In-situ Measurement and Prediction of Hearing Aid Outcomes Using Mobile Phones

Syed Shabih Hasan[1], Ryan Brummet[1], Octav Chipara[1],
Yu-Hsiang Wu[2], and Tianbao Yang[1]

[1]*Department of Computer Science,*
[2]*Department of Communication Sciences and Disorders*
University of Iowa
Iowa City, IA 52242
{*syedshabih-hasan, ryan-brummet, octav-chipara, yu-hsiang-wu, tianbao-yang*}*@uiowa.edu*

***Abstract*—Audiologists have devised a battery of clinical tests to measure auditory abilities. While these tests can help determine the candidacy of patients for amplification intervention, they do not accurately predict the degree to which a patient would benefit from using a hearing aid (i.e., the hearing aid outcome). Measuring hearing aid outcomes in the real-world is challenging as it not only depends on a patient's auditory abilities, but also on auditory contexts that include characteristics of the listening activity, social context, and acoustic environment. This paper explores the problem of creating predictive models for hearing aid outcomes that incorporate information about auditory abilities, hearing-aid features, and auditory contexts. Our models are built on a dataset collected using a mobile phone application that measures auditory contexts and hearing aid outcomes using Ecological Momentary Assessments. The use of a mobile application allowed us to collect fine-grained hearing aid outcome measures in different auditory contexts. The dataset includes 5671 surveys from 34 patients collected over two years. Our analysis focuses on identifying the features necessary for predicting hearing aid outcomes in different clinical scenarios. Most importantly, we show that models that only included measures of auditory ability as features are cannot predict the hearing aid outcome of a patient with accuracy better than chance. Incorporating information about auditory contexts increases the prediction accuracy to 68%. More excitingly, accuracies as high as 90% can be achieved when a small amount of training data is collected from a patient in-situ. These results suggest that audiologists could prescribe a mobile phone application at the time of dispensing the hearing aid in order to accurately predict a patient's likelihood of becoming a successful and satisfied hearing aid user.**

I. Introduction

Hearing aids (HAs) are the primary method for treating the 11.3% of Americans [1] who suffer from sensorineural hearing loss. Regular use of HAs has been shown to improve communication and avoid the negative effects of hearing loss that include anxiety, isolation, paranoia, and depression [2], [3]. Patients that are candidates for amplification intervention, however, experience different levels of satisfaction with the use of HA in daily life. Patients who are dissatisfied tend to use HAs less frequently limiting their effectiveness [4]. A recent survey indicates that only 59% of HA users are satisfied and regularly use their HAs [5].

Providing audiologists with the ability to identify patients at risk of having poor HA outcomes would help improve the low satisfaction rates of HA users. In the best case, HA outcomes should be predicted from standard measures that are already collected during the battery of tests a patient undergoes to determine his/her candidacy for hearing amplification. Such an approach would be reasonable if a strong relation between measures of auditory ability and HA outcomes existed. Unfortunately, this remains an elusive goal as most of the existing literature points towards the existence of only a weak relationship between auditory ability and HA outcomes [6].

Measuring HA outcomes in the real world is particularly challenging since aside from a patient's auditory abilities other factors contribute to a successful HA outcome. HA outcomes are known to depend on *auditory contexts*, which include the type of listening activity, social context, acoustic environment, and HA configuration. Unfortunately, a majority of existing studies do not capture the auditory contexts in which HAs are used since it would be impractical to do so using retrospective self-reports. A key novelty of this work is the improved methodology that we use to assess HA outcomes. We used a mobile phone application called AudioSense to collect data *in-situ* [7]. AudioSense periodically prompts a patient to describe the auditory context in which he/she is and the perceived performance of the HA in that context. Our dataset includes 5671 surveys completed by 34 patients using four HA configurations collected over the past two years. Additionally, the auditory abilities of each study participant are evaluated using two standard hearing assessments —Pure Tone Audiometry (PTA) and QuickSIN — at the time of enrolling in the study. To the best of our knowledge, this is the first study that predicts HA outcomes based on EMA data that includes auditory context information.

Using the collected data, we analyze the accuracy of predicting HA outcomes based on a patient's auditory abilities, HA configuration, and auditory contexts. We show that a successful HA outcome for a new patient cannot be predicted with odds better than chance based on the results of the

PTA and QuickSIN tests. Incorporating information about auditory contexts, however, increases prediction accuracy to 68%. Collecting a small number of surveys from the patient further improves the prediction accuracy to 90%. Additionally, we have also considered the scenario of a patient switching hearing aids. Specifically, we are interested in predicting the HA outcome for the new HA when data from the previous HA is available. In this case, a successful outcome for the new HA can be predicted with an accuracy of 86%.

The above results highlight the importance of collecting patient information in-situ to predict HA outcomes. More importantly, this points to the feasibility of prescribing a mobile phone application along with the HA. Such an application would allow audiologists to accurately predict the likelihood of a patient becoming a successful and satisfied HA user. Based on the feedback from our application, an audiologist may take some remedial actions to improve the likelihood of success including spending additional time to council patients, suggesting HA that include more advanced features to improve HA benefit, or encouraging participation in aural rehabilitation/training programs. We note that the efficacy of these interventions has not been studied in literature as methods for assessing the patient's likelihood of becoming a HA successful user are still in their infancy.

