more likely to be socially engaged.



(a) Activity type importance



(b) Location importance

Figure 3: Importance of activity types and locations

Result: *The importance assigned to hearing well in a context is strongly related to subject's level of social engagement in that context.*

5.2 HA Outcomes Measures

HA outcomes are typically assessed across multiple domains to better understand what factors have a negative impact on the subject's assessment of the HA. Our surveys targeted the following HA outcome dimensions: speech perception, listening effort, loudness, sound localization, HA satisfaction, and activity participation (see Table 3 for details). It is of interest, therefore, to understand the relationships among outcome dimensions. Moreover, if outcomes are correlated, a single aggregated score could be created that would potentially reduce the inherent noise of each dimension. For the analysis presented in this and the following section, we focus on surveys in which subjects reported using a HA and engaging in conversations.

Figure 4 plots the distribution of HA outcome scores using box plots. All scores are continuous variables in the range 1 — 100. A higher score for all variables except listening effort indicates improved HA outcomes Listening effort has an inverse relationship with HA outcomes (i.e., a lower effort indicates better outcomes). For consistency, we adjusted the value of listening effort to be 100 - LE ensuring that higher scores indicate improved HA outcomes for all variables. The median scores were in the range 71 – 86 across all dimensions. The high scores indicate that the subjects

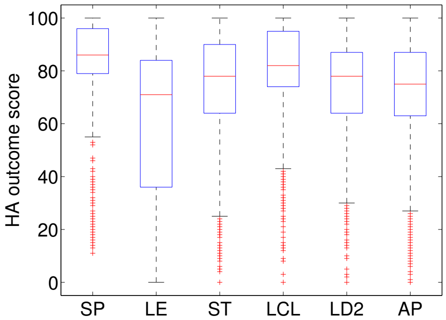


Figure 4: Distribution of HA outcome measures.

	SP	**LE**	**ST**	**LCL**	LD2	AP
SP	1.0000	0.6178	0.6562	0.5847	0.4785	0.5126
LE	0.6178	1.0000	0.5963	0.5029	0.4732	0.6431
ST	0.6562	0.5963	1.0000	0.5477	0.5429	0.5693
LCL	0.5847	0.5029	0.5477	1.0000	0.3451	0.4030
LD2	0.4785	0.4732	0.5429	0.3451	1.0000	0.4989
AP	0.5126	0.6431	0.5693	0.4030	0.4989	1.0000

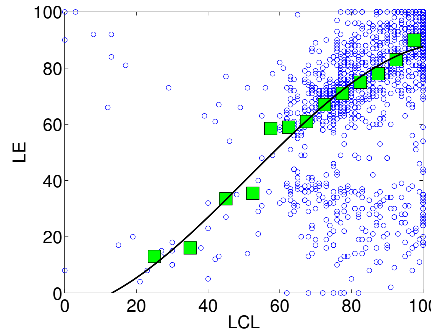
Table 4: Spearman's rank correlation between HA outcome measures. The bolded variables are used to compute a combined HA outcome score.

were overall satisfied with their hearing aids. However, the score variability and presence of outliers indicate that there are contexts in which HA outcomes can be improved.

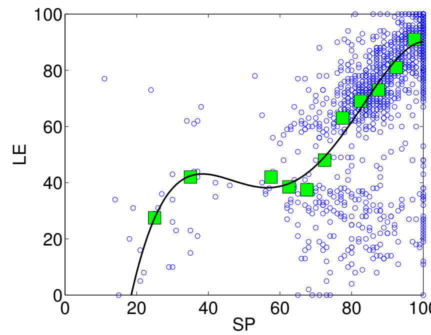
Table 4 shows the Spearman's rank correlation coefficient for the outcome measures. The correlations were computed over the entire dataset (without averaging across patients). Spearman's correlation is used instead of standard Pearson's correlation coefficient, as it does not require variables to have a linear dependence and is less susceptible to outliers. Correlations vary in the range 0.34 – 0.65 indicating fair to moderately strong correlations between outcome measures. This suggests that dimensions measure different underlying aspects of HA outcomes but they are sufficiently well correlated to derive an aggregated score.

