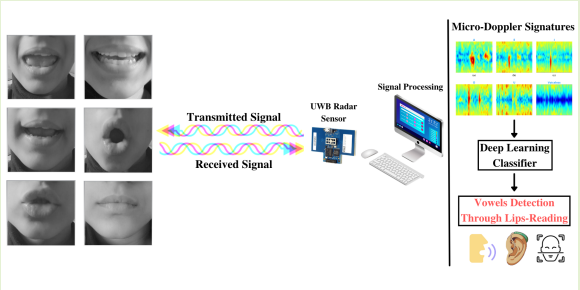
Extracting Visual Micro-Doppler Signatures from Human Lips Motion Using UoG Radar Sensing Data for Hearing Aid Applications

Umer Saeed, Syed Aziz Shah, Yazeed Yasin Ghadi, Muhammad Zakir Khan, Jawad Ahmad, Syed Ikram Shah, Hira Hameed and Qammer H. Abbasi, *Senior Member, IEEE*

***Abstract*— This study proposes a secure and effective lips-reading system that can accurately detect lips movements, even when face masks are worn. The system utilizes radio frequency (RF) sensing and ultra-wideband (UWB) radar technology, which overcomes the challenges posed by traditional vision-based systems. By leveraging deep learning models, the system interprets lips and mouth movements and achieves an overall accuracy of 90% for both mask-on and mask-off scenarios. The study utilized a trusted dataset from the University of Glasgow (UoG), consisting of spectrograms of lips motions stating five vowels and a voiceless class from distinct participants. The cutting-edge deep learning algorithm, Residual Neural Network (ResNet50), was used for the evaluation of the dataset and achieved an 87% accurate detection rate with a mask-on scenario, which is a 14% improvement compared to prior published work. The findings of this study contribute to the development of a robust lips-reading framework that can enhance communication accessibility in applications such as hearing aids, voice-controlled systems, biometrics, and more.**



***Index Terms*— ResNet50; InceptionV3; VGG16; RF sensing; UWB radar; lips-reading; speech recognition.**

I. Introduction

A potential technique that provides significant benefits over conventional methods for speech recognition is radio frequency (RF) sensing. Speech recognition using RF sensing is better for applications that require privacy and convenience than optical-based systems since it does not need a line of sight and can pass through barriers such as walls or masks [1]. Moreover, RF sensing can provide precise and reliable speech recognition in noisy or busy circumstances where other methods would not work. Several industries, including healthcare, smart homes, security and the military, are likely to greatly benefit from this technology. Without cameras or microphones, which can cause privacy problems, it can offer hands-free communication in hospital settings or speech recognition in smart homes [2], [3]. RF sensing-based speech recognition can improve the usability of the technology for those with hearing or speech problems.

Hearing aids are essential tools for people with hearing loss. However, traditional hearing aids have certain limitations. They can intensify background noise, making it difficult for users to concentrate on discussions [4]. Furthermore, in loud surroundings, hearing aids are often ineffective. Researchers are investigating cutting-edge wireless technologies and artificial intelligence to create the next generation of hearing aids to solve these problems [5]. The ability to hear sounds at a volume of 20 dB or above is referred to as normal hearing. Hearing disorders severely hamper effective communication and learning and it is estimated that 700 million people will have hearing impairments by 2050 [6]. In the United Kingdom, there are 11 million people with hearing disability and age associated hearing loss is becoming a major issue. This calls for disruptive multi-modal processing that is not constrained by speech or sound augmentation requirements. Humans need more than just sound to understand spoken words, which means the use of optical information.

An essential component of speech recognition is the use of lips-reading as an optical cue. However, privacy issues arise when hearing aid cameras collect optical data [7]. The legal effect of such devices alone would discourage their broad

Umer Saeed and Syed Aziz Shah are with the Research Centre for Intelligent Healthcare, Coventry University, Coventry CV1 5FB, UK. e-mail: saeedu3@uni.coventry.ac.uk; syed.shah@coventry.ac.uk

Yazeed Yasin Ghadi is with the Department of Computer Science, Al Ain University, Abu Dhabi P.O. Box 112612, United Arab Emirates. e-mail: yazeed.ghadi@aau.ac.ae

Muhammad Zakir Khan, Hira Hameed and Qammer H. Abbasi are with the James Watt School of Engineering, University of Glasgow, Glasgow G12 8QQ, UK. e-mail: m.khan.6@research.gla.ac.uk; hira.hameed@glasgow.ac.uk; qammer.abbasi@glasgow.ac.uk

Jawad Ahmad is with the School of Computing, Edinburgh Napier University, Edinburgh EH11 4BN, UK. e-mail: j.ahmad@napier.ac.uk

Syed Ikram Shah is with the College of Electrical and Mechanical Engineering, National University of Sciences and Technology, Islamabad 44000, Pakistan. e-mail: syed.shah15@ce.ceme.edu.pk