II. Related Work

Historically, studies of HA performance have been either performed exclusively in the laboratory or combined laboratory tests with survey methods. However, 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 [8], [9], [10], [11]. A potential explanation for the observed differences is that the benefit of HA technology is highly contextual. For example, the presence or absence of visual queues during a conversation can significantly affect the perceived benefit of HAs [10]. Since it is impractical to capture such details accurately using traditional survey methods, some audiologists are increasingly interested in Ecological Momentary Assessment (EMA) [12]. EMA is an established alternative to retrospective self-reporting methods that reduces the problem of memory-bias by collecting data *in the moment*. Computer scientists have developed a number of EMA systems [13], [14], [15]. In previous work, we have developed AudioSense [7] – a system that provides similar capabilities to existing EMA systems but emphasizes collecting data relevant to audiologists such as descriptions of auditory environments and sensor data (e.g., audio, GPS). The use of computerized EMA in Audiology is in its infancy – aside from our prior work, only three other studies have used computer-based EMA methods. Henry et al. [16] and Wilson et al. [17] evaluated the impact of tinnitus on daily lives of people and Galvez et al. [18] assessed patient satisfaction with hearing aids.

Audiologists have evaluated the associations between a number of HA performance indices and HA outcomes. A primary focus has been on evaluating the association between measures that audiologists collect as part of standard practice (e.g., PTA, QuickSin, or Acceptable Noise Level (ANL)) and patient satisfaction. Recent studies show that there is no or weak correlation between auditory ability and HA outcomes [6], [19].

In [19] it was shown that PTA had virtually no correlation with the measured HA outcomes and while a statistically significant correlation existed between outcomes and QuickSIN, it was likely attributed to participant age. Additionally, while ANL has been shown by some studies to be an indicator of real world HA success [20], [21], others have found no link [6]. Our analysis further validates that HA outcomes cannot be predicted accurately based on PTA and QuickSIN test scores.

In previous work [22], we characterized the auditory contexts patients encounter in the real-world and made a preliminary analysis of the relationship between contexts and HA outcomes. Since the focus of the prior work was to show the importance of auditory contexts, the models we considered included patient and HA identifiers as features. As a result, these prior models are not applicable to the important clinical scenarios considered in this paper (when one or both of the identifiers are not available). In this paper, we consider for the first time the use of auditory contexts to predict the HA outcomes of novel patients, novel hearing aids, and novel conditions. Moreover, we show that it is possible to achieve prediction accuracies as high as 90% when a small amount of data in-situ is used. In the broader context, our work points to the feasibility of incorporating computer-based EMA as part of standard practice to improve the successful use of HA.

III. Field Study

Participants for the study are recruited in three ways: (1) the Department of Communication Sciences and Disorders maintains a pool of potential participants and those who match the study criteria are invited to participate, (2) through word of mouth from participants of other studies, and (3) hearing screenings. We recruit adults who are native English speakers and at least 65 years old. The hearing loss of participants is mild to moderate. Our participants are further screened for adult-onset, bilateral, and symmetric sensorineural hearing loss. At the time of analysis, 36 participants completed the study. The demographic details are included in Table I.

Each participant completes six one-week sessions as indicated in Table II. The order in which the participants complete the session is randomized. Each participant started by completing a weeklong training session (condition 99) to get accustomed with reporting data using the mobile phone. For hearing aided conditions (conditions 1 – 4), subjects

Variable		Statistics
Gender	Male	50%
	Female	50%
Age(years)	Median: 73, Range: 65 – 88	
Hearing loss onset(years)	Median:8, Range: 1– 54	
Duration of HA use (years)	Median: 7, Range : 0 - 40	

Table I: 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
99	Training	

Table II: Study sessions

wore a HA for 1 month, followed by a one-week EMA. After the EMA week, they start wearing the next HA (i.e., started the next condition). The participants wore either an entry-level model or a premium level model. Both HAs have adaptive directional microphones (DN) and digital noise reduction (DNR) features. In the remainder of the paper, we will refer to the combination of a patient and HA, or lack of HA, as a condition. A subset of the patients volunteered not to use their HAs for a week either in the beginning or end of the study. The study was single blinded: participants did not know which HA they used.

Hearing Assessments: The auditory abilities of each participant were assessed using PTA and the QuickSIN tests. The PTA test is designed to assess the hearing loss of study participants. The test consists of presenting pure tones at different frequencies and amplitudes to determine the hearing threshold for a selected set of frequencies. The participants in our study suffer from mild to moderate hearing loss. Accordingly, using PTA we found that patients had a hearing loss of 25 – 60 dB HL in the speech frequency range ($0.5KHz$ – $4.0KHz$). In addition to PTA, the participant's ability to discriminate speech in noise was evaluated using the QuickSIN test. QuickSIN measures the SNR loss of a patient compared to a normal hearing person. The test works by presenting a set of standardized sentences that are corrupted by varying degrees of noise. The test identifies the SNR threshold at which the study participant is able to identify 50% of the keywords in the presented sentence.

