Materials Horizons

ROYAL SOCIETY OF CHEMISTRY

COMMUNICATION

View Article Online
View Journal | View Issue

Cite this: *Mater. Horiz.*, 2023, 10, 5607

Received 10th July 2023,
Accepted 20th September 2023

DOI: 10.1039/d3mh01062g

rsc.li/materials-horizons

Decoding silent speech commands from articulatory movements through soft magnetic skin and machine learning†

Penghao Dong, [a] Yizong Li,[a] Si Chen,[a] Justin T. Grafstein,[a] Irfaan Khan[b] and Shanshan Yao *[a]

Silent speech interfaces have been pursued to restore spoken communication for individuals with voice disorders and to facilitate intuitive communications when acoustic-based speech communication is unreliable, inappropriate, or undesired. However, the current methodology for silent speech faces several challenges, including bulkiness, obtrusiveness, low accuracy, limited portability, and susceptibility to interferences. In this work, we present a wireless, unobtrusive, and robust silent speech interface for tracking and decoding speech-relevant movements of the temporomandibular joint. Our solution employs a single soft magnetic skin placed behind the ear for wireless and socially acceptable silent speech recognition. The developed system alleviates several concerns associated with existing interfaces based on face-worn sensors, including a large number of sensors, highly visible interfaces on the face, and obtrusive interconnections between sensors and data acquisition components. With machine learning-based signal processing techniques, good speech recognition accuracy is achieved (93.2% accuracy for phonemes, and 87.3% for a list of words from the same viseme groups). Moreover, the reported silent speech interface demonstrates robustness against noises from both ambient environments and users' daily motions. Finally, its potential in assistive technology and human–machine interactions is illustrated through two demonstrations – silent speech enabled smartphone assistants and silent speech enabled drone control.

New concepts

In this study, we introduce a ground-breaking concept in the field of silent speech interfaces. Central to our concept is the utilization of a single soft magnetic skin discreetly positioned in the ramus-temporal junction area, which enables socially acceptable silent speech recognition through precise decoding of articulatory movements. The fabricated magnetic skin exhibits conformability to the human skin while providing a robust magnetic signal strength. Consequently, it achieves great sensitivity to even subtle deformations of the skin. Compared to current methodologies, our approach effectively overcomes concerns associated with face-worn sensor interfaces, minimizes sensor quantity, reduces facial visibility, and eliminates obtrusive interconnections between sensors and data acquisition components. By employing machine learning-based signal processing techniques, we achieve remarkable speech recognition accuracy, with 93.2% accuracy for phonemes and 87.3% accuracy for a list of words from the same viseme groups. Notably, our proposed silent speech interface demonstrates exceptional robustness against ambient noises and users' daily motions. Furthermore, we showcase the potential applications of this novel concept in assistive technology and human–machine interactions through two practical demonstrations: silent speech enabled smartphone assistants and silent speech enabled drone control.

1. Introduction

Spoken communication, being one of the most intuitive means of communication, plays a vital role in conveying information among humans and human–machine interactions (HMI). However, it is susceptible to physiological constraints and environmental interferences.[1] Physiologically, speech generation involves multiple organs such as the lungs, larynx, tongues, lips, teeth, jaws, and ears, which are responsible for phonation, articulations, resonance, and auditory perceptions, respectively.[2,3] Any disruption to these organs can impact speech or hearing abilities and potentially lead to voice disorders or hearing impairments,[4] thus diminishing communication efficiency in both human–human and human–machine scenarios. On the other hand, environmental factors such as noisy surroundings (*e.g.*, acoustically harsh workplaces, crowded gatherings, or background noise from televisions), situations requiring quiet or privacy (*e.g.*, hospitals, public areas, or private communications), and environments lacking an acoustic medium (*e.g.*, underwater or in the space) often impose limitations on voice-based speech communication.[5,6]

[a] *Department of Mechanical Engineering, Stony Brook University, Stony Brook, New York 11794, USA. E-mail: shanshan.yao@stonybrook.edu*
[b] *Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, New York 11794, USA*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3mh01062g

The silent speech interface, which eliminates the need for acoustic speech sounds, emerges as an alternative method to overcome physiological and environmental challenges for vocalized speech. This technique enables speech communication by detecting and interpreting subvocalized articulatory movements.[7,8] Methods for silent speech recognition can be broadly classified into two categories: contactless approaches and contact-based approaches. Contactless approaches are mainly explored through camera-based visual signals,[9–15] ultrasound signals.[16–23] Camera-based visual solutions require external video tracking devices, and users must remain within the camera's line of sight. Despite efforts to develop compact shoulder-mounted devices[9] to enhance portability, visual solutions still face challenges in terms of lighting conditions and angles between users and cameras, thereby limiting their practicality. For ultrasound-based solutions, ultrasonic imaging devices were employed to construct 2D lip and tongue images.[17,24,25] While these systems overcome the issue of visible light intensity, they encounter a similar alignment issue between the device and targeted articulators. As a more portable and user-friendly ultrasound-based solution, the speaker of the cell phone was used to emit ultrasound signals, and the microphone was employed to capture reflected signals from the lips.[16,18–22] This method is not hands-free and is susceptible to multipath interferences caused by bodily movements and surrounding objects.

Contact-based approaches involve attaching sensors to subjects' tongue, facial or neck skin, speech motor cortex, or inside the ear canal to detect signals induced by articulator movements (Table S1, ESI†). These approaches include systems that utilize physiological signals (*e.g.*, Electromyography (EMG), electroencephalography (EEG), and electropalatography (EPG)) and articulatory movement-induced signals (*e.g.*, strain, pressure, acceleration, angular velocity, and magnetic signals). EEG methods[26–28] can interpret speech information but are susceptible to interferences, especially when subjects experience cognitive distractions or mental deviations. Silent speech interfaces based on inertial measurement units (IMU) attached to the temporomandibular joint[3,29] or the chin and neck skin,[30] proximity sensors attached to the ear canal,[18,31] and rigid magnets to the tongue and facial skin[32–34] can achieve high portability and accuracy. These systems implement rigid components and highly visible interfaces on the skin.

