ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446544

QUANTIFYING THE EFFECT OF SIMULATOR-BASED DATA AUGMENTATION FOR SPEECH RECOGNITION ON AUGMENTED REALITY GLASSES

Riku Arakawa

Carnegie Mellon University
Pittsburgh, PA, USA
rarakawa@cs.cmu.edu

Mathieu Parvaix, Chiong Lai, Hakan Erdogan, Alex Olwal

Google Research
Mountain View, CA, USA
{parvaix, chionglai, hakanerdogan, olwal}@google.com

ABSTRACT

Augmented reality (AR) glasses have an immense potential for enhancing conversations by leveraging speech recognition to display real-time transcription or translation, for example, to assist people with hearing impairments or for people conversing in a non-native language. For deployment in real environments, such systems, however, need to be able to separate the speech of interest from noise and other speakers. In this paper, we evaluate the effectiveness of leveraging a room simulator to generate large amounts of simulated training data for such front-end sound separation models, to complement the ideal, but costly, collection of real-world data recorded on the device. Using both recorded and simulated impulse responses (IRs), we demonstrate that the use of simulation data is an effective method for training models that can ultimately enhance speech recognition performance in real-world settings. Furthermore, we show that performance can be further improved by adding microphone directivity in the room simulation, and by fusing synthetic data with a small amount of real IRs. Our results also suggest that existing room simulators would benefit from incorporating the head shadow effect, given its significant impact on multi-microphone recordings on AR glasses.

Index Terms— sound separation, speech recognition, room simulator, augmented reality, head-mounted display

1. INTRODUCTION

Recently, augmented reality (AR) glasses have been gaining attention as a platform for leveraging speech models to enhance communication [1, 2]. For example, *Wearable Subtitles* [3] augments communication through all-day speech transcription, the potential of which is demonstrated in multiple user studies with deaf/hard-of-hearing participants. Such augmentation can be especially helpful in group conversations or noisy environments where people may encounter difficulty distinguishing what others say. Hence, accurate sound separation and speech recognition on AR glasses are key in offering a reliable and valuable user experience.

The quantity and quality of the training data are paramount to the raw performance and generalization capabilities of sound separation models. High-quality data captured on device enables the tailoring of models to specific hardware and acoustic environments. However, data collection on device in a real-world setting at a sufficient scale can be challenging, especially when there is a need to support a range of devices with different microphone configurations and characteristics. Synthetic data, on the other hand, has the potential to generate unlimited amounts of data, but the discrepancies between simulated and measured data may negatively impact model performance. In this paper, we examine the feasibility of using room simulators to synthesize impulse responses (IRs) for training sound separation models to enhance speech recognition on AR glasses. While speech enhancement on augmented reality devices has recently been getting more attention [4], to our knowledge, the use of room simulations to improve the performance of speech recognition on an actual device has not yet been addressed.

The contributions of this work are:

- **Real-world recordings with glasses-mounted microphone**. Audio recordings with a microphone on head-worn glasses from 10 sound sources, captured from 72 directions (platform rotates in 5°increments) for 3 rooms with different acoustic properties. The data was captured with and without a head-and-torso simulator (HATS) to provide insights into the head shadow effect.
- **Comparison of real and simulated IRs**. Using room-wise direct-to-reverberant ratio (DRR) as a comparison metric, we show that, while adding microphone directivity brings the simulation closer to the measured IRs without the HATS, the head shadow causes a significant gap between the simulation and the measured IRs with the HATS.
- **Synthesized datasets**. We developed a data generation pipeline to generate at-scale training datasets for supervised machine learning models: mixes of reverberant speech and noise sources with controlled distributions of levels and temporal overlap leveraging data augmentation methods.
- **Results that show the benefits of utilizing simulated IRs**. Our results show significant improvement (19.7% relative WER reduction) when using automatic speech recognition (ASR) front-end models trained on a small amount of real-world measurements and simulated IRs with microphone directivity.