Auditory Contexts: The participants used the mobile phone application to record the auditory contexts and associated HA performance. 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 by the first time the application is started. The surveys in our field study are delivered on average every 1.5 hours and consecutive 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 burden of 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. Additional details on AudioSense may be found in [7].

The mobile phone application characterized environments in an exhaustive manner along three dimensions: activity context, acoustic context, and social context. Each of the various characteristics of the environment has been previously shown to impact HA performance either in the laboratory or the real-world. Table III summarizes the questions AudioSense asks the user. We note that in order to minimize the reporting burden of our participants, the application only presents the questions relevant to current auditory context of the participant.

HA Outcomes: The application asked participants to evaluate the HA performance in the last 5 — 10 minutes prior to the delivery of the survey. The evaluation is performed across multiple dimensions. In Section IV-A, we show that scores of each dimension may be combined to create a single combined score that characterizes the HA outcome for a given context. We will refer to this HA outcome measure as the *momentary* HA outcome. The *aggregate* HA outcome (or simply the HA outcome) of a condition is measured by the average of the momentary HA outcomes. Existing studies indicate that the relationship between aggregate EMA measures and HA satisfaction obtained via surveys to be inconsistent and variable [12]. The agreement among researchers is that aggregate measures captured via EMA are indicative of the actual experience whereas surveys measure the participant's perception [23].

Data Included: The data analyzed includes only the conditions when the HA were used, excluding data from the training and the unaided conditions. Additionally, as part of every survey (including those delivered during aided conditions) the patient is asked to confirm that they are using their HAs. The surveys in which participants indicated that they did not use a HA are excluded from the analysis. Two participants out the 36 were excluded due to low response rates. The resulting dataset includes 34 patients using four different hearing configurations for a total of 136 conditions. The dataset includes a total of 5671 surveys, each condition including 41.7 surveys on average (range: 7 – 121).

IV. Results

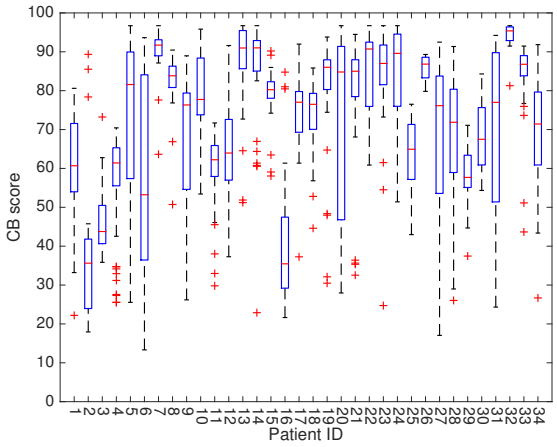
In this section, we characterize the accuracy of predicting HA outcomes based on laboratory test scores, HA configurations, and information about auditory contexts. We are

Dimension	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)?

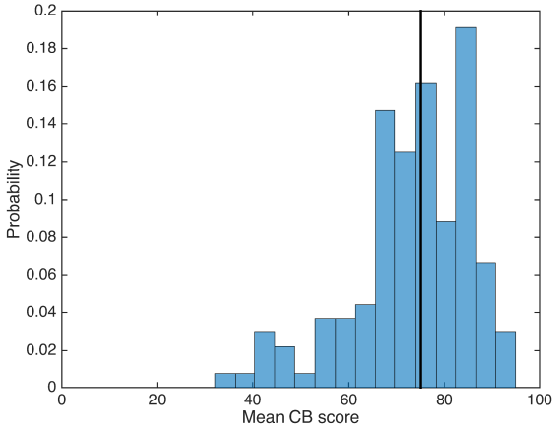
Table III: Features included as part of auditory contexts

Dimension	Variable	Question
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 IV: Measured outcome dimensions



(a) Per patient distributions CB scores (condition=1)



(b) Distribution of $\overline{\text{CB}}$ scores

Figure 1: Statistics of CB and $\overline{\text{CB}}$ scores

	SP	**LE**	**ST**	**AP**	LCL	*CB*
SP	1.00	0.62	0.57	0.47	0.47	0.77
LE	0.62	1.00	0.61	0.64	0.51	0.89
ST	0.57	0.61	1.00	0.64	0.40	0.84
AP	0.47	0.64	0.64	1.00	0.32	0.83
LCL	0.47	0.51	0.40	0.32	1.00	0.48
CB	0.77	0.89	0.84	0.83	0.48	1.00

Table V: Spearman's rank correlation between different domains of HA performance

interested in assessing both the performance of different machine learning algorithms and understanding what are the features that are necessary for making accurate predictions. We consider the following clinically relevant scenarios that differ in the information available for training and predicting HA outcomes:

- **Novel patient:** A new patient is considered for hearing amplification and her/his likelihood of becoming a successful HA user is assessed using data from other patients that use the same or a different HA.
- **Novel HA:** A patient is prescribed a new HA and his/her HA outcome is predicted using the data collected while using the old device. We consider the cases when there are and when there are no other patients that have used the newly prescribed HA.
- **Novel auditory context**: The momentary HA outcome in a novel auditory context is predicted when there is information about the patient's use of her HA. This may help clinicians identify the auditory contexts in which a patient has a difficulty hearing.

