

The VVAD-LRS3 Dataset for Visual Voice Activity Detection

Adrian Lubitz¹^a, Matias Valdenegro-Toro²^b and Frank Kirchner^{1,3}^c

¹Department of Computer Science, University of Bremen, 28359 Bremen, Germany

²Department of AI, University of Groningen, 9747 AG Groningen, The Netherlands

³Robotics Innovation Center, German Research Center for Artificial Intelligence, Bremen, Germany

Keywords: Human-Robot Interaction, Perception, Dataset, Deep Learning.

Abstract: Robots are becoming everyday devices, increasing their interaction with humans. To make human-machine interaction more natural, cognitive features like Visual Voice Activity Detection (VVAD), which can detect whether a person is speaking or not, given visual input of a camera, need to be implemented. Neural networks are state of the art for tasks in Image Processing, Time Series Prediction, Natural Language Processing and other domains. Those Networks require large quantities of labeled data. Currently there are not many datasets for the task of VVAD. In this work we created a large scale dataset called the VVAD-LRS3 dataset, derived by automatic annotations from the LRS3 dataset. The VVAD-LRS3 dataset contains over 44K samples, over three times the next competitive dataset (WildVVAD). We evaluate different baselines on four kinds of features: facial and lip images, and facial and lip landmark features. With a Convolutional Neural Network Long Short Term Memory (CNN LSTM) on facial images an accuracy of 92% was reached on the test set. A study with humans showed that they reach an accuracy of 87.93% on the test set.

1 INTRODUCTION

Technology is integrating more and more into the life of the modern man. A very important question is how are people interacting with technology. The human brain does not react emotionally to artificial objects like computers and mobile phones. However, the human brain reacts strongly to human appearances like shape of the human body or faces (gun Choi and Kim, 2009). Therefore humanoid robots are the most natural way for human-machine interaction, because of the human-like appearance. This hypothesis is strongly supported by HRI Research from (Kanda and Ishiguro, 2017), (Ángel Pascual del Pobil Ferré et al., 2013), (OZTOP et al., 2005) and (Miwa et al., 2003). They see Social Robots as a part of the future society. (Kanda and Ishiguro, 2017) also defines the following three issues which need to be solved to bring social robots effectively and safely to the everyday life:


- a. Sensor network for tracking robots and people
- b. Development of humanoids that can work in the daily environment.


- c. Development of functions for interactions with people.

This paper is located in the field c, as we propose a large scale dataset to train models for the task of Visual Voice Activity Detection (VVAD) which detects whether a person is speaking to a robot or not, given the visual input of the robot's camera.

VVAD is an important cognitive feature in a Human-Robot Interaction(HRI). As we want Robots to integrate seamlessly into our society, Human-Robot Interaction needs to be as close as possible to Human-Human Interaction (HHI). VVAD can be used for speaker detection in the case where multiple people are in the robot's field of view. Furthermore it can be useful to detect directed speech in noisy environments.

In this paper we present a new benchmark for the VVAD task, produced from the LRS3 dataset (Triantafyllos Afouras, 2018) which contains TED Talks, and by using the provided textual transcripts, we can extract parts of the TED Talk video in order to generate positive/negative video samples for the VVAD task. Our dataset contains 37.6K training and 6.6K validation samples, making it the largest VVAD dataset currently. The dataset will be publicly available on the internet. We provide baseline models using commonly

^a <https://orcid.org/0000-0003-0609-2850>

^b <https://orcid.org/0000-0001-5793-9498>


^c <https://orcid.org/0000-0002-1713-9784>



Figure 1: Example of error detection - Person is classified as having a mouth activity, however does not speak (Meriem Bendris and Chollet, 2010).

used neural network architectures. In an experimental setup with a CNN LSTM an accuracy of 92% was reached on the test set. A study with humans showed that humans reach an accuracy of 87.93% on the test set.

This paper contributes a large scale dataset and a simple approach on how to use it for VVAD.

