Microwave Speech Recognizer Empowered by a Programmable Metasurface

Hengxin Ruan, Siyuan Jiang, Hongrui Zhang, Hanting Zhao, Zhuo Wang, Shengguo Hu, Jun Ding, Tie Jun Cui, Philipp Del Hougne, Lianlin Li

Microwave Speech Recognizer Empowered by a Programmable Metasurface

Hengxin Ruan
peking university
Siyuan Jiang
peking university
Hongrui Zhang
peking university
Hanting Zhao
peking university
Zhuo Wang
peking university
Shengguo Hu
peking university
Jun Ding
East China Normal University
Tie Jun Cui
Southeast University https://orcid.org/0000-0002-5862-1497
Philipp del Hougne
CNRS https://orcid.org/0000-0002-4821-3924
Lianlin Li (lianlin.li@pku.edu.cn)
peking university https://orcid.org/0000-0001-9394-3638

Article

Keywords:

Posted Date: June 15th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1682756/v1

Microwave Speech Recognizer Empowered by a Programmable Metasurface

Hengxin Ruan[1,2+], Siyuan Jiang[1+], Hongrui Zhang[1+], Hanting Zhao[1+], Zhuo Wang[1], Shengguo Hu[1], Jun Ding[3], Tie Jun Cui[4,*], Philipp del Hougne[5,*], Lianlin Li[1,*]

[1]State Key Laboratory of Advanced Optical Communication Systems and Networks, School of Electronics, Peking University, Beijing 100871, China;
lianlin.li@pku.edu.cn

[2]Peng Cheng Laboratory, 518000, Shenzhen, Guangdong, China

[3]Key Laboratory of Polar Materials and Devices, School of Physics and Electronic Sciences, East China Normal University, Shanghai 200241, China

[4]State Key Laboratory of Millimeter Waves, Southeast University, Nanjing 210096, China;
tjcui@seu.edu.cn

[5]Univ Rennes, CNRS, IETR - UMR 6164, F-35000 Rennes, France
philipp.del-hougne@univ-rennes1.fr

[+] These authors contributed equally to this work.

Abstract

We present an experimental prototype of a microwave speech recognizer empowered by a programmable metasurface that can recognize voice commands and speaker identities remotely even in noisy environments and if the speaker's mouth is hidden behind a wall or face mask. Thereby, we enable voice-commanded human machine interactions in many important but to-date inaccessible application scenarios, including smart health care and factory scenarios. The programmable metasurface is the pivotal hardware ingredient of our system because its large aperture and huge number of degrees of freedom allows our system to perform a complex sequence of tasks, orchestrated by artificial-intelligence tools. First, the speaker's mouth is localized by imaging the scene and identifying the region of interest. Second, microwaves are efficiently focused on the speaker's mouth to encode information about the vocalized speech in reflected microwave biosignals. The efficient focusing on the speaker's mouth is the origin of our system's robustness to various types of parasitic motion. Third, a dedicated neural network directly retrieves the sought-after speech information from the measured microwave biosignals. Relying solely on microwave data, our system avoids visual privacy infringements. We expect that the presented strategy will unlock new possibilities for future smart homes, ambient-assisted health monitoring and care, smart factories, as well as intelligent surveillance and security.

Introduction

Voice commands are arguably the most natural approach to human-machine interfaces because speech is the most direct communication method between humans. However, the most obvious approach for the machine to capture voice commands, namely the acquisition of the acoustic signals that are the primary information carrier, precludes many important deployment scenarios. On the one hand, voice commands can be drowned in ambient noise under operation in noisy environments such as streets, public transport, or restaurants. On the other hand, it is impossible to operate under silent-speech requirements[1] which may arise to preserve privacy, for use in quiet settings like libraries, or through verbally impaired users (e.g., post-laryngectomy). Therefore, a wide variety of indirect secondary carriers of information about voice commands has been explored. Many of these techniques achieve high accuracy at the price of being highly invasive because they rely on placing sensors (e.g., magnetic[2], surface electromyographic[3], infrared[4], electropalatographic[5], electromagnetic[6,7]) directly on the human's body to detect subtle vibrations that are correlated with the speech production. Obviously, such contact-based approaches are oftentimes inconvenient and, moreover, incompatible with large-scale deployment in our daily lives. Similar limitations apply to contactless radar-like approaches based on the emission and reception of acoustic[8,9] or electromagnetic[10,11] waves in the very close proximity (a few centimeters) of the speaker's face. A popular contactless technique for speech recognition that can operate remotely uses optical image sequences as secondary information carrier to recognize speech by analyzing lip or face motion[12,13]. However, such visual speech-recognition methods fail under unfavorable lighting conditions such as darkness as well as when the line of sight from the camera to the speaker's mouth is obstructed by a wall or, more recently, a face mask. In addition, the acquisition of camera images risks to infringe the user's privacy. Bi-modal speech recognition approaches, combining, for instance, acoustic and visual inputs[14], benefit from richer input information but are also unable to tackle, for instance, the recognition of speech uttered behind a face mask in a noisy environment. An ideal voice-commanded human-machine interface would remotely capture relevant biosignals in a robust, noise-resilient, and privacy-respecting manner while being cheap, consuming little power, and being easy to deploy in our daily lives, even when the speaker's mouth is hidden behind an optically opaque tissue or wall.

The use of microwaves as remote contactless secondary information carrier of voice commands is predestined to meet this formidable challenge. The ability of microwaves as remote, non-ionizing sensing technology to penetrate through visually opaque layers is well known, for example, from airport security checks[15]. However, capturing microwave biosignals that bear sufficient information about the sought-after voice commands is itself very challenging because most signal variations may be due to motion that is not related to speech. Therefore, it is of pivotal importance to focus the microwaves on the speaker's mouth, which in turn requires real-time tracking of the mouth as a pre-requisite. By focusing on the mouth, the weight of reflections from the region of interest (ROI) in the measured signals is drastically increased. Notable results toward that goal were reported in Ref.[16] through a multiple-input-multiple-output (MIMO) beamforming approach at WiFi frequencies in the 2.4 GHz range. However, the underlying multi-channel coherent emission is costly and cumbersome because it requires synchronized sources and individual IQ modulation on each channel. Moreover, using only a few antennas, the setup's degrees of freedom were quite limited, resulting, for instance, in a focal spot that was so large that even winks perturbed the measured signals. A large antenna array emitting coherently controlled wavefronts would be necessary to efficiently localize and focus on the speaker's mouth. Yet this hardware is too costly and power hungry for widespread deployment in human-machine interaction.

In this article, we show that a deep-learning-controlled programmable metasurface fully reaps the benefits of microwave speech recognition with a drastically simpler hardware. Our programmable metasurface[17] is an array of 1024 meta-atoms with individually controllable reflection properties, fed by a single source. Compared to a conventional antenna array comprising a few antennas, we have thus three orders of magnitude more degrees of freedom and, moreover, a much larger aperture. Building on recent results from intelligent computational meta-imaging[18], these advantages allow us to localize and focus on the speaker's mouth with unprecedented efficiency. We use this ability to prototype a voice-commanded human-machine interaction scenario in which a speaker who is hidden behind a wall commands a mechanical hand. Our system is capable of tracking a moving speaker in real time, dynamically generating suitable spatial beams for focusing on the speaker's mouth and interpreting the measured biosignals with a deep-learning technique. We also demonstrate multi-speaker listening. Moreover, we shed new light on the mechanisms through which speech information is encoded in microwave biosignals: we demonstrate that, besides the obvious reflection off the mouth, the probing

microwave signals partially penetrate through the skin and are affected by the tongue and other vocal entities. Finally, we evidence that our system can also be utilized as biometric identification technology because of the individual manner in which each subject utters speech. Our experimental results enable voice-commanded human-machine interaction at minimal cost in a plethora of challenging and to-date inaccessible scenarios such as health care for assisted living (see **Fig. 1a** for an example); our results may also be valuable in security applications requiring intelligent surveillance.

