Proposal: Development of a Visual Voice Activity Detection

Adrian Lubitz

MASTERTHESIS

submitted in the
Master program

Systems Engineering

at the University of Bremen

December 2018

Contents

Chapter 1

Proposal

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 the shape of the human body or faces. Therefore humanoid robots are the most natural way for human-machine interaction, because of the human-like appearance. This hypothesis is strongly supported by Takayuki Kanda and Hiroshi Ishiguro in [10]. They see Social Robots as a part of the future society as shown in figure 1.1. Takayuki Kanda and Hiroshi Ishiguro also define 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.

Figure 1.1: Social Robots in the future society [10]

- c. Development of functions for interactions with people.

This thesis is located in the field c, as we try to implement a 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. The detailed description of the problem setting is given in Section 1.2.

1.1 Motivation

In this Section we want to present why a 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). To achieve this goal we need to understand how humans communicate. In the specific we need to know how humans start a conversation. There is one main concept when it comes to starting an interaction, which is deeply encoded in our DNA, as it is used by newborns and even animals follow this concept. To start a interaction of any kind one needs to increase the attention of the other. In HHI an increased attention comes mainly from the following stimuli:

- 1. auditory stimuli
- 2. tactile stimuli
- 3. visual stimuli

In the context of a conversation all the above stimuli can represent a direct starting point to a conversation. For auditory stimuli a direct starting point can be a name or a word socially known to increase attention like *hey* in the English language. Calling someones name works like a directed message to someone, while words like *hey* work like a broadcast to everyone in the surrounding. The concept of using names as a trigger phrase for virtual assistant systems is implemented in well known assistants like the Amazon Alexa, the Google Assistant or Apples Siri.

A tactile stimulus gives direct attention because the human body is very sensitive to tactile stimuli in a way that every individual has it's own private space. A tactile stimulus is invading a humans personal space and is therefore triggering an increase in attention. Meaning when you touch someone and say something directly afterwards the touched person can directly process the message, because of the increased attention.

For visual stimuli gaze and eye contact have been studied for decades, because they have such a special role in HHI. As Michael Argyle and Janet Dean stated in [5] gaze has a huge influence on the personal space and therefore also increases the attention. Meaning if eye contact is established and one of the two parties speaks it works like a directed message.

As the hypothesis of Takayuki Kanda and Hiroshi Ishiguro is that humans will react more naturally to human-like robots, it is obvious that not only the physical appearance needs to be human-like, also the psychological concepts of interaction need to be implemented.

For the purpose of this thesis we want to enhance the implementation of a Pepper Robot from SoftBank Robotics(a detailed description of the robot can be seen in Section

??).[1] The robot already is able to react to the described stimuli. Nevertheless there is a gap after that. Assuming someone is establishing eye contact with the robot. Which will result in an increase of the robots attention, meaning he will start listening. But the problem here is that the robot does not have the capability to detect if the person, eye contact is established with, is speaking. The robot may respond to something someone else in the surrounding was saying, which will confuse the person, eye contact is established with, and result in a non natural way of communication. To solve this problem we want to develop a VVAD which detects whether a person is speaking to a robot or not, given the visual input of the robot's camera. Thus the robot can decide whether to listen if the person is speaking to it or to proactively act on the invasion of it's personal space if the person is just starring at the robot. We hope that this social skill will improve the quality of HRI in a way that communicating with a robot feels more natural.

1.2 Problem Setting

As stated before this thesis deals with the implementation of a VVAD which detects whether a person is speaking to a robot or not, given the visual input of the robot's camera. We consider an open world setting in which people start an interaction with the robot whenever they want and it can also occur that people are talking near the robot without talking directly to the robot. In such an environment the robot should be able to react on a visual stimulus (described in Section 1.1) as natural as possible. The algorithm to classify whether a person is just starring at the robot or is speaking to the robot will be performed solely on video data from the camera of the robots head. This seems to be unnecessary few input on first sight, as the robot provides a lot more useful information, like audio input and even echo location. But with a closer look to the described cognitive feature of detecting an invasion of personal space on the basis of visual stimuli, it is only logical to reduce the input to only visual information. A fusion of different cognitive features could happen on a higher level of cognition. As described in Section 1.1 we use a Pepper Robot as the target platform. Pepper provides video data in a maximum resolution of 2592x1944 pixels at a framerate of 1 fps or with a maximum framerate of 30 fps with a resolution of 640x480 pixels(for further details of the given Hardware see Section **??** and [17]).

1.3 Approach

In this Section we will describe the approach to develop a VVAD which detects whether a person is speaking to a robot or not, given the visual input of the robot's camera. This goes from the training data over the learning algorithm to the evaluation.

