


Classification and Direction Detection of Ambient Sounds on Microsoft HoloLens to Support Hearing-impaired People

Beauclair Dongmo Ngnintedem¹^a, Eric Mense¹^b, Johannes Rückert¹^c
and Christoph M. Friedrich^{1,2}^d

¹Department of Computer Science, University of Applied Sciences and Arts Dortmund (FHDO), Dortmund, Germany

²Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Essen, Germany
(beauclair.dongmongnintedem001, eric.mense001)@stud.fh-dortmund.de
(christoph.friedrich, johannes.rueckert)@fh-dortmund.de

Keywords: Sensor Substitution, Hearing Loss, Ambient Sound Classification, Sound Source Localization, Mixed Reality.

Abstract: Hearing-impaired people are exposed to greater dangers in everyday life, due to the fact that they are not able to perceive danger and warning signals. This paper addresses this problem by developing an application, that could help by classifying and detecting the direction of ambient sounds using Microsoft HoloLens 2 devices. The developed application implements a client-server architecture. The server-side REST-API supports not only the classification of sounds from audio files via deep-learning methods, but also allows the results of the sound source localization to be saved and read. The sound source localization is performed by a Maix Bit microcontroller with a 6-channel microphone array. For the user integration and interaction with the application, a 3D scene has been designed using Unity and the Mixed Reality Toolkit (MRTK). The implemented application showcases how classification and direction detection of ambient sounds could be done on the Microsoft HoloLens to support hearing-impaired people.

1 INTRODUCTION


According to World Health Organization (WHO), approximately 446 million¹ people suffer from disabling hearing loss (Berra et al., 2020). To help those people, the solution includes the development of new supporting wearable technologies. The following paper describes the development of an application, which can be used for ambient sound source localization and classification on the Microsoft HoloLens 2 to support hearing-impaired people. Therefore, this paper presents the development process of such an application. The resulting application is a proof of concept Mixed Reality (MR) application running on Universal Windows Platforms (UWPs) such as Microsoft HoloLens 2. The aim of the application is twofold: on the one hand, the application records, saves, and classifies sounds from the surrounding environment and presents the results on the 3D scene. On the other


hand, the results of the sound source direction detection, which is performed with a microphone array, are permanently queried and displayed on the scene. Sound classification is performed on a server using a neural network. In Section 2 the fundamentals, including a short overview of sensor substitution, Convolutional Neural Network (CNN) and some related works will be described. After that, the design and implementation of the UWP application will be described in more detail in Section 3. The results of an evaluation will be presented and discussed in Section 4. Section 5 concludes the paper and gives some insights about future works.


2 FUNDAMENTALS AND RELATED WORKS


2.1 Sensor Substitution

Human beings have various senses at birth. These include seeing, hearing, touch, tasting, and smelling. The organs responsible for these senses are the eye, the ear, the skin, the tongue, and the nose. These

^a <https://orcid.org/0000-0002-8291-7124>

^b <https://orcid.org/0000-0003-2748-7958>

^c <https://orcid.org/0000-0002-5038-5899>

^d <https://orcid.org/0000-0001-7906-0038>

¹<https://www.who.int/health-topics/hearing-loss>, accessed 2021-11-26

organs function as sensors, enabling people to perceive their immediate environment using only these senses. A failure or disease of a sensory organ can occur not only at birth, but also during the growing process. This can lead to considerable problems in everyday life, leading to neurodegenerative diseases and depression (Bach-y Rita and Kercel, 2003). Thanks to scientific and technological progress, it is nowadays possible to compensate or treat a damaged or diseased sensory organ by using sensor substitution (Bach-y Rita and Kercel, 2003).

A failure of an organ does not affect the whole sense, but only the part which is responsible for the transmission of the signal to the brain, as it is commonly the case by the retina in the eye, or the *Cochlea* in the ear (Bach-y Rita and Kercel, 2003). Besides, by sensor substitution there is a possibility to replace a failed sensor (sense) with another one, for example *seeing by hearing*, *seeing by feeling and hearing* (Deroy and Auvrey, 2012), *hearing by seeing and reading* or *hearing by feeling or touching* (Cieřla et al., 2019).

To sum up, sensor substitution is about coupling an artificial receptor with the brain via a Human Machine Interface in order to restore a lost sense. Here, in the damaged organ, the signal perceived by the receptor is transmitted directly to the brain, where impressions are created. This is possible due to brain plasticity. The term brain plasticity refers here to the ability of the central nervous system to adapt itself when needed by changing its structural organization and function. This includes neurochemical, synaptic, receptor, and neuronal structural changes (Bach-y Rita and Kercel, 2003).