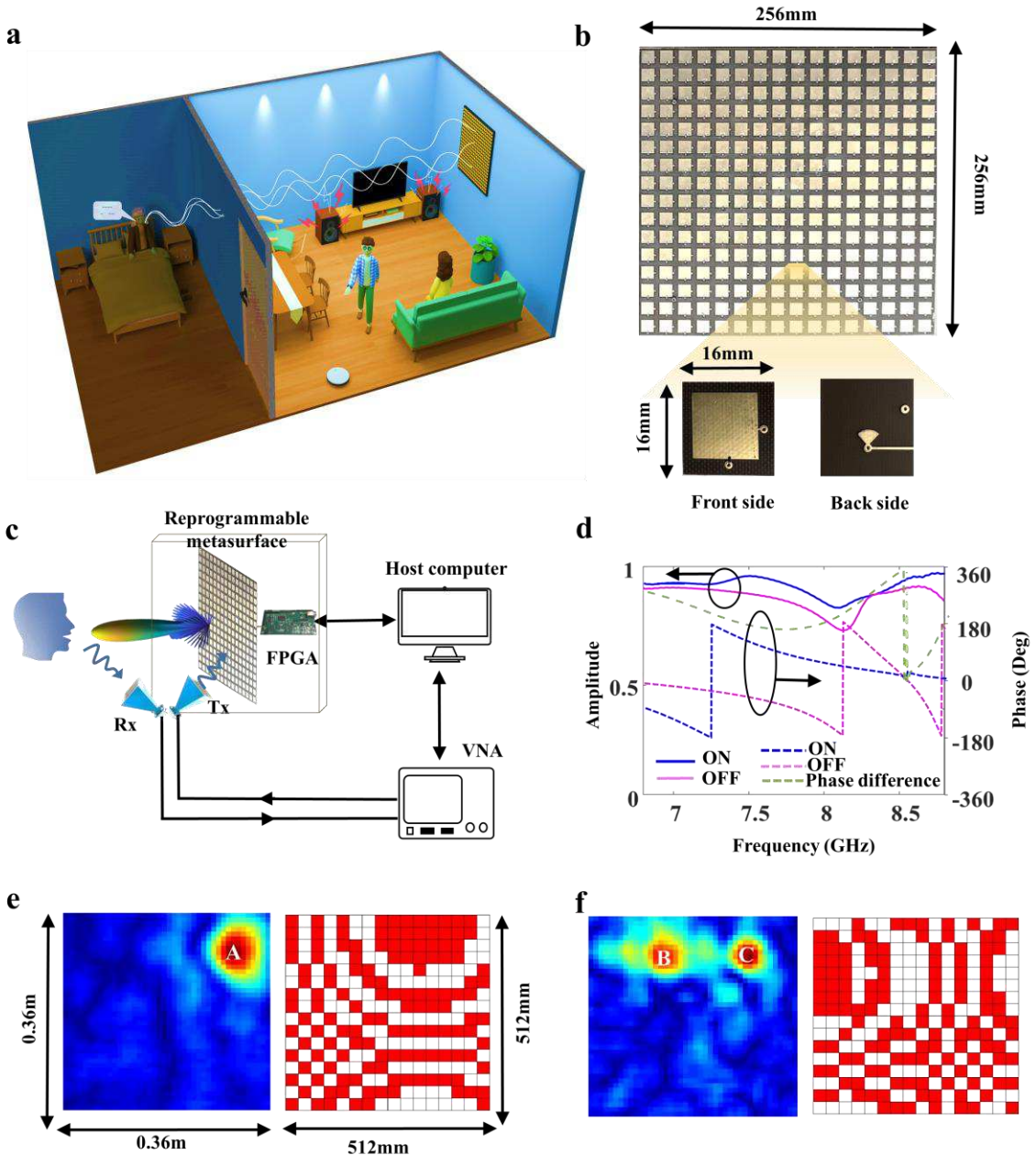

Figure 1. System design of the metasurface-empowered microwave speech recognizer. a) Conceptual illustration of the proposed microwave speech recognizer in a challenging indoor scenario: an elderly person in the sleeping room voice-commands an appliance (e.g., lights) through a wall and despite loud music and motion in the neighboring guest room. **b)** Photographic image of one out of four panels of our one-bit programmable coding metasurface. The insets show the front and back sides of the designed individual meta-atom. **c)** Schematic drawing of the hardware configuration of our proposed microwave speech recognizer. The hardware of our system consists of a large-aperture one-bit reprogrammable metasurface, a pair of horn antennas, a vector network analyzer (VNA) and a host computer. **d)** Experimental characterization of the frequency-dependent phase and amplitude response of our designed meta-atoms in their two possible configurations ("0"/"ON" and "1"/"OFF"). **e**) and **f**) Maps of the spatial distribution of the microwave field magnitude (measured via near-field scans) corresponding to the indicated metasurface configurations that are chosen to focus on point A (**e**) or points B and C (**f**).

Results

System configuration. We start by elaborating on the system configuration of our implemented prototype for our proposed metasurface-empowered microwave speech recognizer. On the hardware level, our system comprises a large-aperture one-bit reprogrammable metasurface (1024 meta-atoms, $51.2\ \mathrm{cm} \times 51.2\ \mathrm{cm}$ aperture), a pair of commercial horn antennas, a vector network analyzer (VNA, Agilent E5071C), and a host computer – see **Fig. 1c**. Our programmable-metasurface-empowered system must accomplish a series of complex tasks on the fly. First, our system must localize the speaker's mouth. This involves imaging the scene and interpreting the resulting image to identify the ROI, that is, the mouth. Second, our system must focus microwaves on the mouth and capture the reflected biosignals. Third, our system must interpret the measured biosignals to extract the sought-after speech content.

Our system's autonomous ROI identification and analysis follows Ref.[19]: during the first step, the scene is imaged with a series of 18 random illuminations, generated through a known series of 18 random metasurface configurations. The measured data is processed by artificial intelligence (AI) tools to generate a 3D skeleton of the speaker. Details of the utilized AI architecture are provided in the **Methods** and **Supplementary Note 4**. Based on the estimated skeleton, the coordinates of the mouth, our ROI, can be deduced. Using a modified Gerchberg-Saxton algorithm, our system then identifies a metasurface configuration that focuses microwaves on the speaker's mouth. A series of probing focused microwave signals is emitted with a period of 70 ms, and the reflected signals are captured by the receiving horn antenna. Note that the use of directive horn antennas, in contrast to omnidirectional antennas, further helps with discriminating ROI signals from multipath scattered signals in the room. The measured biosignals are then interpreted by

an artificial neural network (ANN) that directly maps the acquired microwave data to the desired speech content. Such a direct transcription of measured signals with text, without intermediate steps involving phonetic representations, is known as “end-to-end” speech recognition in the signal-processing literature[20]. Inspired by Ref.[21], we have developed a customized microwave-speech transformer for our system and trained it with supervised learning. To obtain labeled microwave-speech training data, we conveniently used the host computer’s built-in microphone that was synchronized with our proposed microwave speech recognizer. Algorithmic details about our microwave-speech transformer are provided in the **Methods** and **Supplementary Note 4**.

