Assessing the Performance of Hearing Aids using Surveys and Audio Data Collected In Situ

(Invited Paper)

Syed Shabih Hasan, Ryan Brummet, Octav Chipara, Yu-Hsiang Wu†
Dept. of Computer Science †Dept. of Communication Sciences & Disorders
University of Iowa

***Abstract*—Measuring the performance of hearing aids in the real-world is challenging because it depends on both the subject's auditory abilities and the auditory context in which they use their hearing aids. The auditory context includes characteristics of the listening activity, social context, and acoustic environment. Ecological momentary assessment is the state-of-the-art method for evaluating the performance of hearing aids in the real-world. It may be used to assess the auditory context and its performance in that context. However, such techniques a introduces a significant data collection burden on study participants. A potential avenue for reducing the data collection burden is to use audio data collected in-situ to characterize the auditory context automatically. This paper presents our initial attempt at answering this challenge. Our results indicate that it is possible to predict the noise level and listening activity with an accuracy of 68% and 70%, respectively. Additionally, we show that a 4.6% reduction in the mean squared error of predicting the listening effort may also be achieved by using the audio data (without including subjective self-reports).**

I. Introduction

Twenty percent of Americans will be 65 years or older by 2030 [1] out of which between 35% and 50% will report having presbycusis [2], 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 and depression [3], [4]. Unfortunately, many subjects that would benefit from HAs do not use them regularly, 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 encounter 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 [5]–[7].

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 [18]. 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 paper is to explore how the audio the data gathered by AudioSense may be used. We are interested in this problem for two reasons. First, collecting 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. 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 *subjective* 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.

II. 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) [8], [9]. However, studies have shown that HA performance measured in the lab is a poor predictor of the real-world HA performance. [7], [11]. More recently, Ecological Momentary Assessment (EMA) [12] 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 [13]–[16]. The use of computer-based EMA in Audiology is still in its infancy with a few studies evaluating HAs [17], [18] and tinnitus [19]–[21]. The AudioSense system [18] 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 [22], [23].

Despite these advances, to the best of our knowledge, no work exists that utilizes 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 [24] and background environmental information such as signal-to-noise ratio [25]. 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 question that they are asked and (2) it is possible to construct intelligent sampling policies in contexts that may be of interests to audiologists (e.g., low SNR, when conversation are present).

Variable	Statistics
Gender Distribution	51% Female
Age (years)	Range: 64 - 88 Median: 72.5
Hearing Loss Onset (years)	Range: 1 - 54 Median: 8
Duration of HA use (years)	Range: 0-40 Median: 6

TABLE I: Demographic Details of Participants

Condition	HA use	DM/DNR usage
1	Entry level	Off
2	Entry level	On
3	Premium	Off
4	Premium	On

TABLE II: Study conditions

III. Field Study

The participants for the study are recruited in three ways: (i) through the pool of potential participants maintained by the Department of Communication Sciences & Disorders, (ii) through word of mouth from participants of other studies, and (iii) through hearing screenings performed within associated clinics. We only recruit participants older than 64 years of age that have mild-to-moderate hearing loss. Furthermore, we ensure that subjects admit in the study have hearing loss that is symmetric, bilateral, and has an adult-onset. The demographic details are given in Table I.

During the study, each subject participated in four data collection conditions that involved two different HAs each having two possible configurations. We used two types of HAs: entry-level and premium-level. Both the HAs had advanced processing modes like directional microphones (DN) and digital noise reduction (DNR) which could be switched on or off. The details of each condition are summarized in Table II. Every participant started with a training session where the participants became acclimatized with reporting data via the mobile phone application. The order of the subsequent conditions was randomized. Each week long data collection session was preceded by a month-long washout period during which they wore a new HA (from Table II) to (i) familiarize themselves with the HA, and (ii) to minimize reporting fatigue. Once a data collection condition was over the participants began the washout period for the next session. The study was single-blinded: participants did not know which HA they used.

The mobile phone application was used to collect the HA performance from the user via the delivery of assessments in conjunction with surrounding environmental audio. The assessments were delivered in one of two ways: (i) participant-initiated: the participant initiated the assessment to report the performance, or (ii) timer-initiated: we used a semi-randomized timer to schedule assessment delivery. The semi-randomized timer had a fixed component to ensure that two assessments were separated by a minimum time and a random component. Additionally, clinicians configured the start and end times when assessments were delivered during the day. The audio data used in this work was recorded on a sound recorder that subjects carried around their necks. We have opted to use this setup due to concerns that the sound recorded from the phone might not always be representative when a subject does not carry the phone with them. Additional details on AudioSense may be found in [18], [22].