Soft electronics have greatly contributed to contact-based approaches due to their conformable contact with the tissue surface, which allows for high accuracy and sensitivity. EPG utilizes a high-density electrode array placed on the hard palate.[35] Though effective, it is an invasive method and requires wiring to connect EPG electrodes in the mouth to an external circuit. EMG-based systems[5,36–41] are very promising, and researchers have developed soft conformal dry EMG electrodes[42,43] to improve their signal quality and long-term wearability. However, EMG-based interfaces often require multiple electrodes placed on the face or the neck, increasing system complexity and reducing user acceptance. An ionic hydrogel-based pressure sensor was developed to track throat pressure and translate signals to speech using Morse code.[44] Another approach involves translating sign language to speech by measuring finger strain.[45] This system has good wearing comfort and offers high accuracy. However, they are not based on natural speech. Recently, another approach is attaching soft resistive or triboelectricity-based strain sensors to the facial skin[46–52] for measuring skin strains induced by lip and jaw movements. More efforts are needed to improve speech recognition accuracy of this approach and alleviate the obtrusiveness of sensors and interconnects placed on the facial skin. Overall, the soft skin-worn sensors have superior wearing comfort and/or sensitivity than conventional rigid electronics, these silent speech systems face several critical challenges, including a large number of sensors, obtrusive and socially inappropriate interfaces on the skin, low accuracy, poor robustness to interferences, inability to handle natural language. In addition to Fig. 1(a)–(l) for describing our work, the comparison is summarized in Fig. 1(m).

In this work, we present an unobtrusive, wireless, and robust silent speech interface that addresses the above challenges through innovations in materials, structural design, sensing location, and signal processing algorithms. Our system tracks speech-relevant magnetic signals induced by the movement of the temporomandibular joint and decodes these signals into speech. Efforts were made to overcome the limitations of traditional magnetic signals-based speech recognition interfaces: (1) the silent speech interface utilizes only one piece of soft magnetic skin placed behind the ear and no cumbersome wires or cables between sensors and data acquisition components, allowing for a wireless, unobtrusive, user-friendly system for daily use. (2) With the optimized polymer matrix, magnetic particle loading ratio, and magnetization direction, the magnetic skins possess skin-like softness and can precisely track subtle skin movements in all three axes without affecting natural skin movements. (3) Displacement and strain changes in the temporomandibular joint area were measured using the digital correlate image (DIC) technique to facilitate the selection of optimal sensing locations. (4) The signal processing was facilitated by machine learning (ML) methods, which enable the recognition of phonemes, word pairs with similar pronunciations, and sentences/phrases with high accuracy. (5) With a reference magnetometer and advanced signal processing algorithms, the developed silent speech interface exhibited robustness against environmental acoustic noises, lighting conditions, and daily motion induced interferences, which are top concerns for acoustic-, visual-, and many sensor-based systems. Building upon the silent speech interface, two demonstrations, including a silent speech enabled smartphone assistant and drone control, were developed. These systems demonstrate the potential of the developed silent speech interfaces in assistive technology and human–machine interactions.

2. Results and discussions

2.1. Overview of the wireless silent speech interface based on soft magnetic skin

Fig. 1(a) provides the conceptual overview of the wireless silent speech interface. The interface consists of a magnetic skin