The remainder of the section is organized as follows. In Section IV-A, we consider the problem of creating a single combined score from multiple HA performance measures. The score is then used to determine whether a patient will become successful a HA user or not. The different models used for predicting HA outcomes are described in Section IV-B. The results of applying the models in the context of the above scenarios are presented and discussed in Section IV-C.

A. Measuring HA Outcomes

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 measure HA outcomes along six dimensions: speech perception, listening effort, loudness, sound localization,

HA satisfaction, and activity participation (see Table IV). The correlations between performance domains are included in Table V. Most performance domains have moderate correlation indicating that they may be combined to create a single momentary HA outcome score. An advantage of this approach is that by combining scores the inherent noise associated with measuring each dimension is reduced.

In prior work [7], we have proposed a method for creating a combined score (CB). CB is computed in two steps using the most correlated measures: SP, LE, ST, and AP. The first step in creating a combined score is to construct the following three mappings: $f_1 : \texttt{SP} \mapsto \texttt{LE}$, $f_2 : \texttt{ST} \mapsto \texttt{LE}$, and $f_3 : \texttt{AP} \mapsto \texttt{LE}$. We map SP, ST, and AP onto LE because it has the widest score distribution, which allows for better discrimination between HA outcomes. The combined score (CB) is computed by taking the average of the LE score and $f_1(\texttt{SP})$, $f_2(\texttt{ST})$, and $f_3(\texttt{AP})$. The functions f_1, f_2, and f_3 are third degree polynomials whose coefficients are determined using robust fitting.

Audiologists do not have an objective standard for differentiating between successful and unsuccessful HA users. Different methods have been used in the field such as defining a minimum HA usage period per day [24], [25] or using a threshold over an aggregate score [6]. CB is a measure of the *momentary* HA outcome of a patient, wearing a HA, in a specific auditory context. We consider a condition (i.e., a patient using a given HA configuration) to be successful if the *mean* CB scores of that condition is higher than a threshold that is determined such that the top-half of conditions are successful while the bottom-half unsuccessful. We will use the notation $\overline{\texttt{CB}}$ to denote the mean CB score of a condition.

A key challenge to accurately predicting the HA outcome is the high variability of CB scores. Figure 1a plots the distribution of CB scores per patient for condition 1. The boxplots clearly indicate that the distribution of CB scores varies significantly between patients, many patients having a wide distribution of scores. The significant variability in HA outcome scores may be partially explained by the differences in the auditory context. Figure 1b plots the distribution of $\overline{\texttt{CB}}$ scores (mean 73.2, standard deviation 12.3). The distribution suggests that it might be easy to discriminate the outcome of conditions at opposite ends of the scale, but this task would be particularly challenging close to the threshold $\overline{\texttt{CB}} \approx 76$ (indicated in the Figure 1b as a black vertical bar) that separates successful and unsuccessful conditions.

B. Models and Algorithms

We have evaluated the use of linear models, mixed models, and bagged trees to predict HA outcomes. The choice of model is motivated by our desire to explore models of different complexity and modeling assumptions.

The linear models that we use have the general form:

$$\texttt{CB}_i = \beta_0 + \sum_{f \in F} \beta_f I[f] + \epsilon_i$$

where i is the index of observation, F represents the set of features included in the model, and I is the indicator function. The residuals ϵ_i are normally distributed with zero mean and variance σ^2 ($\epsilon_i \sim \mathcal{N}(0, \sigma^2)$). The fitting process determines the β parameters. A key challenge to fitting the linear model is to determine what features to include in the model. The set of features F is determined through step-wise regression by incrementally adding features to the model until no further improvement is possible. The quality of the models is evaluated using t-tests.

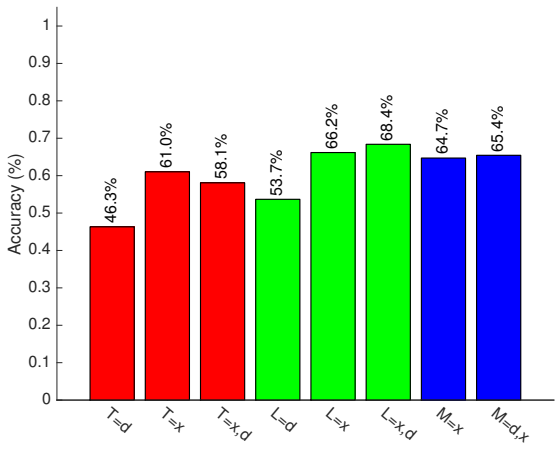
Mixed models have been successfully applied to characterize multi-level data. We may view the dataset as having two levels that cluster data within patients and patients within conditions. Mixed effect models allow us to construct models that reflect the dependencies of the data associated within the same statistical unit. The model has the general form:

$$\texttt{CB}_{i,p,h} = \sum_{f \in F} \beta_f I[f] + \sum_{p \in P} a_p \Pi_p + \sum_{(p,h) \in C} b_{p,h} \Gamma_{p,h} + \epsilon_i$$

