Quantifying Hearing Aid Users' Auditory Ecology with a Deep Neural Net

Erik Jorgensen[1], AuD, Dhruv Vyas, MS[2], Yumna Anwar[2], Justin Jensen[1], MA, Octav Chipara[2], PhD, and Yu-Hsiang Wu[1], MD, PhD

[1]Department of Communication Sciences and Disorders, University of Iowa; [2]Department of Computer Science, University of Iowa

Objective

- There is increasing interest in quantifying the listening environments of hearing aid users to better understand the effectiveness of hearing aids in the real world (e.g. Andersson et al., 2020; Jensen & Nielsen, 2005; Wagener et al., 2008; Wu & Bentler, 2012; Wu et al., 2018).

- Common methods used include ecological momentary assessment (EMA), audio recording, and hearing aid classification. EMA lacks acoustic information and the required sampling can be burdensome for participants. Manual analysis of audio recordings is time and labor intensive. Hearing aid classification is proprietary and the accuracy is unknown. **Methods that can provide reliable, transparent, and efficient analyses of large amounts of audio data from hearing aid users are of interest.**

- **This study investigated the use of a widely available deep neural net to quantify a large data set of audio recordings collected from adult hearing aid users**. The purposes of this study were to: 1) explore the feasibility and usefulness of analyzing a large data set of audio recordings from hearing aid users with a deep net; 2) determine whether results of the deep net analysis aligned with prior work using other methods and with participant self-report.

Design

- Data came from a larger study (Wu et al., 2018). 54 adult hearing aid users participated (26 males, 38 females; age range=65-88 years; mean age=73.6 years). 30 were experienced users, 24 were new users. Most were retired: 1 was employed full-time and 7 were employed part-time.

- Participants wore a Language Environment Analysis (LENA) device during waking hours and completed Ecological Momentary Assessments (EMA) on a smartphone for five weeks. The LENA recorded continuously during wear time. These audio recordings provided the data set for this study. A subset of recordings of 5-minutes prior to the delivery of an EMA were extracted and paired with their respective EMA responses.

- **The artificial neural network YAMnet was used to classify recordings**. YAMnet is a deep net that runs on TensorFlow and predicts 521 audio event classes. YAMnet was pre-trained using the AudioSet-YouTube corpus, an ontology of 632 audio event classes and 2,084,320 10-second human-labeled sound clips from YouTube. The output of YAMnet contains a specified number of classifiers and confidence estimates for each audio analysis window, in descending confidence order. The confidence quantifies the proportion of the analysis window associated with each identified audio class. Audio recordings were analyzed in non-overlapping 5-minute windows. For this study, only the top 3 classifiers were examined.

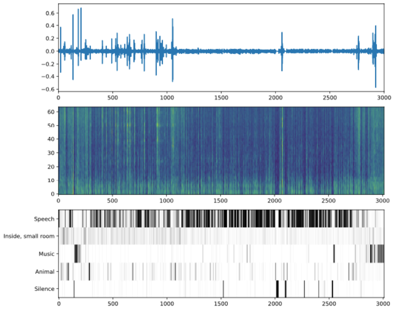

Figure 1. YAMnet output example for 5-miutes of audio (3000 analysis frames).

References: Anderson et al. AJA. 2020. Preprint. Jensen & Nielsen. 21th Danavox Symposium. 2005. Klein et al. JAAA. 2018. 29(4), 279-291. Wagener et al. JAA. 2008. 19(4), 248-370. Wu & Bentler. JAAA. 2012. 23(9), 697-711. Wu et al. EH. 2018. 39(2), 293-204.

Funding: NIH/NIDCD F32DC018980 (Jorgensen); NIH/NIDCD R01DC015997 (Wu), NSF SCH 1838830 (Chipara).

Contact: Erik Jorgensen, University of Iowa, erik-j-jorgensen@uiowa.edu.

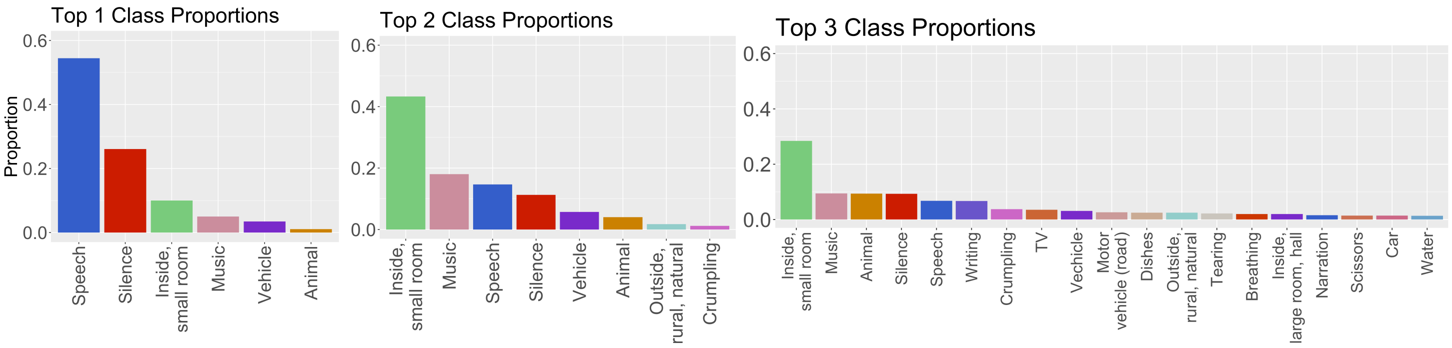

Figure 2 (above). Proportion of top 1, 2, and 3 sound classes across all recordings.

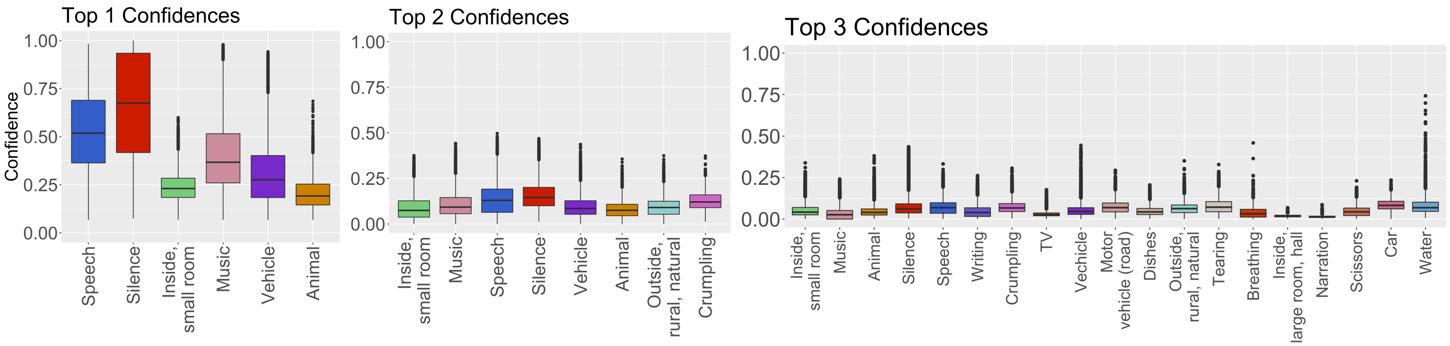

Figure 3 (above). Confidences for top 1, 2, and 3 sound classes across all recordings.

Results

- In total, **24,552 hours of audio recordings were analyzed** by YAMnet. The top 1 classifier set contained 143 classes, the top 2 classifier set contained 312 classes, and the top 3 classifier set contained 374 classes. A 5-minute example (3000 analysis frames) is shown in Figure 1, where the waveform is shown in the top panel, the spectrogram in the middle panel, and the YAMnet classification in the bottom panel (band darkness signifies confidence level). Because of the large number of classes returned, classes accounting for less than 1% of audio events were removed.

- Top 1, 2, and 3 classes are shown above in Figure 2. Each figure shows the proportion of sound classes comprising at least 1% of all classes. **Speech, silence, inside small room, music, vehicle, and animal were primary classes among all of the top 3 classifiers**.

- Top 1, 2, and 3 confidences are shown in Figure 3. The mean confidence for top 1 classes was .52 (range=.07-1.0, sd=.25); the mean confidence for top 2 classes was .11 (range 0-.5, sd=.07); the mean confidence for top 3 classes was .05 (range=0-.7, sd=.04).

- Comparisons of the results obtained in this study with prior studies are shown in Table 2. Studies using varying methods showed broad consensus for proportion of speech-related events in the lives of hearing aid users.

- To determine whether participant report was aligned with YAMnet classification, questions on the EMA were retrospectively identified where participants and YAMnet reported on the same sound classes. These classes were: speech (speech or not speech), room size (small or large), and location (inside, outside, or vehicle). Match rates were calculated by comparing YAMnet classification to participant response. The match rate is the percent of instances where the YAMnet returned a classifier that agreed with the EMA. 894 recordings paired with EMA were analyzed to calculate match rates between participant report on EMA and YAMnet output. These results are shown in Table 1.

Class	Match Rate
Speech	95.7
Small Room	78.3
Large Room	99.7
Outside	99.3
Inside	95.8
Vehicle	95.5

Table 1. Math rates between EMA and YAMnet classes.

Study	Method	# Subjects	# Samples	% Speech	% Silence	% Music
This study	Audio + YAMnet	54	24,552 hours of recording	53	25	5
Andersson et al., 2020	EMA + Hearing Aid Classification	11	~1,930 EMAs + HA class	46	50	NA
Klein et al., 2018	Audio + LENA analysis	22	3,461 hours of recording	54	40	NA
Wu & Bentler, 2012	Daily Diaries	27	~1,268 entries	61	NA	NA
Wagener et al., 2008	Audio	20	349 ~5 minute recordings	51	NA	<9
Jensen & Nielsen, 2005	Audio + EMA	18	330 EMAs	60	NA	5

Table 2. Results comparing speech-related event proportions across studies using varying methods.

Conclusions

- This study used a large data set of unbiased (continually recorded) audio recordings from adult hearing aid users to assess the feasibility of using machine learning to quantify auditory ecology among hearing aid users. YAMnet results were consistent with prior studies using manual methods, based on the top 1 classifiers. Match rates between YAMnet output and participant report were high for the classes evaluated in this study.

- Results support the potential for using machine learning in the evaluation of acoustic scenes in the real world. Machine learning may offer an efficient approach to analysis of large data sets, which are needed to better understand auditory ecology in the hearing loss population. YAMnet also offers more fine-grained acoustic classification than has been possible with other methods. However, there are significant limitations. In particular, important distinctions are not made by YAMnet. For example, based on prior work on these data, it is likely that the "speech" class returned by YAMnet includes speech in quiet, group conversation, speech in noise, and television watching. For comprehensive analyses of auditory ecology, it is likely that multiple methods need to be combined.