2 RELATED WORK

The classic approach to solve VVAD is to detect lip motion. This approach is taken by F. Luthon and M. Liévin in (F. Luthon, 1998). They try to model the motion of the mouth in a sequence of color images with Markov Random Fields. For the lip detection they analyze the images in the *HIS* (*Hue, Intensity, Saturation*) color space, with extracting *close-to-red-hue prevailing regions* this leads to a robust lightening independent lip detection. A different approach was taken by Spyridon Siatras, Nikos Nikolaidis, and Ioannis Pitas in (Spyridon Siatras and Pitas, 2006). They try to convert the problem of lip motion detection into a signal detection problem. They measure the intensity of pixels of the mouth region and classify with a threshold, since they argue that frames with an open mouth have a essentially higher number of pixels with low intensity. In (Meriem Bendris and Chollet, 2010) Meriem Bendris, Delphine Charlet and Gérard Chollet propose a method, which measures the probability of voice activity with the optical flow of pixels in the mouth region. In (Meriem Bendris and Chollet, 2010) the drawback of lip motion detection based approaches is already discussed. As shown in Figure 1, what makes the problem difficult is that people move their lips from time to time although they are not speaking.

This issue is tackled by Foteini Patrona, Alexandros Iosifidis et al. (Patrona et al., 2016). They use a Space Time Interest Point (STIP) or the Dense Trajectory- based facial video representation to train a Single Hidden Layer Feedforward Neural Network. The features are generated from the CUAVE dataset

(Patterson et al., 2002). This erases the implicit assumption (of the approaches above) that lip motion equals voice activity. A more robust approach, which uses Centroid Distance Features of normalized lip shape to train a LSTM Recurrent Neural Network is proposed by Zaw Htet Aung and Panrasee Ritthipravat in (Aung and Ritthipravat, 2016). This method shows a classification accuracy up to 98% on a relatively small dataset. In conclusion all of the mentioned methods use some kind of face detection and some also use mechanics to track the face. This is needed if there is more than one face in the image. From the facial images features are created in different ways. From that point the approaches divide into two branches. The first and naive approach is to assume that lip motion equals speech. This is obviously not always the case, which is why the later approaches do not rely on this hypothesis. The latter approach uses learning algorithms to learn the real mapping between facial images and the speech/no speech. This approach is strongly relying on a balanced dataset to learn a good performing model.

While datasets like the LRS3 or CUAVE (Patterson et al., 2002) provide a good fit for lipreading they lack the negative class for VVAD. There seems to be not many datasets for the VVAD task. The only competitive state of the art dataset for VVAD that we found was the WildVVAD (Guy et al., 2020). WildVVAD is not only 3 times smaller than the VVAD-LRS3 it is also more prone false positive and false negative because of the loose assumption that detected voice activity and a single face in the video equals a *speaking* sample and every detected face in a video sequence without voice activity is a *not speaking* sample. Furthermore the source WildVVAD is drawn from makes it less diverse. Table 1 shows a comparison of state of the art datasets. The VVAD-LRS3 that we propose in this paper is $\sim 3 \times$ larger than WildVVAD.

3 DATASET CAPTURE

To create the large scale VVAD dataset we took the Lip Reading Sentences 3 (LRS3) Dataset introduced by Afouras et al. in (Triantafyllos Afouras, 2018) as a basis. The LRS3 is a dataset designed for visual speech recognition and is created from videos of 5594 TED and TEDx talks. It provides more than 400 hours video material of natural speech. The LRS3 dataset provides videos along with metadata about the face position and a speech transcript. In the LRS3 metadata files the following fields are important for the transformation to the VVAD dataset:

Table 1: Overview of state of the art datasets for VVAD.

Dataset	Samples	Diversity	Pos/Neg Ratio
VVAD-LRS3 (this work)	44,489	Very high	1-to-1
WildVVAD (Guy et al., 2020)	13,000	High	1-to-1
LRS3 (Triantafyllos Afouras, 2018)	>100,000	Very high	1-to-0
CUAVE (Patterson et al., 2002)	~7,000	Low	1-to-0

Table 2: Number of samples for training, validation and test splits of the VVAD-LRS3 dataset.

Training Set	Validation Set	Test Set
37646 Samples	6643 Samples	200 Samples

Text. contains the text for one sample. The length of the text or respectively the sample is defined by length of the scene. That means one sample can get as long as the face is present in the video.