The first step of our system’s pipeline is a conventional instance of compressive imaging by leveraging the configurational diversity of a programmable meta-imager, first reported in Ref.[22] (see also Sec. II.B in Ref.[18] for a balanced review of the field). Nonetheless, our system’s pipeline in its entirety qualifies as an instance of “intelligent meta-imaging” according to the taxonomy from Ref.[18] (see Sec. III.A therein) because AI tools influence the choice of task-specific hardware configurations for our metasurface in the second step of its complex sequences of tasks. We note that related techniques for autonomous ROI identification were put forward in optical ghost imaging based on analytical rather than AI tools[23–25]. However, our AI-driven sensing pipeline is distinct from another class of intelligent meta-imagers which integrates the programmable meta-atoms as trainable physical weights directly into an end-to-end pipeline comprising both the physical and digital layers[26,27] (see Sec. III.B-C in Ref.[18]). The latter could be a future extension for the first step of our scheme.

The pivotal hardware ingredient of our microwave speech recognizer is an inexpensive one-bit reprogrammable coding metasurface composed of 32×32 electronically controllable meta-atoms. A photographic image of a 16×16 panel of meta-atoms is shown in **Fig. 1b**. Each meta-atom has dimensions of $16\ \mathrm{mm} \times 16\ \mathrm{mm} \times 1.88\ \mathrm{mm}$ and consists of a five-layer structure. A PIN diode (MADP-000907-14020x) is embedded on the top layer. By controlling the bias voltage of the PIN diode, we can electronically switch between two distinct meta-atom reflection properties in the microwave domain: The 0^{o}-phase (denoted as digit ‘0’) and 180^{o}-phase (denoted as ‘1’) states are achieved when the PIN diode is biased with an externally applied DC voltage of 5V (ON) or 0V (OFF), respectively. The frequency-dependent magnitude and phase responses of our designed meta-atom are measured experimentally and plotted in **Fig. 1d**. Our designed meta-atom

can efficiently manipulate the reflected electromagnetic field within the frequency range from 7.49 GHz to 8.3 GHz: the reflection phase of the meta-atom shows a roughly 180^{o} phase difference when the embedded PIN diode is switched from the ON state to the OFF state, while the reflection magnitudes are almost the same and close to unity. Thus, the radiation pattern can be flexibly manipulated by suitably biasing the PIN diodes of the metasurface through a field-programmable gate array (FPGA). More details about the reprogrammable coding metasurface are provided in **Methods** and **Supplementary Note 1**.

To examine the crucial role of our large-aperture programmable metasurface in discriminating between reflections from the mouth and undesired clutter as well as to improve the signal-to-noise ratio (SNR), we have conducted a series of preliminary experiments which seek to focus the microwave field at one or two desired location(s). The spatial distributions of the microwave field are obtained through near-field scans. Two representative results are shown in **Fig. 1e** and **Fig. 1f** for focusing on a single point A (0.12m, 0.14m, 1m) or simultaneously on two points B (-0.05m, 0.14m, 0.65m) and C (0.1m, 0.14m, 0.65m), respectively. The corresponding metasurface configurations are identified via a modified Gerchberg-Saxton algorithm and also displayed in **Fig. 1e** and **Fig. 1f**. These results demonstrate the ability of our system to reallocate the microwave energy to one or multiple desired spot(s) in a dynamic manner using a suitable configuration of our programmable metasurface. The signal level at the desired spots is enhanced by a factor of around 20 dB. This focusing on the ROI is crucial in order to drastically increase the weight of reflections from the ROI and to suppress the influence of undesired clutter, noise and multipath reflections.

Encoding of speech information in microwave biosignals. Having described our system configuration, we now wish to investigate the mechanisms through which speech information is encoded in the microwave biosignals that our system acquires. The acoustic sounds that constitute a voice originate from air being exhaled by the lungs and causing vocal cord vibrations. The sound is additionally shaped through a series of further articulatory entities such as tongue position and mouth opening. Clearly, the more the acquired signals of a remote sensing technique interact with all of the involved articulatory entities, the more speech information is expected to be encoded in the biosignals. Contactless visual speech recognition techniques solely rely on mouth motion because optical frequencies cannot penetrate through the skin and teeth. Similarly, Ref.[16] based

on a remote contactless microwave approach at 2.4 GHz suggests it relies solely on mouth motion. In contrast, Ref.[6] reported that interior articulatory entities like the tongue decisively impact their contact-based measurements in the microwave regime. In fact, Ref.[6] observed less microwave interaction with interior articulatory entities when the speaker had metallic tooth fillings, a clear indicator that the microwaves penetrated through the skin. These findings from Ref.[6] are in line with our observations. As displayed in **Fig. 2a**, for our working frequency around 8 GHz, we find in full-wave simulations (CST Microwave Studio) that a significant portion of the microwave signal penetrates through the skin and teeth and interacts notably with the subject's articulatory entities such as the tongue. In addition, the spatial resolution is decent because the wavelength at 8 GHz is 3.75 cm in free space and even less inside the biological tissue. These results suggest that our technique is remarkably different from existing non-microwave-based strategies in that it can efficiently encode speech information not only from the mouth but also from interior articulatory entities such as the tongue.

To test this hypothesis experimentally, we next conduct a series of speech experiments with our microwave speech recognizer system. Therein, the subject pronounces alternately two syllables, an alveolar |s| and a cacuminal |ʃ|, and repeats them with a period of two seconds for five minutes (see **Fig. 2b**). The participant is asked to keep the mouth as close as possible during the whole pronunciation process so that only tongue motion inside the mouth is involved. Thereby, we minimize the encoding of speech information in the microwave biosignals through the mouth. If our system is nonetheless capable of extracting speech information, this proves that it efficiently probes articulatory entities other than the mouth, too. The amplitudes of the acquired microwave biosignals as a function of the Doppler frequency are plotted in **Fig. 2c**; each curve is averaged over 100 repeated acquisitions, and the experiment was performed for five different distances between the speaker and the metasurface (d = 0.9 m, 1.2 m, 1.5 m, 1.8 m, and 2 m). The corresponding sizes of the focal spot on the mouth are estimated to be on the order of $O(\frac{\lambda d}{D})$, where $D = 51.2$ cm is the metasurface aperture, yielding values between 6.6 cm and 14.6 cm. As expected, the amplitude responses show the peaks at the Doppler frequency of around 0.5 Hz in all five cases. It is evident that the microwave signals can capture the hidden vocal vibrations and motions even though the motion of the subject's skin (lip and face) cannot be visually perceived. These microwave signals could be further processed to infer the speech content.

