Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie
AGH University of Science
and Technology

Quality Indicator - Lip Sync

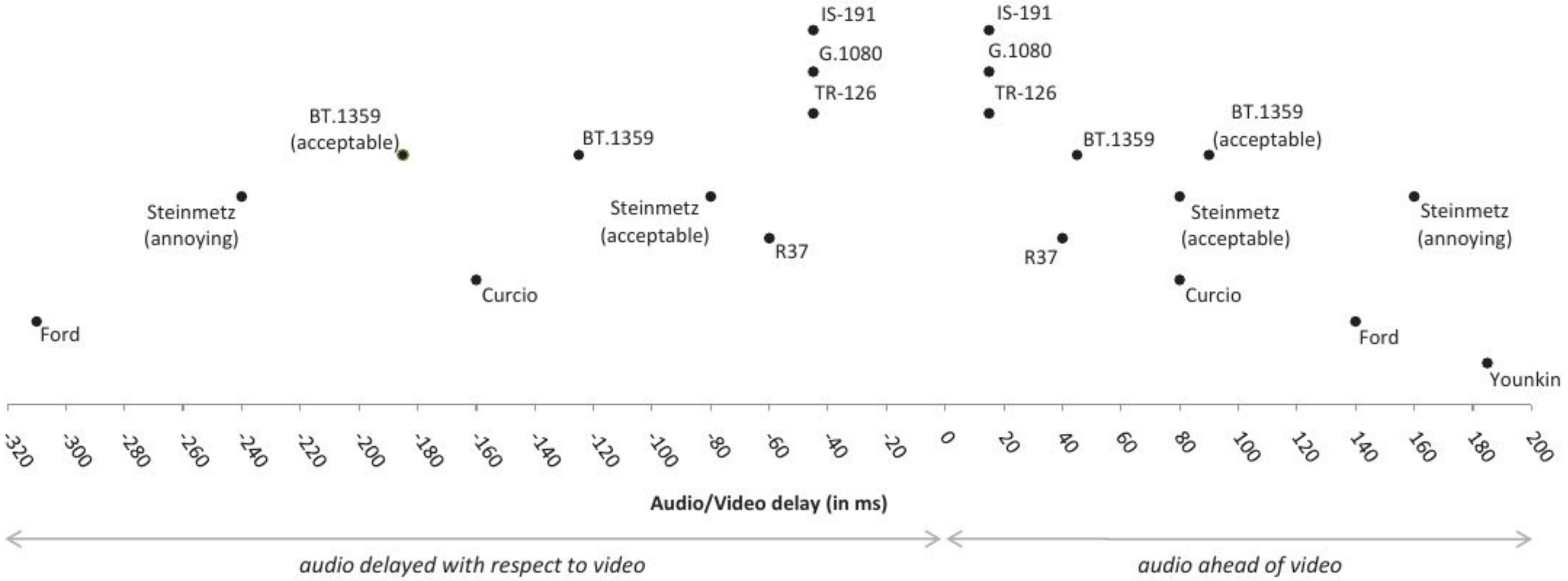
Mikołaj Leszczuk, Lucjan Janowski, Jakub Nawała, Wiktoria Rewer,
Wojciech Zima, Łukasz Pułka, George Heston

Lip Sync

- Also known as:
  - **Audio-to-video synchronization**
  - **Lip sync error (lack of)**
  - **Lip flap (lack of)**
- Issue in:
  - **Television**
  - **Videoconferencing**
  - **Film**

