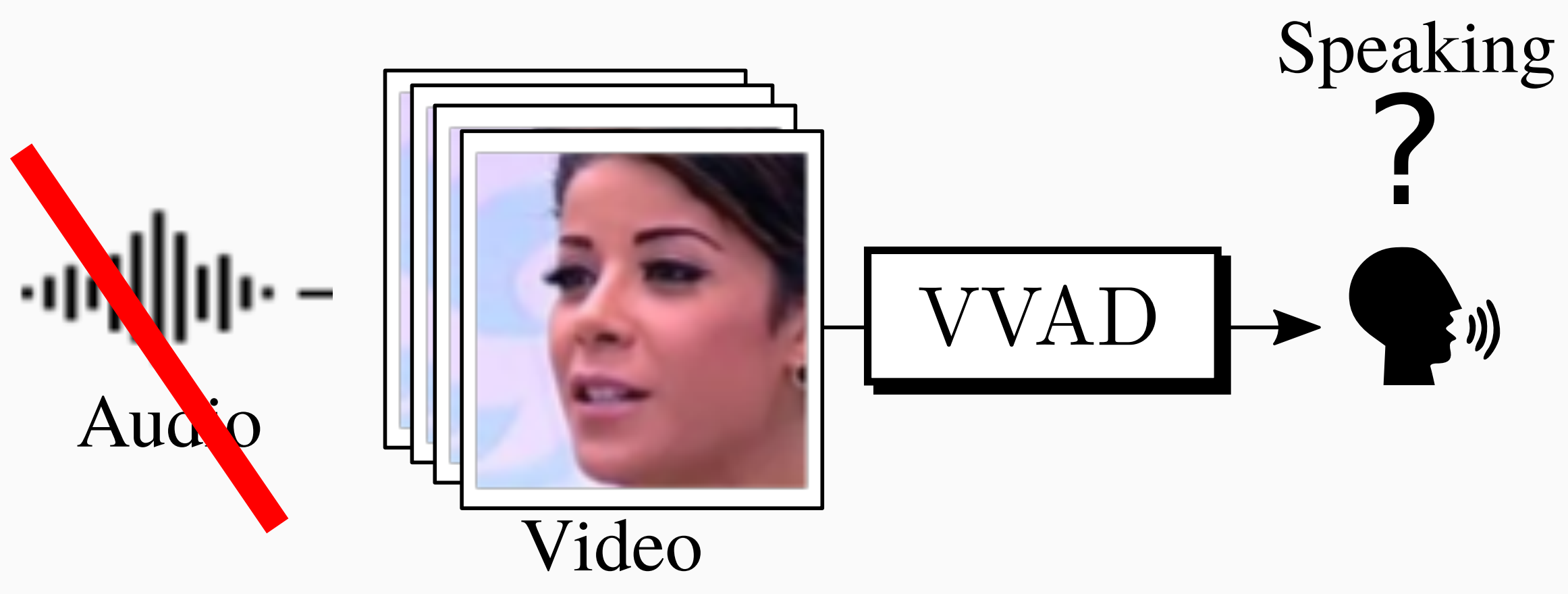
Learning Visual Voice Activity Detection with an Automatically Annotated Dataset

Sylvain Guy[1], Stéphane Lathuilière[2], Pablo Mesejo[3] and Radu Horaud[1]

[1] Inria Grenoble Rhône-Alpes, [2] *Télécom Paris, IPP*, [3] University of Granada

Visual Voice Activity Detection (VVAD)

VVAD consists in detecting whether a person is speaking without using audio signal.



