Autoprezentacja Piotr Czarnecki 21.10.2019

Autoprezentacja

PROFIL

Inżynier, lider projektów badawczych, zainteresowany analizą i rozpoznawaniem audio, IoT, oraz aplikacji mobilnych.

EDUKACJA

Politechnika Warszawska
2006 - 2007
Dyplom Magistera Inżyniera

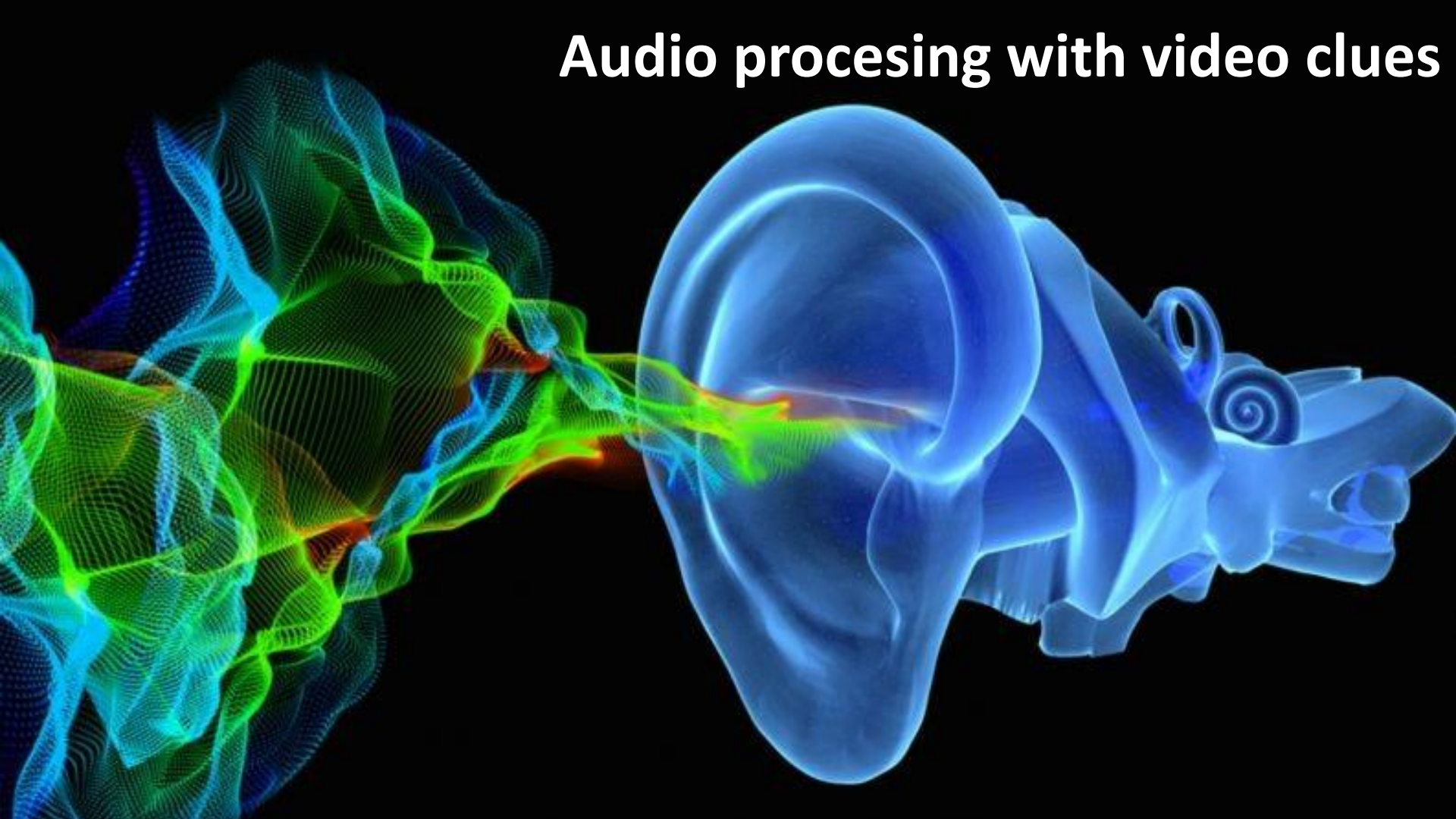
Publikacja: P. Czarnecki, L. Danko and R. Szabatin, "One channel capacitance tomograph with hardware implementation of image reconstruction algorithm," *2009 IEEE International Workshop on Imaging Systems and Techniques*, Shenzhen, 2009, pp. 242-246. doi: 10.1109/IST.2009.5071642
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5071642&isnumber=5071585

