APA Citation

Hearing Aid Audio Processing Model Benchmarking with Binary Environmental Classification

Team ECHO:

Rajit Mukhopadhyay, Bhargav Tumkur, Lily Li,

Samuel Waters, Chelsea Reyes, Ronoy Sarkar,

Rahul Nair, Pruthav Patel, & Perfect Sare

University of Maryland, College Park

GEMS 497: Team Dynamics and Research Methodology

Dr. Sahil Shah

May 1, 2025

Review Committee:
Dr. Sahil Shah
Dr. Jonathan Simon
Dr. Eric Hoover
Dr. Matthew Goupell
Dr. Brian Beaudoin

Thesis submitted in partial fulfillment of the requirements of the Gemstone Honors Program, University of Maryland, 2025

Table of Contents

1. Abstract

Current hearing aid audio processing models are trained to filter out noise and amplify specific sounds such as speech. However, real-world audio environments contain a multitude of sounds that users may want to hear that cannot exhaustively be defined within the model. Team ECHO's project proposes a method for hearing aid audio processing models to account for these unknown sounds without explicit definition. The team developed an open source audio dataset containing unfiltered-environmental audio that is categorized by a set of 4 binary features (indoors, crowded, walking, speaking) which are applicable to any audio environment. The team collected audio data using over-the-ear microphones at different locations centered around the UMD College Park campus and in Washington DC. The team also constructed a benchmarking platform for researchers to compare the performance and efficiency of different models in different environments based around this dataset. The platform evaluates models on 6 different metrics (e.g., Noise Floor, Signal-to-Noise Ratio, Dynamic Range, and Crest Factor). Results from testing various audio processing models demonstrated significant differences in performance metrics across unique noise environments. The developed and tested benchmark serves as a groundwork for future audio processing research tailored specifically for hearing aid comfort and realism.

2. Introduction

2.1. Understanding hearing loss and the hearing aid experience

Hearing loss (HL) affects more than 1.3 billion people globally, and around 15% of adults have experienced some form of hearing loss (Tsimpida et al, 2019). Although HL is one of the leading causes of disability, the efforts of many countries, especially less-developed ones, to find HL treatments regressed in order to tackle other health issues (Wilson et al., 2017). Consequently, HL has remained a significant problem globally. Beyond its physical effects, HL had negative emotional, cognitive, and social impacts (Tsimpida et al., 2019), which affected different populations in a variety of ways. For instance, children who lost hearing before the age of three often found it difficult to learn to speak, resulting in an educational barrier between them and their classmates with hearing (Wilson et al., 2017). HL further impacted adults, as they often felt extremely isolated and withdrew from social situations, including interactions with their families (Brody, 2015). It is common for those with HL to attempt to conceal their condition due to the stigma often associated with it. There was a high probability that those with HL risked developing lower self-esteem and other forms of mental illnesses due to the social barrier those with HL experienced with others. Due to barriers in communication, those with HL also had fewer opportunities, resulting in receiving half the pay and having double the unemployment rate compared to those with normal hearing in some countries (Wilson et al., 2017).

Hearing aids help those with hearing loss by amplifying the audio of speech in comparison to other types of noise. People who use hearing aids have a higher quality of life (QoL), and

more specifically, a higher health-related quality of life (HRQoL). Hearing aids lessen the negative emotional impact that those with HL might feel by helping to comprehend speech and identify where sounds are coming from (Chisolm et al., 2007). However, despite the significant impacts of HL, the processes involved in acquiring hearing aids and maintaining them could be very tedious for many users. Acquisition tended to be noticeably expensive, and more importantly, modern hearing aids were largely ineffective in noisy environments and lacked adaptability. The average cost in 2020 ranged between one thousand and four thousand dollars in the United States (Gillard & Shannon, 2022). Beyond the cost of the devices themselves, obtaining a hearing aid involved multiple audiology appointments, which made the process of acquiring a hearing aid even more expensive and less accessible. Appointments to maintain and calibrate the devices also exacerbated the cost, so long-term users of hearing aids could end up paying significantly more beyond the initial cost of the devices. Recent developments, such as the FDA's approval of over-the-counter (OTC) hearing aids, offer a promising avenue to address the accessibility issue (NIDCD, 2024). By leveraging open-source solutions, it is possible to develop adaptable and cost-effective hearing aids.

Given the challenges in acquiring, maintaining, and using hearing aids, it is no surprise that over two-thirds of Americans with HL do not use them (Chisolm et al., 2007). Stigma also seemed to play a large role in why people with HL refuse to use hearing aids. Having to wear a hearing aid could come with the challenges of being treated differently because of it. Older adults with HL might have been afraid that the appearance of a hearing aid would increase their chances of being discriminated against. People in positions of power may

have felt like they had to overcompensate because wearing a hearing aid could diminish their authority (McCormack & Fortnum, 2013). Additionally, there were individuals with hearing loss who chose not to use hearing aids because of the difficulty that came with utilizing them. They had a reputation for being too uncomfortable for all-day, everyday use, or they may have been too difficult to take care of or maintain, which made people shy away from the idea of using them (McCormack & Fortnum, 2013). On the other hand, there are members of the Deaf and hard of hearing communities who embrace this part of their lives and avoid hearing aid usage to preserve Deaf culture and sign language. Additionally, many members of this community were Deaf or hard of hearing from a very young age, influencing their neural capabilities to process audible information and rendering hearing aids to have no actual benefit (Cox et al., 2005)

Team ECHO strove to be equitable in its research by acknowledging the difference between people who choose not to use hearing aids because of their own preference, and people who choose not to use hearing aids because of the issues with acquisition, maintenance, and quality. Since the team could not solve each of these problems, it focused on improving the adaptability and effectiveness of hearing aids. By improving these aspects of hearing aids, the team hoped to reduce the need for repetitive manual adjustments while also creating a more comfortable experience for hearing aid users. Beyond pure signal processing research, accessibility was an important consideration for the team's research, given the fact that hearing aids had historically been very expensive. The average cost of prescription hearing aids in the United States is $5,000 a pair, compared to OTC hearing aids ranging from $200 to over $2,000 per pair (Knoetze et al., 2024). With the FDA's recent

approval of over-the-counter hearing aids, a new avenue was opened up for the team to create an effective hearing aid solution at lower costs (NIDCD, 2024).

In the past, various systems had been implemented and utilized in hearing aids. Originally, hearing aids focused on detecting sound in one sole environment, especially one where the environment was quieter and speech was cleaner to detect. However, due to a variety of sound environments that involved different needs and preferences (Dillon, 2012), more research had been conducted on algorithms that could distinguish and adapt between these environments. The most generic types of speech environments were quiet and noisy. Hearing algorithms that focused on detecting speech in quiet environments were much simpler since these environments had larger Signal-to-Noise ratios. However, systems for noisy environments were more complex because they needed to include additional features, such as noise suppression (Nordqvist & Leijon, 2004). Although past hearing aid research made significant progress in amplifying sound, specifically speech, this was only one of the audio elements that users listened for in a complex sound environment. Being able to successfully filter desired sound from other noise in the environment was a gap the team sought to fill in past research. Team ECHO aimed to determine how to evaluate platforms which successfully filter noise in a variety of busy environments.

Threshold-based sound classification systems had been used in the past to implement particular features in noisy environments (Xiang et al., 2010). However, to distinguish between the different types of noisy environments, utilizing classification systems based on percentiles was not enough because the modulation of different noisy environments might

still be the same, which could lead to the same results. Therefore, to ensure that the different features preferred for different types of noisy environments could be implemented, hidden Markov models (HMMs) were developed. One HMM could represent different Signal-to-Noise ratios for a single type of noisy environment, which could lessen computational complexity (Nordqvist & Leijon, 2004).

As technology continued to improve, hearing devices also became more modernized, and their potential features increased. When it came to what exactly would make the hearing aid experience better for those with HL, several factors needed to be considered in the development process. The team hoped to add to the continuous evolution of modern-day hearing aids while improving both the flexibility and accessibility of hearing aids for patients. Ultimately, the team strived to increase accessibility for the HL community and hopefully serve as a model for other audio-processing systems. To achieve their broader goals of improving hearing aid usability and accessibility, Team ECHO recognized the need for a more comprehensive and adaptable solution through the development of a benchmarking system.

2.2. Benchmarking in audio processing models

Benchmarking in audio processing models is crucial because it provides a standardized way to evaluate and compare the performance of different models across various conditions and metrics (Zöller & Huber, 2021). In the context of hearing aids, benchmarking enables researchers to assess how well different audio processing algorithms can detect and enhance desired sounds in diverse environmental settings (Park et al., 2020). Without

benchmarking, it would be difficult to determine which models perform best in challenging real-world conditions, such as noisy or dynamic environments. It ensures that the models are not only optimized for ideal conditions but can also handle complex, real-world scenarios that users experience. By using a consistent dataset for comparison, benchmarking also allows researchers to track improvements over time, identify gaps in current technologies, and guide future advancements in audio processing techniques (Stapelberg & Malan, 2020).

2.3. Gaps in existing datasets and evaluation metrics

Existing datasets for hearing aid research have notable gaps, primarily in their lack of comprehensive coverage of real-world sound environments. For instance, one study highlighted that the temporal resolution of certain datasets is limited, making it challenging to link hearing aid usage patterns to specific sound environments and listening conditions (Christensen et al., 2021). Many of these datasets rely on controlled conditions with minimal noise variability, which limits their applicability to actual use cases. Additionally, the evaluation metrics used in these datasets tend to focus more on assessing speech intelligibility rather than evaluating overall sound quality or the user experience. Research has shown that sound-quality judgments can differ between listener groups, indicating that speech intelligibility and sound quality are distinct yet interconnected factors (Van Buuren et al., 1996). Furthermore, a study noted that the relationship between sound quality and speech intelligibility is complex and warrants systematic research (Sockalingam et al., 2009). This narrow focus has led to a clear need for datasets that capture a wider range of environmental conditions, sound interactions, and user preferences, factors essential for

training and evaluating adaptable hearing aid models.

Past research in hearing aids has also often concentrated on directional noise, which required users to adjust their hearing aids to focus on sounds from specific directions. However, this approach has limitations, as it relies on user control and does not effectively account for surrounding noise. By narrowing the team's focus on information to help distinguish between surrounding and desired noise, ECHO's contributions could improve the accuracy of hearing aids, allowing them to automatically calibrate to various environments. Combining passive and interactive techniques in tracking auditory ecology can enhance hearing aid performance by adapting to the wearer's environment (Glista et al., 2020). This would reduce the need for users to continuously readjust their hearing aids, providing a more seamless and intuitive experience. The team's data could aim to identify quantifiable factors that determine which sounds are most "desired" in certain contexts, ensuring users are more aware of their surroundings and can be alerted to critical situations, such as emergencies. Ultimately, team ECHO's goal is to bring the experience of hearing aid users closer to natural hearing perception, reducing the burden of manual adjustments while enhancing overall auditory awareness.

2.4. The goal of Team ECHO's work and the contributions of the dataset

Team ECHO aimed to improve the usability of hearing aids by developing a unique environmental audio dataset with binary classifications that can be used as a benchmarking tool for research regarding hearing aid satisfaction. The purpose of the dataset is to enable benchmarking of audio processing models along environmental variables. To begin, the

team identified specific environmental factors and began recording audio samples for data collection. Environmental factors were divided into binary classifications, and audio data was collected through specialized equipment and organized by environment-based classifiers, including location, movement, speech, and crowdedness. Once collected, the data was organized and stored for future research accessibility. Unlike existing datasets, which distinguish between clean audio clips and added noise components, this dataset includes full environmental audio clips without such distinctions. It contains a broader set of binary features and classifications to enable audio processing models to optimize for more realistic and varied noise environments. The dataset was run through a simple audio-processing model and turned into an open-source dataset that other researchers can contribute to and run their own models on. This provides a way for future researchers to compare different audio processing models across different metrics and types of environments, thus enhancing research in the field of hearing aid technology.