1.3.1 Training data

To train our classifier we need a dataset which has labeled video data. As we try to solve a binary classification problem the data needs to have two labels. In our case these

[1] All further mentions of *the robot* refer to the Pepper Robot

would be *speaking* as the positive class and *not speaking* as the negative class on every frame of the video data. We also need a Region Of Interest (ROI) on every image. So we end up with a label for every image determining whether the person is speaking or not and a face bounding box as our ROI. The University of Oxford developed three datasets(LRW, LRS2, LRS3) for lipreading in cooperation with the BBC and TED [1], [2],[6]. These need to be slightly adjusted to the purpose of the thesis. The video data is labeled with words and the corresponding timestamps. It also provides a face bounding box for every frame. The positive class can be easily extracted from this datasets, as we know which frames correspond to speech. The negative class needs to be extracted from the phases where the person is not speaking, which can be slightly more challenging because the phases are not stored in the dataset's labels. As we know there is no speech in those phases we follow the given face from the ROI to construct a negative sample if the number of frames is over a given threshold.

1.3.2 Learning algorithm

To choose a good fit for the learning algorithm it is important to exploit the knowledge of the underlying data. In the case of a VVAD, we are dealing with sequences of video frames. As described in 1.3.1 every frame is annotated with a positive (*speaking*) or a negative (*not speaking*) label y and a ROI as a face bounding box from which we can extract a vector of features(these can be the pixels directly or features from the face detection) $\mathbf{x} = \langle x_0, x_1, \cdots, x_{n-1}, x_n \rangle$. What we want is a classifier that maps a sequence of those frames to a class.

$$f : \langle \mathbf{x}_t, \mathbf{x}_{t-1}, \cdots, \mathbf{x}_{t-k} \rangle \rightarrow y_t \quad | \; y \in \{0, 1\}, \; k \in \mathbb{N} \mid 0 < k \leq t \tag{1.1}$$

This classifier uses a temporal sliding window, to always perform the classification on the last $k + 1$ frames. The hyperparameter k needs to be chosen small enough to make the classification fast enough, but needs to be chosen big enough to secure accuracy. For $\mathbf{x}$ we will evaluate four different approaches. These can be split in 2x2 categories. The first categories are *Focus on facial features* or *End-to-End Learning*. The second categories are *only mouth* or *whole face*. This will end up in the following four models, which should be tested against each other:

- Focus on facial features
  - only mouth features
  - whole face features
- End-to-End Learning
  - only mouth image
  - whole face image

For the facial features we want to use dlib[16] pretrained Convolutional Neural Network (CNN) for face detection which gives 68 features for a face including 20 features for the mouth. Independent of which model we use, we will need to apply some kind of normalization to the scale, rotation and translation of the features. The face or the mouth can appear in any translation, rotation and scale in the dataset, with the normalization we make sure that we concentrate on the important features of the image. Otherwise it

could happen, that the classification is based on the position of the face in the image, which is obviously not correct. As we want to know the label for a ongoing sequence as fast as possible it is considered to be a problem of early classification of time series data [7]. As described earlier we use a fixed size temporal sliding window to solve this issue. Our Approach to solve the classification problem is a Long Short Term Memory (LSTM) Fully Convolutional Network (FCN). As shown in [4] [3] Recurrent Neural Networks (RNNs) and in specific LSTM-FCNs are the state of the art method for Time Series Classification.

1.3.3 Evaluation

To evaluate the classification we will use different test sets. One will be from the dataset mentioned in 1.3.1. To take a portion of the whole dataset to evaluate the classification is a standard approach in the field of data science. But since our real data, meaning the video data coming from the robot's camera, may differ a little, we also want to manually label a small test set consisting of videos from the robot camera. Further on we want to evaluate the four different approaches in terms of the input features described in 1.3.2. For the End-to-End approach we also consider making a visual evaluation using algorithms like Grad-CAM [14]. Another aspect is the tuning of the hyperparameter k, as it needs to be small enough to allow fast classification but big enough to sustain the accuracy.

1.4 Related Works

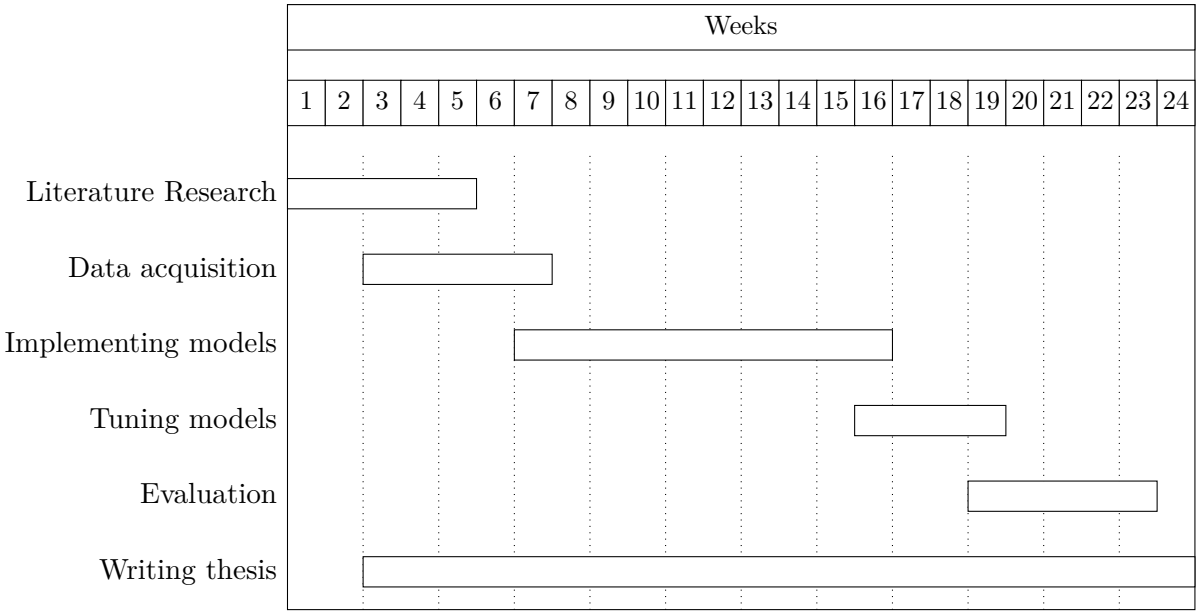
The classic approach to solve Visual Voice Activity Detection is to detect lip motion. This approach is taken by F. Luthon and M. Liévin in [9]. They try to model the motion of the mouth in a sequence of color images with Markov Random Fields. For the lip detection they analyze the images in the *HIS (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 [15]. 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 [11] 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 this paper the drawback of lip motion detection based approaches is already discussed. As shown in Figure 1.2 the problem 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. [12]. 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 [13]. 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 [8].

