We Can "Hear" You With Radar

Daan Kerkhof *s2590611*
University of Twente

June 2023

Abstract

Detecting speech with a radar has several benefits over a regular microphone, i.e., the radar is tolerant to ambient noise. In addition to this, a radar based microphone could locate the exact location of the speaker, which is not possible with a single microphone. In this thesis we study the implementation of detecting speech with a Frequency Modulated Continuous Wave (FMCW) radar. A spectrogram of a viseme is extracted using the Short Time Fourier Transform (STFT). These visemes are the input of a Convolutional Neural Network (CNN), which can classify the visemes. Eventually, it was shown that there is uniqueness in the spectrograms of visemes. The proposed CNN has an accuracy of 70% in detecting the right visemes.

1 Introduction

This thesis is about using Frequency Modulated Continuous Wave (FMCW) radar to detect what a person is saying. The recent advancements of semiconductor devices and signal processing have improved the performance and reduced the price and size of FMCW radars [1]. FMCW radars are used in a wide variety of applications, i.e. weather forecasting, surveillance systems and automotive vehicles [1].

FMCW radars can achieve operating frequencies of extremely high frequencies (between 30 and 300 GHz). As the wavelength of these frequencies are in the order of millimeters, these radars are also called millimeter wave (mmWave) radars. Because of the small wavelengths, it is possible to detect small movements. Lots of research has already been done in monitoring vital signs (heart rate and breathing rate) [2], [3] and [4]. These sources show that it is possible to detect the small movements of a person's chest and retrieve the vital signs from this.

There are already studies showing possibilities in "hearing" with antennas. WiHear [5] is a system that utilizes WiFi signals to detect movements of the lips. WiHear can recognize a set of vowels and consonants. So, when a word is said, it combines the detected vowels and consonants into words. It is able to have an accuracy of 91% in detecting words on one person speaking not more than 6 words [5]. WiHear did not make use of FMCW radar, however there is research that does make use of FMCW for speech detection. In [6] and [7], a FMCW radar is used to detect the vibrations in vocal folds. When speaking, these vocal folds vibrate at different frequencies, which can be converted to speech.

The main question of this thesis is: *Is it possible to "hear" with radar?*. The goal is to aim the radar at the face of the user and convert the facial movements to speech. Facial movements could be the movements of lips, jaws, tongue, etc.

A system like this would be applicable in a scenario where there is a lot of ambient audible noise. An audio based microphone is based on sound waves, which are disturbed with this noise. A system which can detect speech by means of FMCW radar, has immunity to this audible noise, making them able to detect speech, even in such environments. In addition, the radar based system would be able to detect exactly where the sound was coming from. This is not possible with using a single microphone.

In this thesis, theory about FMCW radar and about visemes will be covered. Visemes are what the radar is going to detect. In the following section the processing methods will be explained. How the data is stored and what algorithms are applied to go from raw radar data into a spectrogram that can be used for classification. After that the results will be shown, followed with a discussion and conclusion.

2 Theory

2.1 FMCW Radar

In FMCW radar, the transmitted signal is frequency modulated by so-called chirps. These chirps increase the frequency linearly over time. In Fig. 1 both the transmitted and received chirps are displayed. When the signal gets reflected by a target, a frequency shift (f_D) and/or time shift (Δt) occurs. This signal is picked up by the receivers. From the frequency shift, the velocity of the target can be determined. This is due to the Doppler effect. From the time shift, the range can be determined. Since a greater distance

means more time is passed before the signal hits the target and reflects back to the receiver. Both the Doppler shift and time shift contribute to a combined frequency shift (Δf). The transmitter sends out both an in-phase and a quadrature-phase signal. This is necessary to determine the direction of the target. If only an in-phase signal was transmitted, the frequency difference will still be the same if the velocity would change sign. However, for a quadrature phase signal, there will be a phase shift of π relative to the in-phase signal, when this sign change occurs. Meaning the in-phase signal will lead or lag depending on if the target moves closer or away. We obtain the frequency difference (Δf), by mixing the transmitted and received signals. This difference is called the beat frequency. The beat frequency is important since it contains both information on the velocity and range.

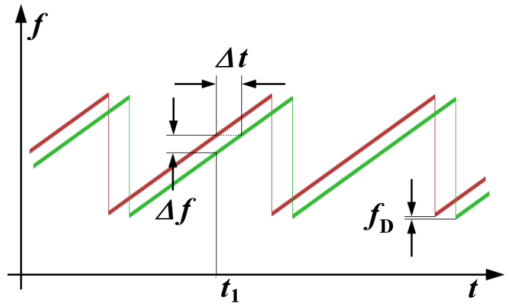


Figure 1: Time-frequency plot of chirps [8].

The chirps have certain parameters which can be tuned and adjusted to the scenario in which the radar is used. The chirp bandwidth (B) can be adjusted. This is the bandwidth of the linear sweep. This sweep happens with a frequency slope (S), which determines the rate in which frequency increases. The chirps duration (T_C) and the amount of chirps (N_C) can also be adjusted.

