Iowa Research Online
The University of Iowa's Institutional Repository

University of Iowa
Iowa Research Online

Theses and Dissertations

Spring 2017

Mobile ecological momentary assessment for hearing aid evaluation

Syed Shabih Hasan
University of Iowa

Copyright © 2017 Syed Shabih Hasan

This dissertation is available at Iowa Research Online: https://ir.uiowa.edu/etd/5494

Recommended Citation
Hasan, Syed Shabih. "Mobile ecological momentary assessment for hearing aid evaluation." PhD (Doctor of Philosophy) thesis, University of Iowa, 2017.
https://ir.uiowa.edu/etd/5494.

Follow this and additional works at: https://ir.uiowa.edu/etd


Part of the Computer Sciences Commons

MOBILE ECOLOGICAL MOMENTARY ASSESSMENT FOR HEARING AID EVALUATION

by

Syed Shabih Hasan

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Computer Science
in the Graduate College of
The University of Iowa

May 2017

Thesis Supervisor: Assistant Professor Octav Chipara

Copyright by
SYED SHABIH HASAN
2017
All Rights Reserved

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Syed Shabih Hasan

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Computer Science at the May 2017 graduation.

Thesis Committee: _____
Octav Chipara, Thesis Supervisor

_____
Yu-Hsiang Wu

_____
Alberto Segre

_____
Joesph Kearney

_____
Juan Pablo Houracde

In memory of Chacha-Abba

ACKNOWLEDGEMENTS

On December 28, 2011 I sent an email to a new faculty member at The University of Iowa's Computer Science Department where I had been admitted in the Ph.D. program expressing my interest in working with him. Little did I know that this email was the beginning of a remarkable journey that is nearing its conclusion with this thesis. I want to thank Professor Octav Chipara, my advisor, for his support, patience, and calm wisdom over the years that have helped me reach goals that I set for myself. We have had our ups and downs but his constant faith in me always helped me, as he put it once, move towards the light. I would also like to thank Professor Yu-Hsiang Wu, my collaborator, for introducing me to the fascinating world of hearing science and his constant support and willingness to indulge all my requests for data. I never imagined that a simple meeting at the annual AMBI symposium in March 2012 would blossom into such an intensive collaboration resulting in my Ph.D. I want to thank my committee members Professors Segre, Kearney, and Hourcade for their advice and support. Everyone in the Mobile Systems Lab at Iowa contributed in some way in my thesis: from critiquing to helping me with my results. I would like to thank Farley, Austin, Ryan , Moosa, and Dhruv. I could not have asked for better lab-mates. I would like to thank the administrative staff of the CS Department: Catherine, Sheryl, and Matthieu thank you for making my Ph.D. journey as comfortable as possible.

Of course, my time at Iowa would not have been the same without the support of my friend here: Vivek, Rahil, James, Sameer, Naveen, Amit, Piyush, Prasad, Jayant,

Akshay, and Shivam. Our discussion made my time at Iowa really memorable. I truly hope that we can someday continue where we left off.

I would also like to thank Starkey for generously supporting some of my endeavours and giving me the opportunity to spend a highly productive summer which gave me very deep insights into hearing science. I would like to thank Dr. Jason Galster, Dr. Sridhar Kalluri, and Swapan Gandhi for being extremely helpful whenever I needed them.

I would like to thank my mentors from Aligarh, Professor. M. Sarosh Umar, and Professor Abdul Qadeer for their guidance and inspiration.

A special thank you is due to my closest friends Arham, Ehraz, Mayank, and Sumaiyah. I cannot describe how much I value your friendship and support over the years. Soon, you too shall have your doctorates and all of us can pretend to be intelligent.

I would like to thank my aunts, uncles, and cousins on both the coasts for opening their hearts and homes to me. You have always made me feel like I never left home.

From the bottom of my heart I would like to thank my parents Drs. Seemin and Abrar Hasan. Their love, joy on my success, consolation on my failures, and always giving me a listening ear have spoilt me beyond repair. There are days when my responsibilities seem overwhelming, talking to the both of you in those moments has given me the strength to keep moving forward. Amma, Abba words cannot describe the magnitude of my gratitude for always being there for me. I would like to thank my

elder brother Dr. S. Saif Hasan for being the rock in my life that I could always lean on. Your encouragement, guidance, and research insights have all made me a better researcher and a better person. With sincerity I pray that Purdue will have a winning football season this year. After this we can all argue in Aligarh on the intended recipient when an invitation for Dr. Hasan shows up.

I want to thank Sarah Zaidi, my sister-in-law, for being the most amazing sister than one could ask for. Thank you for always making me feel welcome and loved, partaking in all my adventures, celebrating my achievements, and just being simply awesome. Massive doses of joy come in little packages. Zoya, you have played a very interesting role in my life. The simple call of "*Moomoo*" makes all the worries of my life disappear and I thank you for that.

Finally I would like to thank my wife, Anam. She has changed me and my life in ways that I never imagined. We have shared laughs, adventures, cooked-in, geeked-out, and made outrageous plans together. She has always listened to my rants, my crazy opinions, supported me when I was down, encouraged me when I was successful. She has made our life, with the limited resources that we have, the most beautiful and comfortable that I could imagine. In this road-trip on the highway of life, nothing gives me more joy than knowing that we are travelling together.

ABSTRACT

Hearing loss can significantly hinder an individual's ability to engage socially and, when left untreated, can lead to anxiety, depression, and even dementia. The most common type of hearing loss is sensor-neural hearing loss that is treated using hearing aids (HAs). However, a significant fraction of individuals that may benefit from using HA do not use them and, the satisfaction of those that do, is only between 60–65%. Today, we have only a limited understanding regarding the factors that contribute to the low adoption and satisfaction rates. This is a limitation of existing laboratory-based assessment methods that cannot accurately predict the performance of HAs in the real-world as they do not fully reproduce the complexities of real-world environments.

There four core contributions of my PhD thesis: i) the development new computer-based methods for assessing HAs in the real-world. Our approach is based on the insight that HA performance is intrinsically dependent on the context in which a HA is used. A context includes characteristics of the listening activity, social context, and acoustic environment. To evaluate this hypothesis, we have developed AudioSense, a system that uses mobile phones to jointly characterize the context of users and the performance of HAs. ii) We provide the first instance of characterization of the auditory lifestyle of hearing aid users, and the relationships that exist between the context and hearing aid outcomes. iii) We utilize the subjective data collected using AudioSense to build novel models that can predict the success of hearing aid prescriptions for new and experienced users. We also quantitatively prove the importance of collecting contextual

information for evaluating hearing aids. iv) We use the objective audio data collected with AudioSense to predict contextual information like acoustic activity and noise level. This provides us a way to intelligently infer contextual information automatically and reduce the burden on the study participants.

PUBLIC ABSTRACT

Hearing loss can significantly hinder an individual's ability to engage socially and, when left untreated, can lead to anxiety, depression, and even dementia. The most common type of hearing loss is sensor-neural hearing loss that is treated using hearing aids (HAs). However, a significant fraction of individuals that may benefit from using HA do not use them and, the satisfaction of those that do, is only between 60–65%. Today, we have only a limited understanding regarding the factors that contribute to the low adoption and satisfaction rates. This is a limitation of existing laboratory-based assessment methods that cannot accurately predict the performance of HAs in the real-world as they do not fully reproduce the complexities of real-world environments.

There four core contributions of my PhD thesis: i) the development new computer-based methods for assessing HAs in the real-world. Our approach is based on the insight that HA performance is intrinsically dependent on the context in which a HA is used. A context includes characteristics of the listening activity, social context, and acoustic environment. To evaluate this hypothesis, we have developed AudioSense, a system that uses mobile phones to jointly characterize the context of users and the performance of HAs. ii) We provide the first instance of characterization of the auditory lifestyle of hearing aid users, and the relationships that exist between the context and hearing aid outcomes. iii) We utilize the subjective data collected using AudioSense to build novel models that can predict the success of hearing aid prescriptions for new and

experienced users. We also quantitatively prove the importance of collecting contextual information for evaluating hearing aids. iv) We use the objective audio data collected with AudioSense to predict contextual information like acoustic activity and noise level. This provides us a way to intelligently infer contextual information automatically and reduce the burden on the study participants.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1

INTRODUCTION

Hearing is an integral part of our being that allows us to interact and experience our environment in a comprehensive way. A decline or loss of this sense can lead to significant changes in a person's lifestyle ranging from minor difficulty in interactions to complete social isolation [9, 41, 61]. The World Health Organization estimates that untreated hearing loss costs $750 billion annually [5]. A common way of treating hearing loss is with the use of hearing aids. Although improvements in performance and benefit have been achieved in with new hearing aids, there still exist issues that need to be addressed. A key issue in this regard is the low adoption rate (approximately 1 in 4), and dissatisfaction with hearing aids [8]. Gaining insights about the underlying reasons for the dissatisfaction requires evaluating the performance of hearing aids. These performance evaluations have typically been conducted in laboratory based settings and sometimes have been augmented by interviews. There are two major drawbacks with such methodologies *viz.* the inability of the laboratory setting to reproduce real-world acoustic contexts, and the introduction of noise in the performance data due to memory bias. The goal of this dissertation is to develop modern tools using mobile technologies for evaluating hearing aids to overcome the aforementioned problems and utilizing the collected data to offer insights into areas of auditory lifestyle, hearing aid prescription success, and identifying contextual information that could potentially lead to the

development of next-generation tools for evaluating and tuning hearing aids.

1.1 Limitations of Traditional Hearing Aid Methodologies

The traditional hearing aid evaluation methodologies have limited to laboratory settings where a hearing impaired individual is exposed to numerous standardized tests such as the Speech-In-Noise (SIN), Hearing In Noise Test (HINT) to evaluate their hearing loss. Tests like the SIN and HINT require the individual to be in a special sound treated room, known as the sound booth. In the booth clean speech signals mixed with different levels of noise to achieve specific Signal-to-Noise Ratios (SNRs) are presented and the individual is asked to perform certain tasks like recall the contents of the presented speech, or doing a secondary task etc. Based on the person's performance on these tests at different SNRs and frequency inputs the hearing loss patterns are recognized and hearing aids are prescribed. Sometimes, these tests are also augmented with interviews about auditory lifestyle where the individual is asked to remember specific details like noise level, source of speech, location etc. which might be helpful in tuning the hearing aid for the person. The two major drawbacks with this methodology are:

1. *Non-representative testing*: The testing conditions presented in the standardized tests create synthetic environments with clean speech signal mixed with synthetic noise like babble, pink noise etc. These synthetic environments might be representative of some real-world scenarios but are incapable of representing a number of auditory contexts that the person might encounter in their daily life. Hence, the results obtained from these tests might not be representative of what the per-

son being tested experiences and hence do not translate well into their daily life through hearing aid prescriptions.

2. *Memory Bias*: Sometimes the standardized tests are augmented with interviews regarding the auditory lifestyle to overcome the non-representation problem. Once hearing aids are fitted based on the laboratory tests, the individuals are asked to come back and report their experience after a few weeks or months. When the user returns, interviews are conducted. The purpose of these interviews is to gain a better understanding of the individual's acoustic lifestyle and use these insights in conjunction with testing results to tune the hearing aid. During these interviews the individual is asked to remember details of events that happened long ago. For example, if the individual reports that they were at a social gathering, the follow-up questions can be related to describing the crowd size, location of noise, reverberation etc. Since a significant amount of time has passed since the event being reported occurred, the individual might not be able to remember exact details and would introduce noise within their responses leading to incorrect tuning of the hearing aids. This phenomenon of misremembering details about events that happened in the past is known as *memory bias*.

1.2 Mobile Ecological Momentary Assessment

An alternative methodology that can be used to evaluate the hearing aids is Ecological Momentary Assessment (EMA). EMA involves the repeated sampling of

the subject's current state and experience in real-time [60]. Specifically, EMA has two important properties associated with the data collection:

- *Ecology*: The data is collected within the ecology of the subject i.e. in their daily life and hence is representative of their real-world experiences.
- *Momentary*: The assessment is delivered *in-the-moment* to capture the data in-situ. This reduces the error caused by memory bias.

EMA, even though a novel idea, had limited effectiveness within the traditional study design. The data collection involving pen-and-paper diary methods were, like traditional studies, difficult to scale and execute precisely. In addition to this, the issue of latency, data being acquired by the researchers only during lab visits of the participants, remains unresolved. However, if we combine the EMA with mobile phone (smartphone) driven sensing, we potentially open new doors for clinical research in the form of mobile EMA. The ubiquitous nature of mobile phones allows the studies to be viable while the presence of embedded sensors helps in enriching the collected contextual data. Scaling the study to many users also becomes a trivially solvable issue.

The overall design of mobile EMA (*m*EMA) studies are fairly straightforward: when the study participant is in the relevant context, an alarm is delivered using the mobile device. While the participant provides their input (usually in the form of filling out a survey), the device also captures sensor data that enriches the survey based contextual data. Once the data collection has ended, the data is uploaded to a remote server

at the earliest opportunity.

In this thesis we create a new *m*EMA system called AudioSense for evaluating hearing aids. The EMA methodology helps us overcome the limitations associated with traditional methodologies and the mobile phone allows us to collect data at unprecedented levels from the perspective of hearing aid research. This system combines the collection of contextual and hearing aid performance data in the form of electronic surveys with objective sensor data like in-situ audio and location. Using this system we have conducted, to the best of our knowledge, the largest academic study on evaluating hearing aids with *m*EMA. Based on the real-time and in-situ data collected with AudioSense were able to make several contributions, details of which are mentioned in the following section. We also present the next-generation of AudioSense capable of streaming objective information from many more sources, and providing greater flexibility for subjective measurements.

1.3 Research Contributions

This thesis makes the following research contributions:

1. **Create a comprehensive data collection platform**: In Chapter 2 we describe our *m*EMA system called AudioSense. This system is capable of collecting hearing aid performance data and contextual information in the form of adaptive electronic surveys in conjunction with recording the acoustic and location information in-situ and in real-time. We show that AudioSense, in contrast with existing *m*EMA systems for hearing aid evaluations, collects much more data due to a

flexible timer and user initiated data collections scheme.

2. **Characterize the auditory lifestyle of hearing aid users**: Chapter 3 describes using the in-situ contextual data to characterize the auditory lifestyle of hearing aid users for the first time. We established that hearing aid users spent most of their time in socially engaging contexts and attached higher importance to listening well in unfamiliar contexts.

3. **Predict the success of hearing aid prescriptions**: In Chapter 4 we utilize the subjective assessments to predict the success of a hearing aid prescription with high accuracy for new and experience hearing aid users. In addition to this we quantitatively prove the importance of collecting contextual information for evaluating hearing aids.

4. **Utilize in-situ audio data to identify subjective responses**: In Chapter 5 we use the objective audio data collected during our study to predict contextual information reported by the study participants *viz.* the activity context and noise level. We further utilize these predictions to test whether they can predict hearing aid outcomes accurately.

5. **AudioSense+, next generation *m*EMA for HA evaluations**: In Chapter 6 we highlight the limitations of the AudioSense system and present AudioSense+ which is the next-generation *m*EMA platform. We highlight how AudioSense+ is capable of capturing a wider variety of data from multiple sources that were

not available previously.

CHAPTER 2

AUDIOSENSE: A MOBILE ECOLOGICAL MOMENTARY ASSESSMENT APPLICATION FOR REAL-TIME HEARING AID EVALUATION

A 2008 MarkeTrak survey estimates that 11.3% of Americans (approximately 34.25 million) suffer from hearing loss [57]. Hearing loss often leads to social isolation that has significant deleterious effects on one's health. For example, hearing loss in older adults has been associated not only with communication difficulties, but also with decreased health and reduced engagement in physical activities [14]. The primary intervention for sensorineural hearing loss and related psychosocial consequences is hearing aid amplification. However, in spite of significant advancements in hearing aid technology during the past decade, hearing aids use is not prevalent among people with hearing loss [40, 57] and only half of those using hearing aids are satisfied with their performance in noise [36]. Moreover, several recent clinical studies indicate that laboratory assessments of hearing aid performance are not predictive of their real world performance [12, 55, 68, 69]. Therefore, in order to improve hearing aids, there is a critical need to develop assessment techniques that allow engineers and clinicians to understand the factors that affect hearing aid performance in the real world.

Measuring the performance of hearing aids in the real world poses significant challenges as it depends on the patient's *listening context* which includes characteristics of listening partners, listening activities, location of conversation partners, and environment. Audiologists currently measure hearing aid performance either through self-

reporting methods or speech-in-noise tests. Self-reports are commonly used to assess the auditory handicap and patient satisfaction with hearing aid performance. Unfortunately, self-reports are plagued by memory biases, as patients are required to remember the circumstances in which hearing aids performed poorly long after they occurred. Speech-in-noise laboratory tests are used to assess the benefits of hearing aids, configure parameters of amplification algorithms, and compare different hearing aid technologies. During a test, a patient placed in a sound booth is presented segments of speech under different noise conditions. As these tests are usually focusing on showcasing the various aspects of hearing aid technology (e.g., use of omnidirectional vs. directional microphones) they fail to be representative of the listening contexts that patients encounter during their daily life. Accordingly, neither self-reporting nor speech-in-noise tests are effective in describing the listening contexts observed by patients in the real world.

In this chapter, we present AudioSense, a novel system for evaluating hearing aid performance in the real world that integrates mobile phones and web technology. The novelty of AudioSense is that it *combines subjective and objective measures of hearing aid performance and listening contexts*. AudioSense uses Ecological Momentary Assessment (EMA) methods. EMA involves the repeated sampling of the subject's current state and experiences in real-time [60]. We make use of EMA with electronic surveys to evaluate both the perceived hearing aid performance as well as to characterize the listening environment (e.g., listening activity, room size, and location of speak-

ers). This is accomplished by delivering electronic surveys either at randomized intervals or when triggered by patients. Compared to other forms self-reporting, EMA has the advantage of reducing memory bias since patients report on their recent experiences (in the previous 5 - 10 minutes). Concurrently with the delivery of surveys, AudioSense further characterizes a patient's listening context by recording their GPS location and sound samples. Standard sound analysis techniques (e.g., computing SNRs) are used to analyze the sound samples after upload to a web server. GPS locations could be used to determine whether the subject is indoors or outdoors. By creating a time-synchronized record of listening performance and listening contexts, AudioSense opens significant opportunities to understand the relationship between listening contexts and hearing aid performance.

The implementation of AudioSense has been evaluated across three dimensions: reliability, energy consumption, and errors in SNR estimation. Experimental results indicate that 100% of the surveys were successfully collected in spite of intermittent network connectivity. Moreover, AudioSense can deliver surveys at 1.5 hours intervals for two days without requiring the mobile phone to be recharged. Finally, we have evaluated the ability of estimating SNR from sound files when various levels of Gaussian noise were added. Preliminary results indicate that the average SNR estimation error was 0.62 dB.

The remainder of the chapter is organized as follows: in Section 2.1 we provide the related work comparing AudioSense with the current methods used for hearing aid

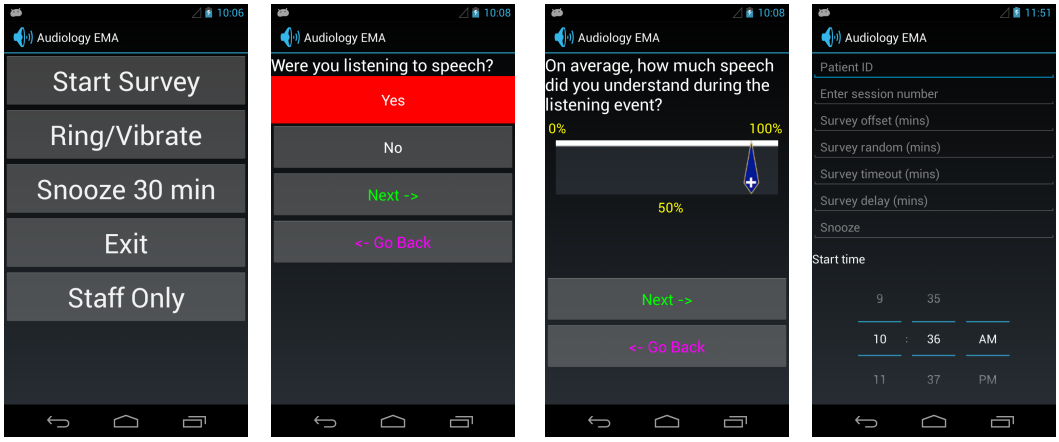
evaluation. Section 2.2 describes the software architecture of the AudioSense system. In Section 2.3 we test the performance of AudioSense from energy consumption and data reliability perspectives. The concluding remarks are made in Section 2.4.

2.1 Related Work

EMA has been proposed as an alternative to retrospective self-reporting methods that suffer from memory bias. A PubMed literature search indicates that only two audiology studies have used computer-based EMA to date. Henry et al. [26] used EMA to evaluate the impact of chronic tinnitus[1] on the day-to-day activities of patients. Galvez [19] used EMA to evaluate patient satisfaction with hearing aid performance. In contrast to the tools used in these studies, AudioSense can track of patient compliance in real-time using a web portal. Galvez reports a compliance rate of 77% in his study. We expect that by tracking patient compliance in real-time, AudioSense may achieve higher compliance rates. More importantly, neither study collects any sensor data to characterize the patient's context.

While audiologists continue to use relatively simple versions of EMA, computer scientists have proposed to combine experience sampling and collection of sensor data to capture contextual information [18, 30]. However, clinicians have not adopted these techniques since they do not include domain-specific measures of contextual information that are necessary to assess their medical relevance. AudioSense addresses this limitation by providing an extensible environment for using algorithms for characteriz-

[1] Tinnitus is the perception of sound in the ear and may interfere with hearing.



(a) Home screen (b) MCQ (c) Scale rating question (d) Settings screen

Figure 2.1. EMA component: Screen shots of (a) the first screen seen by the user, (b) example of the multiple choice questions, (c) example of a continuous scale question, (d) settings screen used by clinicians.

ing the listening context.

Speech-in-noise tests are widely used to assess the benefits of hearing aid noise reduction technologies. Such tests including QuickSIN and Hearing in Noise Test (HINT) present speech and noise at different SNRs. Among the contextual factors that would affect hearing aid users' speech understanding, SNR is probably the most important one. AudioSense already includes algorithms to characterize the SNR of collected speech. In the future, we plan to integrate AudioSense with other algorithms to further classify and characterize listening contexts. We will leverage on the significant body of work on sound classification (e.g., [39]); many of such algorithms are already implemented in MATLAB allowing for a simple integration with AudioSense.