Ref. is the reference to the corresponding YouTube video. The value of this field needs to be appended to <https://www.youtube.com/watch?v=>

FRAME. corresponds to the face bounding box for every frame, where `FRAME` is the frame number, `X` and `Y` is the position of the bounding box in the video and `W` and `H` are the width and height of the bounding box respectively. It is to mention that for the frame number a frame rate of 25 fps is assumed and the values for `X`, `Y`, `W` and `H` are a percentage indication of the width and height of the video.

WORD. maps a timing to every said word. Here `START` and `END` indicate the start and end of the word in seconds respectively. It is to mention that the time is in respect to the start of the sample given by the first frame and not to the start of the whole video.

The LRS3 dataset comes with a low bias towards specific ethnic groups, because TED and TEDx talks are international and talks are held by men and women as well as small children. It also comes with the advantage that it depicts a large variety of people because the likelihood of talking in multiple TED or TEDx talks is rather small. This is a big advantage over the LRS2 and LRW dataset that are extracted from regular TV shows, which brings the risk of overfitting to a specific persons. LRS3 makes learning more robust in that sense. Since natural speech in front of an audience includes pauses for applause and means to structure and control a speech as described in (Nikitina, 2011), the LRS3 dataset provides *speaking* and *not speaking* phases.

To transform LRS3 samples to VVAD ones the given text files are analyzed for these *speaking* and *not speaking* phases. In (Zellner, 1994) Brigitte Zellner shows that pauses occur in natural speech and explicitly in speech in front of an audience. This leads to two constants we need to define in the context of pauses. The first is `maxPauseLength` which defines the maximal length of a pause which is still considered to be an inter speech pause. In consideration of the different types of pauses mentioned in (Zellner, 1994) `maxPauseLength` is set to 1 s. The second constant is `sampleLength` which defines the length of a sample. In other words this defines how long a pause should be to be considered as a negative (*not speaking*) sample or how long a speech phase needs to be to be considered a positive (*speaking*) sample. It shows that most of the pauses have a length between 1.5 s and 2.5 s, therefore `sampleLength` is set to 1.5 s to get the most out of the LRS3 dataset. The extraction of positive and negative samples for the VVAD starts only on textual basis. Theoretically the whole extraction of the data could work on this basis but the given bounding boxes where very poor. To overcome this problem face detection and tracking was remade using dlib's (King, 2009) correlation filter based tracker and face detector. We provide for different kinds of features derived from the tracked face and facial features:

- **Face Images.** The whole image resized and zero padded to a specific size.
- **Lip Images.** An image of only the lips resized and zero padded to a specific size.
- **Face Features.** All 68 facial landmarks extracted with dlib's facial landmark detector
- **Lip Features.** All facial landmarks concerning the lips extracted with dlib's facial landmark detector

For the face images the input image only needs to be resized and zero padded to a given size. As depicted in Figure 2 the predictor extracts facial shape given by 68 landmarks, while 20 of these landmarks describe the lips. The predictor is trained on the ibug 300-W face landmark dataset ¹. For the *lip images* the minimal

¹Available at <https://ibug.doc.ic.ac.uk/resources/facial-point-annotations/>

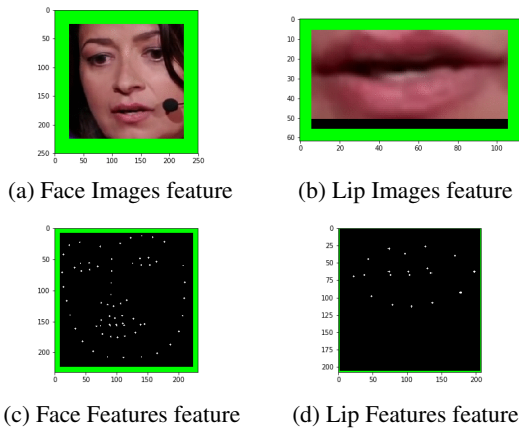


Figure 2: Visualization of one frame of different features.