As a last set of preliminary experiments, the speaker is asked to pronounce a larger number of syllables and repeat them for five minutes with a frequency of 0.5 Hz per syllable. The corresponding microwave responses are collected by the developed microwave speech recognizer. **Figure 2d** shows amplitude and phase of the microwave responses corresponding to five different syllables. The corresponding sound signals obtained from the built-in microphone of the host computer are also plotted for comparison. It can be seen from **Fig. 2d** that the microwave responses are correlated with the corresponding sound signals in the time domain. The frequency spectrum of the microwave responses has distinct properties for each syllable**.**

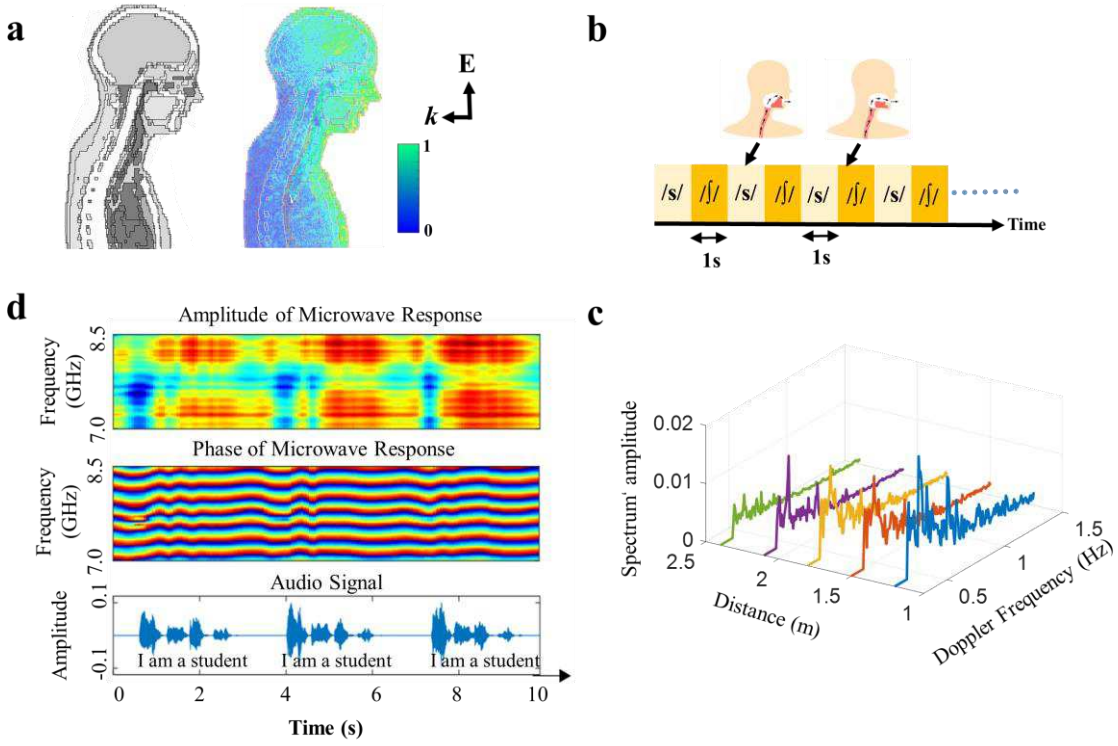

Figure 2. Physical mechanisms of speech encoding in the proposed microwave speech recognizer. a) Full-wave simulation of the interaction of the chosen microwave signal with the vocal organs, which is achieved by using a full-wave simulator, CST Microwave Studio 2012. In our simulation, the subject (Laura, a model of a 43-year-old female from Ref.[28]) is illuminated with a linearly polarized plane wave. **b)** The experimental scheme according to which the subject is asked to alternately pronounce an alveolar |s| and a cacuminal |ʃ|, with a period of two second for a duration of five minutes. **c)** The experimental results corresponding to the scheme presented in **b**: the amplitudes of the microwave biosignals are plotted as a function of the Doppler frequency. The results are shown for five separations between the speaker and the programmable metasurface (0.9 m, 1.2 m, 1.5 m, 1.8 m, and 2 m). **d)** Selected microwave responses when the subject is asked to pronounce a short sentence, i.e., 'I am a student', three times. In addition, the corresponding sound signal acquired by the in-built microphone of our host computer is plotted for comparison.

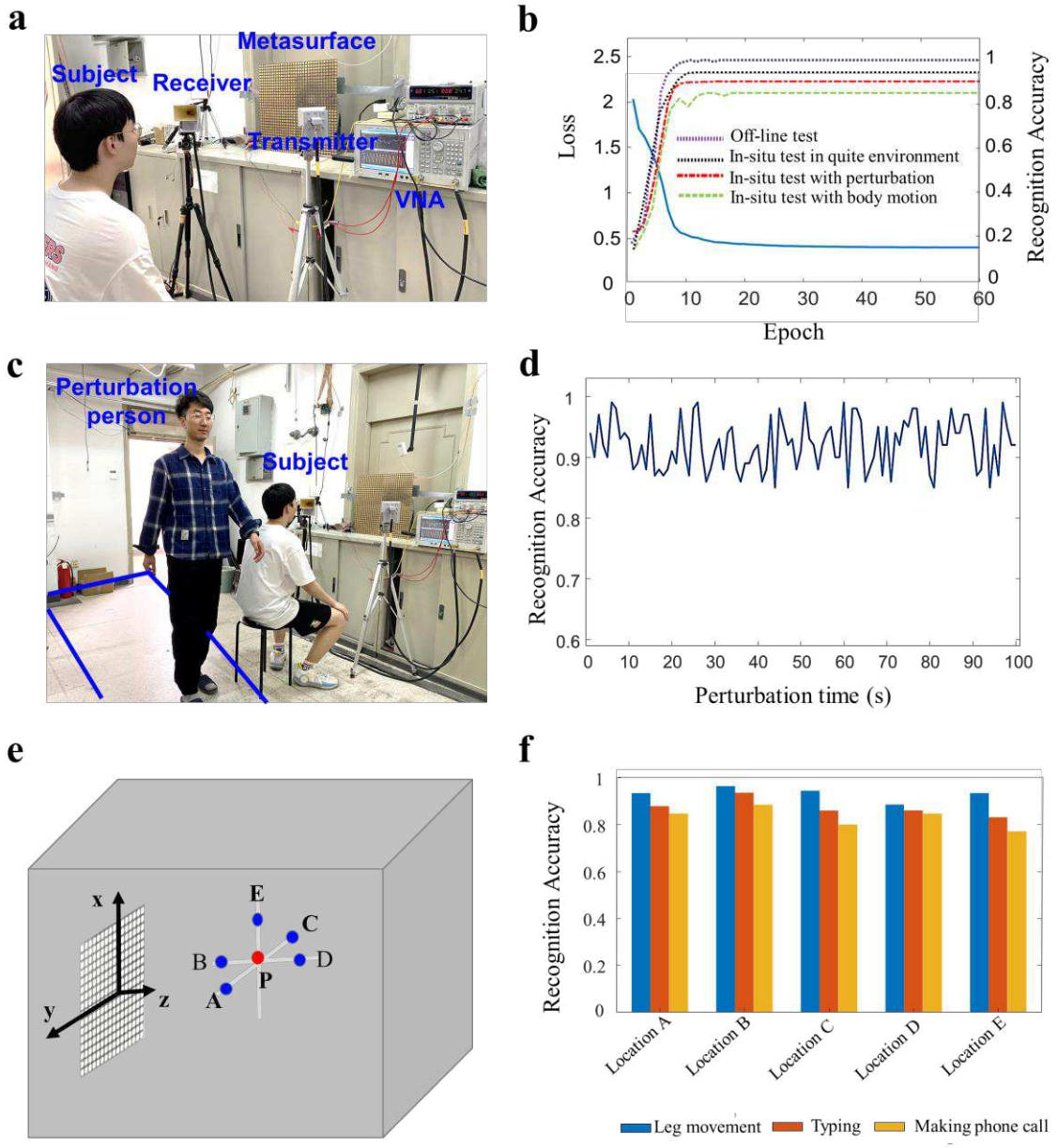

Figure 3. Experimental results of the microwave speech recognition in a line-of-sight setting. a) Experimental setting, where the subject is wearing or not wearing a face mask and sitting in front of the reprogrammable metasurface. **b)** The training and test behaviors of the microwave speech recognizer as a function of the training epoch. The logarithm of the loss function is plotted as blue solid line (left axis), and the test behavior is examined in terms of the recognition accuracy (right axis). Here, we consider four test cases with the microwave speech recognizer trained in the quiet environment: the simple test with the off-line collected test samples (called off-line test), the in-situ test with a subject in the same quiet environment as that for the training (called in-situ test in quiet environment), the in-situ test with a subject disturbed by an additional person freely acting in room (called in-situ test with perturbation), the in-situ test with a subject with different body motions (called in-situ test with body motion). **c)** Experimental setting for the investigation of the microwave metasurface speech recognizer's robustness to the disturbances of the ambient environment: an additional person acts freely within the region marked by the blue box while subject reads out loud the assigned material. **d)** Experimental speech recognition results for the setting from **c**. **e)** Experimental setting for examining the robustness when the speaking subject makes three