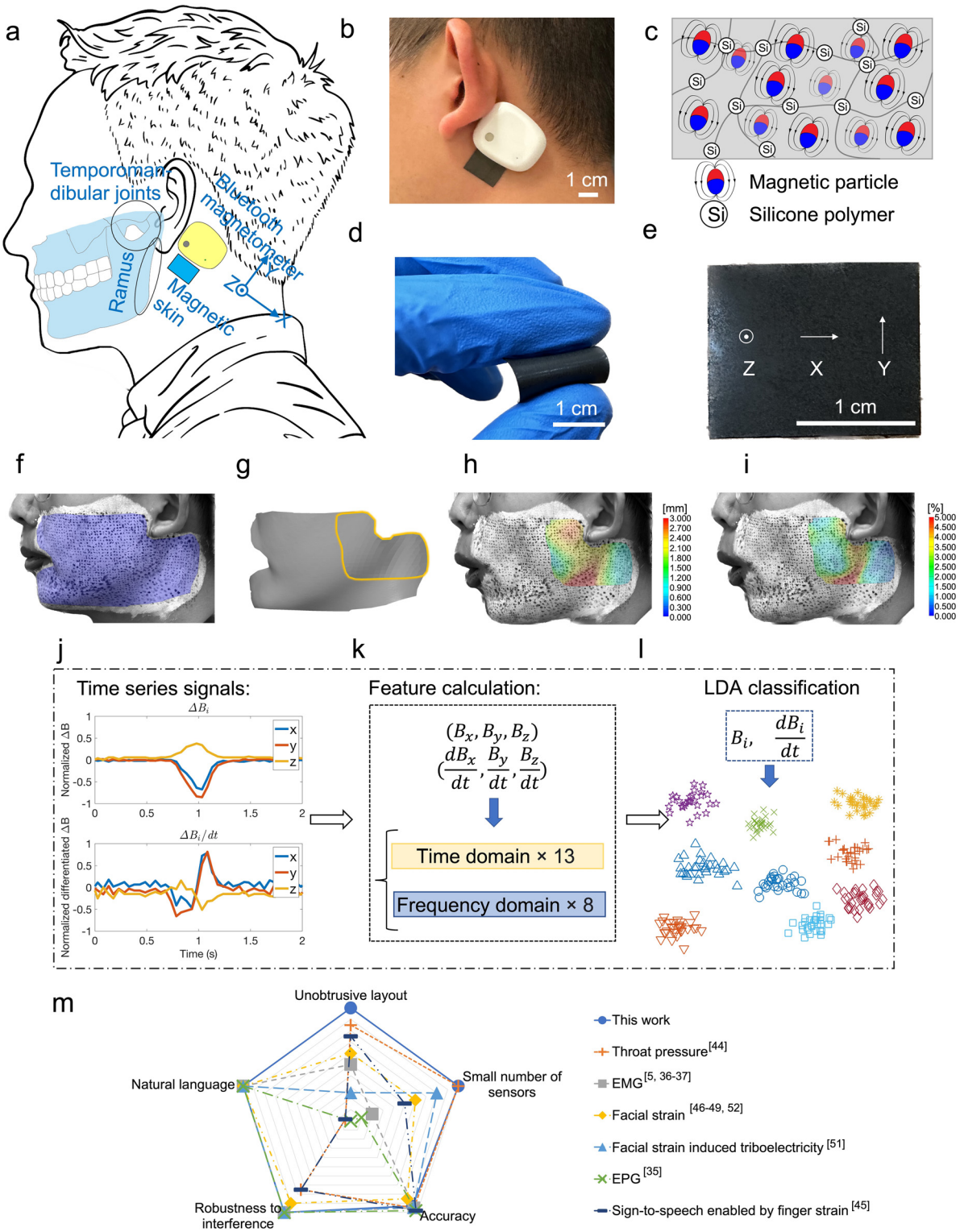

Fig. 1 Overview of the wireless silent speech interfaces based on soft magnetic skin. (a) Conceptual overview of the silent speech interface. (b) Photograph of a subject wearing the silent speech interface. (c) Structure of the magnetic skin. (d) Photograph showing the softness of the magnetic skin. (e) Illustration showing the magnetization direction of the magnetic skin. (f) DIC image of a face region painted with black dots. (g) 3D constructed model of the facial skin where the region of interest is indicated by the yellow outline. (h), (i) DIC images showing the displacement (h) and major strain (i) profiles when the subject is silently speaking the phoneme/o/. (j) One trial signal of the phoneme/m/. (k) Feature calculations in both the time and frequency domains. (l) LDA classifications based on the calculated features. (m) Comparison of different silent speech interfaces based on soft skin-worn sensors.

affixed to the skin area between the ramus and temporal bone, a working Bluetooth magnetometer attached to the temporal bone on one side of the head behind the ear, and another reference Bluetooth magnetometer attached to the temporal bone on the other side of the head (Fig. S1, ESI†). Soft magnetic skin is a composite material composed of small magnetic particles embedded in a soft polymer matrix.[53–56] The magnetic skin is on the skin beside the ramus, which is the junction of the mandible and sternocleidomastoid muscle. When a subject (Fig. 1(b)) attempts to speak by opening the mouth, the working magnetometer remains stationary, while the movement of temporomandibular joints causes displacement and strain changes in the magnetic skin. Consequently, the magnetic flux density captured by the magnetometer changes, which is highly correlated with the speech content. The reference magnetometer is included to minimize unwanted environmental and motion-induced noises, such as signal changes induced by geomagnetic fields and walking. The weight of one magnetic skin is approximately 0.2 g and one Bluetooth magnetometer is about 16 g. Thus, the weight of the total system is around 32.2 g (two magnetometers and one magnetic skin).

As depicted in Fig. 1(c), the magnetic skin is composed of magnetic particles dispersed in a matrix of silicone polymers. The softness enables it to conform to the contour of human skin and enhances its sensitivity to skin deformations (Fig. 1(d)). The magnetic skin (18 mm × 12 mm) consists of three units (6 mm × 12 mm each) with three different magnetization directions (Fig. 1(e)). The diverse magnetization directions within one single magnetic skin provide strong signals in all x, y, and z directions, thereby offering more valuable information for silent speech analysis. The DIC technique (Fig. 1(f)–(i)) is employed to analyse the displacement and strain changes during speech and determine the optimal location for the device placement. A three-dimensional model of the human facial skin is constructed (Fig. 1(f) and (g)). This allows for the measurement of displacement (Fig. 1(h)) and strain (Fig. 1(i)) by tracking the position and shape changes of small dots painted on the face. The region enclosed within the yellow line (Fig. 1(g)) was selected for analysis due to its unobtrusiveness. Attaching the speech interface within this region is much more socially acceptable compared to the area surrounding the lips.

Fig. 1(j)–(l) outline the brief process of ML-based silent speech recognition. The tri-axis working magnetometer effectively captures magnetic flux densities in three directions. To enhance the signal quality, the captured signals are first denoised using the signals acquired by the reference magnetometer. These signals are then differentiated with respect to time, yielding three additional signal channels. Thus, six channels of time series signals can be acquired. Based on six channels of signals, which are (B_x, B_y, B_z) and (dB_x/dt, dB_y/dt, dB_z/dt), multiple features related to speech recognition can be calculated and labelled for the following supervised learning. Linear discriminant analysis (LDA) is employed here to classify different silent speech contents based on the calculated features and labels (Fig. 1(l)).

2.2. Design, fabrication, and characterization of the soft magnetic skin

The fabrication process of the magnetic skin is depicted in Fig. 2(a), and magnetization directions are shown in Fig. 2(b)–(e).