2. RELATED WORK

Prior work has investigated the effectiveness of room simulators in specific speech processing tasks, such as in ASR [5, 6, 7, 8, 9, 10, 11]. For example, Tang *et al.* [7] proposed low-frequency compensated synthetic IRs and showed an improvement in far-field ASR tasks using the LibriSpeech dataset. The efficacy of such data augmentation has been demonstrated in other tasks as well; for instance, Koyama *et al.* [12] used a room simulator for the sound event localization task. Srivastava *et al.* [13] showed that making the simulation more realistic leads to enhanced blind acoustic parameter

979-8-3503-4485-1/24/$31.00 ©2024 IEEE

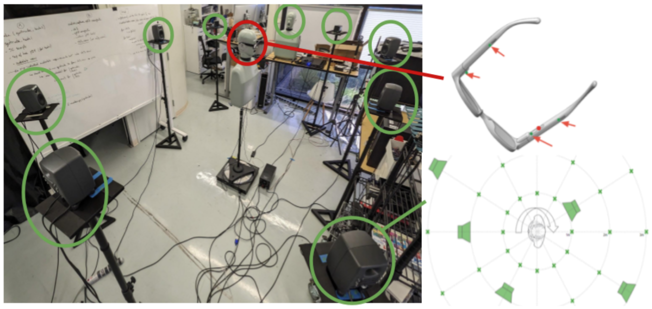

Fig. 1. Data collection setup. The glasses-mounted microphone is positioned on a rotating platform, surrounded by 9 speakers, in addition to a 10th mouth simulation speaker. 720 IR recordings were made in 3 different rooms with and without a head-and-torso simulator (HATS).

estimators. Similarly, Bezzam *et al.* [14] showed that a more realistic room simulation that incorporates factors such as air absorption and frequency-dependent surface improves the performance of far-field keyword spotting. StoRIR [15] is a stochastic room impulse response generation method to simulate acoustic parameters of rooms, and its effectiveness was shown in the speech enhancement task. Aralikatti *et al.* [16] confirmed the effectiveness of combining synthetic IRs with real IRs in training sound separation model using the VOiCES dataset (Voices Obscured in Complex Environmental Settings) [17].

These studies demonstrate the benefits of using room simulators for different tasks, as well as, the efficacy of making the simulation more realistic, such that it matches real-world conditions. In this work, we examine such approaches for training a sound separation model used as a front-end to speech recognition on AR glasses, where unique challenges like the device being head-worn are expected. Our work does not only open up a practical method for speech applications on an emerging computing platform but also provides further evidence of the effectiveness of simulator-based data augmentation, in addition to directions for future work.

3. DATASET

To investigate the efficacy of simulated IRs in comparison to measured IRs for training sound separation models as front-end to ASR, we recorded IRs on an AR glasses prototype in multiple environments, and we also generated synthetic IRs using a room simulator.

3.1. Real-World Recording: 2160 IRs with and without HATS

We use the exponential sine sweep method [18] to measure the IR from a speaker in the room to the microphone on AR glasses. The glasses are positioned on the HATS, which is placed on a motorized turn-table that we programmatically rotate in 5°steps to vary the angle between the microphone and the sound sources across recordings. Nine loudspeakers are placed around the HATS, at different heights and distances, and the mouth simulator speaker provides the 10th sound source, as shown in Figure 1. For each angle, a sine sweep is played from each speaker and recorded on the glasses microphone, resulting in 720 recordings for each room ($360^\circ/5^\circ \times 10$ sound sources). While this study keeps the sound sources and microphone stationary during the IR recordings, we are interested in exploring moving configurations in future work, to better represent dynamic environments.

Table 1. Characteristics of rooms used for our IR recording. M2L is the microphone to loudspeaker distance, which varied across the 720 recordings for each room.

	RT60	L	W	H	M2L
room_1	0.70s	4.0m	6.0m	5.2m	1.0–3.5m
room_2	1.00s	6.9m	7.7m	5.4m	1.2–4.2m
room_3	0.82s	3.8m	5.0m	5.4m	1.1–1.9m

We conducted recordings in three different room environments, varying the position of the nine speakers around the HATS. The chosen rooms have relatively high RT60 values (*i.e.*, *Reverberation Time 60* is the time for a sound to decay by 60 dB from its original level) such that we can evaluate the sound separation model in realistic and thereby challenging situations. The characteristics of each room are presented in Table 1.

We also recorded IRs on AR glasses without the HATS, where the glasses were attached to a tripod. All the other elements of the setup were kept the same as when the HATS was present. These measurements are meant to provide data without the effect of head shadow caused by the HATS since the room simulator used does not model this effect.