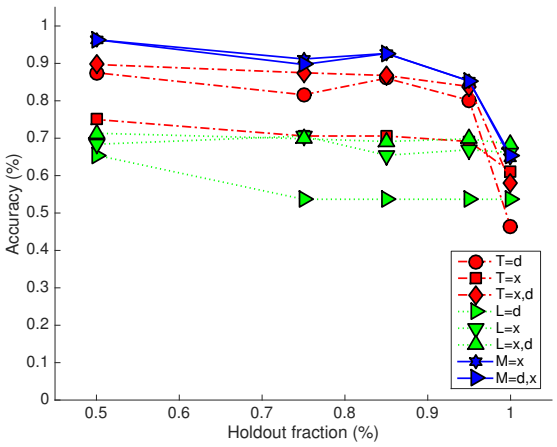
where i is the observation index and indices p and h represent the patient and HA configuration of the i^{th} observation. The sets P and C include the patients and conditions of the study, respectively. In addition to the fixed effects coefficients β_f that are fitted similarly to the linear regression, a mixed model also includes random effects. The matrix Π represents the patients and matrix Γ the conditions that have patient p nested in HA configuration h. The fitting procedure determines the random effect coefficients a_p and $b_{p,h}$. The procedure constrains the parameter vectors a_p and $b_{p,h}$ to be normally distributed such that $a_p \sim \mathcal{N}(0, \sigma_p^2)$ and $b_{p,c} \sim \mathcal{N}(0, \sigma_{p,c}^2)$. A similar procedure to the one described for linear models is used to select the features that will be included in the model. Specifically, new features are added to F as long as the model is improved while the random structure of the model is fixed. For a review of linear mixed models, we refer the reader to [26].

The last learning algorithm considered is bagged ensemble of regression trees. An advantage of bagged regression trees is that unlike the linear models they have built-in feature selection. The bagging algorithm improves the overall performance of regression trees by repeatedly sampling the training data and constructing multiple regression trees. We iteratively add more trees to the model until the improvement of out-of-bag error falls below 1%. The out-of-bag error has been shown to be a good indicator of the generalization error of the algorithm.

The three algorithms predict the CB score as a continuous response variable. To simplify the interpretation of results, in the case of novel patients and HA, the continuous predictions



(a) Classification accuracy for novel patients



(b) Accuracy improvements when some novel patient surveys are used for training. Holdout fraction of 1 is equivalent to 2a.

Figure 2: Accuracy of predicting HA outcomes for a novel patient using data collected from other patients

are discretized. This is accomplished by computing the mean of all predictions associated with a condition (i.e., the predicted $\overline{\texttt{CB}}$). The condition is predicted to be successful if the predicted $\overline{\texttt{CB}} \geq 76$; otherwise the condition is unsuccessful. The reader may refer to Section IV-A for the methodology used to determine the threshold value.

Each model is fit using different information to assess which features must be included to achieve high accuracy. Laboratory tests include the results from the PTA and QuickSIN tests. The contextual information includes all the survey information collected using AudioSense (see Table III). We note that both the laboratory tests and the auditory contexts include 6 continuous variables and 40 dummy variables that encode contextual information, respectively. Additionally, some models include statistically relevant interaction terms to capture the interaction between pairs of features. Models are labeled using the convention `model=features`, where the `model` may be linear `L`, mixed model `M`, or bagged regression tree `T`. The features may include laboratory tests (`d`), auditory context features (`x`), or both. Baseline models may also include the patient (`p`) and condition (`c`) identifiers when predicting novel context.

C. Empirical Results

In the following, we present the results of applying the models to the three previously discussed scenarios.

1) Novel patient: The most common scenario is that of predicting the HA outcome of a novel patient based on historical information collected from other patients. We evaluate the performance of the machine learning algorithms and models using leave-one-patient-out cross-validation. Accordingly, we consider a patient p and train the model on all the data that does not involve patient p. Using the constructed model, we predict the aggregate HA outcome of patient p using the four HA configurations available in the dataset. This process is repeated for all patients in the dataset. During training, there are $N-1$ patients having information for each of the conditions. We note that the models cannot include features that depend on patient identifiers since directly estimating these features for the novel patient is impossible (as none of its data is included in the training set).

Figure 2a plots the accuracy of predicting the outcome of patients for the different models. The worst performing models are `T=d` and `L=d` that achieve prediction accuracies of 46.3% and 53.7%, respectively. These models include only the results of PTA and QuickSIN tests along with potential interactions between these variables. For these two models, we can predict with odds close to chance whether or not a condition is successful. This result shows that measures of auditory abilities are not predictive of real-world outcome measures of HA success adding to the growing body of evidence that support this conclusion.