Table 3: Dimensionality for the different features with **ts** as the number of timestamps, **d** as the image dimensions.

	ts	d	dtype
Face Images	38	200 × 200 × 3	uint8
Lip Images	38	100 × 50 × 3	uint8
Face Features	38	68 × 2	float64
Lip Features	38	20 × 2	float64

values in x- and y-direction are taken as the upper left corner of the lip image, while the lower right corner is defined by the maximal values in x- and y-direction. The *face features* are taken directly from the landmarks given from dlib’s facial landmark detector. For the *lip features* only landmark 49 to landmark 68 are taken into account, because they fully describe the lip shape as seen in Figure 2. It is to mention that it is useful to normalize the features for *face features* and *lip features* when applied to a learning algorithm.

With this approach we could create 22,245 negative (*not speaking*) samples and 22,244 positive (*speaking*) samples which is equal to 18.5 h of learning data in total. While the theoretical number of positive samples is way higher we were aiming for a balanced dataset and experimental results show that this is sufficient. Table 2 shows the number of samples on the training, validation and test sets. Figure 3 shows a random selection of 10 positive and negative images from the training set. Table 3 shows the dimensionality of one sample for the different features.

We evaluate two important hyper-parameters of our dataset to examine their relation with learning performance:

Image Size. The optimal image size is evaluated for MobileNets (Howard et al., 2017) using image sizes starting from 32×32 to the maximal image size of 200 with a step size of 32. Figure 4a shows that the maximal accuracy in the spatial domain can be

reached using a image size of around 160×160 .

Number of Frames. Figure 4b shows how accuracy improves over the number of frames for a *TimeDistributed* MobileNet on 96×96 pixel images (limited by available GPU memory). These results show that the VVAD task requires many frames for an accurate prediction and speaking cannot be inferred from a low number of frames.

Taking Figure 4b and 4a into account the optimal values for the image size and number of frames are 160×160 and 36 respectively.

Dataset Construction. To test the dataset with different models we created the following four features that are available directly on our dataset:

Face Images used for the most sophisticated model. These *Face Images* come in a maximal resolution of 200×200 pixels and with a maximal number of 38 frames. So the maximum shape of one sample of the *Face Images* feature is 38 frames × 200 pixels × 200 pixels × 3 channels = 4.56 MB. Pixel values range between 0 and 255 which can be represented with one byte.

Lip Images are also used for an end-to-end learning approach but they obviously concentrate on a small subset of the *Face Images*. *Lip Images* are RGB images with a maximum of 38 frames but they have a maximal resolution of 100×50 pixels. This resolves to 38 frames × 100 pixels × 50 pixels × 3 channels = 0.57 MB.

Face Features are used for the learning approach which focuses on facial features.

We provide 68 landmarks with a (x,y) position for a single face as depicted in Figure 2. A single feature is given as float64 (8 bytes), given by 38 frames × 68 features × 2 dimensions × 8 bytes = 41.4 KB.

Lip Features are a small subset of *Face Features* that only take the features of the lips into account. dlib’s facial landmark detector reserves 20 features for the lips as shown in Figure 2. This results in the size of 38 frames × 20 features × 2 dimensions × 8 bytes = 12.1 KB for a single sample in the lip features flavor.

Test Set and Human-Level Accuracy. To test the VVAD-LRS3 dataset a human accuracy test was performed. The test is built with a randomly seeded subset of 200 samples that is not part of the train/validation splits, and we used 10 persons to produce predictions for this set. The overall human accuracy level was 87.93%, while the human accuracy level on positive

