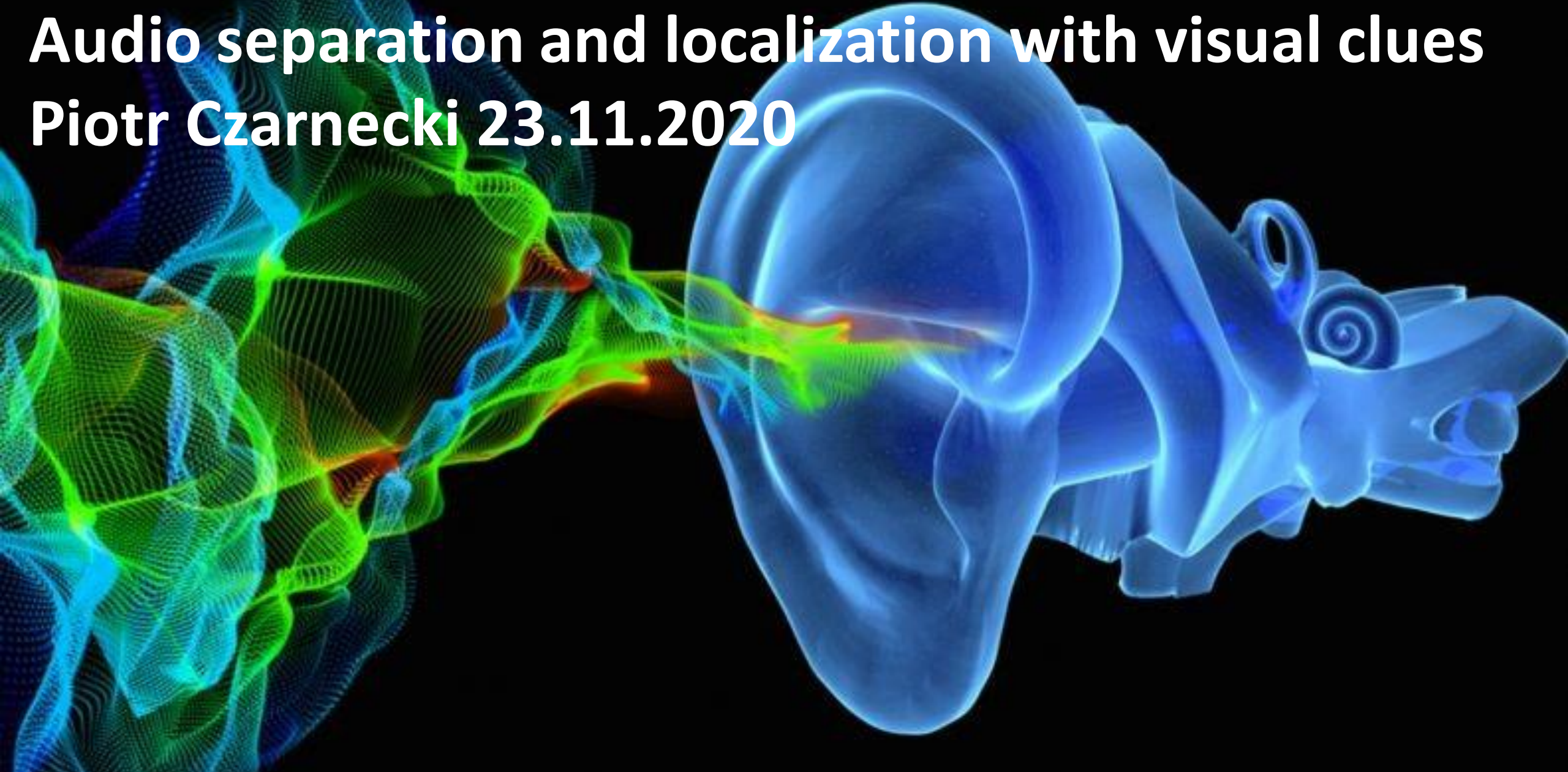


Looking to listen:

Audio separation and localization with visual clues

Piotr Czarnecki 23.11.2020



Sound Sources Localization by audio only

- Binaural unmasking: mostly important to improve perception if source is from different direction than noise, e.g. cocktail party effect.
 - Most important for low frequencies
- Localization clues:
 - Azimuth – Inaural time difference: phase delay below 1000Hz/interaural level differences (shadow effect) above 1500Hz also spectral reflections by torso, shoulders, pinnae.
 - Distance – loss of amplitude, loss of high frequencies, ratio of direct to reverberated signal.
- Selective attention
 - Ability to focus on single voice, however able to redirect attention where is urgently required.

Audio-Video background

Human perception

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



Auditory Illusion

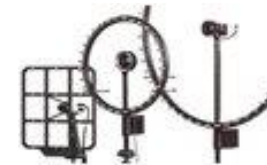
Ventriloquists



audio perception overridden by visual clues. McGurk effect: <https://www.youtube.com/watch?v=2k8fHR9jKVM>

Audio-only HW

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



Audio-Video problem

- Wydaje się że dobrze działające przetwarzanie binaural samo w sobie daje bardzo dużo możliwości, więc..