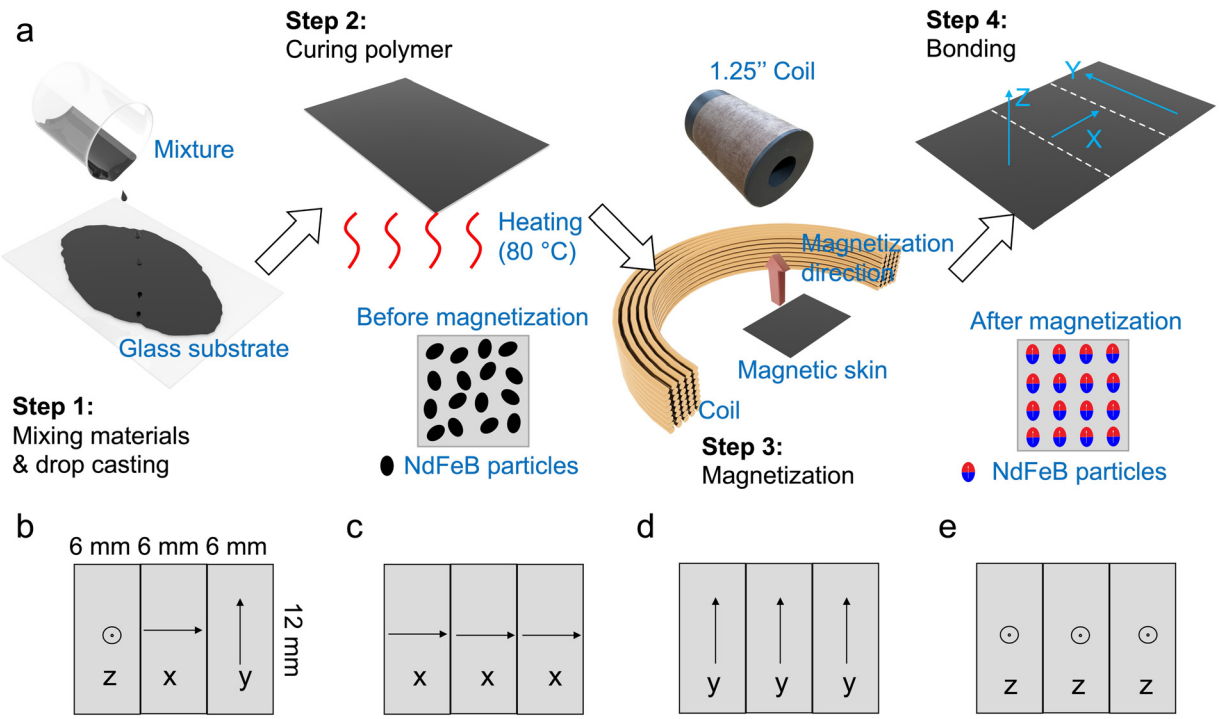

Fig. 2 Fabrication process and magnetization directions of the magnetic skins (a) Schematic illustration showing the fabrication process of the magnetic skin with different magnetization directions. (b)–(e) Schematic illustrations of magnetic skins composed of units with different magnetic directions: (b) *XYZ* sample, (c) *XXX* sample. (d) *YYY* sample. (e) *ZZZ* sample.

In brief, the magnetic skin consists of the magnetic layer and the adhesive layer. The magnetic layer is composed of NdFeB magnetic particles and the silicone polymer matrix. The mixture was first poured onto a glass substrate and heated to cure. The resulting composite thin film was then magnetized by an impulse magnetizer. Finally, three pieces of thin films with the same or different magnetization directions were assembled by a thin layer of Ecoflex Gel. The Ecoflex Gel also serves as the adhesive layer for attaching the sample to the skin.

The design goal of the magnetic skin is to achieve skin-like softness and optimal sensitivity and signal amplitude in the x, y, and z directions. We optimized the design of the magnetic skin in the following aspects: (1) selection of the silicone polymer; (2) weight ratio between the silicone polymer and magnetic particles; (3) magnetization directions. The silicone polymer serves as the matrix of the magnetic skin and plays a crucial role in the softness of the magnetic skin. Two different silicones are employed to fabricate the polymer matrix. Ecoflex 00-30 is introduced into Liveo MG 7-9900 with a weight ratio of 1 : 4 to render the elastomer free-standing while maintaining superior stretchability and softness.

To determine the optimal weight ratio of magnetic particles to the silicone polymer, the strain–stress curve (Fig. 3(a)) and magnetic flux density (Fig. 3(b)) were measured at different weight ratios. As can be expected, increasing the proportion of magnetic particles leads to an increase in the magnetic flux density. But in the meantime, the elastic modulus of the magnetic layer is increased. When the weight ratio (magnetic particles to silicone polymer) reaches 7 : 1 and 8 : 1, the magnetic skin becomes brittle, with fracture strains at 33% and 22%, respectively (Fig. 3(a)). The elastic modulus of the epidermis layer of human skin is approximately 1 MPa.[57] Considering that the elastic modulus of the sample at the ratio of 6 : 1 is approximately 0.84 MPa, this weight ratio is used for the following experiments to achieve a stretchable and skin-like magnetic skin, without sacrificing much of the magnetic flux density.

Magnetic skins composed of units with different magnetization directions (Fig. 2(b)–(e)) can provide different signal amplitudes in the x, y, and z directions. The ideal scenario is to obtain decent signal amplitudes in all three directions, which provides comprehensive information on speech-induced movements. Performances of samples composed of units with different magnetization directions (Fig. 3(c)–(k)) were tested using the setup shown in Fig. 3(l) and Fig. S2, ESI.† The magnetic skin was attached to a skin replica that was stretched from 0% to 10% strain along the y direction to mimic the skin deformation at the junction area of the mandible and sternocleidomastoid muscle during speech. The influence of the stretchability of the magnetic skin on the measured magnetic flux density was first evaluated. In the first set of experiments (Fig. 3(c), (e), (g), and (i)), as-prepared stretchable magnetic skins with different magnetization directions were tested. When stretching the skin replica, the magnetic skin experiences both displacement and strain changes, leading to variations in the magnetic flux density. In the second set of experiments (Fig. 3(d), (f), (h), and (j)), the stretchability of the as-prepared magnetic skin was constrained using a non-stretchable tape attached below it. When stretching the skin replica, the magnetic skin experiences only displacement change, while its strain is minimized by the strain-limiting layer. Notably, the signal amplitudes of magnetic skins undergoing both displacement and strain changes are larger than that of samples experiencing only displacement changes, indicating the advantage of stretchable magnetic skins. Among magnetic skins with different magnetization directions, the *XXX* sample shows minimal signal changes in the z direction (Fig. 3(c) and (d)), while *YYY* and *ZZZ* samples display extremely small signal amplitudes in the x direction. Only the *XYZ* sample demonstrates a decent signal amplitude in all three directions, although with a slightly reduced maximum amplitude compared to other samples. As depicted in Fig. S3, ESI,† the magnetic signals obtained from the magnetic skin (*XYZ* sample) exhibit good repeatability during over 1100 cycles of stretching/releasing at 10% strain. The adhesive layer maintains consistent adhesion to the skin replica during repeated stretching and releasing cycles. No delamination and fracture of the magnetic skin were observed visually or from signal changes. It is worth noting that the skin deformations during speech within the ramus area are typically within 5%, a strain level lower than the applied strain during testing. In addition, rigid permanent magnets (z-direction magnetized) assembled into a similar size to the magnetic skin (Fig. S4, ESI†) were tested for comparison. The signals measured from rigid magnets are much higher in amplitude in the y direction (Fig. 3(k)) compared to that from soft magnetic skins, due to a larger thickness and magnetic material density. However, the rigid magnet is unable to conform to the human skin, leading to a lower recognition accuracy compared to the magnetic skin (discussed in detail in Section 2.4).