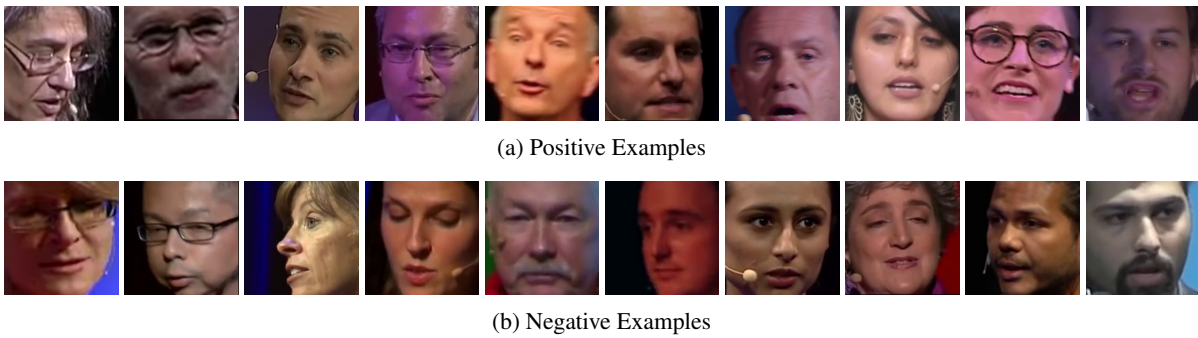
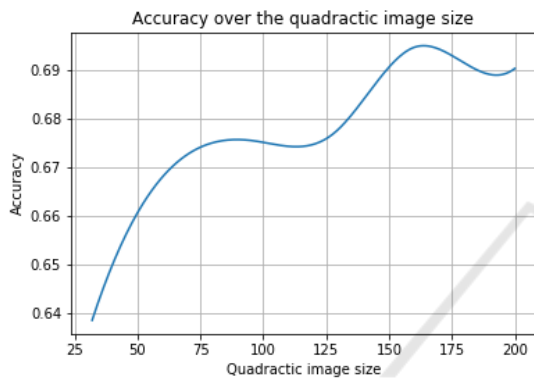
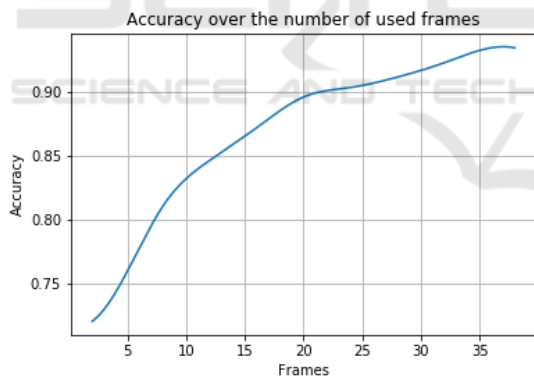


Figure 3: Random selection of speaking (positive) and negative (not speaking) samples from the VVAD-LRS3 dataset.



(a) Evaluation of the optimal image size with a single frame.



(b) Validation Accuracy for CNN LSTM over the number of timesteps/frames used.

Figure 4: Comparison of performance as image size and number of timesteps/frames is varied on MobileNet.

samples is 91.44%, and the human accuracy level on negative samples is only 84.44%.

This shows, that the automatic extraction of the negative samples is more prone to errors than the automatic extraction of positive samples. This is due to the purpose of the LRS3 as a lipreading dataset which obviously offers more positive samples than negative samples for a VVAD dataset.

In the human accuracy test some of the samples

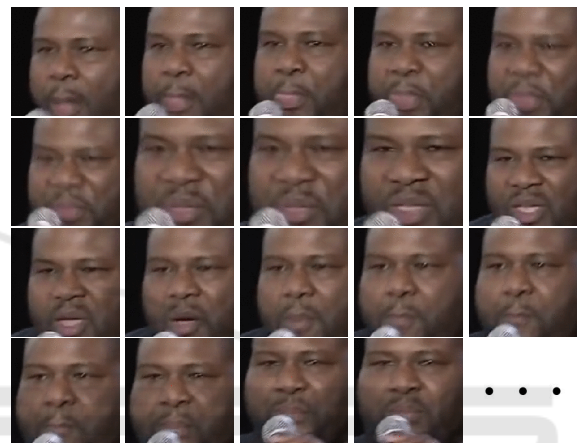


Figure 5: Sample 6178 is labeled as a negative (*not speaking*) sample by the automatic transformation from LRS3 to VVAD dataset. On the human accuracy level test 100% of the subjects classified the sample as positive (*speaking*) sample. Beat boxing is not considered speech in the LRS3 dataset, which causes the wrong label.

were labeled incorrectly or at least were classified with the opposite class label. A closer look is taken into four of these samples from the test set. While the samples 31366 and 42768 are labeled positive from the automatic transformation from the LRS3 sample to the VVAD sample they were classified as negative by all the subjects in the human accuracy level test. For the samples 14679 and 6178 the opposite is the case. On further investigation it was seen that sample 14679 and 42768 are obviously wrong labeled while sample 31366 and 6178 have some special properties that make them perform very bad on the human accuracy level test. Sample 31366 has a very quick head movement which makes it very hard to see the very little movements of the mouth to produce speech. Sample 6178 shows a person obviously producing sound with his mouth. But the sound here is no speech but *beat boxing* which is not considered speech in the original LRS3 dataset. Sample 6178 and 31366 are depicted in Figure 5 and 6 respectively.