Politechnika Warszawska
2002 - 2006
Dyplom Inżyniera

DOŚWIADCZENIE ZAWODOWE

Samsung RD Poland, Inżynier ds produkcji oprogramowania
06.2009–Obecnie
Lider projektów, programista, architect rozwiązań w dziedzinie Audio, IoT, lokalizacja, kontekst użytkownika.
Patent: P. Czarnecki, A.Kedzia, M. Wnorowski, S.Kang, "Devices and methods of providing response message in the devices", United States Patent US9553981B2, 2017
URL: https://patentimages.storage.googleapis.com/dc/34/61/217a2f6d776ef3/US9553981.pdf
- kontrybucja Open Source (Tizen)
- Samsung Best Paper Award – nagroda wewnętrzna na poziomie całej organizacji Samsung-a (Lokalizacja wewnątrz budynkow)

Audio procesing with video clues



Audio procesing with video clues - concept

Human Perception

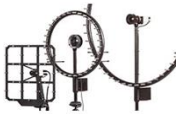
When looking on talking people, human can recognise who the speaker is. Humans, by supplementing hearing with visual clues, can locate more presicely sound source than just by hearing.

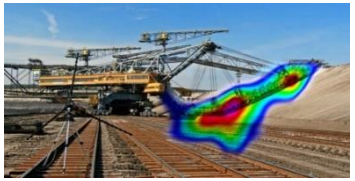
Perception trick

Human system of audio recognition with visual clues can be manipulated. Localization by hearing is less accurte than by audio supplemented with visual clues. Voice can be emitted from different direction, but because strong visual clues, humans can match it with manipulated direction.

HW Limitation

In order to achieve high accuracy in localization sound sources based only on audio, sofisticated microphone arrays are required. That is not feasible to be used for mobiles.





Audio procesing with video clues - concept

Audio-Visual approach

- ✓ **It is feasible to build system that use visual clues** (among audio analysis) **in order to separate and localize sound sources.**
- ✓ **It is feasible to separate sound sources even if single microphone is used (mono recording).**

Web demo on https://www.youtube.com/watch?v=rVQVAPiJWKU

Speech Separation

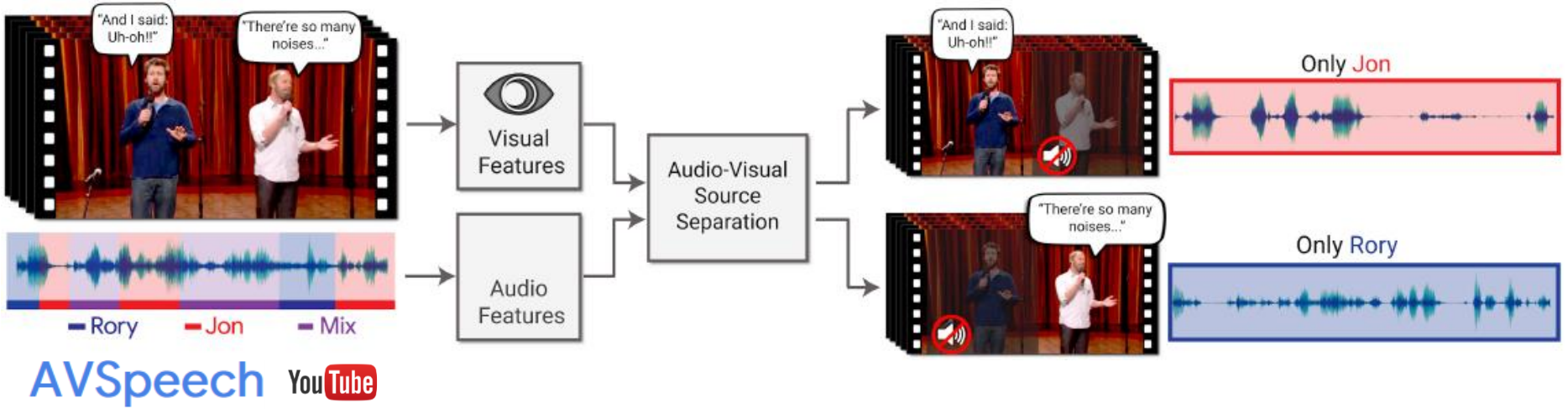
Web available demos

https://www.youtube.com/watch?v=rVQVAPiJWKU

http://www.robots.ox.ac.uk/~vgg/demo/theconversation/.