2.3. Optimization of sensing location and speech dictionary for verification

The location to attach the magnetic sensor is optimized using the DIC technique. The process of speech generation can be divided into three sub-processes related to the lung, vocal cord, and articulator.[3,29] Initially, the air is inhaled by the lungs. The subsequent air pressure generated by the lungs causes the vocal cords to vibrate and produce sound. The sound is then shaped into recognizable speech through the movement of articulators such as the tongue, lips, teeth, and jaw. The process for silent speech is similar except that the vocal cord does not vibrate, resulting in the absence of audible sound production. The articulatory movements remain active during silent speech. Studying the movement of the tongue and teeth typically requires implantable devices. Additionally, the skin area around the lips is unsuitable for developing a socially acceptable device for daily use. Therefore, the skin area related to jaw movement was selected for our silent speech interface. The DIC technique was employed to measure displacement and strain changes to identify the region with the largest movements. This step is especially important for detecting jaw movement, as the

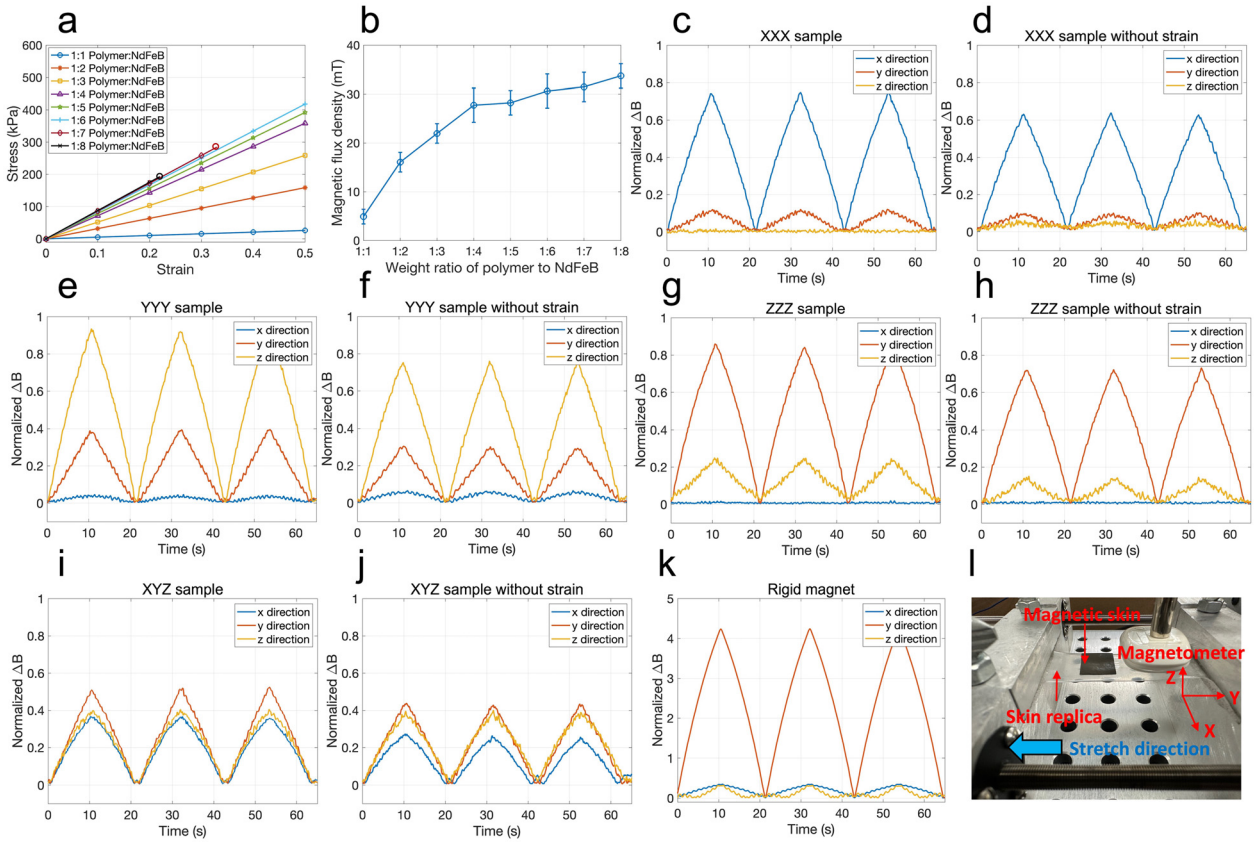

Fig. 3 Characterizations and optimizations of the magnetic skin. (a), (b) Stress–strain curves (a) and magnetic flux densities (b) for samples with different weight ratios between the polymer and magnetic particles. The error bar in (b) is due to the magnetic flux density differences between the edges and central points of the magnetic skin. (c)–(g) Magnetic flux density changes when the skin replica was stretched with a strain between 0 to 10% along the y axis three times. Magnetic skins composed of units with different magnetic directions were tested: (c) *XXX* sample, (d) *XXX* sample with strain-limiting tape, (e) *YYY* sample, (f) *YYY* sample without strain, (g) *ZZZ* sample, (h) *ZZZ* sample without strain, (i) *XYZ* sample, (j) *XYZ* sample without strain. (k) Magnetic flux density changes of the sample with rigid magnets. (l) Photograph of the setup for measuring magnetic flux density changes showing in (c)–(k). Changes of magnetic flux density in (c)–(k) are normalized by dividing all values by 100 μT and taking the absolute value.

skin deformation and displacement are more subtle compared to the skin around the lips.