These parameters together influence the range resolution, maximum unambiguous range, velocity resolution and maximum unambiguous velocity of the radar. The formulas for these are given in (1) and (2) from [9]. Where λ is the wavelength, which is the speed of light divided by the operating frequency of the radar, $\lambda = c/f_c$. And, f_{max} is the maximum frequency supported by the radar, which is is 0.9 times the ADC sampling frequency (f_s) [9].

$$d_{res} = \frac{c}{2B} \quad \text{and} \quad d_{max} = \frac{f_{max} c}{2S} \tag{1}$$

$$v_{res} = \frac{\lambda}{2N_C T_C} \quad \text{and} \quad v_{max} = \frac{\lambda}{4T_C} \tag{2}$$

The radar module also has 4 transmitters and 3 receivers, creating a virtual array of 12 antennas. This creates the possibility to determine the angle of the target. In this thesis the virtual array will not be used to estimate the angle of the target. This is because the location and the angle of the target are already known Therefore the principle on the angle estimation will not be covered. However, it is convenient to be aware of the fact that there are multiple transmit-receive pairs from which the data can be processed.

The radar module used in this thesis is the Texas Instruments' IWR1443BOOST. It is a FMCW radar operating at a frequency ranging from 71 GHz up to 81 GHz [10]. The IWR1443BOOST is connected to Texas Instruments' DCA1000EVM capture card. This device captures and saves the data, such that it can be processed later on.

2.2 Visemes

We take inspiration from lip readers for speech detection. Lip readers make use of visemes. These are the visual representation of phonemes. Lips, jaws, tongue and other facial features will move differently for different phonemes. Visemes are common utilized in lip-reading detection techniques [11]. Note that there is no one-to-one relation between visemes and phonemes. There are more phonemes than visemes. Several phonemes can fit into one viseme. For instance, the viseme for pronouncing the letter *b* is identical to the letter *p*. Therefore it becomes a complex task to translate visemes back to words. There are algorithms available to convert a sequence of visemes into a sentence by making use of machine learning [11]. However, converting visemes to speech is outside the scope of this thesis.

Compared to the phonemes, for which an extensive library and proper documentation is available, visemes are less common and therefore have not much clear documentation. A vocabulary of visemes is created, based on [12]. These include the visemes and phonemes displayed in Table 1. For simplicity, not every viseme is used in this thesis. A selection has been made, which can be seen in Table 1.

Viseme	Phoneme	Sound	Used?
P	p, b, m	octo**b**er	✓
F	f, v	**f**inal	
T	t, d, s, z, th, dh	**t**ea	✓
W	w, r	**w**ater	✓
CH	ch, jh, sh, zh	**ch**ange	
K	k, g, n, ng, hh, y	**c**ase	✓
IY	iy, ih	b**i**t	
EH	eh, ae	**s**econd	✓
AA	aa	**g**arden	✓
AH	ah	gard**e**n	✓
ER	er	b**i**rd	✓
AO	ao	**o**ver	✓
UH	uh, uw	b**oo**k	

Table 1: Viseme Vocabulary.

There are more approaches to detecting what a person is saying. Another approach could be to monitor entire words, instead of each viseme. However, this would create a large database and make processing more complex. Visemes can be seen as building blocks of words, with a small set of visemes a lot of words can be made. Resulting in a larger range of words that can be detected.

Previous research shows that it is possible to detect sections of words and combine them into words. In [5], a system was created that detects consonants and vowels and converts those into words. This shows that there are possibilities in detecting smaller portions of words, and converting them into words.

3 Methods

The following section covers the processing of the data. The measurements itself are done using the mmWave Studio software from Texas Instrument. In this program, the radar can be connected and the parameters can be set. After doing the measurements in mmWave Studio, the data is exported to a *.bin* file and processed in MATLAB. In Fig. 2 the pipeline of the data processing is displayed.

3.1 Radar Cube

Every chirp has its own amount of samples. This amount is a parameter that can be adjusted. Meaning for one chirp, the received data is a one dimensional array of these samples. However, there is more than one chirp used in a frame. The samples of all different chirps are then stored in a two dimensional array. And since there are multiple transmit-receive pairs because of the virtual array, which all have their own chirps sequence, the final data is stored in a three dimensional array. The size of this three dimensional array is determined by the number of ADC samples per chirp, the number of chirps, and the amount of transmit-receive pairs. This way of arranging data is referred to as a radar cube. In Fig. 3 a radar cube is displayed. The dimension over which all the samples per chirp is given is called the fast time, the dimension for each chirp is called the slow time and the sensor space is for each transmit-receive pair. A MATLAB code will transform the *.bin* file from mmWave Studio into a radar cube.

The radar sends out these bursts of chirps in frames. The duration of the frames (T_f) and the amount of frames (N_f), have influence on how long the measurements goes on. Important is that each frame gets its own radar cube.