Speech Separation

Speaker-Independent, language-independent, Audio-Visual Model, for Speech Separation



150k distinct speakers, 4700 hours of video segments (~6.5 months of speech), from a total of 290k YouTube videos, clean speech (one user) segments for training. Web demo on https://www.youtube.com/watch?v=rVQVAPiJWKU. During training, series of feces (not whole video) from two videos was put as input with mixed audio from both videos on audio input. Model was trained to separate each speaker on its output (ground true was known as dataset is based on clean speech segments). Much effort was put to create clear speech of single speaker dataset (AVSpeech).
Similar technology developed by other team http://www.robots.ox.ac.uk/~vgg/demo/theconversation/.

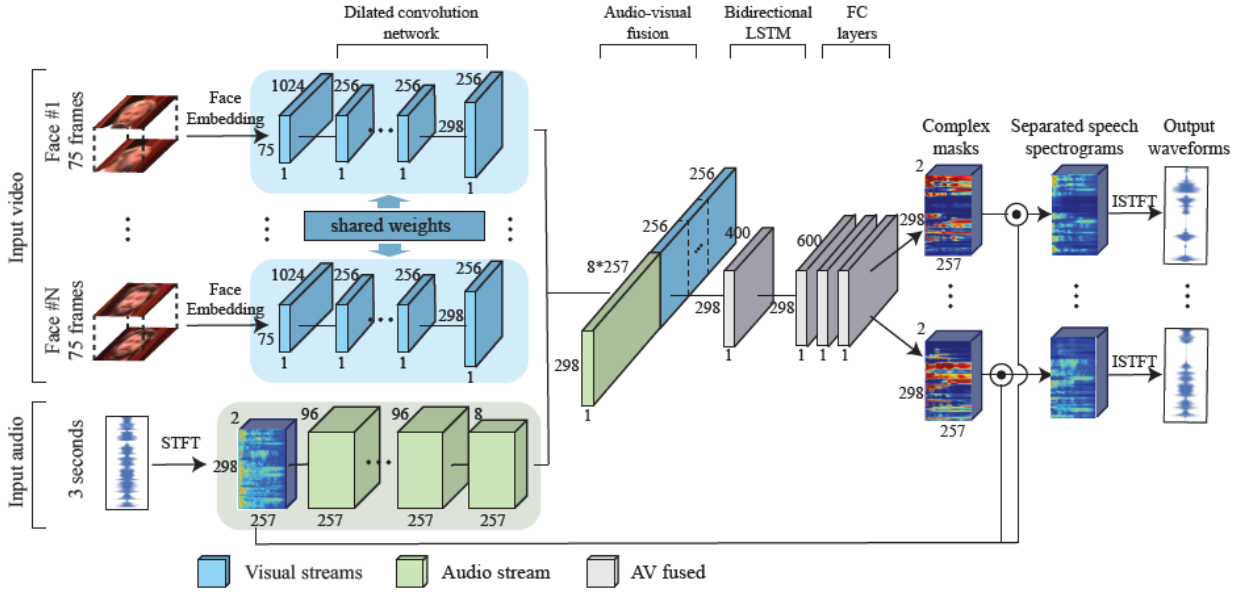
Speech Separation

Core components – baseline model development

The visual streams take as input thumbnails of detected faces, the audio stream takes as input the video's soundtrack, containing a mixture of speech and background noise.

The visual streams extract face embeddings for each thumbnail using a pretrained face recognition model, then learn a visual feature.

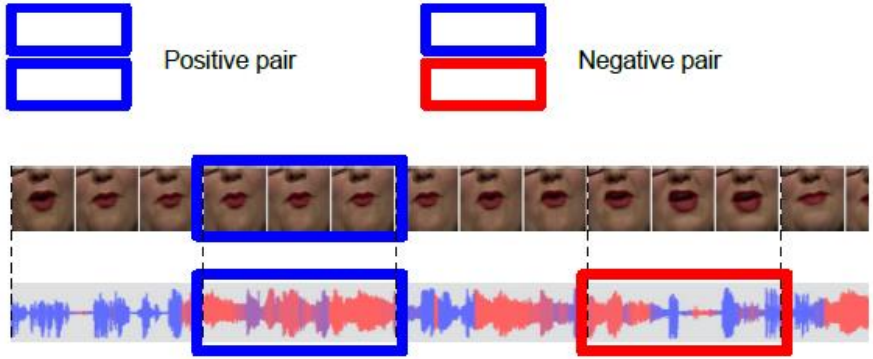
The audio stream first computes the STFT of the input signal to obtain a spectrogram, and then learns an audio representation.