2.2 AudioSense System

AudioSense is designed to collect objective measures of hearing aid performance and listening contexts in the real world. The design of AudioSense must address four key requirements:

1. must facilitate compliance with data collection protocols over multi-week deployments,
2. must ensure the reliability of data collection,
3. must provide an extensible software architecture to enable signal processing and audio analysis on collected sensor measurements, and
4. support concurrent data collection from multiple users.

In the following, we present the system architecture and software components of AudioSense, focusing on how the system addresses these requirements.

2.2.1 System Architecture

AudioSense is a two-tier system that is composed of mobile phones and a backend server. The mobile phones are carried by patients and are used to deliver surveys and collect sensor measurements. The server backend includes three components: a web server, a database, and a speech analysis component. 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 patient compliance with data

collection regiment. The speech analysis component allows the uploaded data to be automatically processed in the MATLAB environment. We opted to integrate with MATLAB to provide a flexible and extensible environment for signal processing and speech analysis. This choice is motivated by the availability of several speech analysis algorithms as open-source components implemented in MATLAB (e.g., VoiceBox).

The communication between mobile phones and the web server is accomplished using HTTP over Wi-Fi or a cellular network. As patients in our studies are mobile and may live in rural parts of Iowa, wireless connectivity may be intermittent. AudioSense is designed to tolerate intermittent network connectivity by having each mobile phone cache the collected data aggressively. Periodically, the mobile phone attempts to establish connections 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.

2.2.2 Software Components

The client-side of AudioSense running on mobile phones is implemented on top of Android OS. Android OS is available on numerous mobile phones and tablet computers. AudioSense can be deployed on any Android device. The backend server is portable and can be deployed on Mac OS, Linux, and Windows. The web portal is implemented using the Django web framework. SQLite is used to store data and manage metadata associated with the collected sensor readings and surveys. MATLAB is used as a computing environment for analyzing collected sensors measurements.

Next, we describe each software component.

2.2.2.1 EMA Component

The EMA component runs on mobile phones and is responsible for managing activities associated with the delivery of electronic surveys. The EMA component addresses the needs of both software developers and patients.

A software developer can create new surveys using a simple API. A survey is modeled as a set of questions. To keep track of the patients' choices at run-time, we associated with each question a variable to which we assign a value based on the response of the patient to each question. A patient may navigate through the survey both forwards and backwards. They may revise their answers as necessary. The next question presented to the patient depends on his previous answers, thus allowing for adaptive surveys.

While the EMA component has an extensible architecture, we currently support two types of questions: multiple-choice questions (MCQ) and scale rating. Multiple-choice questions are rendered as a sequence of buttons whose text can be specified by the programmer (see Figure 2.1(b)). The patient is allowed to select a single option out of those presented. Scale rating questions are rendered using seekbars and the programmer can provide labels to be rendered for the middle and ends of the bar (see Figure 2.1(c)).

The delivery of electronic surveys may be alarm triggered or patient-initiated. The EMA component supports the delivery of surveys using either fixed or a random-

ized schedules. If a survey was just delivered, the time offset until the next survey will be delivered is computed by adding a constant time offset T_{offset} and a random number picked uniformly from the time interval $[0, T_{rand}]$. This method allows for the generation of both fixed schedules (i.e., by setting $T_{rand} = 0$) as well as randomized schedules. Typically, our audiology surveys are delivered on average every 1.5 hours and consecutive surveys are 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 patients, clinicians can select the time interval during a day when surveys can be delivered. An alarm outside the delivery interval will be postponed until the next day.

Appropriate user interface (UI) design can have a significant impact on the compliance of patients with the data collection protocols. This is particularly problematic given that patients with hearing loss also tend to be older. Accordingly, they do not only suffer from hearing loss but also may have impaired vision and potential loss of fine motor control. These considerations influenced our UI designing choices. We refined our initial user design based on patient feedback. Accordingly, we opted for large font sizes and a color scheme that has colors, which are easy to distinguish. Similarly, we opted for a large buttons and overrode the default seekbar provided by Android OS with one that provides a larger area that is sensitive to touch. The most consequential 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 patients, we decided to deliver survey alarms by vibrating the phone, playing loud ringtones,

and turn on/off the flash of the camera. An alarm sounds for 30 seconds. Our choice for an alarm that can be quite intrusive and irritating is balanced by the ability to easily dismiss it: the patient may press the power button to stop the alarm. Moreover, we have added a *Snooze* option that allows the patient to postpone completing the survey by 30 minutes.

2.2.2.2 Sensor Data Collection

While surveys are administered, AudioSense records audio at 16 KHz and GPS locations at 0.1 Hz. The data collection is triggered either by an alarm or when the user opens the application. The data collection is stopped after a timeout configured by the developer.

Unlike the EMA component that utilizes only a fraction of the phone's resources, the design of sensor data collection must minimize resource utilization. To this end, AudioSense implements a simple but effective pipeline abstraction: in a pipeline, the data flows from the source to the sink and is transformed by the intermediary components. Each pipeline is executed in a different thread in order to isolate the data collection from different sources. An additional concern is the need to minimize the number of times the garbage collector is invoked on Dalvik Virtual Machine (DVM). Each time when the garbage collector identifies objects that are no longer used by an application it reclaims the allocated memory. The garbage data collection operation interrupts the execution of the application between 10 – 100 ms depending on the number of objects freed. While most applications would not be affected by this delay, when

high rate audio is recorded, such delay may lead them to drop audio frames. We ensure that the objects used in data collection never need to be garbage collected using the following approach. Each pipeline sources manages a shared buffer pool that contains a number of frames preallocated when the application starts. When a source has data to write, it retrieves a frame from the buffer pool and writes the data into the frame. Frames are pushed down the pipeline through each intermediary component, which receives a reference to that frame. Upon reaching a sink, the frame is put back into the buffer pool of its source. This mechanism of cycling the frames between sources and sinks prevents the frames from being garbage collected since they are always in use.

AudioSense includes three pipelines: audio processing, GPS processing, and file upload. The audio and GPS pipelines have a similar behavior: they collect data from their respective sensor source and save it to a file. Upon the completion of the data collection, the names of files containing the sensor data are passed to the file upload pipeline. The file upload pipeline maintains a queue of the files that are to be uploaded. The content of the queue is saved to disk in order recover from application crashes without losing information according to the following policy. When a new file is added to the queue, the content of the queue is saved immediately to disk to avoid data loss in case of an application crash. In contrast, when a file is removed from the queue, this operation does not result in an immediate write to disk as in the worst case this would lead to a file being uploaded twice. The file upload pipeline dequeues the names of the files and creates HTTP POST request to be sent to the server that includes the file and

additional metadata. The metadata includes a patient identifier, a phone identifier, a session identifier, and the time when the data was collected. Upon a successful upload, the uploaded file is removed from the queue.

AudioSense uses the power-lock interface provided by Android OS to manage its power usage. The EMA component acquires a lock that maintains an active screen at the start of a survey. If the EMA component does not receive any user input for one minute, the survey component is stopped and the screen power lock released. This indicates to the OS that it may turn off the screen if no other application has acquired a power lock on the screen. AudioSense maintains a CPU lock during the collection of sensor data. During the delivery of alarms AudioSense also turns on the camera to access the flash, but it turns it off after the 30 seconds alarm is delivered. In a typical deployment, AudioSense is, on average, active for 10 minutes every 1.5 hours resulting in an 11.11% duty cycle.

2.2.2.3 Web Server Backend

The AudioSense web application is implemented using the Django web framework. Django provides basic facilities for secure website login and user management. The AudioSense web application takes advantage of these capabilities to provide a simple user portal. The primary goal of the user portal is to provide clinicians access to real-time data for monitoring patient compliance.

The web application is responsible for handling the HTTP POST requests from clients. Each HTTP request includes identifiers for the patient and phone from where

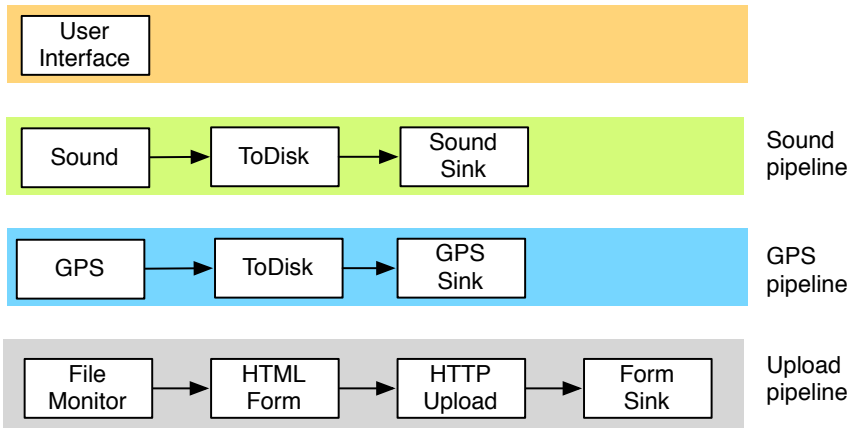


Figure 2.2. AudioSense Data Pipelines: Each horizontal level represents a data pipeline that is used in AudioSense. Individual blocks represent the components used in the pipeline as the data flows from the source to the sink.

the data is uploaded along with the actual data. The metadata is stored in a database for easy querying while the files are stored on the local hard drive. For security purposes, the local hard drive is encrypted. A request also results in a new processing job being added to speech analysis component. The web server may serve multiple clients concurrently.

The speech analysis component integrates with MATLAB environment on the server. This allows AudioSense to be an extensible environment in which many back-end algorithms can be implemented. Currently, we have implemented a number of algorithms for estimating the SNR from collected speech segments. Our focus on SNR is justified by the fact that it is a good indicator listening context.

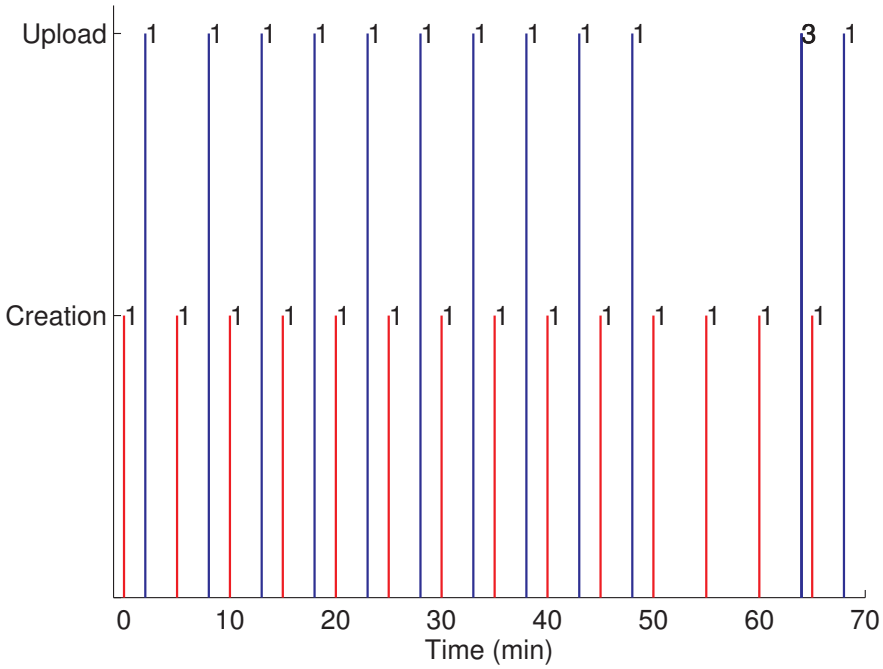


Figure 2.3. Reliability Measurements: The short red bars indicate the creation of the set of assessment files (audio, GPS, and survey) while the following tall blue bars indicate the reception of the files on the server. The numbers next to each bar indicates the number of sets of files created/uploaded.

2.3 Performance Evaluation

The key to successfully deploying AudioSense is to ensure reliable and energy efficient data collection. Accordingly, this section measures the reliability and power consumption of AudioSense under a realistic deployment scenario. These results are complemented by preliminary results from actual field deployments. Additionally, we also evaluated AudioSense's capability of estimating SNR using the MATLAB back-end.

We configured AudioSense to deliver surveys every five minutes. AudioSense

operated as follows: during the first three minutes of each data collection round, AudioSense recorded sound samples and GPS locations. One minute within each data collection round, AudioSense triggered an alarm for the user to complete the surveys. During the experiments, AudioSense recorded sound and GPS locations at 16 KHz and 0.1 Hz, respectively. Under these settings, for a data collection round, approximately 5.46 MB have been recorded and uploaded to a web server.

2.3.1 Reliability

For evaluating data collection reliability, we collected data for 70 minutes during which a total of 15 surveys were delivered. The evaluation was performed inside a home using a Wi-Fi connection to upload the data. Multiple walls attenuated the Wi-Fi connection, which is realistic setup for what we expect in patient homes. Additionally, to evaluate the tolerance of AudioSense to network disconnections, we turned off the wireless adapter on the phone at the 48 minute mark for approximately 12 minutes.

Figure 2.3 captures the reliability of the system during the 70-minute evaluation. The short red bars indicate when the data collection was initiated. As expected, consecutive bars are separated by 5 minutes, which is consistent with the experimental setup. The tall blue bars indicate the time when the data was successfully uploaded on the server. The overall reliability was 100% – all files containing the sound and GPS data have been successfully uploaded to the server.

During the first 48 minutes of the experiment, the phone had connectivity to the server. During this time interval, the average of the time from when data collec-

tion started until it was successfully uploaded was 184.48 seconds. In Figure 2.3, this interval is captured as the distance between consecutive short and long lines. Two factors contribute to the observed delay: a total of 180 seconds were spent collecting the data (per our setup) and the remainder of 4.48 seconds was spent upload the data. On average, the phone uploaded data at a rate of 9.756 Mbit/s.

The phone's wireless interface was turned off during the interval [48, 61] minutes.Without network connectivity, AudioSense cached data from 3 data collection rounds.Upon turning the network interface back on at minute 61, AudioSense proceeded to upload the cached files. The number on top of bar indicates the number of audio files created/uploaded within a 60 second interval. Accordingly, the number 3 on top of the penultimate tall bar indicates that the three sound files that were cached, have been uploaded within a minute.

2.3.2 Power consumption

The power consumption was tracked using the Power Tutor [72]. Figure 2.4 plots the CPU and LCD power consumption that can be attributed to AudioSense. AudioSense records data for 3 minutes during each 5 minute data collection round. This pattern is clearly visible in the figure for both the energy consumed by CPU and LCD: periods of high-energy consumption alternate with periods of no energy consumption. The LCD is used for a shorter period of time than the CPU since AudioSense starts collecting data one minute prior to delivering an alarm and turning on the LCD. During the interval [48, 61] minutes, additional energy is spent by the CPU trying to reestablish

connectivity to the server. Under the considered experimental setup, the AudioSense operates at a duty cycle of 60% and the phone does not need to be recharged for at least a day.

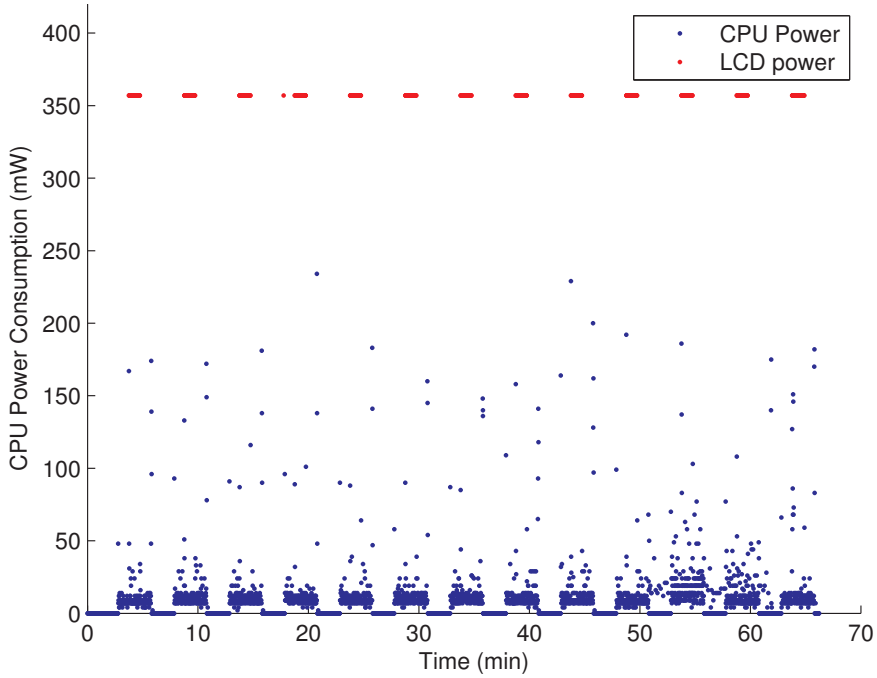


Figure 2.4. Power Consumption: This figure shows the power consumed by AudioSense during testing. The blue dots represent the power consumed by the CPU (responsible for Audio and GPS recording) while the red dots represent the power consumed by the LCD screen (responsible for the survey).

2.3.3 Deployment

AudioSense is being used as a clinical trial that aims at evaluating the effectiveness of hearing aid technology. Currently, AudioSense has been deployed as part of

three weeklong data collection sessions with 5 subjects. In contrast to the experimental setup discussed above, during the field deployment, AudioSense uploads data over the cellular network. Moreover, AudioSense operates at an 11.1% duty cycles being active (on average) for 10 minutes every 1.5 hours. Under this lower duty cycle, AudioSense operates without recharging in excess of three days.

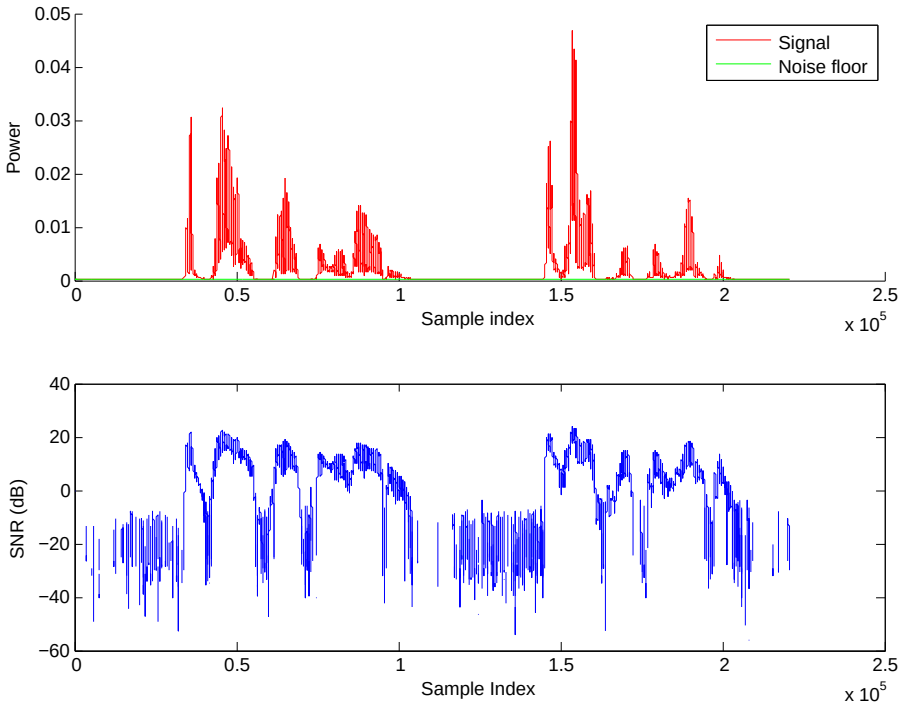


Figure 2.5. Inferring SNR: The figure on the top shows the instantaneous composite signal power in red while the noise floor power is represented in green. The bottom figure shows the SNR calculated from the instantaneous powers.

Actual (dB)	Predicted (dB)
11.52	11.59
10.56	9.76
9.54	9.76
8.56	8.54
7.56	7.750
6.59	7.17
5.57	6.01
4.56	5.11
3.60	5.17
2.56	4.29

Table 2.1. SNR Estimation Accuracy: The left column indicates the actual SNR and the right column indicates the predicted SNR from the captured audio.

2.3.4 Computing SNR

A key factor that determines the difficulty of the listening task is the SNR. We evaluated the ability of AudioSense to compute the global SNR for noisy sound files. The noisy files were generated from a clean sound file to which Additive Gaussian Noise was added.

The SNR was estimated using the signal level and the noise floor from the power spectrum. Figure 2.5 plots the power of the signal and noise levels (red and green curve) for a file where the SNR was 10 db. The instantaneous SNR (computed of 0.65 ms segments) is plotted in the lower part of the graph. Figure 2.1 compares the actual and the estimated SNR values. On average, the SNR error was 0.62 dB but there is a clear trend of increasing error for smaller SNR values. This is expected since for low SNR values it is difficult to distinguish between signal and noise.

2.4 Conclusions

This chapter presented AudioSense a novel system for evaluating the performance of hearing aids in the real world. AudioSense combines EMA techniques with the collection of sensor data to characterize a patient's listening context of the user. To this end, AudioSense integrates mobile phone technology with web applications. AudioSense is capable of delivering customized surveys at fixed or randomized time intervals. User feedback was integrated to refine the design of elements of user interfaces and alarms. Empirical studies show that AudioSense provided 100% reliability, supported the delivery of surveys 1.5 two hours without requiring recharging the mobile phone for two days, and provide facilities to integrate sound analysis techniques. As part of the future work, we plan to infer the listening context automatically in real-time from the audio signals and the GPS locations collected. This would help in making the electronic surveys more intuitive and shorter for the subjects.

CHAPTER 3

EVALUATING AUDITORY CONTEXTS AND THEIR IMPACTS ON HEARING AID OUTCOMES

The auditory lifestyle of hearing aid (HA) users and the corresponding performance of their devices can provide the clinicians with useful insights that can guide the fitting process. Traditionally hearing aid performance has been evaluated in laboratory settings and in some cases has been augmented by interviews or diary methods. The success of these methodologies is severely limited by two factors *viz.* the inability of laboratory tests to accurately present the hearing aid user with real-world scenarios, and the unreliability introduced due to memory bias in the collected data using diary methods. Laboratory tests typically present the hearing aid user with signals consisting speech mixed with different levels of noise. Such tests might be able to partially reproduce certain types of real-world situations within the laboratory such as conversations. It is, however, difficult to introduce a sense of realism within these due to the lack of secondary cues such as lip movement, or the user's familiarity with the surrounds and the people. Hence, it becomes extremely difficult to comprehensively recreate an immersive environment within the confines of a laboratory that is representative of the real-world. Audiologists sometimes augment these tests with interviews and questionnaires to measure real-world HA outcomes. The interviews are generally conducted once in several weeks and are negatively affected by memory bias as users are asked to recall circumstances in which their HAs performed poorly. Accordingly, neither self-

reporting nor laboratory-based tests are effective in describing the auditory contexts observed by patients in the real world.