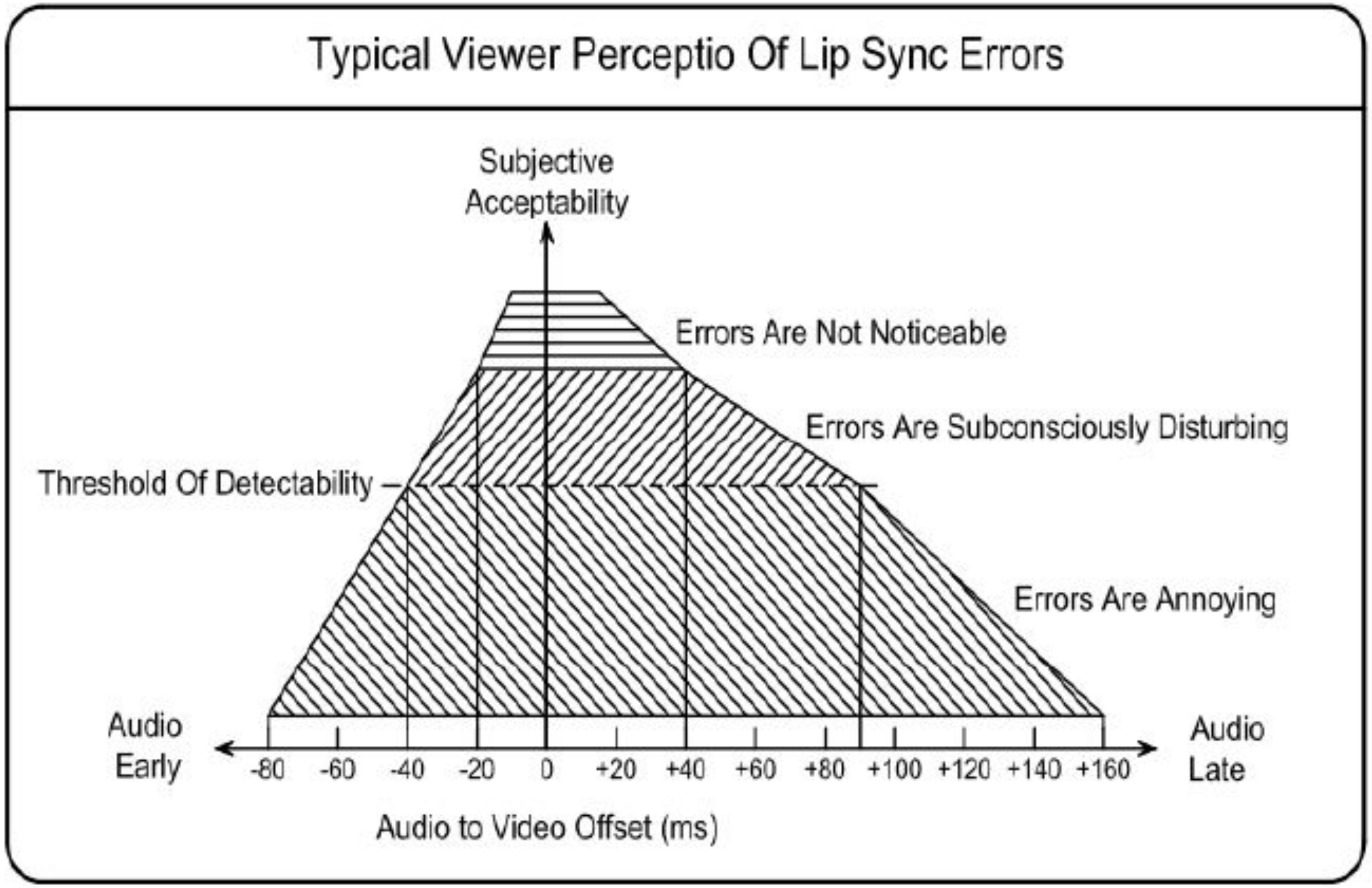
- Referring to relative timing of **audio** (sound) & **video** (image) parts during:
  - **Creation,**
  - **Post-production (mixing)**
  - **Transmission**
  - **Reception**
  - **Play-back processing**

Lip Sync Influence (1/2)



Staelens, Nicolas, Jonas De Meulenaere, Lizzy Bleumers, Glenn Van Wallendael, Jan De Cock, Koen Geeraert, Nick Vercammen, et al. 2012. “Assessing the Importance of Audio/video Synchronization for Simultaneous Translation of Video Sequences.” Multimedia Systems 18 (6): 445–457.