....do czego jeszcze video. -> motywacja jest taka że w wielu aplikacjach nagrania są mono, np.. mowa w nagraniach stereo/również 5.1 zwykle jest mono (w obu kanałach takie samo), lokalizowane są jedynie 'sound efekty'.

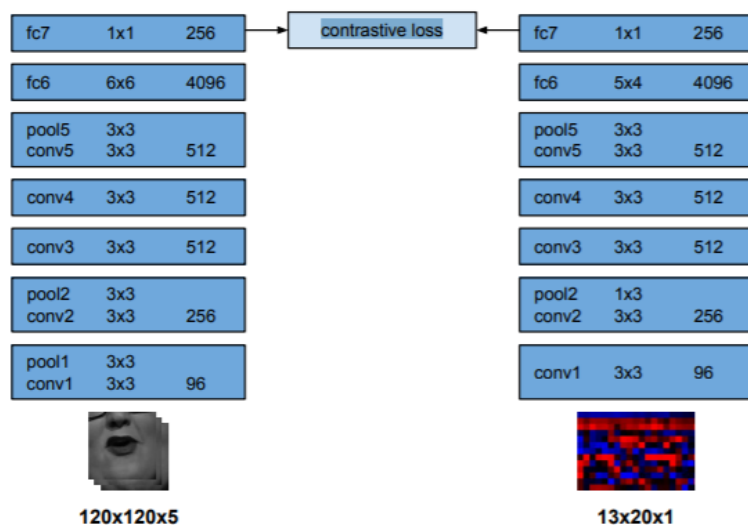
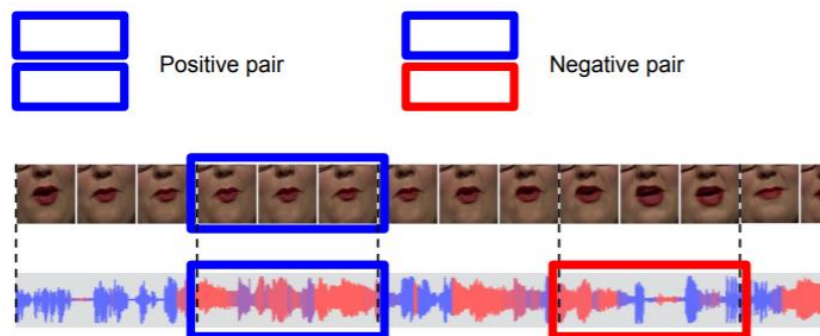
- Generalny problem w technice: jak nagrać dźwięk i jak go zreprodukować żeby podczas odtwarzania był taki jakby człowiek był na miejscu.
- A więc problem to mając mono audio, zlokalizować i odseparować źródła

Audio-Video processing

- Video: 30 FPS -> 1 frame every 33.3ms
- Audio: 48 kHz -> 1 video frame takes 1600 samples, processing could be e.g. 16 frames each 200 samples with 100 overlap.
 - Other are: 16kHz (as voice is up to 8kHz. 3 kHz in telecom)
 - STFT/mel features
- Main approaches:
 - Synchronization – regression but usually binary classification
 - Correspondence (cooccurrence) – binary classification
 - Separation

Main approaches

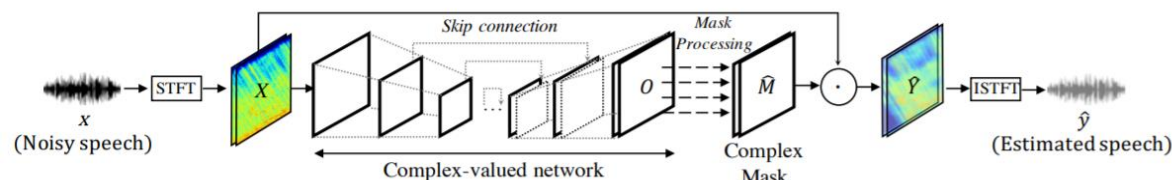
classification



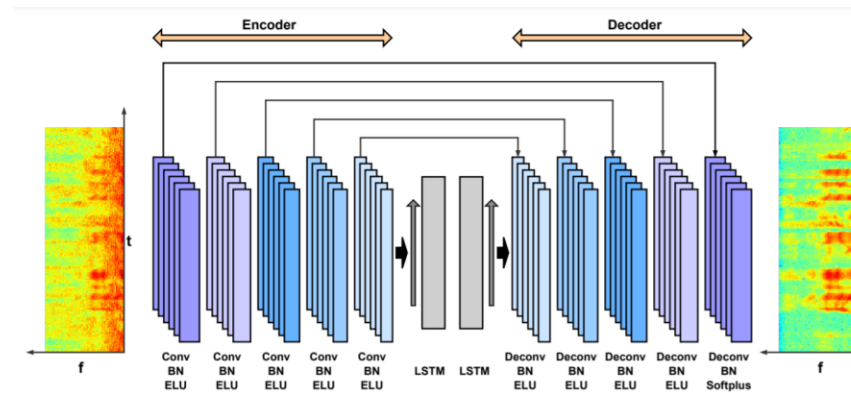
Out of time: automated lip sync in the wild, University of Oxford, ACCV2017

Separation Audio-Only

Deep Complex U-Net



PHASE-AWARE SPEECH ENHANCEMENT WITH DEEP COMPLEX U-NET, Seoul National University, Clova AI Research, NAVER Corp., 2019



A Convolutional RNN for Real-Time Speech Enhancement, The Ohio State University, 2018

Audio-Video existing solutions

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 voices even if single microphone is used (mono recording).