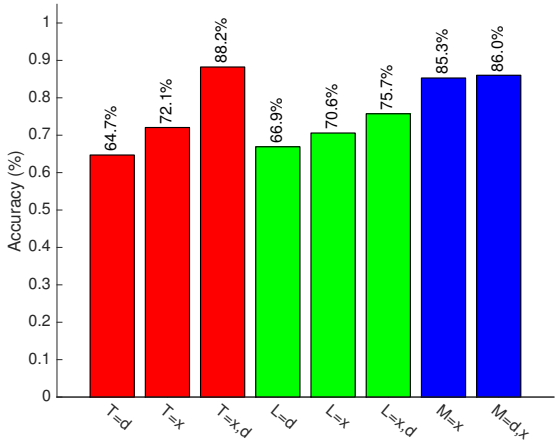
Including information about the different contexts a patient experiences during her/his daily routine significantly improves the prediction accuracy. The prediction accuracy of models `T=X`, `L=X`, and `M=X` is in the range 61% – 66%. A slight increase in prediction accuracy of 1 – 3% may be achieved by combining lab results and context information. These results highlight that HA outcomes cannot be evaluated without understanding the auditory context in which they are measured. Accordingly, audiologists must transition from retrospective surveys measurements to using computerized EMA to capture such information. Furthermore, from a clinical perspective, there is a significant benefit to collect data from a patient in-situ to accurately predict her HA outcome.

To understand the importance of collecting data from a patient, we allowed a small fraction of the patient's data to be used for training the models. The results are shown in Figure 2b. The amount of data withheld for testing varies from 50 – 100%; when the holdout fraction is 100%, the results are the same as the ones discussed above and are shown in Figure 2a. The graph clearly indicates that even a small fraction of patient information can significant increase performance. By moving from including no patient data to including a mere 5% of the data for that patient, the best prediction accuracy jumps from 68.4% to 85%. 5% of the data translates to an average of 2 surveys (range: 1 – 6) that must be completed by the patient. This highlights the importance of collecting personalized information.

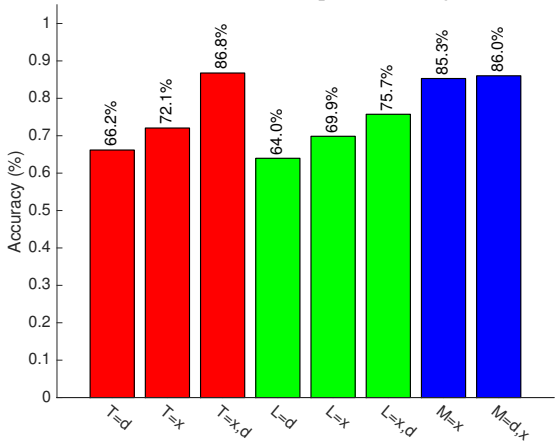
The models that perform best in the case when no patient information is available are the simple linear regression models. However, the performance of these models remains relatively flat as more patient information is used for training. This is because the linear models compute global parameters that ignore grouping the data per patient or per condition. The linear mixed models perform the same as linear mixed models when making predictions for groups that have no data included in the training set. This explains the similar performance of linear and mixed models when the all data of a patient withheld. However, as additional information about patients becomes available, mixed models may incorporate this information to make increasingly accurate predictions. Similarly, bagged tree models can increase the number of trees used in the model to achieve slightly worse performance than mixed models.

2) Novel HA: Another important clinical case is what happens when a patient changes their HA device. We consider both the case when there is and when there is no information associated with the new HA device in the training set. The case when no information is available is evaluated through leave-one-HA-out cross-validation. Accordingly, the data associated with a HA configuration is retained for testing while the remaining data is used for testing.

Figure 3a plots the accuracy of predicting HA outcomes when no patient information is available for that patient. We note that this case differs from the novel patient scenario in that the training set includes some data for the considered patient (i.e., when they used the other conditions). As previously observed, the worst performance is that of models that rely solely on laboratory test information. Their best accuracy is 66.8%. Models that include auditory context information perform overall better with a best accuracy of 85.3%. Including the both contextual and demographic information results in increases in accuracy for all three models. However, this increase can be significant: the trees models have an increase of 16.1% to achieve the best accuracy of 88.2%. The higher accuracy in predicting novel HA than novel patients may be attributed to the fact the training set



(a) Novel HA, without other patients using the HA



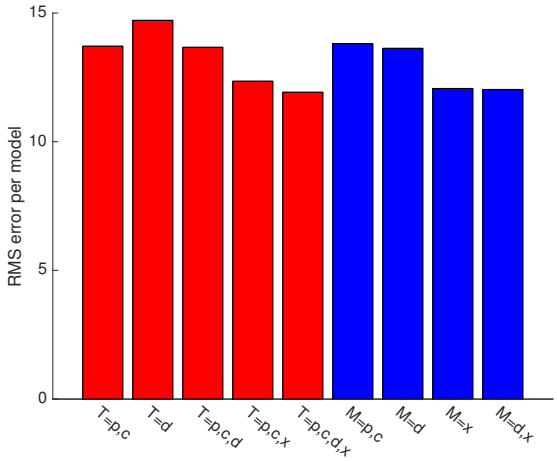
(b) Novel HA, with the other patients having used the HA