In the previous chapter, we developed AudioSense [25], 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 [60]. 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 chapter, we make the following contributions:

1. We present one of the first empirical studies that use mobile phones to assess the auditory contexts and their impact on HA outcome. We analysed 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.

2. Using these subjective assessments, we characterize the common properties of auditory contexts and the importance subjects associate with hearing well in a given context.

3. Audiologists evaluate HA outcomes using correlated measures. 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.

The chapter is organized as follows: Section 3.1 introduces the background literature and compares them with AudioSense. This is followed by Section 3.2 that describes the basic setup of the field study from which AudioSense collects the data.

3.1 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 [12,55,68,69]. As a result, there is an increased interest in measuring the prevalence of auditory contexts and HA outcomes in the real world [19, 26, 68, 69]. Most in-situ studies of HA use paper-based surveys methods. Unfortunately, these methods limit both the accuracy of the collected data and the scale of the studies significantly. Ecological Momentary Assessment (EMA) [60] is an established alternative to retrospective self-reporting methods that reduces the problem of memory-bias by collecting data *in the moment*. More importantly, EMA techniques can be implemented using computer technology to

also alleviate the scalability concerns to date. Audiologists have conducted only two computer based EMA studies to date [19, 26]. In [26], the authors evaluate the impact of tinnitus[1] on daily lives of people using EMA, while in [19] the authors use EMA to assess patient satisfaction with hearing aids. This chapter describes utilizes approximately 3400 surveys, exceeding the scale of the previous computer-based EMA studies in Audiology.

Computer scientists have developed a number of EMA systems [18, 27, 49]. 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 [70]), 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 chapter 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 wider context, our work contributes to the growing body of literature establishing computer-based EMA as a reliable method for assessing HA technology.

[1] Tinnitus is the perception of sound in your ear and may interfere with your hearing.

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 3.1. Demographic information of subjects. All participants within our study are older adults from the state of Iowa. All of them have mild-to-moderate hearing loss.

3.2 Field Study

For this chapter we considered nineteen participants. 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 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 3.1.

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

Table 3.2. Different types of hearing aids used in our study. We used two hearing aids each of which had Directional Microphones and Digital Noise Reduction modes which could be turned on/off. Condition 99 represents the practice sessions to familiarize the subject to the data collection procedure.

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 3.2). 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 with a low-end adaptive directional microphone (DM) and digital noise reduction (DNR) and (2) a premium level hearing aid 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

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 3.3. Contextual and outcome measures captured in the AudioSense subjective assessments.

(see Table 3.3). We leverage on AudioSense's capability to dynamically 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 focuses on describing the elements that affect noise level, location, and degree of reverberation (as determined by room size and presence of carpeting). (3) The social context evaluates 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 chapter are based on 3437 surveys collected from the 19 subjects. This showcases the feasibility of mobile phones as a data collection platform in field and clinical studies.

3.3 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 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.

3.3.1 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

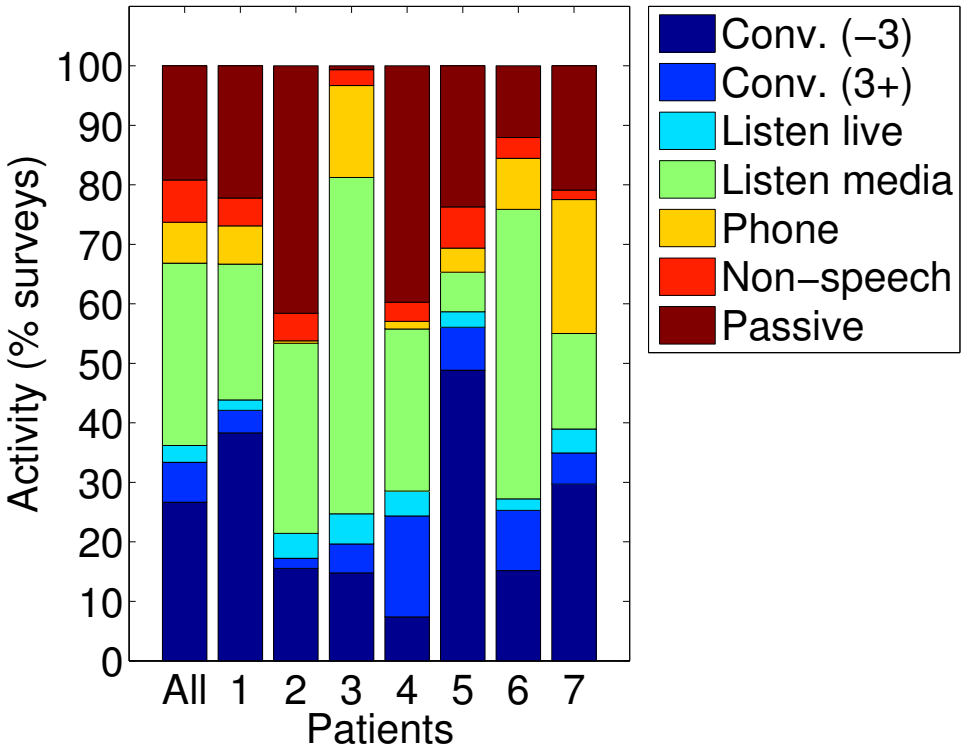


Figure 3.1. Distribution of activity types for a subset of study participants. The first bar indicates the trend across all participants.

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 3.1 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 lis-

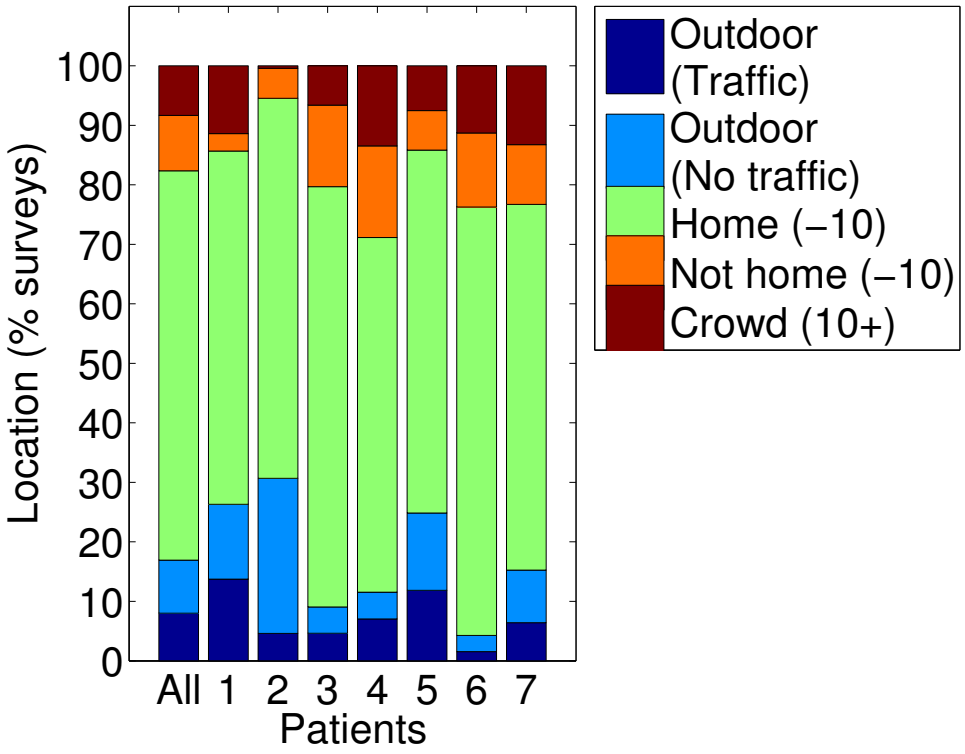


Figure 3.2. Distribution of locations for a subset of study participants. The first bar indicates the trend across all participants.

tening 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 3.2 shows that subjects spend 16.9% of their time outdoors and 83.1%

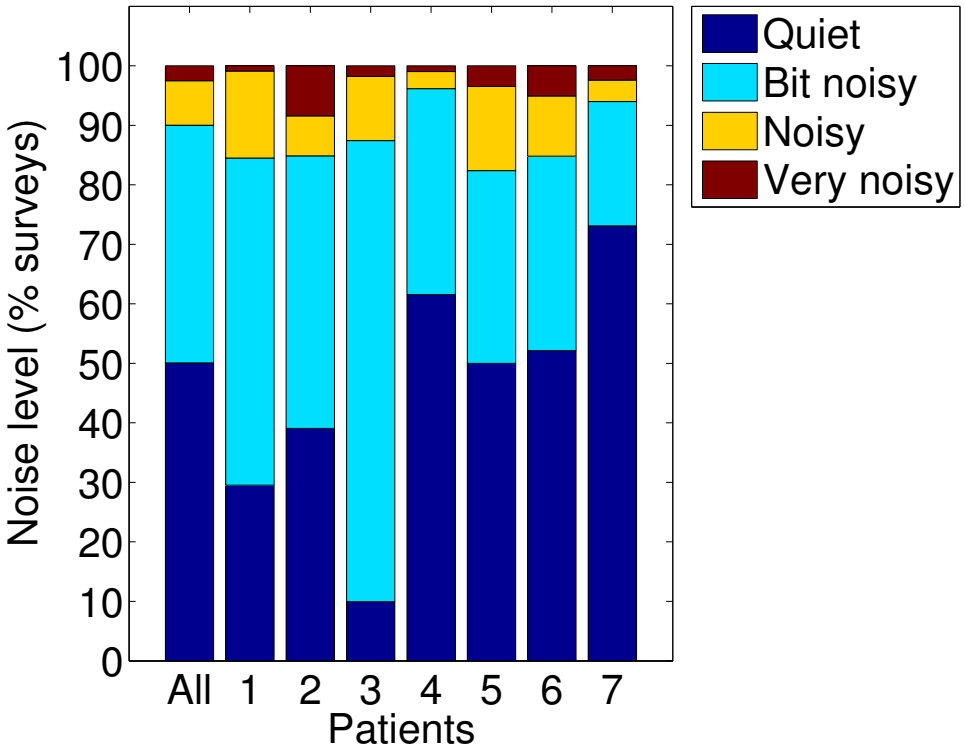


Figure 3.3. Distribution of noise level for a subset of study participants. The first bar indicates the trend across all participants.

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 3.3 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 justified by

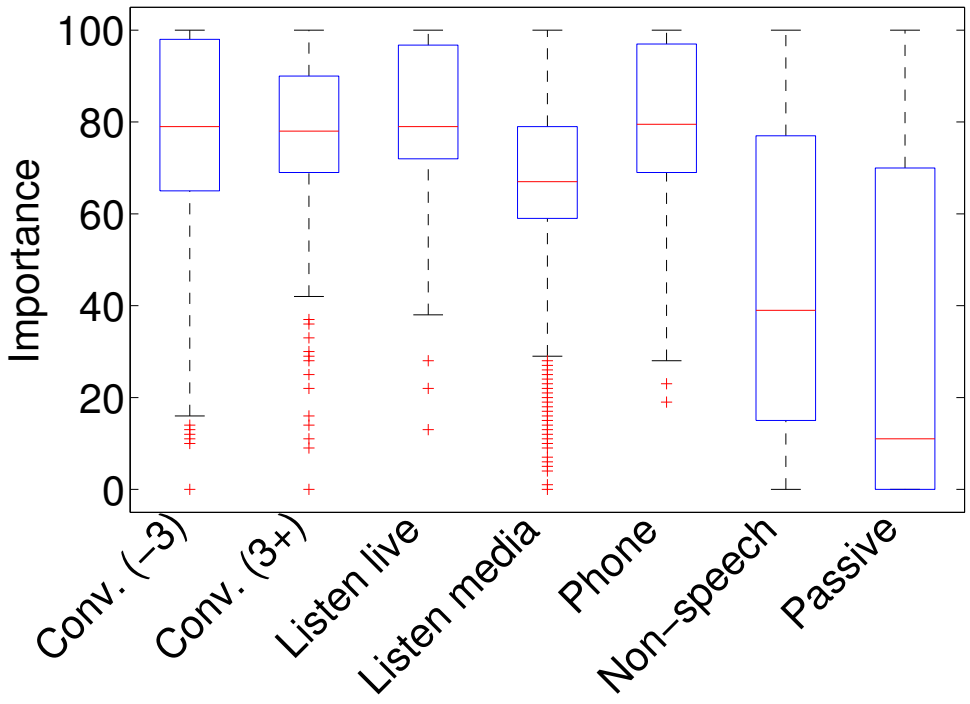


Figure 3.4. Importance of listening well in different activity contexts. Non-engaging activities are relatively less important.

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 3.4 and 3.5, 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

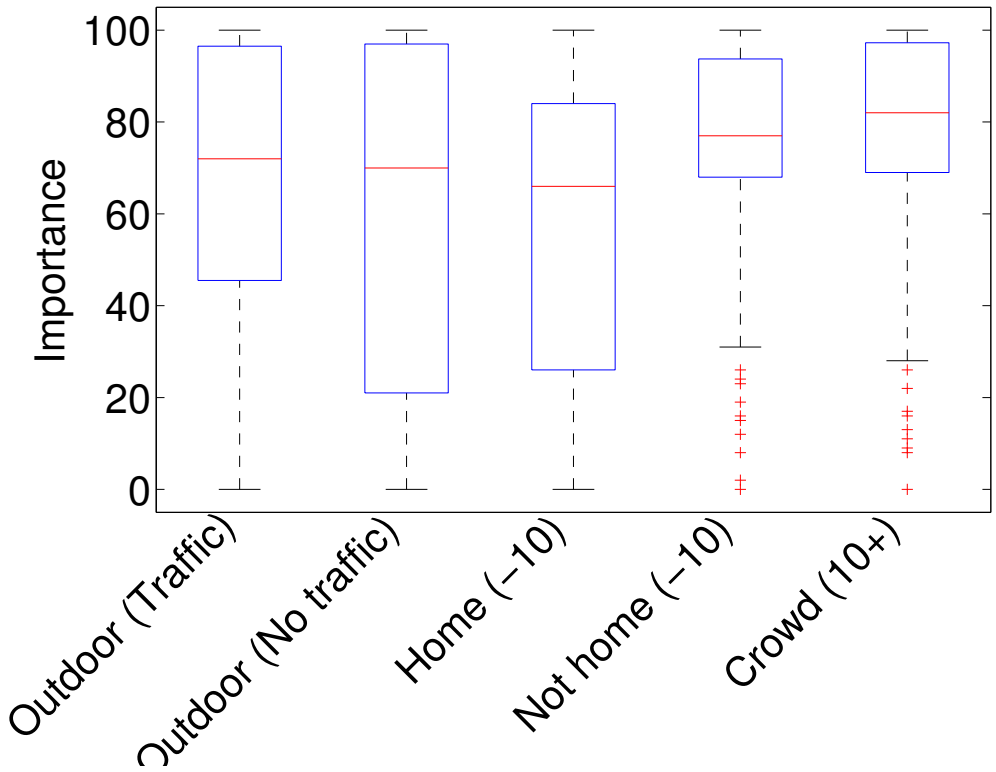


Figure 3.5. Importance of listening well in different locations. Unfamiliar locations are relatively more important.

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.

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

3.3.2 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 3.3 for details). It is of interest, therefore, to understand the relationships between 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 3.6 plots the distribution of HA outcome scores using box plots. All scores are continuous variables in the range 1 — 100; a higher score indicates improved HA outcomes. The median scores were in the range 71 – 86 across all dimensions. The high scores indicate that the subjects 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 3.4 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 low-

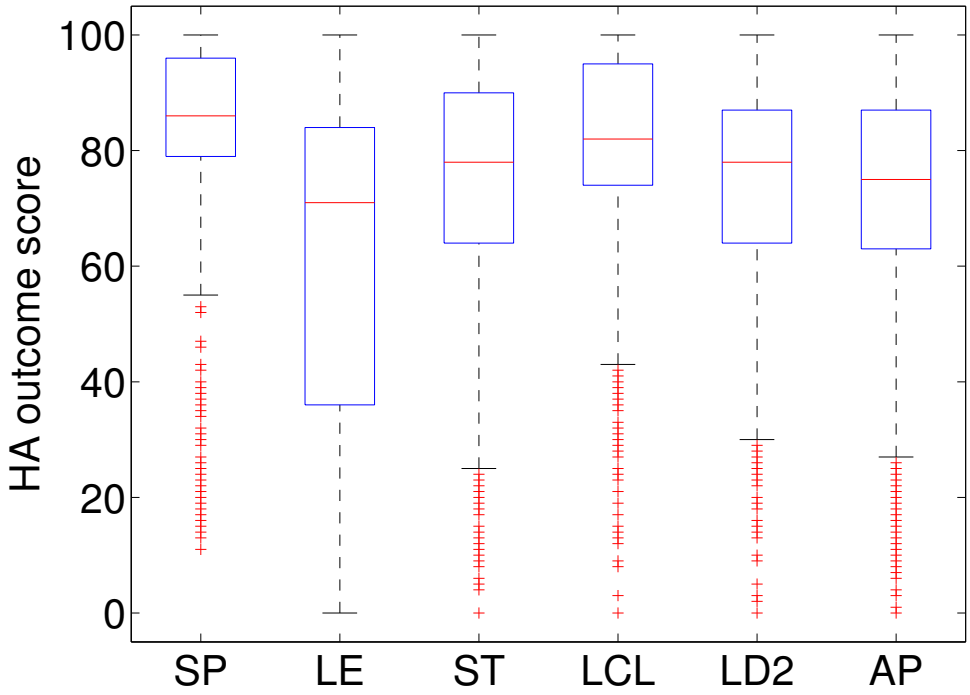


Figure 3.6. Distribution of HA outcome measures.

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

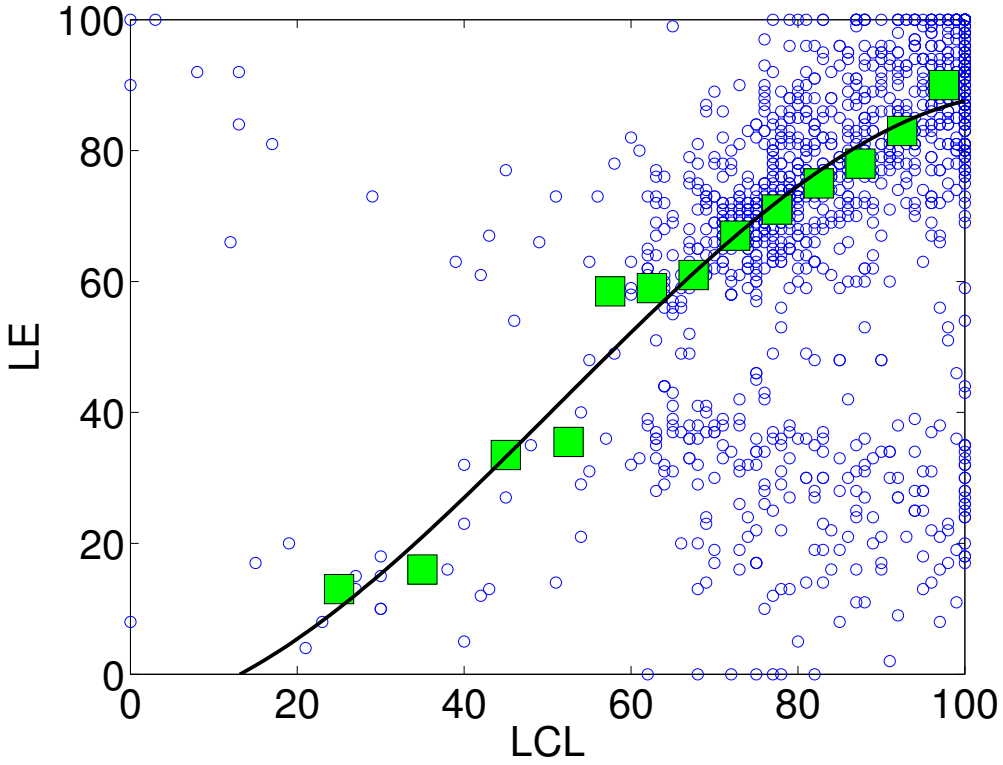


Figure 3.7. f_1 : LCL $\mapsto$ LE, mapping the ability to localize the sound to listening effort.

medium to medium-high 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 : LCL $\mapsto$ LE, f_2 : SP $\mapsto$ LE, and f_3 : ST $\mapsto$ LE. We map LCL, SP, and ST onto LE because it has the widest score distribution (as shown in Figure 3.6), which allows for better discrimination between HA outcomes. The combined score is computed by

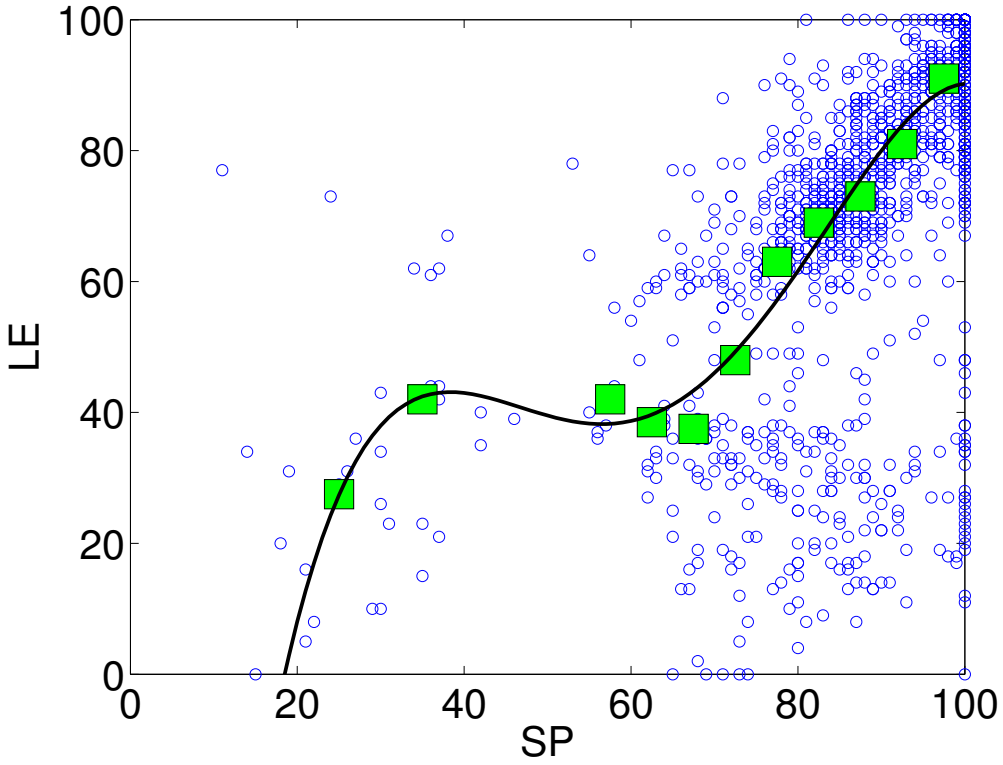


Figure 3.8. f_2 : SP $\mapsto$ LE, mapping the speech perception to listening effort.

