# Egocentric Auditory Attention Localization in Conversations: Supplementary Material

Fiona Ryan[1,2], Hao Jiang[2], Abhinav Shukla[2], James M. Rehg[1,2], Vamsi Krishna Ithapu[2]
Georgia Institute of Technology[1], Meta Reality Labs Research[2]
{fkryan, rehg}@gatech.edu, {haojiang, ithapu}@meta.com, abhinav.shukla.research@gmail.com

## Appendix

In this appendix, we provide further details about elements of our work. We organize the content as follows:



## A. Visualizations

Visualizations of our model's output heatmaps on videos in the test split of our dataset can be viewed on our project page http://fkryan.github.io/saal, or in the included video file **video_examples.mp4**. Yellow bounding boxes denote the ground truth attended speakers and blue bounding boxes denote people who are speaking but not attended to by the camera wearer. These examples illustrate the complexity of the conversation environments in our dataset; we see that most frames contain multiple people within the FOV, and there are typically multiple people speaking at once. There is lots of head motion as people engage in head-nodding behaviors and look between different people and around the room while listening. Additionally, cases occur where multiple people within the camera wearer's conversation group speak at the same time and are both considered to be attended. Our model is able to determine the target(s) of attention effectively in many of these difficult cases and identifies temporal attention shifts between attended speakers.

These visualizations additionally give insight into failure modes. Our model has difficulty in some cases where attended and non-attended speakers are close together, and sometimes falsely identifies people as attention targets while the camera wearer is speaking and is not attending to any of the visible people. These failure modes reflect the challenging nature of our evaluation dataset and give insight into how future work may improve upon our approach.

## B. Implementation Details

### B.1. Model

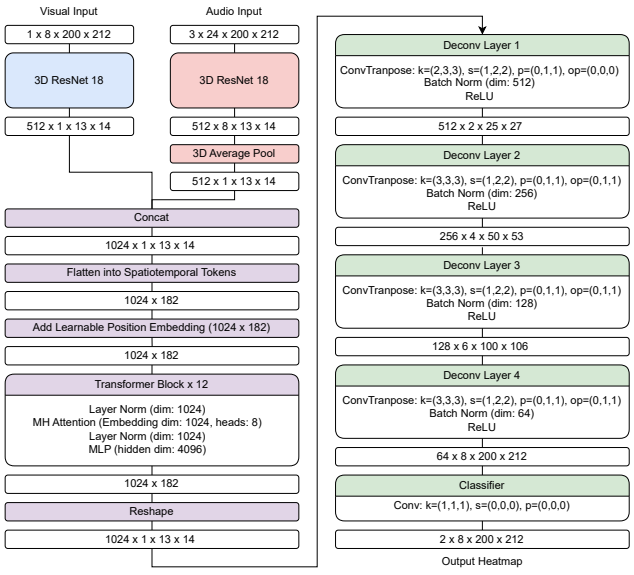


Figure 1. Architecture details. Shapes are denoted as channels × frames × height × width.

We provide architecture details for our model in Figure 1, including the output shapes after each component and individual layer details. Our best model (**Heads** & **Audio Corr + Spectogram**) converges after 5 epochs and takes approximately 1 hour per train epoch on 4 GPU's with batch size 32, using our input clip size of 8 visual frames with temporal stride 3 and 24 audio frames.

### B.2. Input Representations

**Visual Input Representation** In our experiments, we consider 3 visual input representations: **Image** (the raw image),

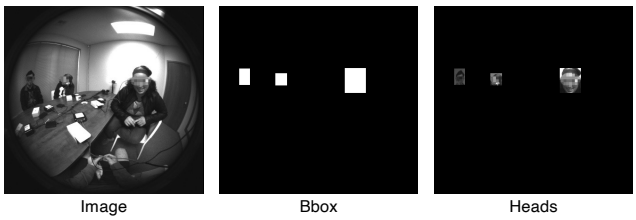


Figure 2. Visual input representations

**Bbox** (a binary map of the bounding boxes), and **Heads** (the cropped heads from the raw image on a black background). Visualizations of these 3 representations are shown for a sample frame in Figure 2.