Subjective Data: The mobile phone collects several subjective measures that assess the performance of the HA and the context in which the HA is used. Participants were asked to evaluate the HA performance in the last 5 — 10 minutes before the delivery of the survey. The context in which the HA is used is characterized using multiple metrics. However, in this paper, we will focus on the reported noise level and activity type as audiology studies have indicated that these factors may affect a subject's satisfaction with their HA.

Objective Data: The audio data used in this work was recording on a sound recorder that subjects carried around their necks. We have opted to use this setup due to concerns that the sound recorded from the phone might not always be representative when a subject does not carry the phone with them. For each survey, we have identified a five-minute audio clip recorded prior to the delivery of a survey. The audio data was recorded at 16KHz. The data is divided into a 10-millisecond frame from which low-level features were extracted. The low-level features include the following time- and frequency-domain features: Zero-Crossing Rate, Pitch, Spectral Entropy, and Mel-Frequency Cepstral Coefficients. The low-level features were then used to compute high-level features by using the following summary statistics: min, max, mean, median, standard deviation, skewness, and kurtosis. We have also included three additional high-level features which measure the signal-to-noise ratio in the captured sound clip using different algorithms.

Data Included: The data analyzed includes only the conditions when the HA were used, excluding data from the training and the unaided conditions. Additionally, as part of every survey (including those delivered during aided conditions), the subject 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.

IV. 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, due to the space constraints, we results obtained using extremely randomized trees [26]. 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 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. 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.

A. 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 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 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.

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 [25] 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

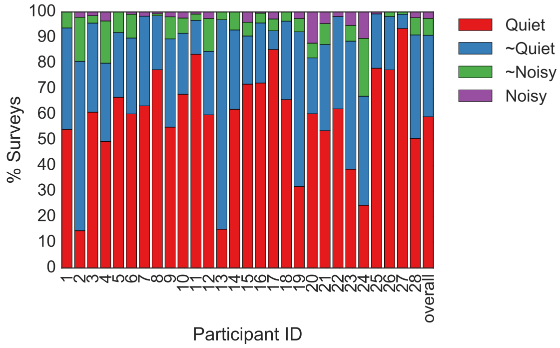
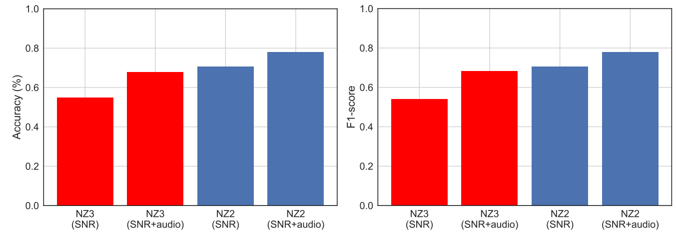


Fig. 1: Noise level per subject



(a) Accuracy (b) F1-score

Fig. 3: Predicting NZ2 and NZ3 from SNR and audio data.

[27], [28]. The noise is assumed to be Gaussian. VAD SNR [29] applies the same SNR estimation only to those segments where the presence of speech is detected. Figure 2 plots the distribution of estimated SNR for of the noise levels. All three estimator 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.

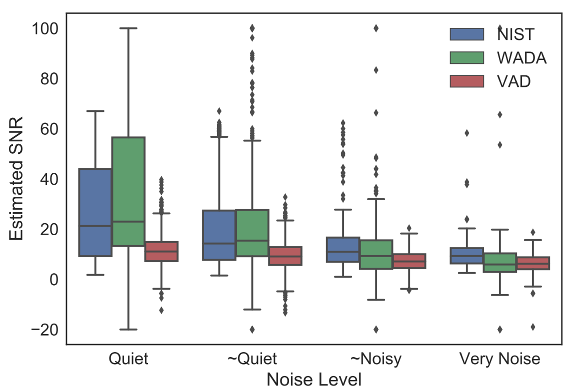


Fig. 2: Estimated SNR

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 3 plots the accuracy and F1-score for the NZ2 and NZ3 estimators when using SNR 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.

B. 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 4 plots the distribution of activities in which the subjects engage in. Subjects spent about 18% of the time listening passively. 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.

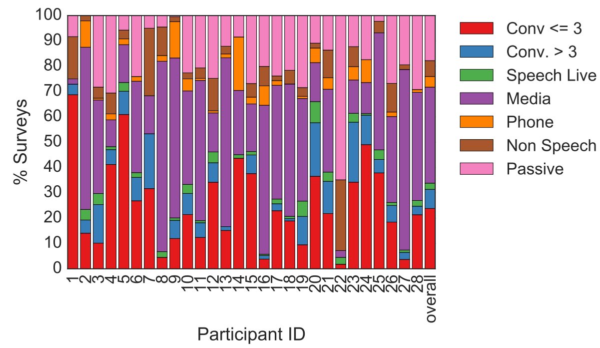


Fig. 4: Distribution of listening activites

Figure 5 shows the confusion matrix for the classifier. Overall, the classifier is 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.