taking the average of the LE score and $f_1(\texttt{LCL})$, $f_2(\texttt{SP})$, and $f_3(\texttt{ST})$. Figures 3.7, 3.8, 3.9 show 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 Figures 3.7, 3.8, 3.9. 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

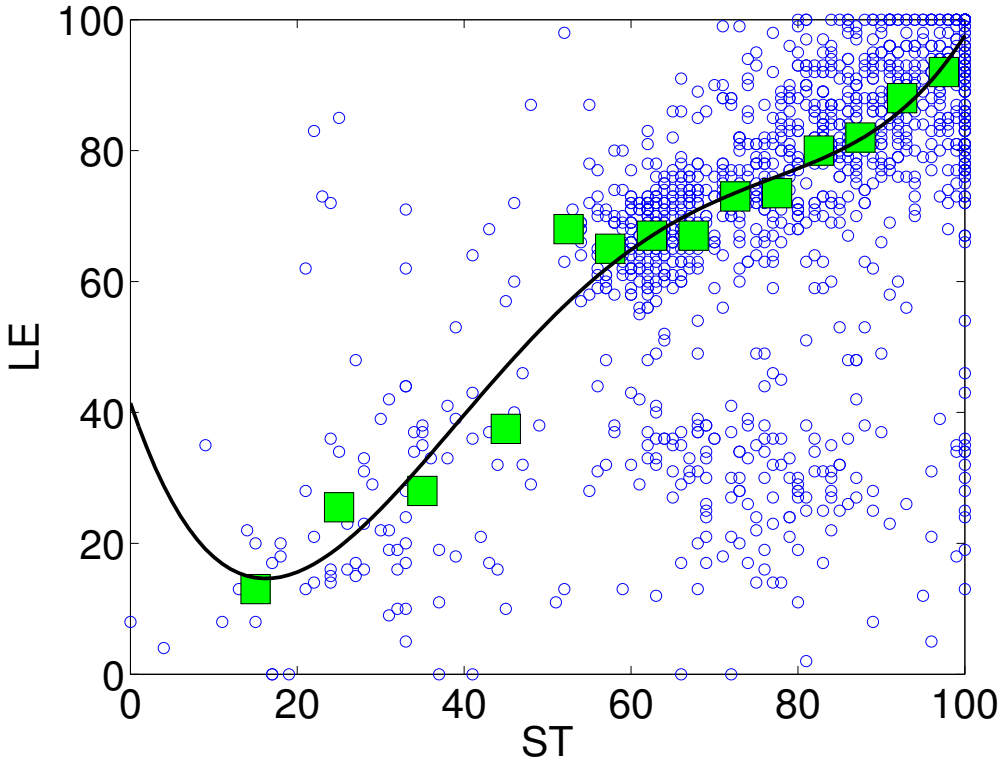


Figure 3.9. $f_3 : \texttt{ST} \mapsto \texttt{LE}$, , mapping satisfaction with the HA to listening effort.

of each bin and median LE scores (the green squares). The degree of the polynomial was selected to improve the accuracy of predicting the combined score given auditory contexts and HA features.

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

3.3.3 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 overfitting.

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 3.2. 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 mod-

els. A model that models the entire dependence between subjects, sessions, and the variables characterizing the auditory context may be easily defined:

$$Y = \beta + subject \cdot session \cdot \sum_{x \in D} x \tag{3.1}$$

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:

$$Y = \beta + subject \cdot \sum_{x \in D} x + session \cdot \sum_{x \in D} x \tag{3.2}$$

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 3.2 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 test 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.

Figures 3.10, 3.11 plot the results on the final model obtained from using the stepwise procedure. Figure 3.10 plots the actual versus the predicted combined scores. The line of best fit (plotted in black) clearly indicates a linear relationship between the

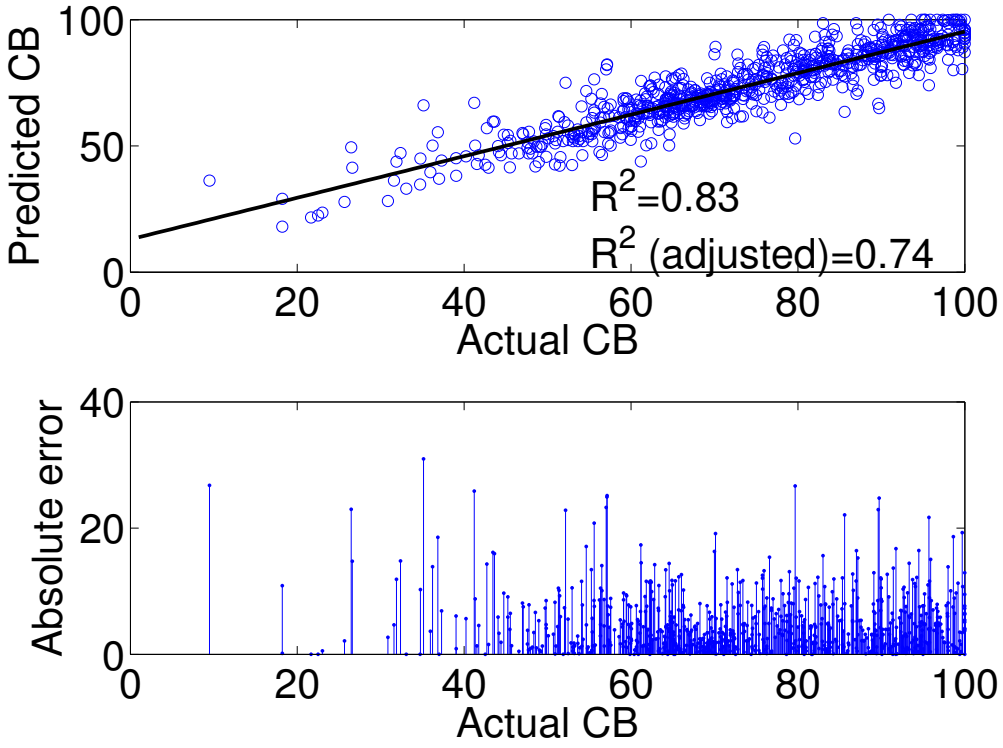


Figure 3.10. Predictions using linear mixed model. The top figure plots how well the regression fits the data. The bottom figure indicates the corresponding errors.

actual and predicted scores. The high R^2 value supports the goodness of fit of the model to the data. The plot of absolute error against observed combined HA outcome does not indicate any additional unaccounted associations. The cumulative distribution of errors is shown in Figure 3.11. 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. We have also investigated the use of non-linear models include 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

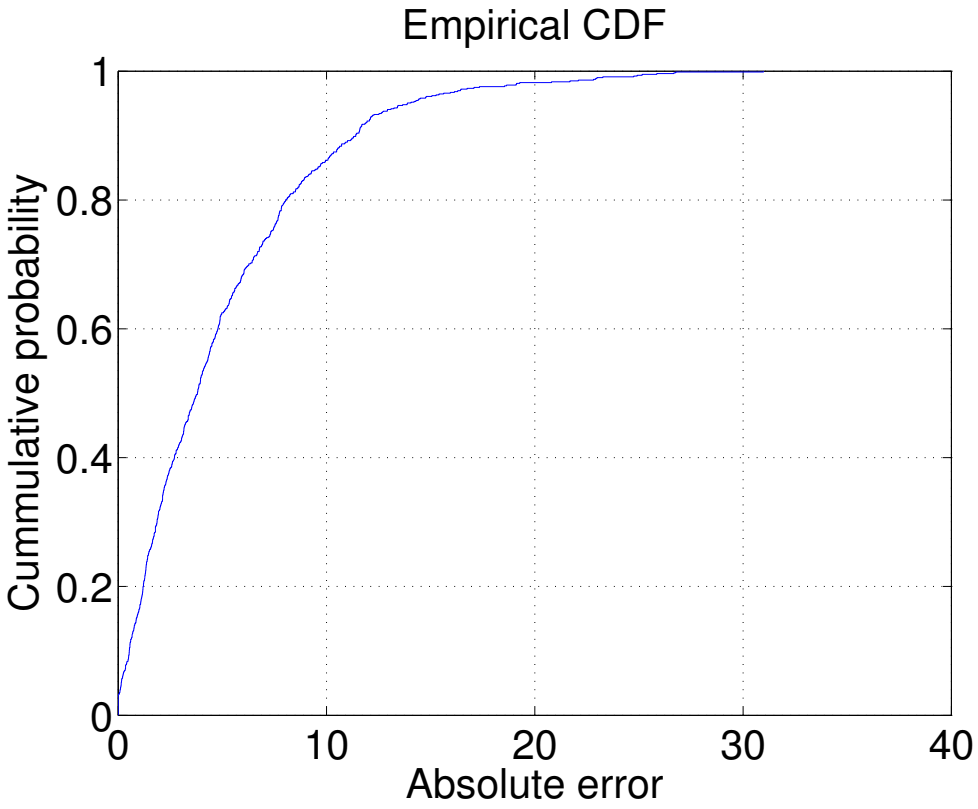


Figure 3.11. CDF of the absolute errors in predicting the combined score. Approximately 85% of the predictions have an error less than 10 points.

yield improvements in accuracy. 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.

To further underline 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 val-

idation, 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%.*

3.4 Conclusion

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. In this chapter we used data from the first ten months through 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 chapter 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%.

CHAPTER 4

IN-SITU MEASUREMENT AND PREDICTION OF HEARING AID OUTCOMES USING MOBILE PHONES

Hearing aids (HAs) are the primary method for treating the 11.3% of Americans [57] 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 [63, 64]. 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 [11]. A recent survey indicates that only 59% of HA users are satisfied and regularly use their HAs [36].

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 [29].

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* [25]. 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. For this chapter, we use 5671 surveys completed by 34 patients using four HA configurations collected over the first two years of AudioSense's deployment. 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.

4.1 Data Utilized

For this chapter we analyzed 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 of 36 were excluded due to low response rates. The resulting dataset includes 34 participants using

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 4.1. Demographic information of subjects included in Chapter 4. All participants within our study are older adults from the state of Iowa. All of them have mild-to-moderate hearing loss.

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). The details of the participants are given in Table 4.1.

4.2 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 [12, 55, 68, 69]. 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 [68]. Since it is impractical to capture such details accurately using traditional survey methods, some audiologists are increasingly interested in Ecological Momentary Assessment (EMA) [60]. 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 [18, 27, 49]. In a previous chapter, we have developed AudioSense [25] – 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. [26] and Wilson et al. [66] evaluated the impact of tinnitus on daily lives of people and Galvez et al. [19] 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 [29, 65].

In [65] 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 [17, 62], others have found no link [29]. Our analysis further validates that HA outcomes cannot be predicted accurately based on PTA and QuickSIN test scores.

In the previous chapter, we characterized the auditory contexts patients encounter in the real-world and made a preliminary analysis of the relationship between contexts and HA outcomes [24]. 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 chapter (when one or both of the identifiers are not available). In this chapter, 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.

4.3 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 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 4.3.1, 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 4.3.2. The results of applying the models in the context of the above scenarios are presented and discussed in Section 4.3.3.

4.3.1 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 ef-

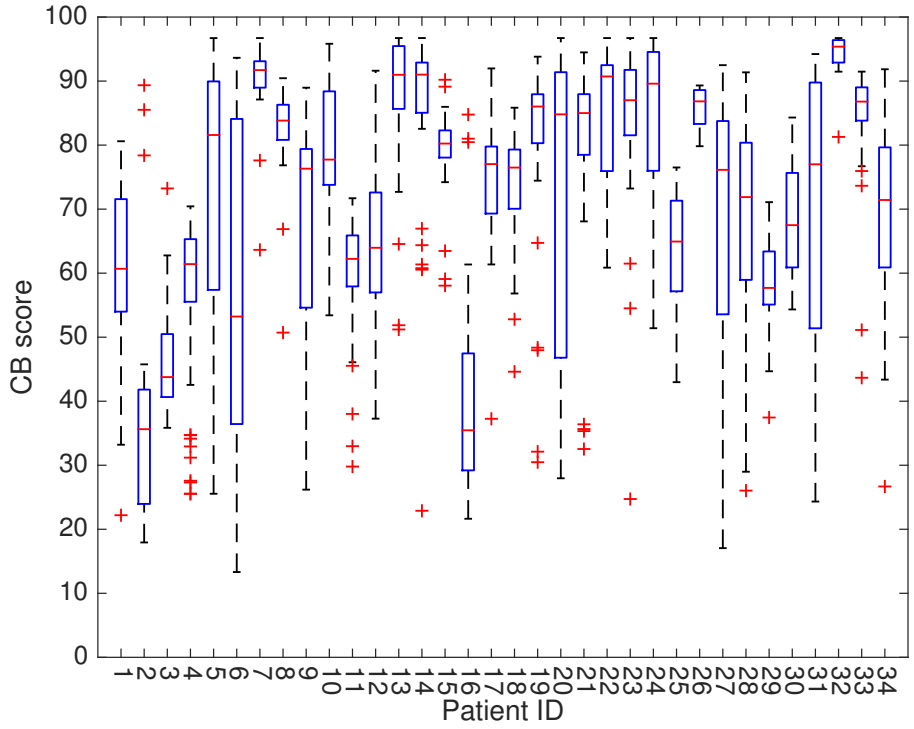


Figure 4.1. Per patient distributions of the combined score scores. The figure only shows the score for Condition 1.

fort, loudness, sound localization, HA satisfaction, and activity participation (see Table 3.3). The correlations between performance domains are included in Table 4.2. 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 [25], 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

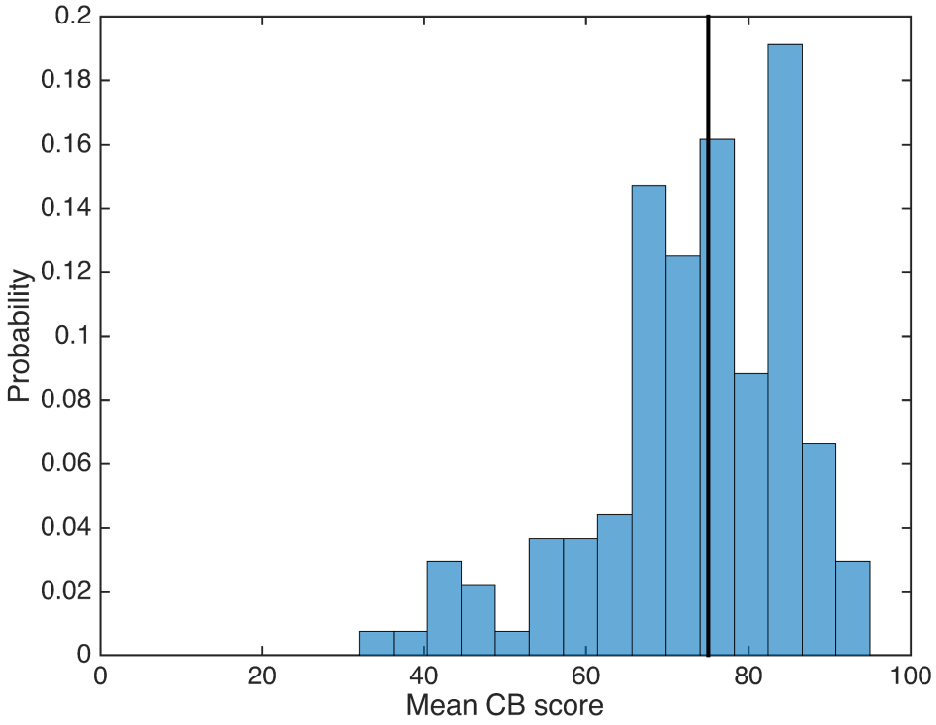


Figure 4.2. Distribution of the combined score across participants. The black line indicates the median score.

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 [28, 47] or using a threshold over an aggregate score [29]. CB is a measure of the *momentary* HA outcome of a patient,

	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 4.2. Spearman's rank correlation between different domains of HA performance for 34 participants. The outcome measures in bold were the most correlated scores and were used for constructing the combined score.

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 $\overline{\text{CB}}$ 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 4.1 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 4.2 plots the distribution of $\overline{\text{CB}}$ 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 $\overline{\text{CB}} \approx 76$ (indicated in the Figure 4.2 as a black

vertical bar) that separates successful and unsuccessful conditions.

4.3.2 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:

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

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:

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

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 [20].

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 con-

tinuous predictions 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 4.3.1 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 3.3). 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.

4.3.3 Empirical Results

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

4.3.3.1 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 4.3 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 ac-

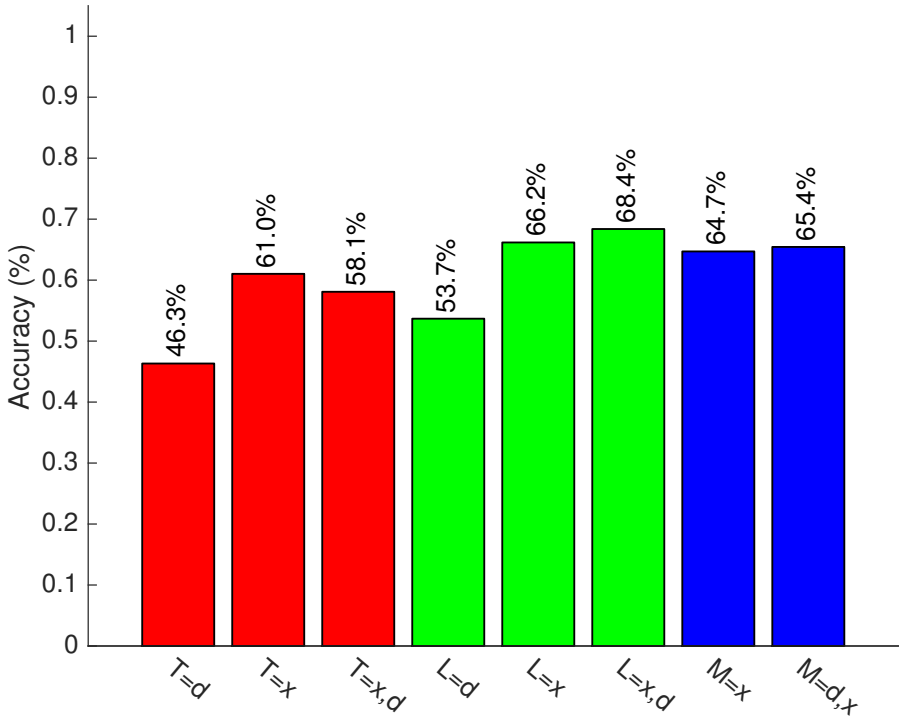


Figure 4.3. Classification accuracy for different models in the novel patient domain. The naming is Model = features, with the models being (T)rees, (L)inear Models, and (M)ixed Effect Models. The features are laboratory tests (d), and contextual information (x).

curacy 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.

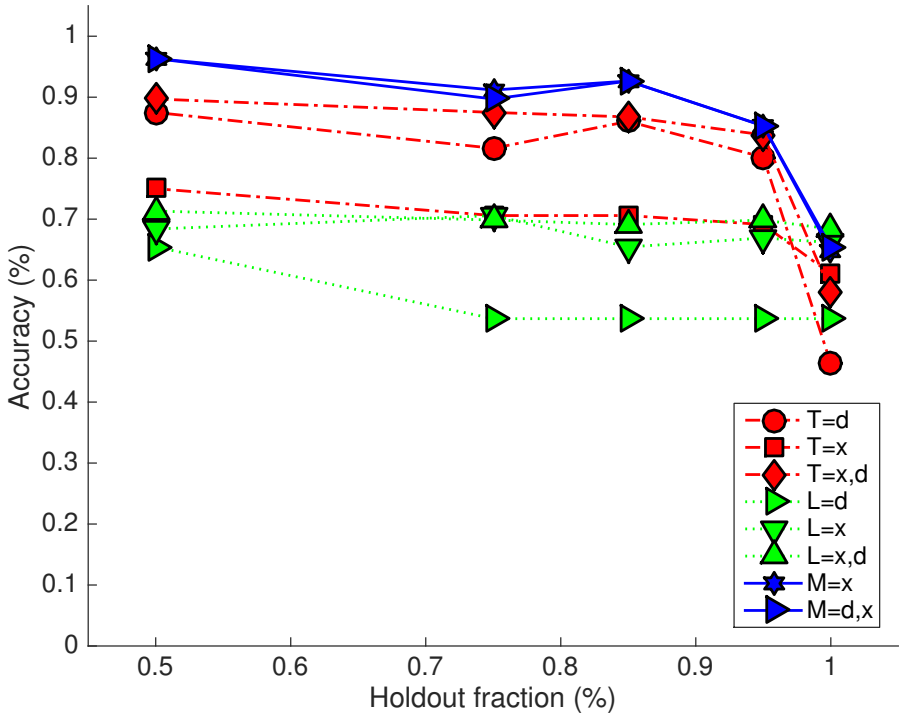


Figure 4.4. Accuracy improvements when some novel patient surveys are used for training. As more information about the participant's lifestyle is introduced in the training, higher accuracies are achieved. Holdout fraction of 1 is equivalent to 4.3.

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 4.4. 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 4.3. The graph clearly indicates that even a small fraction of patient information can have a significant impact on increasing 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.