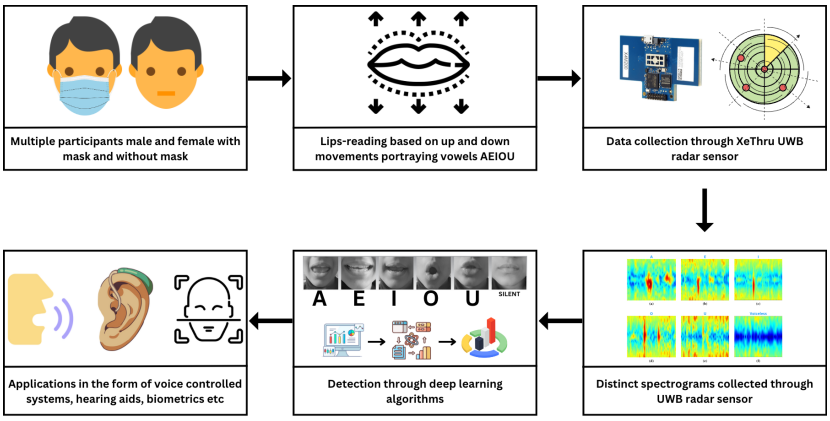


Fig. 1: Block diagram of the proposed scheme. A system using radar technology interprets lips and mouth movements with deep learning models. The implementation of this system has the potential to enhance communication accessibility in various devices such as hearing aids, voice-controlled systems, biometric devices, and other similar technologies.

usage in private and public settings since it is against the law to record someone without their consent in many places around the world. Modern hearing aids with optical information have limitations, with the face mask used during the COVID-19 pandemic period being a significant one [8]. RF sensing can be used to monitor the movements of the lips and mouths, helping to meet the need for next-generation hearing aids. Lips-reading using RF detection can provide hearing aids with precise indications for distinguishing spoken sounds and finding speech patterns utilizing deep learning and machine learning methods that have been effectively used in the past for different applications [9]. In contrast to visual-based systems, RF-based lips-reading is unaffected by face mask restrictions. RF signals can pass through the mask to pick up optical cues like lips and mouth movements that optical hearing aids would normally miss. This offers a unique chance to include RF sensing in next-generation multi-modal hearing aids, which could only need one antenna to be added to the device [10].

Lips-reading is gaining popularity due to its uses in interacting with the deaf population and biometric identification. It has also been investigated in the contexts of voice augmentation and visual speech recognition. Nevertheless, camera-based lips-reading systems have disadvantages, including privacy problems, inadequate illumination and trouble with face masks [11]. RF sensing is presented as a remedy to these restrictions, especially in the period of COVID-19 when face coverings are prevalent. The intensity of the wireless signals fluctuates in response to lips and mouth movement. As these signals are analyzed, machine learning algorithms look for patterns representing spoken sounds like words or characters [12].

In this work, a practical method for hearing through face masks using RF sensing has been proposed. The suggested RF sensing device can work independently or assist in detecting hearing aids via lips-reading when face masks are worn, which often obscure optical indications for hearing aids in visual-based techniques. The system, based on ultra-wideband (UWB) radar, has been tested for lips-reading applications using deep learning models that has been effectively used in the past for distinct healthcare applications [13]–[15]. To detect different lips motions, the radar-based device recognizes Doppler shift spectrograms. This lips-reading framework can find usage in a number of devices such as voice-controlled systems, automobile systems, biometric security systems and hearing aids. A block diagram of the proposed system is presented in Figure 1. Moreover, the contributions of this study are outlined as follows.

- A state-of-the-art radar sensing system has been introduced for transforming human lips motions into visual micro-Doppler signatures for applications such as hearing aids, voice-controlled systems and biometrics.
- Multiple deep learning algorithms such as Residual Neural Network (ResNet50), InceptionV3 and VGG16, have been utilized on the novel radar dataset for lips-reading spectrogram classification.
- An advanced model, ResNet50, achieved a reliable detection rate of 87% with a mask on scenario and 93% with a mask off scenario.
- Compared to the work in [16], which achieved the highest accuracy of 73% with a mask on and 85% with a mask off, the accuracy has improved by 14% for the mask on scenario and 8% for the mask off scenario.
- Additionally, the classification has been performed for each class of vowels and voiceless class and the results are presented in the form of a confusion matrix.

II. Related Work

RF sensing has been used in the past for various applications [17]–[19]. To detect lips and mouth movement, the majority of techniques analyzed visual information. For instance, in [20], the authors created a laptop-based lips-reading system. This method uses a number of algorithms to identify lips, faces and other characteristics before extracting features and using dynamic programming to identify lips motions. By matching input instructions with pre-registered databases, it can allow real-time interactions and manage various camera positions. Another visual-based method that compared face motions from various perspectives was suggested by authors in [21]. Also, by identifying frames in a film, authors in [22]

developed a useful lips-reading method. The method described in [23] employs visual cues and a small number of visemes to identify a large variety of languages and words that were not trained in the system. It is a neural network-based lips-reading system. The study's contributions include a newly developed transformer for identifying visemes in continuous speech, utilizing visemes as a classification schema for lips-reading sentences and converting visemes to words using perplexity analysis to improve accuracy. While these visual-based methods can achieve excellent identification accuracy, they are constrained by their sensitivity to lighting conditions and are thus ineffective in low-light situations. The possible applications for such systems are severely constrained by this restriction [24], [25].