We created an aggregated HA outcome score from the four most correlated features: `SP`, `LE`, `ST`, and `LCL`. The first step in creating a combined score is to compute the following three mappings: $f_1 : \texttt{LCL} \mapsto \texttt{LE}$, $f_2 : \texttt{SP} \mapsto \texttt{LE}$, and $f_3 : \texttt{ST} \mapsto \texttt{LE}$. We map `LCL`, `SP`, and `ST` onto `LE` because it has the widest score distribution (as shown in Figure 4), which allows for better discrimination between HA outcomes. The combined score is computed by taking the average of the `LE` score and $f_1(\texttt{LCL})$, $f_2(\texttt{SP})$, and $f_3(\texttt{ST})$.

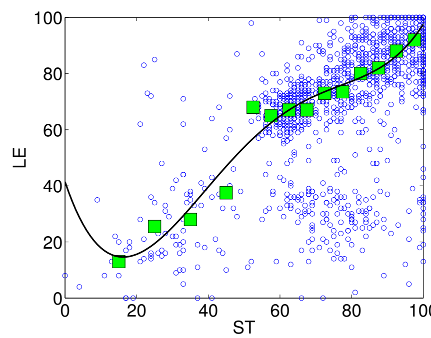
Figure 5 shows the three mappings that we constructed. Each circle represents the `LE` value corresponding to the input (`LCL`,`SP`, and `ST`) in a survey. A key challenge to building such a mapping is to handle the large variability in test scores. The large variability is clear in Figure 5. The mappings were constructed by first dividing the scores into bins over the domain 1 – 100. For each bin, the median `LE` score was determined as indicated by the green squares in the figure. A third degree polynomial was fitted to go through the (x, y) coordinates of the middle of each bin and median `LE` scores (the green squares). Bins that contain only a few values are omitted from the calculation of the best-fit polynomial to achieve robustness to noise. The degree of the polynomial was selected to improve the accuracy of predicting the combined score given auditory contexts and HA features.



(a) $f_1 : \texttt{LCL} \mapsto \texttt{LE}$



(b) $f_2 : \texttt{SP} \mapsto \texttt{LE}$



(c) $f_3 : \texttt{ST} \mapsto \texttt{LE}$

Figure 5: Creating a combined (`CB`) outcome score from `LE`, `SP`, `ST`, and `LCL`

Result: *HA outcome measures are moderately correlated allowing for the computation of a combined HA outcome score.*

5.3 Predicting HA outcomes

In this section, we consider the problem of predicting HA outcomes based on auditory contexts and HA features. An accurate model would highlight the importance of auditory contexts to understanding HA outcomes. Moreover, there are other factors that affect HA outcomes that are not measured in our study (e.g., the comfort of wearing a HA) or not included as part of the model (e.g., level of education). Factors that are not modeled affect the error rates in our model. Therefore, the accuracy of the HA outcomes also quantifies the degree to which the auditory context is characterized well by the selected variables.

The accurate prediction of HA outcomes faces several challenges: (1) The model should incorporate data from all subjects. This is only feasible if we are able to account for individual differences among subjects, some of whom may consistently have more negative evaluations than others. (2) The model must account for the interplay between HA features and auditory contexts. However, the model must be parsimonious to avoid over fitting.

The HA outcome (Y) is evaluated using the combined score introduced in the previous section. All independent variables are nominal. The auditory context is represented by ten nominal variables. The HA features are represented by the nominal variable *session* whose values are given in Table 2. All nominal variables are encoded using dummy coding. The variable D is used to denote the set of dependent variables. We start by modeling the problem as a regression problem where the HA outcome is a continuous variable. Later, we discretize the HA outcome to evaluate the ability of the model to discriminate between poor and good HA outcomes.

The collected data set can be analyzed in the framework of linear model models. A model that models the entire dependence between subjects, sessions, and the variables characterizing the auditory context may be easily defined:

$$Y = \beta + subject \cdot session \cdot \sum_{x \in D} x \qquad (1)$$

where β is the intercept term. However, this model introduces a high number of variables to model the Cartesian product of subjects, sessions, and auditory contexts. As a result a significant number of surveys would be necessary to fit the model. Motivated by this insight, we opted for a more parsimonious model:

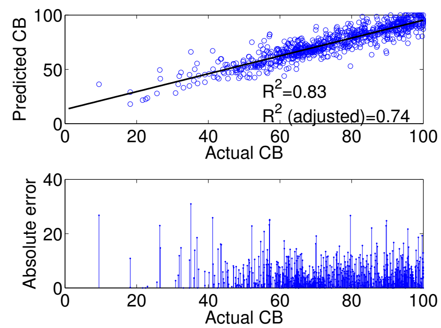
$$Y = \beta + subject \cdot \sum_{x \in D} x + session \cdot \sum_{x \in D} x \qquad (2)$$

The term $subject \cdot \sum_{x \in D} x$ accounts for variations in auditory contexts among patients. Similarly, the term $session \cdot \sum_{x \in D} x$ accounts for variations between HA features.