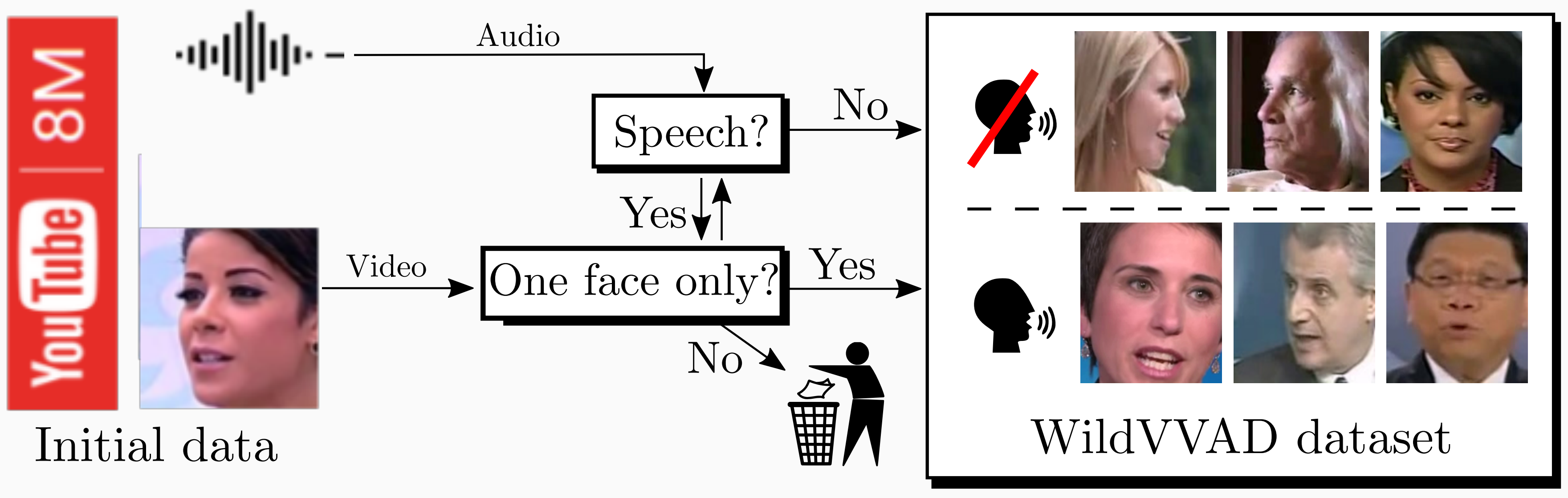
Why do we need VVAD?

Case 1: Audio unavailable

Case 2: Noisy Audio

Automatic Dataset Annotation

We introduce a novel algorithm to automatically collect a dataset for VVAD:



We collect 13000 video clips with high diversity.