3. Related Work

3.1 Audio Processing Models

Several audio processing techniques exist and can be applied to improve audio quality. Some of these are tailored to optimize particular sounds or signals, while others are designed for environments as a whole.

3.1.1 Filters

Filters are at the base level of audio processing. Audio filtering techniques take in a particular audio and modify it by selectively muting and amplifying particular frequencies. Audio filters range from simpler static filters to more complex adaptive filters and are tailored for a particular purpose.

Audio filters can be static or adaptive. Static filters are much more basic and have a consistent, predictable performance. They are not affected by the input audio file and instead apply the same filtering mechanism regardless of the input. One common example of a static filter is a low-pass filter, which allows signals up to a cut-off frequency to “pass” through while attenuating higher frequencies (Rezmeriță et al., 2023).

Adaptive filters, by contrast, are affected by the input audio and can remove different frequencies based on the signal and environment it is given. The team examined the performance of several filters on a variety of input audios to determine optimal use cases.

One adaptive filter the team examined is the Least Mean Squared (LMS) Filter. This model receives a reference audio as well as an audio that will be filtered (Mendiratta & Jha, 2014).

The LMS filter tries to minimize the disparity between the reference audio and the desired audio to reduce background noise. This is particularly helpful in environments where isolating a specific signal, such as a conversation, is crucial.

3.1.2 Algorithms

A step up from basic filtering, which typically applies uniform changes across frequencies regardless of static or adaptive, are noise reduction and audio enhancement algorithms that respond more intelligently to the characteristics of the input signal. One of the earliest of such techniques is Spectral Subtraction. First introduced in 1979, it operates by estimating the noise profile from silent segments of a signal and then subtracting them from a signal's frequency spectrum (Boll, 1979). Qualitatively, this reduces low stationary noises, but was observed to introduce other distortions and artifacts to audio files when noise shifts and changes.

Another such filter is Dynamic Range Compression (DRC). Developed for analog systems in radio broadcasting in the 1930s, DRC attempts to reduce the gap between the loudest and quietest parts of an audio signal and is used commonly in hearing aids today (May et al., 2020). This was utilized to improve audio intelligibility in noisy and constrained playback environments, but could sometimes suppress the diverse nature of noise environments too aggressively.

Although both of these algorithms demonstrate clear advances and advantages over filters, they are not without limitations. Many algorithms are extremely environment and signal-characteristic dependent. Investigating how these methods complement one another

in specific contexts and spaces can guide more robust and adaptive audio processing strategies.

3.1.3 Trained Models

Building on these foundational techniques, recent advancements in machine learning have introduced models trained to identify and enhance specific audio features. Many such models offer a more adaptive and context-aware approach to audio processing.

One notable approach to this, specifically tailored to speech enhancement, is SEGAN (Speech Enhancement Generative Adversarial Network) which pioneered the use of adversarial networks for denoising speech (Pascual et al,, 2017). An extension of this is publicly accessible as the MetricGAN+, implemented within the SpeechBrain toolkit. This model uses adversarial training to optimize perceptual speech clarity metrics and is trained on large speech datasets such as VoiceBank-DEMAND.

An alternative approach, focused on noise reduction, is the Noise2Noise method, which leverages deep learning techniques to perform noise reduction in scenarios where clean data is not available. This technique is beneficial in extremely noisy environments with low signal-to-noise ratios and demonstrates how denoising models can be trained without clean reference data (Kashyap et al., 2021).

While these models may outperform traditional algorithms in their specific tailored tasks for diverse noise environments, they also heavily depend on the training data they are provided. In many of these cases, it requires specific environments or factors built into an audio signal. Investigating their generalization across different defined acoustic

environments is an area of research that still requires more data and development, which team ECHO aimed to fill.

3.2 Benchmarking

3.2.1 Audio Datasets

There are many methods currently used to evaluate the efficacy of audio processing models. For audio classification, the authors of (Abbasi et al., 2022) preprocess audio via data framing before extracting features through methods such as Principal Component Analysis (PCA) and calculating Mel-frequency Cepstral Coefficients (MFCC) among other features and then comparing the accuracy, precision, etc. of different classification models.

Although recent years have seen an increase in the generation of audio processing benchmarks, there is a significant gap in the accessibility and quantity of audio classification benchmarks compared to visual or image classification in machine learning. MetaAudio sought to remedy this by utilizing seven datasets (five for training and evaluation, two for testing) to determine the classification accuracy of metric, baseline, and joint-training meta-learning algorithms (Heggan et al., 2022). MetaAudio's datasets focused on various environmental sounds, such as urban noises, as well as diverse human speech.

Similarly, AudioLLM is also a benchmarking tool that evaluates current audio LLMs based on their ability to understand speech, non-speech audio, and other human-related data (emotion, gender, and accent). It was designed to find a comprehensive way to evaluate LLMs based on their instruction-following abilities, ie, understanding user speech inputs, transcribing them into text, and following the requested queries appropriately, regardless

of environmental or other factors (Wang et al., 2024). Through AudioBench, a comprehensive benchmark model consisting of twenty-six datasets, each specializing in one of eight audio-related tasks, the tool evaluated five audio LLMs and failed to find a singular LLM that excelled in all categories: automatic speech recognition, speech question answering, speech instruction, audio captioning, audio-scene question answering, accent recognition, gender recognition, and emotion recognition. AudioBench's datasets focus primarily on speech comprehension with captioning, answering spoken queries, and recognizing certain sentiments from speech. It uses metrics such as word error rate and other NLP metrics (ie METEOR) to evaluate accuracy.

The evolution of audio benchmarking has provided future opportunities for further expansion of datasets and metrics. Many large datasets in currently existing benchmarks have been processed (Chen et al., 2021), meaning there is a need for more datasets that provide real-world inaccuracies that occur in speech, such as stumbles over words, fillers, and pauses, which Team ECHO aims to rectify. Additionally, current audio benchmarking metrics often utilize metrics that focus more on classification or speech recognition on a higher level. On the other hand, more traditional metrics that are utilized to evaluate audio are often overlooked, such as SNR, Crest Factor, etc., which are used as metrics in Team ECHO's benchmarking system. However, high-level metrics may not provide scores that necessarily correspond well to the human perception of audio, leading to a need for no-reference metrics as a replacement (Manocha et al., 2022). The relationship between higher-level metrics and lower-level metrics can also be further studied.

3.2.2 Machine Learning and Other Benchmarks

Benchmarking is an integral part of the iterative machine learning development process. More specifically, they are key in evaluating machine learning model performance across domains by providing a standardized framework for comparison. When models have similar objectives, they can reveal insights that discern strengths, weaknesses, and opportunities for improvement for different implementations. In fields like computer vision and natural language processing (NLP), widely adopted benchmarks such as ImageNet and GLUE (Wang et al., 2019) have driven sustained progress through clear evaluation protocols, representative datasets, and leaderboard systems. These benchmarks emphasize not only performance metrics but also reproducibility and fairness. Moreover, leaderboards drive the spirit of competition that sparks further improvements. Although benchmarking itself does not involve training, the insights often guide refinements for the evaluated model through architectural changes, hyperparameter tuning, or domain adaptation.

Audio benchmarks face unique challenges in representing the diversity of real-world environments and capturing spontaneous speech patterns or background noise. Datasets that include a variety of speaker identities, accents, and settings can support these goals while also enabling more nuanced evaluations of model robustness (Babel, 2022). Given these challenges, the goal of the team's research was to bridge the gap in audio benchmarking by creating a diverse catalog of data collected from a wide range of environments. The team sought to design a dataset that reflected the variability of real-world environments. This approach ensures that models are tested for robustness and adaptability, establishing a more comprehensive and holistic audio benchmark.

4. Dataset Design

In order to fill the gaps between existing research, datasets, and benchmarking, the team decided to construct a new environmental audio dataset for evaluating models. The dataset was structured with the goal of representing most noise environments without relying on features that only apply to certain locations.

4.1 Dataset Structure

There were 4 primary binary features the team uses to classify audio, which resulted in 16 different audio environment classifications *(see Table 1)*. These features were: Indoors vs. Outdoors, Crowded vs. Not Crowded, Walking vs. Not Walking, and Speaking vs. Not Speaking. The Crowded vs Not Crowded factor was defined by the presence of 2 or more competing foreground audio sources (e.g., music, speech, traffic). We defined foreground as an audio source which in recording yields an audio level equal or above 75% of the intensity of a recording speaker. The Speaking vs. Not Speaking feature indicated the presence of another individual aside from the user that roughly read from a set script to the user (*see Figure 3)*; the recorder never spoke in any instances of the data. The dataset's audio clips consist of various lengths, all greater than 2 minutes. Voice type (male/female), individual speaker IDs, and location type (hallway, restaurant, garden, etc.) were also included as tags in the dataset alongside audio clips.

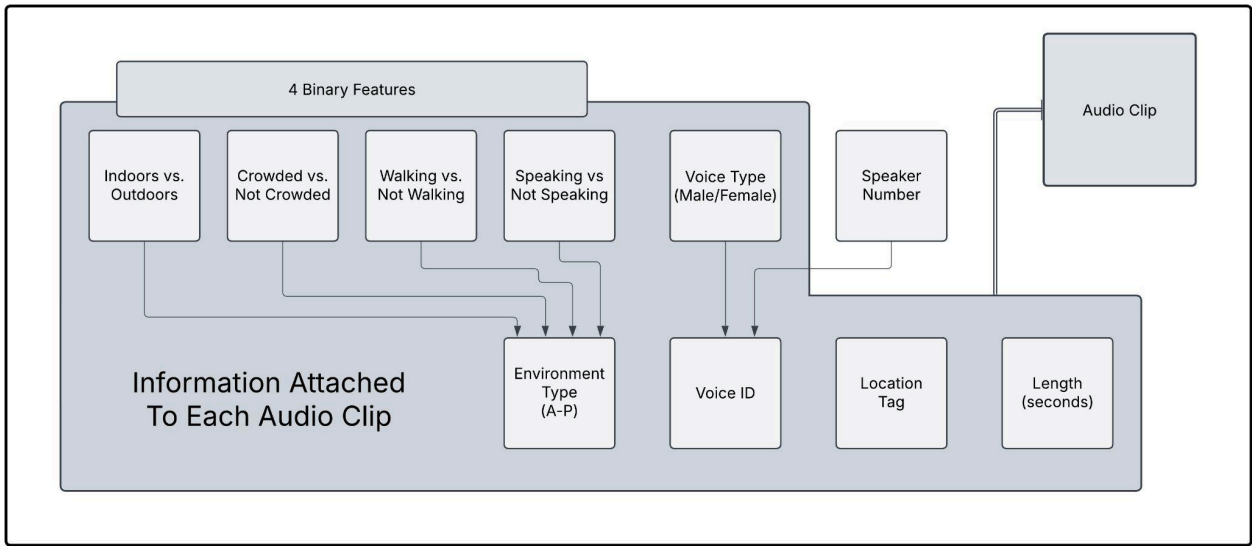

Figure 1: Dataset Information Structure

Environment Type	Indoor	Crowded	Speaking	Walking
A	Yes	Yes	Yes	Yes
B	Yes	Yes	Yes	No
C	Yes	Yes	No	Yes
D	Yes	Yes	No	No
E	Yes	No	Yes	Yes
F	Yes	No	Yes	No
G	Yes	No	No	Yes
H	Yes	No	No	No
I	No	Yes	Yes	Yes
J	No	Yes	Yes	No
K	No	Yes	No	Yes
L	No	Yes	No	No
M	No	No	Yes	Yes
N	No	No	Yes	No
O	No	No	No	Yes
P	No	No	No	No

Table 1: 16 Environmental Classifications based on selected factors

The audio was recorded using an over the ear microphone and flash based recorder at 20 kHz (*see Figure 2*). The team recorded audio on the University of Maryland College Park campus, with select recordings taken in Washington DC. In assigning this dataset structure and general recording pattern, the dataset was able to be transitioned to open source, enabling a broader range of geographical locations to be represented alongside more voices with further expansions.