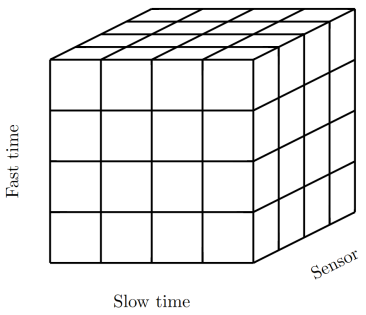


Figure 3: Radar cube dimensions

3.2 Locating The Target

After the radar data is transformed into radar cube, the target needs to be located. The way to do this is by making a range profile and implementing a Constant False Alarm Rate (CFAR) detection. For this, some assumptions are made. First of all, the target is speaking right in front of the camera with his or her mouth at the height of the antennas. This makes sure that the target is in the region of the antenna where it has the most gain. This results in a strong received signal. With these assumptions it is possible to find the target using a one dimensional range profile.

To get the range profile, we need to take the FFT of the fast time axis of the radar cube. After this, we choose one transmit-receive pair from the sensor space. Then, the average over the slow time will be taken resulting in a one dimensional range FFT.

To locate the target, a cell averaging CFAR detector is used. What CFAR does is taking a cell, often called the cell under test (CUT), and comparing it to a threshold. If it exceeds this threshold, a detection is made. However, this threshold is not a fixed value. In cell averaging CFAR, the threshold is based on the average noise power of the cells around the CUT. By setting a training window size and guard window size, the cells of which the noise power will be calculated can be selected. The training window size determines how many cells around the CUT will be used. However, the guard cells directly next to the CUT will be neglected to remove interfering from the power of the CUT. An schematic of this is given in Fig. 4.

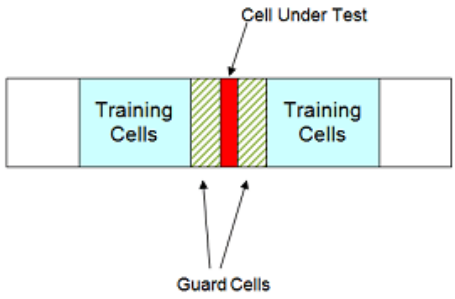


Figure 4: Cell averaging CFAR [13]

The average noise power for the CUT is given with (3), where x_m are the training cells.

$$P_n = \frac{1}{N} \sum_{m=1}^{N-1} |x_m|^2 \tag{3}$$

The algorithm goes through every cell from the range FFT as a CUT, and compares the noise power to a set threshold. If this threshold is exceeded, a detection is made.

3.3 Short Time Fourier Transform

We locate the range bin of the target for every frame and concatenate the slow time arrays of all the frames.

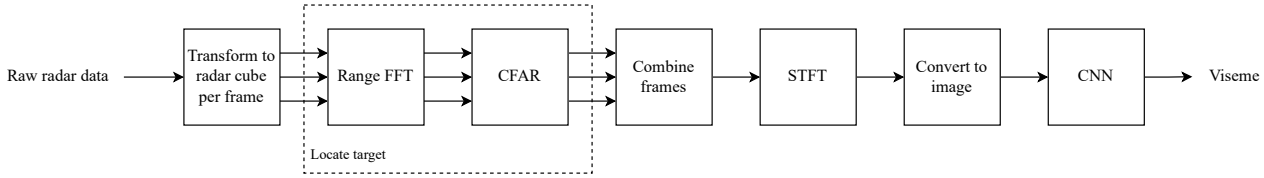


Figure 2: Pipeline of the data processing

This results in a slow time axis of the entire measure time. Now the data needs to be processed in such a manner that different visemes can be distinguished. Movements in the face cause frequency shifts in the received signal of the FMCW radar. Therefore we use the Short Time Fourier Transform (STFT) to capture these movements. The advantage of the STFT over the FFT, is that it shows the time at which the frequency shifts occur. We chose the STFT over the FFT since not all the visemes take the same amount of time. One may take a few milliseconds longer than the other. The STFT shows therefore more information than a FFT.

$$X[m,\omega] = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{j\omega n} \tag{4}$$

In (4), the mathematical formulation for the discrete STFT is given. Where $x[n]$ is the input signal and $w[n]$ is a windowing function. The window functions is centered around a certain time m, which means only the frequency spectrum around this time instance gets computed. Then the spectrum is computed, the window shifts forwards in time a little and then another spectrum is calculated. This procedure is then repeated until there are spectra of the entire time duration of the signal. When all the spectra are calculated, they are combined into a spectrogram. In Fig. 5 the principle of the STFT is shown.

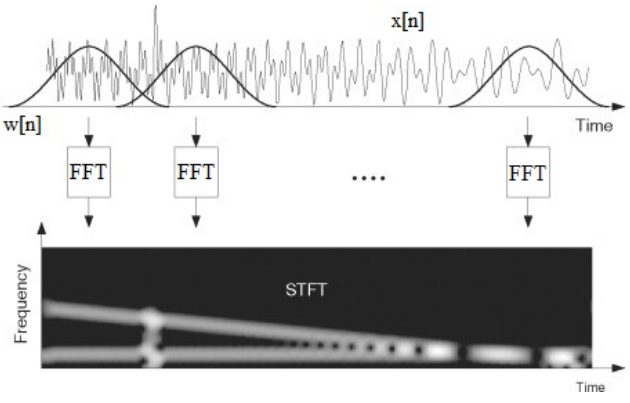


Figure 5: Principle of the STFT [14].