Proposed models for VVAD

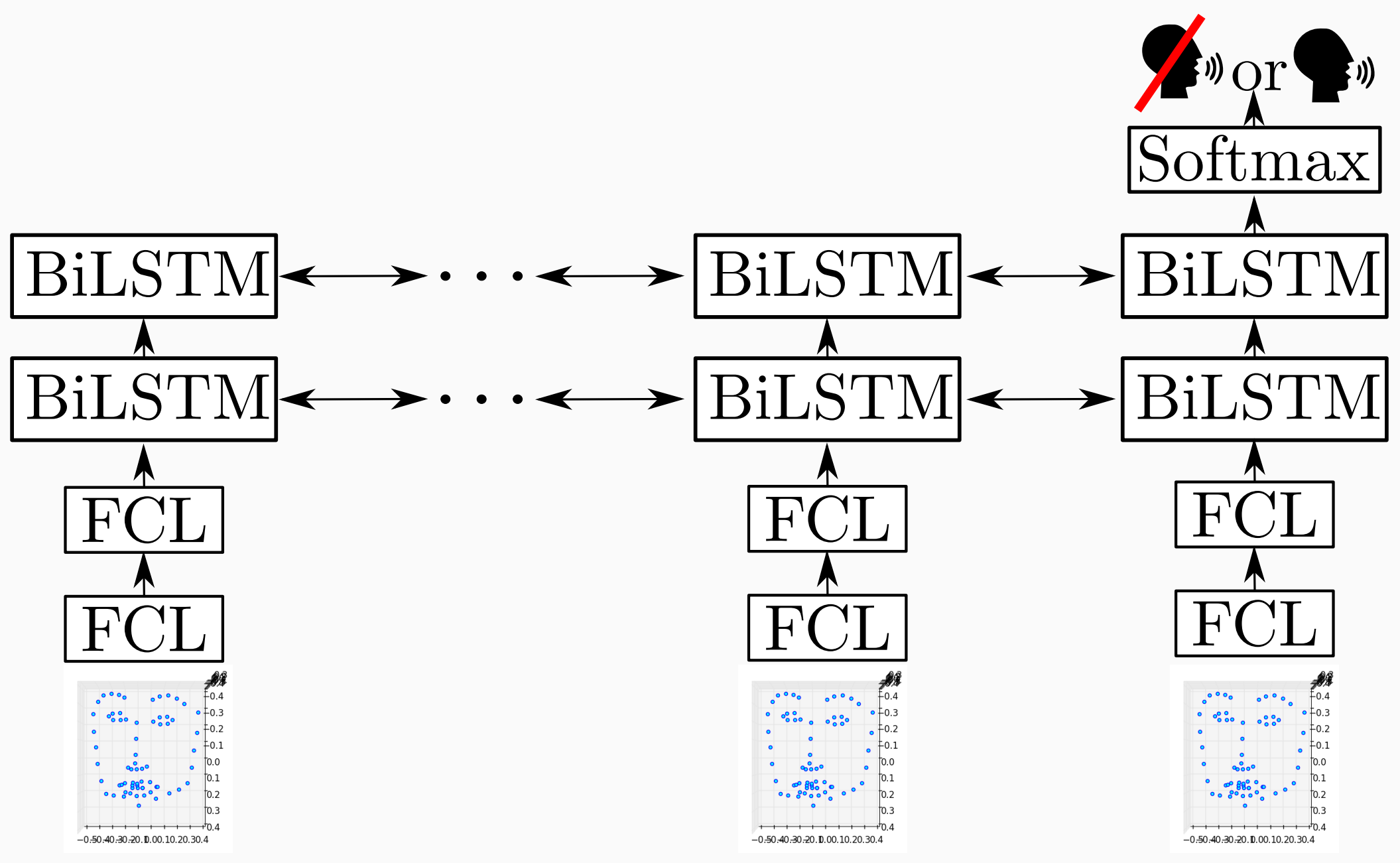


Figure 4: Land-LSTM

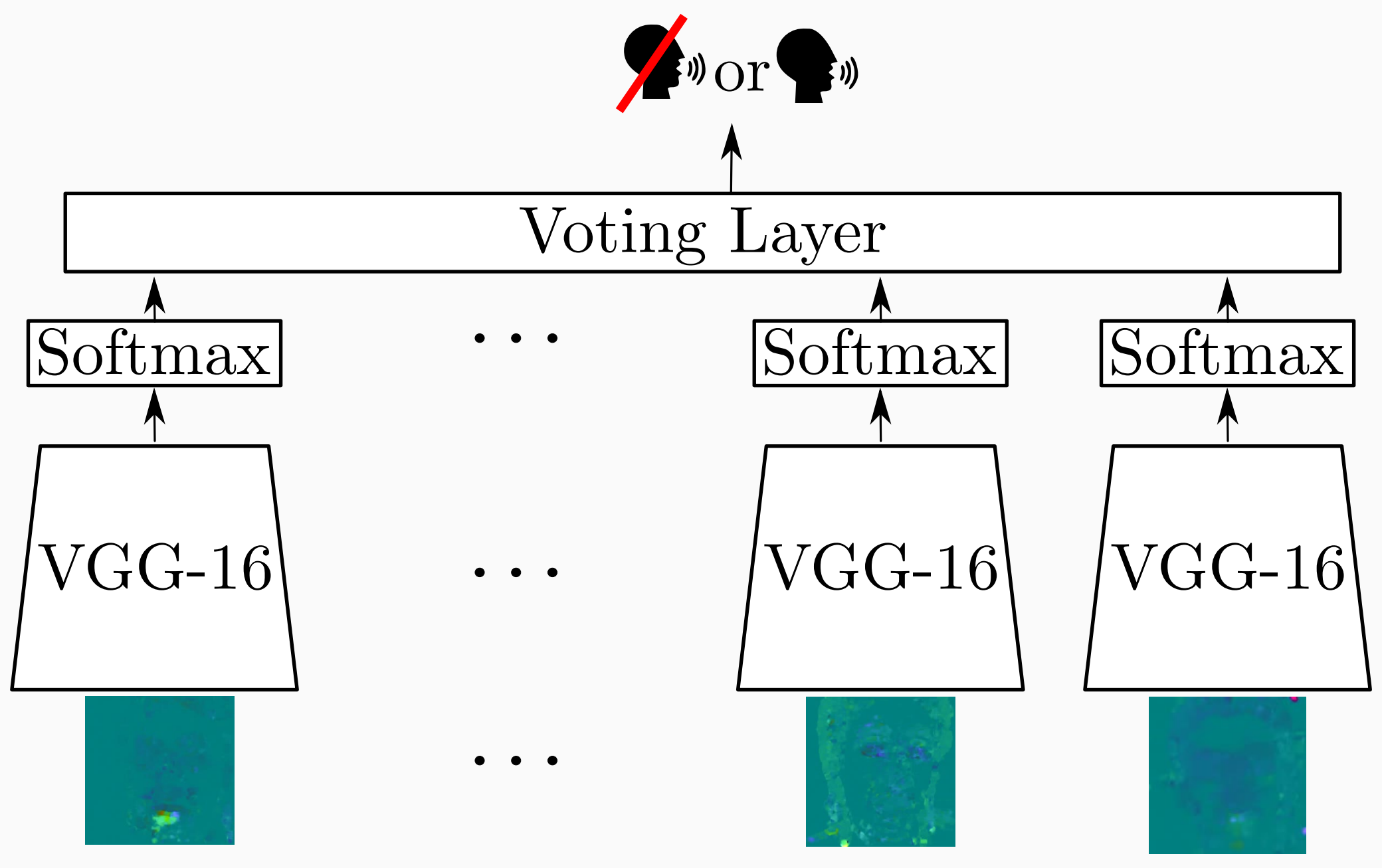


Figure 5: OF-ConvNet