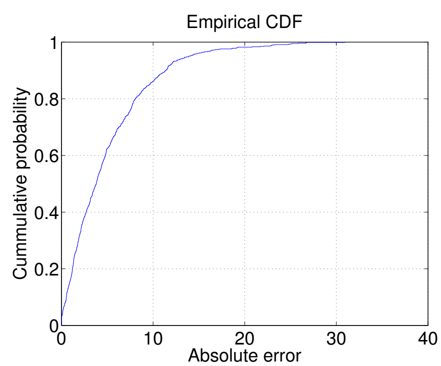
The model described in Equation 2 was further refined using a stepwise procedure to remove terms that are not statistically significant. The procedure removes terms in a greedy manner until the sum of squared errors cannot be further improved. In each iteration, the procedures considers each term in the model and uses an F-statistic test to assess the quality of the model with or without a term. The null hypothesis is that the term has a zero coefficient. If there is insufficient evidence to reject the null hypothesis, the term is removed.

Figure 6 plots the results on the final model obtained from using the stepwise procedure. Figure 6(a) plots the actual versus the predicted combined scores. The line of best fit (plotted in black) clearly indicates a linear relationship between the actual and predicted scores. The high R^2 value supports the goodness of fit of the model to the data.

The cumulative distribution of errors is shown in Figure 6(b). The graph indicates that an absolute error of less than 5 and 10 is achieved 65% and 85% of the time, respectively. This is a positive result as measurements are on a scale 1 – 100. The validity of the model was further evaluated using 10-fold cross validation. The average and standard deviation of the median absolute error across the 10 folds is 6.2 and 1.0882, respectively. Furthermore, we have also investigated the use of non-linear models including support vector machines and neuronal networks. In both cases, the same features as the ones in the linear model were used. The non-linear models did not yield significant improvements in accuracy.



(a) Predictions using linear mixed model



(b) Distribution of absolute errors

Figure 6: Predicting HA outcomes based on auditory context and HA features

To further assess the model's goodness, we considered the problem of discriminating poor versus good hearing outcomes. To this end, we discretized the combined HA score into two classes: good outcomes and bad outcomes. The classes were determined by comparing each score with the median value. Using 10-fold cross validation, linear models were able to discriminate between classes with an accuracy of 78%. Achieving an accuracy of 78% (well above chance) suggest that HA features are indeed essential to accurately predicting HA outcomes. However, it also indicates that there is the potential room for improvement by incorporating other factors in the model.

Result: *The auditory contexts and HA features are essential to understanding HA outcomes. A linear model based on* *auditory contexts and HA features can predict HA outcomes with an accuracy of 78%.*

6. CONCLUSIONS

Hearing aid outcomes depend on both auditory contexts and hearing aid features. Evaluating this relationship in the real world has been tremendously difficult due to the limitations of traditional survey methods. Over the past ten months, we have used AudioSense – a novel hearing-aid evaluation tool – to collect 3437 surveys from nineteen patients. AudioSense uses EMA to characterize auditory contexts and hearing aid outcomes given a hearing aid configuration. The primary contribution of this paper is the empirical analysis of the collected dataset.

Our analysis indicates that most frequent listening activities were conversations and listening to media. These activities commonly occurred at home in a predominantly quiet environment. The results indicate a significant variation in listening activities and locations among subjects. More importantly, subjects associate different levels of importance to hearing well to contexts. We showed that the degree of social engagement given a context determines the importance a subject associates with hearing well in that context. Hearing outcomes are measured across multiple dimensions to understand what factors affect a subject's assessment of HA performance. Our analysis indicates that these measures are moderately correlated. We propose a method that creates a combined outcome score by creating mappings between dimensions using polynomial fitting. The method is designed to tolerate the significant noise observed in real outcome measures. Finally, we show that it is feasible to predict the HA outcomes (measured by the combined scores) based on the auditory context and HA features. A linear model discriminates between good and poor HA outcomes with an accuracy of 78%.

In future work, we will explore approaches to incorporating other variables in the model. We are particularly interested in combining the subjective measures obtained using EMA with objective measures captured from sensors such as microphones and GPS.