It is feasible to separate voice of same person (same speech) mixed with delayed copy

All above is feasible, however not robust and real-time, also difficult to compare solutions as no standard benchmark (no pretrained models, no exact testset)

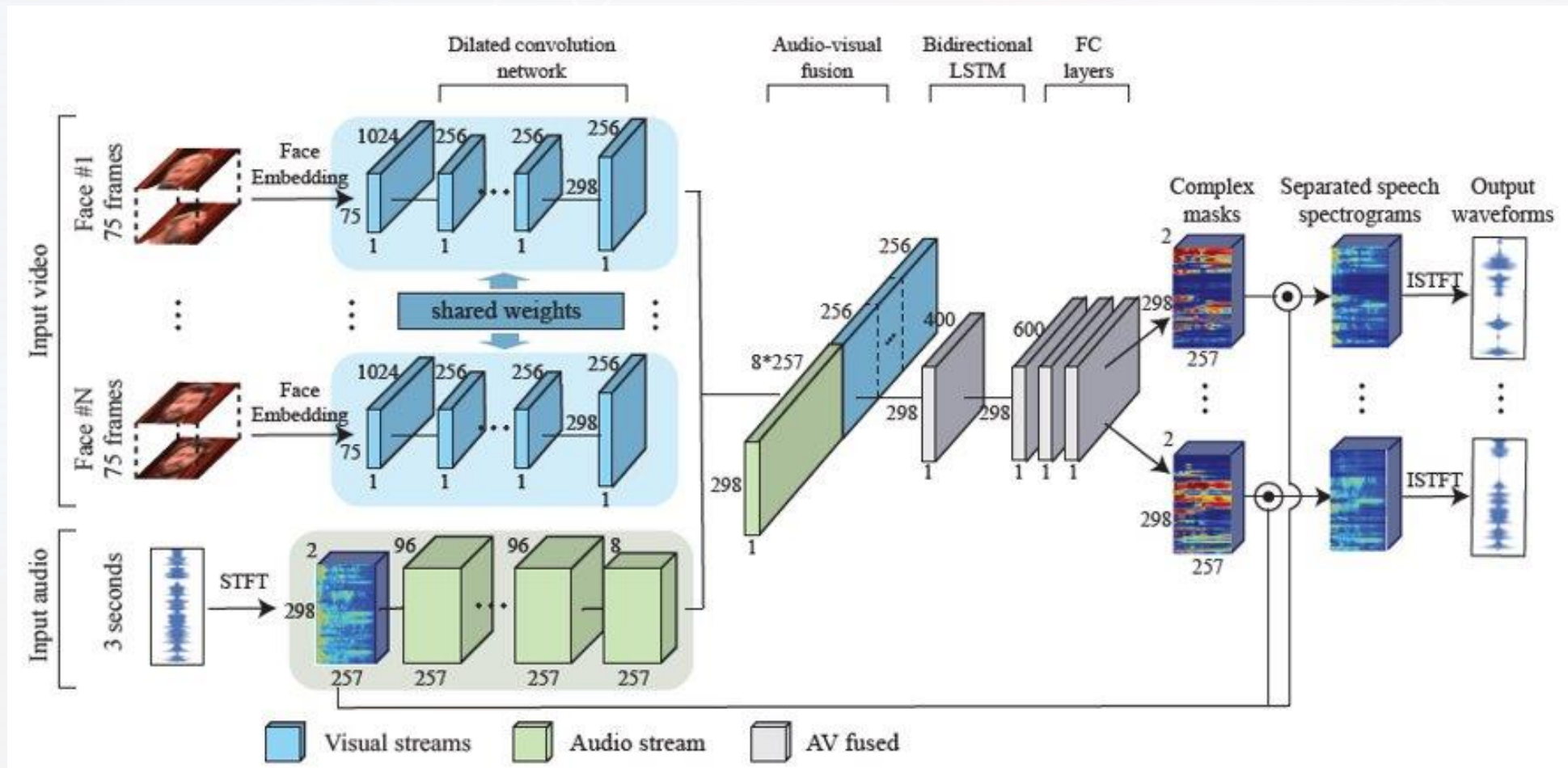


<https://www.youtube.com/watch?v=rVQVAPiJWKU>



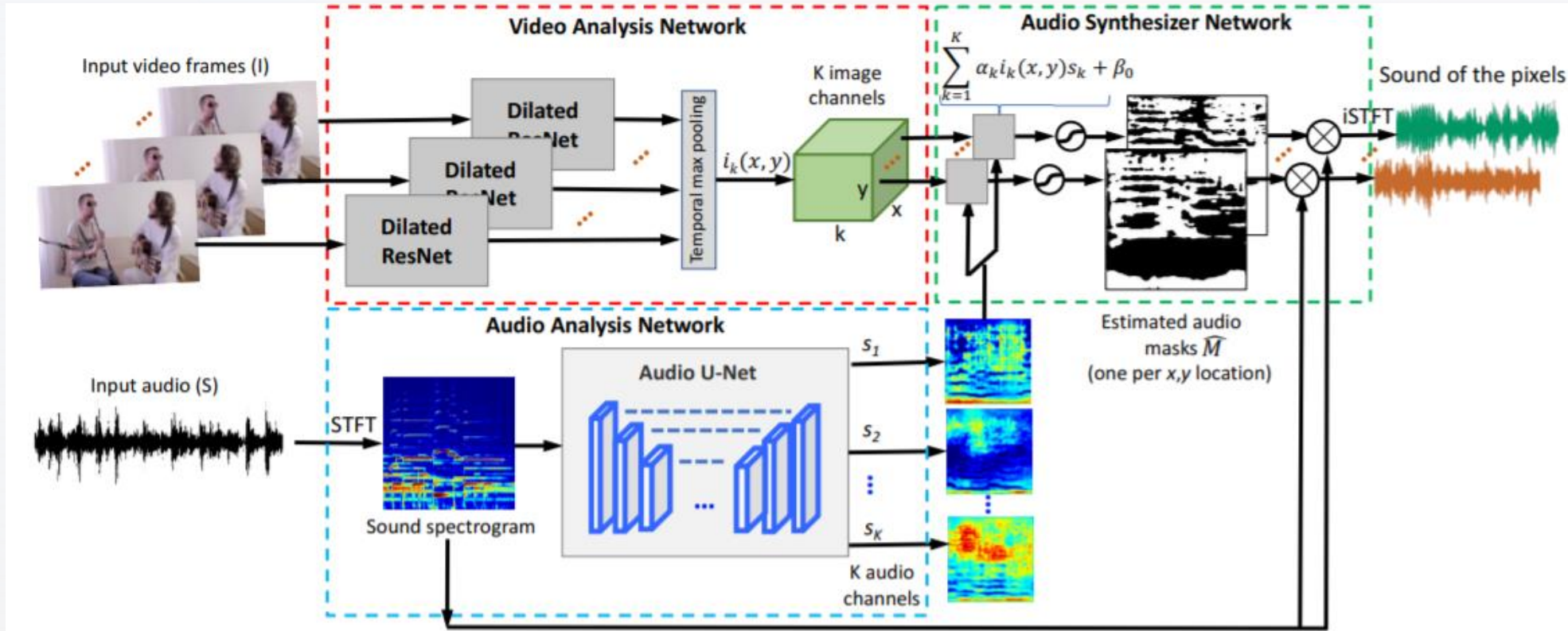
<http://www.robots.ox.ac.uk/~vgg/demo/theconversation/>

Model – late fusion/face detector (2018)



Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation, Google Research and The Hebrew University of Jerusalem, Israel, SIGGRAPH

Model – no object detector (2018)



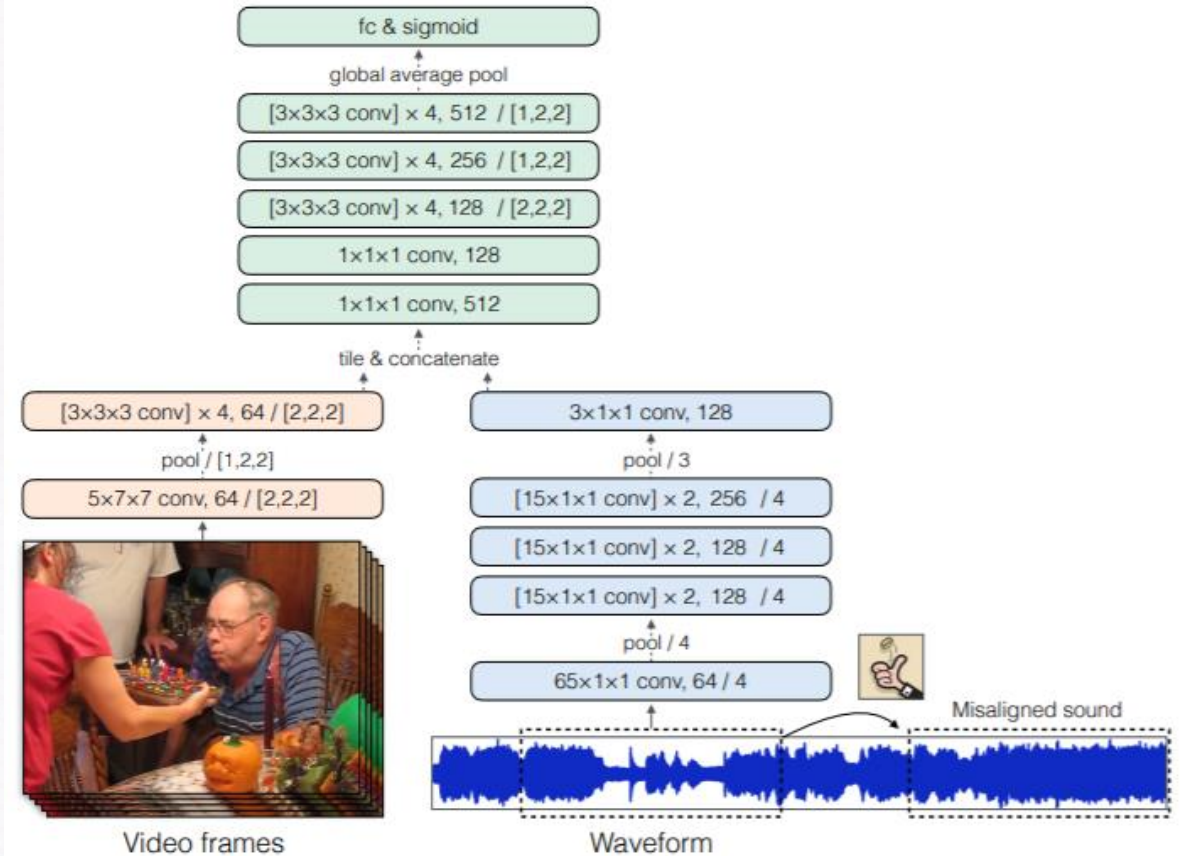
For an input video of size $T \times H \times W \times 3$, the ResNet model extracts per-frame features with size $T \times (H/16) \times (W/16) \times K$. After temporal pooling and sigmoid activation, we obtain a visual feature $i_k(x, y)$ for each pixel with size K .