different kinds of body motions (making phone call, typing, rhythmical leg movement) at five different locations A (0.04 m, 0.1 m, 1.15 m), B (0.04 m, 0 m, 1.0 m), C (0.04 m, -0.1 m, 1.15 m), D (0.04 m, 0 m, 1.3 m), and E (0.08 m, 0 m, 1.15 m), while reading out loud. (**f**) Experimental speech recognition results for the setting from **e**.

Microwave speech recognition in line-of-sight setting. We begin to examine the performance of the developed microwave speech recognizer in a line-of-sight scenario, where the subject sits in front of the metasurface, as seen in **Fig. 3a**. To train our metasurface speech recognizer, we have recruited 22 participants (7 graduate students and 15 undergraduate students; 5 females and 17 males) and aim at recognizing 100 daily used English words. In the experiments, all participants were asked to read out loud the designated material five times at a normal speed (half a word per second on average) in a quiet environment. As mentioned above, the built-in microphone in our host computer is utilized for collecting the labeled voice data for supervised learning. In this way, we have acquired a total of 11000 pairs of labeled microwave-voice samples per location, in which 70% of the samples are randomly selected for training the microwave-voice transformer, and the rest are used for testing. Selected samples are provided in **Supplementary Note 2.**

First, we consider the simple case in which a single subject with or without wearing a face mask sits at P (0.04 m, 0 m, 1.15 m) in front of the reprogrammable metasurface in a static and quiet environment, as shown in **Fig. 3a**. **Figure 3b** demonstrates the learning and test performances of the developed microwave speech recognizer in terms of how the loss function evolves over the course of the training epochs. In addition, the dependence of the speech recognition accuracy over the training samples as the epoch index increases is plotted in **Fig. 3b**. We mix cases in which the subject wears a face mask or not because the results are almost identical in both cases. **Figure 3b** shows that the developed speech recognizer can be effectively trained to achieve near-perfect speech recognition, and that the trained system works very well on the ‘unseen’ test samples, even when the speaker wears a mask. In other words, these results indicate that the proposed microwave speech recognizer can ‘hear’ what people say without audio and visual clues and ‘see’ the speaker’s mouth in a remote sensing manner.

Next, we evaluate the robustness of our speech recognition procedure in a dynamically changing environment. Body motion or a changing environment can lead to parasitic variations of the acquired microwave signals that deteriorate the speech recognition

performance. We conducted a set of experiments in which a second person, referred to as the perturbation person, acts freely within the indicated region in **Fig. 3c** while the subject reads out loud the designated material. To interpret the microwave signals we use the ANN previously trained in a quiet environment without any disturbances. The results in this dynamically changing environment are shown in **Fig. 3d** and demonstrate the robustness of our approach. This robustness can be attributed to the efficiency with which our programmable metasurface focuses the microwave beam on the subject's mouth such that reflections off the perturbation person are very weak and hence their disturbing impact is efficiently suppressed.

We also examined the robustness with respect to body motion of the speaking subject. Here, we consider a realistic scenario in which the subject walks around freely while speaking. This implies that the distance and orientation of the subject's mouth relative to the programmable metasurface are varying. Specifically, the subject spoke at the five different locations around the point P, indicated in **Fig. 3e** while performing three different kinds of body motion (making a phone call, typing, rhythmical leg movements). The achieved recognition accuracies, again based on the microwave-voice transformer network trained in the quiet environment, are reported in **Fig. 3f**. These results indicate that the speech recognition performance is almost unchanged when the subject makes phone calls or types while reading the designated text. Again, this robustness can be attributed to our efficient programmable-metasurface system that tracks the subject's mouth and ensures that the microwave focal spot follows the subject's motion. These characteristics are very encouraging with regard to real-life convenient and robust speech sensing, irrespective of the speaker's distance and orientation.

Microwave speech recognition in through-a-wall setting. Furthermore, we tackle microwave multi-speaker speech recognition in a more challenging through-a-wall scenario where two subjects sit behind a 5 cm-thick wooden wall and talk with each other. The wall has a higher dielectric constant than air and possibly an additional microstructure, which previously motivated the conception of special wall-compensation algorithms to mitigate artefacts due to reflections and wavefront distortions in through-a-wall computational meta-imaging[29]. The multi-speaker problem now requires that the metasurface is configured such that it simultaneously focuses microwave energy on both speakers' mouths – like the example from **Fig. 1f**. To train the microwave metasurface speech recognizer, 22 participants read the assigned English reading material five times

behind the 5 cm-thick wooden wall. As before, 70% of the samples are randomly selected for training the microwave-voice transformer, and the rest are used for testing. The corresponding experimental results are reported in **Supplementary Note 3**. The recognition accuracies of above 80% from **Supplementary Fig. 3** confirm that our system performs well also for the very challenging multi-speaker through-a-wall speech recognition task, even if a third person acts freely in the room while the two subjects talk to each other.