Figure 2: Recording materials used for initial dataset construction. Handheld flash-based recorder (Left) and SonicPresence SP15 Binaural Over-Ear Microphone

4.2 Design Considerations

4.2.1 Audio Classification and Uniqueness

The four binary classification features the team used were selected in order to represent a total of 16 major noise environments in which an audio processing model might be used without relying on specific location descriptions. Many existing publicly available audio datasets, such as (Piczak, 2015) and (Huwel et al., 2020), only label their environmental audio with descriptions similar to the 'location type' label. This method of classification would require the team to account for every conceivable type of location for the use of a dataset to be equally useful in every variety of different audio scenarios. The team considered that it may be more practical to classify audio environments using a set of features that can be applied to every type of noise environment. In order to keep the data collection process simple, the team decided to include a total of 4 different binary features rather than exponentially increasing the number of different environments that needed audio to be recorded in by adding more features.

4.2.2 Speech Script

The script used for the "Speaking" classification was designed to include as many different phonetic sounds that appear in the English language as possible. Since real-world use of language includes speech errors and different variations of the same sentences, the audio recorded only generally followed the script. Every word was spoken, but some samples will include speech errors or additional conjunction words such as 'and', 'but' that appear in between individual sentences. The script is intended to be read in both forward and reverse

orders between lines and/or sections. This was done to create more variety in how lines were articulated, especially when joining multiple lines with a conjunction. However, the elicitation paragraph taken from the Speech Accent Archive (Weinberger, 2015) never had its sentences reordered since the extensive phonetic transcription data from the archive could provide some use to others who might use the data. The full script with the Speech Accent Archive paragraph at the bottom of the page is shown in Figure 3.

That quick beige fox jumped in the air over each thin dog. Look out, I shout, for he's foiled you again, creating chaos.

Are those shy Eurasian footwear, cowboy chaps, or jolly earthmoving headgear?

The hungry purple dinosaur ate the kind, zingy fox, the jabbering crab, and the mad whale and started vending and quacking.

With tenure, Suzie'd have all the more leisure for yachting, but her publications are no good."

Shaw, those twelve beige hooks are joined if I patch a young, gooey mouth.

The beige hue on the waters of the loch impressed all, including the French queen, before she heard that symphony again, just as young Arthur wanted.

The birch canoe slid on the smooth planks.
Glue the sheet to the dark blue background.
It's easy to tell the depth of a well.
These days a chicken leg is a rare dish.
Rice is often served in round bowls.
The juice of lemons makes fine punch.
The box was thrown beside the parked truck.
The hogs were fed chopped corn and garbage.
Four hours of steady work faced us.
A large size in stockings is hard to sell.

The boy was there when the sun rose.
A rod is used to catch pink salmon.
The source of the huge river is the clear spring.
Kick the ball straight and follow through.
Help the woman get back to her feet.
A pot of tea helps to pass the evening.
Smoky fires lack flame and heat.
The soft cushion broke the man's fall.
The salt breeze came across from the sea.
The girl at the booth sold fifty bonds.

The small pup gnawed a hole in the sock.
The fish twisted and turned on the bent hook.
Press the pants and sew a button on the vest.
The swan dive was far short of perfect.
The beauty of the view stunned the young boy.
Two blue fish swam in the tank.
Her purse was full of useless trash.
The colt reared and threw the tall rider.
It snowed, rained, and hailed the same morning.

Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

Figure 3: Audio-Recording Speech Script

4.2.3 Benchmarking and Evaluation

The structure of the team's dataset supports benchmarking across a comprehensive range of environments. By using four binary features, the team enabled representation of 16 environment types (*see Table 1)*. This allows for the evaluation of models in controlled performance in scenarios that differ by one or more factors, such as Crowded or Walking. The benchmarks are expected to reveal whether a model struggles with speech in the presence of background noise or if the speaker is walking. This structure can uncover how individual features impact model performance in a way that datasets with location-based labels cannot. Although location is not used as a primary classification feature, it is still included as a tagged attribute. This can be useful to researchers looking to perform secondary analysis but doesn't pigeonhole them into overfitting models on specific locations. Rather, classification is generalized to the acoustic scene of the present environment, which is more generalizable.

Furthermore, the dataset is suitable for benchmarking since the team accounted for robustness with annotations for variations in speakers and gender. With a script that maximized phonetic diversity, recordings were prone to natural variations that the team captures with a wide range of voices from various backgrounds. This emphasis on phonetic diversity is key to possible improvements to the models run against the team's benchmarking infrastructure.

4.3 Data Collection Process

The system the team used for data collection was as follows: data was collected for 4 out of the 16 defined noise environments throughout each recording session based on time constraints. For all recording sessions, the same script was used. The speakers were instructed to read the script line-by-line, and to then read the same lines in reverse order (*see Figure 3*).

Team members grouped into different pairs and recorded in a location based on whether or not it was Indoors/Outdoors and Crowded/Not Crowded. At that location, the pairs recorded for each combination of the Walking/Not Walking and Speaking/Not Speaking features. For the Not Speaking recordings, pair members would take turns recording the environment in silence for roughly 5 minutes each. When recording the Speaking environments, one member would record in silence while the other member would read from the speech script at a volume appropriate for the location. In the case of environments involving both Speaking and Walking, both members would walk side by side, where the member recording with the device would guide the member who was speaking around the area to allow them to more safely read the script while walking.

Either after or during the recording session, pairs would upload their recorded audio to a shared cloud storage folder and create entries in the spreadsheet for the newly recorded audio. Each clip was tagged with a location taken from a list of previously visited location types, with any new types that were significantly different from the others being added to the list as needed. This was done to avoid any confusion among the team with using different names for very similar location types (e.g., a plaza and a square). Each member tagged audio which contained them speaking with their assigned voice ID.

5. Metrics

With this dataset constructed, the next step would be to determine set metrics for evaluating audio across the defined environments. The team decided to utilize six different metrics to evaluate audio processing performance within a given benchmark: Total Harmonic Distortion (THD) , Noise Floor, Signal-to-Noise Ratio (SNR), Dynamic Range, Crest Factor, and Waveform Complexity Index (WCI).

5.1 Total Harmonic Distortion

Total Harmonic Distortion (THD) in the context of hearing aid audio refers to the distortion introduced by non-linearities in the system. These cause additional harmonic multiples of the primary input (fundamental) frequency to appear in the output. These harmonics are not present in a clean signal and can degrade audio fidelity, especially in speech-critical applications like hearing aids.

To approximate this for audio benchmark, the team used the following equation:

$$Total\ Harmonic\ Distortion\ = \frac{\sqrt{\sum_{k=2}^{K} A_k^2}}{A_1}$$

A_1 = The RMS Amplitude of the fundamental frequency,

A_k = The RMS Amplitude of the k-th harmonic frequency (k = 2 to K)

K = the total number of significant harmonics considered

Higher THD in Audio indicates that the audio contains more noise and distortions. This could happen due to input hardware limitations, background noise, or other sources of distortion within the hearing aid.

Lower THD suggests that the signal has a cleaner, more accurate representation of an original sound. When processed audio has a lower THD compared to the unfiltered version, it means that the filtering or processing has successfully reduced distortion, providing a clearer and more natural sound. This is crucial for hearing aid users, as it means better speech understanding and a more authentic listening experience.

Lowering THD through processing can demonstrate a model's capability to refine the original signal, improving both the user's comfort and their ability to comprehend speech in noisy environments.

5.2 Noise Floor

The Noise Floor in the context of hearing aid audio refers to the background noise level present in a recording, even when no significant sound input occurs. It represents the lowest level of noise that the hearing aid captures, which can be due to internal electronic noise from the hearing aid's circuitry or ambient noise.

To approximate this for audio, the team used the following equation:

$$Noise\ Floor_{dB} = 20 * log_{10}\left(\frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{M}\sum_{j=1}^{M}x_{i,j}^{2}}\right)$$

N = number of audio segments,
M = number of samples per segment,
$x_{i,j}$ = the j-th sample of the i-th segment of

This can also be described as the average RMS signal power across small segments of audio, estimating the overall signal floor.

If the unfiltered audio has a high Noise Floor, it means there's significant background noise throughout the recording, affecting overall sound clarity. If the processed audio has a lower Noise Floor, the processing effectively reduces background noise, making softer speech parts or desired signals clear.

Lowering the Noise Floor is typically desirable for clarity in more complex or noisy audio environments. Intentionally raising the Noise Floor can improve comfort, naturalness, and sound awareness, especially in hearing aids, telephony, and broadcast applications.

5.3 Signal-to-Noise Ratio

Signal-to-Noise Ratio (SNR) is a measure of the level of a desired signal to the level of the background noise. It is typically expressed in decibels (dB) and calculated as:

$$SNR_{dB} = 20 * log_{10}(\frac{RMS\ Audio\ Signal}{Noise\ Floor})$$

SNR is a key metric in audio and signal processing because it quantifies how much of the useful signal is distinguishable from the unwanted noise.

Quantitatively, a high SNR is the RMS of the signal is much higher than the RMS of the noise. A low SNR is the opposite where the RMS of the noise is higher than the RMS of the signal, resulting in a lower ratio within the log function.

Qualitatively, a higher SNR in audio leads to clear, crisp sound with minimal background interference. A lower SNR can make speech unintelligible, introduce static in radio transmissions, or degrade the quality of music playback.

Scenarios where an increase in SNR is beneficial is high quality music recording, voice communication in noisy environments, and even precision measurement tools. Any scenario where having a high output signal within any environment is important needs high SNR.

5.4 Crest Factor

Crest Factor is a metric that is used to determine the amount and concentration of "peaks" within a waveform. This is determined by calculating the ratio of peak amplitude and root mean squared values. Root mean squared is the square root of the mean signal value and is used as an indicator of the average signal present within the waveform.

$$Crest\ Factor\ = \frac{Peak\ Amplitude}{RMS\ Audio\ Signal}$$

Within signal processing, a higher concentration of peaks can cause a higher degree of stress on output devices. It is a useful benchmark to consider as it can be used to determine the power demands necessary to adjust for peak signals.

Quantitatively, a higher Crest Factor represents a higher disparity between the peak values and the overall average signal within a waveform. Having a higher Crest Factor is indicative of having sharper peaks. Qualitatively, this means that most of the signal contains heavily concentrated peaks which could indicate the presence of speech. A lower value indicates a more consistent sound that may not be picking up extensively on one particular signal.

5.5 Dynamic Range

Dynamic Range is used to determine the variability between the largest and smallest signals within a waveform. This can be calculated by converting the ratio between the maximum and minimum signals to decibels.

$$Dynamic\ Range_{dB} = 20 * log(\frac{Peak\ Signal}{Minimum\ Signal})$$

A higher Dynamic Range can support a more precise sound and indicates the ability of the system to represent a wide range of signals. It is a good benchmark to consider as it can show the variability of noise present within a waveform and thus indicate the range in conditions within the encompassing environment.

A higher Dynamic Range represents a higher range of signals present within a waveform. A smaller range means that the signals are closer together, although not necessarily uniform. Qualitatively, a smaller Dynamic Range could indicate an environment where noises are similar in value, such as in a crowded area, or a limitation in recording, such as a compressed audio file or a lower quality microphone.