Lip Sync Influence (2/2)



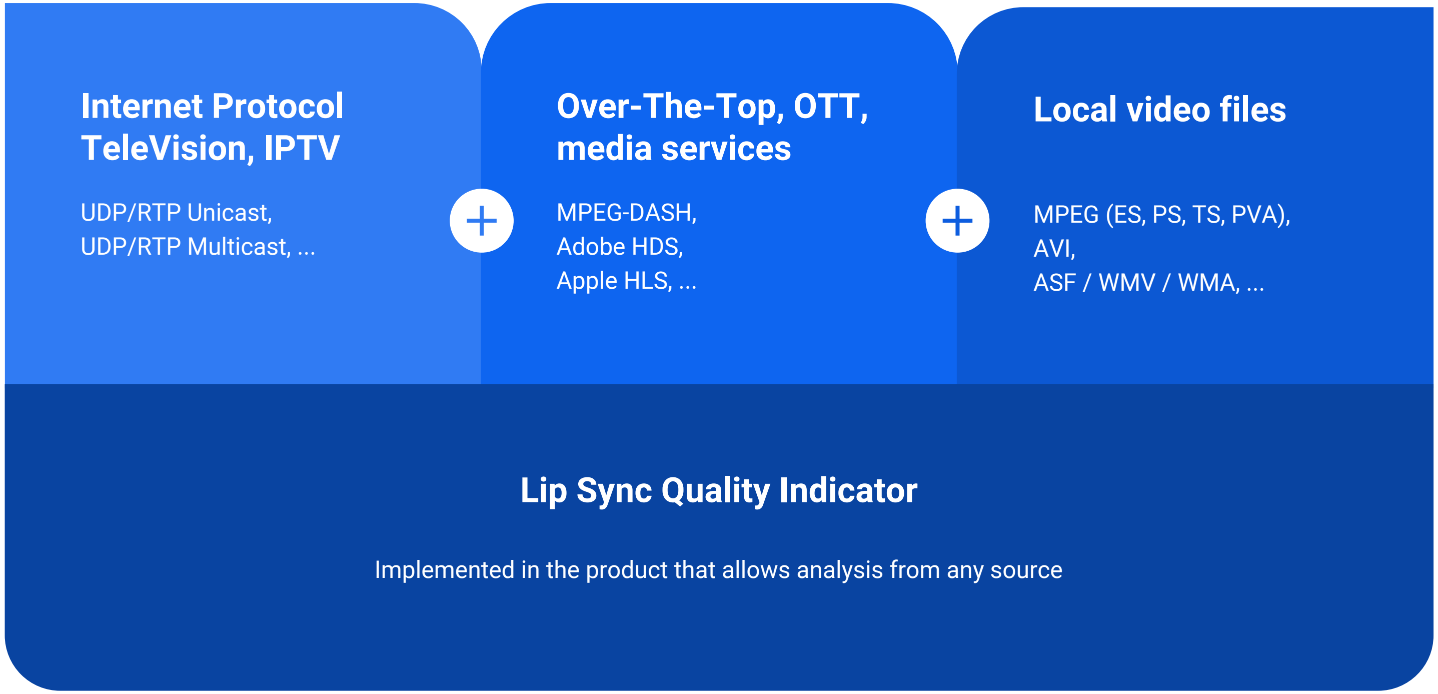
S. Kunić, Z. Šego and B. Z. Cihlar, "Analysis of audio and video synchronization in TV digital broadcast devices," 2017 International Symposium ELMAR, Zadar, 2017, pp. 67-72.

“Quality Indicator - Lip Sync” Project

Ministry of Science
and Higher Educati
Republic of Poland

NeT
RESEARCH

No-Reference Measurement



Internet Protocol TeleVision, IPTV

UDP/RTP Unicast,
UDP/RTP Multicast, ...

+

Over-The-Top, OTT, media services

MPEG-DASH,
Adobe HDS,
Apple HLS, ...