Recently, lips-reading recognition based on wireless sensing techniques has gained the attention of researchers due to its clear advantages over visual-based approaches. In [26], the authors discussed MIMO technology where reflected signals can be extracted and received by Wi-Hear. Wi-Hear uses beamforming technology and wavelet analysis to amplify and concentrate on the properties of oral motion, allowing fine-grained detection of lips and tongue motions. Speaking activities normally cause negligible Doppler shift and amplitude variation. Similar to this, the authors in [27] developed a Wi-Fi-based method to precisely predict human postures even when obstacles like walls and occlusions are present. The Wi-Fi technology has been successfully used in the past for several healthcare applications such as abnormal respiratory detection [28]–[30]. To identify various mouth motions, in [31], the authors analyze the Doppler shift in the reflected ultrasonic signals from a smartphone. A unique method of biometric identification for mobile devices that relies on lips-reading and acoustic signals has been explored by the authors in [32].

Furthermore, the authors in [33] proposed HearMe, a real-time lips-reading system based on commercial radio frequency identification (RFID) readers that can quickly and effectively identify words from a predefined vocabulary list. To increase recognition accuracy, the system makes use of an efficient data-gathering method and feature extraction techniques. By improving model resilience in cross-environment situations, HearMe's transfer-learning-based technique lowers training costs and increases accuracy in identifying speaking gestures and distinguishing between words in a lexicon, as shown by experimental findings. The authors in [16] from University of Glasgow explored a novel lips-reading approach that inspired our research. We utilized their publicly available dataset to conduct our analysis. The paper explored an RF-based lips-reading framework that employs Wi-Fi and radar technologies to identify vowels from lips movements. The framework overcomes the constraints of camera-based systems such as occlusion and privacy problems and is functional even when individuals are wearing face masks.

III. Approach

A. Experimental Setup

The complete block diagram of the paper is depicted in Figure 1, while Figure 2 showcases the experimental setup

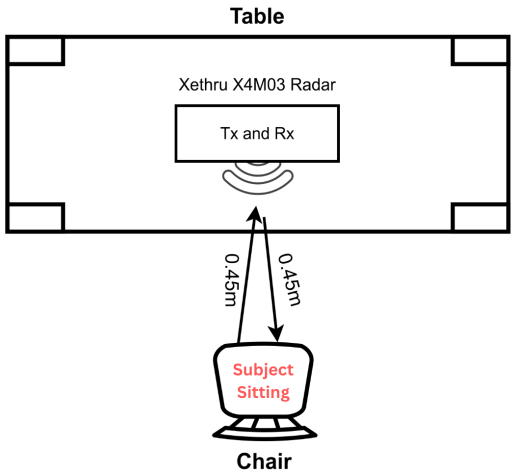


Fig. 2: Data collection experimental setup [16].

TABLE I: XeThru X4M03 UWB radar sensor parameters configuration.

Parameter	Value
Range	9.6 m
Subject's Distance	0.45 m
Operating Frequency	7.29 GHz
Transmitter Power	6.3 dBm
Activity Period	6 sec
Acquired Samples Per Activity	50

utilized for data collection. The data was gathered from both male and female participants, with and without masks, while enunciating vowels. The UWB radar sensor XeThru X4M03 was employed for experimentation that has been efficiently used in the past for other applications [34]. The configuration parameters for the radar sensor are provided in Table I. The radar sensor is equipped with a built-in receiver and transmitter antenna, with an utmost detecting range of 9.6 meters. The experiment was carried out by placing the radar sensor on top of the computer screen.

During the experiment, the participant sat 0.45 meters away from the radar and pronounced different vowels. The participant maintained a normal body position with minimal movements such as lips movements and slight head motions commonly associated with speech. Each activity, representing the data acquisition of an individual vowel from an individual participant, lasted approximately 6 seconds. During this time, the radar transmitted and received the RF signal. To extract features from the radar, we utilized the short-time Fourier transform (STFT), which provided spectrograms indicating the radar Doppler shift resulting from mouth and lips gestures. By analyzing these spectrograms, we were able to identify distinct differences among vowels due to variations in lips and mouth movements. To recognize the vowels, we employed various deep learning algorithms such as ResNet50, InceptionV3 and VGG16.

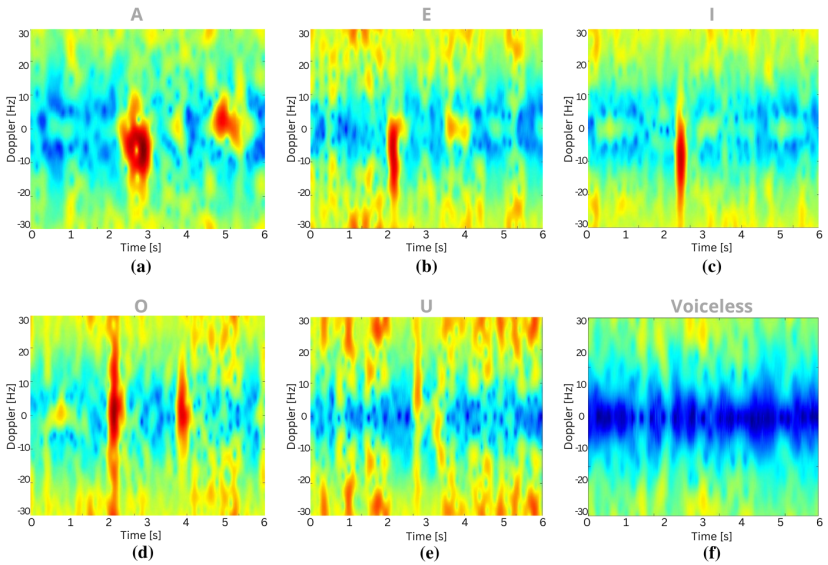


Fig. 3: Sample spectrograms of lips-reading vowels acquired using UWB radar sensor.