4.3.3.2 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 4.5 plots the accuracy of predicting HA outcomes when no patient infor-

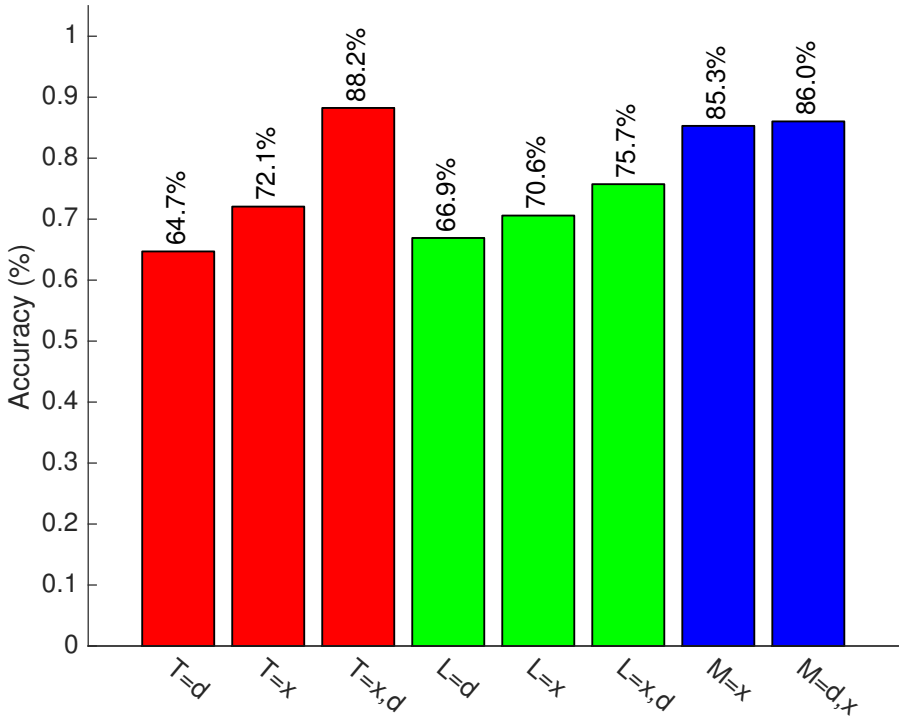


Figure 4.5. Accuracy of the different models for the Novel Hearing Aid domain without any information from other patients who used the Hearing Aid under consideration. The best performance is achieved by the Trees modelled using the laboratory and contextual data.

mation 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

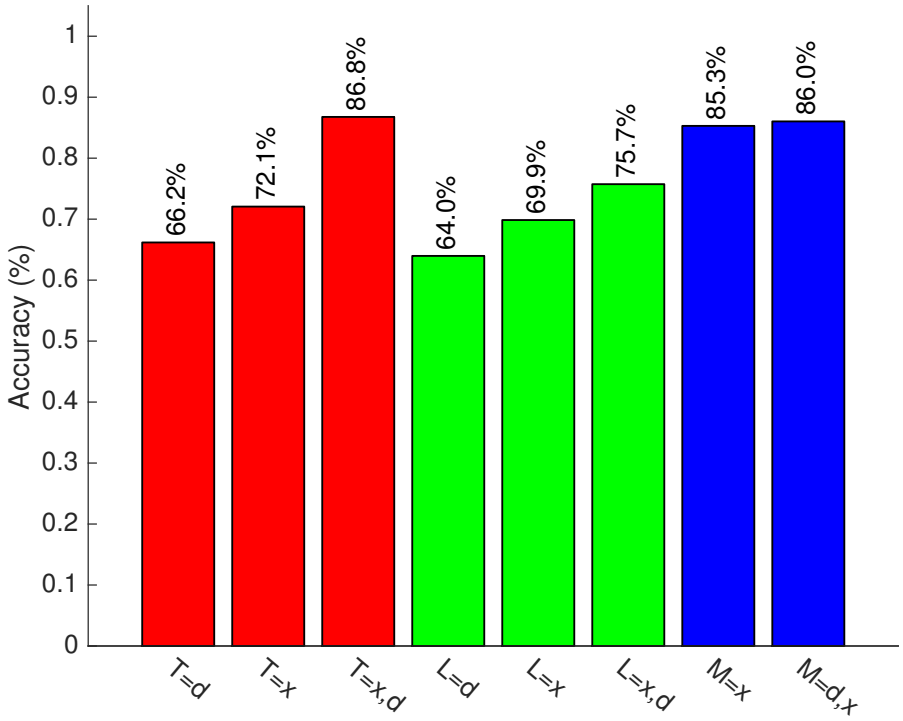


Figure 4.6. Accuracy of the different models for the Novel Hearing Aid domain with information from other patients who used the Hearing Aid under consideration. The best performance is achieved by the Trees modelled using the laboratory and contextual data.

higher accuracy in predicting novel HA than novel patients may be attributed to the fact the training set 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 4.6 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

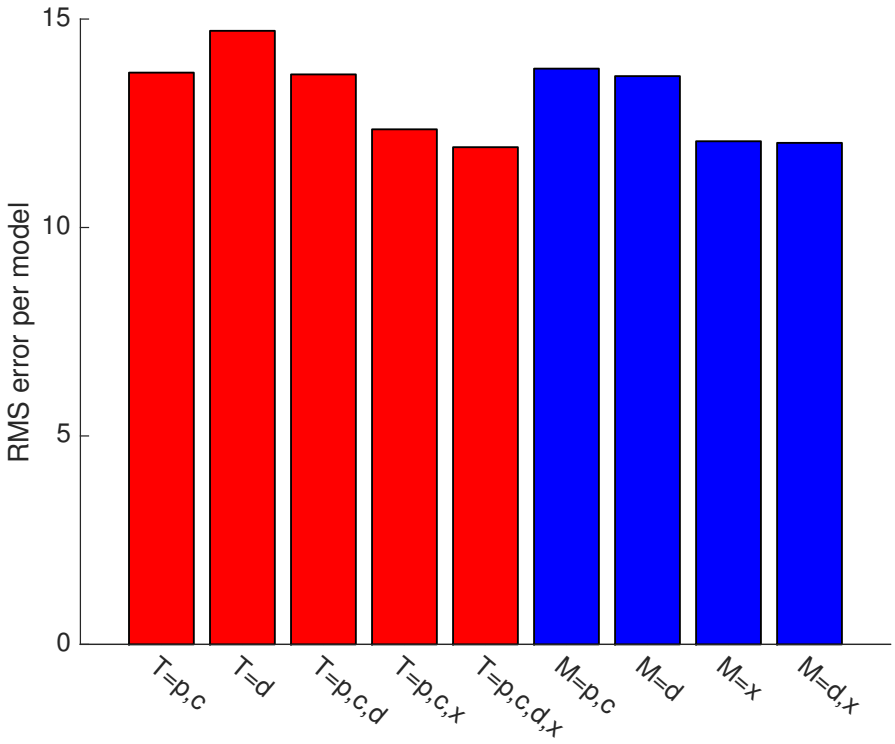


Figure 4.7. RMS error for different models

performance of the models between Figures 4.5 and 4.6 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.

4.3.3.3 Novel Contexts

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

rately 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 4.7 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 4.8 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

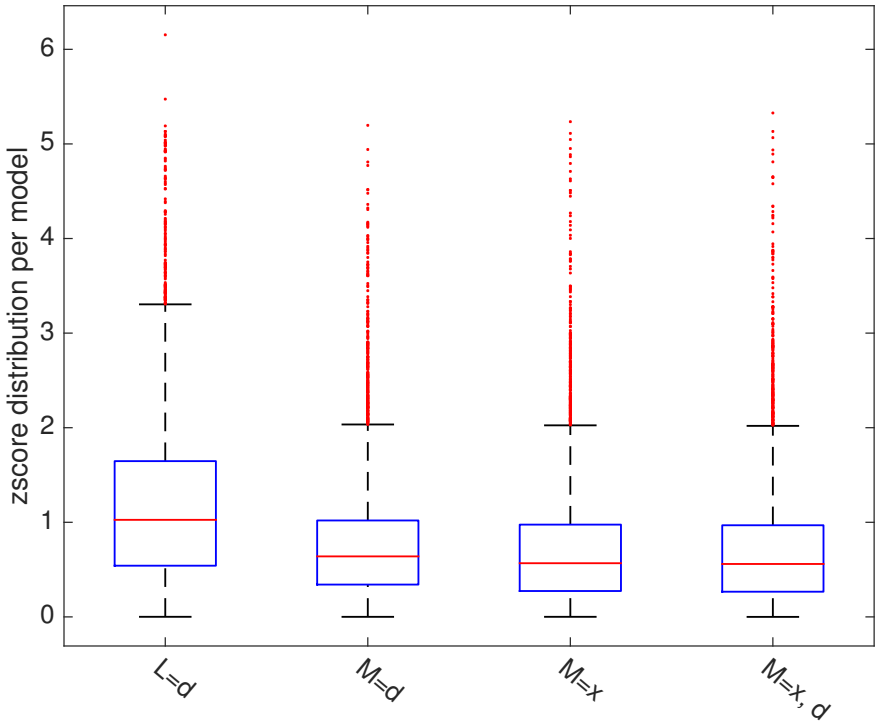


Figure 4.8. Distribution of zscores per model

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.

4.4 Conclusion

This chapter 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 capabil-

ities, 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.

CHAPTER 5

ASSESSING THE PERFORMANCE OF HEARING AIDS USING SURVEYS AND AUDIO DATA COLLECTED IN SITU

Twenty percent of Americans will be 65 years or older by 2030 [31] out of which between 35% and 50% will report having presbycusis [13], an age-related hearing impairment that is primarily treated with hearing aids (HA). Regular use of HAs has been shown to improve communication and avoid the negative effects of hearing loss that include an increased risk of social isolation, depression, and even dementia [6, 63, 64]. Unfortunately, many subjects that would benefit from HAs do not use them regularly [7, 8], as they are often unsatisfied with the performance that their HA provides in the real world. Therefore, there is a critical need to develop clinical tools that can effectively assess the satisfaction of subjects with the performance of HAs in situ to improve the HA technology.

Measuring the performance of HAs poses significant challenges since it depends on the subject's *auditory context*. The auditory context includes characteristics of the listening activity, listening partners, and acoustic environment. Laboratory assessments such as speech recognition tests have been used extensively to evaluate the performance of HAs. During a speech recognition test, a subject is placed in a sound booth and presented segments of speech under different noise conditions. As it is difficult to recreate real world listening conditions in the sound booth, laboratory-based assessments usually fail to be representative of the listening contexts that subjects en-

counter during their daily life. An alternative to using laboratory experiments is to rely on interviews and questionnaires to assess the performance of HAs. Unfortunately, the accuracy of data collected using survey methods is negatively affected by memory biases as subjects are asked to remember the circumstances in which HAs performed poorly long after they occurred. Thus, neither laboratory-based tests nor self-reports are effective in describing the auditory contexts observed by subjects in the real world as clearly demonstrated in several recent studies [12, 55, 68].

An alternative methodology is Ecological Momentary Assessment (EMA) that can jointly characterize the auditory context as well as the HA performance in that context. EMA has the advantage of reducing recall bias and capturing a rich description of auditory contexts that includes the type of listening activity, social context, or the acoustic features of the environment. We have developed a novel mobile phone application called AudioSense that allows audiologists to evaluate the performance of HAs in the real-world [25]. Two hypotheses guided the design of AudioSense: (1) The satisfaction of subjects with their HAs in the real world is best quantified by measuring it repeatedly, in the moment, and in situ. (2) The real-world performance of HAs is intrinsically linked to the auditory context in which the HA is used. An AudioSense assessment combines subjective that that characterizes a subject's perception of the auditory context and HA performance as well as objective audio data.

The goal of this chapter is to explore how the audio the data gathered by AudioSense may be used. We are interested in this problem for two reasons. First, collect-

ing data using AudioSense introduces a significant burden on study participants. Part of this burden may be alleviated by having the application automatically infer characteristics of the auditory context without requiring user input. Specifically, we are interested in whether it is possible to predict the noise level and listening activity reported by subjects. Second, audiologists are interested in understanding the impact that the acoustic environment has on the subject's performance for a given HA. We will focus on exploring the impact that the noise level and listening activity have on the self-reported listening effort. Audiologists have extensively studied this relationship in laboratory conditions. Laboratory experiments clearly show that the listening effort required to understand speech sharply increase with a reduction in SNR [22, 32, 58]. However, little is known about the relationship between listening effort and SNR in the real-world.

We start by considering the problem of predicting the perceived noise level poses. This poses unique challenges since the noise level reported by a subject does not only depend on the acoustic environment but also on the HA used and their subjective perception. Our results indicate that classification algorithms that use only signal-to-noise ratio (SNR) estimates achieve low accuracy. When the SNR features are augmented with other audio features, the classification accuracy increased to 68%. Similarly, the listening activity may be predicted with an accuracy of 70%.

Next, we will evaluate the impact that noise level and listening activity have on the listening effort reported by subjects. Our results show that when we use the *subjec-*

tive noise level and listening activity, we achieve an 18% reduction in the mean squared error (MSE) compared to a baseline model that do not include this information. It is possible to build a model for predicting the listening effort from *objective* audio data using a hierarchical model. The low-level of the model uses the previously developed classifiers to predict the noise level and listening activity from audio data. The predictions of the higher-level classifier are then used to train a classifier the predicts the listening effort. Our results show that using this approach we achieve a reduction of 4.8% in MSE compared to the baseline model. In other words, using the audio data, we can recover about 21.9% of the information contained in the subjective reports.

5.1 Data Utilized

For this chapter we analyzed only the conditions when the HA were used, excluding data from the training and the unaided conditions. The dataset that we consider includes data from 58 subjects within 4 conditions. From this initial dataset, we have removed all the subject-condition pairs that did not include at least 20 surveys. 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. The details of the participants are given in Table 5.1.

5.2 Related Work

Hearing loss is typically evaluated with laboratory tests like Pure Tone Average (PTA), Quick Speech-In-Noise (QuickSIN), and Acceptable Noise Level (ANL)

Variable		**Statistics**
Gender	Male	49%
	Female	51%
Age(years)	Median: 72.5, Range: 64 – 88	
Hearing loss onset(years)	Median:8, Range: 1 – 54	
Duration of HA use (years)	Median: 6, Range : 0 - 40	

Table 5.1. Demographic information of subjects included in Chapter 5. All participants within our study are older adults from the state of Iowa. All of them have mild-to-moderate hearing loss.

[34, 47]. However, studies have shown that HA performance measured in the lab is a poor predictor of the real-world HA performance. [68, 69]. More recently, Ecological Momentary Assessment (EMA) [60] has been proposed as a methodology for assessing HAs. EMA is an attractive alternative to the memory-bias prone retrospective self-report based evaluations. Computer scientists have developed several EMA systems which make use of embedded sensors in mobile devices to collect data in real-time [10,18,27,53]. The use of computer-based EMA in Audiology is still in its infancy with a few studies evaluating HAs [19, 25] and tinnitus [26, 54, 66]. The AudioSense system [25] is more customizable than the existing systems in terms of delivery schedules, adaptive assessments, and collecting multiple dimensions of objective data like audio and GPS. We have shown that using data gathered by AudioSense it is possible to characterize the auditory lifestyle of HA users and predict whether they will be successful users of HA technology [23, 24].

Despite these advances, to the best of our knowledge, no work exists that uti-

lizes audio data to predict a subject's perception of noise level and listening effort. Individual works do exist that use acoustic signals to predict individual activities [42] and background environmental information such as signal-to-noise ratio [35]. The use of audio data to automatically characterize the properties of the auditory context and linking the auditory context to subjective assessment of HA performance has several potential benefits: (1) it can potentially reduce the burden of evaluation on study participants by reducing the number of questions that they are asked and (2) it is possible to construct intelligent sampling policies in contexts that may be of interest to audiologists (e.g., low SNR, when conversation is present).

5.3 Empirical Study and Analysis

The goal of the study is to evaluate whether it is possible to use audio data to predict information about the auditory context and the performance of the HA. Specifically, we will answer the following questions:

- Can the noise level be predicted from audio features?
- Can the listening activity be predicted from audio features?
- Can the listening effort be predicted from audio features?

Our approach to answering the three questions involves the following steps. First, we will empirically characterize the distribution of noise levels and speech activities in the collected dataset. We will highlight the challenges associated with constructing predictors for these subjective measures. Next, we will construct models that will

be used to predict these features. We have experimented with a number of classifiers including support vector machines, decision trees, random forests, and extremely randomized trees. The classifiers provided similar performance and we report the results obtained using extremely randomized trees [21]. Extremely randomized trees are an ensemble method that has been successfully applied to both classification and regression problems. The hyper-parameters of the classifiers are optimized over a manually refined using grid search.

The dataset that we consider includes data from 55 subjects within 4 conditions. From this initial dataset, we have removed all the subject-condition pairs that did not include at least 20 surveys. The results that are reported are obtained using 8-fold cross validation. The folds are generated such that an approximatively equal number of samples for each subject-condition are included in each fold. Due to the significant imbalance in the dataset (some subjects provided significantly more reports than others), we weighted as sample such that each subject-condition pair has an equal weight.

5.3.1 Predicting the Noise Level

New algorithms and technologies for HAs are primarily evaluated in the laboratory using carefully controlled experiments. A common setup is to present speech under different SNR conditions. Laboratory experiments show that the SNR is correlated with the listening effort required for correctly understanding speech. In our study, the subjects report the noise level as a proxy for SNR. It is important to realize the noise level does not depend only on the actual noise level in the environment (which

can be assessed using audio data) but also on the behavior of the HA and the subjective preferences of the subject.

Figure 5.1 plots the number of reports pertaining to each noise level as reported by a subject. A few trends are clear: (1) The subjects spend most of their time in quiet or somewhat quiet conditions. (2) There is a significant variation between subjects when they are exposed to different noise levels. These trends make the problem of classifying the perceived noise level particularly difficult due to the imbalance between classes and the high variation between subjects.

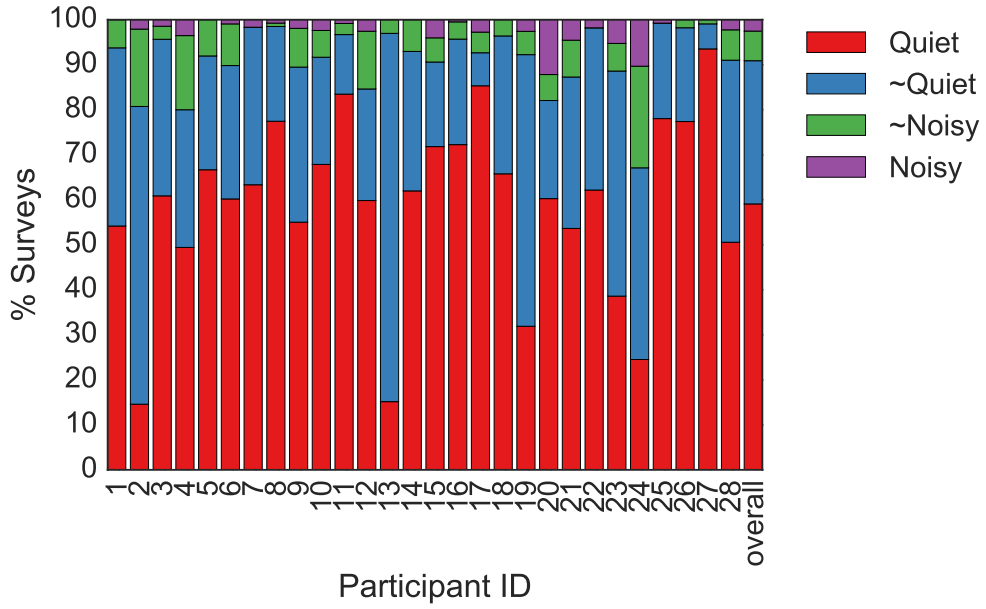


Figure 5.1. Distribution of noise level per participant. The figure shows only a subset of all the participant in the study. The rightmost bar indicates the overall trend.

The starting point for predicting the noise level is to use off-the-shelf algorithms that have been designed for assessing the SNR. NIST SNR evaluates the SNR by computing the RMS power histogram of the audio signal. The method estimates the noise power by fitting a raised cosine to the histogram. The noise power is then subtracted from the composite signal power histogram to obtain the clean signal power. WADA SNR [35] estimates the clean signal by modeling it as a Gamma distribution. The Gamma distribution has been shown to be a good approximation of amplitude distribution of speech [50,56]. The noise is assumed to be Gaussian. VAD SNR [4] applies the same SNR estimation only to those segments where the presence of speech is detected. Figure 5.2 plots the distribution of estimated SNR for of the noise levels. All three estimators show a similar trend: as the noise level increases the median and interquartile range of the estimated SNRs decreases. However, it may be hard to discriminate the noise level when the estimated value is in the range 10 – 20 because of the significant overlap between the estimated SNR distributions for different noise levels.

We have built two classifiers that address the challenge of the imbalanced data by reducing the levels of the noise variable in different ways. The `NZ3` classifier has three classes: quiet, somewhat quiet, and merged class including somewhat noisy and noisy. Similarly, the `NZ2` classifier has two classes: quiet and non-quiet which includes the remainder of the data. We have fit the model using only the data from the SNR estimators and the SNR estimators in conjunction with the other audio features. Figure 5.4 plots the accuracy and F1-score for the `NZ2` and `NZ3` estimators when using SNR

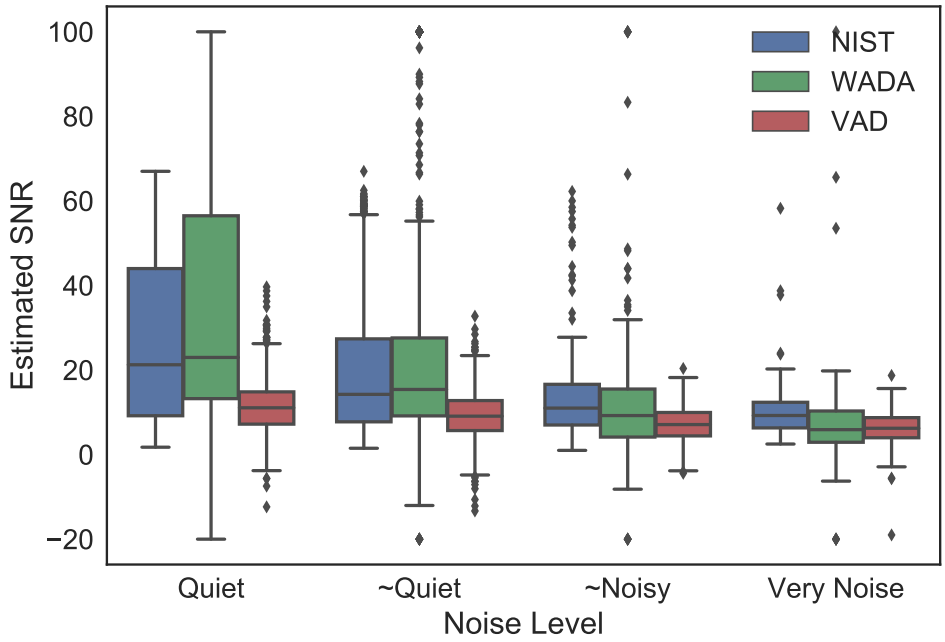


Figure 5.2. Distribution of the SNR calculated by automated algorithms in different reported noise levels. Higher noise levels have lower SNRs with less variability.

and audio features. The figure indicates that `NZ2` has higher accuracy and F1-score than `NZ3`. This indicates that is relatively easy to identify quiet conditions with accuracy as high as 78%. The figure also indicates that including audio features increase the accuracy by about 10% for both classifiers over the case when only the SNR features are used.