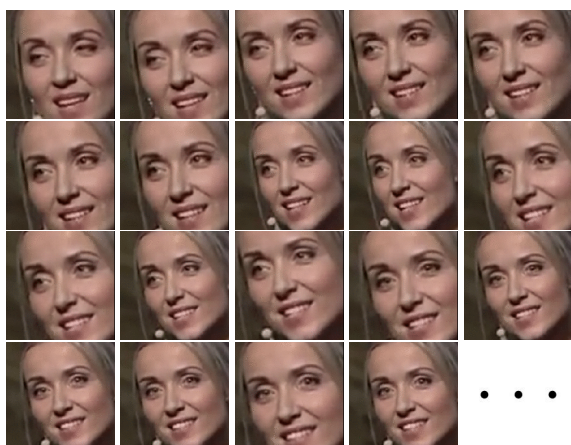


Figure 6: Sample 31366 is labeled as a positive (*speaking*) sample by the automatic transformation from LRS3 to VVAD dataset. On the human accuracy level test 100% of the subjects classified the sample as negative (*not speaking*) sample. The fast movement of the head while producing only a small movement of the lips causes the wrong label.

4 INITIAL EXPERIMENTAL RESULTS

Pre-trained Models. To show that the dataset can be efficiently used to train a VVAD, we implemented and trained CNN-LSTM models with our dataset as baselines. As described earlier speech cannot be effectively classified with a single image, which motivates the use of recurrent neural networks.

We evaluate the use of LSTM cells, as described in (Hochreiter and Schmidhuber, 1997). We use standard architectures as a backbone, which are wrapped by a *TimeDistributed* wrapper in order to transform them into a recurrent network that can process a sequence of images. *TimeDistributed* is a wrapper provided by Keras (Chollet et al., 2015) which basically copies a model for all timesteps, to effectively handle time series and sequences. The sequence can be processed by a LSTM Layer to make temporal sense, while the last Dense Layer is used to make the classification. Experiments have shown that a single Dense layer with 512 units on top of a LSTM layer with 32 units show good results. We use a 200×200 pixels input image size on one or two frames for initial testing.

We use DenseNet (Huang et al., 2018), MobileNet (Howard et al., 2017) and VGGFace (Parkhi et al., 2015) as backbone networks in the *TimeDistributed* wrapper. These models are pre-trained and used as is from the *keras-applications* library.

All models were trained using Stochastic Gradient Descent, with a starting learning rate $\alpha = 0.01$ and decaying as needed. Models were trained until con-

vergence, which varied between 80 to 200 epochs. A binary cross-entropy loss is used, and each network has an output layer with a single neuron and a sigmoid activation. All architectures and hidden layers use a ReLU activation.

Our results are presented in Table 4. It shows that DenseNet, MobileNet and VGGFace improve by around 2.3% using one more frame. Our results also shows that MobileNetV1 and DenseNet121 perform better than the corresponding model alteration. We will refer as MobileNet and DenseNet to MobileNetV1 and DenseNet121 respectively.

End-to-End Learning. In this section we evaluate end-to-end models trained from scratch, using not just face images but also other features such as lips and their features. Since evaluating for all 38 frames is not always possible (depending on access to GPUs with large amounts of RAM), only the MobileNet as the smallest of the base models is taken further into consideration. For this experiment we use 96×96 input image sizes for image features.

In comparison MobileNet contains approximately 4.2 million parameters while DenseNet requires around double the amount with 8 million parameters and VGGFace has over 50 million parameters. Knowing this the MobileNet is a good compromise between performance and size, because it is able to consider more timesteps, which in the end can lead to even higher accuracy.