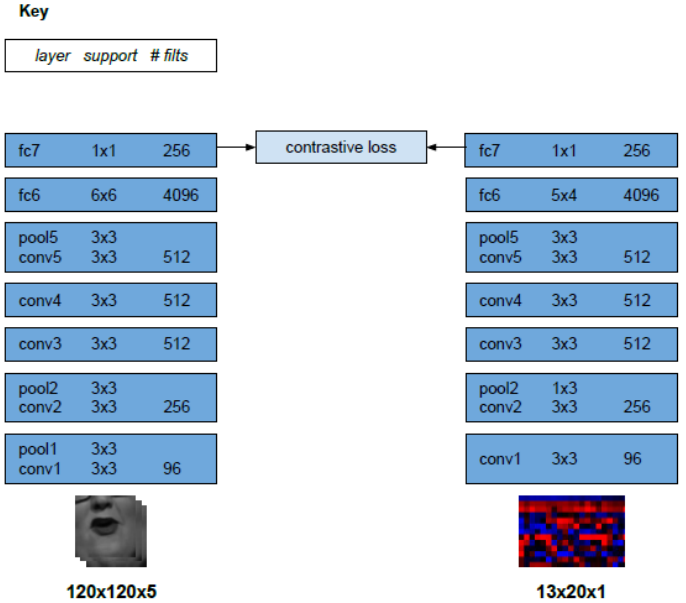


The network outputs a complex spectrogram mask for each speaker, which is multiplied by the noisy input, and converted back to waveforms to obtain an isolated speech signal for each speaker.