TABLE II: Dataset description.

Subject	Sex	Class	Training Samples	Testing Samples	Total Samples
1	Male	With Mask	210	90	300
		Without Mask	210	90	300
2	Female	With Mask	210	90	300
		Without Mask	210	90	300
3	Female	With Mask	210	90	300
		Without Mask	210	90	300
			Overall: **1260**	Overall: **540**	Overall: **1800**

B. Data Collection

The experimental setup used for data collection is depicted in Figure 2. The experiments were conducted using UWB radar technology. The data collection focused on five vowels: A, E, I, O, U, along with a blank letter where the participants remained voiceless with their lips in a normally closed position. A block diagram in Figure 1 illustrates the lips movements required to pronounce all the vowel classes. Figure 3 showcases the spectrograms of lips motion while pronouncing vowels and the voiceless state. To ensure a diverse and realistic dataset, three participants, consisting of two females and one male, took part in the data acquisition. The inclusion of multiple participants aimed to introduce variations in the dataset.

In total, 1800 data samples were composed during the experimentation, covering six classes: A, E, I, O, U and voiceless (representing silence). Each class consisted of 50 samples. The data collection involved participants wearing face masks and participants without face masks, resulting in a total of 900 samples in each scenario. The radar collected data while each participant performed the speaking task 50 times for each vowel, once while wearing a mask and another time without a mask. As a result, each participant contributed a total of 600 data samples for the six classes. Table II provides a description of the acquired dataset.

C. Data Processing

Initially, the radar chip was programmed via the XEP interface using the x4driver. Information was then gathered from the module at a rate of 500 frames per second (FPS) in the form of floating-point message data. A loop was implemented to read the data file and store it in a data stream variable, which was subsequently transformed into a matrix representing complex ranges and time intensities. Following this, a moving target indication (MTI) was employed to obtain the Doppler range map. Subsequently, a Butterworth fourth-order filter was employed as the second MTI filter. This filter was used to create spectrograms by adjusting parameters such as overlap percentage, window length and padding factor of fast Fourier transform (FFT). For this objective, a padding factor of 16 and a window length of 128 samples were deliberately selected.

The process of creating a range profile involved converting each chirp and then performing an FFT. Subsequently, a second FFT was performed on a predefined number of sequential chirps within a specified range bin. In order to obtain information about both time and frequency, spectrograms were generated using STFT, which includes dividing the data into segments and applying Fourier transforms to each segment. The frequency and temporal resolutions are inversely influenced by modifying the length of the window, so increasing one will decrease the other. The amount of Doppler information present in the radar data is resolute by the sampling capability of the hardware. The maximum

unambiguous Doppler frequency in the radar is determined by $F_d, \max = \frac{1}{2}t_r$, where t_r represents the chirp time. In this study, we investigate lips-reading recognition from a specified location such as the mouth, at a distance $D(t)$. The variable $V(t)$ denotes the movement of the target in front of the radar, while T_s represents the transmitted signal.

$$T_s(t) = A\cos(2\pi ft) \tag{1}$$

$Rs(t)$ is the signal that has been received.

$$R_s(t) = \dot{A}\cos\left(2\pi f\left(t - \frac{2D(t)}{c}\right)\right) \tag{2}$$

The coefficient of reflection is denoted by A and the speed of light is represented by c. The signal that bounces back from the target at an angle θ relative to the radar's direction can be mathematically expressed as $R_s(t)$.

$$R_s(t) = \dot{A}\cos\left(2\pi f\left(1 + \frac{2v(t)}{c}\right)\left(t - \frac{4\pi D(\theta)}{c}\right)\right) \tag{3}$$

Its corresponding Doppler shift can be expressed as.

$$f_d = f\frac{2v(t)}{c} \tag{4}$$

The signal that is received back is combined with various moving parts, including the head and lips. Each component travels at its acceleration and speed. The signal that was received can be expressed as follows if we assume that i represents the different moving parts of the lips.

$$R_s(t) = \sum_{i}^{N} A_i\cos\left(2\pi f\left(1 + \frac{2vi(t)}{c}\right)\left(t - \frac{4\pi D_i(0)}{c}\right)\right) \tag{5}$$

The frequency shift brought on by movement is the outcome of a complicated interaction between a number of frequency changes brought on by the movement of various face components. The accurate identification of lips movements for lips-reading depends on the exact features presented by the frequency signatures. The spectrogram dataset that illustrated the distinguishing features of different vowels and voiceless was acquired after the subject's activity. The recommended pre-trained deep learning classification techniques were then used on the dataset to identify vowels.

D. Deep Learning Models

1) ResNet50: The deep learning technique known as ResNet50, or Residual Network with 50 layers, is commonly employed in computer vision applications, notably in image categorization. The authors in [35] first discussed it in their publication. On several benchmark datasets, ResNet50 has achieved state-of-the-art performance, demonstrating its high degree of efficacy. The fundamental idea of ResNet50 is the use of residual learning, which aids in addressing the degradation issue that deep neural networks experience. Due to the difficulties of deep learning deep networks, the degradation problem develops when increasing the network depth results in lower accuracy. Skip connections, often referred to as residual connections, are included in ResNet50 to overcome this problem and allow the gradient to pass across the network without degrading.