Figure 3: Accuracy of predicting HA outcomes when using a novel HA

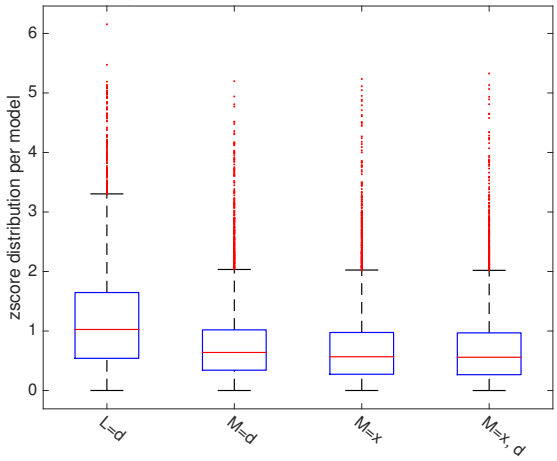
includes patient information that characterizes the auditory style of the patient irrespective of the HA they use. An alternative explanation is that the better accuracy is the result of lower variability induced by different hearing aids compared to the variability induced by different patients.

Figure 3b plots the accuracy of predicting the outcomes for a patient and HA combination. In each experiment, a patient and HA pair is withheld for testing while the remaining data is used for training. Somewhat surprising, the differences in the performance of the models between Figures 3a and 3b are very small. This suggests that in our study there is little that can be gained by considering the scores of other patients that have used the same HA. This result further bolsters the theme that there are significant differences between patients.

3) Novel Contexts: The previous two sections focused on predicting the aggregated HA outcomes ($\overline{\mathrm{CB}}$) for a condition for novel patients or conditions. In this section we turn



(a) RMS error for different models



(b) Distribution of zscores per model

Figure 4: Accuracy of predicting the momentary HA outcomes in novel contexts

our attention to the problem of predicting the momentary rating (CB) that a patient would give to a HA used in an auditory context. For this learning task, it is not sufficient to accurately predict the mean CB score but instead to explain the variability across different auditory contexts. We evaluate the performance of different models and algorithms by using 5-fold cross validation. Each fold is constructed to ensure that data from 4/5 of data of each condition is used for training while the remaining 1/5 is used for testing.

Figure 4a plots the root mean squared error (RMS) for different models. The results indicate that the models that include just information about the patient and condition performs the worst. This is because these models can only predict accurately the average CB scores and are included in the graph as baselines. The models that include only the results of laboratory tests have similar performance to the baselines since they do not characterize the contexts in which HAs were assessed. The models that include contextual information overall achieve better performance showing that it essential to include contextual information if we want to accurately predict momentary HA outcomes. The models that combine both laboratory tests and auditory context information achieve the lowest RMS error.

To get a better understanding of the size of the errors observed for a given patient and condition, we standardize the errors with the respect to the mean and standard deviation of the samples associated with that patient and condition. This is necessary to allow us to aggregate the results across different patients and conditions since these distributions differ significantly in their means and standard deviations. Figure 4b plots the distribution of z-scores for each mixed effect model. Consistent with the RMS errors, the worst performance is observed when only demographic information is included. In this case, the median z-score error is 1 indicating that on average the model makes an error equal to one standard deviation. In contrast, the best performing model that includes information from both lab tests and auditory contexts reduces almost in half. This highlights the need to integrate both features from lab tests and contextual information to achieve high performance.

V. Conclusions

This paper considers the problem of measuring and predicting HA outcomes in the real-world in order to provide audiologists a new method to improve the low satisfaction rates of HA users. Measuring HA outcomes in the real-world is particularly challenging as it is affected by multiple factors including a patient's auditory capabilities, HA configuration, and auditory context. This is the first audiology dataset that jointly measures the auditory context and the associated HA outcomes. Computerized EMA enables us collect fine-grained information about auditory contexts including the type of listening activity, characteristics of the acoustic environment, and their social context. The collected dataset includes 5671 surveys collected from 34 patients using four different HA configurations. The surveys are complemented by laboratory assessments of hearing loss for each patient.

We have analyzed the ability to predict HA outcomes in three clinically relevant scenarios: novel patient, novel HA, and novel contexts. In order to identify the features that are important to achieve high prediction accuracy, we built models with different features and fit them using linear models, mixed models, and bagged trees. Our analysis indicates that we cannot predict the HA outcome of a novel patient with likelihood better than chance using only laboratory measurements of hearing loss. In contrast, incorporating information about the auditory contexts that characterize the auditory lifestyle of the patient increase prediction accuracy to 68.4%. It is possible, however, to achieve accuracy rates as high as 90% when some information about a patient is collected in-situ. We can predict the HA outcome of a patient using a novel HA with an accuracy of 85% leveraging

information about her auditory lifestyle collected using the previous HA. We also provide results for predicting the momentary HA outcome after collecting some data from the user. Our best model can predict the combined HA score with a median error of a half a standard deviation from the condition's mean.

