
There may be differences between this version and the published version. You are advised to consult the publisher’s version if you wish to cite from it.

http://eprints.gla.ac.uk/301924/

Deposited on 03 July 2023

Enlighten – Research publications by members of the University of Glasgow
http://eprints.gla.ac.uk
Integrating RF-Visual Technologies for Improved Speech Recognition in Hearing Aids

Zikang Chen  
James Watt School of Engineering  
University of Glasgow  
Glasgow, UK  
2429295c@student.gla.ac.uk

Chong Tang  
James Watt School of Engineering  
University of Glasgow  
Glasgow, UK  
chong.tang@glasgow.ac.uk

Yao Ge  
James Watt School of Engineering  
University of Glasgow  
Glasgow, UK  
2288980g@student.gla.ac.uk

Muhammad Imran  
James Watt School of Engineering  
University of Glasgow  
Glasgow, UK  
muhammad.imran@glasgow.ac.uk

Qammer H. Abbasi  
James Watt School of Engineering  
University of Glasgow  
Glasgow, UK  
qammer.abbasi@glasgow.ac.uk

Abstract—Traditional hearing aids solutions are based on audio and visual information, which has limited effectiveness in challenging scenarios, such as noisy environments and obstacles. Additionally, their level of comfort and privacy protection is often unsatisfactory. In this case, there has been increasing interest in recent years, aiming to develop a contactless and privacy-preserving alternative using radio frequency (RF) signals. However, the RF-based approach requires a large amount of training data to support its reliability and accuracy, which is undoubtedly very time-consuming and labour-intensive. To address these limitations and benefit future hearing aids with RF sensing, the fusion of multiple modalities provides us with a solution. In this paper, we propose an RF-Visual based speech recognition system based on the fusion of visual and RF information, which is based on a multi-input convolutional neural network (CNN) and can achieve up to 87.55% recognition accuracy. We have comprehensively compared and evaluated the system performance with single and multiple modalities, and can conclude the proposed RF-Visual-based SR system has great potential for advancing hearing aid technology.

Index Terms—speech recognition, RF-based speech recognition, Radio frequency sensing

I. INTRODUCTION

According to a report from the World Health Organization [1], the global population is expected to reach 10 billion by 2050, and the number of people suffering from hearing loss is predicted to reach 2.5 billion. This highlights the need for effective and robust hearing aid systems. Moreover, the COVID-19 pandemic shows a negative impact on people with hearing loss in understanding others’ speech. Since masks obstruct the lip reading and affect the performance of audio-based hearing aids [2]. In recent years, the emergence of RF-based lip-reading systems [3] provides a promising alternative solution for the future of hearing aids. Unlike audio-based systems, RF signals are unaffected by acoustic conditions, and their penetrability enables them to work through obstacles and walls. Consequently, hearing aid systems that incorporate RF signals can better resist the variations of environmental factors than traditional solutions. However, the robustness and accuracy of RF-based speech recognition often require the support of vast amounts of measurement data, which can be time-consuming and laborious. Developing effective signal processing and denoising algorithms is also necessary to mitigate interference from other movements during lip-reading, which could be quite challenging.

To address these limitations, the fusion of visual and RF information could provide a more robust and accurate solution than relying on just a single modality. As a result, we propose a novel RF-Visual-based SR system for hearing aids in this paper that can gather information across visual and RF modalities and achieve considerable recognition improvements over single-modal SR systems. Especially, the RF sensor used in this work is impulse radio ultra-wideband (UWB) radar, which has been extensively deployed in various applications, such as activity recognition [4], [5], localization [6], and vital sign detection [7]. It features low power consumption, high accuracy, high data acquisition rate and high noise resistance, which plays an important role in this work. In the end, we have conducted a series of experiments based on our multi-modal dataset [8] to validate the performance of the proposed system, where the experimental results have demonstrated the effectiveness of the proposed fusion system.

Fig. 1: Data collection setup
II. METHODOLOGY

The procedure of the whole system is shown in Fig. 1. This section will introduce the principle of the sensing methods, pre-processing method and the feature extraction method involved.

A. Data pre-processing

UWB radar is a type of impulse-based radar that uses very short and low-power pulses to transmit the radar signal. UWB is defined as where the bandwidth is larger than 500 MHz which allows UWB radar to detect objects with high resolution and low power [9]. Therefore UWB radar can be used in multiple detecting applications such as through-wall object detection and tracking, and human motion analysis. The distance of the object is gathered by measuring the Time of Flight (ToF) of the reflected signal $d = c \times \tau^2$, where $d$ is the distance, $c$ is the speed of light and $\tau$ is the time delay of the received signal. The distance distinguish resolution is given by $R_{\Delta} = \frac{c}{2B}$, where $R_{\Delta}$ is the resolution, $B$ is the bandwidth of the impulse signal. Larger bandwidth provides more precise distance detection.