7. ACKNOWLEDGMENTS

This work supported by the National Science Foundation (Grant No. 1144664), by the Roy J. Carver Charitable Trust (Grant No. 14-4355), and National Institutes of Health/National Institute on Deafness and Other Communication Disorders (Grant R03 DC012551).

8. REFERENCES

[1] T. H. Chisolm, C. E. Johnson, J. L. Danhauer, L. J. Portz, H. B. Abrams, S. Lesner, P. A. McCarthy, and C. W. Newman. A systematic review of health-related quality of life and hearing aids: final report of the american academy of audiology task force on the health-related quality of life benefits of amplification in adults. *Journal of the American Academy of Audiology*, 18(2):151–183, 2007.

[2] R. M. Cox and G. C. Alexander. Maturation of hearing aid benefit: objective and subjective measurements. *Ear and Hearing*, 13(3):131–141, Jun 1992.

[3] J. Froehlich, M. Y. Chen, S. Consolvo, B. Harrison, and J. A. Landay. Myexperience: A system for in situ tracing and capturing of user feedback on mobile phones. In *MobiSys '07*, pages 57–70, 2007.

[4] G. Galvez, M. B. Turbin, E. J. Thielman, J. A. Istvan, J. A. Andrews, and J. A. Henry. Feasibility of ecological momentary assessment of hearing difficulties encountered by hearing aid users. *Ear Hear*, 33(4):497–507, Jul-Aug 2012.

[5] S. S. Hasan, F. Lai, O. Chipara, and Y.-H. Wu. AudioSense: Enabling real-time evaluation of hearing aid technology in-situ. In *CBMS '13*, 2013.

[6] J. A. Henry, G. Galvez, M. B. Turbin, E. J. Thielman, G. P. McMillan, and J. A. Istvan. Pilot study to evaluate ecological momentary assessment of tinnitus. *Ear and Hearing*, 32(2):179–290, Mar. 2012.

[7] J. Hicks, N. Ramanathan, D. Kim, M. Monibi, J. Selsky, M. Hansen, and D. Estrin. AndWellness: an open mobile system for activity and experience sampling. *Wireless Health*, 2010.

[8] S. Kochkin. Customer satisfaction with hearing instruments in the digital age. *Hearing Journal*, 58(9):30–39, 2005.

[9] F. R. Lin, R. Thorpe, S. Gordon-Salant, and L. Ferrucci. Hearing loss prevalence and risk factors among older adults in the united states. *Journals of gerontology. Series A*, 66(5):582–590, May 2011.

[10] Ohmage. `http://www.ohmage.org`.

[11] J. L. Punch, R. Robb, and A. H. Shovels. Aided listener preferences in laboratory versus real-world environments. *Ear and Hearing*, 15(1):50–61, Feb 1994.

[12] K. S. Marketrak VIII: 25-year trends in the hearing health market. *Hearing Review*, 16(11):12–31, 2009.

[13] S. Shiffman, A. A. Stone, and M. R. Hufford. Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4:1–32, April 2008.

[14] G. Takahashi, C. D. Martinez, S. Beamer, J. Bridges, D. Noffsinger, K. Sugiura, G. W. Bratt, and D. W. Williams. Subjective measures of hearing aid benefit and satisfaction in the nidcd/va follow-up study. *Journal of the American Academy of Audiology*, 18(4):323–349, 2007.

[15] The National Council on Aging. The consequences of untreated hearing loss in older persons, May 1999. Study conducted by the Seniors Research Group.

[16] R. F. Uhlmann, E. B. Larson, T. S. Ress, T. D. Koepsell, and L. G. Duckert. Relationship of hearing impairment to dementia and cognitive dysfunction in older adults. *JAMA*, (261):1916–1919, 1989.

[17] Y.-H. Wu and R. A. Bentler. Impact of Visual Cues on Directional Benefit and Preference: Part I—Laboratory Tests. *Ear and Hearing*, 31(1):22–34, Feb. 2010.

[18] Y. H. Wu and R. A. Bentler. Impact of Visual Cues on Directional Benefit and Preference: Part II—Field Tests. *Ear and Hearing*, 2010.

[19] Y.-H. Wu and R. A. Bentler. Do Older Adults Have Social Lifestyles That Place Fewer Demands on Hearing? *Journal of the American Academy of Audiology*, 23(9):697–711, 2012.