Fig. 4 presents the displacement and major strain of the phoneme '/o/' and word 'pay' over time as examples. The images from 0 s to 1 s illustrate the process of mouth opening and closing. Two regions exhibit significant deformations and strains: the temporomandibular joint area and the ramus area (Fig. 1(a) and 4). The skin in the temporomandibular joint area experiences larger deformations due to joint rotation during silent speech. Similarly, the skin on the ramus, which is a part of the mandible, shows substantial deformations as the ramus rotates around the joints, moving toward the back of the head. In addition to determining the position of the magnetic skin, the placement of the magnetometer must also be considered. To minimize the noise caused by skin deformations, the magnetometer is attached to the skin above the temporal bone, close to the ear (Fig. S1, ESI†). The skin in that area is relatively unaffected by articulatory movements since the temporal bone is part of the skull. When ranking skin deformations from high to low, the order is as follows: ramus area close to the chin, ramus area close to the ear, temporomandibular joint area, and the remaining area. The magnetic skin is attached to the ramus area close to the ear (between the ramus and the temporal bone shown in Fig. S1, ESI†). The ramus area close to the chin was not selected because it is too far from the magnetometer (placed on the temporal bone), resulting in a significant attenuation of the magnetic signal due to increased distance. The size of the magnetic skin was determined as 18 mm × 12 mm (length × width) to sufficiently cover the ramus area close to the ear. The magnetic skin is aligned parallel to the magnetometer along its length, allowing for the closest possible distance between the magnetometer and the magnetic skin.

The recognition of different phonemes is of great significance for silent speech recognition, given that English is a language composed of a sequence of phonemes. A phoneme is a unit of sound that distinguishes the pronunciation of words.[58] To evaluate the effectiveness of our silent speech interface, the nine most frequently used phonemes in the English language, [3] namely/m/,/k/,/i/,/a/,/j/,/p/,/u/,/n/, and/o/, were selected for the study. Additionally, a very challenging list of words (Table S2, ESI†) was chosen to further test the

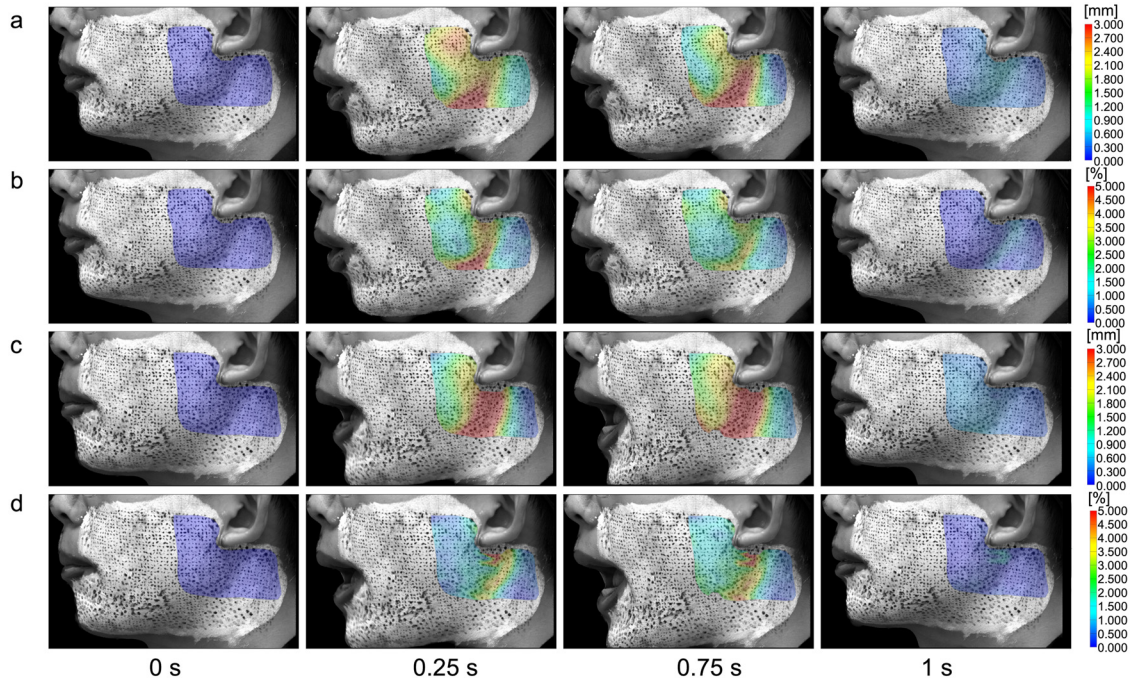

Fig. 4 DIC images of a subject showing the displacement and strain during the speech from the beginning to the end. The displacement (a) and major strain (b) of four frames when speaking the phoneme/o/. The displacement (c) and major strain (d) of four frames when speaking the word 'pay'.

capabilities of the developed silent speech interface. The list contains word pairs with similar pronunciations. For instance, although "pay" and "bay" contain different phonemes/p/and/b/, these two phonemes have similar pronunciations as they belong to the same viseme group (known as the Bilabial viseme).[5] The term "viseme" refers to a visual speech unit that includes phonemes with identical visual representations.[47] Essentially, when a subject attempts to articulate "pay" and "bay," the lip gestures and muscle movements will be quite similar, leading to comparable jaw movements. Moreover, in this word list, several words from different viseme groups contain the same element. For example, the words "pay", "bay", "kay", "gay", "way" all have "ay" as the ending phoneme. Successfully recognizing subtle differences in this word list is a difficult task for speech recognition.

2.4. Silent speech recognition by machine learning

Silent speech data was collected from five subjects (3 males and 2 females). Each phoneme was repeated fifty times by each subject to generate the training data. Detailed signal processing using ML methods can be seen in the experimental section. The results for different subjects are presented separately, as given in Fig. 5 and Fig. S5–S16, ESI.† This section elaborates on the results for subject 1. Statistical analysis for all five subjects can be seen in Table S3, ESI† and the Experimental section. Fig. 5(a) and (b) present a trial signal measured from the *XYZ* magnetic skin for all nine phonemes, including both the magnetic flux density signals and the signals after differentiation. The signals acquired from magnetic skins with other magnetization directions are shown in Fig. S5a–f, ESI,† which only have strong amplitudes in one or two directions. Signals from the *XYZ* magnetic skin exhibit good amplitude in all three directions. The resulting confusion matrix for the nine phonemes (Fig. 5(c)) demonstrates that the LDA model can effectively classify the phonemes, achieving an overall classification accuracy of 92.7%. The micro-average Receiver Operating Characteristic (ROC) curve is commonly used to evaluate the performance of a classification model by aggregating the true positive rate and false positive rate across all classes into a single metric.[59] Our system achieves a remarkable value of 0.994 (Fig. 5(d)), indicating the model's exceptional ability to discriminate between classes. Moreover, the optimal operating point at (0.02, 0.96), corresponding to the threshold that maximizes the overall classification performance across all classes,[59] further illustrates the model's high accuracy and reliability. It should be noted that the algorithm utilized in this study is user dependent due to the users' different ways of speech generation. Efforts were made to examine the accuracy of the model across various subjects. Sequential incorporation of data from subjects 2, 3, 4, and 5 into the training set was executed while using the data set from subject 1 as the testing data. The resulting recognition accuracy is notably low (Table S4, ESI†). One potential solution to achieve a universal model is to include a much larger training data from significantly more subjects and use deep learning techniques for speech recognition.