The Sound of Pixels, Massachusetts Institute of Technology, MIT-IBM Watson AI Lab, Columbia University, CVPR

Model – early fusion, no detector (2018)

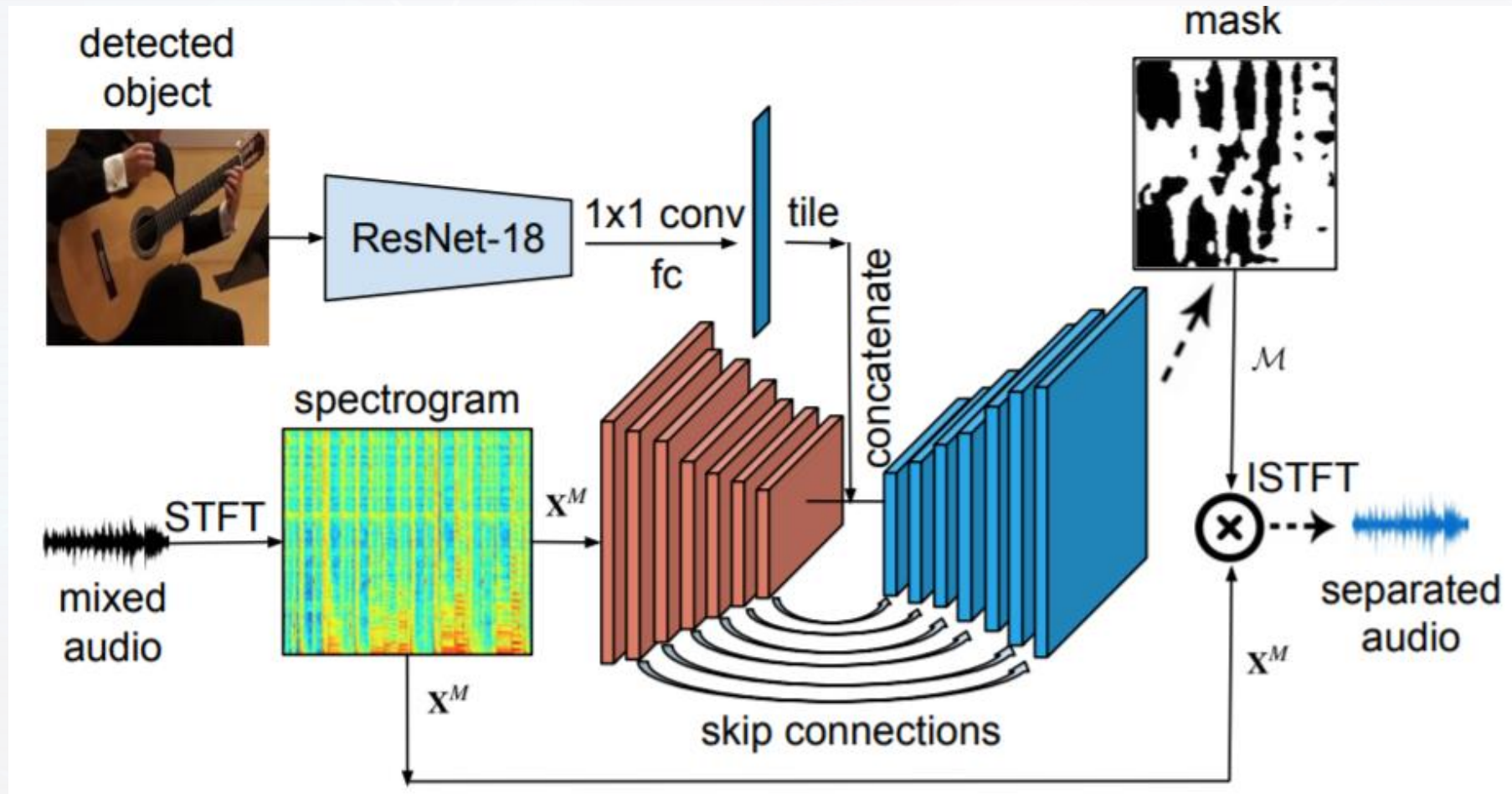


On/off-screen audio-visual source separation



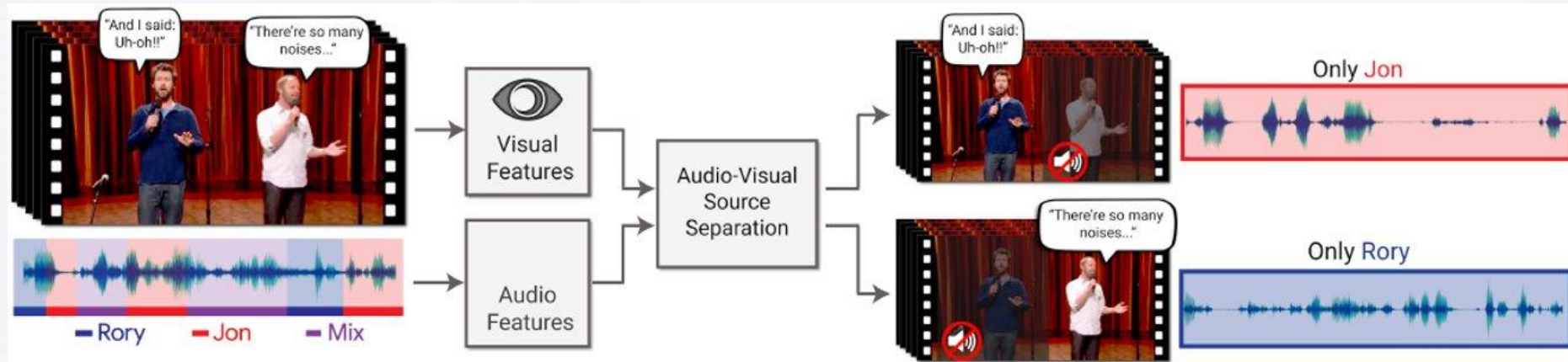
