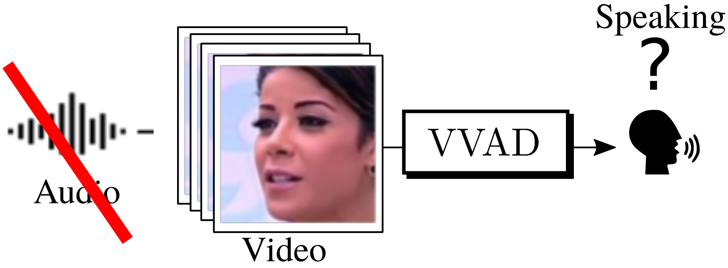
Learning Visual Voice Activity Detection with an Automatically Annotated Dataset

Sylvain Guy, Stéphane Lathuilière, Pablo Mesejo and Radu Horaud

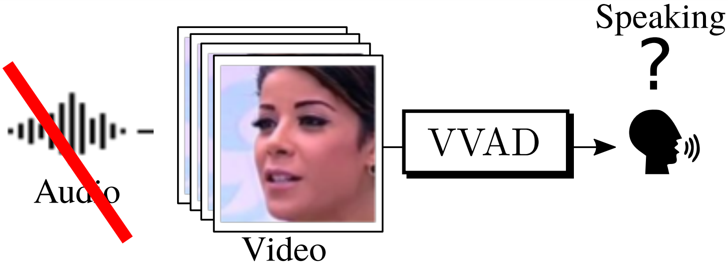
Inria Grenoble Rhône-Alpes and Univ. Grenoble Alpes, France
LTCI, Télécom Paris, Institut polytechnique de Paris, France
Andalusian Research Institute in Data Science and
Computational Intelligence (DaSCI), University of Granada, Spain.

ICPR 2021

Visual Voice Activity Detection(VVAD)



Visual Voice Activity Detection(VVAD)

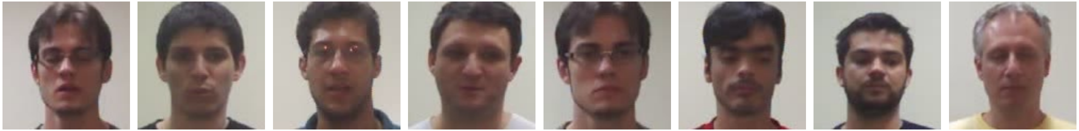


Why do we need VVAD?

(c) Audio unavailable

(d) Noisy Audio

Existing Datasets



(e) Speaking examples **(f)** Silent examples

Figure: MVAD dataset.

(a) Speaking examples **(b)** Silent examples

Figure: CUAVE dataset.

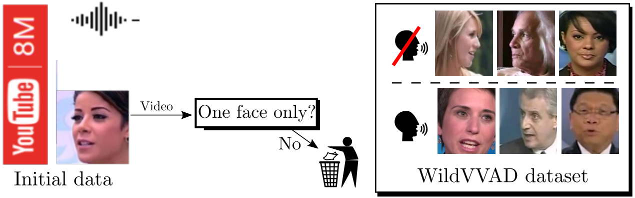
Existing datasets are too simple and too constrained.

Automatic Dataset Annotation

Initial data

WildVVAD dataset

Automatic Dataset Annotation



Initial data

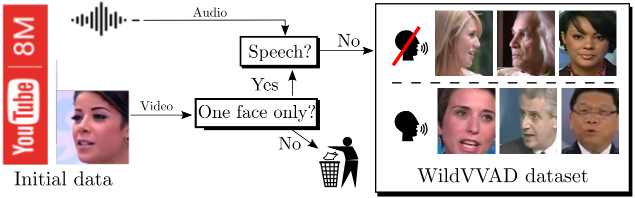
Video

One face only?

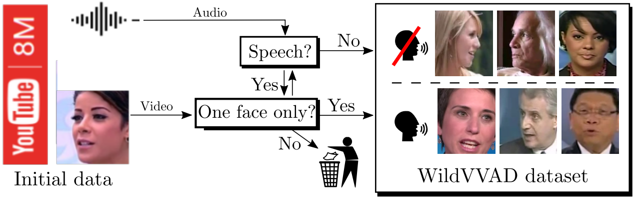
No

WildVVAD dataset

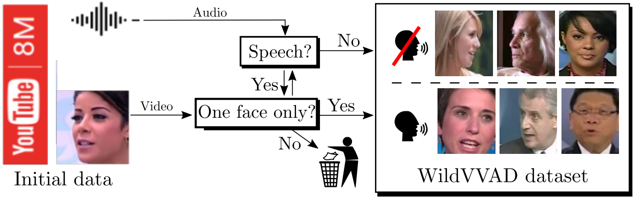
Automatic Dataset Annotation



Automatic Dataset Annotation



Automatic Dataset Annotation



8M

Audio

Speech?