Audio – visual synchronization

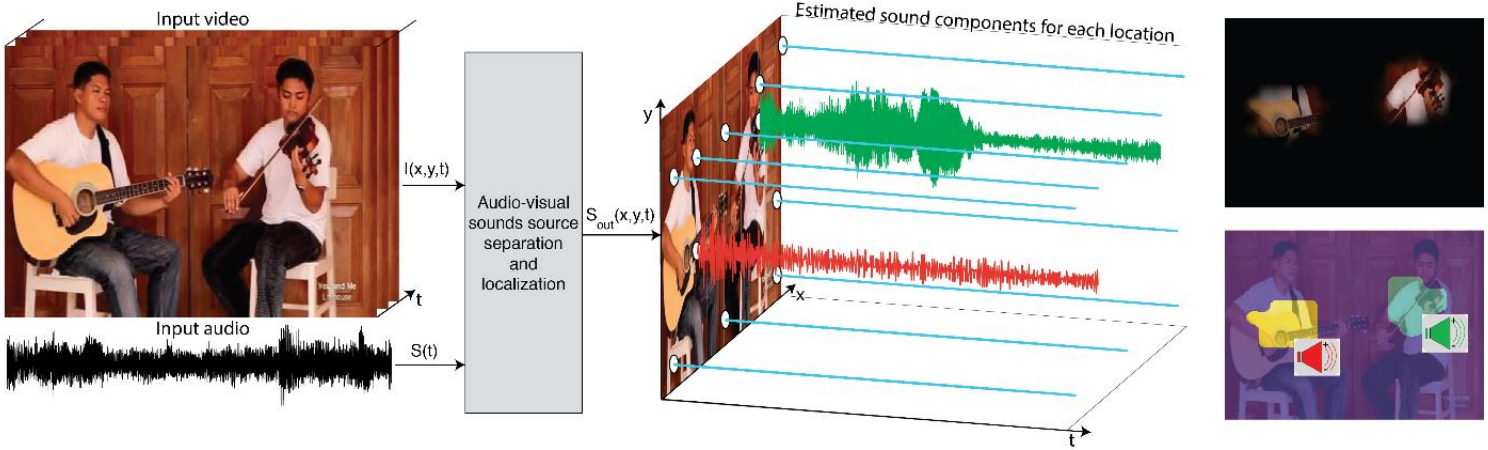
Speaker detector & lip reading





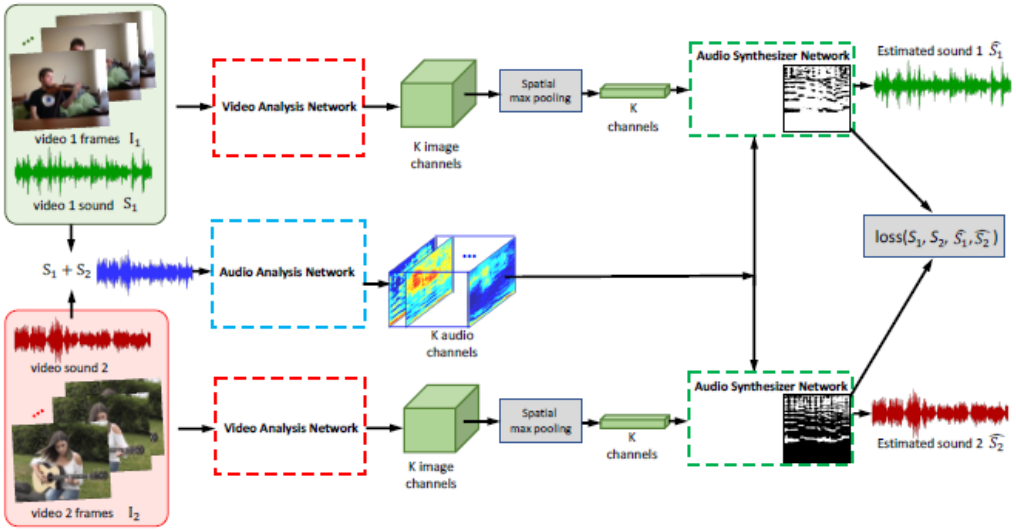
Training is based on synchronization of audio and video. There are positive examples and negative, negative are taken as delayed in time audio in relation to video frames. Dataset contains 800 hours of speech from BBC videos. [http://www.robots.ox.ac.uk/~vgg/research/deep_lip_reading/]

Music instruments separation



PixelPlayer, a system that, by watching large amounts of unlabeled videos, learns to locate image regions which produce sounds and separate the input sounds into a set of components that represents the sound from each pixel. The system is trained with a large number of videos containing people playing instruments in different combinations, including solos and duets. No supervision is provided on what instruments are present on each video, where they are located, or how they sound. During test time, the input to the system is a video showing people playing different instruments, and the mono auditory input. The system performs audio-visual source separation and localization, splitting the input sound signal into N sound channels, each one corresponding to a different instrument category.

Music instruments separation

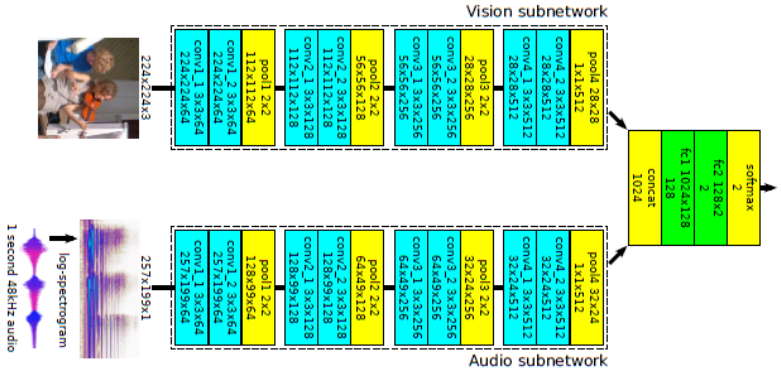


The audio signals from the two videos are added together to generate an input mixture with known constituent source signals. The network is trained to separate the audio source signals conditioned on corresponding video frames, its output is an estimate of both sound signals. There is no assumption about number of sound sources in each video. Moreover, no annotations are provided. The system thus learns to separate individual sources without traditional supervision.

Environmental sounds localization

Sound Sources Localization with visual clues

Rochster model (based on pre-trained audio and video sub networks with attention layer). Audio sub network is pretrained on Audioset, video subnetwork is pretrained on ImegeNet. Attention layer was trained with use of so called weakly labeled data. Weakly labeled data are much less expensive than strongly labeled data (strong labels requires not only audio event class but also exact time when event occurred). Overall our model works only for selected videos of the classes used during training attention layer. At this stage it requires improvements to cover robot scenario, however works fine for PC based scenario (*please click on video attached below*).