The final spectrogram shows the magnitude of the FFT for different frequencies and times. Now we can see at what times the frequency changes, giving us more information on the actual movement of the face compared to using only one FFT.

3.4 Radar Parameters

For the detection of the small lip movements a high range resolution is necessary. The range resolution determines how close two objects can be together, whilst still being two distinguishable targets. Since the lips of a human are closely spaced, we need the highest range resolution possible. In addition, we want a high resolution STFT spectrogram. Therefore, the slow time axis of the radar cube needs to be long. This can be achieved by having many chirps per frame with many samples. Taking these two requirements into account resulted in the parameters displayed in Table 2. The number of frames is not given in Table 2, since it only determines the total time duration of the measurement, which may vary depending on what is being measured.

Parameter	Value
Frequency slope	39.976 $MHz/\mu s$
Idle time	200 μs
ADC start time	6 μs
ADC samples	280
Sample rate	3 $Msps$
Ramp end time	100
Bandwidth	3997.6 MHz
Number of frames	600
Frame duration	80 ms
Number of chirps	255

Table 2: Used FMCW radar parameters.

These parameters yield a range resolution of 3.75 cm and an unambiguous maximum range of 10.1 m. As for the velocity, the resolution is 0.0254 m/s and the unambiguous maximum is 3.2445 m/s.

3.5 Convolutional Neural Network

After the measurements with the radar have been performed, we need to have a method to check which viseme corresponds to the STFT spectrogram. To do this, a Convolutional Neural Network (CNN) is used. CNNs are commonly used in image recognition. And after training the network, it can classify images and attach a label to it. One of the many advantages of a CNN is that it is tolerant to translations or rotations of the image [15]. This makes it ideal to be used for spectrogram classification.

The CNN is a process that consist of several layers. An overview of the CNN architecture is given in Fig.

6. It starts with an input layer and ends with an output layer. In between, there are multiple hidden layers. The hidden layers contain a repetition of a convolution filter, a Rectified Linear Unit (ReLU) layer and a pooling layer. The convolution layer convolves the input. It does this by having a adjustable filter and applying that on the input. The layer does not apply the filter directly on the entire image. It divides the image up into smaller sections and then applies the filter on those sections. This is why the CNN is tolerant to translations and rotations as mentioned before. The purpose of the convolutional layers is to extract features. Then, the data gets centered and normalized in a normalization layer. After this, there is a ReLU layer, which makes all the negative values equal to zero. This layer is also called an activation layer. This decreases the computation necessary. Next, the data goes through a pooling layer. In this layer the data gets downsampled, which reduces the computational effort. These layers (convolutional, normalization, ReLu and pooling layer) are repeated several times before going to the output layer. In this layer there will be classifications made with a fully connected layer and a softmax layer. Resulting eventually in probabilities of the features present in the image.

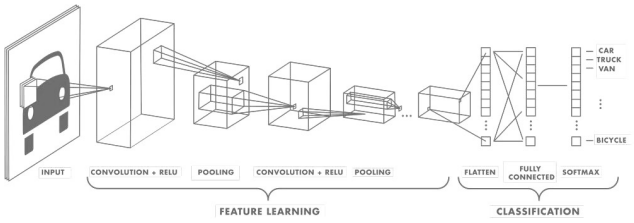


Figure 6: CNN Architecture [15].

After the CNN is setup, it needs training before it can classify images. There are multiple training options. One which can be used for spectrogram classification is the stochastic gradient descent [16]. With this method, parameters such as the learn rate can be chosen. This value determines how much the network should change if it occurs an error during training. Also the amount of epochs can be specified. An epoch is the amount of times the network runs through the training data.

3.6 Implementation

For the range FFT, a length of 256 was used. The CA-CFAR was implemented using the phased array toolbox from MATLAB. The training window was set to 32 cells, with a guard window of 8 cells. The noise power threshold was set to 18 dB. These values are based on how the mmWave studio typically processes the data. For the STFT, a periodic Hann window is used with a length of 450 samples. These windows overlap with 400 samples. The length of the FFT is 2048.

The setup for the CNN algorithm in this thesis is based on [16] and [17]. The used CNN in this thesis has 4 convolutional layers, 4 pooling layers, 4 ReLU activation layers and batch normalization. The kernel size of each convolutional layer is 3 by 3. The kernel numbers of the convolutional layer are 8, 16, 32 and 64 respectively. After each convolutional layer there is a maximum pooling layer with a 2 by 2 size and a stride of 2. The input is an image of size 656 by 875, which is the size of the spectrogram when exporting it as an image in MATLAB.