2.2 Hearing Enhancement with Augmented Reality

According to (Mehra et al., 2020), people with hearing loss may carry more cognitive load to deal with complex acoustic environments. They have to spend more effort in order to fully understand speech in these situations. The treatment of Hearing Loss (HL) is commonly undertaken with hearing aids. This can be described as a multichannel wide dynamic range compression, which enhances the perception of soft sounds while keeping louder sounds within a comfortable range. However, hearing aids present some considerable limitations. Firstly, sufficiently increasing the intelligibility of speech in noisy environments may be challenging. Secondly, the current hearing technologies seem not to always match the user's needs in complex everyday situations. The most advanced devices provide only modest additional benefits, even with additional features. Cur-

rent ear-centered, multi-microphone hearing aid solutions have limited spacing between microphones, and current state-of-the-art beamforming and machine-learning technologies do not allow for the required source separation and sound enhancement. This ear-centric form factor also puts tight constraints on the computation and memory resources available due to limited battery capacity and power budget.

Due to the above-mentioned limitations, the authors stated that there is a real need for new technologies that would help hearing-impaired people by giving them additional support in problematic listening situations. So, the authors suggest that an Augmented Reality (AR) platform, which is described as an interdependent hardware, software, and algorithmic system consisting of a collection of constituent technologies, would give such additional support. According to the authors, an AR platform can be a single device or a collection of interlinked wearable devices working together, which could serve as a frontend to current and future hearing solutions.

To help solve some problems of listening situations like the cocktail-party problem for example, the authors introduced and described one potential configuration of an AR platform named AR hearing-enhancing device, which combines AR glasses, cloud, hearing aids, and input device.

Another way to support people with hearing loss in problematic listening situations would be to give real-time speech-to-text captioning displayed in the AR glasses display system. This has been successfully demonstrated by (Slaney et al., 2020) with a mobile accessibility app designed for the deaf and people with hearing loss, where speech and sound are transcribed to text and displayed on the screen.

To sum up, the authors stated that a combination of multimodal egocentric sensing, a Machine Learning (ML) backbone, and a socially acceptable form factor point toward a future where an AR platform could become the ideal choice to help overcome challenges in compensating for hearing loss.

2.3 Support of Hearing-Impaired People with Mixed Reality

To our knowledge, the present work is the first to deal with classification and detection of direction of ambient sounds on *Microsoft HoloLens* to support hearing-impaired people. However, two projects are related to the sensor substitution, with the focus on ML, using image classification and object detection on the *Microsoft HoloLens*. In the first one, a system for object detection on the *HoloLens* to assist the blind has been developed (Eckert et al., 2018). The approach here

was similar to that of the present work. By this system, using a speech command such as *Scan* or simply the *HoloLens Clicker* enables the user to take a picture of all objects present in the field of view and output the results of object recognition as sound.

The second project deals with the direct use of *Deep Learning* models on the *Microsoft HoloLens 2*². In that work, a MR application has also been implemented, which allows the classification of captured images directly on the *Microsoft HoloLens 2* using its embedded chip. To perform classification of the image, a CNN model trained based on *EfficientNetB0* has been used.

2.4 Ambient Sound Classification

Convolutional Neural Networks (Krizhevsky et al., 2012) refer to a particular type of neural network that has become widely used in *computer vision* for image classification or detection tasks. The success of CNNs in image recognition and classification has led to their adoption for audio data classification (Piczak, 2015).

Like classical neural networks, a CNN consists of inputs, hidden layers, and output layers. In CNNs, a different number of convolutional layers come after the input layer, depending on the network architecture. For example, after the image matrix is passed to the input layer it undergoes a sequence of operations named *convolutions* followed by dimension reduction (*sub sampling*), and the class or category of the image is conveyed through a multilayer perceptron to the output layer. Any number of filters with different filter masks can be used in the convolutional layers. The filters extract features from the images that are learned by the network throughout the training process. In the Sound Event Detection (SED) field, CNNs are used specifically for the classification of so-called Mel scaled spectrograms generated from single audio files. There are also works applying advanced architectures from the *Natural Language Processing* (NLP) field, such as *Bidirectional Encoder Representations from Transformer* (BERT) (Devlin et al., 2019) for the same task. Increasingly, attention-based mechanisms are added on top of CNNs to form CNN-attention hybrids, which help sound classification because they better capture the global context (Kong et al., 2020). These models were also used in the DCASE 2021 challenge on detection and classification of sound events (Nguyen et al., 2021b). (Gong et al., 2021) went further by introducing Audio Spectrogram Transformer (AST),

²<https://github.com/doughtmw/HoloLens2-Machine-Learning>, accessed 2021-11-26

a convolution-free, purely attention-based model inspired by the success of such models in vision tasks, which provides state-of-the-art performance on some popular audio classification datasets.