3.2. Simulation

3.2.1. Room simulator extended with frequency-dependent reflections and microphone directivity

The room simulator [19] is based on a frequency domain implementation of the image source method [20]. We extend the simulator to consider frequency-dependent reflections by representing each wall with an FIR causal wall filter, for a more realistic simulation. More specifically, we decompose each delay into integer and quantized fractional parts. For the quantized fractional part, we calculate the FFT of a windowed sinc function corresponding to that fractional delay. We also extended the capabilities of the room simulator to simulate microphone directivity, which is critical to simulate microphones embedded in wearable devices adequately. We used a cardioid microphone pattern based on the microphone's direction in our prototype.

3.2.2. Geometry-matching simulated IRs: 2160 IRs

We synthesized IRs that matched the recording conditions from the real-world recording, matching room geometry, microphone position, speaker positions, as well as the room RT60. We thus had 2160 (720×3 rooms) synthetic IRs that corresponded to the real-world recordings. The simulator uses the given RT60 to match the overall value of the generated IR.

3.2.3. Synthesized IRs: 8000 IRs with and without directivity

We also synthesized IRs based on room characteristics randomly sampled from predefined ranges for room length, height, width, and RT60 to use in the experiments (Section 4). The microphone position, orientation, and speaker position were also randomly sampled so that they were all within the room, and the distance between the microphone and the speaker was within the range of 0.2–4.0 m. The height of the speakers was constrained to the range of 1.3–2.1 m to represent average human talkers. See Table 2.

Table 2. Characteristics of the generated rooms for the synthesized data set. M2L is the microphone to loudspeaker distance. S_{height} is the height of the speakers.

RT60	L / W	H	M2L	S_{height}
0.2–1.5s	2–6m	2–5m	0.2–4.0m	1.3–2.1m

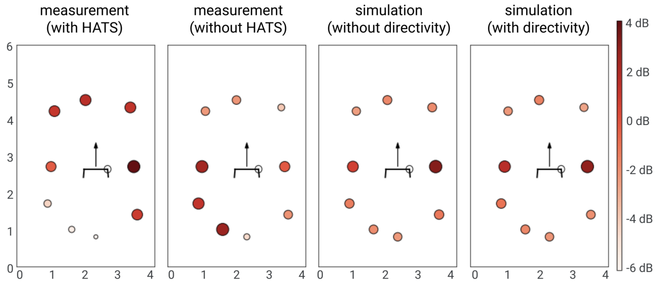

Fig. 2. Example plot of DRRs from different speakers in a room. The circle color and size indicate the speaker's DRR, normalized within each plot. The two leftmost plots highlight the impact of the head shadow, captured with HATS. The two rightmost plots show how microphone directivity provides a more realistic simulation.

3.3. Comparison

We first did a low-level comparison of the measured and simulated IRs before assessing the effectiveness of simulated IRs to generate training data for speech separation models in the next section. Here, we sought to understand how similar synthetic IRs are to real ones. Among several metrics existing to compare IRs [21], we used the DRR, in dB, to assess the reverberation characteristics of different source-to-microphone positions.

Figure 2 shows an example plot of the DRR values for different speakers in the room and compares the measured and simulated IRs. In this plot, the glasses face the direction represented by the black arrow, and the white dot indicates the target microphone we used to calculate the DRRs, which faces inward. The color and size of the circles around the glasses represent the DRR value. First, by comparing the measurements with and without the HATS, it is observed that the head shadow plays a key role in affecting the DRR values. For example, the DRR for the speaker directly to the left of the glasses is weaker in the measurement with the HATS than without the HATS. This is because the energy of the direct path is significantly reduced in the presence of the head. Second, when we compare the simulation with and without the microphone directivity, it is observed that the DRR value on the left speaker is stronger due to the simulated directivity. This is because the reverberation path becomes weaker by more than 75% due to the directivity pattern while the direct path remains the same, thus increasing the DRR value. Similarly, for the right speaker, its DRR becomes smaller since the energy of the direct path is weaker due to the directivity. As a result, the DRR distribution in the simulation is closer to the measurement without the HATS. To summarize, adding microphone directivity can improve the realism of the simulation, but there are still differences between the measured and simulated IRs due to the head shadow effect. Similar behavior is observed when the glasses are facing different angles.