5.6 Waveform Complexity Index

Waveform Complexity Index (WCI) examines shapes present in non-overlapping parts of waveform and determines the level of diversity between these shapes. WCI is calculated by segmenting a signal into these non-overlapping sections, calculating the diversity distribution based on a given correlation metric (r) and then calculating the median of this.

$$Waveform\ Complexity\ Index\ (WCI)\ =\ median\ (distribution\ of\ 1\ -\ |r|)$$

The WCI gives insight into the variability present across the entire waveform and can highlight large scale changes that show different signals being picked up that deviate from the normal patterns.

A higher WCI means there is a greater degree of change present within the waveform, which correlates with a more complex signal. A lower value, by contrast, is a more consistent signal. Qualitatively, a higher value could represent more transitions and variations, which can be divided into distinct parts. A lower value would indicate that the signal represents a more uniform concept and is consistent all the way through.

6. Benchmarking Utility Architecture

Once audio was processed, the team moved to build a benchmark that combined both the dataset's environmental features and recordings with the metrics used to evaluate signal and audio impacts of different models.

6.1 JSON Benchmark Structure

For the proposed benchmarks, the team decided to evaluate audio metrics independently for each audio file rather than aggregated across entire datasets. This per-file evaluation enables more specific comparisons across environmental conditions such as the indoor or outdoor setting, movement classification, and presence of speech alongside other informative tags. Maintaining this level of granularity for evaluation design allows for filtering during analysis enabling applications of this dataset and benchmarking tool outside the explicit scope of the 16 provided and lettered environments.

To support this structure, the team's benchmark employs the JSON (JavaScript Object Notation) format for data storage. JSON's hierarchical key-value structure is well-suited for representing both metadata and computed metrics associated with each recording. It enables encapsulation of structured yet heterogeneous information—such as categorical labels and continuous audio metrics—without imposing a rigid schema.

JSON is a widely adopted data storage format across the machine learning community for benchmark design and annotation tasks due to its legibility and compatibility with major programming languages and toolchains. Notably, the COCO dataset for object detection and segmentation (Lin et al, 2014) utilizes a JSON schema for storing object annotations, image

metadata, and evaluation results. Similarly, the natural language understandability benchmark SuperGLUE (Wang et al., 2019) stores benchmark results in JSON-compatible formats for standardization. JSON is also found in MLCommons benchmarking infrastructure (Mattson et al., 2020), which stores metadata to capture experiment configurations and performance results across a wide range of ML models and tasks.

The JSON structure selected by the team is split into model-level factors and file-level factors. Model-level factors include the model name or type, as well as runtime for processing the entire dataset. Then, at an audio-file level, it includes tags for the four environmental classification factors in addition to voice-type for if a speaker present is male or female, voice-identifier for which specific individual is speaking, and the six metrics for audio quality recorded for each audio file (see Figure 4).

Benchmark JSON	
Model	
“audio_model” - Name	String
“runtime” - Time to Process Dataset (s)	Integer

File 1	
“ID” - Full File Identifier	String
“Length” - Duration (s)	Integer - Seconds
“Location” - More Specific Location Identifier	String
“Indoors” - Indoor (T) or Outdoor (F)	Boolean (Yes or No)
“Crowded” - Loud Environment (T) Quiet (F)	Boolean (Yes or No)
“Speaking” - Script Read File (T) No Script Read (F)	Boolean (Yes or No)
“Walking” - Recorder Walking (T) Recorder Stationary (F)	Boolean (Yes or No)
“Voice_Type” - Category of speaker in record (Male/Female)	String or NA
“Voice_ID” - Specific speaker in record (MXX or FXX)	String
“Total_Harmonic_Distortion” - Calculated Metric	Float
“Signal_Noise_Ratio” - Calculated Metric	Float
“Noise_Floor” - Calculated Metric (db)	Float
“Dynamic_Range” - Calculated Metric (db)	Float
“Crest_Factor” - Calculated Metric	Float
“Waveform_Complexity_Index” - Calculated Metric	Float
...	

Figure 4: Benchmark File Architecture with Associated Values

6.2 Benchmark Generation

To generate a benchmark, a user of the benchmarking first imports the entire environmental audio dataset. From this, each contained .wav file would be processed under the audio processing model for evaluation, yielding a corresponding output .wav file with the same title and file name identifier. Utilizing one such identifier, the generation code populates a JSON structure with each environmental factor and associated voice tag. Following this, each of the six metrics is computed for the file and stored in its respective JSON structure (*see Figure 4*). This process of populating tags and metric calculations is repeated and stored for each audio file under the model JSON. From this point, the benchmark can be used for comparison and further data analysis (*see Figures X and X*).

This benchmarking system allows researchers easy access to edit and add metrics for specific projects. The JSON structure promotes reproducibility through its adaptability, allowing the new metrics to be easily inserted within the JSON generation code. The data visualization can also easily be modified to incorporate graphs for these new metrics. Additionally, if users want to add more datasets to the system, they only need to ensure that the datasets follow the correct naming convention, without needing to worry about other aspects of the pipeline. Overall, Team ECHO's benchmarking system provides a structured pipeline that allows the user to compartmentalize changes.

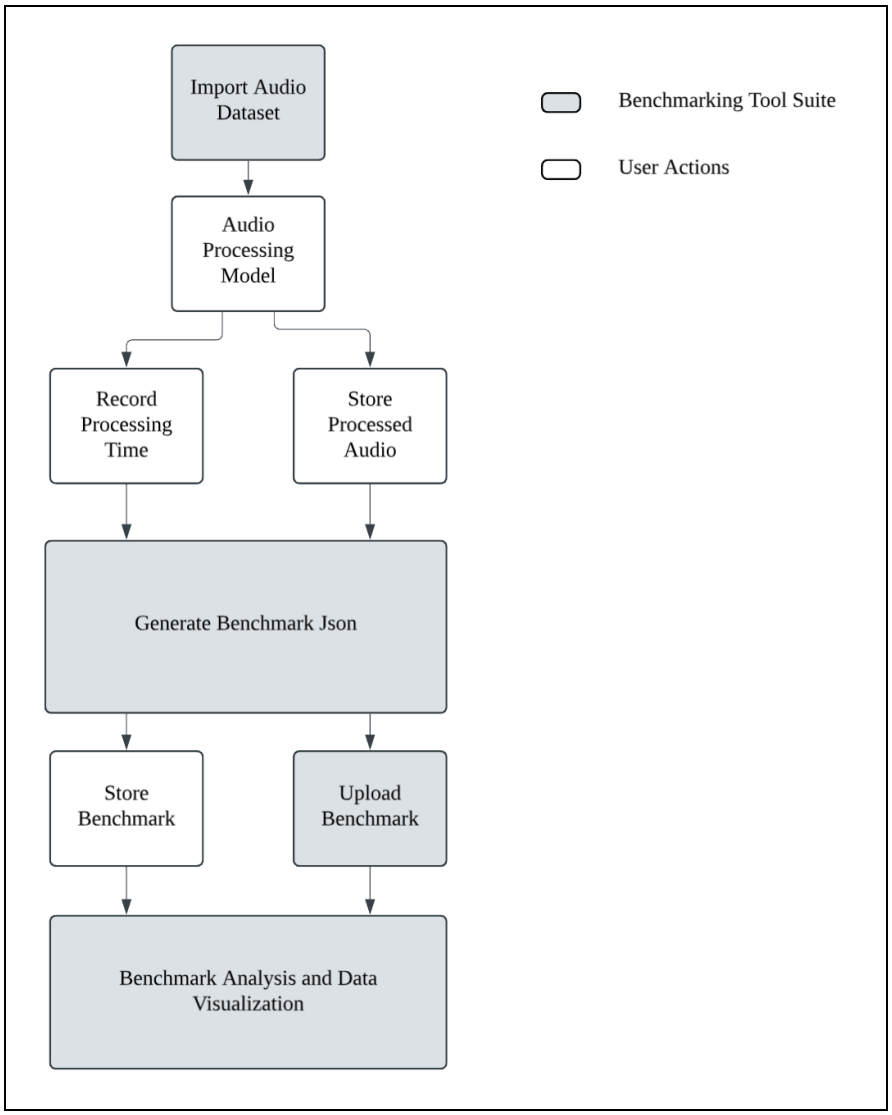

Figure 5: High-Level Architecture of the Benchmarking Tool Suite

6.3 Data Comparison and Visualization

After the JSON benchmark files are created, the next step in the benchmarking process is to analyze these files to evaluate how the different models are performing based on each metric (Total Harmonic Distortion, Noise Floor, Signal-to-Noise Ratio, Crest Factor, Dynamic Range, Waveform Complexity Index). The team designed a Python library to process multiple JSON benchmark files, analyze their contents, and produce comparison graphs.

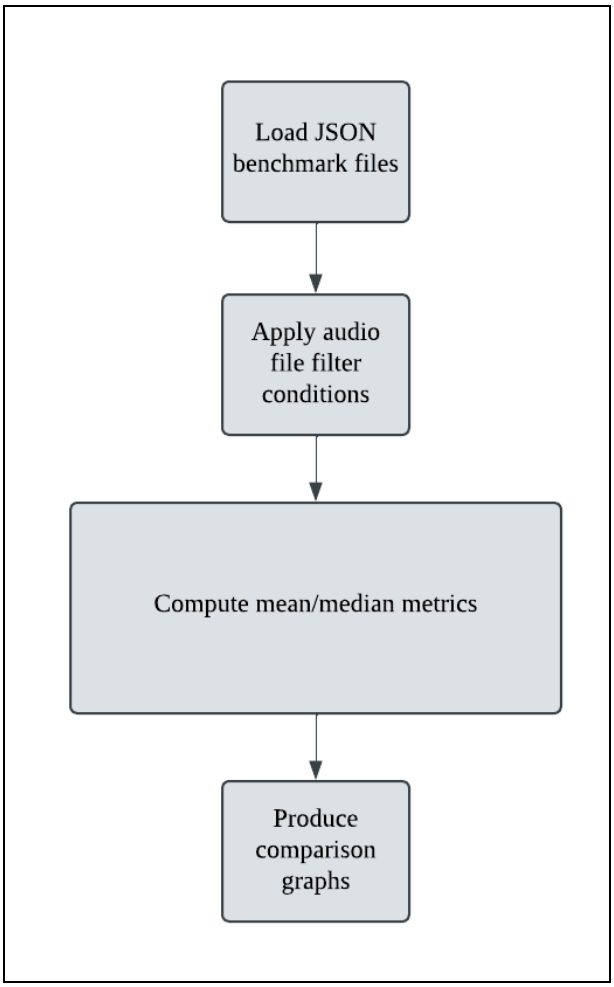

Figure 6: Data Analysis and Filtering Flow

The first step (*see Figure 6*) is to load each of the JSON benchmarking files into a list of readable files. After these files are loaded, a group of filter conditions is applied to each JSON's list of audio files. These filter conditions indicate what data will be analyzed. Two types of filters are applied. The first filter controls how many and which metrics will be analyzed. The second filter determines which specific subset of environments will be analyzed. Both of these filters are optional, and if they are not populated, all metrics and audio files are analyzed. After the filters are applied, the functions will, by default, calculate the mean of the selected metrics. If explicitly specified, the functions can also calculate the median of the selected metrics. Once all the metrics are calculated, the functions will

produce comparison graphs that display how all the models perform on the selected metrics.

Two main functions perform this full process. The main difference between the two functions is the type of comparison graphs that are produced. The first function produces a snapshot of one environment with comparison graphs for each selected metric. Here, each bar represents the mean/median value for a single model (*see Figure 7*). The second function produces comparison graphs for each selected metric across multiple specified environments *(see figure 8)*. Each subsection of bars represents a selected environment, and the set will display the mean/median value for each model within that selected environment and metric.