The CNN has been trained with a dataset. This dataset contains the STFT spectrograms of the 9 different visemes from Table 1. For each viseme, a measurement has been made where every 4 seconds, a viseme is pronounced. The person pronouncing the viseme was in front of the radar at about 80 centimeters and on the same height as the antennas of the radar. The STFT spectrogram is then divided into smaller sections such that every plot contains one viseme. In total every viseme has been repeated 90 times. Since there were measurement for 9 visemes, the total dataset is 810 images. From the 90 images per viseme, 70 of them are for training. The remaining 20 images are for validation. The network itself is trained using the stochastic gradient descent algorithm with an initial learn rate of 0.01 and the maximum amount of epochs is 20.

4 Results and Analysis

4.1 Separate Visemes

In Fig. 7, the STFT spectrograms of three of the nine visemes are displayed. On the horizontal axis, the time is displayed. The entire measurement took 204 seconds in total. The plot shows for clearness only a section of this total time. The frequency is displayed on the vertical axis. The movements of the face cause changes in frequencies, which can be seen in the plots. The color indicates the magnitude of an included frequency. The horizontal yellow line in the center indicated the non moving parts, which is present during the entire measurement, because non-moving objects do not contribute to a change in frequency.

As it can be seen in Fig. 7, there is a repeating pattern in the spectrogram for every viseme. This pattern is unique for every viseme. This shows indeed possibility in classifying visemes. However, not all the viseme patterns are as distinguishable. In Fig. 8, the visemes for AH and EY are shown. These visemes both have a big peak going down with a smaller peak going up at the same time, followed by a smaller peak going up. Because of this, they are hard to distinguish by observation.

The frequency range of all the visemes is roughly between -600 and 600 Hz. The lower frequencies come from the facial movements of moving the lips, jaws, tongue, eyebrows. These move at relatively slow velocities. The higher frequencies come from the vocal folds in the throat. This results matches with the results of [6], in which vocal folds vibrations are extracted using FMCW radar. These frequencies were between -500

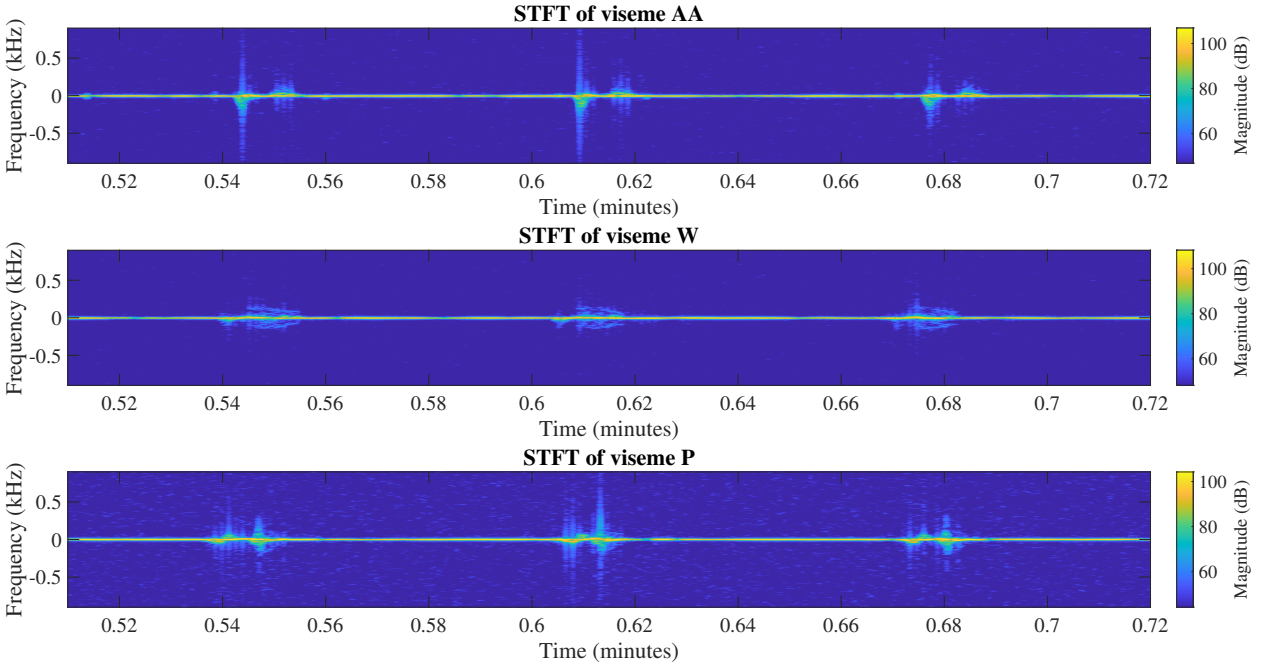
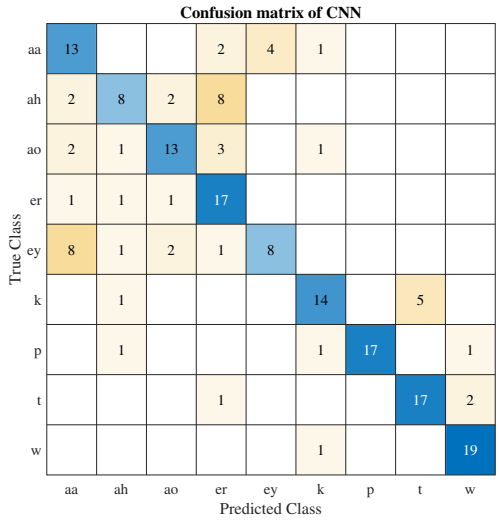
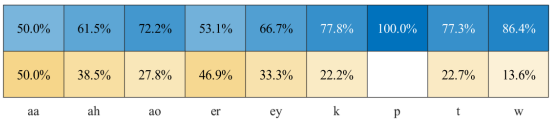