5.3.2 Predicting the Listening Activity

The degree to which an HA benefits a subject may also depend on the type of listening activity in which they engage. Figure 5.5 plots the distribution of activities in which the subjects engage in. Subjects spent about 18% of the time listening passively.

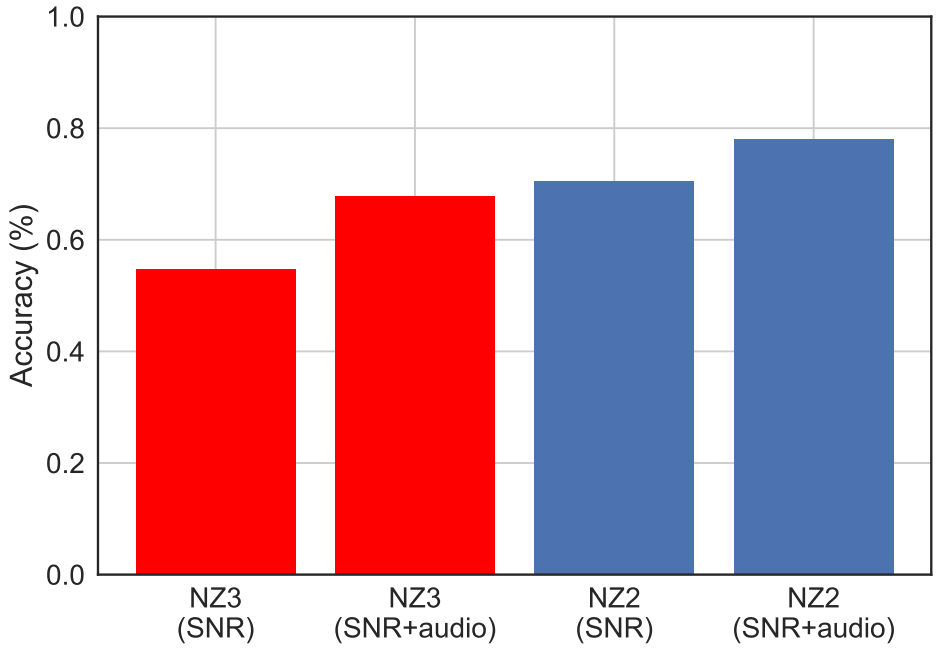


Figure 5.3. Accuracy of the machine learning model for predicting the noise levels (3 level = NZ3, 2 level = NZ2) based on SNR and audio data.

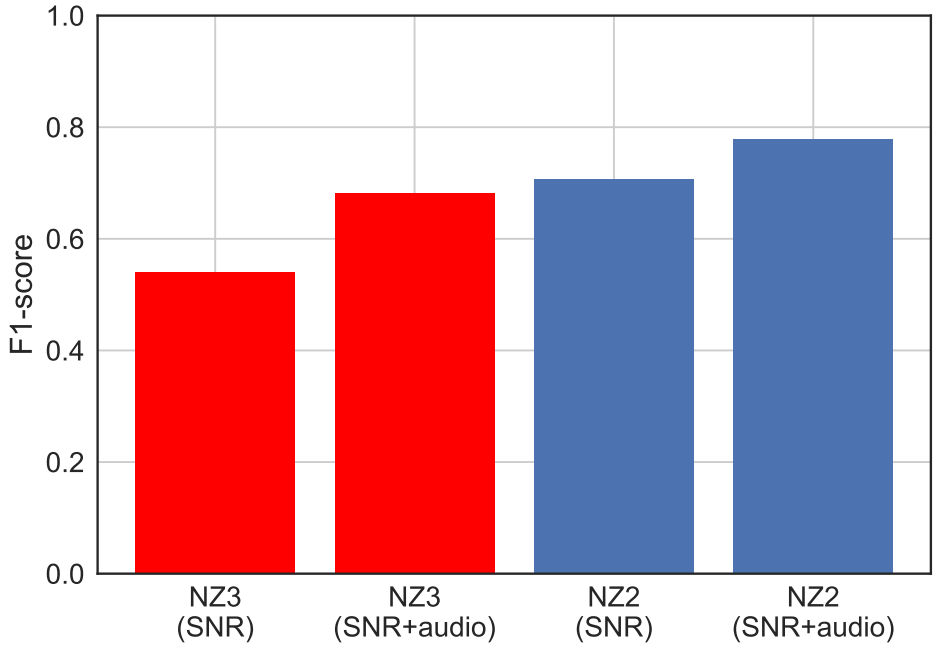


Figure 5.4. F-1 Score of the predictions made by the machine learning model for noise levels (3 level = NZ3, 2 level = NZ2) based on SNR and audio data.

The most prevalent activities were listening to media (35%) and speaking to fewer than three people (25%). The figure also highlights a wide range of variations between subjects. A challenge to building a classifier is that several activities have similar auditory characteristics. For example, the two conversation classes (`Conv <= 3` and `Conv > 3`) and `Phone` involve people talking. Accordingly, to simplify and improve the accuracy of the classification, we collapse these listening activities in a single class. The classifier is trained using the audio and SNR features.

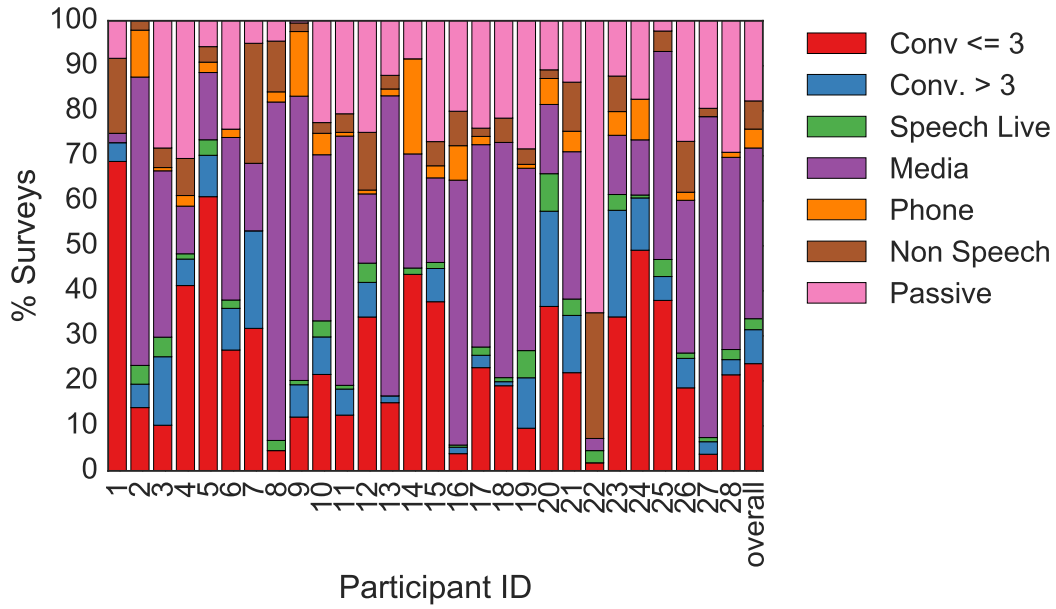


Figure 5.5. Distribution of listening activities across different participants. The figure shows only a small subset of all participants in the study. The rightmost bar indicates the overall trend.

Figure 5.6 shows the confusion matrix for the classifier. Overall, the classifier is

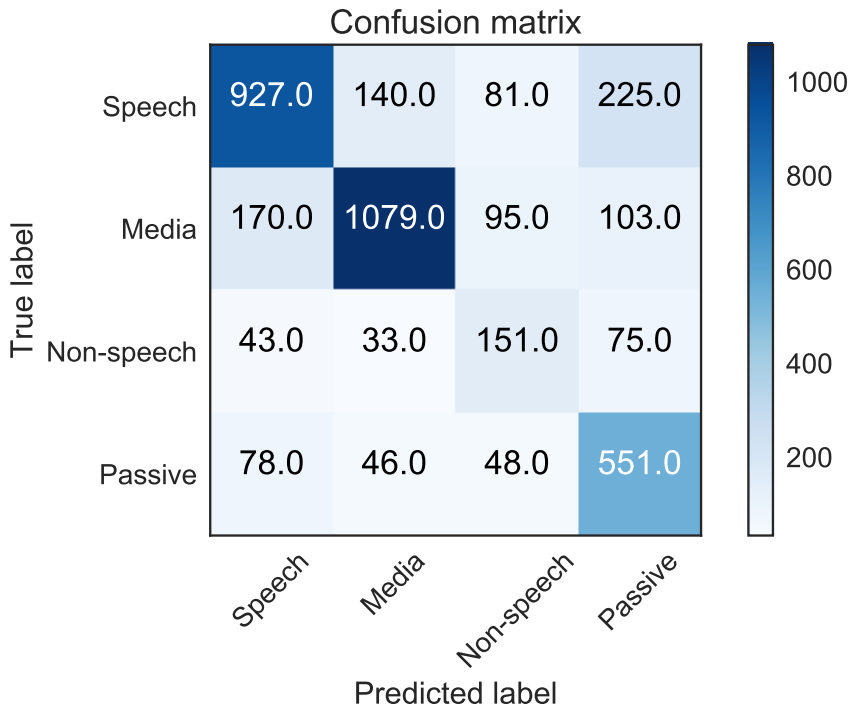


Figure 5.6. Confusion matrix for listening activity. It can been seen that all classes can be discriminated quite well save for Non-speech activity. This can be because of the wide latitude that the label non-speech provides the user.

reasonably accurate having mean accuracy and F1-score of 70% and 0.71, respectively. The most common misclassification is between speech and media. This is expected since speech is usually present when subjects are watching TV or listening to media.

5.3.3 Predicting the Listening Effort

Listening effort is a sensitive measure of the performance of the HA, particularly in speech. Figure 5.7 plots the relationship between the noise level and listening effort. In order to account for the differences in how subjects may rate and the impact of HAs, we group samples according to their subject and condition. For each one of those

samples, we subtract the mean of the group. This scaling allows us to interpret positive values as requiring more effort than the average. Conversely, negative values indicate they require less effort than the average. In quiet, the subjects require a less listening effort to hear well. This is clear from the slightly below zero median and the narrow interquartile range in quiet. In contrast, the lower quartile of the listening effort is about zero in somewhat noisy environments. This indicates that subjects require significantly higher listening effort to cope with higher noise. Our results are consistent with survey results that show that a significant fraction of subjects are unsatisfied with the performance of their HAs in noise. Figure 5.8 plots the relationship between the listening activity and listening effort. A subject requires higher effort to listen to speech than media or non-speech sounds. This seems to point towards these conditions being less demanding for HA technology. However, unlike with the noise levels, the difference in the listening effort scores between various listening activities is less pronounced.

The open research question that we consider here is whether audio measures are predictive of their listening effort. While such a relationship has been studied before in the laboratory, this is the first time it is evaluated using a large-scale dataset collected in situ. In order to evaluate this question, we will build a hierarchical classifier. The bottom-level consists of the classifiers that we have described in the previous sections to predict the noise level and the listening activity from audio features. The top-level consists of a classifier that combines the predicted noise level and listening activity with information about the identity of the subject and the HA they are using to predict their

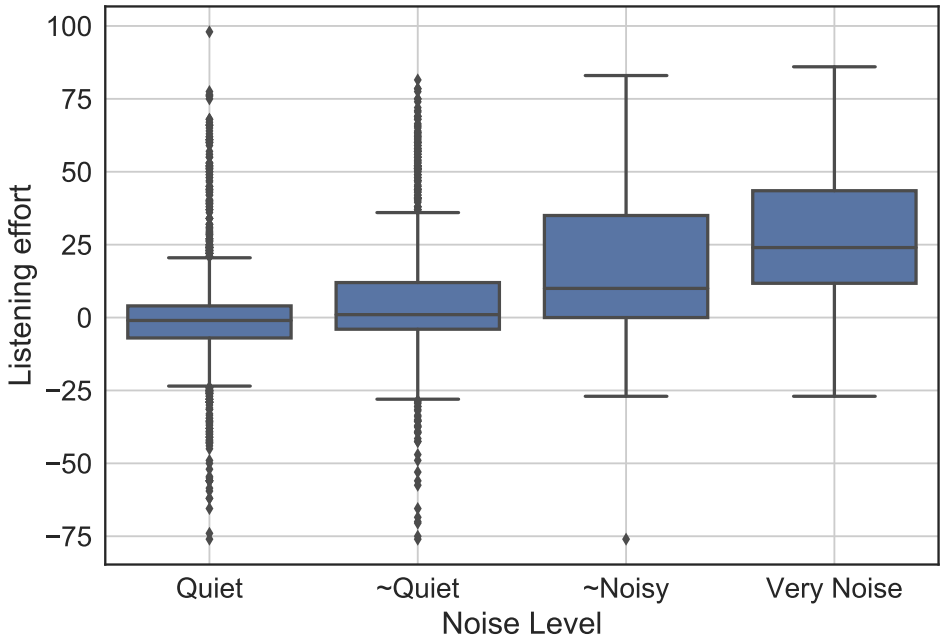


Figure 5.7. Impact of the level of noise reported by the user on the effort invested by the user in listening well. As the noise level increases, the effort also increases. The listening effort has been normalized across users by subtracting the mean.

satisfaction. The baseline is a classifier that uses the subjective values of the noise level and listening activity as reported by the user.

Figure 5.9 plot the predictions of listening effort based on different subsets of features: subject identifier `p`, condition identifier `c`, the subjective noise level and activity (`nz` and `ac`) and their objective counterparts (`onz` and `oac`). The results obtained using subjective and objective data are colored in red and blue, respectively. The baseline performance is the classifier that uses only the subject and condition identifiers. This classifier essentially predicts the mean listening effort of each subject for the considered HA. The performance of the classifiers may be improved by considering

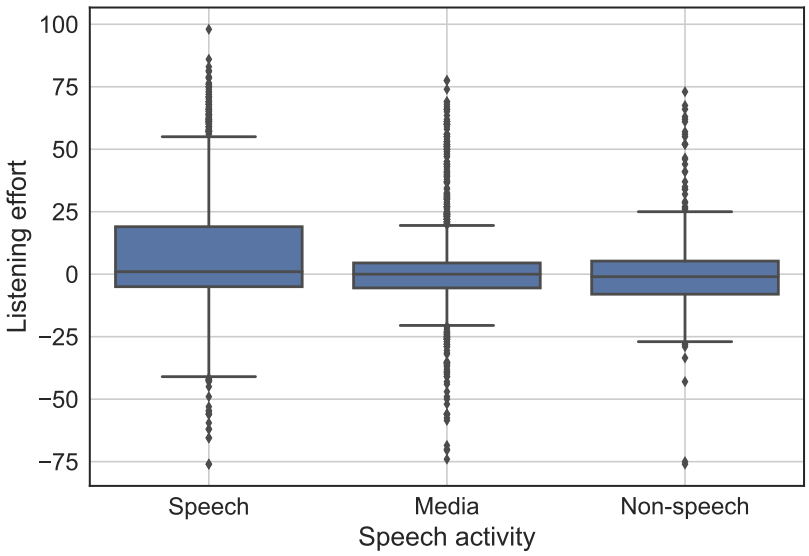


Figure 5.8. Impact of the listening activity reported by the user on the effort invested by the user in listening well. The listening effort has been normalized across users by subtracting the mean.

additional subjective measures. For example, the MSE is reduced from 493 when only subject and condition are available to 385 when all the subjective features are used. This is a reduction of 21.9% in MSE. Using objective data is not as effective in improving the prediction accuracy. The classifier that uses a combination of predicted noise level and activity type performs the best achieving an MSE of 518. This is an improvement of 4.8% over the baseline.

5.4 Conclusion

Effective tools for assessing the performance of HAs are essential to developing novel HA algorithms and technology. A key challenge to building such tools is the need to reduce the data collection burden on the subject. In this chapter, we make an initial

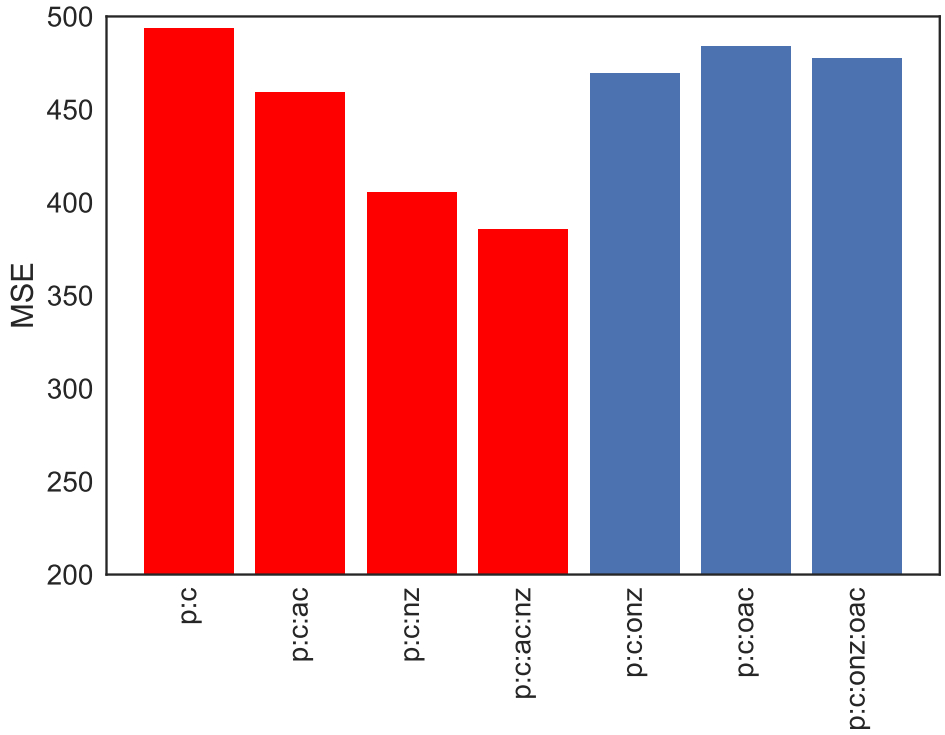


Figure 5.9. Performance of the 2-level model in predicting the outcome scores (LE) using the reported information (ac, nz), and the inferred information (oac, onz).

attempt at evaluating the potential of reducing the burden of data collection on the user. Our results show that audio features may be used to predict the perceived noise level with an accuracy of 78%. This is remarkable given that the noise level reported by a subject depends on both the subject's hearing abilities and the performance of the HA. Additionally, we also show that it is possible to predict the listening activity with an accuracy of 70%. This suggests that some aspects of the auditory context could be automatically inferred from audio data without involving the user. More importantly, we show that the listening effort depends on both noise level and listening activity.

Using subjective information regarding the noise level and the listening activity, the predicted MSE can be reduced by as much as 20% over a baseline model that includes information about the subject and HA. In contrast, a hierarchical classifier to predict the listening effort from audio data can reduce the MSE by a mere 4%. The significant gap between the prediction made using audio data, and those made using the subject's self-reports suggests that there may be significant room for developing novel machine learning models to tackle this problem.

CHAPTER 6

AUDIOSENSE+: NEXT-GENERATION MOBILE-EMA FOR HEARING AID EVALUATIONS

In previous chapters we saw the creation of AudioSense which utilized ideas of *m*EMA to sample data in real-time and characterize the performance of the HAs jointly with the auditory context experienced by the HA user. Through the various stages of data collection, analysis, and insight generation we found a number of avenues for expanding AudioSense and found instance where AudioSense lacked the capability of fully capturing the auditory context. Some of these issues relate to the design of the survey with inability of a user to report more than one listening activity, or the wording of the options which leads to confusion. In addition to this, from the data source perspective, we only concerned ourselves with contextual data, and HA outcomes. A key piece of data that was missing was information coming from the HAs themselves. This is a very important piece of information because it serves as a link that allows us to understand how contextual information relates to the HA performance in a much more comprehensive manner. In this chapter we shall first describe the limitations associated with AudioSense in detail and then present the next generation of the system called AudioSense+ and show how it overcomes those issues while retaining the flexible nature of the preceding system.

6.1 Limitations of AudioSense

AudioSense has been highly successful in collecting real-world data related to HA performance. However we discovered several domains where extensions can be made to refine the process of the data collection, and increase the capability of AudioSense. We shall focus on the following three domains:

1. Survey design
2. Objective data sources
3. Assessment delivery and data collection system

6.1.1 Survey Design

The current AudioSense survey design asks the user to report the details of the acoustic activity that occurred *most of the time*. Such a methodology can lead greater noise within the data because study participants experiencing multiple listening activities will not be able to fully report the details of their context. Additionally, it is very difficult to delineate the details of individual contexts while experiencing multiple ones.

Figure 5.6 shows the confusion matrix of the activity classifier. The misclassification between Speech and Media classes stems from two reasons: (i) media recordings contain speech, and (ii) in a number of samples study participants were engaged in conversations with the media device playing in the background. Due to the limited nature of the current study design the participants were unable to report instances of the latter

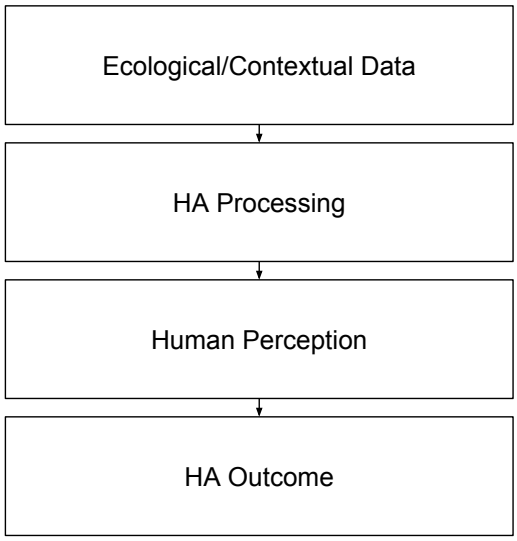


Figure 6.1. The process of generating HA outcomes from the human perspective. The environmental data is captured by the HA and is processed. The processed data is then fed to the human ear for processing which in turn leads to the development of the perception of performance in the form of HA outcomes

.

kind and this led to increased misclassification within the model.

The HA outcomes are captured on a 0–100 scale. The disadvantage of using a wide scale is the variation that creeps in when the participants are reporting events of the similar intensity. For example, if a participants puts in a significant effort into listening well and they report the LE to be, say, 75, the next time they experience something very similar they might report their effort to be 79. The difference does not necessarily mean that the participants put in more effort, but could very well be an effect of the high resolution of the scale.