3 DESIGN AND IMPLEMENTATION

In Figure 1 the architecture of the overall system is shown. At the center of the system is the *Microsoft HoloLens 2* on which the implemented MR application will be deployed. The user interacts with the application through speech commands and by clicking on configured buttons. The implemented MR application accesses an implemented REST interface on the server-side, that classifies sounds recorded using the embedded microphone on the *Microsoft HoloLens 2* device. The reason for outsourcing the classification on a server is that the preprocessing of the recorded clips currently cannot be done locally on the client. Sound source localization is performed by the microcontroller *Maix Bit*. The results from this will be written to the *ESP32 Thing* via the Universal Asynchronous Receiver Transmitter (UART) interface and finally sent by Bluetooth to a connected device. Direct communication between the *ESP32 Thing* and Microsoft HoloLens proved to be complex. Instead, the data is read using the Android application from (Mense, 2020) and sent to the server using the Representational State Transfer (REST)-API.

3.1 Training of a Classification Model

Ambient Sound classification is the first task the developed application should perform. To solve this task, a classification model is needed. For training, we used the Environment Sound Classification Dataset (ESC-50) dataset (Piczak, 2015), which is a collection of 2000 audio clips grouped in 50 categories of everyday sounds. Each clip was recorded with a sample rate of 44.1kHz in mono. The dataset provides 5 splits for comparable cross-validation. First, we trained a CNN similar to the one in (Kumar et al., 2018) from scratch, which only achieved a mean accuracy of 71%, so we instead decided to focus on state-of-the-art models for audio classifications. At the time of writing, the best performing model on the ESC-50 dataset is Audio Spectrogram Transformer (AST)³ (Gong et al., 2021). AST is a convolution-free, purely attention-based model which reaches an

³<https://github.com/YuanGongND/ast>, accessed 2021-11-25

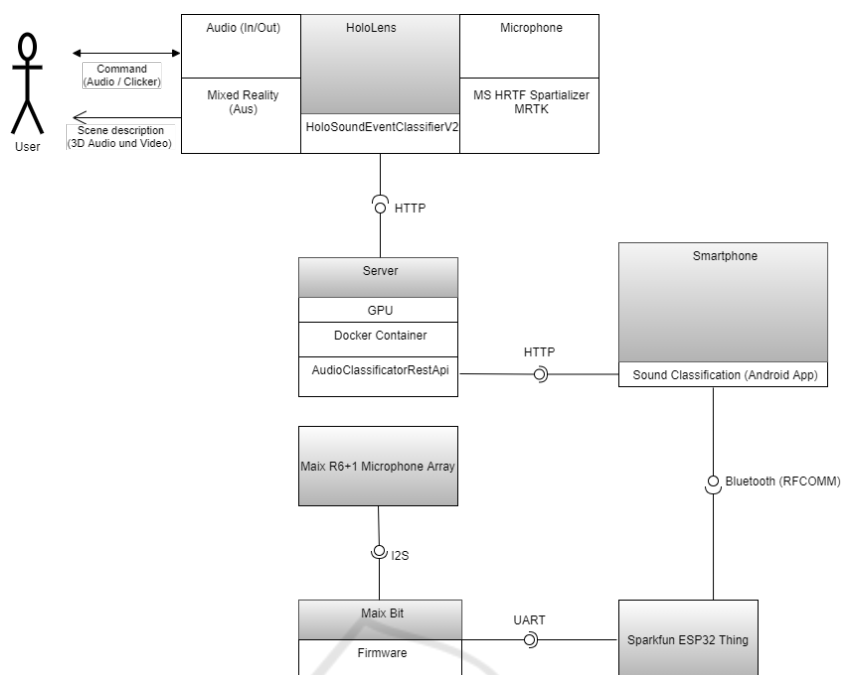


Figure 1: Diagram of the technical architecture of the implemented application.

accuracy of 88.7% (ImageNet pre-training) and 95.6% (ImageNet and AudioSet pre-training) on ESC-50. 5-fold cross-validation training was performed for 25 epochs based on the splits provided by the dataset, and based on a model pre-trained on both ImageNet and AudioSet (Gemmeke et al., 2017), making use of cross-modality transfer learning from images to audio data. Table 1 summarizes the overall achieved accuracies by the trained model from scratch *ESC50Net* and the *AST*-architecture.