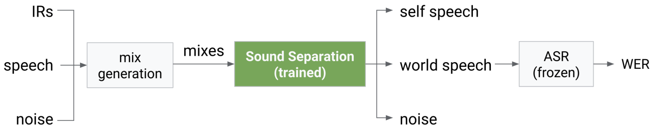

Fig. 3. Speech processing setting. For comparison, we only changed the type and number of IRs used to train sound separation models.

4. EXPERIMENTS

4.1. Settings

In our scenario, a *user* is wearing the AR glasses while having a conversation with one or more persons (*world talkers*). Our goal is to transcribe or translate the *world speech*, while separating out noise and the user's own speech (*self speech*). Note, in this work, we focus on single-channel sound separation and speech recognition, and using multiple microphones on the glasses remains an important future work. The microphone used is on the right side of the frame, facing the template, as illustrated in Figure 2.

We first created audio mixes that mix together *self speech* and *world speech*, as well as background noise. The reverberant speech and noise signals are created by convolving measured or simulated IRs with prerecorded audio datasets. Specifically, we used the LibriSpeech dataset [22] for speech, fsd50k dataset [23] for noise, and ambient noise recorded in our environment. All the parameters used to create the reverberant noisy mixtures are kept constant across all datasets (*e.g.*, same distribution of speech and noise levels, or temporal overlap between signals). Only the IRs differ across datasets. We developed a data generation pipeline that allows us to adjust the level of each sound source and their temporal onset and offset to match pre-defined distributions of signal-to-noise and signal-to-interference ratios, as well as temporal overlap. The availability of separate audio sources composing the mixture allows us to create labels for our supervised learning process. Training and evaluation datasets are created using this method. The duration of generated mixes was 150 hours and 30 hours for training and evaluation, respectively, when we used real-world measured IRs or geometry-matching simulated IRs. For the synthesized IRs, we created 375 hours and 75 hours for training and evaluation. We randomly chose room_2 as a test room and used the other rooms for training.

We used a sound separation model based on an improved time-domain convolutional network (TDCN++) described in [24]. We used an STFT basis for the analysis and synthesis with a 32ms window (512 samples at 16kHz) and 16ms hop. The model is trained to separate three output channels: *self speech*, *world speech*, and noise. For ASR, we used an on-device model of Google's text-to-speech API [25], a pretrained model without finetuning.

We compared the following training conditions for the sound separation model. The *Baseline* model uses no sound separation model, thus applying ASR to the unprocessed mixed audio. Note that in this case, we use the available onset/offset for the *world speech* sources to only send the relevant portions of the mix signal to the ASR. The models are denoted by *M* and *S* to represent the use of the measured and simulated IRs, respectively. Their subscript value represents the number of IRs used in their training and corresponds; for example, M_{1440} means 1440 measured IRs. The subscript G indicates that they are geometry-matching simulated IRs. Moreover, the subscript D means that their simulation involves microphone directivity. Hybrid models with a small amount of measured and simulated IRs are also included in the comparison, such as $M_{720}S_{4000}$.

Table 3. Comparison of different data augmentation approaches with measured and simulated IRs. *M=Measured; S=Simulated; subscript=number of IRs, +D for microphone directivity in simulation, +G for simulations that use matching geometry to the measured rooms*. The WER reduction is the improvement in WER compared to the baseline (without sound separation).

		MEASURED		SIMULATED			SIMULATED WITH DIRECTIVITY			HYBRID			HYBRID WITH DIRECTIVITY		
Model	Baseline	M_{720}	M_{1440}	S_{1440G}	S_{4000}	S_{8000}	S_{1440GD}	S_{4000D}	S_{8000D}	$M_{720}S_{1440G}$	$M_{720}S_{4000}$	$M_{720}S_{8000}$	$M_{720}S_{1440GD}$	$M_{720}S_{4000D}$	$M_{720}S_{8000D}$
WER (mean %)	37.6	34.8	31.8	36.3	34.4	34.0	34.5	32.2	31.8	32.3	31.9	30.5	31.7	31.6	30.2
Rel. WER reduction (%)	-	7.4	15.4	3.5	8.5	9.6	8.2	14.4	15.4	14.1	15.2	18.9	15.7	16.0	19.7
Measured IRs		720	1440							720	720	720	720	720	720
Simulated IRs				1440	4000	8000	1440	4000	8000	1440	4000	8000	1440	4000	8000

Smallest WER (30.2%) → Largest WER (37.6%)

Largest relative WER reduction (19.7%) → Smallest relative WER reduction (3.5%)

Table 4. Results when using test data generated by IRs recorded without the HATS. $M_{1440NO-HATS}$ is trained on the IRs captured without HATS and, therefore, do not take head occlusion into account. *M=Measured; S=Simulated; subscript=number of IRs, +D for microphone directivity in simulation, +G for simulations that use matching geometry to the measured rooms*. The WER reduction is the improvement in WER compared to the baseline (without sound separation).