C. Predicting the Listening Effort

Listening effort is a sensitive measure of the performance of the HA, particularly in speech. Figure 6a 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

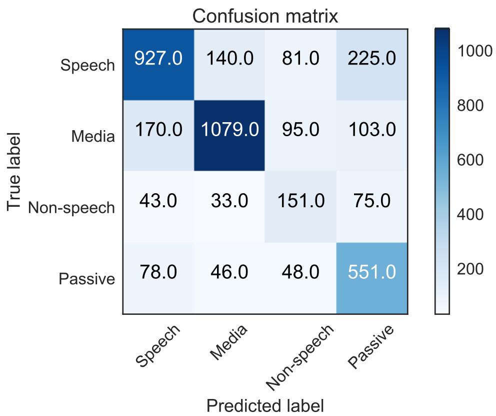
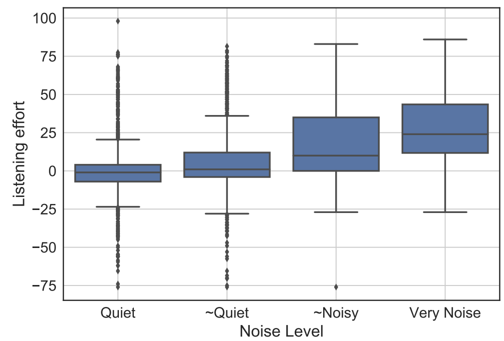


Fig. 5: Confusion matrix for listening activity

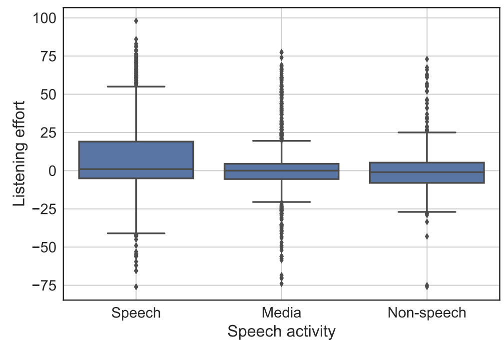
of the group. This scaling allows us to interpret positive values as requiring more effort than the average. Conversely, negative values indicate they require less effort than the average. In quiet, the subjects require a less listening effort to hear well. This is clear from the slightly below zero 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 is unsatisfied with the performance of their HAs in noise. Figure 6a 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 satisfaction 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 satisfaction. The baseline is a classifier that uses the subjective values of the noise level and listening activity as reported by the user.

Figure 7 plot the predictions satisfaction 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 use only the subject and condition identifiers. This classifier



(a) Noise Level



(b) Listening activity

Fig. 6: The impact of noise level and listening activity on listening effort

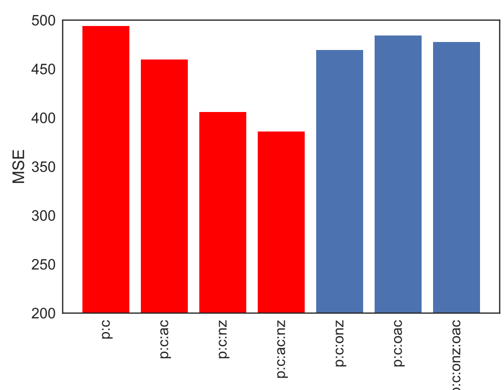


Fig. 7: Prediction of satisfaction using different features

essentially predicts that mean of satisfaction of that subject for the considered HA. The performance of the classifiers may be improved by considering 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.

V. Conclusions

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 paper, we make an initial 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 68%. 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.

Acknowledgements

The contents of this paper were developed under a grant from the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR grant number 90RE5020). NIDILRR is a Center within the Administration for Community Living (ACL), Department of Health and Human Services (HHS). The contents of this (insert type of publication; e.g., book, report, film) do not necessarily represent the policy of NIDILRR, ACL, HHS, and you should not assume endorsement by the Federal Government

References

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

[2] 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*, vol. 148, no. 9, pp. 879–886, Nov 1998.

[3] R. F. Uhlmann, E. B. Larson, T. S. Ress, T. D. Koepsell, and L. G. Duckert, "Relationship of hearing impairment to dementia and cognitive dysfunction in older adults," *JAMA*, no. 261, pp. 1916–1919, 1989.

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

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

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

[7] Y.-H. Wu and R. A. Bentler, "Impact of visual cues on directional benefit and preference: Part ii field tests," *Ear and hearing*, vol. 31, no. 1, pp. 35–46, 2010.