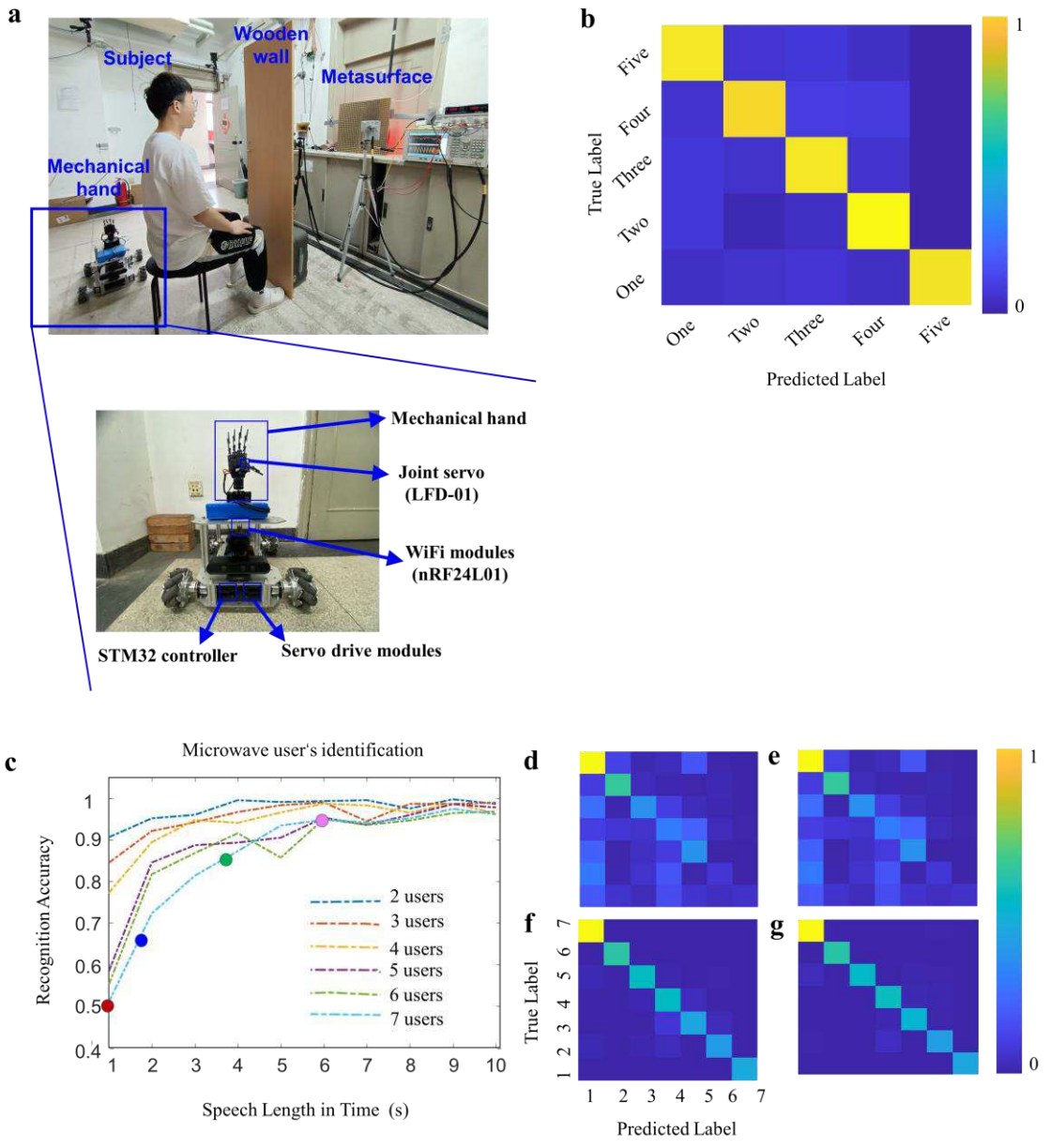

Figure 4. **Voice-commanded through-a-wall human-machine interaction based on our metasurface-empowered microwave speech recognizer. a)** Experimental setting. Details about the mechanical hand are provided in the inserted figure. Further details can be found in **Supplementary Video 1**. **b)** Classification confusion matrix for the five different speech commands: 'one', 'two', 'three', 'four', and 'five'. **c)** Results of the dependence of the recognition accuracy on the speech length for the different number of speakers. **d)** Classification confusion matrices corresponding to the four red points marked in **c.**

Voice-commanded human-machine interaction. Besides recognizing the content of voice commands which is the problem we have studied so far, ideally, a voice-commanded human-machine interface should additionally be able to recognize the speaker's identity. Voice is well-established as acoustic "fingerprint" of a user's identity because unique vocal features are encoded by unique properties of the individual's lung, vocal cords, vocal tract, and other articulatory entities. Therefore, even though multiple people pronounce the same words, the uttered sounds include distinctive features for each person. We now explore whether user identification is also possible with our acquired microwave biosignals. Interestingly, this variation across different users can be interpreted as a form of noise with respect to speech recognition whereas it is the salient feature for user identification. This double-sided-sword nature of signal variations as being either noise or the crucial feature is reminiscent of microwave-based complex localization problems[30].

We consider the through-a-wall microwave speech data of seven subjects, labelled with classes from 1 to 7 according to which subject they correspond to. The data processing problem is now a multivariate classification problem for which we have developed a deep convolutional neural network (see **Supplementary Note 4** for more details). We explore two factors that we expect to have a major influence on the identification performance: the speech sample length and the number of subjects to be distinguished. Indeed, the results plotted in **Fig. 4c-g** demonstrate that as more subjects must be distinguished, acceptable classification accuracy can be achieved by analyzing longer speech samples. Using only six-second-long speech samples, we can distinguish between all seven individuals based on the microwave biosignals, without any vocal and visual clues.

Finally, we now discuss our demonstration of a voice-commanded human-machine interface in which a mechanical hand is controlled based on through-a-wall vocal speech that is recognized by our microwave speech recognizer. The corresponding experimental setup is depicted in **Fig. 4a**. The system now recognizes the speech content and subsequently sends out the recognized speech commands to a mechanical hand in order

to control the motion of the latter in real time. The mechanical hand is integrated into a mobile vehicle and each finger is fitted with an anti-blocking joint servo (LFD-01) for the finger retraction control. An on-board Wi-Fi module (nRF24L01) is used to wirelessly receive control commands from the host computer based on the recognized vocal commands. The vehicle is equipped with an STM32 controller to process the received control commands into the control quantities for the corresponding finger servos (see **Supplementary Note 5** for details). Five different speech commands are involved in this experiment: 'one', 'two', 'three', 'four', and 'five'. We have evenly collected 1000 pairs of microwave and acoustic samples for the five commands, and utilized 70% and 30% of samples to train and test the microwave speech recognizer, respectively. We report the classification confusion matrix in **Fig. 4b**, showing that near-perfect speech recognition of the five commands can be achieved by using our microwave speech recognizer. More details have been recorded in **Supplementary Video 1**. These primary experimental results demonstrate the important potential of our microwave speech recognizer for microwave-based contactless voice-commanded human-machine interfaces.

Conclusions

To summarize, we have proposed and experimentally prototyped the concept of a microwave speech recognizer empowered by a programmable metasurface, including a demonstration of a voice-commanded human-machine interface. Our system answers two fundamental questions in speech recognition – "what is being said?" and "who is speaking?" – based on voice-modulated microwave biosignals. The unique advantages of a large-aperture programmable metasurface enable us to implement microwave-based speech recognition with unprecedented accuracy because we can dynamically track the speaker's mouth and focus microwaves on it with high efficiency. Our work is particularly timely in the current pandemic context: people always wear masks in public places such that their lip movements cannot be seen, and neither can their voice be heard given loud ambient noise sources. The demonstrated ability to implement contactless voice-commanded human-machine interfaces without reliance on optical or acoustic cues will enable numerous important but to-date inaccessible applications of human-machine interfaces such as in smart health care or industrial settings, as well as intelligent surveillance and security.

Methods

Design of the one-bit reprogrammable coding metasurface. The reprogrammable coding metasurface is an ultrathin planar array of meta-atoms that are individually reconfigurable via electronic commands[17,31,32]. Thanks to its unique capabilities to manipulate electromagnetic wavefields in a reprogrammable manner, it has elicited many exciting physical phenomena (e.g., nonreciprocal reflection effects[33]) and versatile functional devices, including computational imagers[18,22,26,19,27,34–40], dynamic holography[41], wireless communications[42–49], analog wave-based computing[50–52], and dynamic cloaks[53].

We have designed a one-bit reprogrammable metasurface that is composed of 32×32 meta-atoms. The meta-atom is a five-layer structure as shown in **Fig. 1b** and **Supplementary Note 1**. The top layer is a square copper patch with dimensions of $11\ \mathrm{mm} \times 11\ \mathrm{mm}$, which contains a PIN diode to control the reflection phase of the meta-atom. The second layer has a thickness of 1.58 mm and is made of Taconic TLX-8 which has a relative permittivity of 2.55. The fourth layer has a thickness of 0.3 mm and is made of FR-4 with a dielectric constant of 4.3. The third and fifth layers are ground planes made of copper, and a via hole is introduced on the third layer to isolate the bias voltage coming from the fifth layer. For the sake of easy fabrication, the entire reprogrammable coding metasurface is designed to be composed of 2×2 metasurface panels, and each panel consists of 16×16 electronically controllable digital meta-atoms. One such panel is depicted in **Fig. 1b**. Each metasurface panel is equipped with eight 8-bit shift registers (SN74LV595APW), and eight PIN diodes are sequentially controlled. The adopted clock rate is 50 MHz, and the ideal switching time of the PIN diodes is 10 μs.