The following is a mathematical illustration of ResNet50. The intended underlying mapping is designated as $H(x)$, while the input to the network is denoted as x. By stacking residual construction pieces, the ResNet50 network comes close to $H(x)$. Following batch normalization and ReLU activation, each construction block is composed of a number of convolutional layers. It can be expressed as $F(l, x)$, where l stands for the block's index to indicate the output of the l-th construction block. The output $Y(l)$ of each construction block is added to the input x to create a residual connection.

$$Y(l) = F(l, x) + x \tag{6}$$

The network attempts to maximize the following goal in order to understand the residual mapping.

$$\text{minimize } \|H(x) - Y(L)\|^2 + \lambda\Sigma\|W(i)\|^2 \tag{7}$$

$W(i)$ represents the weights of the ith layer, L represents the total number of building blocks and λ regulates the degree of load decay normalization. ResNet50 simplifies deep network optimization by using residual connections to help the network distinguish between the intended mapping and the input. Skip connections allow gradient flow throughout the network, solving the degradation issue and making it easier to train more complex models.

2) InceptionV3: Developed by Google researchers, InceptionV3 is a deep learning model that includes a number of architectural improvements to boost performance and productivity in computer vision tasks. The model achieves effective multi-scale feature capture by using inception modules, which consist of parallel convolutional branches. It also utilizes dimensionality reductions and factorized convolutions to simplify computations while maintaining accuracy. To improve convergence and reduce overfitting, InceptionV3 includes auxiliary classifiers. The model has demonstrated outstanding performance on benchmarks for picture categorization and is widely used in both academic and real-world settings.

3) VGG16: VGG16 is a renowned deep learning model developed by researchers at the University of Oxford. Its consistent design and ease of use have made it widely used in computer vision studies and applications. The model learns hierarchical representations of images effectively using narrow receptive fields and ReLU activation functions. It has 16 layers, including convolutional and fully connected ones. VGG16 has shown to be exceptionally effective on picture classification tests like the ImageNet dataset despite its processing requirements. It is a well-liked option in computer vision research due to its straightforward architecture and deep layer structure.

IV. Results and Discussion

Data was gathered, processed, and spectrograms generated for further analysis. The spectrogram samples used to train three different deep learning algorithms, ResNet50, InceptionV3, and VGG16, are displayed in Figure 3. The dataset

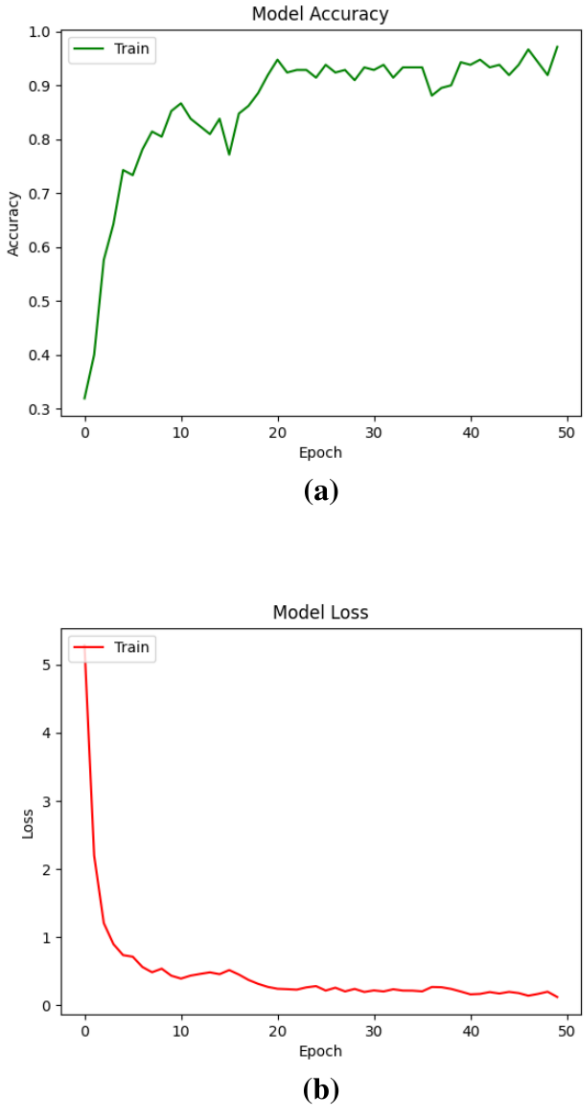


Fig. 4: ResNet50 model (a) accuracy and (b) loss during training on with mask lips-reading scenario.

used for testing and training purposes is described in Table II. Training was carried out across a variety of epoch counts, including 20, 30, 40, and 50, to evaluate the effectiveness of deep learning algorithms. The algorithms' performance did not improve with the number of epochs, and steady accuracy was only attained at about 20 epochs. This is due to the fact that the dataset used in this paper is not huge and even with a certain number of epochs, better performance can be achieved. This is demonstrated in Figure 4 (a) for ResNet50 model accuracy and Figure 4 (b) for ResNet50 model loss.