**Audio Input Representation** In our experiments, we consider 4 audio input representations, where the representation for each audio frame corresponds to the 6-channel audio segment associated with an egocentric video frame. The input representations are **Channel Corr** (the channel correlation features), **Channel Corr + Spectogram** (the channel correlation features concatenated with the real and complex parts of the multichannel spectogram), **ASL$_{\text{real}}$** (ASL output maps from MAVASL [9] trained on the the active speaker labels for our dataset), and **ASL$_{\text{synthetic}}$** (MAVASL ASL maps trained on a synthetic dataset for our microphone array). We visualize these representations for a sample audio frame in Figure 3. The channel correlation features capture spatial audio information by representing the cross correlation between each pair of channels in the microphone array at each time. They are calculated in the same way as Jiang et al. [9], which finds them to be an effective spatial audio input representation for ASL.

We augment these features with the real and complex parts of the spectogram to include finer grained details about the speech signals. To construct the multichannel spectograms we calculate the real and complex parts of the spectogram for each channel individually, $\mathcal{S}_1^R ... \mathcal{S}_6^R$ and $\mathcal{S}_1^C ... \mathcal{S}_6^C$, and concatenate these vertically to form the combined real spectogram $\mathcal{S}^R$ and complex spectogram $S^C$. We then concatenate $\mathcal{S}^R$ and $S^C$ along the channel dimension with the channel correlation features $\mathcal{C}$ to form the $3 \times 200 \times 212$ audio input feature for each frame. We additionally tried concatenating $\mathcal{S}_1^R ... \mathcal{S}_6^R, \mathcal{S}_1^C ... \mathcal{S}_6^C$ along the channel dimension instead of vertically along with $\mathcal{C}$ to form a $13 \times 200 \times 212$ input feature for each frame. We found this slightly reduced performance for our best model (81.68% mAP as opposed to 82.94% mAP).

**ASL$_{\text{synthetic}}$ Training** For the ASL map audio input representations, we include the **ASL$_{\text{synthetic}}$** input representation in addition to **ASL$_{\text{real}}$** to cover both the case where an ASL model is tuned to the dataset and when it is not. Because microphone array setups can vary widely between systems, an existing ASL model that uses multichannel audio may not be immediately applicable or available.

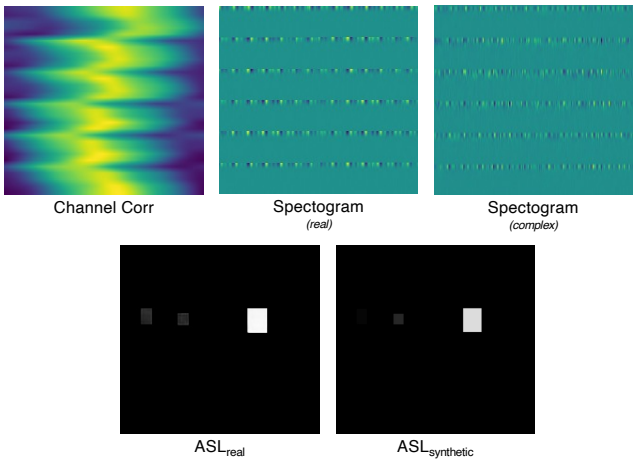


Figure 3. Audio input representations

The synthetic multi-channel audio training data is generated using the VCTK speech dataset [15] and the far field ATFs (audio transfer functions) of the microphone array on the headset. We uniformly sample all the possible sound source directions and for each direction, we randomly select several speech samples from the VCTK dataset and apply the corresponding ATFs to generate the audio signal for each microphone. This is equivalent to putting a virtual speech sound in each direction. We use clean speech; we do not introduce noise and room acoustics effects. The image data is generated using a cut-and-paste method. We paste a "speaking" head, randomly selected from the speaking heads in the EasyCom dataset [3], corresponding to the direction of the audio signal, on a black background. We also paste five more "non-speaking" heads from the EasyCom dataset in randomly selected positions in the image. We randomize the size of the heads to simulate people at different distances from the wearer. The ground truth 360-degree voice activity map and the ground truth of the voice activity map in the FOV can be easily generated using the known direction of the speech source and placement of the head. We use a total of 85,277 audio-visual training samples. The end-to-end training converges in 50 epochs with learning rate 1e-4 and the Adam optimizer.

## C. Dataset Details

### C.1. Comparison to Prior Datasets

To our knowledge, there is no existing dataset and accompanying labels that support our Selective Auditory Attention Localization task. EgoCom [13] and EasyCom [3] capture small group conversations with egocentric video and multichannel audio (binaural in the case of EgoCom) and include speech activity labels. However, both focus on single group conversations, with EgoCom containing con-

versations among 3 people and EasyCom containing conversations among groups of 3-6. In these single-group conversation scenarios, there are rarely cases of more than one person speaking nor are there speakers in the background (EasyCom does include background noise played through speakers, but there are not actual, visible people speaking in the background), so determining auditorily attended speakers can be effectively reduced to audiovisual Active Speaker Localization. However, this reduction is not suitable for realistic noisy conversation environments where there are multiple speakers present. Training a model for SAAL on these datasets therefore would not generalize to complex conversation environments, and evaluating a model for SAAL on these datasets cannot reflect the model's ability to identify selective auditory attention among competing speakers. However, competing speaker environments like restaurants and large group social settings are a main target for downstream sound source enhancement applications. In this work, we specifically seek to investigate selective auditory attention in the presence of multiple speakers, where a person must selectively attend to certain speaker(s) and tune out others. We therefore choose to collect and evaluate on a dataset that explicitly captures overlapping speech, multiple simultaneous conversations, and visible background speakers, where ASL alone cannot determine auditory attention.

The AV Diarization & Social Interactions benchmark subset of Ego4D [6] contains a broader array of conversation scenarios captured by egocentric video and, in limited cases, binaural audio. The dataset includes some cases with multiple speakers and background speakers such as grocery stores, outdoor dining areas, office hours, and group board games. However, determining auditory attention labels for such a dataset is subjective and ambiguous, and there are relatively few cases that capture multiple conversations occurring at once. Quantitatively, only 1.9% (65,261 frames) of the Ego4D train dataset contains at least 2 visible speakers. Ego4D does include "Talking to me" labels which identify which speaker(s) are talking to the camera-wearer and may give insight into who the camera wearer is interacting with and who is in the background. However, this label is inherently different than auditory attention, which identifies listening behavior. Within the Ego4D train dataset, only 0.24% (8,444) frames have a visible "talking to me" speaker as well as another visible speaker. While the"talking to me" labels could indicate which speakers are in a conversation group with the camera-wearer, it is clear there are few cases that capture competing speakers where we could potentially determine from the labels which speaker the camera wearer is listening to. Additionally, the Ego4D dataset largely does not contain multichannel audio, which we find to be a critical aspect of our modeling approach. We evaluate our approach on a dataset we design to capture multi-conversation scenarios on a large scale with selective auditory attention labels.

## C.2. Conversation Layouts

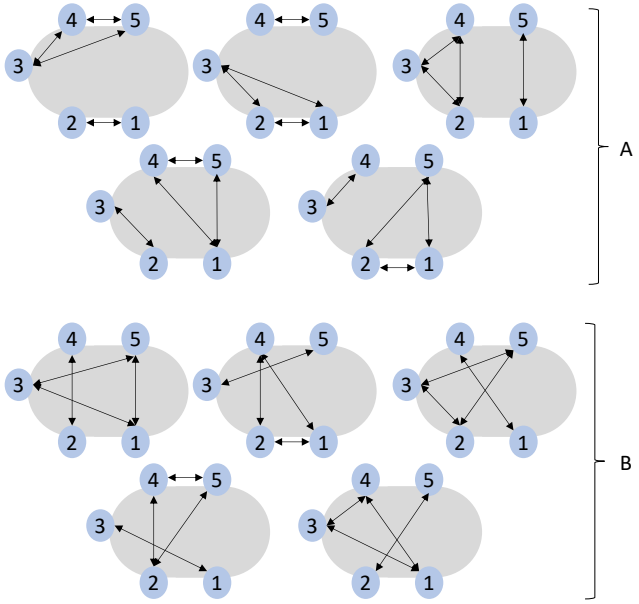


Figure 4. Conversation layouts: Group A represents common scenarios where people converse only with those adjacent or across to them, while Group B represents more challenging cross-talk scenarios. Group A comprises approximately 2/3 of our dataset.

We train and evaluate on a dataset of 5-person scenarios where people converse in 2 separate conversation groups. We illustrate the different conversation group assignments we use in Figure 4. We design the dataset to encompass two kinds of conversational scenarios, which are denoted as Group A and Group B. The conversation layouts in Group A reflect situations where separate social groups converse in close proximity, such as at a coffee shop. In these settings, there is spatial separation between the 2 groups and people converse with those adjacent or directly across from them. The conversation layouts in Group B represent more challenging scenarios where the people in the different conversation groups must talk across each other. These are representative of situations like a large dinner table where multiple conversations occur simultaneously. Group A layouts comprise approximately 2/3 of our dataset and Group B layouts comprise approximately 1/3. In this way, the majority of our dataset is Group A layouts, which are simpler and more likely occur, but the dataset also encompasses challenging cases like those in Group B.

## C.3. Person Tracking, Head Bounding Box, & Active Speaker Labels

To label our auditory attention dataset, we need to track each person throughout the video. We take advantage of

the camera pose output from the Intel SLAM camera for robust tracking. If people are perfectly still and the wearer only rotates their head, we can track each person's location by back-projecting to 3D, rotating, and then projecting to the 2D image. In reality, this does not work because people move and the wearer's head not only rotates but also translates. Our tracking-by-detection method back-projects head detection to the surface of a 2-meter sphere. The head bound box detector is Yolo-v3-tiny [1] trained on images from the Open Images dataset [12]. By using such a representation, the tracking algorithm generates a matching cost matrix using the current object observations and previous position estimations. Using min-cost matching, we extend each target's trajectory frame-by-frame.

We use the wearer voice activity classification network from [9] to determine voice activity for each person, giving us the set of active speakers at each time. Because the headset microphones can easily pick up the sound of the camera wearer speaking, we find this to be a reliable method for determining active speakers. The model is trained synthetically for our headset's microphone array using speech data from VCTK. The near-microphone ATFs are used to generate positive training examples and the far-field ATFs for negative training examples. We manually inspected the tracking, bounding box, and speaker labels to ensure high quality.

### C.4. Unseen Environment Data Subset

To test the generalizability of our model under different visual and acoustic conditions, we collected a small subset of data in a different room, which resembled an open kitchen area. The structure of this subset followed that of the main dataset, with 5 participants conversing in 2 simultaneous conversation subgroups. The participants are different than those included in the main dataset. Conversation layouts from both Group A layouts and Group B layouts were included (see Section C.2), and ground truth labels were constructed in the same manner as in the main dataset. However, only the 3 people in one of the conversation subgroups wore headsets. In total, this subset included 87,977 frames, or 49 minutes of data. The result of 80.43% mAP on this dataset was obtained by running our best model trained on the main dataset on the unseen environment subset with no finetuning. This result shows that our model can generalize to a different environment as well as to cases where people are not wearing glasses.

## D. Additional Baselines

In addition to the baselines described in section 4.2 we provide two further groups of baselines. A full comparison of all competing methods, including those described in the main paper, is shown in Table 1.

| Method | mAP (%) |
|---|---|
| Perfect ASL* | 47.99 |
| CP–I* | 63.55 |
| CP–II* | 51.48 |
| CS–I* | 53.86 |
| CS–II* | 49.47 |
| LS–I* | 27.78 |
| LS–II* | 30.63 |
| FCN+$ASL_{real}$ + CP | 72.33 |
| FCN+$ASL_{synthetic}$ + CP | 74.30 |
| FCN+$ASL_{synthetic}$ + CP + WVA* | 73.32 |
| MAVASL–I | 59.11 |
| MAVASL–II | 75.20 |
| MAVASL–III | 72.90 |
| Ours–Bbox & $ASL_{synthetic}$ | 75.93 |
| Ours–Bbox & $ASL_{real}$ | 74.97 |
| Ours–Bbox & Channel Corr | 80.41 |
| Ours–Bbox & Channel Corr + Spectogram | 80.31 |
| Ours–Image & $ASL_{synthetic}$ | 72.20 |
| Ours–Image & $ASL_{real}$ | 70.04 |
| Ours–Image & Channel Corr | 76.52 |
| Ours–Image & Channel Corr + Spectogram | 76.95 |
| Ours–Heads & $ASL_{synthetic}$ | 76.72 |
| Ours–Heads & $ASL_{real}$ | 77.11 |
| Ours–Heads & Channel Corr | 82.35 |
| **Ours–Heads & Channel Corr + Spectogram** | **82.94** |

Table 1. Comparison results for all methods on the multi-speaker conversation dataset. (*) denotes methods that use inputs that are not given to our model including ground truth active speaker labels, the camera wearer's speaker activity label, and other people's headset audio.

**(1)** Fully convolutional network (FCN) combined with ASL (FCN+ASL): We investigate the extent to which a simple convolutional architecture can solve our task with different types of inputs by adapting the FCN AV Network architecture from MAVASL [9] to predict auditory attention from a concatenation of the raw image and a pre-predicted ASL map (either $\mathbf{ASL_{real}}$ or $\mathbf{ASL_{synthetic}}$), and include an additional center distance map channel to embed the center in which each pixel's value equals the normalized distance to the center of the image, denoted as (CP). We additionally include a variation where the wearer's ground truth speech activity is represented as an extra input channel in which each pixel is 0 or 1 depending on wearer's voice activity label (WVA), which is the label for whether the wearer is speaking or not.

**(2)** Selecting attended speaker based on close-microphone speech activity: We additionally estimate the loudness of each person relative to the camera wearer's position by using the audio energy of each person's worn microphones on their headset and their distance from the camera wearer, as calculated by the SLAM camera. We

calculate loudness for each visible person as $L = \frac{A}{d^2}$ where $A$ is the short-time energy from the person's wearable microphone array (averaged across the channels) for the given audio frame, and $d$ is the distance between the person and the camera-wearer, as estimated by the SLAM camera. We construct 2 baselines that use this loudness measure: LS–I selects the loudest speaker as attended. LS–II selects the loudest speaker as attended unless the wearer is speaking per the ground truth voice activity labels, in which case it selects no people as attended. We note that this baseline uses inputs not given to our model: the audio signals from the headsets of the other participants, the distance as calculated by the SLAM camera, and the ground truth voice activity label for the camera wearer. It is thus not a fair comparison to our model, but we include it to illustrate that in complex conversation environments, assuming that the loudest speaker is attended is not sufficient to solve SAAL.

## E. Real-time Applications

Our problem is motivated by the application of selective sound source enhancement, or developing devices that can enhance certain sound sources while suppressing others. Such a setting demands algorithms that can be run in real-time. While we do not implement our architecture specifically to run in real-time on a mobile device in the scope of this work, our problem formulation reflects this downstream application in two ways: (1) In contrast to the AVA-Active Speaker detection problem formulation which classifies a single head bounding box track at a time, our model reasons about the full scene and all people at once. Not only does this modeling choice reflect the need to reason holistically about the scene to determine SAAL, but this is also conducive to efficient real-time applications. (2) Our architecture is a clip-based video model that runs on a short temporal window at a time (approximately 1 second). Our architecture can be adapted to produce frame-level predictions using this short temporal history. Future work may explore implementing our architecture as part of a real time sound source enhancement system.

## F. Societal Impact

Our work is motivated by developing wearable computing devices that can help people communicate naturally in noisy environments, and can especially assist individuals with hearing difficulties in day to day conversations. Researchers have explored hearing enhancement systems that allow a user to select certain sound sources to enhance using controls like head orientation, eye gaze, and hand controls [2, 4, 5, 7, 8, 10, 11, 14]. By using egocentric cues to automatically determine attended speakers as sound sources to enhance, our work may allow more naturalistic behavior while using such a system. We acknowledge that selective sound source enhancement brings about important privacy considerations. The ability to enhance certain sounds may change notions of conversational privacy in public places, and care must be taken in implementing devices with such capabilities. We note that our work does not apply our algorithm to an end-to-end selective sound source enhancement system.

Additionally, in our dataset design, we avoid constructing scenarios such as intentional, covert eavesdropping that could be used to implement systems with the intent to violate privacy. We instead focus on modeling listening behaviors in multi-group conversation scenarios where participants expect to be heard by others. We believe our work has great potential for the development of devices that assist people with everyday social communication, and can especially help individuals with hearing loss.

## G. Limitations & Future Work

While our work takes an important step towards selective sound source enhancement by addressing modeling selective auditory attention from an egocentric audiovisual perspective for the first time, our approach has limitations. In our dataset design, we constrain our ground truth auditory attention labels to considering speakers within the camera wearer's conversation group as being attended. In reality, this does not encompass attentional dynamics like getting distracted by speakers in the other conversation or sounds occurring elsewhere in the room (which falls under bottom-up auditory attention), or simply zoning out of the conversation. Because auditory attention is covert, sourcing true labels for auditory attention that encompass all such cases is impossible. We believe we take a practical approach to generating objective labels for selective auditory attention, and our labels are appropriate for downstream selective sound source enhancement applications in conversational settings.

There are several opportunities for future work in this direction to improve upon our approach. First, our method may be applied to settings beyond what we collect in our dataset including larger group settings, diverse physical environments, conversing while doing other activities, changing conversation groups over time, and scenarios where people move around the space. Additionally, we constrain our attention modeling to in-FOV cases. While this works well given the wide 180-degree FOV camera we use, future work may explore expanding our model's capabilities to handle cases where people move beyond the wearer's FOV. Further work may also explore explicitly modeling longer-term context, such as conversational turn taking and social groupings over time, to improve predictions. We hope our work will inspire further research in this exciting direction.

## References


[1] Pranav Adarsh, Pratibha Rathi, and Manoj Kumar. Yolo v3-tiny: Object detection and recognition using one stage improved model. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 687–694. IEEE, 2020. 4

[2] Virginia Best, Elin Roverud, Timothy Streeter, Christine R Mason, and Gerald Kidd Jr. The benefit of a visually guided beamformer in a dynamic speech task. *Trends in Hearing*, 21:2331216517722304, 2017. 5

[3] Jacob Donley, Vladimir Tourbabin, Jung-Suk Lee, Mark Broyles, Hao Jiang, Jie Shen, Maja Pantic, Vamsi Krishna Ithapu, and Ravish Mehra. Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments. *arXiv preprint arXiv:2107.04174*, 2021. 2

[4] Antoine Favre-Felix, Carina Graversen, Renskje K Hietkamp, Torsten Dau, and Thomas Lunner. Improving speech intelligibility by hearing aid eye-gaze steering: Conditions with head fixated in a multitalker environment. *Trends in hearing*, 22:2331216518814388, 2018. 5

[5] Michele Geronazzo, Luis S Vieira, Niels Christian Nilsson, Jesper Udesen, and Stefania Serafin. Superhuman hearing-virtual prototyping of artificial hearing: a case study on interactions and acoustic beamforming. *IEEE transactions on visualization and computer graphics*, 26(5):1912–1922, 2020. 5

[6] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 3

[7] Jamie Hart, Dumitru Onceanu, Changuk Sohn, Doug Wightman, and Roel Vertegaal. The attentive hearing aid: Eye selection of auditory sources for hearing impaired users. In *IFIP conference on human-computer interaction*, pages 19–35. Springer, 2009. 5

[8] Ľuboš Hládek, Bernd Porr, Graham Naylor, Thomas Lunner, and W Owen Brimijoin. On the interaction of head and gaze control with acoustic beam width of a simulated beamformer in a two-talker scenario. *Trends in Hearing*, 23:2331216519876795, 2019. 5

[9] Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu. Egocentric deep multi-channel audio-visual active speaker localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10544–10552, 2022. 2, 4

[10] Gerald Kidd Jr. Enhancing auditory selective attention using a visually guided hearing aid. *Journal of Speech, Language, and Hearing Research*, 60(10):3027–3038, 2017. 5

[11] Gerald Kidd Jr, Sylvain Favrot, Joseph G Desloge, Timothy M Streeter, and Christine R Mason. Design and preliminary testing of a visually guided hearing aid. *The Journal of the Acoustical Society of America*, 133(3):EL202–EL207, 2013. 5

[12] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 4

[13] Curtis Northcutt, Shengxin Zha, Steven Lovegrove, and Richard Newcombe. Egocom: A multi-person multi-modal egocentric communications dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[14] Todd Ricketts and Sumit Dhar. Comparison of performance across three directional hearing aids. *Journal of the American Academy of Audiology*, 10(04):180–189, 1999. 5

[15] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017. 2