+

Local video files

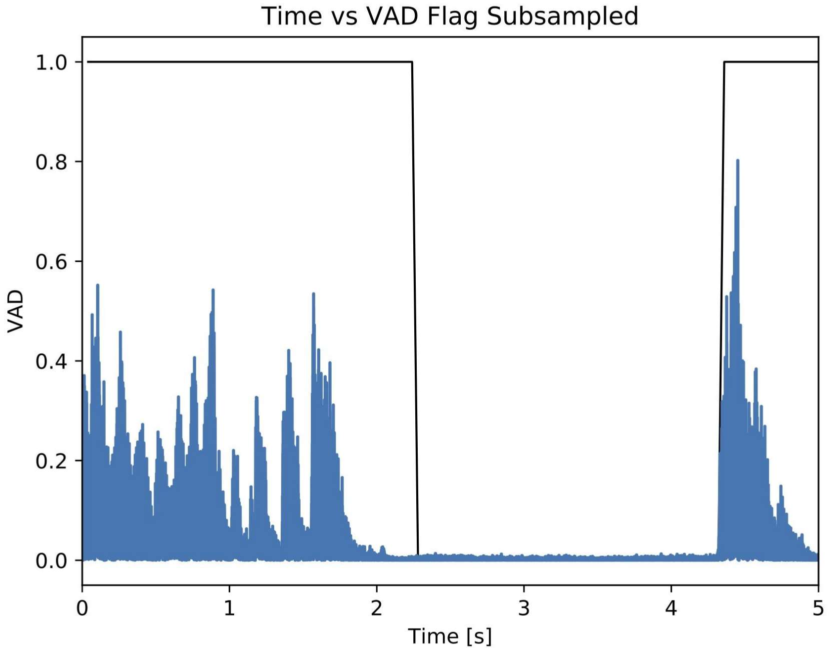
MPEG (ES, PS, TS, PVA),
AVI,
ASF / WMV / WMA, ...

Lip Sync Quality Indicator

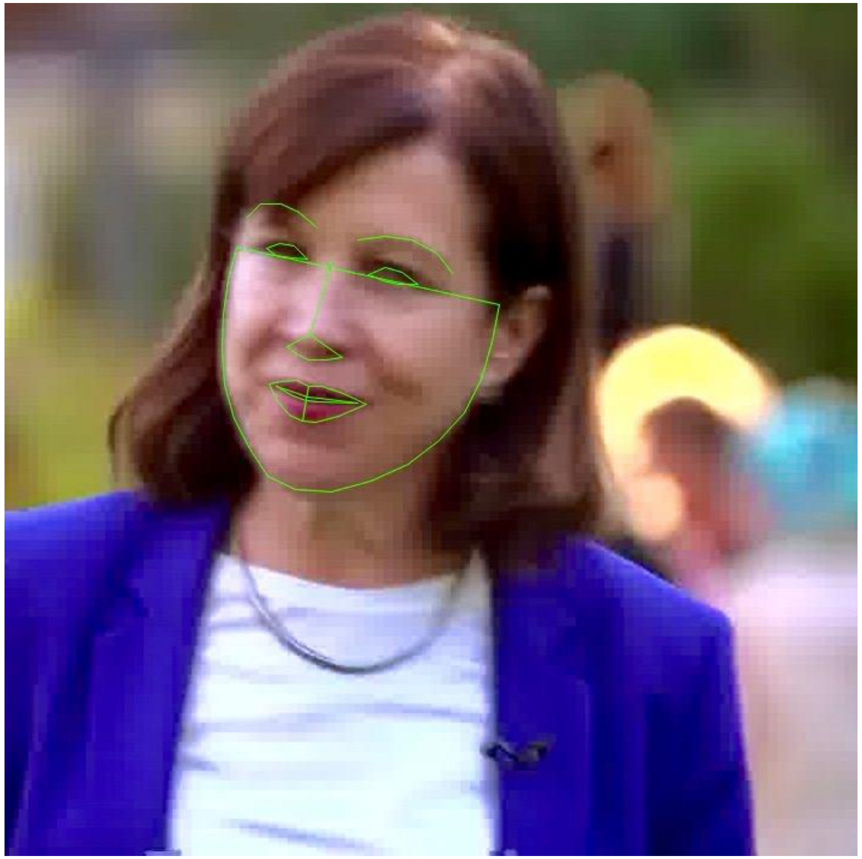
Implemented in the product that allows analysis from any source

Detection and Tracking

Voice Activity



Lip Motion



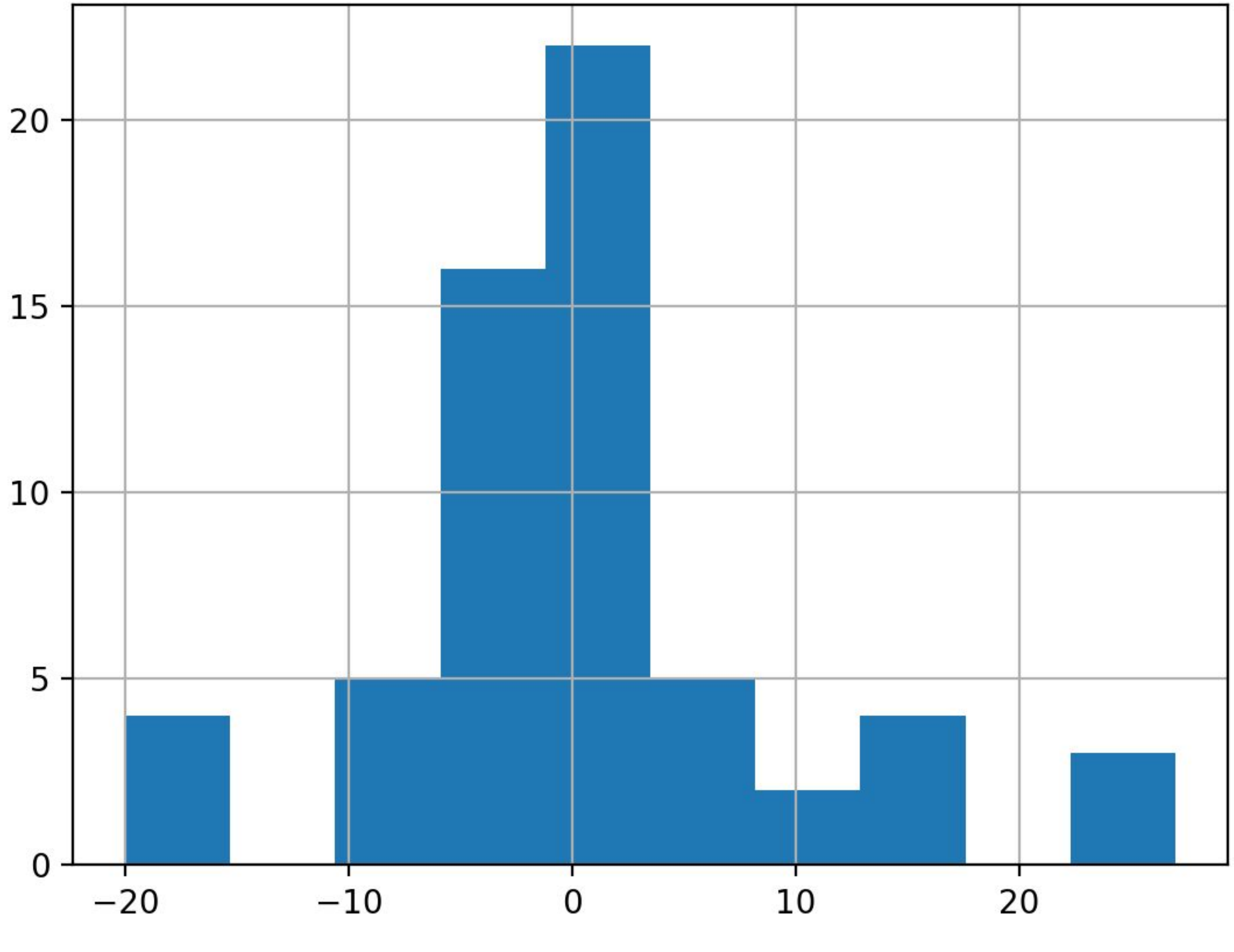
Background Technologies Used

- » VLC media player
- » Google
  Voice Activity Detector
- » Dlib

Current Test Results

- Tested on:
  - **8** SRC, each **~30** min long (mainly YouTube channels of news TV)
  - **~600** shorter video clips extracted:
    - Exactly **1** face visible for whole duration
    - Each having **300** frames (**12** s)
- Further filtering: minimum **~20-25%** of silence
- Artificially introduced Lip Sync search window of **+/- 25** frames (**+/- 1** s)

Histogram of Errors



Quantitative Results

- Mean error:
  - **Min: 40 ms (1 frame)**
  - **Max: 240 ms (6 frames)**
- Median error:
  - **Min: 40 ms (1 frame)**
  - **Max: 160 ms (4 frames)**
- Share of sequences with Lip Sync possible to be determined:
  - **Min: ~2%**
  - **Max: ~25%**

Conclusions & Further Work

- » More algorithmically challenging than expected
- » First results to be delivered by end of this year
- » Looking forward to advancing our solution (Artificial Intelligence?)
- » Collaboration?

AGH

Thank You for Your Attention!