[8] M. C. Killion, P. A. Niquette, G. I. Gudmundsen, L. J. Revit, and S. 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*, vol. 116, no. 4, pp. 2395–2405, 2004.

[9] A. K. Nabelek, M. C. Freyaldenhoven, J. W. Tampas, S. B. Burchfield, and R. A. Muenchen, "Acceptable noise level as a predictor of hearing aid use," *Journal of the American Academy of Audiology*, vol. 17, no. 9, pp. 626–639, 2006.

[10] J. Johnson, J. Xu, K. Schwartz, and R. Cox, "Development of the sound acceptability test (sat)," *American Auditory Society Convention*, 2012.

[11] Y.-H. Wu and R. A. Bentler, "Impact of visual cues on directional benefit and preference: Part i laboratory tests," *Ear and hearing*, vol. 31, no. 1, pp. 22–34, 2010.

[12] S. Shiffman, A. A. Stone, and M. R. Hufford, "Ecological momentary assessment," *Annu. Rev. Clin. Psychol.*, vol. 4, pp. 1–32, 2008.

[13] S. Consolvo, D. W. McDonald, T. Toscos, M. Y. Chen, J. Froehlich, B. Harrison, P. Klasnja, A. LaMarca, L. LeGrand, R. Libby *et al.*, "Activity sensing in the wild: a field trial of ubifit garden," in *Conference on Human Factors in Computing Systems (SIGCHI)*, 2008.

[14] J. Froehlich, M. Y. Chen, S. Consolvo, B. Harrison, and J. A. Landay, "MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones," in *International conference on Mobile Systems, Applications and Services*. ACM, 2007, pp. 57–70.

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

[16] L. Pina, K. Rowan, A. Roseway, P. Johns, G. R. Hayes, and M. Czerwinski, "In situ cues for adhd parenting strategies using mobile technology," in *International Conference on Pervasive Computing Technologies for Healthcare*, 2014.

[17] G. Galvez, M. B. Turbin, E. J. Thielman, J. A. Istvan, J. A. Andrews, and J. A. Henry, "Feasibility of ecological momentary assessment of hearing difficulties encountered by hearing aid users." *Ear and Hearing*, vol. 33, no. 4, pp. 497–507, Jul-Aug 2012.

[18] S. S. Hasan, F. Lai, O. Chipara, and Y.-H. Wu, "AudioSense: Enabling real-time evaluation of hearing aid technology in-situ," in *International Symposium on Computer-Based Medical Systems*, 2013.

[19] T. Probst, R. Pryss, B. Langguth, and W. Schlee, "Emotional states as mediators between tinnitus loudness and tinnitus distress in daily life: Results from the "trackyourtinnitus" application," *Scientific reports*, vol. 6, 2016.

[20] M. B. Wilson, D. Kallogjeri, C. N. Joplin, M. D. Gorman, J. G. Krings, E. J. Lenze, J. E. Nicklaus, E. E. Spitznagel, and J. F. Piccirillo, "Ecological momentary assessment of tinnitus using smartphone technology a pilot study," *Otolaryngology–Head and Neck Surgery*, 2015.

[21] J. A. Henry, G. Galvez, M. B. Turbin, E. J. Thielman, G. P. McMillan, and J. A. Istvan, "Pilot study to evaluate ecological momentary assessment of tinnitus," *Ear and Hearing*, vol. 32, no. 2, pp. 179–290, Mar. 2012.

[22] S. S. Hasan, O. Chipara, Y.-H. Wu, and N. Aksan, "Evaluating auditory contexts and their impacts on hearing aid outcomes with mobile phones," in *International Conference on Pervasive Computing Technologies for Healthcare*, 2014.

[23] S. S. Hasan, R. Brummet, O. Chipara, Y.-H. Wu, and T. Yang, "In-situ measurement and prediction of hearing aid outcomes using mobile phones," in *International Conference on Healthcare Informatics (ICHI)*, 2015.

[24] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell, "Soundsense: scalable sound sensing for people-centric applications on mobile phones," in *Proceedings of the 7th international conference on Mobile systems, applications, and services*. ACM, 2009, pp. 165–178.

[25] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis." in *INTERSPEECH*, 2008.

[26] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006. [Online]. Available: http://dx.doi.org/10.1007/s10994-006-6226-1

[27] M. Paez and T. Glisson, "Minimum mean-squared-error quantization in speech pcm and dpcm systems," *IEEE Transactions on Communications*, vol. 20, no. 2, pp. 225–230, 1972.

[28] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Prentice Hall, 1978.

[29] "Snr evaluation tools," https://labrosa.ee.columbia.edu/projects/snreval.