Figure 1.2: Example of error detection - Person is classified as having a mouth activity, however does not speak [11]

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 are 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.

1.5 Time table



References

Literature

[1] J. S. "Chung and A." Zisserman. “"Lip Reading in the Wild"”. In: *"Asian Conference on Computer Vision"*. "2016" (cit. on p. 4).

[2] J. S. "Chung et al. “"Lip Reading Sentences in the Wild"”. In: *"IEEE Conference on Computer Vision and Pattern Recognition"*. "2017" (cit. on p. 4).

[3] Houshang Darabi" "Fazle Karim Somshubra Majumdar. “"LSTM Fully Convolutional Networks for Time Series Classification"”. In: *"arXiv:1709.05206v1 [cs.LG] 8 Sep 2017"*. "2017" (cit. on p. 5).

[4] Houshang Darabi" "Fazle Karim Somshubra Majumdar. “"Multivariate LSTM-FCNs for Time Series Classification"”. In: *"arXiv:1801.04503v1 [cs.LG] 14 Jan 2018"*. "2018" (cit. on p. 5).

[5] Janet Dean" "Michael Argyle. “"Eye-Contact, Distance and Affiliation"”. *"Sociometry"* "28" ("1965"). "Issue 3", "289–304". URL: http://www.columbia.edu/~rmk7/HC/HC_Readings/Argyle.pdf (cit. on p. 2).

[6] Andrew Zisserman" "Triantafyllos Afouras Joon Son Chung. “"LRS3-TED: a large-scale dataset for visual speech recognition"”. In: *"arXiv:1809.00496v2 [cs.CV] 28 Oct 2018"*. "2018" (cit. on p. 4).

[7] Philip S. Yu" "Zhengzheng Xing Jian Pei. “"Early Classification on Time Series"” ("2011") (cit. on p. 5).

[8] Zaw Htet Aung and Panrasee Ritthipravat. “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. Auckland, New Zealand: Springer-Verlag New York, Inc., 2016, pp. 380–391. URL: http://dx.doi.org/10.1007/978-3-319-29451-3_31 (cit. on p. 5).

[9] M. Liévin F. Luthon. “Lip Motion Automatic Detection”. In: 1998 (cit. on p. 5).

[10] T. Kanda and H. Ishiguro. *Human-Robot Interaction in Social Robotics*. CRC Press, 2017. URL: https://books.google.de/books?id=2ghEDwAAQBAJ (cit. on p. 1).

[11] Delphine Charlet Meriem Bendris and Gerard Chollet. “Lip activity detection for talking faces classification in tvcontent”. *3rd International Conference on Machine Vision (ICMV)* (2010), pp. 187–190 (cit. on pp. 5, 6).

[12] F. Patrona et al. “Visual Voice Activity Detection in the Wild”. *IEEE Transactions on Multimedia* 18.6 (June 2016), pp. 967–977 (cit. on p. 5).

[13] Eric K. Patterson et al. “CUAVE: A new audio-visual database for multimodal human-computer interface research”. *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing* 2 (2002), pp. II-2017-II-2020 (cit. on p. 5).

[14] Ramprasaath R. Selvaraju et al. “Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization”. *CoRR* abs/1610.02391 (2016). arXiv: 1610.02391. URL: http://arxiv.org/abs/1610.02391 (cit. on p. 5).

[15] Nikos Nikolaidis Spyridon Siatras and Ioannis Pitas. “VISUAL SPEECH DETECTION USING MOUTH REGION INTENSITIES”. *14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP* (2006) (cit. on p. 5).

Software

[16] Davis King. *Dlib*. Python library. URL: http://dlib.net/ (cit. on p. 4).

Online sources

[17] SoftBankRobotics. *Technical overview of Pepper*. 2018. URL: http://doc.aldebaran.com/2-7/family/pepper_technical/index_pep.html (cit. on p. 3).