Experiments

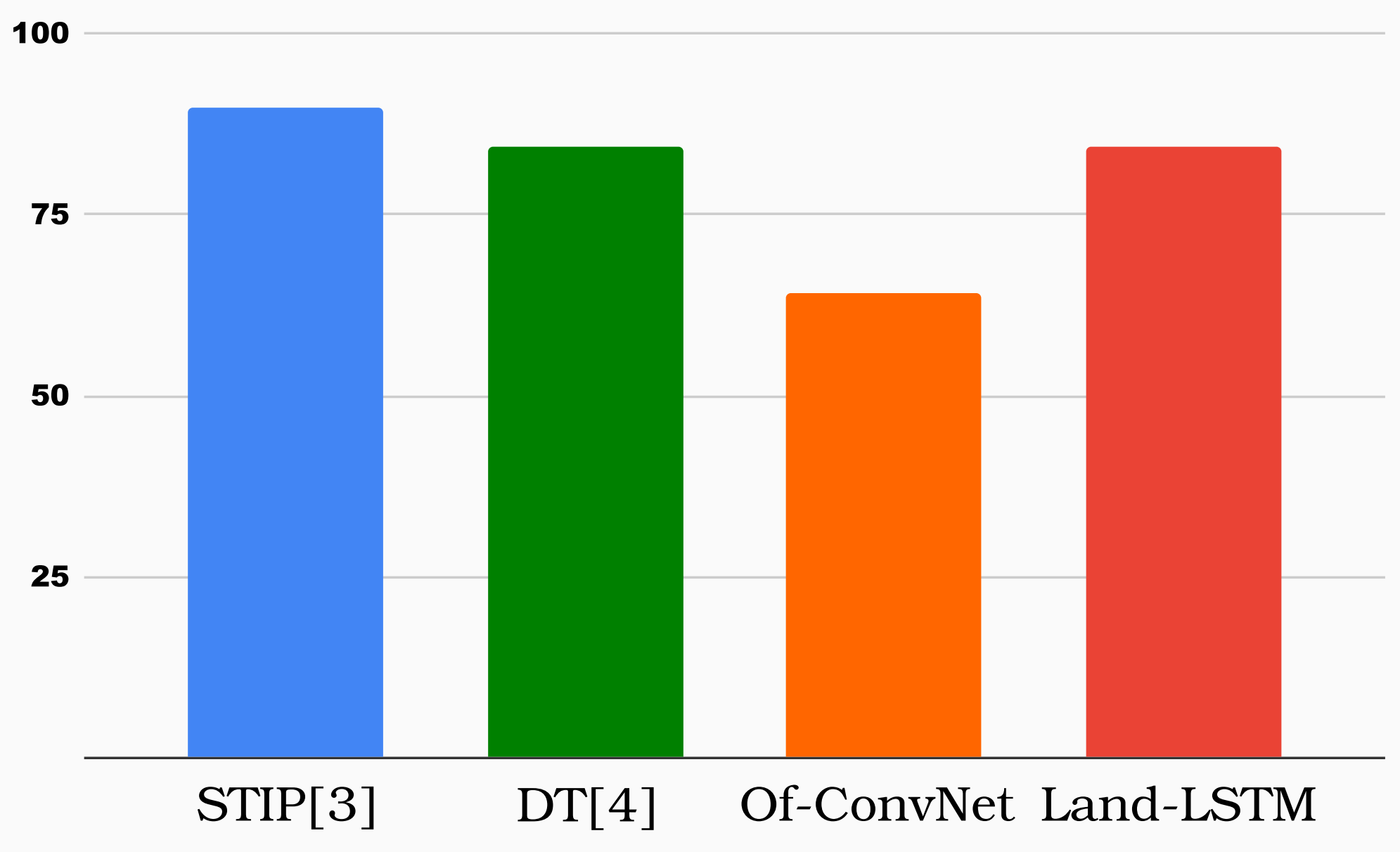


Figure 6: Cuave Dataset

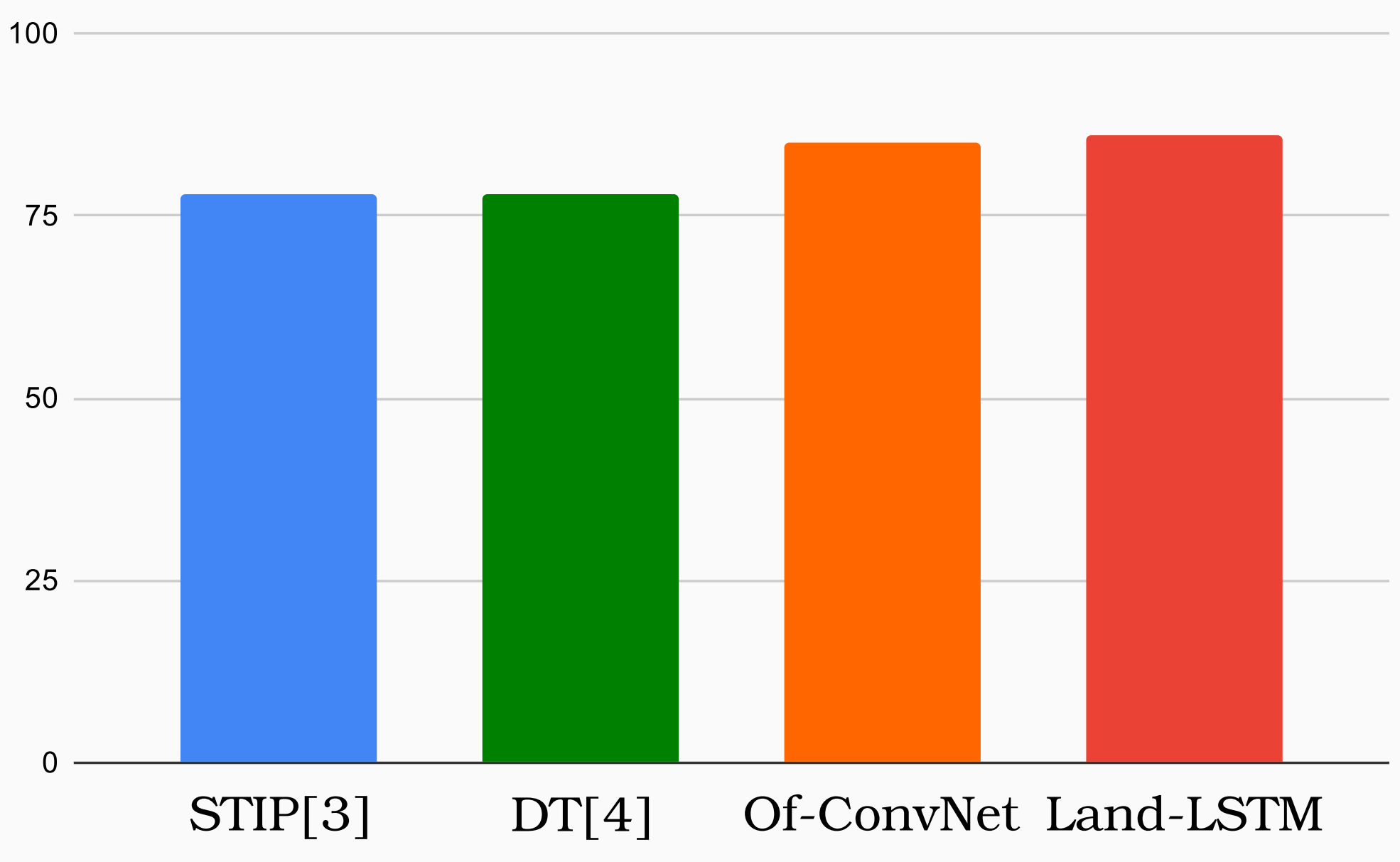


Figure 7: WildVVAD

References