Figure 7: STFT for different visemes.

and 500 Hz.



(a) Confusion Matrix of CNN.



(b) Accuracy of the CNN.

Figure 9: Validation of the CNN.

The separate visemes have been put through the CNN with a testing set. The results of this are plotted in the confusion matrix in Fig. 9a. The total accuracy of the CNN was 70%. The accuracy is lower than the accuracy of a CNN for gesture recognition. In [16], they have an accuracy of around 90%. However, since gestures with hands and arms are more distinct and unique movements compared to facial movements. Therefore 70% is a good accuracy. In Fig. 9a, the rows of the matrix corresponds the true viseme and the columns to the predicted viseme. The numbers on the diagonal show correct classifications and the numbers not on the diagonal are incorrect classifications. Fig. 9b gives the precision per viseme. The visemes that were overall the most distinguishable were the AO, P, T, K and W. The least distinguishable visemes were AA, AH, ER and EY. There are a few weak points in the classification, where there is a high amount of incorrect classifications. One of them is the viseme for ER getting classified as AH (8 times). This is explainable, since these visemes correspond both to almost the same sound. The AH was pronounced as the *e* in garden and the ER as the *i* in bird. Therefore, it is no coincidence that the algorithm has trouble with classifying these visemes. Another weak point is the AA viseme getting classified as an EY viseme (8 times). These visemes belong to a different phoneme, but they appear to have a similar STFT pattern. As it can be seen in Fig. 8. They both have a big peak going down first, with a smaller peak going up at the same time. Followed by a small peak going up.

4.2 Combining Visemes Into Words

In addition to analyzing single visemes, the detection of words was also investigated, to see if it was possible to split up a word into its visemes and deduce what word is said. In Fig. 10a, the spectrogram for the word *time* is given. The white lines indicate the the end and start of the visemes, which have been

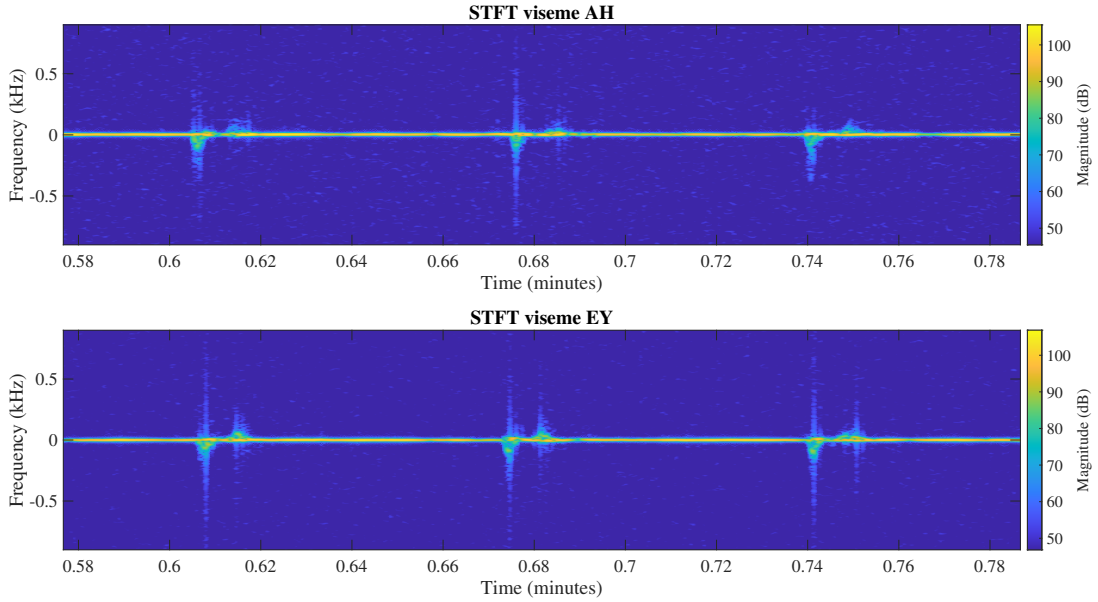


Figure 8: STFT of viseme for AH and EY.