For the face and lip images a *TimeDistributed* MobileNet is used, while the approaches learning on the vector features (facial and lip features) we use a single LSTM layer with 32 units and a single Dense layer with 512 units. Training methodology is the same as pre-trained models.

Our results are presented in Table 5 It shows that even with the substantially smaller face features a validation accuracy of 89.79% can be reached which is still higher than the human accuracy level. With this end-to-end learning approach on face images we were able to reach a very high accuracy of 92% on the test set. This is higher than the reported human-level accuracy on the same dataset.

One interesting remark from our results is that learning from image data, even if it is from scratch, seems to outperform the use of facial or lip features by approximately 3%, which we believe makes sense since an image might contain additional information that the pure facial or lip features do not contain. This shows the importance of using visual models for this problem.

Prediction Analysis. The classifications of the samples from the test set can be seen in Figure 7. The first 100 samples are negative samples while the last

Table 4: Validation accuracies for the different baseline models using only one or two frame in full resolution (200×200 pixels). Increased accuracy highlights the importance of the temporal domain in VVAD.

Baseline Model	1-Frame Acc	2-Frame Acc
DenseNet201	73.08 %	-
DenseNet121	73.17 %	75.34 %
MobileNetV2	67.45 %	-
MobileNetV1	69.56 %	72.11 %
VGGFace	71.96 %	74.36 %

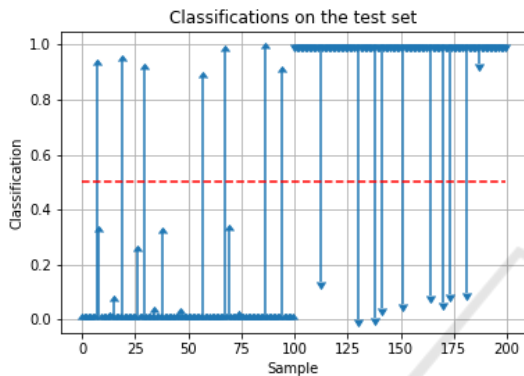


Figure 7: Visualization of predictions on the test set for MobileNet trained on face images. Each arrow represents the prediction confidence, with the first 100 samples being negative (not speaking), and the remaining 100 samples being positive (speaking).

100 samples are positive ones and the arrows show the probability, given by the model, that this sample belongs to the positive class. The red dashed line is the decision boundary on which the model decides its classifications. This visualizes how certain the model is with its predictions. Many predictions are incorrect with a high confidence, indicating overconfidence, which also motivates the use of properly calibrated and Bayesian neural network models (Matin and Valdenegro-Toro, 2020)

5 CONCLUSIONS AND FUTURE WORK

In this work we present the construction of the VVAD-LRS3 dataset using an automated pipeline to construct VVAD samples from LRS3 samples, we also show that these samples are not labeled perfectly, but they can still be used to learn a robust VVAD system. The VVAD-LRS3 dataset provides four kinds of features: facial and lip images, and facial and lip landmark features.

We provide baselines on our dataset using pre-

Table 5: Overview of the validation performance of the different features on MobileNet using all 38 frames at 96×96 pixels.

Feature	Validation Acc	Test Acc
Face Images	94.05 %	92.0 %
Lip Images	93.98 %	92.0 %
Face Features	89.79 %	89.0 %
Lip Features	89.93 %	89.0 %
Human Level	-	87.93 %

trained and end-to-end neural network architectures on all feature kinds. Face images with end-to-end architectures seem to perform best with a validation accuracy of up to 94%, while landmark features on face and lips seem to perform the worse at around 89% validation accuracy. We also show that up to 38 frames are required to obtain the highest predictive performance for this task.

Although the performance shows to be better than human accuracy and the presented solutions seem to be robust enough to handle outliers it may be possible to improve the results with a cleaned dataset. The cleaning can be done by manually testing all labels and correct or remove wrong labeled samples or by enhancing the algorithm to reduce the number of wrong labels.

Due to the comparability of the test results with the human accuracy level it was only possible to use the 200 randomly seeded samples used for the human accuracy test as the test set for the trained models, although it was described as best practice to hold back at least 10% of the data for testing. If the test set would be bigger and comparability to the human performance can be secured the test results would have an even stronger meaning than right now. A larger amount of samples that were tested on humans would make it possible to examine the relationship between DNNs for VVAD and the human brains approach to VVAD more closely. Furthermore it is hard to determine a ground truth for the data because human classification varied for some of the samples. But in general the human classification and the data creation through the automatic pipeline have a significant similarity which allows us to use the data effectively as is. Experiments with trained models in real human-robot interaction can hopefully be conducted in the future. We hope that the community benefits from our dataset and is able to produce learning algorithms that can produce a robust VVAD system for social robots.

The dataset is publicly available under <https://tinyurl.com/mucfmfyx>. With a large scale publicly available dataset for VVAD the research on this topic can be massively accelerated.

Furthermore we were able to publish some of the trained models on PyPI (PSF, 2022) under <https://pypi.org/project/vvadlrs3/> to make it easier to develop applications.

REFERENCES

- Aung, Z. H. and Ritthipravat, P. (2016). Robust visual voice activity detection using long short-term memory recurrent neural network. In *Revised Selected Papers of the 7th Pacific-Rim Symposium on Image and Video Technology - Volume 9431*, PSIVT 2015, pages 380–391, New York, NY, USA. Springer-Verlag New York, Inc.
- Chollet, F. et al. (2015). Keras. <https://keras.io>.
- F. Luthon, M. L. (1998). Lip motion automatic detection.
- gun Choi, J. and Kim, M. (2009). The usage and evaluation of anthropomorphic form in robot design. In *Undisciplined! Design Research Society Conference 2008*.
- Guy, S., Lathuilière, S., Mesejo, P., and Horaud, R. (2020). Learning visual voice activity detection with an automatically annotated dataset.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications.
- Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2018). Densely connected convolutional networks.
- Kanda, T. and Ishiguro, H. (2017). *Human-Robot Interaction in Social Robotics*. CRC Press.
- King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758.
- Matin, M. and Valdenegro-Toro, M. (2020). Hey Human, If your Facial Emotions are Uncertain, You Should Use BNNs! In *Women in Computer Vision @ ECCV*.
- Meriem Bendris, D. C. and Chollet, G. (2010). Lip activity detection for talking faces classification in tvcontent. *3rd International Conference on Machine Vision (ICMV)*, pages 187–190.
- Miwa, H., Okuchi, T., Itoh, K., Takanobu, H., and Takanishi, A. (2003). A new mental model for humanoid robots for human friendly communication introduction of learning system, mood vector and second order equations of emotion. In *2003 IEEE International Conference on Robotics and Automation (Cat. No.03CH37422)*, volume 3, pages 3588–3593 vol.3.
- Nikitina, A. (2011). *Successful Public Speaking*. bookboon.
- OZTOP, E., FRANKLIN, D. W., CHAMINADE, T., and CHENG, G. (2005). Human-humanoid interaction: Is a humanoid robot perceived as a human? *International Journal of Humanoid Robotics*, 02(04):537–559.
- Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In *British Machine Vision Conference*.
- Patrona, F., Iosifidis, A., Tefas, A., Nikolaidis, N., and Pitas, I. (2016). Visual voice activity detection in the wild. *IEEE Transactions on Multimedia*, 18(6):967–977.
- Patterson, E. K., Gurbuz, S., Tufekci, Z., and Gowdy, J. N. (2002). Cuave: A new audio-visual database for multimodal human-computer interface research. *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2:II–2017–II–2020.
- PSF, P. S. F. (2022). The python package index (pypi). Python package repository.
- Spyridon Siatras, N. N. and Pitas, I. (2006). Visual speech detection using mouth region intensities. *14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP*.
- Triantafyllos Afouras, Joon Son Chung, A. Z. (2018). Lrs3-ted: a large-scale dataset for visual speech recognition. In *arXiv:1809.00496v2 [cs.CV] 28 Oct 2018*.
- Zellner, B. (1994). Pauses and the temporal structure of speech.
- Ángel Pascual del Pobil Ferré, Bou, M. D., Anna Stenzel, Eris Chinellato, Markus Lappe, and Roman Liepelt (2013). When humanoid robots become human-like interaction partners: Corepresentation of robotic actions. page 18.