6.1.2 Objective Data Sources

Currently AudioSense only collects location and acoustic data in terms of objective data streams. The performance of the HA or HA outcome is the perception that the user generates over the processed input from the HA as shown in Figure 6.1. In the current version of AudioSense we do not capture the HA processing step at all which is critical in the development of perception. Finding a method of capturing such outputs can not only give us the ability to understand the HA outcomes better, but also help us understand which factors within the DSP of the HA affect the perceived performance of HAs.

6.1.3 Assessment Delivery and Data Collection System

A major draw back of the data collection scheme within AudioSense is the lack of data between the delivery of assessments. Outside of research domains it is not be possible to carry LENA like devices capable of capturing data continuously and streaming continuously on the phone is not energy efficient. We need to develop a way to address this issue and decrease our reliance on additional hardware while minimally affecting the quality of the data collected.

6.2 AudioSense+: A Comprehensive Mobile EMA System for HA Evaluations

We developed the AudioSense+ system with the aim of providing a solution to the issues raised in Section 6.1 and extending the existing capabilities of the AudioSense system. The high-level architecture of the system is shown in Figure 6.2.

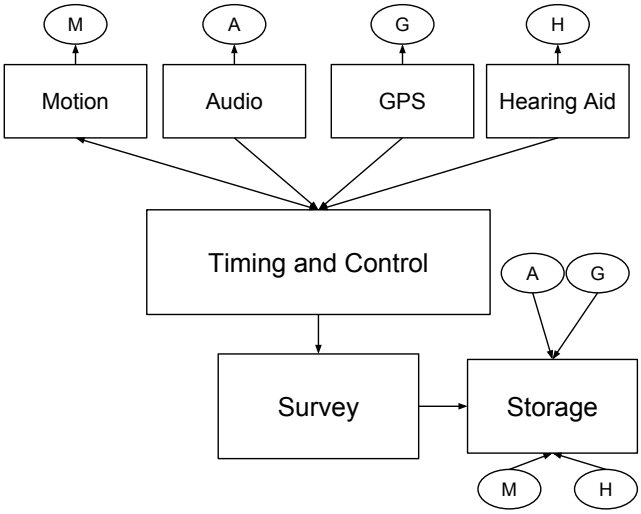


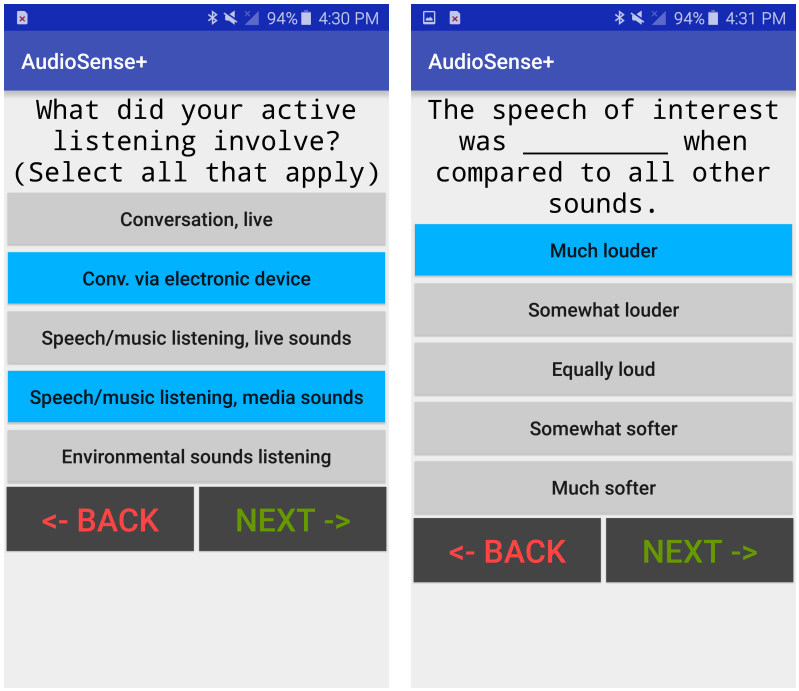
Figure 6.2. High level architecture of the AudioSense+ system. The objective data is collected in the form of motion (acceleration), location (GPS), and HA parameters. The subjective data is collected in the form of surveys. All the data collection is controlled by the timing and control system. The collected data is stored in interpretable format by the Storage system. The complete system is confined within the mobile phone.

6.2.1 Survey System

The survey system has been entirely redesigned to incorporate greater flexibility in defining the types of questions that could be asked. The system now supports multiple selections for allowing users to report multiple listening activities (Figure 6.3(a)). In addition to this we have reduced the outcome score responses from a 101 point scale to a 5 point Likert scale response (Figure 6.3(b). This should allow us to eliminate the variance within the responses that represent the same intensity.

6.2.2 Objective Data Collection

The AudioSense+ system is very versatile in terms of the types of objective data it collects (see Figure 6.2. In addition to collecting the location and audio data, the



(a) Multiple Selections (b) 5 point outcomes

Figure 6.3. Redesigned survey system capable of capturing data in multiple ways. The current system is capable of presenting questions where multiple options can be selected, and we have also redesigned the outcome score questions by making them 5 point Likert scale responses.

system also collects motion data in the form of acceleration readings along the X, Y, and Z axes of the phone and also streams the HA's DSP parameters computer internally by the HA. The acceleration is extracted from the onboard accelerometer on the phone. The connection with the HAs is maintained via Bluetooth Low Energy (BLE) and data is streamed at a frequency of approximately 5Hz. The collection of HA parameters allows us to fill the gap of HA processing that was present in the previous system (see Figure 6.1).

6.2.3 Timing System

The timing system of AudioSense+ retains all aspects of AudioSense (see Section 2.2.2.1). In addition to the timer-initiated assessment delivery, we also collect short snippets of objective data between the assessments, the duration can be configured by the clinicians. This allows us to have a more contiguous stream of objective data while not overwhelming the disk space on the mobile phone. The major advantage of using such a scheme is that we can reduce our reliance on other hardware, like LENA, in future studies.

6.2.4 Privacy

A key new feature within AudioSense+ is the introduction of privacy settings. The study participant can define which objective measure they do not want the researchers to record at the beginning of the study. This feature has been implemented with the hopes of increasing the real-world adoption of our platform.

6.3 Conclusion

In this chapter we touched on the limitations of the current AudioSense system, specifically the lack of flexibility in the survey for selecting multiple activities and the high variance in outcome scores due to continuous 101 point scales, the lack of any data from the HA, and the absence of data between the delivery of assessments on the phone. We addressed these challenges by building the AudioSense+ system, a comprehensive mobile EMA platform capable of handling the aforementioned issues and capturing much more data at a higher frequency. We plan to deploy this system in the next few months (mid 2017) as a part of a multi-site study aimed at evaluating HAs and understanding the effects of various factors like contextual information and HA processing data on HA performance.

CHAPTER 7

CONCLUSION & FUTURE WORK

7.1 Concluding Remarks

This thesis reported the development of AudioSense, a *m*EMA system, for collecting real-time and in-situ hearing aid evaluation data.

In Chapter 2 we began by providing the limitations associated with the traditional methodologies of hearing aid evaluation and then presented AudioSense for overcoming the problems of non-representative data collection and memory bias. The key contribution of the system as a whole is the flexibility in the assessment delivery scheme which was unavailable in previous *m*EMA systems for hearing aid evaluations. We also showed that AudioSense performs with a 100% reliability and demonstrated the low power consumption that is built into it.

Chapter 3 used the data collected using AudioSense to characterize the auditory lifestyle of hearing aid users and extract the relationship between the reported context and hearing aid outcomes. We found that, based on the reported data, study participants spent most of their time in conversations or watching television in low noise environments. We also found that participants reported higher importance of listening well in socially unfamiliar environments like crowds. We also proposed a unique outcome metric (`CB`) for measuring hearing aid performance based on a combination of the individual outcome measures like satisfaction, listening effort, speech perception etc. Using this metric we showed that we can discriminate between good and bad out-

comes based on contextual information with an accuracy of 78% thereby indicating that contextual information is helpful in predicting outcome scores. This idea was further explored in Chapter 4. Using the idea of successful hearing aid users, as defined by traditional methodologies, we created models to predict hearing aid prescription success for new and experienced users. We built several models using the laboratory scores, the contextual information, and a combination of both and found that the introduction of contextual information greatly aids the prediction with accuracies around 70%. This quantitatively proved the importance of collecting contextual information in addition to laboratory measures for hearing aid evaluations. We also observed that if we introduce some data about the user's context into the training of the models it can help in boosting the prediction accuracy to 90%.

Chapter 5 explored the use of objective data for inferring some of the subjective contextual information for reducing the burden of response on the participant. We experiment with using off-the-shelf Signal-to-Noise Ratio (SNR) algorithms and custom spectral and time-domain features for predicting activity types and noise level. We are able to achieve accuracies as high as 78% for discriminating between noise levels, and 70% for discriminating between activity types. We further build a hierarchical model to predict the outcome scores and establish a relationship between the acoustic data and the performance of the hearing aids. We do not see a significant improvement against the baseline predictor while using our predicted objective data. These results, however, provide the first steps towards building a context-sensitive sampling schemes.

Finally, in Chapter 6 we explored the limitations with the current design of the AudioSense system in terms of survey design, data collection sources, and timings. We introduced the next-generation *m*EMA platform AudioSense+ which is designed to overcome the limitations and extend the capabilities by streaming novel objective information like HA parameters.

7.2 Future Work

I believe that this thesis lays the groundwork for very high-impact future research. There are four directions that I believe can be explored based on my work:

7.2.1 Context sensitive sampling

A common issue with using a semi-randomized data collection protocol in *m*EMA systems is the misfiring of alarms *i.e.* delivering assessments in non-informative contexts. Using the results of Chapter 5 we can build machine learning models to identify activities and suppress assessments in situations where it is well established that information, in the context of hearing aid evalutaion, does not exist. Using GPS data can also be helpful in identifying new locations where the study participant has not been before. This information can then be utilized to deliver assessments for capturing a much wider range of environments than would be possible with a semi-randomized approach. There have been preliminary studies in the area of Context-Sensitive EMA (CS-EMA) [30, 45, 51] exploring the domain but no comprehensive real-world study has been conducted so far that utilizes CS-EMA at its core to evaluate hearing aids. A secondary gain from pursuing this direction would be increasing the cost-effectiveness

of clinical studies employing *m*EMA. Study designs often include incentive schemes that depend on the number of assessments responded to by the study participants. By improving the effectiveness of the assessments by delivering them in relevant contexts the costs associated can either be reduced or the quantity of the relevant data collected can be increased while keeping costs level. With these new developments, new challenges are certain to arise. In my opinion the key challenge in this regard would be privacy awareness within the system. For example, rather than storing raw audio the device should store processed features. Providing the study participants with the option of choosing the level of privacy granularity can help in increased adoption.

7.2.2 Exploring the relationship between physiological measures and hearing aid outcomes

The presence of variability within the hearing aid outcomes in the same acoustic context, given the current design of evaluation studies like AudioSense, is non-trivial to explain. The variability in outcome scores might be caused due to variability in the physiological state that is not being captured by the subjective assessments. With the explosion of wearable technology capable of measuring physiological factors like skin conductance, pupil dilation, heart-rate, blood-oxygen levels etc. this is no longer an infeasible task. Recent studies have shown that some physiological measures like skin conductance, heart rate variability, and pupil dilation have relationships with outcomes like listening effort [33, 37, 43, 44]. Capturing these signals in real-time in addition to the subjective assessment using wearable devices like the the Empatica E4 [1] can

potentially help bridge the gap that exists within the current study settings.

7.2.3 Bringing the hearing aid into the research loop

So far the hearing aid evaluation studies, including AudioSense, have used the hearing aids as black boxes where the clinicians configure them and their effects are evaluated in the field. The next-generation hearing aids like Starkey's Halo series are capable of communicating with other devices over Bluetooth Low Energy (BLE), and open source research tools like the Open Speech Platform [3] allow researchers to investigate custom algorithms. This opens a completely new area of research where, potentially, researchers can gain access to the internal parameters of the device thereby bringing it into the research loop and adding more diversity to the collected data. Such parameters can also help validate the efficacy of the objective audio signals that are being captured on the mobile phones. This is of particular importance because the current study designs do not factor in the effects of the hearing aid's internal functioning (or only include abstract representations) on the hearing aid outcome. This information in addition to physiological measures mentioned in Section 7.2.2 can improve the overall understanding of outcomes.

7.2.4 Cloud based hearing aid tuning

We can further utilize the connectivity of hearing aids with mobile phones, as mentioned in Section 7.2.3, to improve the user experience via crowd based learning. This system would consist of two stages. The first stage would consist of collecting acoustic information of the surroundings, corresponding hearing aid parameters,

the user's preferences, and hearing loss information in real-time from multiple users should allow researchers to build a comprehensive cloud-based knowledge base. Using this dataset we can build generic machine learning models that can learn the optimal hearing aid configuration for different acoustic environments similar to Auditeur [48]. The second stage would involve implementing these models on the mobile devices and issuing local updates via active learning while also reporting the changes to the back-end for updating the global models. Two major challenges that, I can foresee, with building such systems would be inferring the amount of data needed to make a reliable prediction and the choice of machine learning models such that they are easy to update on limited resource devices like mobile phones. The aim of such an endeavour would be two-fold: i) to improve the quality of life of hearing aid users by automatically inferring the optimal hearing aid configuration, and ii) to increase the synergy between the hearing aid, the cloud, and the user.

APPENDIX A

AUDIOSENSE SUBJECTIVE ASSESSMENT FLOW

The AudioSense system captured the acoustic context details and hearing aid performance by delivering electronic surveys. The survey captured details of the acoustic context by asking the user to report details of their activity context, noise level, level of familiarity with the talker, presence of visual cues etc. Additionally, the survey also asked the user to report their perception of the performance of their hearing aids across several dimensions like Speech Perception (SP), Listening Effort (LE), Satisfaction with their device (ST) etc. Each report was associated with a measure of importance which represented how important was listening well in the reported context for the user. The following tables represent the questions that are asked in our AudioSense survey.

1. Table A.1: The assessment begins with instructing the user to report the listening event that was happening around them in the past 5 – 10 minutes. If the assessment was user initiated, the survey asks the user to report the time around which the event being reported occured.

2. Table A.2, A.3, A.4, A.5, A.6: Once the instructions have been acknowledged by the user, the survey delivers questions asking about the i) activity, ii) location, iii) presence of visual cues, iv) details about the location of the talker, v) noise level, vi) noise location, vii) room size and carpeting for estimating reverbera-

Q. No.	Conditions	Question Text	Options
I1		When did the event end?	1) Less than 1 hour ago 2) More than 1 hour ago
I2		Choose the activity and condition that occurred most of the time in the event.	

Table A.1. Initiation of the survey and instructions.

Q. No.	Conditions	Question Text	Options
AC1		Were you listening to speech?	1) Yes 2) No
AC2	AC1 = 1	What were you listening to?	1) Conversation, 3 or fewer 2) Conversation, 4 or more 3) Speech listening, live 4) Speech listening, media 5) Conversation, phone
AC2	AC1 = 2	What were you listening to?	6) Non-speech sound listening 7) Not actively listening

Table A.2. Survey questions about acoustic activity of HA user.

tion. Questions within this group depend on how previous questions have been answered.

3. Table A.7: Depending on what responses were given to the previous question group, the survey delivers questions about the performance of the HA. All of the responses are recorded on a 100 point scale to capture the perception.

Q. No.	Conditions	Question text	Options
LC1		Where were you?	1) Outdoor/Traffic 2) Indoors
LC2	LC1 = 1	Please be more specific	1) Outdoor, moving traffic 2) Outdoor, other than traffic
LC2	LC1 = 2	Please be more specific	3) Home, 10 or fewer people 4) Other than home, 10 or fewer people 5) Crowd of people, >10

Table A.3. Survey questions about location of HA user.

Q. No.	Conditions	Question Text	Options
TF	AC2 = 1 - 5	Where you familiar with the talker?	1) Unfamiliar 2) Somewhat unfamiliar 3) Somewhat familiar 4) Familiar
VC	AC2 = 1 - 5	Could you se the talker's face?	1) No 2) Yes, but only sometimes 3) Almost always
TL	AC2 = 1 - 5	Where was the talker most of the time?	1) Front 2) Side 3) Back

Table A.4. Survey questions about details of talker and visual cue availability.

Q. No.	Conditions	Question Text	Options
NZ1		How noisy was it during the listening event?	1) Quiet 2) Somewhat noisy 3) Noisy 4) Very noisy
NZ2	NZ1 = 2 - 4	Where was the noise most of the time?	1) Front 2) Side 3) Back 4) All around

Table A.5. Survey questions about details of noise level and location.

Q. No.	Conditions	Question Text	Options
RS1	LC1 = 2	Compared to an average living room, how large was the room?	1) Smaller 2) About average 3) Larger
RS2	LC1 = 2	Was there carpeting?	1) Yes 2) No

Table A.6. Survey questions about details of room size and carpeting for estimating reverberation.

Q. No.	Conditions	Question Text	Options
SP	AC2 = 1 - 5	How much speech did you understand?	0 = 0% 100 = 100%
LE	AC2 = 1 - 6	How much effort was required to listen effectively?	0 = Very easy 100 = Very effortful
LD1		How would you judge the level of loudness of the sound?	0 = Very soft 50 = Comfortable 100 = Uncomfortably loud
LD2		Were you satisfied with the loudness?	0 = Not good at all 100 = Just right
LCL		Could you tell where the sounds were coming from right away?	0 = Not at all 100 = Perfectly
ST		Were you satisfied with your hearing aids?	0 = Not al all 100 = Very satisfied
AP		With your hearing aids, how much have your hearing difficulties affected what you wanted to do during the listening event?	0 = Not at all 100 = Very much
QoL		Were you happy during the listening event?	0 = Not at all 100 = Very much
IM		How important was it for you to hear well during the listening event?	0 = Not at all 100 = Very much

Table A.7. Survey questions to capture the user's perception of their device's performance. All the answers are on a 100 point scale.

APPENDIX B

ON THE COLLECTED DATA

Using AudioSense we collected subjective and objective data. The subjective data was collected in the form of electronic surveys with a dynamic structure, as described in Appendix A. The objective data was collected in the form of contextual audio on the phone and LENA, and as GPS coordinates on the phone. This appendix gives details about:

1. The amount of data collected
2. The format in which the various data streams were stored
3. Data anomalies and how they were handled

B.1 Amount of data collected

We collected real-world data from two primary sources *viz.* the mobile phone and the LENA device. The mobile phone sampled data using either:

- User-Initiated Protocol: The study participant initiated the assessment which led to the recording of responses from the electronic surveys, audio, and GPS information from the time the assessment was initiated.
- Timer-Initiated Protocol: The internal timer of the AudioSense app delivered assessments that the participant could respond to or ignore. If the participant did not respond to the delivered alarm the phone only collected audio and GPS

Source	Data Type	Data Points Collected
Phone	Survey	14946
	Audio	20409
	GPS	17150
LENA	Audio	1718

Table B.1. Amount of data collected from the phone and LENA device

data. In case of a response from the participant the phone also collected survey responses.

In addition to this, the study participants had the option of wearing the LENA device around their neck everyday. The device recording audio continuously from the time it was switched on until it was switched off. The participants were given one device for each day they were part of the study.

With this setup, to the best of our knowledge, we collected the largest dataset of its kind for evaluating HAs using mEMA. The details of the amount are given in Table B.1. The amount of audio and GPS data is more than the number of surveys because the participants had the option of not responding to the assessments, or pressing the *Snooze* button. Sometimes the phone was unable to get a GPS lock and hence was unable to collect GPS coordinates. In terms of the audio recorded on LENA devices, some participants (i) declined to wear the LENA devices, (ii) forgot to wear them, or (iii) their devices malfunctioned resulting in less than the maximum possible number (approximately 2300) of recordings from LENA.

B.2 Data Format

Each data stream is stored in a specific format balancing two criteria *viz.* data has to be interpretable, and consume less disk space since lack of network connectivity can lead to data being cached on the phone.

B.2.1 Phone

B.2.1.1 Survey

The responses to the survey questions were stored as a plain text CSV with timestamps associated with each response. The CSV was formatted as `< Response Type, Response >`. The *Response Type* were keywords representing each question asked (refer to Appendix A for details) and the responses were the option number that was selected or the value selected on the 0–100 scale for outcome scores.

B.2.1.2 Audio

The audio data was streamed at 16000 Hz on the phone since it covers most of the human sounds, which tend to lie within the [0, 8000] Hz range. Since a significant amount of data is generated per-second from the microphone (16000 samples) we extracted it as 16-bit PCM and stored it as raw `short` data type in the little-endian format.

B.2.1.3 GPS

The GPS location is stored as a CSV in the format `< Latitude, Longitude, Accuracy >`. We sample the location once every 10s i.e. at 0.1 Hz. The low sampling rate of the GPS sensor is motivated by the high power consumption of the sensor

which requires acquiring a satellite lock for identifying location. The noise within the GPS coordinates varies depending on the location of the device i.e. whether it is indoors or outdoors, in a dense building environment with skyscrapers or open fields etc. Dense environments like urban skyscraper settings, or being indoors generally have poor GPS location resolution, to account for this we also stored the *accuracy* of the location (in meters) that android determines based on the lock.

B.2.2 LENA

The LENA device is worn around the neck and records data continuously for the duration it is switched on (8-10 hours). The data was streamed at 16000 Hz and stored as a `.wav` file. In addition to the audio, the LENA also annotated the data with conversations, speaker identification etc. which were stored as an XML with the extension `.trs`. During the post-processing and analysis of the dataset we only utilized the audio file completely. The transcriber files were used to align the timings of the audio recordings from the phone with the continuous LENA recordings.