Table 1: Achieved accuracy (%) by *ESC50Net* and the *AST*-architecture using *cross-validation*.

Fold	<i>ESC50Net</i> top1 acc. (%)	<i>AST</i> top1 acc. (%)
1	71	94.75
2	67	98.25
3	71	95.00
4	77	96.25
5	69	95.00
Mean	71	95.80

3.2 Server-side Ambient Sound Classification

The ambient sound classification is performed by a *Flask*⁴ (Grinberg, 2018) (version 1.1.2) application

⁴<https://palletsprojects.com/p/flask/> accessed 2021-11-26

that was developed and hosted on the server. This is due to the fact that the supporting libraries for audio conversion have been missing on the HoloLens. The *Flask* application provides a REST Interface, which supports the client-server communication through Application Programming Interface (API) endpoints. For this application, four endpoints are needed:

1. **POST */api/classifier***: Send audio files to the server. The result contains the top 3 predicted classes, including probabilities as JSON data.
2. **POST */api/direction***: Send sound localization results from the Android App to the server.
3. **GET */api/direction***: Get the latest sound localization results from the server as JSON data.
4. **GET */api/amplitude***: Get the maximum amplitude value from the localization results as JSON data.

In order to perform inference of received audio files on the server, the best trained model is loaded into the developed *Flask* application.

3.3 Detection of Direction of Sound

Detecting the direction of sound is made difficult by factors like reflection, polyphony, interference, and non-stationary sound sources (Nguyen et al., 2021a). Modern datasets used for sound event localization and detection, like the one used in DCASE 2021 Task 3 (Politis et al., 2021), use either first-order Ambisonics

or microphone arrays for the recordings. To approach this, we decided to go with a microphone array (see Figure 2) with integrated direction detection.



Figure 2: Portability of the hardware used for sound direction detection.

The sound source localization is performed on a *Maix Bit*⁵ microcontroller with a microphone array. The module includes 7 microphones of which 6 are arranged in a circle around a central one. The microphone module communicates with the Inter-IC-Sound (I2S) protocol⁶. To provide a Bluetooth interface for the localization data, a *Sparkfun ESP32 Thing*⁷ is attached to the *Maix Bit* via UART. The *ESP32 Thing* forwards the UART data to a Bluetooth interface. The above-mentioned hardware together with an Android application for sound classification and direction determination has been developed in (Mense, 2020). An image of the solution is displayed in Figure 2. This application was integrated as a proxy in the developed version of the MR application, as the data from the sound source localization could not be read directly from the hardware using the MR application. Figure 3 shows the results of the application in a test.

3.4 User Interaction

To help the user in the interaction with the application, a 3D scene was designed using the game engine *Unity* and the *MRTK*⁸. The *MRTK* provides a

⁵<https://dl.sipeed.com/shareURL/MAIX/HDK/Sipeed-Maix-Bit/Specifications>, accessed 2021-11-25

⁶https://wiki.sipeed.com/soft/maixpy/en/develop_kit_board/maix_bit.html, accessed 2021-11-25

⁷<https://github.com/sparkfun/ESP32-Thing>, accessed 2021-11-25

⁸<https://docs.microsoft.com/en-us/windows/mixed-reality/develop/unity/mrtk-getting-started>, accessed 2021-11-26

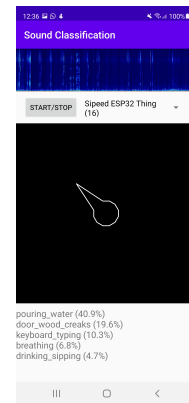


Figure 3: Presentation of the result from direction determination. The direction of the sound is displayed as coming from west-north direction in the middle.

set of components and features that accelerate cross-platform development of MR applications⁹.

In this work, the version 2019.4.8f1 of *Unity* is used in addition to the *MRTK*¹⁰ (version 2.7.0).

The designed user interface is displayed in Figure 4. The most confident class of the top 3 predicted classes is displayed with its probability and inference time at the top. In the center, the direction of the loudest ambient sound is shown by a 3-dimensional arrow. The buttons for starting and stopping the application can also be triggered by voice commands.

While the application is running, an audio clip is recorded and sent to the server every five seconds to determine the class of the sound.

4 EVALUATION AND DISCUSSION

In order to evaluate the performance of the implemented application, several experiments were performed. The experimental setup consisted of:

1. A Bluetooth speaker *PPA401BT-B*,
2. A smartphone *Samsung Galaxy S20 FE*,
3. A laptop *Acer Aspire E5-573G*,
4. The hardware device containing the *Sparkfun ESP32 Thing* and *Maix Bit*, and
5. The module *Maix R6+1 Microphone Array*.

⁹<https://docs.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/>, accessed 2021-11-26

¹⁰<https://docs.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/release-notes/mrtk-27-release-notes?view=mrtkunity-2021-05>, accessed 2021-11-26

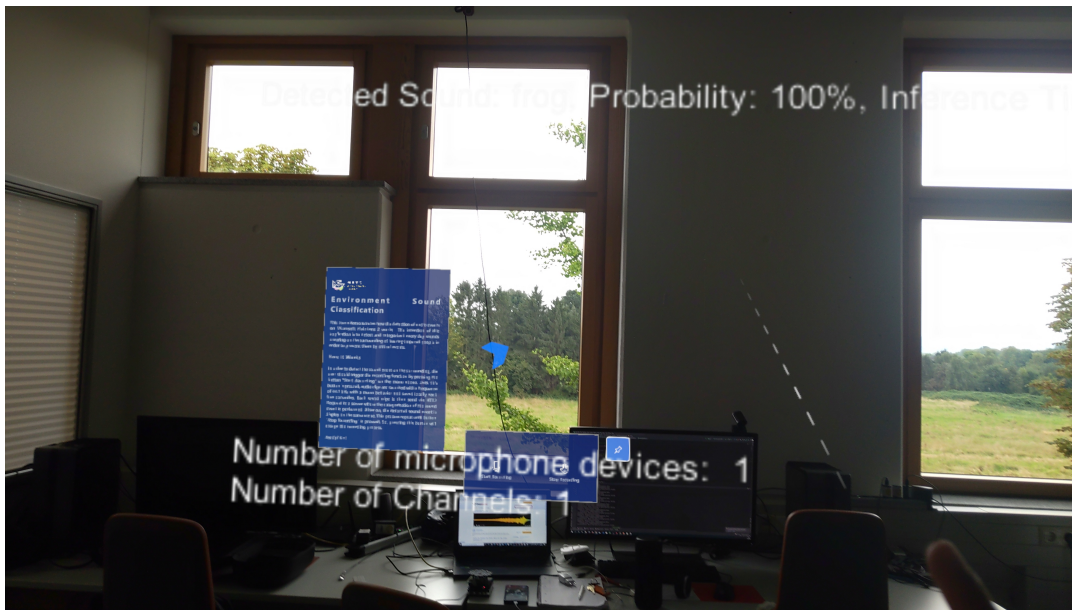


Figure 4: Screenshot of the application running on the Microsoft HoloLens 2.

The speaker used as the sound source was placed around the microphone array for the direction detection. Audio recording was done using the embedded microphone on the devices.

For the experiments, a sound clip containing a mix of several sound classes was played using the speaker, which was placed in a distance $d \leq 1.5$ meter left, right, above and behind the microphone array or *Microsoft HoloLens 2*. Direction detection and sound classification were then performed both using the Android application and the MR application. Classification and direction detection was performed 10 times over the course of 50 seconds, the results for the direction detection are summarized in Table 2. During this experiment, 40 inferences have been conducted on a NVIDIA Titan Xp GPU with a mean preprocessing and inference time of 308 ms.

Table 2: Direction detection results. Mean detection from 10 classifications for each direction.

Direction	Result	Expected	Deviation
Right	82.5°	90°	7.5°
Front	346.5°	0°	13.5°
Left	279°	270°	9°
Behind	174°	180°	6°
Mean			9°

Similar to (Eckert et al., 2018) the voice command recognition for the two configured commands “start recording” and “stop recording” has been evaluated. The results are shown in Table 3.

Table 3: Voice command recognition using the HoloLens.