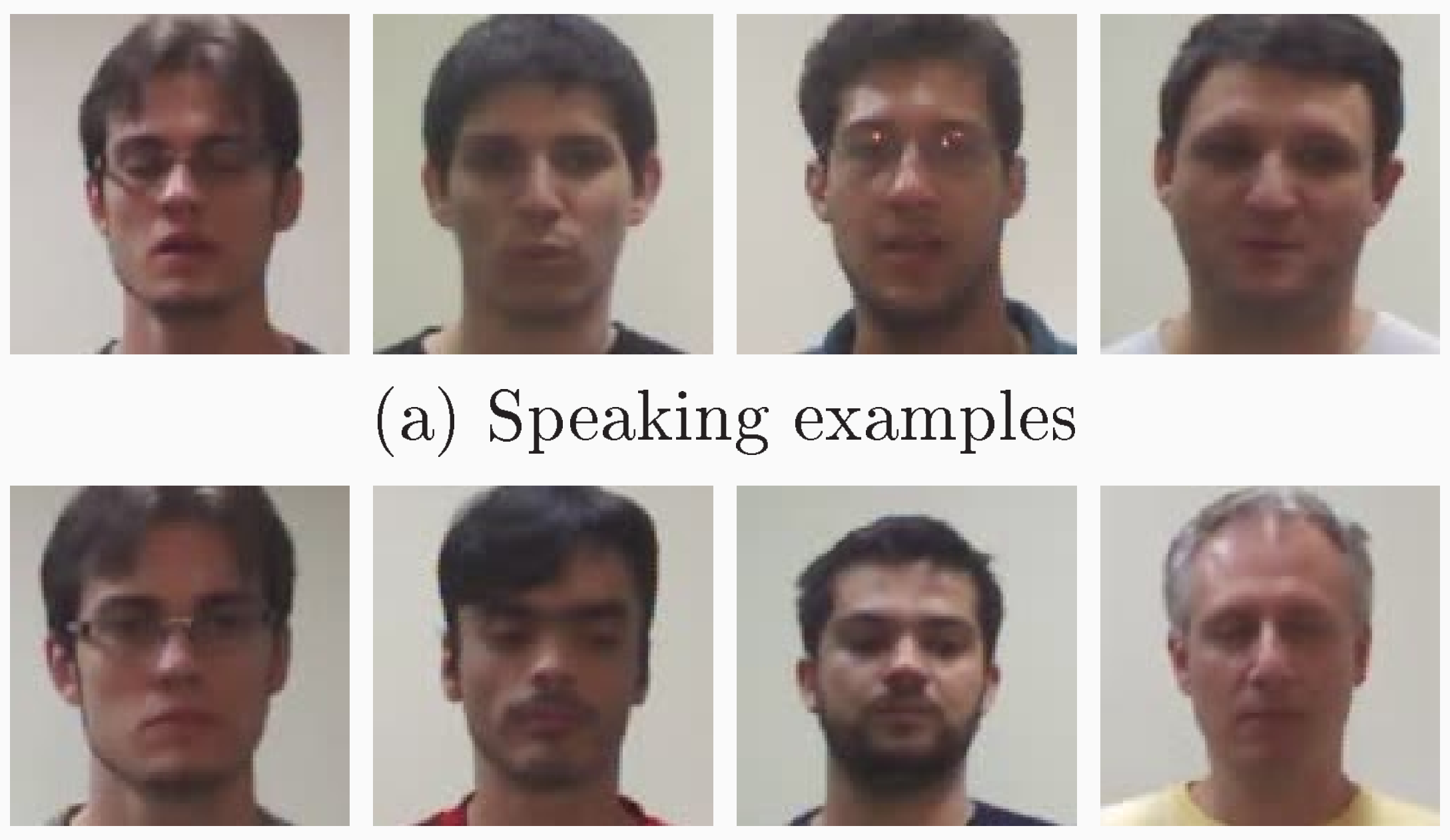
[1] Minotto *et al.*, Simultaneous-speaker voice activity detection and localization using mid-fusion of svm and hmms, IEEE TMM, 2014

[2] Patterson *et al.*, CUAVE, A new audio-visual database for multimodal human-computer interface research. IEEE ICASSP, 2002