The presented results demonstrate the feasibility of predicting HA outcomes with high accuracy. However, this requires that patients collect in-situ information about their auditory lifestyle (i.e., the auditory contexts) and the associated HA performance. This suggests that a mobile phone application should be prescribed to HA users to determine whether they will become successful HA users. AudioSense is designed for research and, as a result, it introduces a significant data collection burden that cannot be justified outside this setting. In the future, we will explore methods of reducing the data collection burden to enable the development of an application that clinicians may use.

ACKNOWLEDGEMENTS

This work supported by the Roy J. Carver Charitable Trust (grant number 14-4355), by NIH/NIDCD (grant number R03 DC012551), and by the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR, grant number 90RE5020-01-00). NIDILRR is a Center within the Administration for Community Living (ACL), Department of Health and Human Services (HHS). The contents of this paper do not necessarily represent the policy of NIDILRR, ACL, HHS, and the reader should not assume endorsement by the Federal Government.

REFERENCES

[1] K. S, "Marketrak VIII: 25-year trends in the hearing health market," *Hearing Review*, vol. 16, no. 11, pp. 12–31, 2009.

[2] 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*, no. 261, pp. 1916–1919, 1989.

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

[4] M. T. Cord, R. K. Surr, B. E. Walden, and L. Olson, "Performance of directional microphone hearing aids in everyday life." *Journal of the American Academy of Audiology*, vol. 13, no. 6, pp. 295–307, Jun. 2002.

[5] S. Kochkin, "Customer satisfaction with hearing instruments in the digital age." *Hearing Journal*, vol. 58, no. 9, pp. 30–39, 2005.

[6] H.-C. Ho, Y.-H. Wu, S.-H. Hsiao, and X. Zhang, "Acceptable noise level (ANL) and real-world hearing-aid success in Taiwanese listeners," *International Journal of Audiology*, vol. 52, no. 11, pp. 762–770, Nov. 2013.

[7] 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.

[8] J. L. Punch, R. Robb, and A. H. Shovels, "Aided listener preferences in laboratory versus real-world environments." *Ear and Hearing*, vol. 15, no. 1, pp. 50–61, Feb 1994.

[9] R. M. Cox and G. C. Alexander, "Maturation of hearing aid benefit: objective and subjective measurements." *Ear and Hearing*, vol. 13, no. 3, pp. 131–141, Jun 1992.

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

[11] Y.-H. Wu and R. A. Bentler, "Impact of Visual Cues on Directional Benefit and Preference: Part I—Laboratory Tests," *Ear and Hearing*, vol. 31, no. 1, pp. 22–34, Feb. 2010.

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

[13] Ohmage, http://www.ohmage.org.

[14] 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.

[15] 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*, 2007, pp. 57–70. [Online]. Available: http://doi.acm.org/10.1145/1247660.1247670

[16] 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*, vol. 32, no. 2, pp. 179–290, Mar. 2012.

[17] M. B. Wilson, D. Kallogjeri, C. N. Joplin, M. D. Gorman, J. G. Krings, E. J. Lenze, J. E. Nicklaus, E. E. J. Spitznagel, and J. F. Piccirillo, "Ecological momentary assessment of tinnitus using smartphone technology: a pilot study." *Otolaryngol Head Neck Surg*, vol. 152, no. 5, pp. 897–903, May 2015.

[18] 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*, vol. 33, no. 4, pp. 497–507, Jul-Aug 2012.

[19] T. C. Walden and B. E. Walden, "Predicting success with hearing aids in everyday living," *Journal of the American Academy of Audiology*, vol. 15, pp. 342–352, 2004.

[20] M. C. Freyaldenhoven, A. K. Nabelek, and J. W. Tampas, "Relationship between acceptable noise level and the abbreviated profile of hearing aid benefit," *Journal of Speech, Language, and Hearing Research*, vol. 51, pp. 136–146, 2008.

[21] B. Taylor, "The acceptable noise level test as a predictor of real-world hearing-aid benefit," *The Hearing Journal*, vol. 61, no. 9, pp. 39–42, 2008.

[22] S. S. Hasan, O. Chipara, Y.-H. Wu, and N. Aksan, "Evaluating auditory contexts and their impacts on hearing aid outcomes with mobile phones," in *PervasiveHealth*, 2014, pp. 126–133. [Online]. Available: http://dx.doi.org/10.4108/icst.pervasivehealth.2014.254952

[23] T. S. Conner and L. F. Barrett, "Trends in ambulatory self-report: the role of momentary experience in psychosomatic medicine." *Psychosomatic medicine*, vol. 74, no. 4, pp. 327–337, May 2012.

[24] L. Hickson, C. Meyer, K. Lovelock, M. Lampert, and A. Khan, "Factors associated with success with hearing aids in older adults," *International journal of audiology*, vol. 53, no. S1, pp. S18–S27, 2014.

[25] A. K. Nabelek, M. C. Freyaldenhoven, J. W. Tampas, S. B. Burchfield, and R. A. Muenchen, "Acceptable noise level as a predictor of hearing aid use," *Journal of the American Academy of Audiology*, vol. 17, no. 9, pp. 626–639.

[26] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.