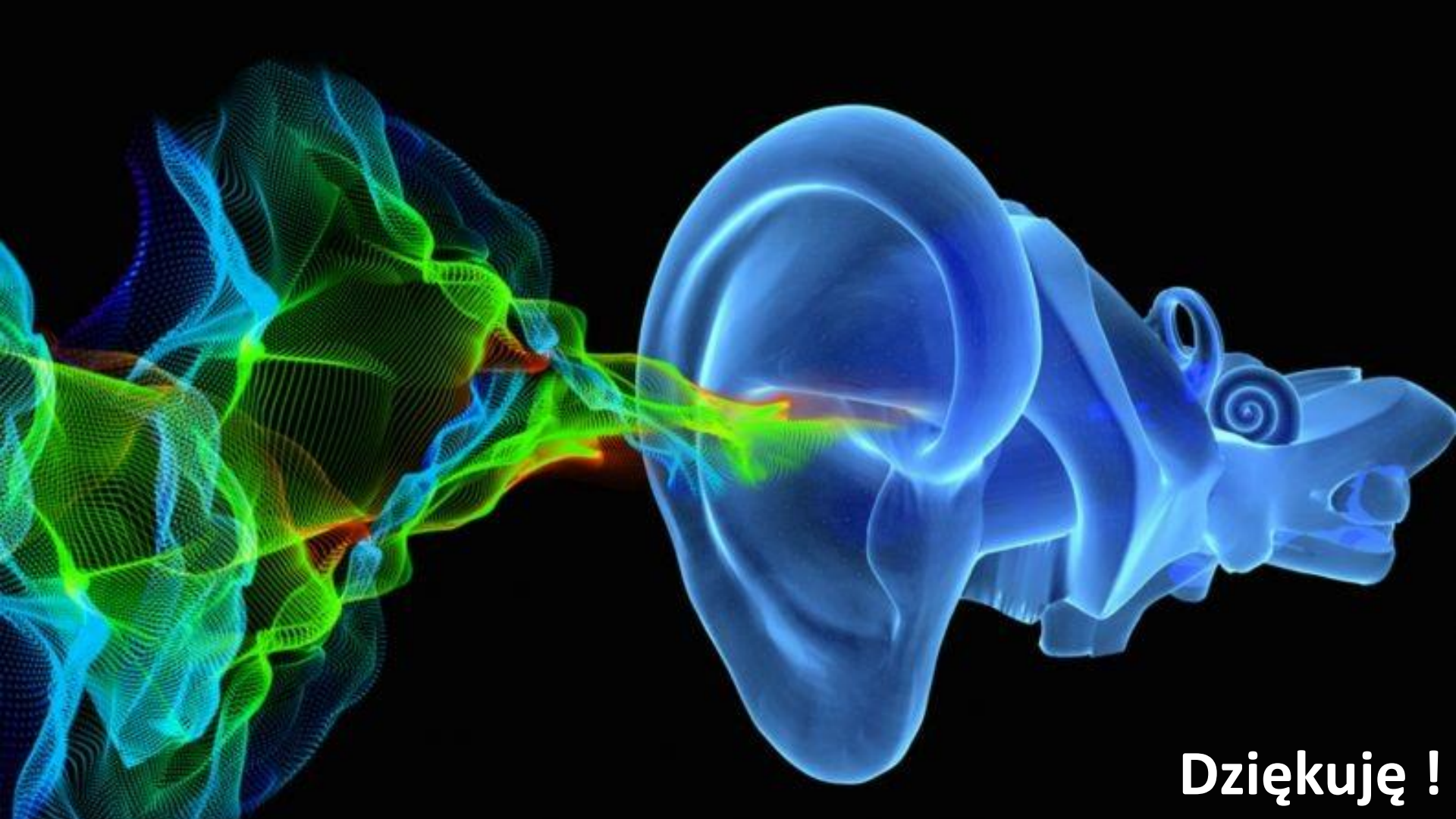
Upcomming plans

Own dataset of objects that sounds

- approach: Youtube high quality preprocessing (object detection, sounds matching)
- goal: audio object editing

Person is speaking, engine will highlight it as it is one of sound source available for editing

Engine will highlight vehicle as its wheels produce sound available for editing (wheels noise can be removed, or its volume lowered)


Dziękuję !