Algorithmic overview. The first step of our microwave speech recognizer is to localize the ROI in the scene, i.e., the speaker's mouth. To this end, the scene is illuminated with 18 random patterns generated by a known series of random metasurface configurations. The acquired data is directly mapped to a 3D skeleton of the speaker using a deep ANN (see details in **Supplementary Note 4**).[54] Based on the 3D skeleton, it is then straightforward to localize the speaker's mouth. These ROI coordinates are needed to identify a metasurface configuration that efficiently focuses microwaves on the speaker's mouth. A suitable metasurface configuration for this focusing task is identified with a modified Gerchberg-Saxton algorithm based on the ROI coordinates. This is the basis for

capturing clutter-resiliant microwave biosignals from which speech information can be extracted. The autonomous ROI identification here differs from that in Ref.[19] in that the acquired data from the first step is mapped to a 3D skeleton as opposed to a full image.

Microwave-speech transformer. The microwave-speech transformer is a deep artificial neural network which directly converts the sequence of microwave signals to the sequence of recognized speech information in an end-to-end fashion. The architecture is inspired by Ref.[21] and detailed in **Supplementary Note 4**. The network adopts the typical Transformer structure and uses an encoder-decoder module structure, which is mainly composed of multi-head attention layer, feed-forward layer, residual connection and layer normalization. The network training is performed using the Adam optimization method[55] with a mini-batch size of 64, an epoch setting of 50, and a learning rate of 3×10^{-4}. The complex-valued weights are initialized randomly with a zero-mean Gaussian distribution of standard deviation 10^{-3}. The training is performed on a workstation with an Intel Xeon E5-1620v2 central processing unit, NVIDIA GeForce GTX 2080Ti, and 128 GB access memory. The machine learning platform TensorFlow is used to define and train the networks.

Speaker identity recognition. The network for recognizing the speaker's identification from the microwave biosignals is based on a simple CNN structure as detailed in **Supplementary Note 4**. It consists of convolutional layers, pooling layers, fully connected layers, and Softmax activations. The network maps the acquired biosignals directly to the user identity class.

Configuration of proof-of-concept system. The experimental setup, as shown in **Fig. 1b**, consists of a transmitting (TX) horn antenna, a receiving (RX) horn antenna, a large-aperture reprogrammable metasurface, and a vector network analyzer. The two horn antennas are connected to two ports of the VNA via two 4m-long 50-Ω coaxial cables, and the VNA is used to acquire the response data by measuring transmission coefficients (S_{21}). In addition, an in-build sound microphone in the host computer has been integrated into our system for acquiring labeled training data. The computer controls the VNA and microphone to acquire the microwave data and voice signal, respectively, using the Python 3.1 software. These two procedures of data acquisition share the same starting time and ending time, but with different sampling intervals, $70\ \mathrm{ms}$ for the microwave data from the VNA and $\frac{1}{22050}\ s$ for the acoustic data from the microphone. Note that the sound

signals are solely used to assist in labeling the microwave data with corresponding text to obtain labeled training data, since our primary goal is to infer the speech information from the microwave data. To that end, we cut the acoustic signal corresponding to a specific text by listening and writing down the start and end times. Since the microwave speech data and the acoustic data are acquired at the same start time, we can easily align each sampling interval. Finally, we can label the microwave speech data with the corresponding text. We input these microwave biosignals into the neural network and train the latter so that it outputs the corresponding speech.

Acknowledgments

This work was supported in part through the National Key Research and Development Program of China (2017YFA0700201, 2017YFA0700202, and 2017YFA0700203), and in part through the National Natural Science Foundation of China (61471006, 61631007 and 61571117).

Author Contributions

L.L. conceived the idea and conducted the theoretical analysis. P.d.H. contributed to the conceptualization of the project. H.R., S.J., Ho.Z., Ha.Z., Z.W., S.H. and L.L. conducted the experiments. All authors contributed to data analysis and interpretation. L.L. and P.d.H. wrote the manuscript, and all authors read the manuscript.

Additional Information

Supplementary Information accompanies this article.

Competing Interests: The authors declare no competing interests.

References

1. Denby, B. *et al.* Silent speech interfaces. *Speech Commun.* **52**, 270–287 (2010).
2. Fagan, M. J., Ell, S. R., Gilbert, J. M., Sarrazin, E. & Chapman, P. M. Development of a (silent) speech recognition system for patients following laryngectomy. *Med. Eng. Phys.* **30**, 419–425 (2008).
3. Meltzner, G. S. *et al.* Silent Speech Recognition as an Alternative Communication Device for Persons With Laryngectomy. *IEEE/ACM Trans. Audio Speech Lang. Process.* **25**, 2386–2398 (2017).
4. Zhang, R. *et al.* SpeeChin: A Smart Necklace for Silent Speech Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **5**, 1–23 (2021).
5. Kimura, N. *et al.* SilentSpeller: Towards mobile, hands-free, silent speech text entry using electropalatography. in *Proc. CHI* 1–19 (ACM, 2022).
6. Birkholz, P., Stone, S., Wolf, K. & Plettemeier, D. Non-Invasive Silent Phoneme Recognition Using Microwave Signals. *IEEE/ACM Trans. Audio Speech Lang. Process.* **26**, 2404–2411 (2018).
7. Wagner, C. *et al.* Silent speech command word recognition using stepped frequency continuous wave radar. *Sci. Rep.* **12**, 4192 (2022).
8. Gao, Y., Jin, Y., Li, J., Choi, S. & Jin, Z. EchoWhisper: Exploring an Acoustic-based Silent Speech Interface for Smartphone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **4**, 1–27 (2020).
9. Zhang, Y., Chen, Y.-C., Wang, H. & Jin, X. CELIP: Ultrasonic-based Lip Reading with Channel Estimation Approach for Virtual Reality Systems. in *Proc. UbiComp* 580–585 (ACM, 2021).

10. Eid, A. M. & Wallace, J. W. Ultrawideband Speech Sensing. *IEEE Antennas Wirel. Propag. Lett.* **8**, 1414–1417 (2009).

11. Shin, Y. & Seo, J. Towards Contactless Silent Speech Recognition Based on Detection of Active and Visible Articulators Using IR-UWB Radar. *Sensors* **16**, 1812 (2016).

12. Movellan, J. R. Visual Speech Recognition with Stochastic Networks. *Proc. NIPS* **7**, 8 (1994).

13. Guoying Zhao, Barnard, M. & Pietikainen, M. Lipreading With Local Spatiotemporal Descriptors. *IEEE Trans. Multimedia* **11**, 1254–1265 (2009).

14. Dupont, S. & Luettin, J. Audio-visual speech modeling for continuous speech recognition. *IEEE Trans. Multimedia* **2**, 141–151 (2000).

15. Gonzalez-Valdes, B. *et al.* Improving Security Screening: A Comparison of Multistatic Radar Configurations for Human Body Imaging. *IEEE Antennas Propag. Mag.* **58**, 35–47 (2016).