placed in manually. The visemes of the word *time* are, according to [12], T-AH-P. These visemes have been displayed in Fig. 10b. After inspection, the resemblance of the viseme for T and AH from Fig. 10b can be seen in Fig. 10a. For the T, there is a small peak going down first and then one peak of approximately the same size going up. The AH viseme has a peak going downwards followed by a smaller peak going up. However, the viseme for P is a little different. In Fig. 10b, this viseme has one peak going up and one going down of about the same amplitude at the same time. Whereas the P in Fig. 10a only has a significant peak going down. The peak going up is less distinct. This can be explained by the fact that the letter pronounced is a *m*. In [12], the pronunciation of the *m* is classified under the viseme P (see Table 1). However, the database of the viseme P is filled with someone repeating the letter *p*, instead of the *m*. The pronunciation of *m* has apparently different facial movements compared to the pronunciation of *p*, which results in a different spectrogram pattern.

The spectrogram of the word in Fig. 10a has been separated into three images indicated by the white lines. The CNN classified these images with the following visemes: T-AH-AA, meaning the CNN deals with the same problem.

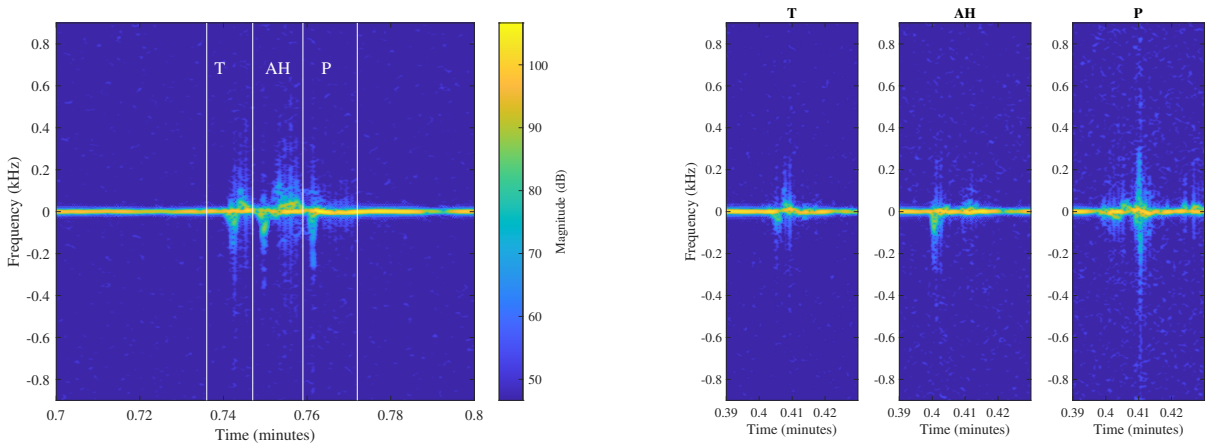
5 Discussion

There are still improvements to be made. The first part that can be improved is the vocabulary. The concept of using visemes might not be accurate enough. The main motive for choosing visemes was because the radar was supposed to do something similar to lip reading. However, as it turned out, the radar captures more than just lip movements. Movements of the jaws, eyebrows and vocal folds, give information beyond just lip movements. Meaning the viseme library might come short. An example of this was seen in the conclusion, where the letter *m* from the word time did not correspond with its viseme. The reason for this is probably that the lips do correspond to approximately the same movements, but the rest of the face does not. Therefore it might be better choice to base the vocabulary on phonemes, instead of visemes.

To increase the accuracy of the CNN, better parameters need to be chosen. The parameters for the networks were inspired on two image recognition networks [15] and [16]. There was some trial and error in seeing which parameters worked best, however more testing and a deeper understanding of the CNN might yield a higher accuracy. More training data would also increase the accuracy.

A final improvement that could be made to the system is regarding its robustness. In this thesis, the location of the target was given, which makes the detection of the target relative easy. In addition to this, the target was in a convenient position right in front of the radar and not that far away. As mentioned briefly in Section 2.1, the used radar has an array of antennas. This feature allows the antenna to distinguish the angle of targets and makes it possible to process multiple targets at once [18]. Combining this with beamforming, which steers the aperture of the antenna, the system could also locate the targets itself.

Next to this, all the patterns of the visemes were measured on a non-moving target. If there were body movements from the target, it may cause interference in the spectrograms. Therefore the system can become more robust if a body movement cancellation was applied. This filters out body movements making the system usable in an environment in which the target has movement [3]. Thus, applying the virtual array



(a) STFT of the word *time*.

(b) STFT the visemes of times.

Figure 10: STFT of time and its visemes.

for angular separation, beamforming for target detection and body movement interference removal makes the system overall more robust.