B.3 Anomalies within subjective data

During the early stages of data collection, on analysis of the outcome scores, we noticed that a significant fraction of the outcome scores possessed the value 50 (Figure B.1). We believed that this was due to the following reasons:

1. All outcome score based questions within the electronic survey were presented with an already selected value of 50.

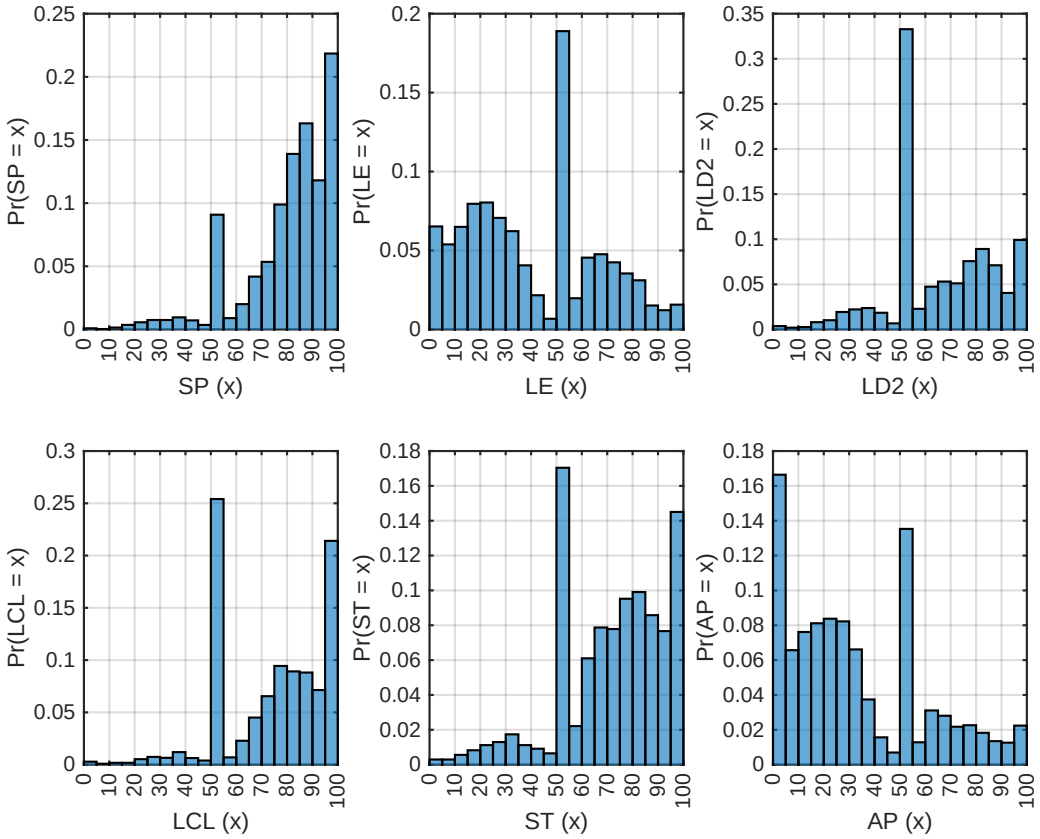


Figure B.1. Distribution of individual outcome scores. The x-axis represents the value of the outcome score and the y-axis represents the fraction of samples containing that value. SP is the speech-perception, LE is listening effort, LD2 is satisfaction with loudness, LCL is localization ability, ST is satisfaction with HA, AP is effect of HA on activity participation. The anomaly is the spike in data at the value 50.

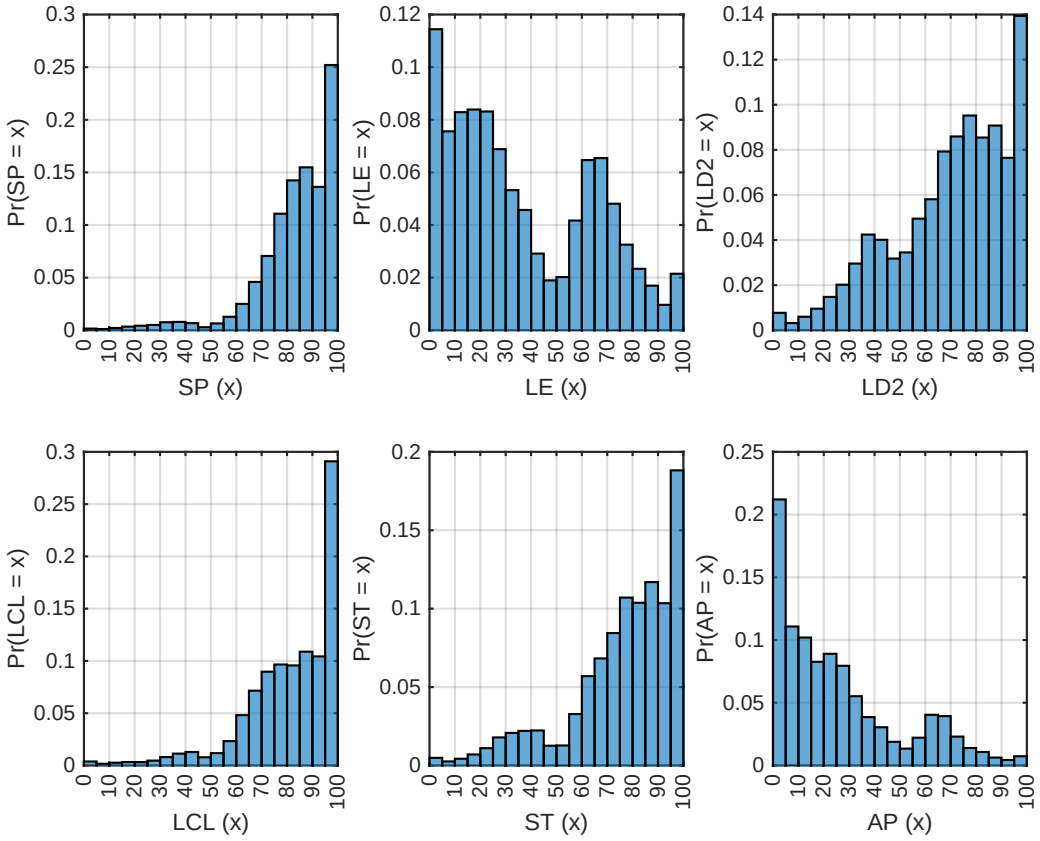


Figure B.2. Distribution of individual outcome scores after the software patch was issued. The x and y axes represent the same values as Figure B.1. The lack of the spike in data at the value 50 indicates that it was an effect of the flaw in the survey design.

2. The user could move to the next question without moving the slider.

We corrected this issue by not showing the slider at the default value and making it necessary for the study participants to move the slider for proceeding to the next question. The outcome scores collected after the issuing of the software patch indicated that the spike was indeed an effect of the flaw in the user-interaction process. The distribution of data collected after the patch is shown in Figure B.2. For all the analysis reported in

this thesis we exclude all samples where a 50 value was present in *any* outcome score before the issuing of the software patch to maintain consistency within the data and remove potential bias.

APPENDIX C

ACOUSTIC FEATURE EXTRACTION

We extracted both time and frequency domain features from the audio data.

C.1 Frame-Level Features

We divided the data into frames of 128ms without any overlap. Each of the features mentioned below were extracted over every frame.

C.1.1 Zero Crossing Rate

The Zero-Crossing Rate (ZCR) is calculated in the time domain and is defined as the number of times the signal changes its sign (or crosses the zero) per frame. The ZCR has been shown to be a good discriminator between speech and music by having high and low values respectively [59]. The ZCR is calculated as follows:

$$ZCR = \frac{\sum_{i=0}^{i=n} abs(sign(x_i) - sign(x_{i-1}))}{2} \tag{C.1}$$

Where x_i, x_{i-1} are signal amplitudes, n is the size of the frame, $sign()$ returns $+1$ for positive values, and -1 for negative values.

C.1.2 Root Mean Squared Amplitude

The Root Mean Square of the amplitude is calculated as a proxy for the signal energy. Since speech contains periods of quiet the RMS value is lower than those signals where quiet periods occur with lower frequency like music. This has been shown to be helpful in discriminating between speech and music signals [16].

C.1.3 Pitch

We calculate the pitch as an input for the estimation of number of speakers using the Crowd++ algorithm [71]. The pitch is calculated using YIN [15] which is simpler, and more robust to noise than competing algorithms like Wu [67] and SAcC [38].

C.1.4 Mel-Frequency Cepstral Coefficients

The Mel-Frequency Cepstral Coefficients (MFCCs) are designed to imitate human hearing by modulating the filter bank width as the frequency under consideration increases. They are calculated in the frequency domain, specifically on the mel-scale which is a logarithmic scale and have been shown to very powerful in auditory scene recognition [52]. We calculated 26 MFCCs per-frame for our context-recognition pipeline.

C.1.5 Spectral Entropy

In order to compute the spectral entropy the probability mass function (PMF) is computed for each frame using Equation C.2.

$$p_i = \frac{X_i}{\sum_{i=1}^{N} X_i} \tag{C.2}$$

Here X_i is the energy of the i^{th} frequency component. We use the PMF to compute the spectral entropy using Equation C.3 [46].

$$H = -\sum_{i=1}^{N} p_i \cdot log_2 p_i \tag{C.3}$$

A peaky spectrum will have a lower entropy and might be representative of voice or music signals while a relatively flat spectrum will have high entropy representing a noisy signal.

C.1.6 Spectral Rolloff

Spectral Rolloff is defined as the frequency bin below which X% of the distribution is concentrated. We used X to be 93 as used in [39, 42]. This measure can be helpful in identifying music as musical signals generally have a greater number of higher frequency components and hence have a greater spectral rolloff.

C.1.7 Sub-band Energy & Entropy

The spectrum for each frame is further sub-divided into frequency bands of 1000Hz to extract fine-grained spectral nuances. For each of these bands the energy and entropy are computed. Sub-band features like energy and entropy have been shown to be informative discriminating between various acoustic activities in areas like speech recognition [46].

C.2 High-Level Features

Once the computation of frame-level features is complete, we are left with multiple vectors of representing frame level details. We reduce the granularity of the features from frame level to the file level by computing a variety of summary statistics over each of the features. This reduces the frame level matrix for a given audio file to a feature vector. In order to capture a comprehensive picture of the variations within the

individual features we compute the following statistics over them:

- **Extremes**: Minimum, Maximum
- **Aggregate**: Mean
- **Variation**: Standard Deviation, Skewness, Kurtosis
- **Percentile**: 1^{st} Quartile, 3^{rd} Quartile, Median

C.3 Signal to Noise Ratio

We computed the signal to noise ratio using off-the-shelf algorithms like the NIST SNR [2], WADA SNR [35], and the VAD SNR [4].

C.3.1 NIST SNR

The NIST SNR evaluates the SNR by computing the RMS power histogram of the audio signal. The method estimates the noise power by fitting a raised cosine to the histogram. The noise power is then subtracted from the composite signal power histogram to obtain the clean signal power.

C.3.2 WADA SNR

The WADA (Waveform Amplitude Distribution Analysis) SNR estimates the clean signal by modeling it as a Gamma distribution. The noise is assumed to be Gaussian.

C.3.3 VAD SNR

The VAD (Voice Activity Detection) SNR identified the portions of the input signals where speech activity is present to calculate the SNR.

We encourage the reader to explore the details of these SNRs in their original publications.

REFERENCES

[1] Empatica e4 wrist band, March. https://www.empatica.com/e4-wristband.

[2] The nist speech snr measurement, September. http://www.nist.gov/smartspace/nist_speech_snr_measurement.html.

[3] Open speech platform, March. http://openspeechplatform.ucsd.edu/.

[4] Snr evaluation tools. https://labrosa.ee.columbia.edu/projects/snreval.

[5] World health organization: Deafness and hearing loss, April. http://www.who.int/mediacentre/factsheets/fs300/en/.

[6] Untreated hearing loss in adults–a growing national epidemic, January 2012. http://www.asha.org/Articles/Untreated-Hearing-Loss-in-Adults/.

[7] Use of hearing aids by adults with hearing loss, September 2014. https://www.nidcd.nih.gov/health/statistics/use-hearing-aids-adults-hearing-loss.

[8] Nidcd statistics about hearing, December 2016. https://www.nidcd.nih.gov/health/statistics/quick-statistics-hearing.

[9] Stig Arlinger. Negative consequences of uncorrected hearing loss-a review. *International journal of audiology*, 42:2S17–2S20, 2003.

[10] Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, et al. Activity sensing in the wild: a field trial of ubifit garden. In *Conference on Human Factors in Computing Systems (SIGCHI)*, 2008.

[11] Mary T Cord, Rauna K Surr, Brian E Walden, and Laurel Olson. Performance of directional microphone hearing aids in everyday life. *Journal of the American Academy of Audiology*, 13(6):295–307, June 2002.

[12] 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.

[13] K J Cruickshanks, T L Wiley, T S Tweed, B E Klein, R Klein, J A Mares-Perlman, and D M Nondahl. Prevalence of hearing loss in older adults in beaver dam, wisconsin. the epidemiology of hearing loss study. *Am J Epidemiol*, 148(9):879–886, Nov 1998.

[14] Dayna S Dalton, Karen J Cruickshanks, Barbara E K Klein, Ronald Klein, Terry L Wiley, and David M Nondahl. The impact of hearing loss on quality of life in older adults. *Gerontologist*, 43(5):661–668, Oct 2003.

[15] Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[16] Yariv Ephraim and David Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6):1109–1121, 1984.

[17] Melinda C. Freyaldenhoven, Anna K Nabelek, and Joanna W. Tampas. Relationship between acceptable noise level and the abbreviated profile of hearing aid benefit. *Journal of Speech, Language, and Hearing Research*, 51:136–146, 2008.

[18] Jon Froehlich, Mike Y. Chen, Sunny Consolvo, Beverly Harrison, and James A. Landay. Myexperience: A system for in situ tracing and capturing of user feedback on mobile phones. In *MobiSys '07*, pages 57–70, 2007.

[19] Gino Galvez, Mitchel B Turbin, Emily J Thielman, Joseph A Istvan, Judy A Andrews, and James A Henry. Feasibility of ecological momentary assessment of hearing difficulties encountered by hearing aid users. *Ear and Hearing*, 33(4):497–507, Jul-Aug 2012.

[20] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

[21] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Mach. Learn.*, 63(1):3–42, April 2006.

[22] Penny Anderson Gosselin and Jean-Pierre Gagne. Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language, and Hearing Research*, 54(3):944–958, 2011.

[23] Syed Shabih Hasan, Ryan Brummet, Octav Chipara, Yu-Hsiang Wu, and Tianbao Yang. In-situ measurement and prediction of hearing aid outcomes using mobile phones. In *International Conference on Healthcare Informatics (ICHI)*, 2015.

[24] Syed Shabih Hasan, Octav Chipara, Yu-Hsiang Wu, and Nazan Aksan. Evaluating auditory contexts and their impacts on hearing aid outcomes with mobile phones. In *PervasiveHealth*, pages 126–133, 2014.

[25] Syed Shabih Hasan, Farley Lai, Octav Chipara, and Yu-Hsiang Wu. AudioSense: Enabling real-time evaluation of hearing aid technology in-situ. In *CBMS '13*, pages 167–172, 2013.

[26] James A Henry, Gino Galvez, Mitchel B Turbin, Emily J Thielman, Garnett P McMillan, and Joseph A Istvan. Pilot study to evaluate ecological momentary assessment of tinnitus. *Ear Hear*, 33(2):179–290, Mar-Apr 2012.

[27] John Hicks, Nithya Ramanathan, Donnie Kim, Mohamad Monibi, Joshua Selsky, Mark Hansen, and Deborah Estrin. AndWellness: an open mobile system for activity and experience sampling. *Wireless Health 2010*, pages 34–43, October 2010.

[28] Louise Hickson, Carly Meyer, Karen Lovelock, Michelle Lampert, and Asad Khan. Factors associated with success with hearing aids in older adults. *International journal of audiology*, 53(S1):S18–S27, 2014.

[29] Hsu-Chueh Ho, Yu-Hsiang Wu, Shih-Hsuan Hsiao, and Xuyang Zhang. Acceptable noise level (ANL) and real-world hearing-aid success in Taiwanese listeners. *International Journal of Audiology*, 52(11):762–770, November 2013.

[30] Joyce Ho and Stephen S Intille. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *SIGCHI*, 2005.

[31] F. B Hobbs. *Population profile in the United States*. US Department of Commerce: US Census Bureau, 2010.

[32] Benjamin WY Hornsby. The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands. *Ear and Hearing*, 34(5):523–534, 2013.

[33] Karen Hovsepian, Mustafa al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. cstress: towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 493–504. ACM, 2015.

[34] Mead C Killion, Patricia A Niquette, Gail I Gudmundsen, Lawrence J Revit, and Shilpi Banerjee. Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 116(4):2395–2405, 2004.

[35] Chanwoo Kim and Richard M Stern. Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In *INTERSPEECH*, 2008.

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

[37] Sophia E Kramer, Charlotte E Teunissen, and Adriana A Zekveld. Cortisol, chromogranin a, and pupillary responses evoked by speech recognition tasks in normally hearing and hard-of-hearing listeners: a pilot study. *Ear and Hearing*, 37:126S–135S, 2016.

[38] Byung Suk Lee and Daniel PW Ellis. Noise robust pitch tracking by subband autocorrelation classification. In *Interspeech*, pages 707–710, 2012.

[39] Dongge Li, Ishwar K. Sethi, Nevenka Dimitrova, and Tom McGee. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 22(5):533 – 544, 2001.

[40] Frank R Lin, Roland Thorpe, Sandra Gordon-Salant, and Luigi 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.

[41] Frank R Lin, Kristine Yaffe, Jin Xia, Qian-Li Xue, Tamara B Harris, Elizabeth Purchase-Helzner, Suzanne Satterfield, Hilsa N Ayonayon, Luigi Ferrucci, Eleanor M Simonsick, et al. Hearing loss and cognitive decline in older adults. *JAMA internal medicine*, 173(4):293–299, 2013.

[42] Hong Lu, Wei Pan, Nicholas D Lane, Tanzeem Choudhury, and Andrew T Campbell. SoundSense: scalable sound sensing for people-centric applications on mobile phones. In *MobiSys '09: Proceedings of the 7th international conference on Mobile systems, applications, and services*, pages 165–178, New York, New York, USA, June 2009. ACM.

[43] Carol L Mackersie and Natalie Calderon-Moultrie. Autonomic nervous system reactivity during speech repetition tasks: Heart rate variability and skin conductance. *Ear and Hearing*, 37:118S–125S, 2016.

[44] Carol L Mackersie, Imola X MacPhee, and Emily W Heldt. Effects of hearing loss on heart-rate variability and skin conductance measured during sentence recognition in noise. *Ear and hearing*, 36(1):145, 2015.

[45] Abhinav Mehrotra, Mirco Musolesi, Robert Hendley, and Veljko Pejovic. Designing content-driven intelligent notification mechanisms for mobile applications. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 813–824. ACM, 2015.

[46] Hemant Misra, Shajith Ikbal, Hervé Bourlard, and Hynek Hermansky. Spectral entropy based feature for robust asr. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I–193. IEEE, 2004.

[47] Anna K Nabelek, Melinda C Freyaldenhoven, Joanna W Tampas, Samuel B Burchfield, and Robert A Muenchen. Acceptable noise level as a predictor of hearing aid use. *Journal of the American Academy of Audiology*, 17(9):626–639.

[48] Shahriar Nirjon, Robert F Dickerson, Philip Asare, Qiang Li, Dezhi Hong, John A Stankovic, Pan Hu, Guobin Shen, and Xiaofan Jiang. Auditeur: a mobile-cloud service platform for acoustic event detection on smartphones. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pages 403–416. ACM, 2013.

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

[50] M Paez and T Glisson. Minimum mean-squared-error quantization in speech pcm and dpcm systems. *IEEE Transactions on Communications*, 20(2):225–230, 1972.

[51] Veljko Pejovic and Mirco Musolesi. Interruptme: Designing intelligent prompting mechanisms for pervasive applications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 897–908. ACM, 2014.

[52] Vesa Peltonen, Juha Tuomi, Anssi Klapuri, Jyri Huopaniemi, and Timo Sorsa. Computational auditory scene recognition. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 2, pages II–1941. IEEE, 2002.

[53] Laura Pina, Kael Rowan, Asta Roseway, Paul Johns, Gillian R Hayes, and Mary Czerwinski. In situ cues for adhd parenting strategies using mobile technology. In *International Conference on Pervasive Computing Technologies for Healthcare*, 2014.

[54] Thomas Probst, Rüdiger Pryss, Berthold Langguth, and Winfried Schlee. Emotional states as mediators between tinnitus loudness and tinnitus distress in daily life: Results from the "trackyourtinnitus" application. *Scientific reports*, 6, 2016.

[55] 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.

[56] Lawrence R Rabiner and Ronald W Schafer. *Digital processing of speech signals*. Prentice Hall, 1978.

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

[58] Anastasios Sarampalis, Sridhar Kalluri, Brent Edwards, and Ervin Hafter. Objective measures of listening effort: Effects of background noise and noise reduction. *Journal of Speech, Language, and Hearing Research*, 52(5):1230–1240, 2009.

[59] John Saunders. Real-time discrimination of broadcast speech/music. In *icassp*, volume 96, pages 993–996, 1996.

[60] Saul Shiffman, Arthur A Stone, and Michael R Hufford. Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4:1–32, April 2008.

[61] William J Strawbridge, Margaret I Wallhagen, Sarah J Shema, and George A Kaplan. Negative consequences of hearing impairment in old age a longitudinal analysis. *The Gerontologist*, 40(3):320–326, 2000.

[62] Brian Taylor. The acceptable noise level test as a predictor of real-world hearing-aid benefit. *The Hearing Journal*, 61(9):39–42, 2008.

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

[64] 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.

[65] Therese C. Walden and Brian E. Walden. Predicting success with hearing aids in everyday living. *Journal of the American Academy of Audiology*, 15:342–352, 2004.

[66] Michael B Wilson, Dorina Kallogjeri, Conor N Joplin, Mitchell D Gorman, James G Krings, Eric J Lenze, Joyce E Nicklaus, Edward E Jr Spitznagel, and Jay F Piccirillo. Ecological momentary assessment of tinnitus using smartphone technology: a pilot study. *Otolaryngol Head Neck Surg*, 152(5):897–903, May 2015.

[67] Mingyang Wu, DeLiang Wang, and Guy J Brown. A multipitch tracking algorithm for noisy speech. *IEEE Transactions on Speech and Audio Processing*, 11(3):229–241, 2003.

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

[69] Yu-Hsiang Wu and Ruth A Bentler. Impact of Visual Cues on Directional Benefit and Preference: Part I—Laboratory Tests. *Ear and Hearing*, 31(1):22–34, February 2010.

[70] Yu-Hsiang Wu and Ruth 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.

[71] Chenren Xu, Sugang Li, Gang Liu, Yanyong Zhang, Emiliano Miluzzo, Yih-Farn Chen, Jun Li, and Bernhard Firner. Crowd++: unsupervised speaker count with smartphones. In *UbiComp '13: Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. AT&T Laboratories Florham Park, ACM, September 2013.

[72] Lide Zhang, Birjodh Tiwana, Zhiyun Qian, Zhaoguang Wang, Robert P. Dick, Zhuoqing Morley Mao, and Lei Yang. Accurate online power estimation and automatic battery behavior based power model generation for smartphones. In *CODES/ISSS '10*, 2010.