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)

Why do we need VVAD?

(a) Audio unavailable

(b) Noisy Audio

S. Guy et al. Learning VVAD
Visual Voice Activity Detection (VVAD)

Why do we need VVAD?

(c) Audio unavailable

(d) Noisy Audio

S. Guy et al. Learning VVAD
Existing datasets are too simple and too constrained.

**Figure:** MVAD dataset.

**Figure:** CUAVE dataset.

S. Guy et al.  Learning VVAD
Automatic Dataset Annotation

(a) Speaking examples
(b) Silent examples

WildVVAD:
- 13,000 videos
- High diversity
- Manually cleaned test set
- Percentage of mislabeled speaking and silent videos are of 12% and 8.6%, respectively.

S. Guy et al. Learning VVAD
Automatic Dataset Annotation

WildVVAD:
13,000 videos
High diversity
Manually cleaned test set
Percentage of mislabeled speaking and silent videos are of 12% and 8.6%, respectively.

Initial data

Video

One face only?

No

WildVVAD dataset

S. Guy et al.
Learning VVAD
Automatic Dataset Annotation

WildVVAD:
- 13,000 videos
- High diversity
- Manually cleaned test set
- Percentage of mislabeled speaking and silent videos are of 12% and 8.6%, respectively.

S. Guy et al. Learning VVAD
Automatic Dataset Annotation

WildVVAD: 13,000 videos
High diversity
Manually cleaned test set
Percentage of mislabeled speaking and silent videos are 12% and 8.6%, respectively.

S. Guy et al. Learning VVAD
Automatic Dataset Annotation

(initial data)

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

(a) Speaking examples
(b) Silent examples

S. Guy et al. Learning VVAD
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

Figure: Architectures of the two proposed models.
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?
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.