6 Conclusion

In this thesis, it was investigated whether it is possible to "hear" what someone is saying by using FMCW radar. The parameters of the radar were set up in such a way that it could detect the small movements of the face. This was achieved by having a high range resolution and high sample rates. After obtaining the location of the target by applying cell averaging CFAR on the range FFT, the corresponding range bin is processed with a Short Time Fourier Transform. This resulted in a spectrogram of the facial movements. A vocabulary of visemes was used to distinguish the different pronunciations, meaning that every viseme has its own spectrogram pattern. These spectrograms were input for a Convolutional Neural Network, which after training, could classify 70% of the visemes correctly. Also it was shown that a word can be separated into its visemes and a CNN can be applied on these visemes. Further improvements are necessary to refine the system. Such as a more extensive vocabulary and a more accurate CNN. Nevertheless, this thesis shows potential in detecting speech with FMCW radar.

References

[1] D. Rogriguez, C. Li, and M.Mercuri, "Recent advances of fmcw-based radar sensors," 2023.

[2] Z. Xu, C. Shi, T. Zhang, S. Li, Y. Yuan, C.-T. M. Wu, Y. Chen, and A. Petropulu, "Simultaneous monitoring of multiple people's vital sign leveraging a single phased-mimo radar," *IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology*, vol. 6, no. 3, pp. 311–320, 2022.

[3] T. K. Vo Dai, "Remote human vital sign monitoring using multiple-input multiple-output radar at millimeter-wave frequencies," 2022.

[4] M. Alizadeh, G. Shaker, J. C. M. De Almeida, P. P. Morita, and S. Safavi-Naeini, "Remote monitoring of human vital signs using mm-wave fmcw radar," *IEEE Access*, vol. 7, pp. 54 958–54 968, 2019.

[5] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We can hear you with wi-fi!" in *Proceedings of the 20th annual international conference on Mobile computing and networking*, 2014, pp. 593–604.

[6] K. Han and S. Hong, "Vocal signal detection and speaking-human localization with mimo fmcw radar," *IEEE Transactions on Microwave Theory and Techniques*, vol. 69, no. 11, pp. 4791–4802, 2021.

[7] H. Li, C. Xu, A. S. Rathore, Z. Li, H. Zhang, C. Song, K. Wang, L. Su, F. Lin, K. Ren *et al.*, "Vocalprint: A mmwave-based unmediated vocal sensing system for secure authentication," *IEEE Transactions on Mobile Computing*, vol. 22, no. 1, pp. 589–606, 2021.

[8] Dipl.-Ing. f. C. Wolff, "Radartutorial," Jun. 2023, [Online; accessed 26. Jun. 2023]. [Online]. Available: https://www.radartutorial.eu/02.basics/Frequency%20Modulated%20Continuous%20Wave%20Radar.en.html

[9] TI, "Programming chirp parameters in ti radar devices," 2017.

[10] TexasInstruments, "Iwr1843 single-chip 76- to 81-ghz fmcw mmwave sensor." [Online]. Available: https://www.ti.com/product/IWR1443

[11] S. Fenghour, D. Chen, K. Guo, B. Li, and P. Xiao, "An effective conversion of visemes to words for high-performance automatic lipreading," *Sensors*, vol. 21, no. 23, p. 7890, 2021.

[12] S. Lee and D. Yook, "Audio-to-visual conversion using hidden markov models," in *PRICAI 2002: Trends in Artificial Intelligence: 7th Pacific Rim International Conference on Artificial Intelligence Tokyo, Japan, August 18–22, 2002 Proceedings 7*. Springer, 2002, pp. 563–570.

[13] "FalseAlarmRateForCFARDetectorsExample," Jun. 2023, [Online; accessed 26. Jun. 2023]. [Online]. Available: https://nl.mathworks.com/help/phased/ug/constant-false-alarm-rate-cfar-detectors.html

[14] N. Kehtarnavaz, "Chapter 7 - frequency domain processing," in *Digital Signal Processing System Design (Second Edition)*, N. Kehtarnavaz, Ed. Burlington: Academic Press, 2008, pp. 175–196. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123744906000076

[15] "What Is a Convolutional Neural Network? | 3 things you need to know," Jun. 2023, [Online; accessed 23. Jun. 2023]. [Online]. Available: https://nl.mathworks.com/discovery/convolutional-neural-network-matlab.html

[16] W. Jiang, Y. Ren, Y. Liu, Z. Wang, and X. Wang, "Recognition of dynamic hand gesture based on mm-wave fmcw radar micro-doppler signatures," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4905–4909.

[17] "Train A Basic Convolutional Neural Network For Classification Example," Jun. 2023, [Online; accessed 24. Jun. 2023]. [Online]. Available: https://nl.mathworks.com/help/deeplearning/ug/create-simple-deep-learning-network-for-classification.html

[18] Z. Xu, C. Shi, T. Zhang, S. Li, Y. Yuan, C.-T. M. Wu, Y. Chen, and A. Petropulu, "Simultaneous monitoring of multiple people's vital sign leveraging a single phased-mimo radar," *IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology*, vol. 6, no. 3, pp. 311–320, 2022.