Command	Recognized	Not recognized	Sum
start rec.	11	20	31
stop rec.	13	2	15

The direction detection works good in a lab scenario, with a mean deviation of 9°. During experiments in the wild, some problems of microphone-arrays could be noticed. For example sound reflections on walls could deviate the direction detection. Inference times are reasonable for a proof-of-concept but overall, the latency of the classification could be improved if the preprocessing and inference could be done on the HoloLens.

For the sound classification, on the other hand, we did not manage to reproduce the AST performance on the development dataset. Possible reasons include the microphone characteristics as well as the fact that the sound was played using a speaker. To improve the microphone performance, several microphones or a 3D microphone (e.g., Ambisonics) could be used in future works.

5 CONCLUSIONS

This paper described the implementation of a proof of concept MR application providing classification and direction detection of ambient sounds on the *Microsoft HoloLens 2* to support hearing-impaired peo-

ple. While the results of the experiments leave room for improvement, especially in terms of the sound classification, we are confident that future work can build on this and improve the performance, for example by using Ambisonics microphones.

REFERENCES

- Bach-y Rita, P. and Kercel, W. S. (2003). Sensory substitution and the human-machine interface. *Trends in cognitive sciences*, 7(12):541–546.
- Berra, S., Pernencar, C., and Almeida, F. (2020). Silent augmented narratives: Inclusive communication with augmented reality for deaf and hard of hearing. *Media & Jornalismo*, 20(36):171–189.
- Cieśla, K., Wolak, T., Lorens, A., Heimler, B., Skarżyński, H., and Amedi, A. (2019). Immediate improvement of speech-in-noise perception through multisensory stimulation via an auditory to tactile sensory substitution. *Restorative neurology and neuroscience*, 37(2):155–166.
- Deroy, O. and Auvrey, M. (2012). Reading the World through the Skin and Ears: A New Perspective on Sensory Substitution. *Frontiers in Psychology*, 3:457.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Eckert, M., Blex, M., and Friedrich, C. M. (2018). Object Detection Featuring 3D Audio Localization for Microsoft HoloLens - A Deep Learning based Sensor Substitution Approach for the Blind. In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies*, pages 555–561. SCITEPRESS - Science and Technology Publications.
- Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio Set: An ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, pages 776–780.
- Gong, Y., Chung, Y.-A., and Glass, J. (2021). AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575.
- Grinberg, M. (2018). *Flask Web Development*. O'Reilly Media, Inc, 2nd edition.
- Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. (2020). Panns: Large-scale pre-trained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12*, page 1097–1105, Red Hook, NY, USA. Curran Associates Inc.
- Kumar, A., Khadkevich, M., and Fügen, C. (2018). Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, pages 326–330. IEEE.
- Mehra, R., Brimijoin, O., Robinson, P., and Lunner, T. (2020). Potential of augmented reality platforms to improve individual hearing aids and to support more ecologically valid research. *Ear and hearing*, 41 Suppl 1:140S–146S.
- Mense, E. (2020). *Sound classification and direction determination with an Android App*. Bachelor thesis, Department of Computer Science, University of Applied Sciences and Arts Dortmund, Germany.
- Nguyen, T. N. T., Watcharasupat, K. N., Lee, Z. J., Nguyen, N. K., Jones, D. L., and Gan, W. S. (2021a). What makes sound event localization and detection difficult? insights from error analysis. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, pages 120–124, Barcelona, Spain.
- Nguyen, T. N. T., Watcharasupat, K. N., Nguyen, N. K., Jones, D. L., and Gan, W. (2021b). DCASE 2021 task 3: Spectrotemporally-aligned features for polyphonic sound event localization and detection. *ArXiv*, abs/2106.15190.
- Piczak, K. J. (2015). ESC: Dataset for environmental sound classification. In Zhou, X., Smeaton, A. F., Tian, Q., Bulterman, D. C., Shen, H. T., Mayer-Patel, K., and Yan, S., editors, *Proceedings of the 23rd ACM international conference on Multimedia - MM '15*, pages 1015–1018, New York, New York, USA. ACM Press.
- Politis, A., Adavanne, S., Krause, D., Deleforge, A., Srivastava, P., and Virtanen, T. (2021). A Dataset of Dynamic Reverberant Sound Scenes with Directional Interferers for Sound Event Localization and Detection. In *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, pages 125–129, Barcelona, Spain.
- Slaney, M., Lyon, R. F., Garcia, R., Kemler, B., Gnegy, C., Wilson, K., Kanevsky, D., Savla, S., and Cerf, V. G. (2020). Auditory measures for the next billion users. *Ear and hearing*, 41 Suppl 1:131S–139S.