No

Yes

Video

One face only?

Yes

No

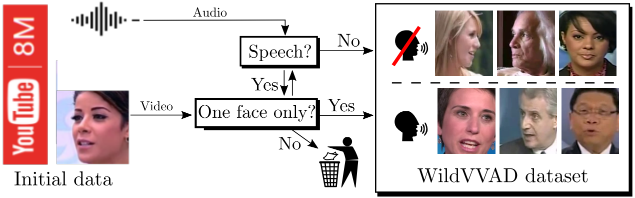
Initial data

WildVVAD dataset

(a) Speaking examples

(b) Silent examples

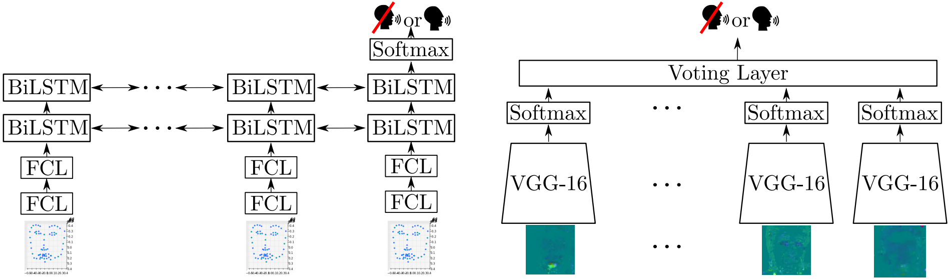
Automatic Dataset Annotation



WildVVAD:

- 13000 videos
- High diversity
- Manually cleaned test set
- Percentage of mislabeled speaking and silent videos are of 12% and 8.6%, respectively.

Proposed architectures

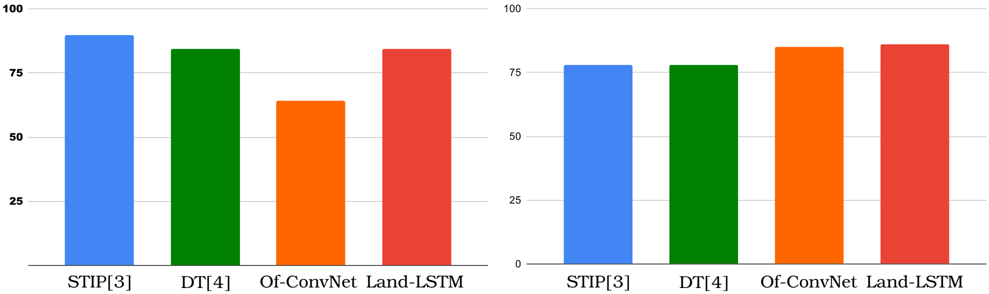


(a) Land-LSTM

(b) OF-ConvNet

Figure: Architectures of the two proposed models.

Comparison with State-of-the-Art



(a) Cuave Dataset

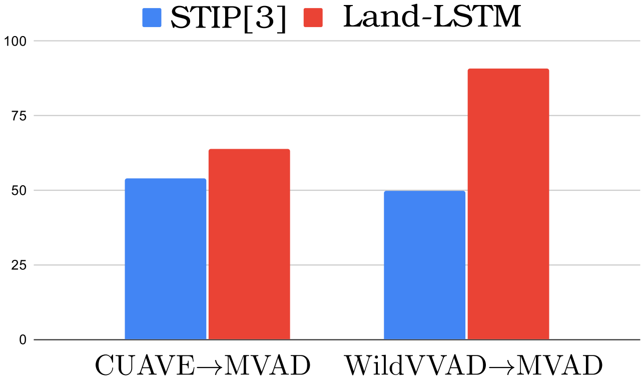
(b) WildVVAD

Figure: Experimental evaluation.

Cross-dataset Experiments

Two questions:

- Which method has better generalization features?
- Which is the best suited dataset to learn a general purpose VVAD model?



Conclusion

Contributions

- We propose a method for automatically collecting a dataset for VVAD.

(a) Speaking examples

(b) Silent examples

- We introduce and compare two deep architectures for VVAD
- We show a better generalization ability of VVAD models when they are trained on our dataset.