Audio-Visual Scene Analysis with Self-Supervised Multisensory Features, UC Berkeley, CVPR

Model – object detector, U-Net (2019)



Co-Separating Sounds of Visual Objects, UT Austin and Facebook AI Research, ICCV

Dataset - AVSpeech



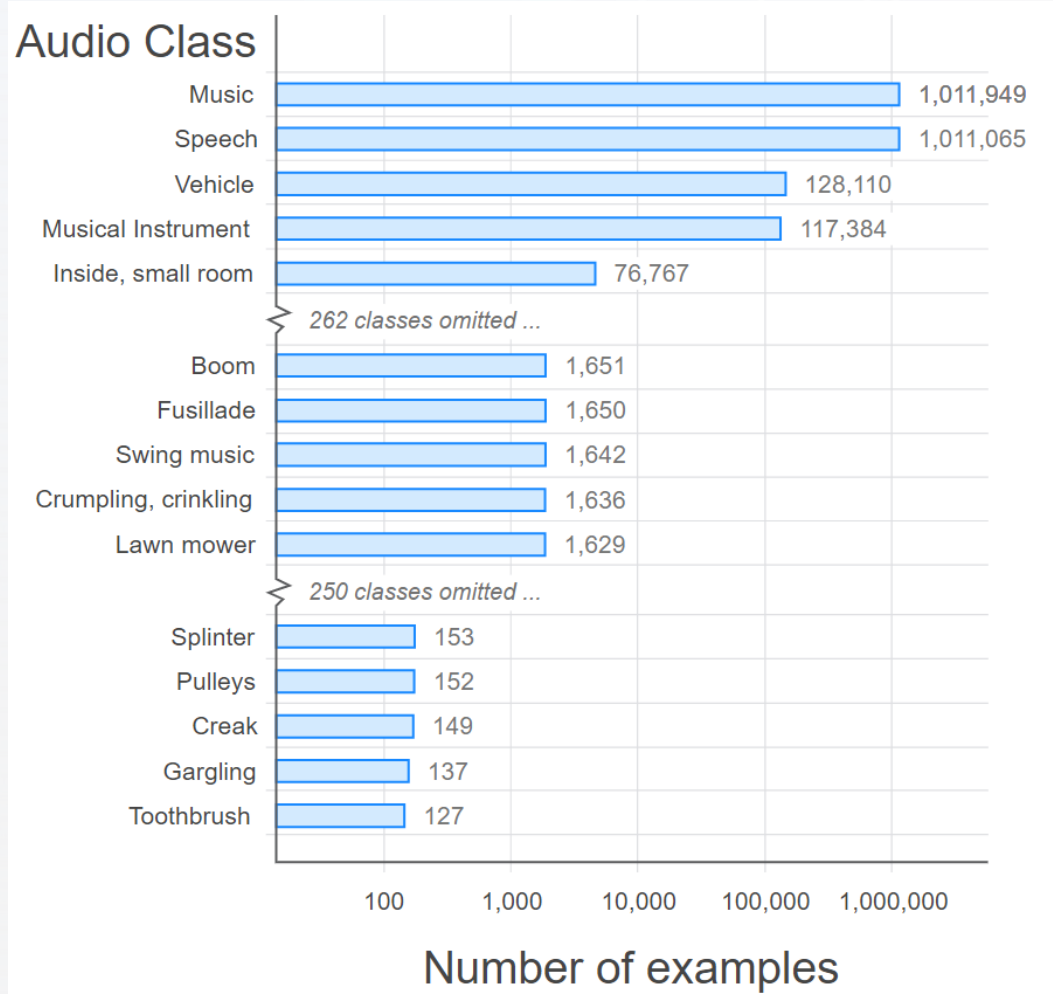
AVSpeech

YouTube

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/>.

Dataset - Audioset/YouTube-8M



8M

Dataset Explore Download Workshop About

YouTube-8M Segments Dataset

The YouTube-8M Segments dataset is an extension of the YouTube-8M dataset with human-verified segment annotations. In addition to annotating videos, we would like to temporally localize the entities in the videos, i.e., find out when the entities occur.

We collected human-verified labels on about 237K segments on 1000 classes from the validation set of the YouTube-8M dataset. Each video will again come with time-localized frame-level features so classifier predictions can be made at segment-level granularity. We encourage researchers to leverage the large amount of noisy video-level labels in the training set to train models for temporal localization.

We are organizing a [Kaggle Challenge](#) and [The 3rd Workshop on YouTube-8M Large-Scale Video Understanding](#) at ICCV 2019.

237K
Human-verified
Segment Labels

1000
Classes

5.0
Avg. Segments /
Video

In addition to annotating the topical entity of the full-video, we want to understand when the entity occurs in videos. Given a 5-second segment and a query class, our human raters are asked to verify whether the entity is identified within the segment. To speed up the annotation process, our human raters do not report presence or absence of non-query classes.

(embeddings only, no raw A-V)

trend

- Model

- U-Net
- RNN (GRU/LSTM/bi-LSTM)
- transformer (2020 CVPR: Listen to Look: Action Recognition by Previewing Audio, The University of Texas at Austin, Facebook AI Research)
- Wavenet (Generative Model for Raw Audio, 2016, 2018, 2019)

- Dataset

- AVSpeech
- Audioset
- Youtube8M (embeddings only)

Our implementation demo run

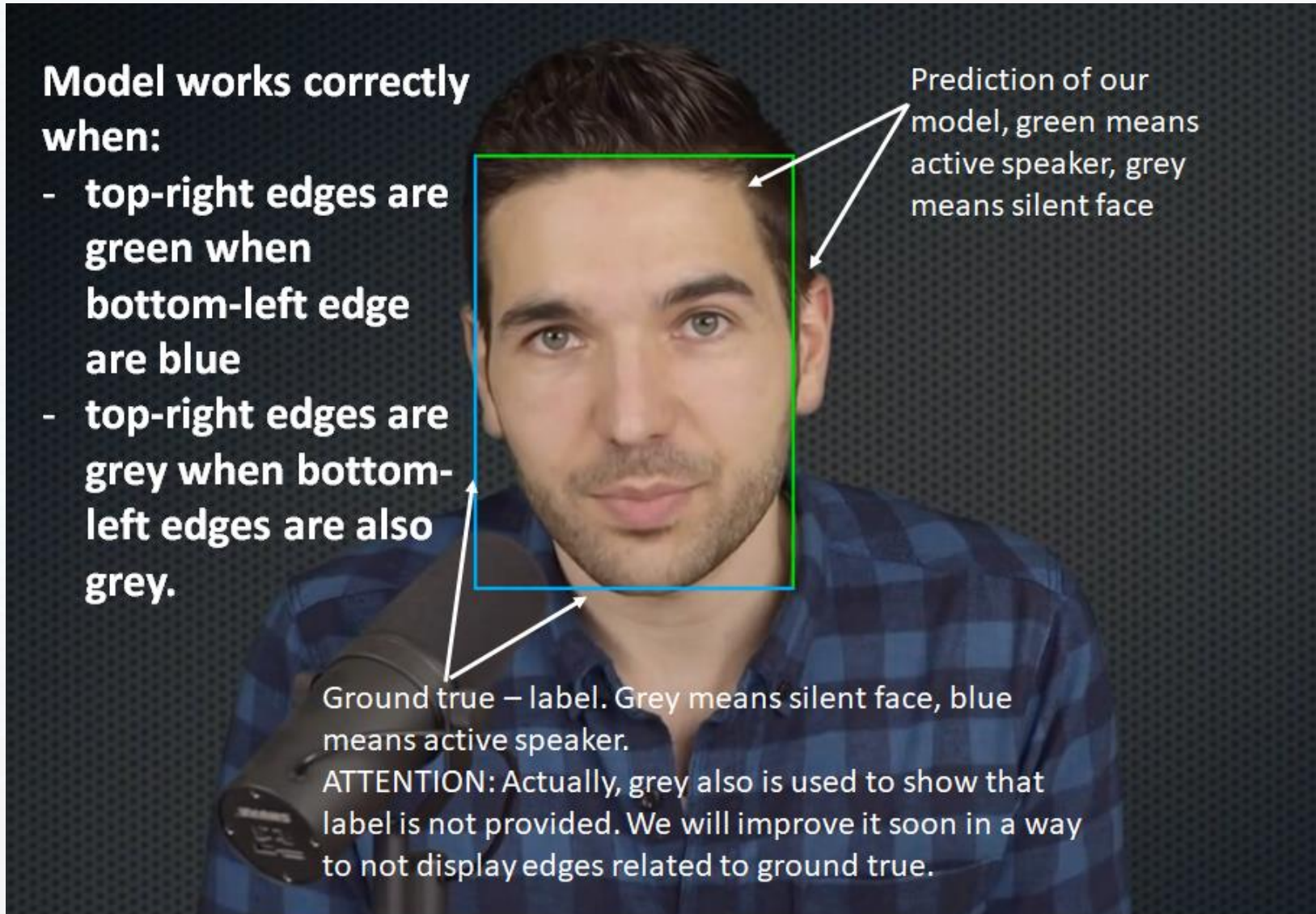
Model works correctly when:

- top-right edges are green when bottom-left edge are blue
- top-right edges are grey when bottom-left edges are also grey.

Prediction of our model, green means active speaker, grey means silent face

Ground true – label. Grey means silent face, blue means active speaker.

ATTENTION: Actually, grey also is used to show that label is not provided. We will improve it soon in a way to not display edges related to ground true.



Our implementation demo run

<https://slack-files.com/T5BNTD7V4-F01FARWEAQJ-499aaac40d>

<https://slack-files.com/T5BNTD7V4-F01FAFCEF5G-1f19a014df>

Dataset improvements

As is: 1 or 2 speaking faces, with different level of noise.

To be:

- 2 and more simultaneous speakers
- not speaking faces (i.e. no lips movement)
- Inaudible faces (removed corresponding voice, add voice of different speaker)
- Speaking face but audio out of synch.
- Speaking face but audio replaced to be looks like in synch e.g. <https://www.youtube.com/user/BadLipReading>

Training dataset improvement by increase cases diversity

As is: 145 clips covering only talk-shows, NEWS

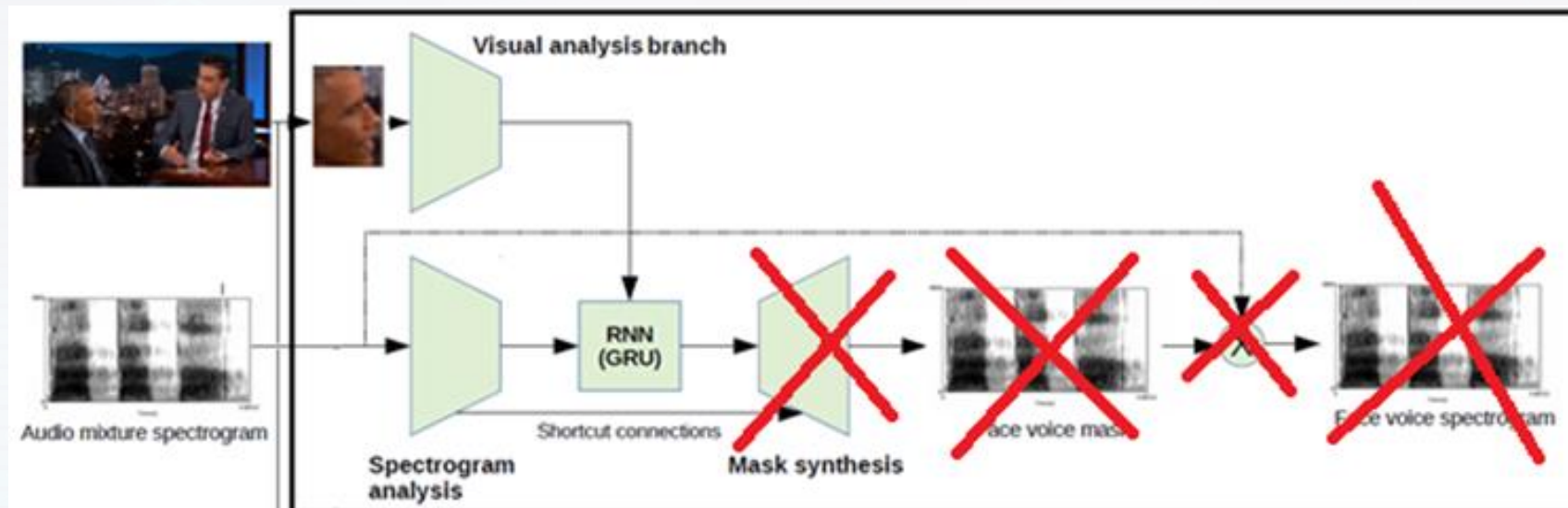


To be: Continuous real 48h TV stream with 5 channels (containing TV shows, TV series, NEWS, ads, etc.) – no annotations yet.



Test dataset improvement by record real TV stream

Active Speaker classification verification



Research in progress

Voice localization and separation

- Pending patent application (hope soon published, submitted 2019...)
 - Benchmark with existing solutions (<https://paperswithcode.com>)
 - Training dataset improvement by increase cases diversity
 - Test dataset improvement by record real TV stream
 - Replace model to lightweight (etc. depth wise separable convolutions)
 - Improve model architecture
-
- Publication of result for A-V speaker separation, also A-V Active Speaker Detector

Dziękuję za uwagę!