16. Wang, G., Zou, Y., Zhou, Z., Wu, K. & Ni, L. M. We Can Hear You with Wi-Fi! *IEEE Trans. Mob. Comput.* **15**, 2907–2920 (2016).

17. Cui, T. J., Qi, M. Q., Wan, X., Zhao, J. & Cheng, Q. Coding metamaterials, digital metamaterials and programmable metamaterials. *Light Sci. Appl.* **3**, e218–e218 (2014).

18. Saigre-Tardif, C., Faqiri, R., Zhao, H., Li, L. & del Hougne, P. Intelligent meta-imagers: From compressed to learned sensing. *Appl. Phys. Rev.* **9**, 011314 (2022).

19. Li, L. *et al.* Intelligent metasurface imager and recognizer. *Light Sci. Appl.* **8**, 97 (2019).

20. Graves, A. & Jaitly, N. Towards End-to-End Speech Recognition with Recurrent Neural Networks. *Proc. ICML* 1764–1772 (2014).

21. Vaswani, A. *et al.* Attention is All you Need. *Proc. NIPS* 11 (2017).

22. Sleasman, T., F. Imani, M., Gollub, J. N. & Smith, D. R. Dynamic metamaterial aperture for microwave imaging. *Appl. Phys. Lett.* **107**, 204104 (2015).

23. Averbuch, A., Dekel, S. & Deutsch, S. Adaptive Compressed Image Sensing Using Dictionaries. *SIAM J. Imaging Sci.* **5**, 57–89 (2012).

24. Aβmann, M. & Bayer, M. Compressive adaptive computational ghost imaging. *Sci. Rep.* **3**, 1545 (2013).

25. Phillips, D. B. *et al.* Adaptive foveated single-pixel imaging with dynamic supersampling. *Sci. Adv.* **3**, e1601782 (2017).

26. del Hougne, P., Imani, M. F., Diebold, A. V., Horstmeyer, R. & Smith, D. R. Learned Integrated Sensing Pipeline: Reconfigurable Metasurface Transceivers as Trainable Physical Layer in an Artificial Neural Network. *Adv. Sci.* **7**, 1901913 (2019).

27. Li, H.-Y. *et al.* Intelligent Electromagnetic Sensing with Learnable Data Acquisition and Processing. *Patterns* **1**, 100006 (2020).

28. *CST Bio Models Library Extension 3.2*.

29. Sleasman, T., Imani, M. F., Boyarsky, M., Trofatter, K. P. & Smith, D. R. Computational through-wall imaging using a dynamic metasurface antenna. *OSA Continuum* **2**, 3499 (2019).

30. del Hougne, P. Robust position sensing with wave fingerprints in dynamic complex propagation environments. *Phys. Rev. Research* **2**, 043224 (2020).

31. Yang, H. *et al.* A programmable metasurface with dynamic polarization, scattering and focusing control. *Sci. Rep.* **6**, 35692 (2016).

32. Huang, C. *et al.* Dynamical beam manipulation based on 2-bit digitally-controlled coding metasurface. *Sci. Rep.* **7**, 42302 (2017).

33. Zhang, L. *et al.* Breaking Reciprocity with Space-Time-Coding Digital Metasurfaces. *Adv. Mater.* **31**, 1904069 (2019).

34. Li, Y. B. *et al.* Transmission-Type 2-Bit Programmable Metasurface for Single-Sensor and Single-Frequency Microwave Imaging. *Sci. Rep.* **6**, 23731 (2016).

35. Sleasman, T., Imani, M. F., Gollub, J. N. & Smith, D. R. Microwave Imaging Using a Disordered Cavity with a Dynamically Tunable Impedance Surface. *Phys. Rev. Applied* **6**, 054019 (2016).

36. Sleasman, T. *et al.* Single-frequency microwave imaging with dynamic metasurface apertures. *J. Opt. Soc. Am. B* **34**, 1713 (2017).

37. Boyarsky, M. *et al.* Single-frequency 3D synthetic aperture imaging with dynamic metasurface antennas. *Appl. Opt.* **57**, 4123–4134 (2018).

38. Li, L. *et al.* Machine-learning reprogrammable metasurface imager. *Nat. Commun.* **10**, 1082 (2019).

39. Ruan, H. & Li, L. Imaging Resolution Analysis of Single-Frequency and Single-Sensor Programmable Microwave Imager. *IEEE Trans. Antennas Propag.* **68**, 7727–7732 (2020).

40. Ruan, H., Wei, M., Zhao, H., Li, H. & Li, L. Programmable Metasurface Based Microwave Gesture Detection and Recognition Using Deep Learning. in *Proc. NEMO* 1–4 (IEEE, 2020).

41. Li, L. *et al.* Electromagnetic reprogrammable coding-metasurface holograms. *Nat. Commun.* **8**, 197 (2017).

42. Yoo, I., Imani, M. F., Sleasman, T., Pfister, H. D. & Smith, D. R. Enhancing Capacity of Spatial Multiplexing Systems Using Reconfigurable Cavity-Backed Metasurface Antennas in Clustered MIMO Channels. *IEEE Trans. Commun.* **67**, 1070–1084 (2019).

43. del Hougne, P., Fink, M. & Lerosey, G. Optimally diverse communication channels in disordered environments with tuned randomness. *Nat. Electron.* **2**, 36–41 (2019).

44. Renzo, M. D. *et al.* Smart radio environments empowered by reconfigurable AI meta-surfaces: an idea whose time has come. *J. Wireless Com. Network* **2019**, 129 (2019).

45. Dai, L. *et al.* Reconfigurable Intelligent Surface-Based Wireless Communications: Antenna Design, Prototyping, and Experimental Results. *IEEE Access* **8**, 45913–45923 (2020).

46. Zhao, H. *et al.* Metasurface-assisted massive backscatter wireless communication with commodity Wi-Fi signals. *Nat. Commun.* **11**, 3926 (2020).

47. Shlezinger, N., Alexandropoulos, G. C., Imani, M. F., Eldar, Y. C. & Smith, D. R. Dynamic Metasurface Antennas for 6G Extreme Massive MIMO Communications. *IEEE Wireless Commun.* **28**, 106–113 (2021).

48. Zhang, L. *et al.* A wireless communication scheme based on space- and frequency-division multiplexing using digital metasurfaces. *Nat. Electron.* **4**, 218–227 (2021).

49. Imani, M. F., Abadal, S. & del Hougne, P. Metasurface-Programmable Wireless Network-on-Chip. *Adv. Sci. (in press), arXiv:2109.03284* (2022).

50. del Hougne, P. & Lerosey, G. Leveraging Chaos for Wave-Based Analog Computation: Demonstration with Indoor Wireless Communication Signals. *Phys. Rev. X* **8**, 041037 (2018).

51. Sol, J., Smith, D. R. & del Hougne, P. Meta-programmable analog differentiator. *Nat. Commun.* **13**, 1713 (2022).

52. Liu, C. *et al.* A programmable diffractive deep neural network based on a digital-coding metasurface array. *Nat. Electron.* **5**, 113–122 (2022).

53. Qian, C. *et al.* Deep-learning-enabled self-adaptive microwave cloak without human intervention. *Nat. Photonics* **14**, 383–390 (2020).

54. Wang, Z. *et al.* High Resolution 3D Microwave Real Time Imaging System Based on Intelligent Metasurface. in *Proc. IMWS-AMP* 379–381 (IEEE, 2021).

55. Kingma, D. P. & Ba, J. L. Adam: A method for stochastic optimization. *arXiv:1412.6980* (2014).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SM.pdf
- Video1.mp4