[3] Laptev, On space-time interest points, IJCV, 2005

[4] Wang *et al.*, Action recognition by dense trajectories, CVPR, 2011

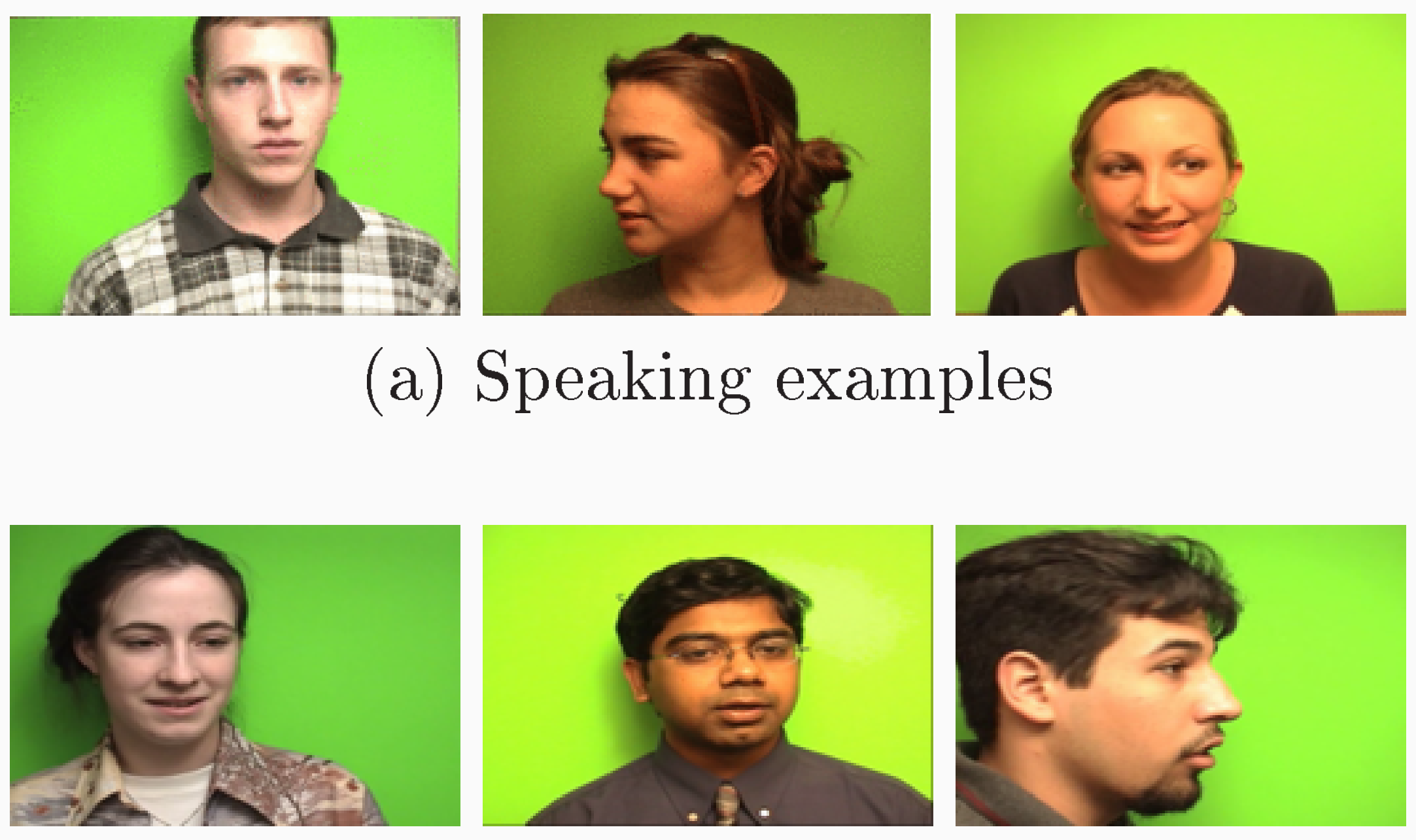
Existing dataset



(a) Speaking examples

(b) Silent examples

Figure 1: MVAD dataset [1].