		MEASURED	SIMULATED	SIMULATED WITH DIRECTIVITY
Model	Baseline	M_{1440}NO-HATS	S_{1440}G	S_{1440}GD
WER (mean %)	38.0	34.8	36.3	34.5
Rel. WER reduction (%)	-	16.6	12.6	15.3
Measured IRs		1440		
Simulated IRs			1440	1440

Lastly, to quantify the impact of the head shadow effect caused by the presence of a head on the effectiveness of the simulated IRs, we tested our pipeline on the mixed audio generated using the IRs measured without the HATS.

4.2. Results

The main results are presented in Table 3. The overall WER, an objective proxy for the final user experience in our use case, is relatively high, highlighting the challenge for speech recognition on AR glasses in realistic conditions (*i.e.*, both reverberant and noisy). Comparing the WER for M_{720} and M_{1440}, it is clear that having more training data improves the performance of the model, as expected. While a similar trend exists for S_{1440G}, S_{4000}, and S_{8000}, their improvement falls short of that of M_{720}, highlighting the gap between measured and simulated IRs. However, when the microphone directivity is added to the simulation, the improvement increases, and S_{8000D} achieves a comparable performance as M_{1440}. This indicates that increasing the realism of the simulation can lead to better model training, which aligns with prior observations in different tasks, as discussed in Section 2. Moreover, the performance is further improved when the simulation is fused with a small amount of measurements; for example, $M_{720}S_{8000}$ outperforms M_{1440}. This finding is particularly beneficial when the resource to collect real-world data is limited and shows the potential of using a room simulator for data augmentation.

As a reference point, we also ran the reverberant target speech signal through our ASR model, without added noise or partially overlapping speech. The WER obtained was 18%, which serves as the upper bound for the performance. The WER of 37.6% without the sound separation (the Baseline model) suggests that our evaluation set is particularly challenging for the ASR model used. Note that this baseline score was computed using ground truth onset/offset information and thus does not precisely reflect real-world user experience where both *self speech* and *world speech* would be transcribed, resulting in an even worse WER. While our best model ($M_{720}S_{8000D}$) reduces the WER to 30.2%, it is still significantly higher than the noise-free, no-speech-overlap case. This gap could be attributed to the limitations of the sound separation model and the distortion of the separated speech signal, to which the ASR model is particularly sensitive.

Table 4 shows the results when the models were tested on the audio created using the IRs without the HATS. Interestingly, the improvement due to the simulated IRs is larger (+12.6% for S_{1440G} and +15.3% for S_{1440GD}) than those tested on the audio generated using IRs measured with the HATS (+3.5% for S_{1440G} and +8.2% for S_{1440GD}), bringing the performance closer to the case using measured IRs ($M_{1440NO-HATS}$). This result implies that the effectiveness of the simulated IRs is limited by the lack of head shadow effect, which the current room simulator does not implement. Thus, future work should focus on incorporating the head shadow effect into the room simulator to further enhance front-end processing for speech recognition on AR glasses.

5. CONCLUSION

While having the potential to unlock many critical applications, speech recognition on AR glasses is challenging, especially in noisy and reverberant conditions. In this work, we quantified the effectiveness of using a room simulator to train a sound separation model used as a speech recognition front-end. Using recorded IRs on an AR glasses prototype in different rooms, we demonstrate that simulated IRs help improve speech recognition by greatly increasing the amount of simulated IRs, by leveraging microphone directivity, and when fused with a small amount of measured IRs. We also highlighted the importance of modeling the head shadow effect, as shown both in the speech recognition results and in the DRR comparison. To date and our knowledge, only a limited amount of work has been done to model the effect using finite element analysis [26], and there is therefore interesting opportunities for future work to address these challenges.

6. REFERENCES

[1] N.A. Basoglu *et al.*, “Exploring adoption of augmented reality smart glasses: Applications in the medical industry,” *Frontiers of Engineering Management*, vol. 5, no. 2, pp. 167–181, 2018.