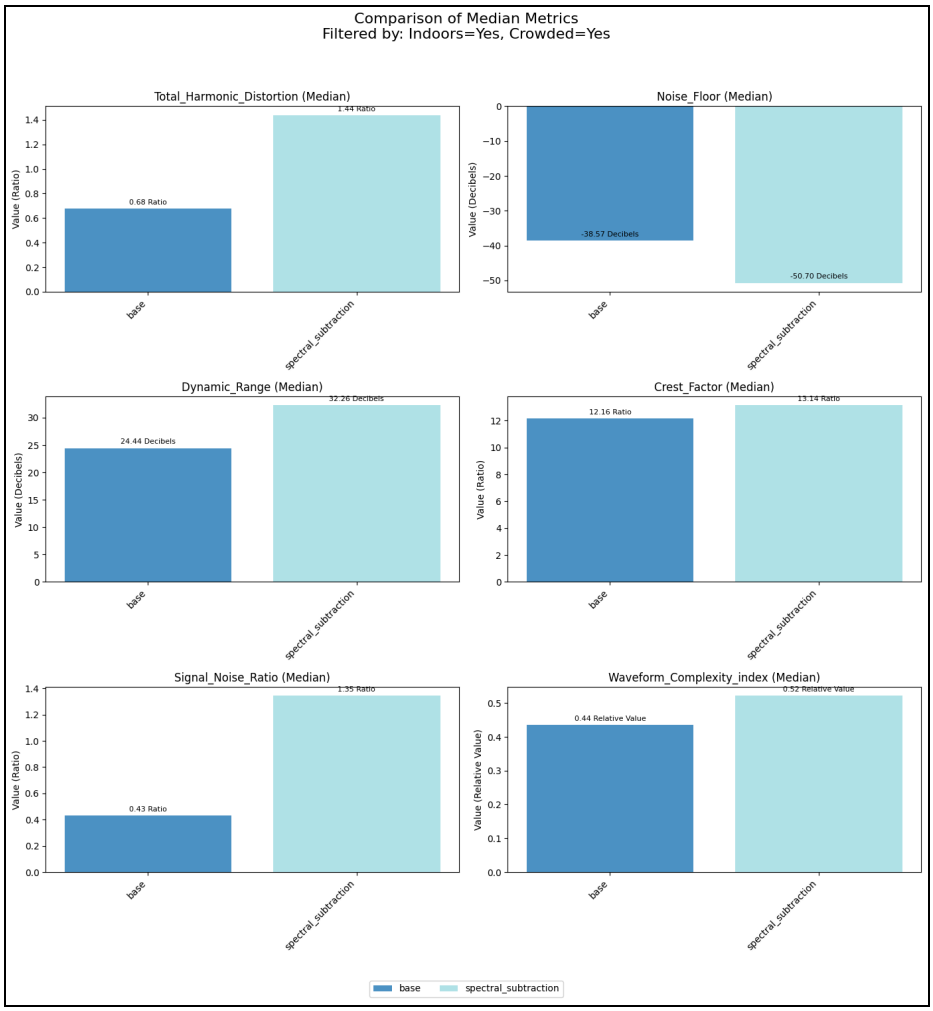

Figure 7: Example use case of the metric snapshot for a subset of factors.

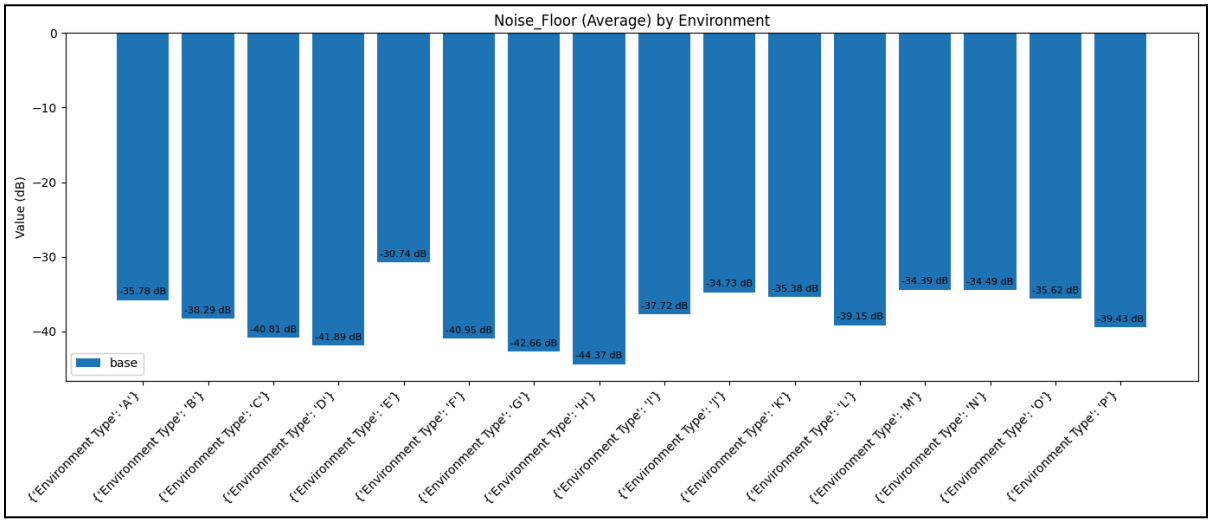

Figure 8: Environment snapshot for the baseline benchmark without any processing for Noise Floor

Both functions are designed to facilitate benchmarking in separate ways. The first function allows for a broader comparison of the models. After filtering, the data is analyzed altogether to extract a singular value for each model-metric pair. This reveals how different models generally perform in comparison to each other over subsets or all environments. The second function provides a more specific, environment-based analysis for benchmarking. Since the function produces separate values for every specified environment, the models can be compared on a more specific level to understand how they perform in specific environments.

These functions provide a straightforward method for benchmarking audio processing models. With this process, several models can be compared simultaneously in order to determine their strengths and weaknesses. This will eventually lead to model improvements by identifying areas where performance can be improved.

7. Demonstration with Existing Models

7.1 Experimental Setup

To demonstrate the utility of the dataset and benchmark analysis in action, the team decided to pass the environmental audio through three separate audio filtering models alongside a baseline: spectral subtraction, Dynamic Range Compression, and a MetricGAN+ speech bank model for speech clarity. Spectral subtraction was selected as a typical noise reduction algorithm utilizing digital signal processing and is used by default in the Python noisereduce package. Dynamic Range Compression was selected as its goal of shrinking; specifically, the Dynamic Range of clips should be visible and consistent compared to a baseline. The MetricGAN+ speech bank model was selected as it is a representation of a machine learning model trained on audio data that is not sorted by environment, and thus the team expected to see large and inconsistent variations in performance across different environment classifications.

Each of these models' outputs from the dataset was stored and used as sources for generating benchmarks titled respectively. From this, an analysis of all metrics across environments A-P was conducted and graphed alongside a single snapshot of the broad subset of environments for crowded and speaking.

7.2 Experimental Results

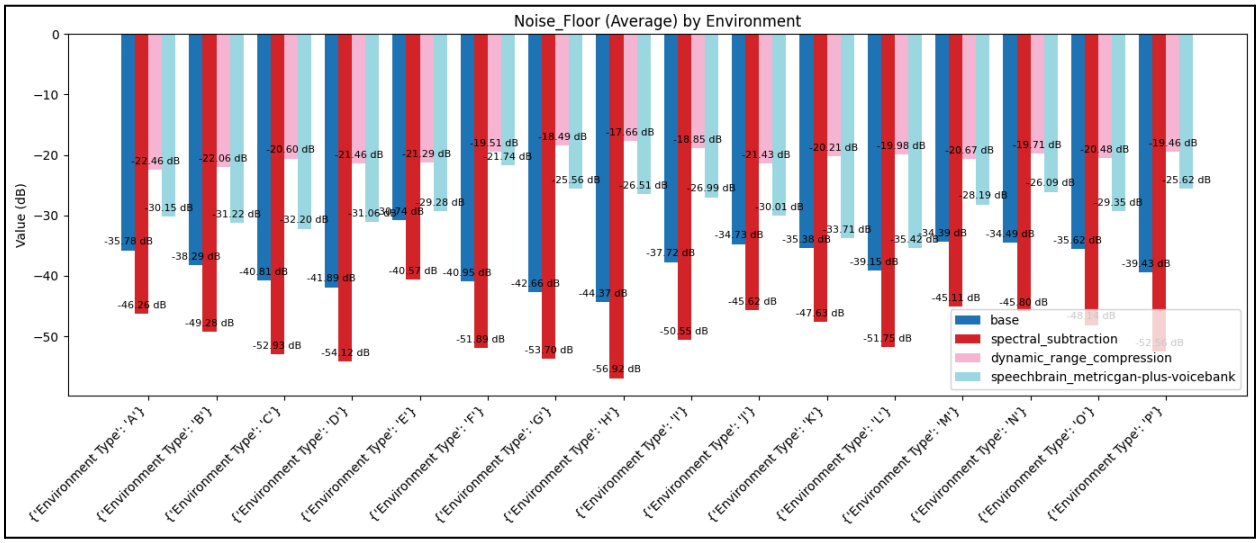

Figure 9: Noise Floor measured (dB) across all environments for all selected audio-processing models

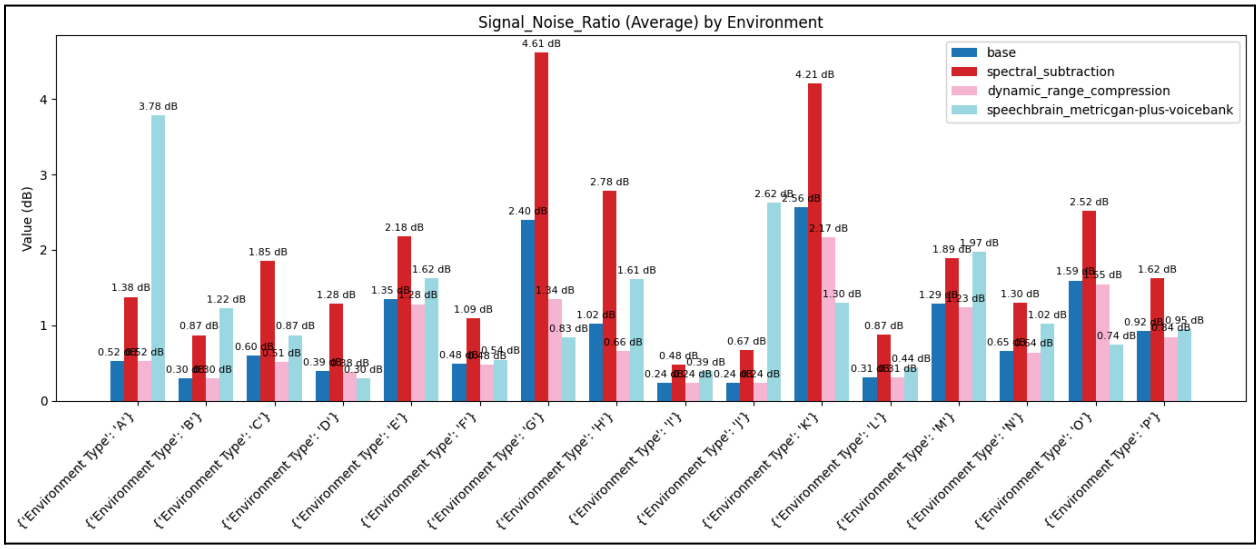

Figure 10: Signal-to-Noise Ratio measured (dB) across all environments for all selected