For the UWB data, the in-phase and quadrature (IQ) data of each pulse is extracted from the radar signal. The response of reflected radar signal depends on impulse delay from detected object distance to radar, which can be represented as the Equation 1:

$$s(\tau, t) = \sum_{i=1}^{N} a_i(\tau, t) e^{-j2\pi \frac{d(i)+d(\tau)}{\lambda}}$$

where $t$ represents frame time, $\tau$ and $i$ represents ToF delay and multipath index, $\lambda$ represents the wavelength of the UWB signal. After raw data collection, the moving target indication (MTI) technique is applied to filter out static targets, leaving only the moving targets. We eliminate the signal of a specific time interval at a particular weight to significantly suppress static fake peaks that are not related to human motion in this case. The Short-Time Fourier Transform (STFT) is then used to create a time-frequency representation of the signal in the form of a spectrogram. By applying this pre-process method, the UWB data is transformed into a spectrogram for feature extraction using deep learning models.

For the concern of the data size and the computation cost, the video data is first processed to extract the lip area of each participant as shown in Fig. 2. The video frames are then resized to the same size and converted into greyscale for training.

The audio data is converted to Mel-Spectrogram to train the network. The Mel-Spectrogram is a type of spectrogram that is closer to how humans perceive sound than the standard spectrogram since the perception of frequency is not linear and is more sensitive to low-frequency signals than high-frequency signals.

III. EVALUATION AND DISCUSSION

A. Experimental Setup

The data collection setup is shown in Fig. 3. In this paper, we utilized three different sensors to collect a multi-modal dataset. The Kinect v2 is used to collect the video and audio data, and the XeThru X4M03 to collect UWB data. The speech data consists of 15 different words, each word was repeated 10 times by 5 participants. Details of the dataset are referred from [8]. The UWB radar setup is shown in Table I.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>UWB Impulse radar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Center Frequency</td>
<td>8.745 GHz</td>
</tr>
<tr>
<td>Sampling Frequency</td>
<td>23.328 GHz</td>
</tr>
<tr>
<td>Frame Rate</td>
<td>300 Hz</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>1.5 GHz</td>
</tr>
<tr>
<td>Antennas</td>
<td>1 Tx / 1 Rx</td>
</tr>
</tbody>
</table>

TABLE I: Parameters of UWB impulse radar

The Kinect v2 has a 1080p resolution RGB camera and a 512 × 424 resolution depth camera. It captures 1920 × 1080 video frames.

Fig. 3: Data collection setup
at 30 fps. And a 4-microphone array to collect the audio data in 256 kbps and 16-bit depth.

B. Result and Discussion

The comparison of results in Fig.4 shows the performance of three modalities and the fusion model, respectively. Specifically, the accuracy of video incorporating UWB data SR has achieved 87.55%, improving uni-modal SR only using UWB data or video by around 1.5% to 6.5%. It can be seen that the proposed system has shown sufficient capability in SR scenarios and has great potential to be implemented in future hearing aids technologies.

From the bar chart Fig. 5, it can be seen that the recognition accuracy of the UWB radar decreases from 87.5% to 80.97% as the number of people increases. This may be due to the differences in individual characteristics and speaking habits between each participant, leading to the decrease of performance in multi-user situations. Which could be the main challenge of UWB-based SR. However, the fusion model of UWB and other sensing sources shows improvement, indicating that the RF-Visual based SR supplements the information for each uni-modal.

IV. Conclusion

In summary, according to experimental comparison and analysis, this paper has comprehensively demonstrated that the fusion of UWB and video modalities can effectively improve SR accuracy compared to uni-modal systems. Further, the mutual complementary of different information can greatly improve the robustness of the system in various complicated environments, which is crucial for real-world hearing aids. Especially, the incorporation of RF sensing demonstrated promising outcomes, which indicate the potential of the RF sensing fusion model in SR application. Overall, the proposed system provides an effective and feasible framework for the development of future hearing aid systems. Meanwhile, there are several main challenges of this model:

- The accuracy of SR based on UWB radar needs improvement to achieve a precise SR.
- The generalization ability of UWB radar for different recognition objects is not adequate. We need to improve the UWB recognition model to enhance its generalization for different targets.
- The use of Deep Neural Network requires computing power, and the model and the data pre-process procedure needs to be improved to achieve real-time recognition.

Our future work will focus on addressing the limitation of this RF-Visual based system, to develop a precise and real-time recognition model involving UWB radar.

ACKNOWLEDGMENT

This work was funded 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