
Reproduced under a Creative Commons License.
https://creativecommons.org/licenses/by/4.0/

This is the author version of the work. There may be differences between this version and the published version. You are advised to consult the published version if you want to cite from it:
https://doi.org/10.1109/RADAR54928.2023.10371189

https://eprints.gla.ac.uk/304502/

Deposited on 11 August 2023
Abstract—Speaker recognition (SR) from speech can help determine the environmental context in multi-talker conversational scenarios to enable the design of context-aware multi-modal hearing assistive technology. In this paper, we argue that the use of wireless sensors such as radars can offer many benefits over conventional audio and visual sensors, such as not being afflicted by privacy and environmental issues, e.g., improper lighting, environmental noise, and potential security concerns of audio and video channels. Radar is relatively less explored and has many advantages over other contactless approaches, such as being more compact compared to RFID and having a better range and resolution than ultrasound and microwave sensors. To this end, we propose the use of ultrawideband radar coupled with a deep learning model for SR from silent speech to enable the design of future context-aware multimodal hearing assistive technology. We collected a dataset from five individuals with origins in Europe, Asia, and the United Kingdom. We obtained an average performance of approximately 82% in recognising an unknown person from a set of known people. This demonstrates that the radar has good potential to be used for privacy-preserving SR in multi-talker environments where audio-visual and other contactless techniques have limited capabilities.

Index Terms—UWB radar, speaker recognition, silent speech recognition, hearing Assistive technologies, contactless sensing.

I. INTRODUCTION

Speaker recognition is the process of identifying or verifying the identity of an individual based on their voice [1]. It has a wide range of applications including security, authentication, and access control. Traditional SR systems rely on audio signals captured by microphones, which can be affected by background noise, distance, and some environmental factors [2], [3]. SR using audio-visual signals, which involves combining information from both audio and visual cues to identify or verify the speaker’s identity, has gained increasing attention in recent years [4], [5]. While SR using audio-visual signals has the potential to improve the accuracy and reliability of SR systems [6]–[8], it is important to consider the limitations of this technology [9]. These limitations include lighting conditions, environmental noise, occlusions, facial expressions, privacy concerns, equipment costs, and algorithmic bias. By understanding these limitations, researchers and practitioners can develop more effective and reliable SR systems [10]–[12]. In recent years, ultra-wideband (UWB) radar technology has emerged as a promising alternative to traditional microphone-based SR systems. UWB radar technology is capable of detecting and tracking human movements and can provide information about the individual’s vocal tract and speech characteristics, which can be used to identify or verify the speaker’s identity. UWB radar technology has several advantages over traditional microphone-based systems. One of the main advantages is its ability to operate in noisy environments, as it is not affected by background noise. It can also penetrate through walls and other obstacles, making it suitable for use in surveillance and security applications.
Another advantage of UWB radar is its ability to detect small movements and vibrations, such as those caused by the human vocal tract during speech. Overall, non-invasive radar-based SR technology offers a number of advantages over traditional invasive methods of SR. It is non-invasive, accurate, contactless, versatile, cost-effective, portable, and capable of providing real-time results, making it an attractive option for a wide range of applications. In this paper, we proposed a modality to adopt UWB radar for human identification. Total of five speakers from Europe, China, Pakistan, and the United Kingdom volunteered to take a data set consisting of fourteen different words commonly used in emergency conditions and five English vowels as given in Table II. To the best of our knowledge, this is the first contribution of this kind to recognise a speaker among other known speakers or recognise an unknown person from other known speakers from their speaking styles and facial features detected by radar. In the summary, there are two main contributions of our work:

1. We adopt a contactless method of human identification using UWB signal. The work frequency band of UWB is allowed from 3 GHz to 10 GHz, less distorted by the 2.4 GHz broad Wi-Fi system and other wireless signals.
2. We collected a multimodal human speech dataset including vowels and words by using RF, audio and video signals, from radar and kinect v2 sensor.

II. RELATED WORK

UWB radar technology uses short pulses of electromagnetic waves with a very wide frequency spectrum to detect different changes in objects. The pulses are typically in the range of a few nanoseconds to a few microseconds and are transmitted from an antenna [19]–[21]. The pulses reflect off objects and are received by a receiver antenna. The time delay between the transmitted and received pulses is used to calculate the distance to the object, while the amplitude and phase of the received pulse provide information about the object’s properties, such as lips and vocal tract movements, shape of face and speech characteristics. While exploring silent speech interfaces based on frequency-modulated continuous-wave (FMCW) radar in [13], the author recognized 13 words spoken by four different speakers of Portuguese origin with 84.5% average accuracy. Another step frequency modulated continuous wave (SFCW) radar used for silent speech recognition, total of forty German words including nouns, adjectives, verbs, and digits spoken by five male native German speakers aged between 28 and 36 years old. Results accuracy for word recognition is 76.50% and 68.18% obtained using the headset and the tape, respectively [14]. While recognizing forty German words and zero to nine German digits from two persons with the help of SFCW radar, author obtained recognition accuracies of 99.17% and 88.87% for the speaker-dependent multi-session and inter-session accuracy respectively (average accuracy 94.02%) [15].

RFID tags have been used for the identification of people for static and dynamic users [17], author collected walking and body information for identification purposes. Other studies includes microwave sensors [18] and Radars [22], [23], details provided in Table I.

III. METHODOLOGY

A. Data Collection

In this section, we discuss our data collection strategy and the setup in which data was collected. We start by first describing our experimental setup.

1) Experimental Setup: This study involves five volunteers from various countries, including Europe, China, Pakistan, and the United Kingdom. Due to the volunteers’ distinct accents and body sizes, an adjustable table was implemented to ensure a consistent distance between the speaker’s head and the sensors. The dataset collected from the volunteers includes Xethru UWB radar signal for lip motion and audio signal, with the setup shown in Figure 1. In this paper we are using only RF signal from radar, audio-visual details from Kinect 2 sensor is for ground truth and future data usage for further results.

2) Data Collection Strategy: To make the data collection process simple, we design a system that can generate the sound of "speaking" and "stop" with timestamps. At the same time, all the other sensors including Kinect and Xethru radar will be activated for recording. While the sensors were running, the volunteer can read the vowels and words shown on the screen. We manually selected five vowels and fourteen words for reading on a computer screen. All volunteers were informed of the potential risks of the experiment and signed a consent form. The details of data collection are referred to
TABLE I: Summary of contactless sensors used for corpus detection and speaker recognition other than audio-visual techniques.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Year</th>
<th>Sensing technology</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>[13]</td>
<td>2022</td>
<td>impulse radio ultra-wideband radar</td>
<td>German words</td>
</tr>
<tr>
<td>[14]</td>
<td>2022</td>
<td>SFCW radar</td>
<td>German words</td>
</tr>
<tr>
<td>[15]</td>
<td>2022</td>
<td>SFCW radar</td>
<td>Forty German words and digits zero to nine</td>
</tr>
<tr>
<td>[16]</td>
<td>2016</td>
<td>Impulse radio ultra-wideband radar</td>
<td>English</td>
</tr>
<tr>
<td>[17]</td>
<td>2018</td>
<td>RFID</td>
<td>N/A</td>
</tr>
<tr>
<td>[18]</td>
<td>2018</td>
<td>Microwave</td>
<td>German</td>
</tr>
</tbody>
</table>

TABLE II: List of Corpus used by five different participants.

<table>
<thead>
<tr>
<th>Type</th>
<th>Corpus</th>
<th>Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Order, Assist, Help, Ambulance, Bleed, Fall, Shock, Medical, Sanitize, Doctor, Rescue, Emergency, Heart, Break</td>
<td>User 1-5</td>
</tr>
<tr>
<td>Vowel</td>
<td>A, E, I, O, U</td>
<td>User 1-5</td>
</tr>
</tbody>
</table>

in [24]. No privacy-related information was collected in our dataset and appropriate anonymization was applied to ensure the confidentiality of the participant’s data.

B. Data Preprocessing

The UWB radar we used is based on impulse modulation. The response of radar signal is completely dependent on impulse delay from transmitting to receiving back, which can be represented as the Eq. 1

\[ s(\tau, t) = \sum_{i=1}^{N} a_i(\tau, t) e^{-j2\pi \frac{a(t) + d_i(t)}{\lambda}} \]  

where the \( \tau \) and \( t \) represent impulse indication in fast-time range and time in frame range, \( N \) is the index of the impulse; \( a_i(\tau, t) \) represents the complex attenuation factor of the related \( t \) slot and \( i^{th} \) impulse; \( \lambda \) represents the wavelength of the UWB signal. After the data collection, we first extract the IQ radar signal from binary file and reshape it to the frame which contains fast-time and slow-time dimensions. The frames are identical to the range-time response image because the fast-time domain represents the indicator of time of flight (ToF), which can be calculated to distances. Then, we adopted the moving target indication of frames in sequence to filter the static object out. Then, we calculated the STFT results in the fast-time dimension which is close to the sensing range, and add them together. The spectrograms can be viewed in Figure 2.

C. Machine Learning-based Data Analytics

We formulated the problem of SR as a classical image classification problem. The objective is to learn a latent function \( f : x \rightarrow y \) to map \( x \) to \( y \), where \( x \) and \( y \) represent the input spectrogram and its corresponding label, respectively. We leveraged a widely used convolutional neural network...
architecture known as VGG16 for learning $f$. Since this model is proposed for 1000 classes, we removed its output layer and stacked a convolutional layer on top of it, having a depth of 32 and a kernel size of $3 \times 3$. We then added a max pooling layer that applies a pooling operation with a kernel size of $2 \times 2$. After that, we added three fully connected layers having 128, 64, and 32 units, respectively. Finally, we used softmax to get the probability vector of size $1 \times 5$. We used pre-training with ImageNet weights that allow the model to extract relevant features from the input spectrograms.

IV. RESULTS AND DISCUSSIONS

A. Data Description

We have data collected from five different persons having native languages from Europe, China, Pakistan, and the United Kingdom. We asked the speakers to speak five English vowels and fourteen different words (a summary is presented in Table II). To ensure a fair comparison and to avoid data bias, we have collected an equal number of samples for each participant. However, note that the resulting signals for each of the speakers vary in terms of time duration (due to demographic variations).

B. Model Training Setup

We partitioned the collected data into training and testing sets using a split of 80% and 20%, respectively. Furthermore, to ensure an efficient training, we applied 3x data augmentation by using widely used techniques such as shear, horizontal flip, and zooming. We then trained the DL models using training data and evaluated their performance using test data. All images were resized to the same size (i.e., $224 \times 224$) prior to training and testing. Since our data size is small, we used pre-training (with ImageNet weights) and fine-tuned the models using our training data. All models were trained using a batch size of 64 and a learning rate of $1e^{-3}$.

C. Results for Speaker Recognition

The performance curves depicting the model training in terms of accuracy and loss for both training and validation (test) data is shown in Figure 3. It is evident from the figure that the underlying model was able to learn latent features from silent speech signal to recognise the speaker. Also, we can see from the figure that the model tends to show an overfitting behaviour (that highlights that the training samples need to be increased for more robust training). We note that the focus of this paper is to demonstrate the feasibility of using non-speech signals to perform SR. The confusion matrix that illustrates the SR performance of the trained model is shown in Figure 4. The figure reveals that the model provided superior performance for “speaker 4” while it provided the worst performance for “speaker 5”. The performance of the model for “speaker 1” and speaker 2 is almost similar and for “speaker 3”, the model provided comparatively lower performance than speaker “1”, “2”, and “4” but is considerably higher than “speaker 5”. A summary and comparison of speaker-wise recognition performance in terms of different performance metrics, including precision, recall, and F1-score is presented in Table III. The table also supports the analysis from Figure 4.

V. CONCLUSIONS

In this paper, we present the feasibility of using data collected through radar to perform speaker recognition. Specifically, we collected a dataset from five different persons speaking different native languages and asked them to speak five English vowels and fourteen different words. We then analyzed the collected dataset using a convolutional neural network (CNN) that provided an overall average performance of more than 82%. While we got a maximum performance of more than 91% for a single speaker (“speaker 4”). Our results demonstrate that the radar possesses great potential to be used for the speaker recognition task that offers a number of advantages over conventional audio-visual signals, including preserving the privacy of users. In our future work, we plan to improve the performance of the developed system along with increasing the size of the dataset.
ACKNOWLEDGMENTS

This work was supported in parts by Engineering and Physical Sciences Research Council (EPSRC) grants EP/T021020/1 and EP/T021063/1, and by the RSE SAPHIRE grant.

REFERENCES