audio-processing models

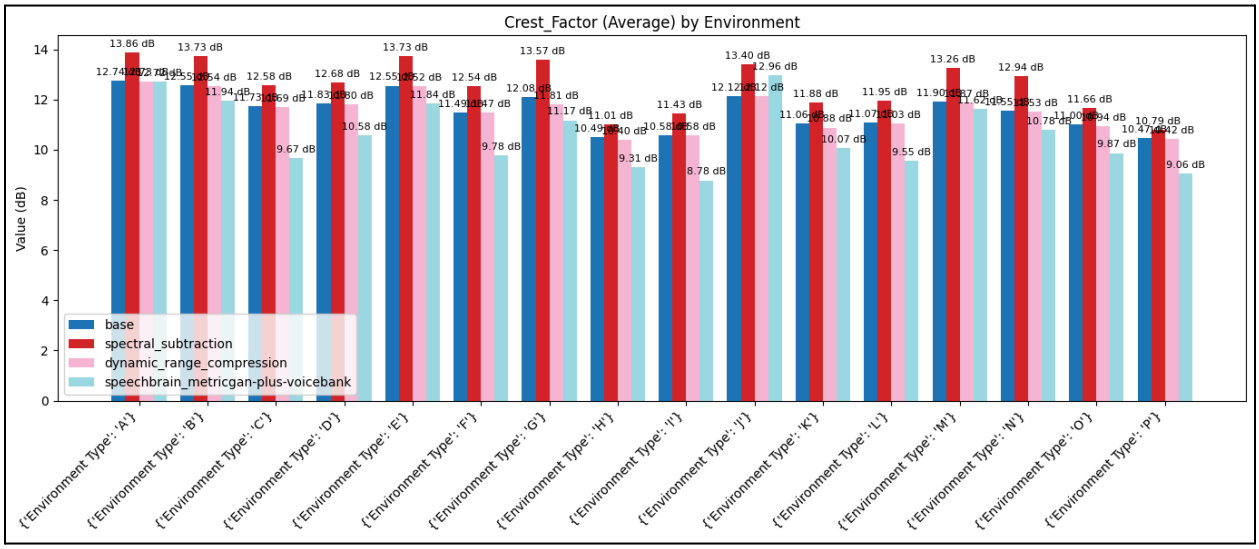

Figure 11: Crest Factor measured (dB) across all environments for all selected audio-processing models

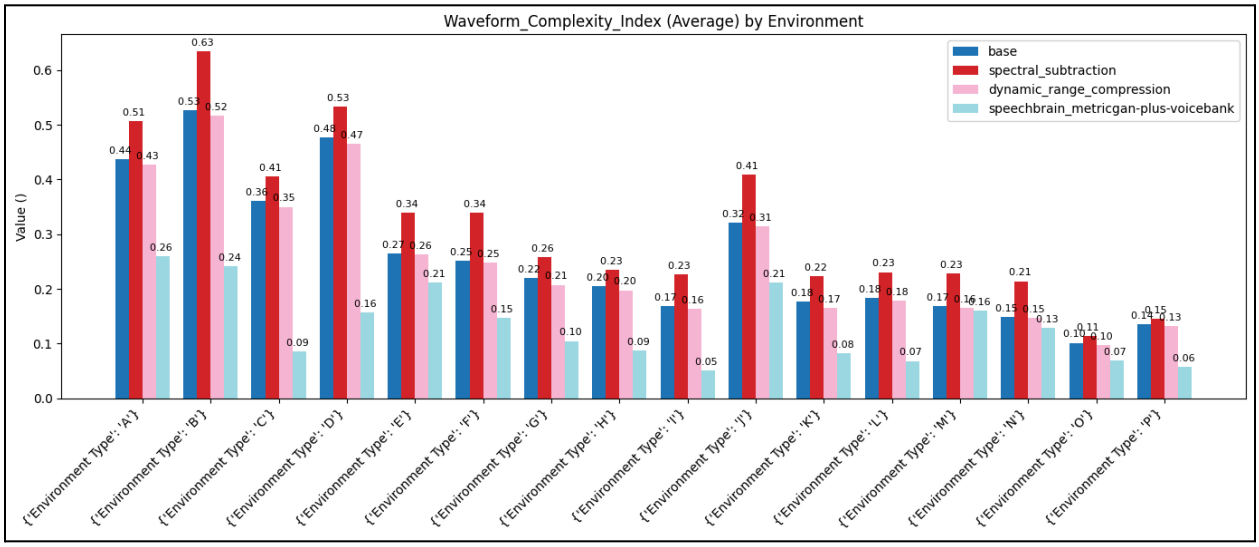

Figure 12: Waveform Complexity Index measured across all environments for all selected audio-processing mode

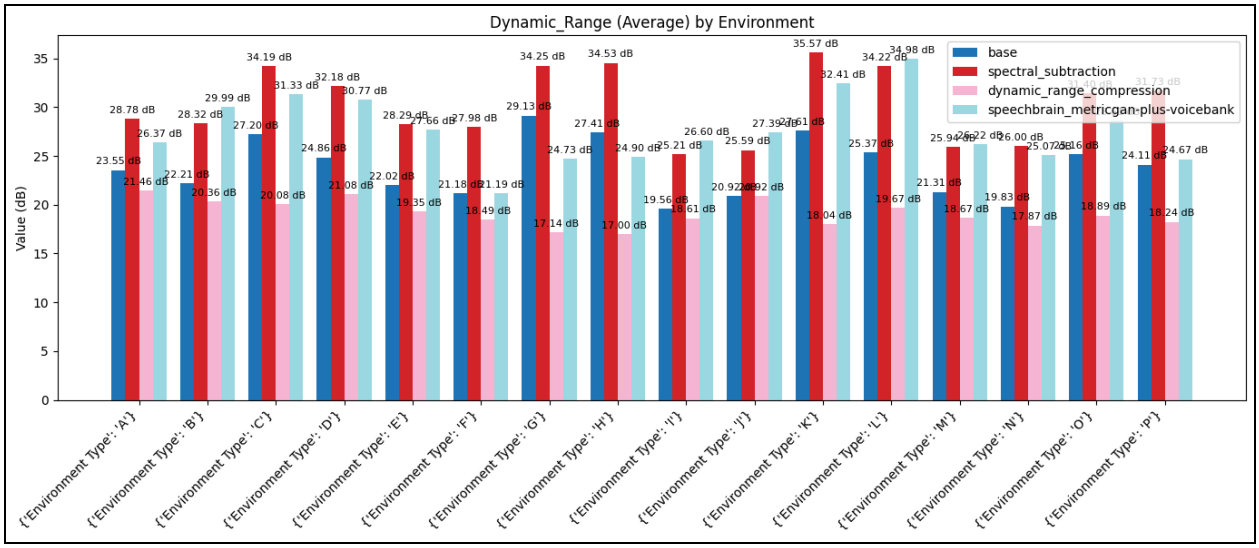

Figure 13: Dynamic Range measured (dB) across all environments for all selected audio-processing models

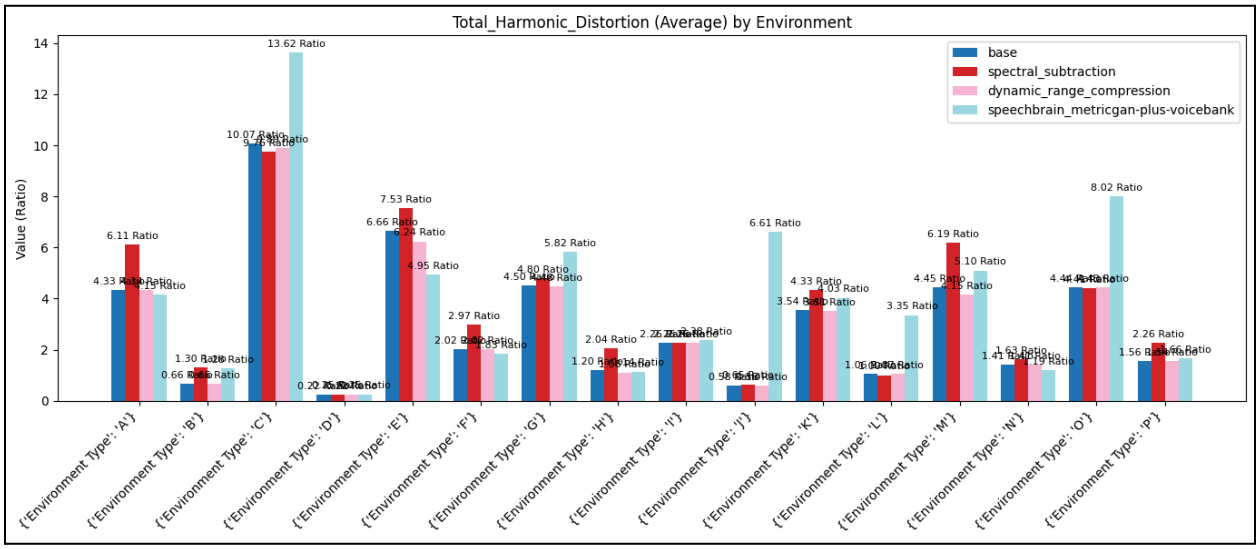

Figure 14: Total Harmonic Distortion ratio measured across all environments for all selected audio-processing models.

7.3 Insights

For this experiment, the team passed this data through the second data comparison function, which produces separate values for each specified environment. Six comparison

graphs were produced, one for each metric, with each graph containing a breakdown of each model's performance in each environment.

Based on selected models, the benchmark should have expected the following results across environments. First, Dynamic Range Compression should have produced the lowest Dynamic Range results across models and, consequently raised the Noise Floor across environments. This was reflected in the outputs as Dynamic Range Compression consistently produced the lowest Dynamic Range values across every environment type (*see Figure 13)*. In addition, Dynamic Range Compression averaged the highest Noise Floor out of all the models for each environment type (*see Figure 9*). Second, as a machine learning model designed to isolate speech, the MetricGAN benchmark was expected to yield a lower Waveform Complexity Index than other models, which was also reflected across environments (*see Figure 12*). Third, the spectral subtraction model is designed to maximize a general signal-noise ratio while lowering the Noise Floor. Across most environments, this model yielded the highest SNR values and resulted in the lowest Noise Floor across all environments (*see Figure 9 and Figure 10*).

This analysis also demonstrated the diversity of the environments included in this research. For the majority of the metrics calculated, there was a very high variance over different environment types. This variance is especially visible in metrics such as Signal-to-Noise Ratio (see *Figure 10)*, Total Harmonic Distortion *(see Figure 14)*, Waveform Complexity Index (see *Figure 12)*, and Noise Floor *(see Figure 9)*. When comparing the environments holistically, the diversity of the dataset and environments is very apparent. This data diversity is essential in ensuring that these benchmarking results are broadly applicable

(Olson, 2017). Providing data from a wide range of environment types helps to mitigate possible biases that may arise from collecting data from a small subset of similar environments. This improves the reliability of the team's research and ensures actionable insights from the benchmarking process.

The insights both through expected behaviors and environmental differences observed by testing these three models alongside a baseline was a strong indicator that the environmental classification model could highlight clear qualitative differences in audio for separate spaces. The team expects that with the expansion of the dataset and generation of further benchmarks, insights will be gained for strategies to optimize hearing models across broadly encountered spaces for an improvement in audio quality.

8. Discussion

The six metrics selected by Team ECHO are easily reproducible and interpretable, and they have been used in past research regarding hearing aids, especially in the search for more objective metrics that limit undesirable features (Agnew, 1998; Kates et al., 2018). However, recent research has shown that objective features alone may not provide an entirely comprehensive audio analysis (Vinay & Lerch, 2022), so Team ECHO's benchmark metrics can be expanded to include subjective features as well. However, this would not only require a clean version of the script for each of the sixteen environment types, but also account for various tonalities, speeds, and voice types.