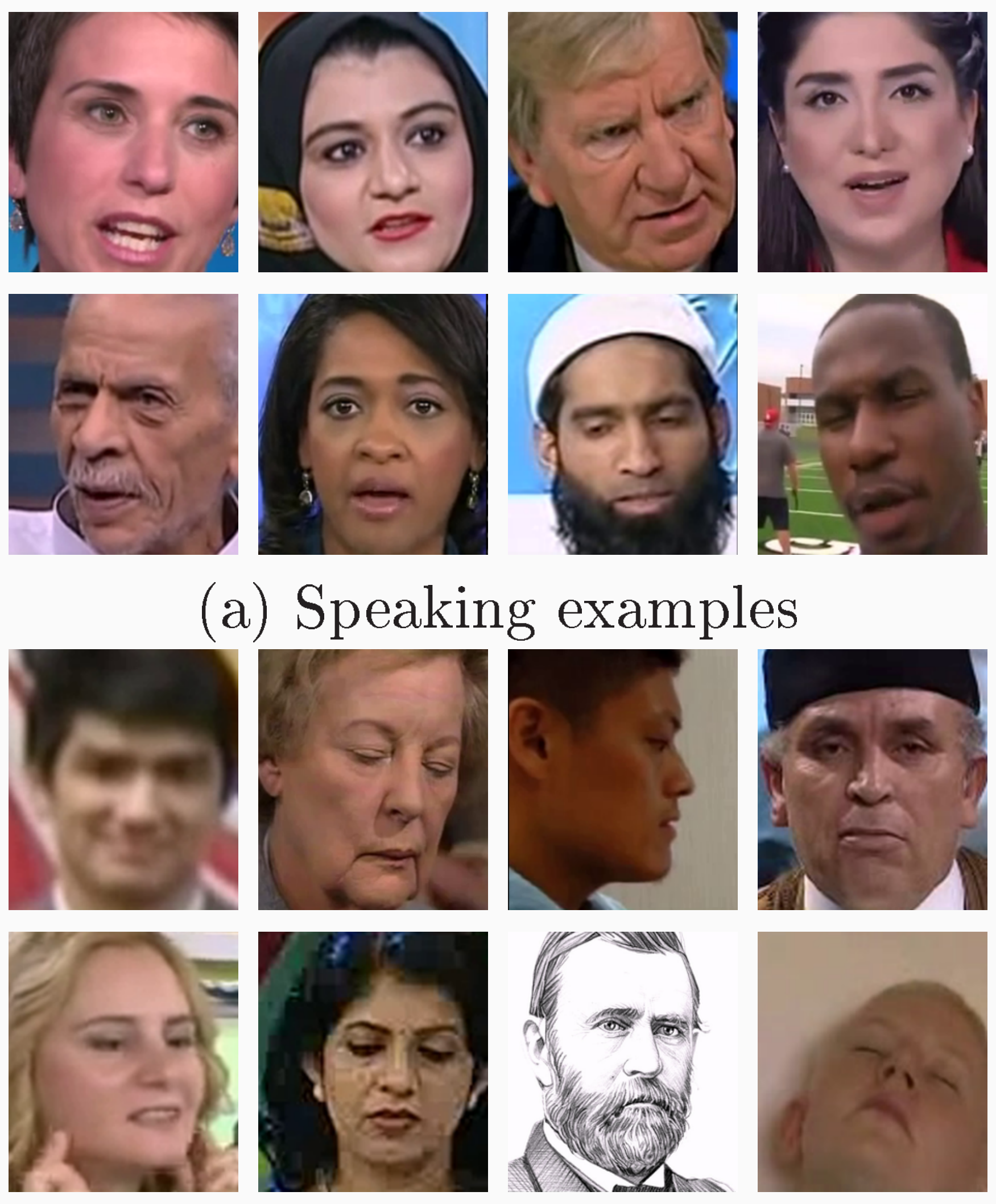


(a) Speaking examples

(b) Silent examples

Figure 2: CUAVE dataset [2].

WildVVAD dataset



(a) Speaking examples

(b) Silent examples

Figure 3: WildVVAD

Cross-dataset experiments

Table 1: Results obtained on the MVAD dataset when training on CUAVE and WildVVAD.

	CUAVE→MVAD		
Method	TPR	TNR	ACC
STIP [3]	76.74%	34.24%	54.78%
Land-LSTM	**81.42%**	**59.79%**	**64.0%**
	WildVVAD→MVAD		
Method	TPR	TNR	ACC
STIP [3]	21.44%	76.65%	49.98%
OF-ConvNet	62.69%	68.71%	65.45%
Land-LSTM	**91.30%**	**90.06%**	**91.01%**

Acknowledgements

This work has been funded by the EU H2020 project #871245 SPRING and by the Multidisciplinary Institute in Artificial Intelligence (MIAI) # ANR-19-P3IA-0003.