[2] A. Miller *et al.*, “The use of smart glasses for lecture comprehension by deaf and hard of hearing students,” in *Proc. CHI EA*, 2017, pp. 1909–1915.

[3] A. Olwal *et al.*, “Wearable subtitles: Augmenting spoken communication with lightweight eyewear for all-day captioning,” in *Proc. UIST*, 2020, pp. 1108–1120.

[4] P. Guiraud *et al.*, “An introduction to the speech enhancement for augmented reality (spear) challenge,” in *Proc. IWAENC*, 2022, pp. 1–5.

[5] C. Kim *et al.*, “Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home,” in *Proc. Interspeech*, 2017, pp. 379–383.

[6] T. Ko *et al.*, “A study on data augmentation of reverberant speech for robust speech recognition,” in *Proc. ICASSP*, 2017, pp. 5220–5224.

[7] Z. Tang *et al.*, “Low-frequency compensated synthetic impulse responses for improved far-field speech recognition,” in *Proc. ICASSP*, 2020, pp. 6974–6978.

[8] A. Ratnarajah *et al.*, “IR-GAN: room impulse response generator for far-field speech recognition,” in *Proc. Interspeech*, 2021, pp. 286–290.

[9] A. Ratnarajah *et al.*, “TS-RIR: translated synthetic room impulse responses for speech augmentation,” in *Proc. ASRU*, 2021, pp. 259–266.

[10] F.B. Gelderblom *et al.*, “Synthetic data for dnn-based doa estimation of indoor speech,” in *Proc. ICASSP*, 2021, pp. 4390–4394.

[11] P. Srivastava *et al.*, “How to (virtually) train your speaker localizer,” in *Proc. Interspeech*, 2023, pp. 1204–1208.

[12] Y. Koyama *et al.*, “Spatial data augmentation with simulated room impulse responses for sound event localization and detection,” in *Proc. ICASSP*, 2022, pp. 8872–8876.

[13] P. Srivastava *et al.*, “Realistic sources, receivers and walls improve the generalisability of virtually-supervised blind acoustic parameter estimators,” in *Proc. IWAENC*, 2022, pp. 1–5.

[14] E. Bezzam *et al.*, “A study on more realistic room simulation for far-field keyword spotting,” in *Proc. APSIPA*, 2020, pp. 674–680.

[15] P. Masztalski *et al.*, “Storir: Stochastic room impulse response generation for audio data augmentation,” in *Proc. Interspeech*, 2020, pp. 2857–2861.

[16] R. Aralikatti *et al.*, “Improving reverberant speech separation with synthetic room impulse responses,” in *Proc. ASRU*, 2021, pp. 900–906.

[17] Colleen Richey, Maria A. Barrios, Zeb Armstrong, Chris Bartels, Horacio Franco, Martin Graciarena, Aaron Lawson, Mahesh Kumar Nandwana, Allen Stauffer, Julien van Hout, Paul Gamble, Jeff Hetherly, Cory Stephenson, and Karl Ni, “Voices obscured in complex environmental settings (voices) corpus,” 2018.

[18] A. Farina, “Advancements in impulse response measurements by sine sweeps,” *Journal of The Audio Engineering Society*, 2007.

[19] Z. Wang *et al.*, “Sequential multi-frame neural beamforming for speech separation and enhancement,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 905–911.

[20] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[21] J. Traer and J. McDermott, “Statistics of natural reverberation enable perceptual separation of sound and space,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 48, 2016.

[22] V. Panayotov *et al.*, “Librispeech: An ASR corpus based on public domain audio books,” in *Proc. ICASSP*, 2015, pp. 5206–5210.

[23] E. Fonseca *et al.*, “FSD50K: an open dataset of human-labeled sound events,” *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 829–852, 2022.

[24] I. Kavalerov *et al.*, “Universal sound separation,” in *Proc. WASPAA*, 2019, pp. 175–179.

[25] Google Inc., “Run google cloud speech ai locally, no internet connection required,” https://cloud.google.com/blog/products/ai-machine-learning/speech-on-device-run-server-quality-speech-ai-locally, 2022, Last accessed: 2024-01-16.

[26] A. Meacham and A. Unruh, “Room impulse response synthesis with device diffraction via image source method and finite element analysis,” *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 2716–2716, 2017.