Team ECHO's environmental approach to benchmarking allows models to be evaluated on real-world scenarios that include variabilities that may not occur in controlled environments. However, a possible limitation is the question of reproducibility across similar but slightly different environments. For instance, recordings in the same urban locations in different weather lead to varying audio spectrograms (Llorca-Bofi et al., 2024). Thus, although environmental datasets help simulate real-world settings and variabilities, they do not necessarily guarantee the same results when minor changes in the environment occur. Both open source contributions for more audio data as well as further examinations in robust benchmarking to add more tags can help mediate these discrepancies (Chen & Revels., 2016).

Although Team ECHO incorporated a diverse set of environments into the dataset, there are still more factors to consider. The main focuses of the environments were indoor/outdoor, crowded/uncrowded, speaking/non-speaking, and walking/stationary, which the team

determined to be critical, as audio algorithms often struggle with the cocktail problem effect (crowded noise), directionality and distortion (walking), and differing noise selection factors (environment type). What is not considered are specific environmental factors like weather types, urban or rural definitions, or even language spoken, which could impact how sound is perceived through different models.

Additionally, the audio recording process revealed some challenges in defining environmental features. The team defined crowdedness as a binary feature, either crowded or uncrowded for an entire audio clip. However, the level of crowdedness in an environment was observed to change significantly across the duration of the clip as a result of people leaving and entering the recording area. From this observation, the team believes that a crowdedness feature may be effective when implemented as a continuous value rather than a binary value in order to more reliably define certain audio environments. If continuous values are implemented, further research is required to determine how to best calculate these continuous values for different features. The team recognizes that although binary classifications do work well to define certain environmental features, they do not work as well for others, indicating a potential need for the inclusion of values that are not binary.

The dataset also faces certain limitations, such as the specificity of surrounding noise factors. To cater the results toward the community of people who experience hearing loss and enhance the dataset overall, a survey could be conducted for associating metrics and audio components with positive and negative effects on audio quality per environment. Utilizing questions centered around each environment defined through the dataset could

allow for the environmental training of future audio processing models around these broader classifications. The team developed a sample survey to receive feedback from those with hearing loss about the most common types of environments and disruptions they hear, which can be further incorporated into the generation and labeling of the dataset for future expansion (*see Appendix B*).

9. Conclusion

Team ECHO's work in benchmarking provides an easily accessible method of evaluating various algorithmic approaches utilized in hearing aids. Through six objective traditional metrics for benchmarking, Team ECHO provided a pipeline that inputs a database of .wav files into a model and generates the metrics and their graphical visualizations. The benchmarking system has been successfully demonstrated via spectral subtraction, Dynamic Range Compression, and a MetricGAN+ speech bank model, where the metric scores for the audio models improve upon the scores for the baseline with no audio filtering. Researchers hoping to test their audio models in the future can quickly import this benchmarking system and utilize its dataset and functionality through the processing and visualization scripts provided by Team ECHO.

Team ECHO's work additionally provides a diverse set of environments that account for the irregularities often experienced in real life, contrasting with many preprocessed audio datasets that currently exist. The benchmarking system's focus on lower-level benchmark metrics helps bridge the gap between the higher-level metrics that are used in current audio benchmarking systems and the lower-level metrics that are more commonly used to evaluate audio signals.

As audio benchmarking and audio clarity modeling are still growing fields, there is still room for future work. Since the dataset produced is labeled to indicate buckets for environments, an algorithmic model can perhaps be trained to accommodate each of the combinations of these factors. Additionally, further research can be conducted to enhance the robustness of the benchmarking system so that the evaluatory context of the

benchmarking system can be generalized beyond the specific environments presented in the database. The database itself can also be expanded to include back-and-forth interactions, rather than just one person speaking within a single recording.

10. References

Abbasi, A., Javed, A. R. R., Yasin, A., Jalil, Z., Kryvinska, N., & Tariq, U. (2022). *A Large-Scale Benchmark Dataset for Anomaly Detection and Rare Event Classification for Audio Forensics*. IEEE Access, 10, 38885–38894. https://doi.org/10.1109/access.2022.3166602

Agnew J. (1998). The causes and effects of distortion and internal noise in hearing aids. *Trends in amplification*, *3*(3), 82–118. https://doi.org/10.1177/108471389800300302

Babel, M. (2022). Adaptation to Social-Linguistic Associations in Audio-Visual Speech. *Brain Sciences*, *12*(7), 845. https://doi.org/10.3390/brainsci12070845

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *27*(2), 113–120. https://doi.org/10.1109/TASSP.1979.1163209

Brody, J. E. (2015, October). *Hearing Loss Costs Far More Than Ability to Hear*. Well. https://archive.nytimes.com/well.blogs.nytimes.com/2015/09/28/hearing-loss-costs-far-more-than-ability-to-hear/

Chen, G., Chai, S., Wang, G., Du, J., Zhang, W.-Q., Weng, C., Su, D., Povey, D., Trmal, J., Zhang, J., Jin, M., Khudanpur, S., Watanabe, S., Zhao, S., Zou, W., Li, X., Yao, X., Wang, Y., Wang, Y., … Yan, Z. (2021). *Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio* (No. arXiv:2106.06909). arXiv. https://doi.org/10.48550/arXiv.2106.06909

Chen, J., & Revels, J. (2016). Robust benchmarking in noisy environments. *arXiv preprint arXiv:1608.04295*.

Chisolm, T. H., Johnson, C. E., Danhauer, J. L., Portz, L. J., Abrams, H. B., Lesner, S., ... & Newman, C. W. (2007). A systematic review of health-related quality of life and hearing aids: final report of the American Academy of Audiology Task Force on the Health-Related Quality of Life Benefits of Amplification in Adults. *Journal of the American Academy of Audiology*, *18*(02), 151-183.

Christensen, J. H., Saunders, G. H., Havtorn, L., & Pontoppidan, N. H. (2021). Real-world hearing aid usage patterns and smartphone connectivity. *Frontiers in Digital Health*, *3*, 722186. https://doi.org/10.3389/fdgth.2021.722186

Cox, R. M., Alexander, G. C., & Gray, G. A. (2005). Who wants a hearing aid? Personality profiles of hearing aid seekers. Ear and hearing, 26(1), 12-26.

Dillon, H. (2012). *Hearing Aids*. Thieme.

Gillard, Danielle & Sharon, Jeffrey. (2022). Understanding the Cost-Effectiveness of Hearing Aids and Surgery for the Treatment of Otosclerosis. Current Otorhinolaryngology Reports. 10. 1-7. 10.1007/s40136-021-00378-y.

Glista D, O'Hagan R, Cornelisse L, Shah T, Hayes D, Doherty S, Gilliland J, Scollie S. (2020, May). Combining passive and interactive techniques in tracking auditory ecology in hearing aid use. *Hearing Review*. https://hearingreview.com/inside-hearing/research/techniques-in-tracking-auditry-ecology-in-hearing-aid-use

Heggan, C., Budgett, S., Hospedales, T., & Yaghoobi, M. (2022, September). MetaAudio: A few-shot audio classification benchmark. In *International Conference on Artificial Neural Networks* (pp. 219-230). Cham: Springer International Publishing.

Huwel, A., Kamil Adiloglu, & Bach, J.-H. (2020). *Hearing aid Research Data Set for Acoustic Environment Recognition.* ICASSP 2022 - 2022 IEEE International Conference on

Acoustics, Speech and Signal Processing (ICASSP), 706–710. https://doi.org/10.1109/icassp40776.2020.9053611

Kashyap, M. M., Tambwekar, A., Manohara, K., & Natarajan, S. (2021). *Speech denoising without clean training data: A noise2noise approach* (No. arXiv:2104.03838). arXiv. https://doi.org/10.48550/arXiv.2104.03838

Kates, J. M., Arehart, K. H., Anderson, M. C., Kumar Muralimanohar, R., & Harvey, L. O., Jr (2018). Using Objective Metrics to Measure Hearing Aid Performance. *Ear and hearing*, *39*(6), 1165–1175. https://doi.org/10.1097/AUD.0000000000000574

Knoetze, M., Manchaiah, V., Oosthuizen, I., Beukes, E., & Swanepoel, D. W. (2024). Perspectives on hearing aid cost and uptake for prescription and over-the-counter hearing aid users. *American Journal of Audiology*, *33*(3), 942-952.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Dollár, P. (2014). Microsoft COCO: Common objects in context. *European Conference on Computer Vision (ECCV)*. https://doi.org/10.1007/978-3-319-10602-1_48

Llorca-Bofí, J., Heck, J., Dreier, C., & Vorländer, M. (2024). Urban background sounds under various weather conditions categorized for virtual acoustics. *Journal of Environmental Management*, *371*, 123081. https://doi.org/10.1016/j.jenvman.2024.123081

Manocha, P., Jin, Z., & Finkelstein, A. (2022). *Audio similarity is unreliable as a proxy for audio quality* (No. arXiv:2206.13411). arXiv. https://doi.org/10.48550/arXiv.2206.13411

Mattson, P., Reddi, V. J., Cheng, C., Coleman, C., Diamos, G., Kanter, D., ... & Zaharia, M. (2020). MLPerf: An industry standard benchmark suite for machine learning performance. *arXiv preprint arXiv:1911.02549*. https://arxiv.org/abs/1911.02549

May, Tobias, Kowalewski, Borys, & Dau, Torsten. (2020). *Scene-Aware Dynamic-Range Compression in Hearing Aids*. 10.1007/978-3-030-00386-9_25.

McCormack, A., & Fortnum, H. (2013). Why do people fitted with hearing aids not wear them?. *International journal of audiology*, *52*(5), 360-368.

Mendiratta, A., & Jha, D. (2014, January). Adaptive noise cancelling for audio signals using least mean square algorithm. In *International conference on electronics, communication and instrumentation (ICECI)* (pp. 1-4). IEEE.

NIDCD. (2024, June 13). *Over-the-Counter Hearing Aids*. Retrieved March 27, 2025 from https://www.nidcd.nih.gov/health/over-counter-hearing-aids

Nordqvist, P., & Leijon, A. (2004). An efficient robust sound classification algorithm for hearing aids. *The Journal of the Acoustical Society of America*, *115*(6), 3033-3041.

Olson, Randal S., et al. (2017) *PMLB : A Large Benchmark Suite for Machine Learning Evaluation and Comparison*. arXiv:1703.00512, arXiv, 1 Mar. 2017. arXiv.org, https://doi.org/10.48550/arXiv.1703.00512.

Park, G., Cho, W., Kim, K.-S., & Lee, S. (2020). Speech enhancement for hearing aids with deep learning on environmental noises. Applied Sciences, 10(17), 6077. https://doi.org/10.3390/app10176077

Pascual, S., Bonafonte, A., & Serrà, J. (2017). *Segan: Speech enhancement generative adversarial network* (No. arXiv:1703.09452). arXiv. https://doi.org/10.48550/arXiv.1703.09452

Piczak K. J. (2015). *ESC: Dataset for Environmental Sound Classification*. Proceedings of the 23rd Annual ACM Conference on Multimedia, Brisbane, Australia.