Figure 5 presents the performance of the cutting-edge deep learning model ResNet50 in terms of a confusion matrix for the classification of spectrogram images of six distinct classes: A, E, I, O, U and voiceless. The scenarios include multiple subjects, either male or female and either wearing a mask or not wearing a mask. The ResNet50 performance in terms of a confusion matrix shows how often the model correctly or incorrectly classified the images. The confusion matrix of the ResNet50 model was analyzed for six scenarios: (a) male subject 1 wearing a mask, (b) male subject 1 not wearing a mask, (c) female subject 2 wearing a mask, (d) female subject 2 not wearing a mask, (e) female subject 3 wearing a mask and (f) female subject 3 not wearing a mask. As can be seen in Figure 5, for all subjects, the scenario without a mask is better detected compared to the scenario with a mask and this is due to the obvious fact that a mask covering can be a resistance towards accurate detection of lips movement through RF signals.

Table III presents the performance of different deep learning algorithms on scenarios with masks and without masks for three subjects, one male and two female. We compared the performance of ResNet50 with InceptionV3 and VGG16. For subject 1, the accuracy of ResNet50 was 98% without a mask and 87% with a mask. For subject 2, the accuracy of ResNet50 was 92% without a mask and 88% with a mask. With and without a mask, ResNet50 accuracy for subject 3 was 90% and 86%, respectively. InceptionV3 was 74% and 58% more accurate than ResNet50 for the subject 1 without and with mask scenarios, respectively. For subject 2, InceptionV3 attained accuracy of 64% without a mask and 41% with a mask. With and without a mask, InceptionV3 was accurate for subject 3 attaining score of 77% and 63%, respectively. For subject 1 without and with a mask scenario, VGG16 outperformed ResNet50 and InceptionV3 with accuracy rates of 89% and 72%, respectively. For subject 2, the accuracy of VGG16 was 83% without a mask and 73% with a mask. For subject 3, VGG16 achieved an accuracy of 88% and 70%, respectively, without and with a mask.

Furthermore, Figure 6 presents a bar chart that displays the average performance for multiple subjects with masks, without masks and overall. As can be seen from the graph, ResNet50 attained the highest accuracy with an 87% accurate detection rate with a mask scenario, 93% without a mask scenario and 90% overall. Compared to InceptionV3 and VGG16 for the lips-reading scenario with masks, ResNet50 enhanced the performance by 33% and 16%, respectively. ResNet50 achieves better performance than InceptionV3 and VGG16 due to its deep architecture with skip connections, residual learning, parameter efficiency and pre-training on large-scale datasets. These factors enable ResNet50 to capture more complex features, optimize weights efficiently and generalize well, resulting in improved performance. However, the most suitable deep learning model selection depends on the trade-offs and constraints of the particular use case.

V. Conclusion and Future Work

This research introduces a lips-reading framework based on RF sensing that utilizes radar technology. The radar component involves the adoption of the XeThru X4M03 UWB radar sensor, which produces Doppler signals that are plotted in frequency-time diagrams. This RF sensing system can function independently or provide assistance to hearing aids by interpreting lips and mouth movements, particularly in scenarios where visual cues are hindered by face masks, as

Fig. 5: ResNet50 model outcome in terms of confusion matrix of (a) subject 1 male with mask scenario (b) subject 1 male without mask scenario (c) subject 2 female with mask scenario (d) subject 2 female without mask scenario (e) subject 3 female with mask scenario (f) subject 3 female without mask scenario.

TABLE III: Lips-reading detection accuracy through distinct deep learning algorithms.

Subject	Sex	Class	ResNet50 Accuracy	InceptionV3 Accuracy	VGG16 Accuracy
1	Male	With Mask	≈ 87%	≈ 58%	≈ 72%
		Without Mask	≈ 98%	≈ 74%	≈ 89%
2	Female	With Mask	≈ 88%	≈ 41%	≈ 73%
		Without Mask	≈ 92%	≈ 64%	≈ 83%
3	Female	With Mask	≈ 86%	≈ 63%	≈ 70%
		Without Mask	≈ 90%	≈ 77%	≈ 88%
			Avg. with mask: **87%**	Avg. with mask: **54%**	Avg. with mask: **71%**
			Avg. without mask: **93%**	Avg. without mask: **71%**	Avg. without mask: **86%**
			Avg. overall: **90%**	Avg. overall: **62%**	Avg. overall: **78%**

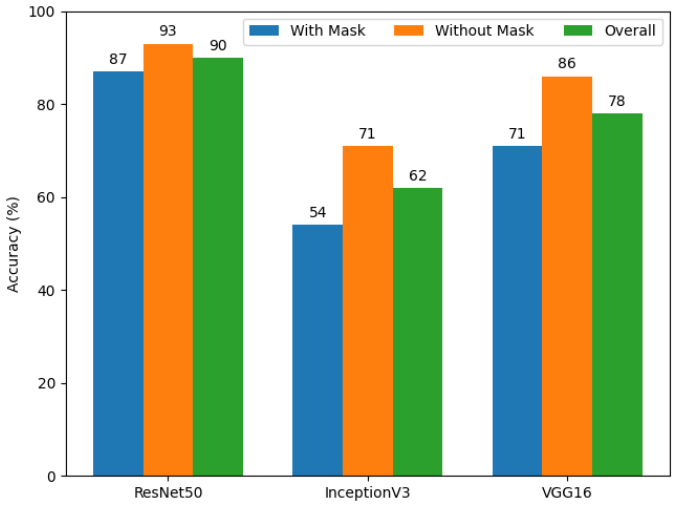


Fig. 6: Distinct deep learning algorithms accuracy comparison for lips-reading. ResNet50 excelled InceptionV3 and VGG16 in terms of detection accuracy.

is typically experienced in vision-based schemes. The team at the University of Glasgow published a publicly available dataset involving three subjects, consisting of two females and one male, which was utilized. This dataset encompassed five vowels (A, E, I, O, U) and a voiceless class where no lips movements occurred. Subsequently, various deep learning algorithms such as ResNet50, InceptionV3 and VGG16 were trained and evaluated using this dataset. The primary objective of this study was to propose a secure lips-reading system capable of accurately identifying lips movements in the presence of masks, employing RF sensing technology and a deep learning model. Compared to InceptionV3 and VGG16, ResNet50, an advanced model, was able to accurately detect lips motions with a mask on scenario with an 87% detection rate and without a mask with a 93% detection rate. Overall, it attained 90% accuracy. This is a significant improvement compared to the results achieved in the study by [16], where the highest accuracy achieved was 73% with a mask on and 85% with a mask off. The accuracy has improved by 14% for the mask on scenario and 8% for the mask off scenario using ResNet50.

The accuracy and quality of the radar sensing system are affected by various limitations. Signal interference and reduced effectiveness in detecting and tracking targets can occur due to coupling with nearby systems. Additionally, external factors

such as weather conditions and electromagnetic disturbances can further compromise the system's ability to provide reliable data.

The current system is a proof-of-concept and is intended to underscore the significance and effectiveness of lips detection using RF sensing technology like radar. Future work will involve real-time detection of different words or sentences from various angles, not directly in the line of sight. Furthermore, the experiments were carried out in a static environment with minimal limb movements. In the future, we aim to perform experimentation that is not completely static. Using advanced deep learning algorithms, other body movements can be discarded, and the RF signal can be focused only on the target activity.

Acknowledgement

This work is supported in parts by EPSRC grant (EP/W037076/1).

References

[1] G. Chen, X. Xiao, X. Zhao, T. Tat, M. Bick, and J. Chen, "Electronic textiles for wearable point-of-care systems," *Chemical Reviews*, vol. 122, no. 3, pp. 3259–3291, 2021.

[2] S. A. Shah and F. Fioranelli, "Rf sensing technologies for assisted daily living in healthcare: A comprehensive review," *IEEE Aerospace and Electronic Systems Magazine*, vol. 34, no. 11, pp. 26–44, 2019.

[3] W. Taylor, S. A. Shah, K. Dashtipour, A. Zahid, Q. H. Abbasi, and M. A. Imran, "An intelligent non-invasive real-time human activity recognition system for next-generation healthcare," *Sensors*, vol. 20, no. 9, p. 2653, 2020.

[4] J. L. Punch, R. Hitt, and S. W. Smith, "Hearing loss and quality of life," *Journal of communication disorders*, vol. 78, pp. 33–45, 2019.

[5] Z. Zhang, F. Wen, Z. Sun, X. Guo, T. He, and C. Lee, "Artificial intelligence-enabled sensing technologies in the 5g/internet of things era: from virtual reality/augmented reality to the digital twin," *Advanced Intelligent Systems*, vol. 4, no. 7, p. 2100228, 2022.

[6] J. Stephenson, "Who report predicts hearing loss for 1 in 4 people worldwide by 2050," in *JAMA Health Forum*, vol. 2, no. 3. American Medical Association, 2021, pp. e210 357–e210 357.

[7] M. Leo, P. Carcagnì, P. L. Mazzeo, P. Spagnolo, D. Cazzato, and C. Distante, "Analysis of facial information for healthcare applications: a survey on computer vision-based approaches," *Information*, vol. 11, no. 3, p. 128, 2020.

[8] U. Saeed, S. Y. Shah, J. Ahmad, M. A. Imran, Q. H. Abbasi, and S. A. Shah, "Machine learning empowered covid-19 patient monitoring using non-contact sensing: An extensive review," *Journal of pharmaceutical analysis*, vol. 12, no. 2, pp. 193–204, 2022.

[9] U. Saeed, S. U. Jan, Y.-D. Lee, and I. Koo, "Fault diagnosis based on extremely randomized trees in wireless sensor networks," *Reliability engineering & system safety*, vol. 205, p. 107284, 2021.

[10] Y. Chen, W. Liu, Z. Niu, Z. Feng, Q. Hu, and T. Jiang, "Pervasive intelligent endogenous 6g wireless systems: Prospects, theories and key technologies," *Digital communications and networks*, vol. 6, no. 3, pp. 312–320, 2020.

[11] D. Lee, G.-Y. Nie, and K. Han, "Vision-based inspection of prefabricated components using camera poses: Addressing inherent limitations of image-based 3d reconstruction," *Journal of Building Engineering*, vol. 64, p. 105710, 2023.

[12] W. Taylor, Q. H. Abbasi, K. Dashtipour, S. Ansari, S. A. Shah, A. Khalid, and M. A. Imran, "A review of the state of the art in non-contact sensing for covid-19," *Sensors*, vol. 20, no. 19, p. 5665, 2020.

[13] U. Saeed, S. Y. Shah, A. A. Alotaibi, T. Althobaiti, N. Ramzan, Q. H. Abbasi, and S. A. Shah, "Portable uwb radar sensing system for transforming subtle chest movement into actionable micro-doppler signatures to extract respiratory rate exploiting resnet algorithm," *IEEE Sensors Journal*, vol. 21, no. 20, pp. 23 518–23 526, 2021.

[14] F. Fioranelli, J. Le Kernec, and S. A. Shah, "Radar for health care: Recognizing human activities and monitoring vital signs," *IEEE Potentials*, vol. 38, no. 4, pp. 16–23, 2019.

[15] U. Saeed, S. Y. Shah, S. A. Shah, J. Ahmad, A. A. Alotaibi, T. Althobaiti, N. Ramzan, A. Alomainy, and Q. H. Abbasi, "Discrete human activity recognition and fall detection by combining fmcw radar data of heterogeneous environments for independent assistive living," *Electronics*, vol. 10, no. 18, p. 2237, 2021.

[16] H. Hameed, M. Usman, A. Tahir, A. Hussain, H. Abbas, T. J. Cui, M. A. Imran, and Q. H. Abbasi, "Pushing the limits of remote rf sensing by reading lips under the face mask," *Nature Communications*, vol. 13, no. 1, p. 5168, 2022.

[17] Z. Wang, A. Ren, Q. Zhang, A. Zahid, and Q. H. Abbasi, "Recognition of approximate motions of human based on micro-doppler features," *IEEE Sensors Journal*, 2023.

[18] A. Zahid, H. T. Abbas, I. E. Carranza, J. P. Grant, M. A. Imran, D. R. Cumming, and Q. H. Abbasi, "Assessing the salt constituents characteristics in aqueous solutions using terahertz waves," in *2020 IEEE International Symposium on Antennas and Propagation and North American Radio Science Meeting*. IEEE, 2020, pp. 1527–1528.

[19] M. U. Rehman, A. Shafique, K. H. Khan, S. Khalid, A. A. Alotaibi, T. Althobaiti, N. Ramzan, J. Ahmad, S. A. Shah, and Q. H. Abbasi, "Novel privacy preserving non-invasive sensing-based diagnoses of pneumonia disease leveraging deep network model," *Sensors*, vol. 22, no. 2, p. 461, 2022.

[20] T. Saitoh, "Development of communication support system using lip reading," *IEEJ transactions on electrical and electronic engineering*, vol. 8, no. 6, pp. 574–579, 2013.

[21] K. Kumar, T. Chen, and R. M. Stern, "Profile view lip reading," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4. IEEE, 2007, pp. IV–429.

[22] Z. Zhou, G. Zhao, and M. Pietikäinen, "Towards a practical lipreading system," in *CVPR 2011*. IEEE, 2011, pp. 137–144.

[23] S. Fenghour, D. Chen, K. Guo, and P. Xiao, "Lip reading sentences using deep learning with only visual cues," *IEEE Access*, vol. 8, pp. 215 516–215 530, 2020.

[24] B. Garcia-Garcia, T. Bouwmans, and A. J. R. Silva, "Background subtraction in real applications: Challenges, current models and future directions," *Computer Science Review*, vol. 35, p. 100204, 2020.

[25] S. I. Shah, S. Y. Shah, and S. A. Shah, "Intrusion detection through leaky wave cable in conjunction with channel state information," in *2019 UK/China Emerging Technologies (UCET)*. IEEE, 2019, pp. 1–4.

[26] 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.

[27] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7356–7365.

[28] U. Saeed, S. Y. Shah, A. Zahid, J. Ahmad, M. A. Imran, Q. H. Abbasi, and S. A. Shah, "Wireless channel modelling for identifying six types of respiratory patterns with sdr sensing and deep multilayer perceptron," *IEEE Sensors Journal*, vol. 21, no. 18, pp. 20 833–20 840, 2021.

[29] C. B. Ali, A. H. Khan, K. Pervez, T. M. Awan, A. Noorwali, and S. A. Shah, "High efficiency high gain dc-dc boost converter using pid controller for photovoltaic applications," in *2021 International Congress of Advanced Technology and Engineering (ICOTEN)*. IEEE, 2021, pp. 1–7.

[30] U. Saeed, Q. H. Abbasi, and S. A. Shah, "Ai-driven lightweight real-time sdr sensing system for anomalous respiration identification using ensemble learning," *CCF Transactions on Pervasive Computing and Interaction*, vol. 4, no. 4, pp. 381–392, 2022.

[31] J. Tan, C.-T. Nguyen, and X. Wang, "Silenttalk: Lip reading through ultrasonic sensing on mobile phones," in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 2017, pp. 1–9.

[32] L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, L. Kong, and M. Li, "Lip reading-based user authentication through acoustic sensing on smartphones," *IEEE/ACM Transactions on Networking*, vol. 27, no. 1, pp. 447–460, 2019.

[33] S. Zhang, Z. Ma, K. Lu, X. Liu, J. Liu, S. Guo, A. Y. Zomaya, J. Zhang, and J. Wang, "Hearme: Accurate and real-time lip reading based on commercial rfid devices," *IEEE Transactions on Mobile Computing*, 2022.

[34] H. Hameed, M. Usman, A. Tahir, K. Ahmad, A. Hussain, M. A. Imran, and Q. H. Abbasi, "Recognizing british sign language using deep learning: A contactless and privacy-preserving approach," *IEEE Transactions on Computational Social Systems*, 2022.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.