EARALITY

AI-based AR Platform for Auditory Situational Awareness in 3D

TEAM

Shivank Vatsal
szv5228@psu.edu
(617)-676-8822
Computer Science
Penn State University Park

Aaryan Patel
Computer Science
Penn State University Park

Shambhavi Chowdhury
Chemical Engineering
Penn State University Park

Nakshatra Sharma
Computer Science
Penn State University Park

Problem Description

The WHO corroborated this year that “over 5% of the world’s population – or 430 million people – require rehabilitation to address their disabling hearing loss (432 million adults and 34 million children).” Moreover, “it is estimated that by 2050 over 700 million people – or 1 in every 10 people – will have disabling hearing loss.”

Disabling hearing loss starts at a hearing loss of greater than 35 db, but even as the world awaits the latest and greatest technology which helps improve listening capabilities, it is worthy to note that this does not necessarily come with accurate sound directionality.

Many hearing aids largely work well on a situation-by-situation basis where one can identify sounds of interest. The reality is, sound conveys a lot of information, and doesn’t tap us on the shoulder before reaching our ears. Responding quickly to sudden sounds can be crucial, especially in dangerous situations. Hearing aids and cochlear implants give access to sound, but situational awareness gives access to understanding.

Understanding can be the difference between life and death, and the reason why someone feels present and/or safe in a situation. Understanding can take many forms, one of which is spatial and situational awareness. Individuals in the DHH (Deaf or Hard of Hearing) community have different levels of hearing impairment. Even in situations where deafness impacts only one ear, a lack of awareness with directionality can make a huge difference when it comes to dealing with traffic. Individuals with hearing impairments are fundamentally much more prone to accidents as a driver or a pedestrian.

Some argue our vision, if used correctly, oftentimes enables us to make decisions that hearing impairments might have blinded us from. Still, however, there are countless situations where sound is the only source of information we have. A fire alarm, the dog barking outside, a baby crying in another room and a public announcement are all examples of factors that contribute to situational awareness. These factors enable humans to take action accordingly and avert dangerous situations.

Aside from danger, feeling present in situations is equally as important. WICHE estimates that “there are over 5 million deaf individuals in the United States who need mental health treatment every year. Only about 2% of these deaf individuals receive appropriate treatment for mental illness due to barriers in the effective diagnosis of mental illness. The prevalence of adults with serious mental illnesses (SMI) and children with serious emotional disturbances (SED) is likely greater in the deaf population than in the hearing population, sometimes estimated to be 3 to 5 times greater." While most DHH individuals are excellent lip-readers and can communicate with ASL easily, this requires their direct attention. Large meetings and group gatherings are only one of the many examples where being able to hear multiple people is difficult. Informally, people do not usually signal before they speak, and in professional settings, even if a translator is obtained, rapid translation of mixed conversations is not an easy task.

A spatial awareness application which accounts for real-life scenarios like these can be incredibly empowering for people within the DHH community by giving them their independence. This is exactly what the goal of Earality is.

Project Overview

Proposed Solution

Earality seeks to augment the reality of DHH individuals with spatial and situational awareness. As an AR application, Earality accomplishes this by providing visual cues for sounds that are detected and classified in three dimensions using normalized spherical coordinates. Our application is also polyphonic by nature and responds to moving audio inputs in real time. In other words, Earality emulates the way we humans hear sounds by detecting multiple sounds at once, all from different directions, while simultaneously updating the location of sounds as their directionality changes in the listener's environment.

Because our mission is to empower DHH individuals by giving them their independence, this also means that Earality seeks to be an all-purpose application. This means not requiring users to carry different devices and resources for every sound medium that they need assistance with. Earality aims to support general sound awareness, speech to text, as well as music which is powered by HapticBeat, the concept presented during the idea phase of the challenge.

Use Case and Features

The use cases of Earality are fairly straightforward. This application will be a daily-use service that is ready to provide visual cues for any sounds that are presented in the user's environment. The interface will be minimalistic such that the cues do not interfere with the daily life and actions of users. Behind the scenes, the interface will consist of a spherical canvas around the user upon which color-coded pulses alongside text will share information about sounds in the user’s environment. For sounds that are not coming from the field-of-view of the user, there will be indicators showing whether there is a sound coming from the back left or right. The class of sounds will always be displayed as text in the front-view of the headset and not with the pulse to ensure users know what the sound source is, even if it is coming from behind them. Furthermore, Earality introduces the concept of residue, which is how we show where sounds come from seconds after they have already stopped being heard. When a color-coded pulse is given on the spherical canvas, that area will fade that color out over a certain period of time to ensure the user can see the directionality of the sound if it was not in their field-of-view.

Additionally, there are two types of inputs that are categorized as different from general sounds: speech and public announcements. Earality will store information on the past 10 seconds of sound in the case there is a discernible speech or public announcement detected. For a short time, after speech/PA is detected, the user will be able to opt-in to text generation which will then rewind Earality to the initial detection of the public announcement and display the associated text, or start text generation from the moment itself if it is speech.

Current Work

Thus far, our team has accomplished and is working on the more challenging components of Earality. We are able to depict sound classification and localization in a three-dimensional format, both to demonstrate accuracy as well as a prototype for our AR application. Earality can currently detect up to 14 classes of sounds including speech which are: alarm, crying baby, crash, barking dog, running engine, female scream, female speech, burning fire, footsteps, knocking on door, male scream, male speech, ringing phone and music. Earality is polyphonic for up to 2 sounds, meaning we can confidently detect and classify up to two sounds which are being played simultaneously. Earality can also detect and display the movement of sounds throughout the 3D space which is also crucial. We have tested some instances of the residue concept mentioned in the previous section, and it is still being worked on. Earality can successfully process any audio, primarily wav files, and output the sound localization/detection information. For real-time, there is noticeable latency that we seek to minimize come the MVP phase.

Technology

Application of AI

Earality processes input audio using a modified version of the SELDnet baseline algorithm, which was featured in the 2020 edition of the DCASE (Detection and Classification of Acoustic Scenes and Events) challenge.

The SELDnet (Sound Event Localization and Detection network) described here is a sophisticated architecture that uses Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to perform two critical tasks: detecting the presence of sound events and pinpointing their direction of arrival (DOA). To accomplish this, SELDnet takes in a spectrogram as input - a time-frequency representation of an audio signal that captures frequency content over time.

The input spectrogram is then fed through multiple layers of CNNs that extract valuable features from the data and learn local patterns. Next, the RNN layers come into play, capturing the temporal dependencies in the data and helping the network understand how sound events change over time.

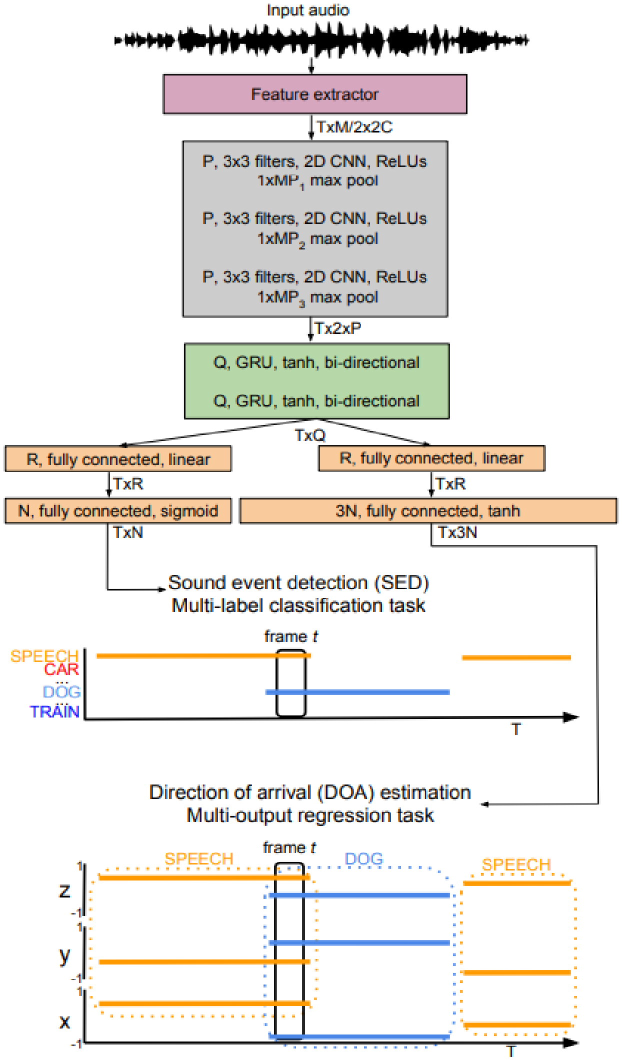

Input audio
Feature extractor
TxM/2x2C
P, 3x3 filters, 2D CNN, ReLUs
1xMP1, max pool
P, 3x3 filters, 2D CNN, ReLUs
1xMP2 max pool
P, 3x3 filters, 2D CNN, ReLUs
1xMP3 max pool
Tx2xP
Q, GRU, tanh, bi-directional
Q, GRU, tanh, bi-directional
TxQ
R, fully connected, linear
TxR
N, fully connected, sigmoid
TxN
R, fully connected, linear
TxR
3N, fully connected, tanh
Tx3N
Sound event detection (SED)
Multi-label classification task
frame t
SPEECH
CAR
DOG
TRAIN
T
Direction of arrival (DOA) estimation
Multi-output regression task
frame t
SPEECH
DOG
SPEECH
z
y
x
1
-1
T

Figure 1: SELDnet

The output of the RNN layers divides into two parallel branches, each consisting of a dense layer followed by a final output layer. The Sound Event Detection (SED) branch determines the presence or absence of various sound events in the audio signal, while the Direction of Arrival (DOA) branch estimates the direction of arrival for the detected sound events. The output layers for each branch use different activation functions; the SED branch utilizes a sigmoid activation function that generates probabilities for the presence of each sound event, and the DOA branch uses a linear activation function that provides continuous values representing the estimated DOAs.

Finally, the outputs from the two branches are combined, resulting in a comprehensive list of detected sound events along with their corresponding DOAs. Overall, the SELDnet architecture is a robust and effective approach to sound event localization and detection.

The modified version of SELDnet employs the same event-independent network, however when polyphonic sounds are to be detected and classified it is thought of as a multiple-track problem. Each track is limited to one event and its corresponding direction-of-arrival (DoA). Instead of relying only on SED and DoA, the network generates three types of output predictions: SED, DoA and event activity detection (EAD). EAD predicts the on-set and off-set times of events more accurately by encompassing feature embedding information from both SED and DoA. This enables the modified version of SELDnet to substantially improve on accuracy.

Data Sources

This featured version of SELDnet from DCASE was pretrained on the TAU-NIGENS Spatial Sound Events 2020 dataset and was retrained with slight adjustments.

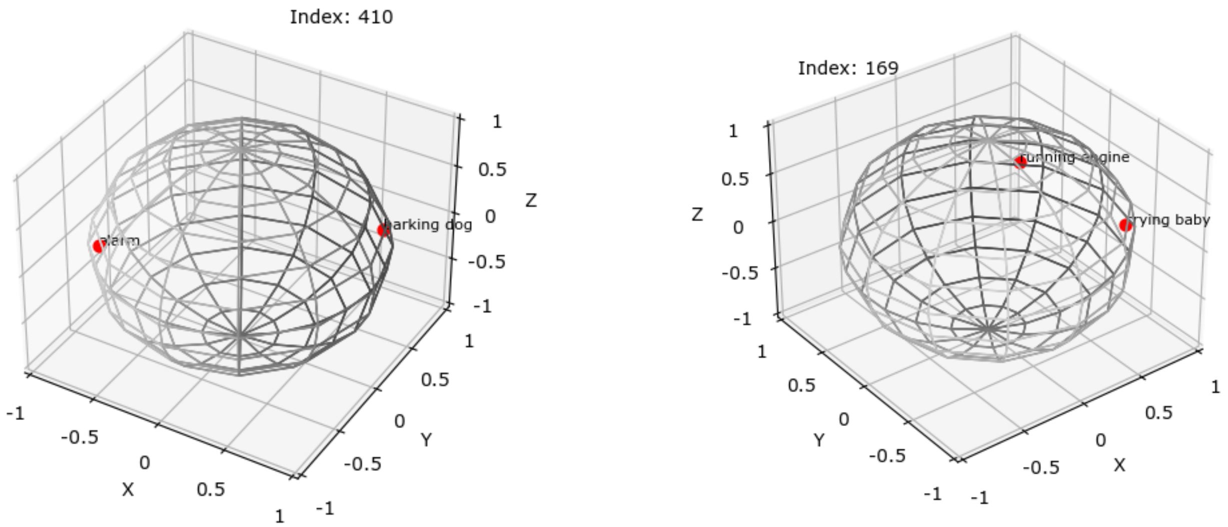
This dataset showcases a wide range of spatial sound-scene recordings, each containing numerous sound events of different categories, integrated into various acoustical spaces, captured from different source directions and distances relative to the recording position. The spatialization of all sound events in the dataset is based on filtering them through authentic spatial room impulse responses (RIRs) obtained from multiple rooms of varying shapes, sizes, and acoustical absorption properties. Moreover, each scene recording is categorized into two different spatial recording formats, either microphone array (MIC) or first-order Ambisonics (FOA). The sound events are spatialized as either stationary sound sources in the room or moving sound sources, where time-variant RIRs are utilized. Each sound event present in the sound scene comes with a trajectory of its direction-of-arrival (DoA) towards the recording point, and a temporal onset and offset time. This provides a comprehensive understanding of each sound event's behavior over time and allows researchers to train robust models.

Earality on Display

Model Performance

In order to demonstrate and test the model performance/accuracy before creating a prototype, we integrated some python scripts with the model which output the localization and classification data in three-dimensions on a spherical plot using matplotlib. We provide some images of the test below.

Figure 2: Model Performance



AR Prototype (Blender)

To prototype our AR application and try to implement some visuals of the residual concept, we decided to use blender. We will shift to ARKit/ARCode when developing the AR application during the MVP phase. Blender gave us the opportunity to exhibit Earality from a first-person perspective. Note that the wireframe sphere is not an actual depiction of what will appear in the AR application, but rather a substitute for the invisible spherical canvas upon which the visual cues will be shown. Of course, there are many other UI/UX considerations to be implemented when the actual application is developed. Here are some images of our prototype:

Figure 3: Third Person View

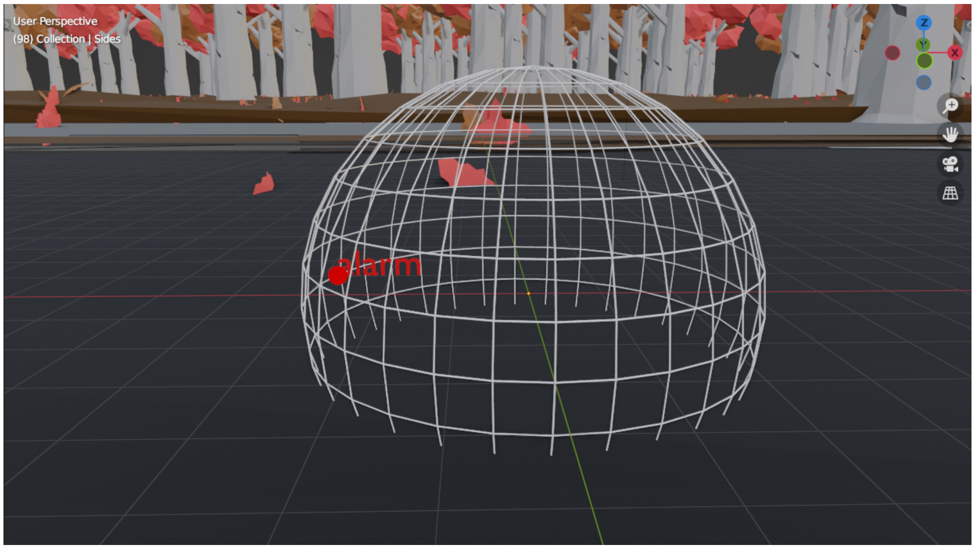
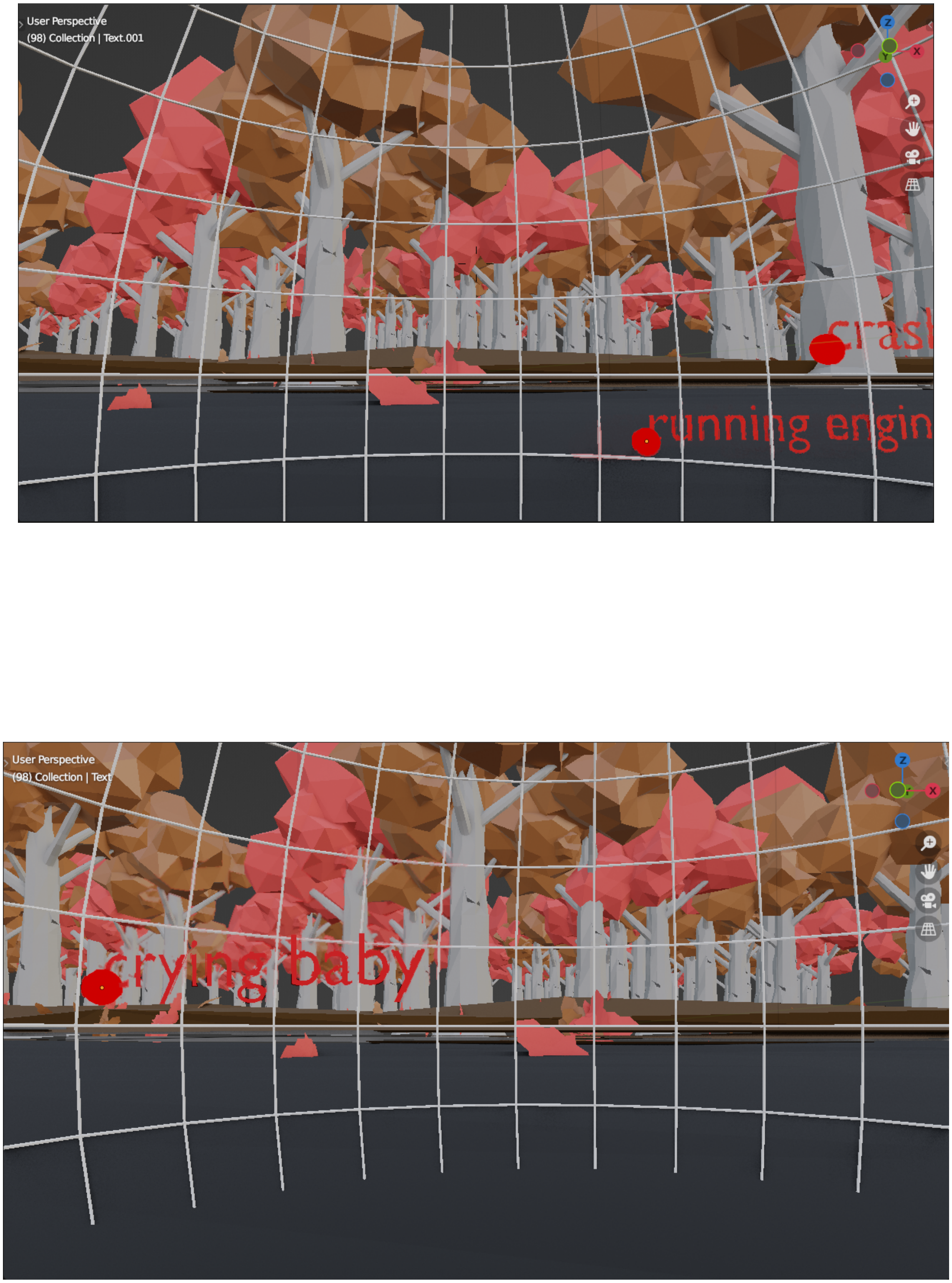


Figure 4: First Person View



Future Scope

Features

There have been many features mentioned across all sections within this documentation. We'd like to clearly establish our goals for the MVP phase here. Ideally we would like to be able to effectively differentiate between speech and public announcements within our model. Once we overcome our goal of minimizing latency and fixing real-time inferencing issues, we will also be able to continuously store the last 10 seconds of data for rewinding and providing TTS (text-to-speech) on discernible speech. If time allows, we also aim to expand the model to accommodate more classes of sounds beyond the 14 we mentioned earlier. This would entail finding or creating new spatial audio data upon which to train the SELDnet. Additionally, we maintain a stretch goal of adding music interpretation (HapticBeat) into Earality to work towards our goal of making an all-purpose application.

Support

Our related projects have received support from the Attack Theatre and Pennsylvania School for the Deaf. They have tested similar technology in the past and fully support projects in this field. They will be our first contacts when it comes to testing Earality in any format.

Development Timeline

Date	Milestone
March 15th	● Wrap up prototype and submit video & documentation
March 19th	● Meet to discuss additional features ● Purchase AR headset
March 20th - April 20th	● Start prototyping AR application with ARKit/ARCore ● Simultaneously work on improving latency for real-time inferencing ● Integrate text-to-speech
April 20th - April 30th	● Consult on UI/UX with users ● Test real-time inference ● Test text-to-speech
May 1st - June 1st	● Implement backend of AR app ● Shift from GCP to AWS ● Get in contact with DHH communities to test early versions of MVP
June 1st - June 15th	● Test MVP with users ● Formulate reports and aggregate feedback
June 16th	● Meet to discuss necessary refinements
June 16th - July 16th	● Make refinements ● Work on stretch goals: ● Expand model to more classes by training other spatial audio data ● Implement HapticBeat for Earality, music detection -> haptics
July 16th - July 23rd	● Final testing and finishing touches
July 23rd - Aug 7th	● Design presentation and practice
Aug 8th	● Final MVP Submission

Team Capabilities

Shivank is a Computer Science and Statistics student at the Penn State University Schreyer Honors College. His past work includes using unsupervised machine learning algorithms to derive insights on the best real-estate investment locations in Texas. His professional experience includes data and software engineering internships at ThoughtSpot, Cisco & Amazon. With additional experience in music production, he brings a variety of audio manipulation skills to the table.

Aaryan is a Computer Science student at the Penn State Schreyer Honors College. He has experience building ML tools in the past. He is also doing research on formal verification of approximate differentially private algorithms where he verifies differentially private-SGD using top of the shelf verifiers. Aaryans past industry experiences include a software development engineering internship at Black Knight.

Shambhavi is a senior majoring in Chemical Engineering with a minor in Engineering Leadership and development. She is enthusiastic to transform the team's ideas into workable enterprises, to reinvent current organizations, and to bring about change in the business environment and the wider world. She intends to oversee the company's documentation and communications to make sure everything is operating as it should. She also completed an R&D internship at Carlisle Construction Materials.

Nakshatra is a senior year student at The Pennsylvania State University majoring in Computer Science. In addition, he is minoring in Cybersecurity and Business Administration. Last summer he interned at a management consulting firm called McKinsey & Company. He was part of the Quantum Black division where he worked with a team of analysts and associates to create a bank statement analysis and turnover estimation tool.

References

Yin Cao, Turab Iqbal, Qiuqiang Kong, Fengyan An, Wenwu Wang, Mark D. Plumbley, "An Improved Event-Independent Network for Polyphonic Sound Event Localization and Detection", submitted for publication

Yin Cao, Turab Iqbal, Qiuqiang Kong, Yue Zhong, Wenwu Wang, Mark D. Plumbley, "Event-Independent Network for Polyphonic Sound Event Localization and Detection", DCASE 2020 Workshop, November 2020

Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. IEEE Journal of Selected Topics in Signal Processing, 13(1):34–48, March 2018.