Rezmeriţă, G., Bordianu, A., & Steliana Valentina Puscasu. (2023). *Low-Pass Filter Analysis*. https://doi.org/10.1109/atee58038.2023.10108156

Sockalingam, R., Beilin, J., & Beck, D. (2009, March 3). *Sound Quality Considerations of Hearing Instruments*. The Hearing Review. https://hearingreview.com/hearing-products/accessories/earmolds/sound-quality-considerations-of-hearing-instruments

Stapelberg, Belinda, & Malan, Katherine M. (2020). A survey of benchmarking frameworks for reinforcement learning. *South African Computer Journal*, *32*(2), 258-292. https://doi.org/10.18489/sacj.v32i2.746

Tsimpida, D., Kontopantelis, E., Ashcroft, D., & Panagioti, M. (2019). Socioeconomic and lifestyle factors associated with hearing loss in older adults: a cross-sectional study of the English Longitudinal Study of Ageing (ELSA). *BMJ Open*, *9*(9), e031030. https://doi.org/10.1136/bmjopen-2019-031030

Tsimpida, D., Kontopantelis, E., Ashcroft, D. M., & Panagioti, M. (2021). The dynamic relationship between hearing loss, quality of life, socioeconomic position and depression and the impact of hearing aids: answers from the English Longitudinal Study of Ageing (ELSA). *Social Psychiatry and Psychiatric Epidemiology*. https://doi.org/10.1007/s00127-021-02155-0

Van Buuren, R. A., Festen, J. M., & Houtgast, T. (1996). Peaks in the frequency response of hearing aids: Evaluation of the effects on speech intelligibility and sound quality. *Journal of Speech, Language, and Hearing Research*, *39*(2), 239–250. https://doi.org/10.1044/jshr.3902.239

Vinay, A., & Lerch, A. (2022). *Evaluating generative audio systems and their metrics* (No. arXiv:2209.00130). arXiv. https://doi.org/10.48550/arXiv.2209.00130

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S. R. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32. https://arxiv.org/abs/1905.00537

Wang, B., Zou, X., Lin, G., Sun, S., Liu, Z., Zhang, W., ... & Chen, N. F. (2024). Audiobench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020.*

Weinberger, Steven. (2015). *Speech Accent Archive*. George Mason University. Retrieved from http://accent.gmu.edu

Wilson, B. S., Tucci, D. L., Merson, M. H., & O'Donoghue, G. M. (2017). Global hearing health care: new findings and perspectives. *The Lancet*, *390*(10111), 2503–2515. https://doi.org/10.1016/s0140-6736(17)31073-5

Xiang, J., McKinney, M. F., Fitz, K., & Zhang, T. (2010, March). Evaluation of sound classification algorithms for hearing aid applications. In *2010 IEEE international conference on acoustics, speech and signal processing* (pp. 185-188). IEEE.

Zöller, M.-A., & Huber, M. F. (2021). Benchmark and survey of automated machine learning frameworks. Journal of Artificial Intelligence Research, 70, 409–472. https://doi.org/10.1613/jair.1.11854

11. Appendix

Appendix A: Echo Codebase + Dataset Link With Full Data Table

Github: https://github.com/rahulnair2003/gemstone/tree/main

ID	Length (S)	Location	Indoors	Crowded	Speaking	Walking	Environment Type	Voice Type	Voice ID
A-M01-001	278	hallway	Yes	Yes	Yes	Yes	A	Male	M01
A-M02-002	292	hallway	Yes	Yes	Yes	Yes	A	Male	M02
B-F04-005	173	restaurant	Yes	Yes	Yes	No	B	Female	F04
B-F04-010	153	restaurant	Yes	Yes	Yes	No	B	Female	F04
B-M01-007	135	restaurant	Yes	Yes	Yes	No	B	Male	M01
B-M01-008	140	restaurant	Yes	Yes	Yes	No	B	Male	M01
B-M02-001	128	lobby	Yes	Yes	Yes	No	B	Male	M02
B-M02-004	245	hallway	Yes	Yes	Yes	No	B	Male	M02
B-M02-006	173	restaurant	Yes	Yes	Yes	No	B	Male	M02
B-M02-009	165	restaurant	Yes	Yes	Yes	No	B	Male	M02
B-M04-002	125	lobby	Yes	Yes	Yes	No	B	Male	M04
B-M05-003	253	hallway	Yes	Yes	Yes	No	B	Male	M05
C-N00-001	301	hallway	Yes	Yes	No	Yes	C	N/A	N/A
C-N00-002	317	hallway	Yes	Yes	No	Yes	C	N/A	N/A

C-N00-003	322	cafeteria	Yes	Yes	No	Yes	C	N/A	N/A
C-N00-004	301	cafeteria	Yes	Yes	No	Yes	C	N/A	N/A
C-N00-005	611	cafeteria	Yes	Yes	No	Yes	C	N/A	N/A
C-N00-006	610	cafeteria	Yes	Yes	No	Yes	C	N/A	N/A
D-N00-001	302	restaurant	Yes	Yes	No	No	D	N/A	N/A
D-N00-002	306	restaurant	Yes	Yes	No	No	D	N/A	N/A
D-N00-003	305	hallway	Yes	Yes	No	No	D	N/A	N/A
D-N00-004	301	hallway	Yes	Yes	No	No	D	N/A	N/A
E-F03-003	330	hallway	Yes	No	Yes	Yes	E	Female	F03
E-M01-001	290	hallway	Yes	No	Yes	Yes	E	Male	M01
E-M01-005	298	lobby	Yes	No	Yes	Yes	E	Male	M01
E-M02-002	258	hallway	Yes	No	Yes	Yes	E	Male	M02
E-M06-004	347	lobby	Yes	No	Yes	Yes	E	Male	M06
F-F01-008	392	hallway	Yes	No	Yes	No	F	Female	F01
F-M01-003	168	library	Yes	No	Yes	No	F	Male	M01
F-M01-004	146	library	Yes	No	Yes	No	F	Male	M01
F-M01-006	275	lab	Yes	No	Yes	No	F	Male	M01
F-M02-001	180	library	Yes	No	Yes	No	F	Male	M02
F-M02-002	180	library	Yes	No	Yes	No	F	Male	M02

F-M03-007	293	lab	Yes	No	Yes	No	F	Male	M03
F-M06-005	315	lab	Yes	No	Yes	No	F	Male	M06
G-N00-001	301	stairwell	Yes	No	No	Yes	G	N/A	N/A
G-N00-002	302	stairwell	Yes	No	No	Yes	G	N/A	N/A
G-N00-003	306	hallway	Yes	No	No	Yes	G	N/A	N/A
G-N00-004	301	hallway	Yes	No	No	Yes	G	N/A	N/A
G-N00-005	304	hallway	Yes	No	No	Yes	G	N/A	N/A
G-N00-006	304	hallway, stairwell	Yes	No	No	Yes	G	N/A	N/A
H-N00-001	305	library	Yes	No	No	No	H	N/A	N/A
H-N00-002	316	library	Yes	No	No	No	H	N/A	N/A
H-N00-003	329	hallway	Yes	No	No	No	H	N/A	N/A
H-N00-004	304	hallway	Yes	No	No	No	H	N/A	N/A
I-F01-001	288	sidewalk	No	Yes	Yes	Yes	I	Female	F01
J-F03-001	240	sidewalk	No	Yes	Yes	No	J	Female	F03
K-N00-001	303	sidewalk	No	Yes	No	Yes	K	N/A	N/A
L-N00-001	304	sidewalk	No	Yes	No	No	L	N/A	N/A
M-F01-008	347	11b parking lot	No	No	Yes	Yes	M	Female	F01
M-F03-003	406	plaza	No	No	Yes	Yes	M	Female	F03

M-M01-004	155	courtyard	No	No	Yes	Yes	M	Male	M01
M-M01-005	155	courtyard	No	No	Yes	Yes	M	Male	M01
M-M01-009	279	sidewalk	No	No	Yes	Yes	M	Male	M01
M-M02-002	246	park	No	No	Yes	Yes	M	Male	M02
M-M02-006	132	courtyard	No	No	Yes	Yes	M	Male	M02
M-M02-007	130	courtyard	No	No	Yes	Yes	M	Male	M02
M-M05-001	250	park	No	No	Yes	Yes	M	Male	M05
M-M06-010	245	sidewalk	No	No	Yes	Yes	M	Male	M06
N-F01-005	384	plaza	No	No	Yes	No	N	Female	F01
N-F03-010	298	11b parking lot	No	No	Yes	No	N	Female	F03
N-M01-006	144	courtyard	No	No	Yes	No	N	Male	M01
N-M01-007	128	courtyard	No	No	Yes	No	N	Male	M01
N-M02-001	127	courtyard	No	No	Yes	No	N	Male	M02
N-M02-002	121	courtyard	No	No	Yes	No	N	Male	M02
N-M02-008	135	courtyard	No	No	Yes	No	N	Male	M02
N-M02-009	142	courtyard	No	No	Yes	No	N	Male	M02
N-M04-003	143	courtyard	No	No	Yes	No	N	Male	M04
N-M04-004	135	courtyard	No	No	Yes	No	N	Male	M04
O-N00-001	301	park	No	No	No	Yes	O	N/A	N/A

O-N00-002	307	park	No	No	No	Yes	O	N/A	N/A
O-N00-003	305	courtyard	No	No	No	Yes	O	N/A	N/A
O-N00-004	301	courtyard	No	No	No	Yes	O	N/A	N/A
O-N00-005	306	plaza	No	No	No	Yes	O	N/A	N/A
O-N00-006	331	courtyard	No	No	No	Yes	O	N/A	N/A
O-N00-007	304	courtyard	No	No	No	Yes	O	N/A	N/A
O-N00-008	302	sidewalk	No	No	No	Yes	O	N/A	N/A
O-N00-009	312	sidewalk	No	No	No	Yes	O	N/A	N/A
P-N00-001	301	courtyard	No	No	No	No	P	N/A	N/A
P-N00-002	302	courtyard	No	No	No	No	P	N/A	N/A
P-N00-003	300	courtyard	No	No	No	No	P	N/A	N/A
P-N00-004	305	courtyard	No	No	No	No	P	N/A	N/A
P-N00-005	308	plaza	No	No	No	No	P	N/A	N/A
P-N00-006	305	courtyard	No	No	No	No	P	N/A	N/A
P-N00-007	302	courtyard	No	No	No	No	P	N/A	N/A
P-N00-008	301	sidewalk	No	No	No	No	P	N/A	N/A
P-N00-009	302	sidewalk	No	No	No	No	P	N/A	N/A
E-F01-006	262	hallway	Yes	No	Yes	Yes	E	Female	F01
E-F03-007	269	hallway	Yes	No	Yes	Yes	E	Female	F03

F-F01-009	253	hallway	Yes	No	Yes	No	F	Female	F01
F-F03-010	267	hallway	Yes	No	Yes	No	F	Female	F03

Appendix B: Surrounding Versus Desired Noise Survey

1. Age
2. Gender
3. Are you Deaf or hard of hearing?
    a. If yes:
        i. How long have you been Deaf/hard of hearing?
        ii. Do you utilize hearing aids?
            1. If yes:
                a. How long have you been using hearing aids?
                b. What type of hearing aids do you use?
                c. I am satisfied with my current experience. (1-5, strongly disagree - strongly agree)
                d. I am able to hear everything that I want to hear
                e. My hearing aids are a positive influence on my daily life
            2. If no:
                a. Why do you not use hearing aids?
        iii. What are some drawbacks of using hearing aids, or aspects of hearing aids that could be improved?
        iv. Are there any specific environments or sounds that are difficult to aurally process even with the use of hearing aids?

b. If no: skip to 4

4. Describe what you would want to hear if you were in the environment shown in the picture:

5. Describe what you would want to hear if you were in the environment shown in the picture:

6. Describe what you would want to hear if you were in the environment shown in the picture:

7. Describe what you would want to hear if you were in the environment shown in the picture:

8. Describe what you would want to hear if you were in the environment shown in the picture:

9. Describe what you would want to hear if you were in the environment shown in the picture:

10. How would you like to see hearing aids become more accessible?

11. Any additional comments: