electronics

MDPI

Article

Doppler Radar-Based Human Speech Recognition Using Mobile Vision Transformer

Wei Li [1,*], Yongfu Geng [1], Yang Gao [1], Qining Ding [1], Dandan Li [1], Nanqi Liu [2] and Jinheng Chen [2]

[1] School of Information, North China University of Technology, Beijing 100144, China
[2] Brunel London School, North China University of Technology, Beijing 100144, China
* Correspondence: lwsar@ncut.edu.cn

Abstract: As one of the important vital features of the human body, the acquisition of a speech signal plays an important role in human–computer interaction. In this study, voice sounds are gathered and identified using Doppler radar. The skin on the neck vibrates when a person speaks, which causes the vocal cords to vibrate as well. The vibration signal received by the radar will produce a unique micro-Doppler signal according to words with different pronunciations. Following the conversion of these signals into micro-Doppler feature maps, these speech signal maps are categorized and identified. The speech recognition method used in this paper is on neural networks. CNN convolutional neural networks have a lower generalization and accuracy when there are insufficient training samples and sample extraction bias, and the training model is not suitable for use on mobile terminals. MobileViT is a lightweight transformers-based model that can be used for image classification tasks. MobileViT uses a lightweight attention mechanism to extract features with a faster inference speed and smaller model size while ensuring a higher accuracy. Our proposed method does not require large-scale data collection, which is beneficial for different users. In addition, the learning speed is relatively fast, with an accuracy of 99.5%.

Keywords: Doppler radar; human speech recognition; Mobile ViT

check for updates

Citation: Li, W.; Geng, Y.; Gao, Y.; Ding, Q.; Li, D.; Liu, N.; Chen, J. Doppler Radar-Based Human Speech Recognition Using Mobile Vision Transformer. *Electronics* **2023**, *12*, 2874. https://doi.org/10.3390/electronics12132874

Academic Editor: Juan-Carlos Cano

Received: 31 May 2023
Revised: 21 June 2023
Accepted: 27 June 2023
Published: 29 June 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the continuous development of smart homes, human–computer interaction has become an integral part, and voice recognition [1] has always been an important part of a smart home. There are two types of speech propagation detection systems now available: air conduction and non-air conduction. The main representative of air conduction detection technology is the traditional microphone sensor; sound waves drive the diaphragm on the microphone to produce vibration, and the diaphragm drives the vibration of the coil set on the magnet, which produces an analog signal of sound, and then the analog signal is converted into a digital signal for research. The main way of contact zoning–air conduction is to put the sensor close to the sounding organ, transmit vibration to the sensor and convert the vibration signal into an acoustic signal through voltage changes. However, the former is highly susceptible to the environment and noise, and the technology is limited by the distance factor, so the final result is easily compromised by noise, distance, obstacles, etc., while the latter requires close skin surface limited by human activity and comfort and safety factors need to be considered. Therefore, the perception of the human voice may become challenging.

In recent years, a new non-contact life detection technology is gradually gaining widespread attention; the detection technology is not affected by the environmental temperature, climate and other conditions, and involves the non-contact, non-invasive, safe, high-sensitivity, high directional detection of remote objects' micro-motion signal. Domestic and foreign scholars call this technology "bio-radar" [2]. Since its introduction, bio-radar technology has been widely used in the detection of human vital signs such as breathing

and a heartbeat, and has achieved good results. It is relatively new research to apply it to voice signal detection.

This paper investigates the use of bio-radar to capture the human voice and identify the detection. Human air can enter the lungs through the normal respiratory system, and during speech, the abdominal muscles contract to bring the diaphragm upward, squeezing out the air from the lungs and creating airflow. The airflow exhale from the lungs is the prime mover of speech production. As the airflow passes through the trachea and bronchi through the pharynx, the tightened vocal cords vibrate due to the impact of the airflow, constantly opening and closing, sending a series of jets upward from the vocal folds. The airflow at this point is truncated into quasi-periodic pulses, generally represented by asymmetric triangular waves. The vibration of the vocal cords depends on their mass. The greater the mass, the smaller the number of vibrations per second; conversely, the smaller the mass, the faster the vocal folds vibrate. The frequency of vocal cord vibration determines the pitch of the sound. The vocal folds vibrate to produce sound, which is the basic source of sound production. The vocal folds are further modulated for passing through the pharynx, mouth or nasal cavity. The opening and closing of the oral cavity, the movement of the tongue and the lifting and lowering of the soft palate create different vocal tract configurations, resulting in different speech sounds. Since the vibrations of the human vocal cords modulate radar waves when irradiated, speech information can be extracted from the returned radar signal based on Doppler frequencies that are linearly proportional to the frequency of the acoustic vibrations.

In 1996, Zongwen Li et al. successfully detected the free-space human speech signal for the first time using 40 GHz millimeter wave radar [3], and in 2002, Staderini et al. evaluated and monitored human heart motion and human vocal function using Ultra WideBand (UWB) radar [4]. To further validate the signal source of vocal vibration measured by EM radar sensors, in 2005, Holzrichter verified that the source of vocal vibration detected by EM radar sensors is mainly the vocal folds with a special set of experiments. In recent years, there are many scholars who are accustomed to using 77 GHz millimeter wave radar to capture the feature acquisition signal inside the human oral cavity; for example, Yudi Dong et al. verified the feasibility of using the signal acquired by millimeter wave radar by experimenting on subjects, extracting the biometric features from the signal and then verifying the feasibility with convolutional neural networks [5].

We used a system for human voice verification by Doppler radar that exploits the stability and reliability of Doppler radar as a tool for detecting a human voice at a distance [6], which not only enables accurate speaker verification by using new biometric features, but also by contactless sensing that does not require any sensors on the user's body. We used Doppler radar and neural networks for the feasibility of speech recognition. A subject's speech is first measured to ensure that the radar recognized the laryngeal vibrations, then the micro-Doppler signals of the words are captured [7], followed by a time–frequency analysis to convert the time–domain signals into spectrograms so that the neural network can recognize the micro-Doppler heat maps of the different words. Currently, most methods for classification detection of target images will use LSTM and CNN neural networks, such as the improved R-CNN-based micro-Doppler image activity for identifying micro-Doppler images and feature maps with CNN neural networks proposed in the literature [5,7]. These methods rely on the use of deep architectures. Some success has been achieved in radar recognition of speech, which can reach more than a 90% success rate. However, there are still two problems: first, when training data are limited, deep learning-based models often encounter a serious problem of overfitting or underfitting, which can lead to a decrease in detection accuracy [8,9]. Speech detection is a branch of natural language processing, so a further step up is desired both in terms of generalization capability to the model and accuracy. On the other hand, these solutions are computationally expensive in terms of processing as well as the memory and storage footprint. Therefore, their use in several applications with computational limitations, such as mobile devices [10], is not feasible in

practice. We hope to find a lightweight neural network with smaller models and a higher accuracy to be implemented in mobile terminals or other display terminals.

In the last 2 years, a lightweight neural network Mobile ViT [11] for mobile devices has been discovered, so we can use Mobile ViT to classify the micro-Doppler features of words by converting the time–domain signal into a Doppler heat map and applying it to Mobile ViT. Then, the first step is to invoke a lightweight general-purpose visualization neural network and optimize its parameters so that its classification accuracy is optimal. Next, we have different subjects collect speech signals in a relatively quiet and smooth environment, and evaluate our neural network architecture through extensive experiments with these subjects. The results show that the network architecture we used is able to identify words more accurately for classification and has a high accuracy for speaker verification.

This paper is organized as follows. We discuss the radar principle and the micro-Doppler signal of the vocal cords in Section 2. Section 3 presents the interpretation of data measurement and processing. Section 4 describes the lightweight neural network Mobile ViT. Section 5 explains the experimental environment as well as the experimental procedure. We summarize the results of this study and make suggestions for future research in Section 6.

2. Principle Description

The human voice organ consists of the larynx and the vocal cords, which are connected to the lower trachea and the upper pharyngeal cavity. The vocal folds are located in the middle of the larynx and are two flexible membranes. The cleft between the two membranes is the vocal fissure, and the size of the vocal fissure can change according to the tension of the vocal folds. When a healthy person makes a sound, they first inhale air, then tighten the vocal cords and temporarily hold their breath. The airflow from the lungs impels the vocal cords to vibrate, and this vibration also causes the air in the larynx to vibrate together, at which point we make our voice. Some of our voices are sharp and thin, some are low, some are silky and some are high and loud, and at least 40 muscles are involved in vocalization, so these are different because of the different frequencies and amplitudes of vocal cord vibration. When a person speaks, the membrane vibrates and this also causes vibrations in the skin of the person's neck [12], as shown in Figure 1. The waves emitted by Doppler radar are reflected by the person's skin. When the skin of the neck vibrates, it can be measured with electromagnetic waves because the received signal is Doppler-modulated by the vibration. Normally, human speech consists of several different frequency components that vary over time, especially harmonics. Then, this means that silent sounds exist when no sound is made, such as deaf people who try to express themselves but fail to make a sound. The lips and tongue also move, so when a person speaks, a micro-Doppler signal is bound to be generated under radar detection. Therefore, we can capture the micro-Doppler signal with the vibration of the vocal cords and the movement of the tongue and lips.

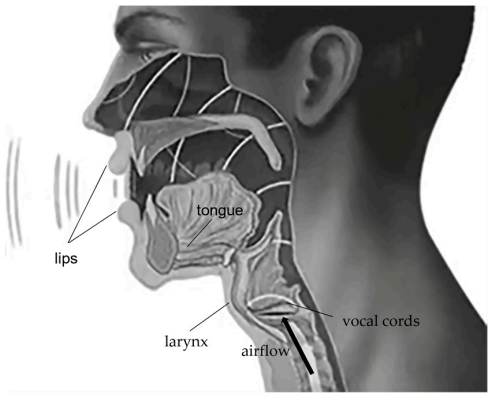


Figure 1. The main vocal organs of the human body.

Radars work by transmitting a number of radio pulses to detect a target and extrapolating the distance to the target by measuring the signal delay of the received echo, i.e., the delay time is the two-way distance traveled by the pulse divided by its propagation speed. In fact, it can be taken a step further; radar can also capture the Doppler frequency of a moving target through the Doppler effect to accurately identify the target characteristics. The difference between the transmitted and received frequencies due to the Doppler effect is called the Doppler shift. The Doppler shift is defined as $f_D = -2v/\lambda$, where λ is the wavelength and v is the velocity of the object. In the case of a vibrating or rotating target, this causes frequency modulation, resulting in a sideband around the main offset of the object's frequency. This offset is known as the micro-Doppler effect. Since most targets are not rigid bodies, other vibrations and rotations are often present in different parts of the target in addition to the platform motion. For example, when a helicopter flies, its blades rotate, or when a person walks, their arms naturally swing. These micro-scale motions produce additional Doppler shifts [13], which are useful when identifying target features. The source of vibration of the vocal organs mentioned above is mainly the vocal folds, and the sound produced by humans is composed of many different frequency components. In order to verify the micro-Doppler effect generated by the laryngeal vibrations during human speech, we investigated the periodic vibrations. The received Doppler from the oscillation point target is modeled as follows:

$$s(t) = A \cdot e^{[j(2\pi f_c t + \beta \sin(\omega_0 t))]} \tag{1}$$

where f_c is the carrier frequency, β is the phase size of the oscillation and ω_0 is the oscillation frequency. The instantaneous Doppler frequency is expressed as

$$f_D = \frac{1}{2\pi} \cdot \frac{d\varphi}{dt} = \frac{1}{2\pi} \cdot \beta\omega_0 \cos(\omega_0 t) \tag{2}$$

Since Equation (1) is a periodic function, it can be written in terms of Fourier series expansions:

$$s(t) = Af(x) = a_0 + \sum_{n=-\infty}^{\infty} c_n e^{[j(2\pi f_c + n\omega_0)t]} \tag{3}$$

The Fourier coefficient c_n is expressed as

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} J_n(\beta) \cdot e^{[j(2\pi f_c + n\omega_0)t]} \tag{4}$$

We can notice that the vibrations produce paired harmonic spectral lines around the fundamental Doppler shift. The amplitude of the micro-Doppler features is highest when $n = 1$, so that the vibration frequencies are usually clearly visible in the spectrogram. It is obvious that the vibrations of the neck skin induced by the vocal cords produce Doppler frequencies corresponding to the stated frequencies.

3. Data Measurement and Processing Methods

This study used Doppler radar to collect data, and the radar block diagram is shown in Figure 2.

First, we conducted a Fourier transform (FFT) [14] on the obtained signal; this step was mainly to detect whether the position of human vocalization matches with the experiment. The next step used the phase change of the radar signal to measure the microdynamic characteristic changes of the human body, and the microdynamic signal was extracted from the phase difference to obtain the corresponding microdynamic characteristic signal of the neck. To represent the time-varying characteristics of the micro-Doppler signal, the measured signal was converted into a spectrogram. We used a short-time Fourier transform (STFT) for a joint time–frequency analysis. A short-time Fourier transform is a commonly used time-frequency analysis method, which represents the signal characteristics at a certain

moment with a segment of signal within a time window. In the process of STFT, the length of the window determines the time resolution and frequency resolution of the spectrogram; the longer the window length, the longer the intercepted signal, the longer the signal, the higher the frequency resolution after the Fourier transform and the poorer the time resolution; on the contrary, the shorter the window length, the shorter the intercepted signal, the worse the frequency resolution and the better the time resolution. When performing STFT, the speech signal is first divided into several time windows, and then the Fourier transform is applied to each time window to obtain the frequency distribution of the signal within that time period. It is easy to know that the short-time Fourier transform is to multiply a function and a window function first, and then perform a one-dimensional Fourier transform. And through the results of sliding the window function to obtain a series of Fourier change, these results are lined up to obtain a two-dimensional representation.

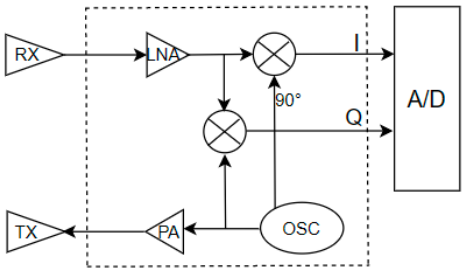


Figure 2. Doppler radar box composition.

Fourier formula:

$$STFTZ(t,f) = \int_{-\infty}^{\infty} [z(u)g*(u-t)]e^{-j2\pi fu}d_u \quad (5)$$

$Z(t)$ is the source signal and $g(t)$ is the window function. In this experiment, we set the window length to 10 ms. The spectrogram of the micro-Doppler feature [7] was obtained by capturing the micro-Doppler signal as the test set we needed.

In the experimental phase, we first tested words of different frequencies, including ourselves, to ensure the reliability of the radar system in capturing micro-Doppler features. For the sake of generalizability in the laboratory, we randomly selected 20 subjects, including 10 males and 10 females between the ages of 20 and 30, who were not professional speakers, so we did not expect the pronunciation of each subject to be standard. When the signal from Doppler radar picked up the micro-motion features of the neck when the experimenter spoke, the corresponding feature signals were extracted for a time–frequency analysis with STFT, and the micro-Doppler feature map of the neck could be obtained, as shown in Figure 3 (we took the spectrum of 10 different words from 0–9). The horizontal axis represents time and the vertical axis represents frequency. In this figure, we can observe different micro-Doppler features. Specifically, we asked each subject to pronounce the words at different frequencies, from 0 to 9; each word was pronounced 20 times at different speech rates, and each word was pronounced for 2 seconds. We successfully created a set of micro-Doppler features about human speech, which we can use for classification recognition in neural networks.

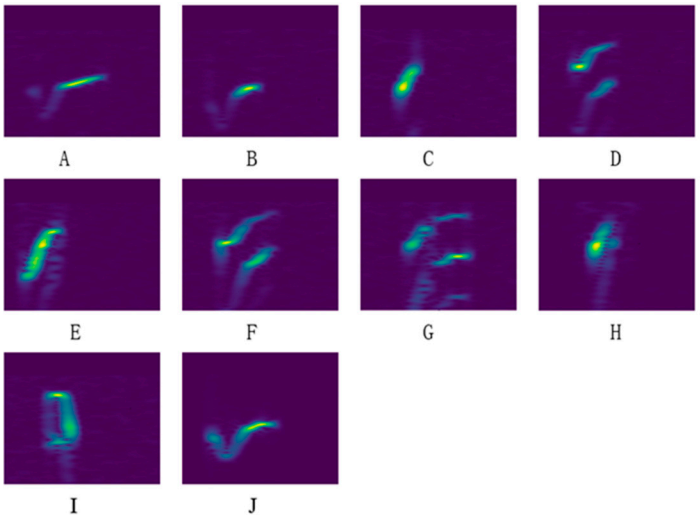


Figure 3. Micro-Doppler heat map representing the 10 words, 0–9, from (**A**–**J**), respectively.

4. Mobile ViT Neural Network

We use a lightweight Vision Transformer network, Mobile ViT, a lightweight neural network [15] visual transformer for mobile devices, to classify and identify the Doppler thermogram of the word under testing. As shown in Figure 4, MobileViT is mainly composed of a normal convolution, MV2, a Light ViT block, global pooling and fully connected layers together. The role of the convolutional layer is to extract features. The MV2 block in a MobileViT network is mainly responsible for down-sampling and the MobileViT block substitutes local processing in convolution with global processing via transformers. This enables the Light ViT block to have qualities comparable to CNN [16] and ViT [17], allowing it to train recipes with fewer parameters and more simplicity. Figure 4 depicts the MobileViT neural network's design.

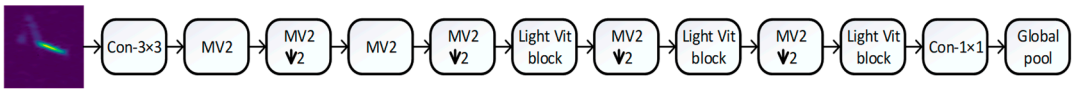


Figure 4. MobileViT structure diagram.

The core module in Mobile ViT is the Light ViT block [11], as shown in Figure 5, which is intended to model local and global information in the input tensor with fewer parameters. The feature map is first modeled locally with a convolutional kernel size of 3×3, then the number of channels is adjusted with a convolutional kernel size of 1×1. In the next step, the global feature modeling is performed with an Unfold–Transformer–Fold structure, and the number of channels is adjusted back to the original size with a convolutional kernel size of 1×1. The shortcut branch is stitched with the original input feature map, and the output is finally fused by a convolutional layer with a 3×3 convolutional kernel size. For a given input tensor $X \in RH \times W \times C$, C, H and W denote the tensor's channel, height and width, respectively, and P = wh represents the number of pixels in a patch with height H and width W. We set H = W = 2 at all spatial levels. The feature map is cut into patches of the same size. For the convenience of calculation, it is assumed that the size of each patch is 2×2 (ignoring the number of channels), and the size of each patch consists of four tokens (each small color block in the graph). When performing self-attention calculation, each token only attends with its own token of the same color as shown in Figure 6, so as to reduce the amount of computational purposes. There is a considerable amount of data redundancy in the image data itself, such as for the shallow feature map, the information between adjacent tokens may not be much different if each token directs attention to the neighboring token.

The increased computational cost is significantly bigger than the advantage on accuracy in order to reduce the waste of computing resources in the higher resolution feature map.

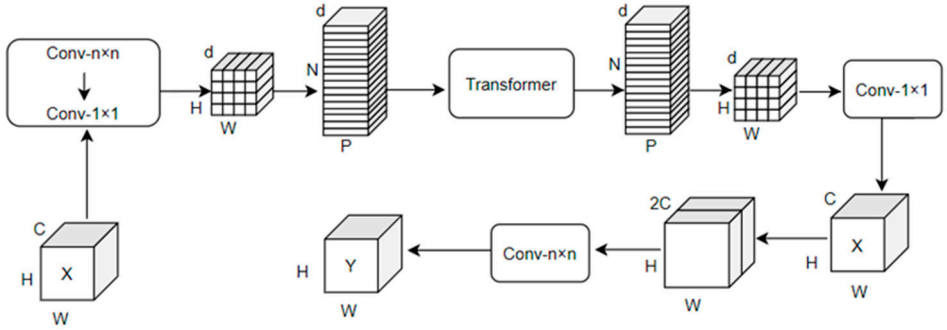


Figure 5. Light ViT block.

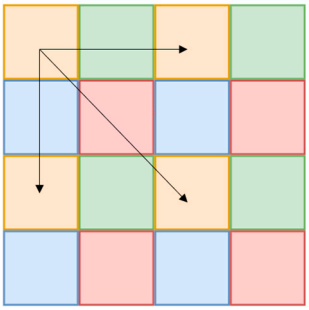


Figure 6. Each token only attends with its own token of the same color.

The advantages of CNN (Spatial Inductive Bias) and ViT (Input Adaptive Weighting and Global Processing) are combined in MobileViT. We introduce Light ViT blocks in particular, which efficiently encode local and global information in tensors. Mobile ViT provides a variety of views on learning global representations. Convolution is typically performed in three steps: expanding, local processing and collapsing. Light ViT blocks use transformers to replace local processing in convolution with global processing. Mobile ViT has a number of advantages. (1) Improved performance: When compared to existing CNN, the Mobile ViT model outperforms them in several mobile vision tasks for a given parameter budget. (2) Generalization capability: The ability of machine learning algorithms to adapt to new examples. Previous ViT variants (convoluted and non-convolutional) performed poorly when compared to CNN; however, MobileViT performed better even with considerable data augmentation. (3) Robustness: The ability of machine learning models to provide very reliable output in unexpected inputs or changing environments.

5. Experimental Procedure and Results

To investigate the feasibility of using Doppler radar to recognize human speech, we measure the micro-Doppler features of human speech. The radar used operates at 24 GHz and has a continuous wave mode. The down-converted signal after I/Q demodulation is digitized by A/D conversion. The data are sent to the computer via a serial protocol at 921,600 bps with a data acquisition time of 2 s each time, and the computer stores the data and the processed feature maps in two files. We use a sampling frequency of 10 KHZ when arranging data acquisition. In the measurement, the radar is kept close to the human neck (about 10–20 cm), as shown in Figure 7, to capture the neck vibration effectively.

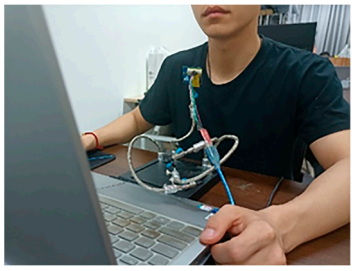


Figure 7. Detecting human voices with radar acquisition.

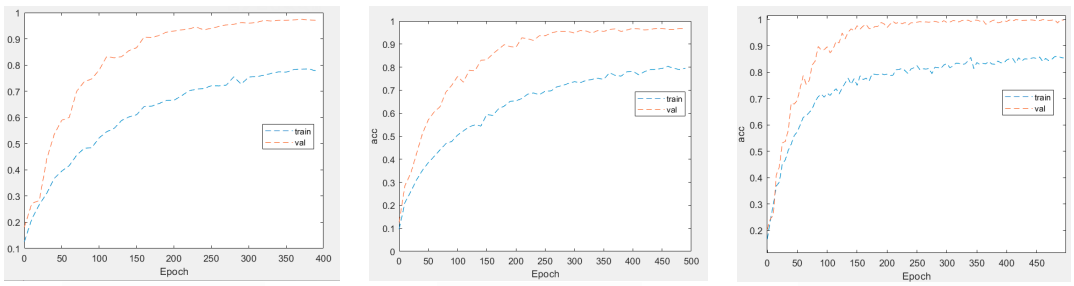
Radar parameters are set to take into account that human voice signal extraction focuses on distance resolution and velocity resolution. In addition to the two resolutions, the sampling time is also an important parameter. If the sampling time is too short, the speech signal features will not be fully captured, and if the sampling time is too long, redundant information other than the target signal will be collected. We chose a miniaturized 24 G radar sensor module of model RKB1102I. The parameter settings for the FM wave together with the data sampling parameters are shown in Table 1.

Table 1. Data sampling parameters.

Parameter	Numeric Value
Operating frequency range (GHz)	24
Phase noise (dBc/Hz)	100
Intermediate-frequency low-frequency cutoff frequency (Hz)	50
Intermediate-frequency high-frequency cutoff frequency (KHz)	100
Sample rate (KHZ)	10

The same digits are measured 200 times, yielding a total of 2000 pictures (10 words × 200 iterations). The number of channels is three, and the image resolution is 461∗369 pixels. We put these micro-Doppler feature maps into the Mobile ViT neural network for training and validation, and for analyzing the experimental results. We split the measurement data into three categories: training (60%), validation (10%) and testing (30%). The words we evaluate are divided into 10 (0–9) categories. In total, 1200 words are utilized as the training set, and 200 words (20 words per number) are randomly chosen as the validation set to ensure their accuracy.

The experimental results are displayed in Figure 8a–c, respectively, for learning rates of 0.005, 0.001 and 0.0001, respectively. The abscissa represents the epoch (number of iterations), the abscissa acc represents the accuracy rate and the blue and red curves, respectively, represent the correct rates for the training set and the verification set.



(**a**). Learning rate 0.005 (**b**). Learning rate 0.001 (**c**). Learning rate 0.0001

Figure 8. Experimental results between different learning rates.

As shown in Table 2, at a learning rate of 0.005, the validation set accuracy of an epoch at 50 times is 58%, and the validation set accuracy at 500 times is 97%; at a learning rate of 0.001, the validation set accuracy of an epoch at 50 times is 59%, and the validation set

accuracy at 500 times is 97%; at a learning rate of 0.0001, the validation set accuracy of an epoch at 50 times is 68%, the accuracy of the validation set at 300 times is 98% and the accuracy of the validation set at 500 times is 99.58%.

Table 2. Validation set accuracy at different learning rates.

	50	100	200	300	400	500
0.005	0.587234	0.806383	0.938298	0.970213	0.972340	0.972789
0.001	0.596372	0.759637	0.900227	0.952381	0.968254	0.970522
0.0001	0.683721	0.878256	0.982101	0.981367	0.989736	0.995863

It is obvious that the accuracy rate, convergence speed and effect are all highest and best when the learning rate is 0.0001. The findings are acquired after training and verification; after around 500 epochs, the accuracy of recognizing each word is about 99%, and the experimental results demonstrate the effectiveness of our method.

In addition to the learning rate, we also check whether the patch size has an impact on the model performance. A larger patch can reduce the computational effort and will be easier to converge. However, if the patch size is too large, more semantic information will be ignored, which will affect the performance of the model. We test the model with two different patch sizes, where the patch size of configuration A is {2, 2, 2} and the patch size of configuration B is {8, 4, 2}, which correspond to the patch sizes used for feature maps with a down-sampling multiplicity of 8, 16 and 32, respectively. In a task like image classification, which does not require high local fine-grained information features, different patch sizes do not have much impact on the accuracy, but can improve the inference speed by a small margin.

In [6], we see that the authors use a VGG-16 [18] convolutional neural network to train the speech feature maps. So, we obtain the difference between the two by conducting a comparison experiment. Firstly, we input the same parameters to the convolutional neural network VGG-16 for training. Secondly, we compare the best data with MobileViT data, as shown in Figure 9. The accuracy of the two is compared on the left, and the loss value of the two is compared on the right. The red line represents Mobile ViT, while the blue line represents VGG-16. We can plainly observe that the accuracy of Mobile ViT at the final convergence is greater than that of VGG-16, and that the Mobile ViT loss value is less than that of VGG-16. As previously stated, the lightweight neural network Mobile ViT cannot be inferior to convolutional neural networks in the case of minimal training parameters, and it clearly outperforms them in the accuracy and loss value. And the transformer model excels in the area of vision. However, the drawback that the model parameters are too vast and the computational power requirement is too high limits its practical application. We have yet to witness the application of lightweight neural networks in the area of voice recognition, and this experiment reveals that Mobile ViT's good performance, stronger generalization ability and excellent durability are sufficient to demonstrate its performance.

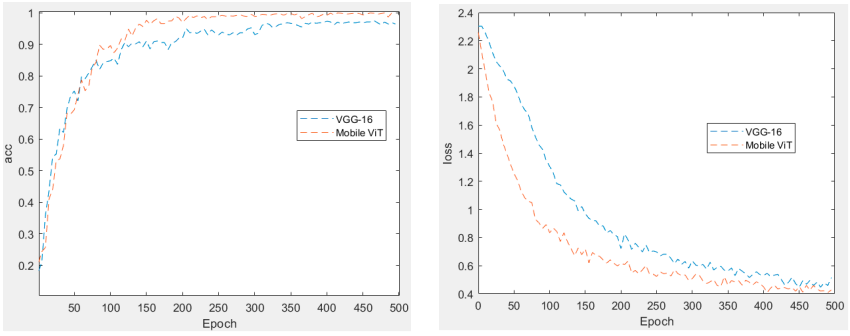


Figure 9. Comparison of Mobile ViT and VGG-16 experiments.

6. Conclusions

In this study, we introduce non-contact Doppler radar to detect a human voice. The micro-motion features of human vocalizations are captured using Doppler radar, and feature maps are generated, which are recognized using the Mobile Vision Transformer neural network. Through experiments, we show that the model we use as a speech word classification can classify about two percentage points higher than the traditional CNN convolutional model, reaching an accuracy of 99.5%.

As this study, our work is limited in the following aspects. In this experiment, we cannot guarantee that the speech feature maps of other subjects can be accurately recognized in the same training mode. Secondly, in the radar detection stage, it is not possible to detect a human voice at a long distance. In future experiments, we hope to find a general method that can recognize the speech signal of all people and is not limited to words only.

Author Contributions: Methodology, Y.G. (Yongfu Geng); Software, N.L. and J.C.; Formal analysis, Y.G. (Yang Gao); Investigation, Q.D. and D.L.; Supervision, W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Muda, L.; Begam, M.; Elamvazuthi, I. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. *arXiv* **2010**, arXiv:1003.4083.
2. Seflek, I.; Yaldiz, E. A comprehensive study about low-cost and limited bandwidth FMCW bio-radar: Detailed analyses on vital signs measurements. *Frequenz* **2022**, *76*, 505–520. [CrossRef]
3. Li, Z.-W. Millimeter Wave Radar for detecting the speech signal applications. *Int. J. Infrared Millim. Waves* **1996**, *17*, 2175–2183. [CrossRef]
4. Staderini, E.M. UWB radars in medicine. *IEEE Aerosp. Electron. Syst. Mag.* **2002**, *17*, 13–18. [CrossRef]
5. Dong, Y.; Yao, Y.D. Secure mmWave-Radar-Based Speaker Verification for IoT Smart Home. *IEEE Internet Things J.* **2021**, *8*, 3500–3511. [CrossRef]
6. Khanna, R.; Oh, D.; Kim, Y. Through-Wall Remote Human Voice Recognition Using Doppler Radar with Transfer Learning. *IEEE Sens. J.* **2019**, *19*, 4571–4576. [CrossRef]
7. Sun, Z.; Liu, K.; Tan, A.; Yan, T.; Yang, F. Activity Recognition Through Micro-Doppler Image Based on Improved Faster R-CNN. In *Proceedings, Part III, Intelligent Equipment, Robots, and Vehicles: 7th International Conference on Life System Modeling and Simulation, LSMS 2021 and 7th International Conference on Intelligent Computing for Sustainable Energy and Environment, ICSEE 2021, Hangzhou, China, October 22–24, 2021*; Springer: Singapore, 2021; pp. 289–298. [CrossRef]
8. Wang, Z.; Zhang, Z.; Dong, L.; Xu, G. Jitter Detection and Image Restoration Based on Generative Adversarial Networks in Satellite Images. *Sensors* **2021**, *21*, 4693. [CrossRef] [PubMed]
9. Han, Y.; Huang, L.; Hong, Z.; Cao, S.; Zhang, Y.; Wang, J. Deep Supervised Residual Dense Network for Underwater Image Enhancement. *Sensors* **2021**, *21*, 3289. [CrossRef] [PubMed]
10. Córdova, M.; Pinto, A.; Hellevik, C.C.; Alaliyat, S.A.-A.; Hameed, I.A.; Pedrini, H.; Torres, R.d.S. Litter Detection with Deep Learning: A Comparative Study. *Sensors* **2022**, *22*, 548. [CrossRef] [PubMed]
11. Mehta, S.; Rastegari, M. MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Trans-former. *arXiv* **2021**, arXiv:2110.02178.
12. Kirov, C. A Bayesian Approach to Speech Production. Ph.D. Thesis, Johns Hopkins University, Baltimore, MD, USA, 2014.
13. Qin, X.; Deng, B.; Wang, H. Micro-Doppler Feature Extraction of Rotating Structures of Aircraft Targets with Terahertz Radar. *Remote Sens.* **2022**, *14*, 3856. [CrossRef]
14. Massaro, A.; Dipierro, G.; Cannella, E.; Galiano, A.M. Comparative Analysis among Discrete Fourier Transform, K-Means and Artificial Neural Networks Image Processing Techniques Oriented on Quality Control of Assembled Tires. *Information* **2020**, *11*, 257. [CrossRef]
15. Savchenko, A.V. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In Proceedings of the 2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, Serbia, 16–18 September 2021; pp. 119–124. [CrossRef]
16. Xu, H.; Chen, B.; Qin, J. A CNN-Based Length-Aware Cascade Road Damage Detection Approach. *Sensors* **2021**, *21*, 689. [CrossRef] [PubMed]

17. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 9992–10002.
18. Gumelar, A.B.; Yuniarno, E.M.; Anggraeni, W.; Sugiarto, I.; Mahindara, V.R.; Purnomo, M.H. Enhancing Detection of Pathological Voice Disorder Based on Deep VGG-16 CNN. In Proceedings of the 2020 3rd International Conference on Biomedical Engineering (IBIOMED), Yogyakarta, Indonesia, 6–8 October 2020; pp. 28–33.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.