In addition to phonemes, a challenging list of words containing words from eleven visemes groups,[5] as discussed in Section 2.3, was also selected to test the developed silent speech interface. Each word was repeated fifty times. The resulting confusion matrix for these words (Fig. 5(e)) and micro-average ROC curve (Fig. 5(f)) demonstrate that the employed algorithms

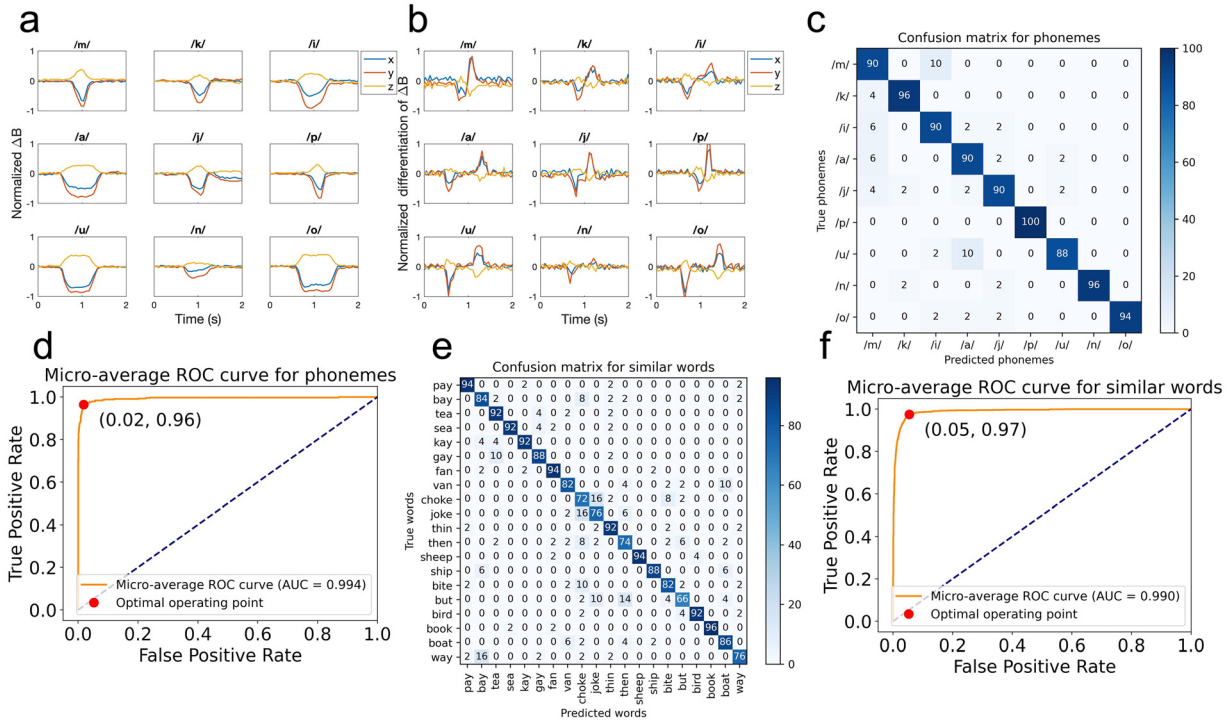

Fig. 5 Results of silent speech recognition (subject 1). (a) Time series signals of nine phonemes. The signals are normalized by dividing all values by 30 μT. (b) Time-series signals of the nine phonemes after differentiation. The signals after differentiation are normalized by dividing all values by 150 μT s^{-1}. (c) Confusion matrix for the nine phonemes using the LDA classifier. (d) Micro-average ROC curve for nine phonemes. (e) Confusion matrix for a list of words containing word pairs with similar pronunciations (from the same viseme group) using the LDA classifier. (f) Micro-average ROC curve of words with similar pronunciations.

can successfully recognize the words with an accuracy of 85.6%. This speech recognition accuracy is comparable to the previous system based on 8-channel EMG sensors placed around lips and on the neck and tested using the same list of words.[5] This demonstrates that the selected sensing area (the ramus area near the ear) contains rich information of speech articulation, comparable to the commonly used EMG methods. In despite of similar accuracies, only a single magnetic skin is needed in this work and the sensor placement location is much more socially acceptable. Additionally, the word list is expanded to 54 words (containing 20 words with similar pronunciations) to assess the performance as the word count increases. The silent speech recognition accuracy for 54 words is 85.7% (Fig. S6 and S7, ESI†), with only a marginal increase of 0.1%. The classification accuracy can be affected by several factors, including the increased number of classes and the difficulty level of the classification task. Here only signals from a single magnetic skin were used and the sensor placement is much more socially acceptable. These results demonstrate the effectiveness of the silent speech interface in recognizing both phonemes and words.

Furthermore, a comparison was made between the classification results obtained using the magnetic skin and the rigid magnet with a similar size (as mentioned in Section 2.2 and Fig. S4, ESI†). Despite that the signal strength of the rigid magnet (Fig. S5g–h, ESI†) is stronger than that of the magnetic skin, the speech recognition accuracy for nine phonemes is only about 75.6% (Fig. S8, ESI†), which is 17.1% lower than that of the magnetic skin. Due to its rigidity, the magnet could not conform to the skin topology and capture the subtle skin deformations during the speech, thereby resulting in reduced accuracy.

The developed silent speech interface also exhibits good robustness against interferences. In an environment with ambient noises of 80 dB, the classification accuracy is about 93.3% for the nine phonemes (Fig. S9, ESI†). Besides, under a dark environment, the classification accuracy is maintained (92.8%) (Fig. S10, ESI†). The variation is small, within 0.6%. Although these three sets of data were acquired from the same subject and the signals are highly similar, the subject could not control his/her speech muscle movements exactly the same when signals were obtained for the normal condition, noisy environment, and dark environment. Consequently, it is reasonable to have small variations when processing these three datasets for speech recognition. Similarly, during daily motions, such as walking at a speed of 0.8 m s^{-1}, the interface achieves an accuracy of approximately 87.8% (Fig. S11a and b, ESI†), after a calibration process using the data from the reference magnetometer attached beside the other ear. Without the calibration process, the speech recognition accuracy is only 54.4% due to the motion-induced interference (Fig. S11c and d, ESI†). The Kabsch algorithm[60] was utilized to perform calibration. The rotation matrix between the working and reference magnetometers was first calculated and then the motion-related signals captured by the reference magnetometer were subtracted from

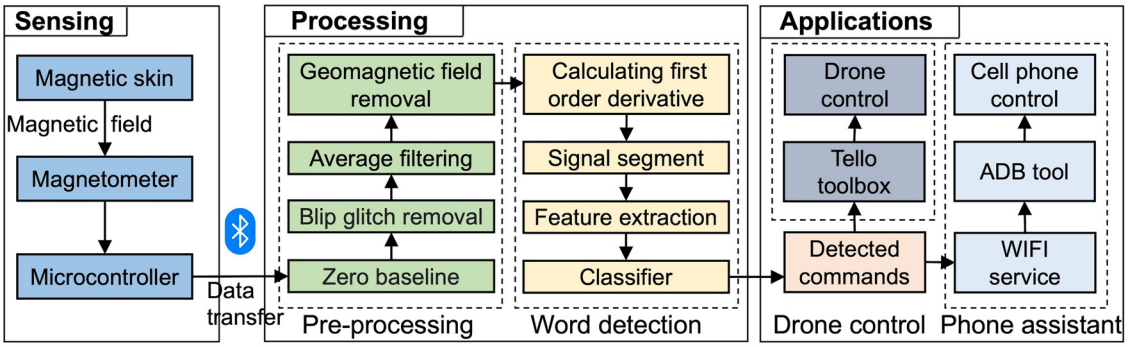

Fig. 6 Flowchart of data acquisition, signal processing, and application development.

signals detected by the working magnetometer. In this way, the influence of motion-induced interferences can be significantly reduced. The detailed calibration process is shown in the Experimental section and ESI.† These experiments conducted under normal, noisy, dark, and motion conditions collectively illustrate the insensitivity of the developed silent speech interfaces to acoustic noises, lighting conditions, and daily motions, which are top concerns for acoustic-based, visual-based, and many sensor-based speech recognition methods.

Experiments were conducted to examine the device variations. Two distinct magnetometers were consecutively affixed at the same position, measuring the magnetic flux density alteration as the magnetic skin underwent stretching. As depicted in Fig. S12a, ESI,† the signal variations were found to be minor, and the largest difference is approximately 1 μT (2% variation) in the y direction. In order to assess the influence of device variations on the recognition accuracy, we obtained silent speech signals using different magnetometers and magnetic skins. Another set of phonemes (/m/,/k/,/i/,/a/,/j/,/p/,/u/,/n/,/o/) data was acquired and added to the training set for subject 1. The final confusion matrix is presented in Fig. S12b, ESI,† revealing an accuracy of 93.1%, a value slightly higher than the original accuracy of 92.6%.

2.5. Applications in phone assistant and drone control

Two demonstrations were developed to illustrate the potential of the developed wireless silent speech interface in assistive technology and human–machine interactions. Fig. 6 presents the flow chart of the development process, and detailed descriptions can be found in the experimental section. In the first demonstration (silent speech-based phone assistant, shown in Fig. 7(a), (b), (d), (e), (g), and (i)), the silent speech is utilized as an alternative input modality to replace the voiced speech to assist in the cell phone control. The silent speech assistant is based on the Android system, and six sentences/phrases (Fig. 7(d) and (e)) were pre-trained to interact with the phone as examples. The confusion matrix (Fig. 7(g)) and ROC curve (Fig. 7(i)) demonstrate the remarkable performance of the silent speech interface in recognizing diverse sentences/phrases, achieving an overall accuracy of 96.7%. The entire application is in real-time. When the user speaks the specified sentences/phrases silently, the intended speech information is interpreted from the acquired magnetic signals by the pre-trained ML model and sent to the Android phone. The corresponding tasks are then executed on the smartphone. With the silent speech assistant, users can perform various operations on their smartphones, such as playing music or opening apps (see Supporting Video 1, ESI†).

In the second demonstration (silent speech-based drone control, shown in Fig. 7(b), (c), (f), (h), and (j)), silent speech interfaces are used for human–machine interactions. A Tello drone is used as an example to receive the commands delivered by silent speech and execute corresponding movements. Drones have been widely used in various inspection tasks and voice control has been implemented to enable intuitive and hand-free communications between the operator and the drone.[61] Eight commands (shown in Fig. 7(f)) for drone control were pre-trained. The confusion matrix (Fig. 7(h)) and the ROC curve (Fig. 7(j)) indicate a good accuracy of 93.5%. Supporting Video 2, ESI† shows the process of drone control by the silent speech interface.

3. Conclusions

In conclusion, this study presents a wireless, unobtrusive, robust, and accurate silent speech interface through comprehensive explorations of materials, structural design, sensing location, ML Methods, and noise reduction algorithms. The cost of a single magnetic skin is approximately $0.6 and the Bluetooth magnetometer for data acquisition is around $130 each. The costs can be further reduced for mass production. An average recognition accuracy of 93.2% was achieved for phonemes and 87.3% for a list of words containing words from the same viseme group. Two proof-of-concept applications were developed that demonstrate the system's capability to decode silent speech signals in real time and enable interactions with external devices. The silent speech interface provides a novel communication interface, which can find broad applications in assistive technology for voice disorders, robot control, and human–machine collaborative systems.

4. Experimental section

Materials

NdFeB magnetic particles (MQP-15-7-20065) were provided by Magnequench. The toluene solvent was purchased from

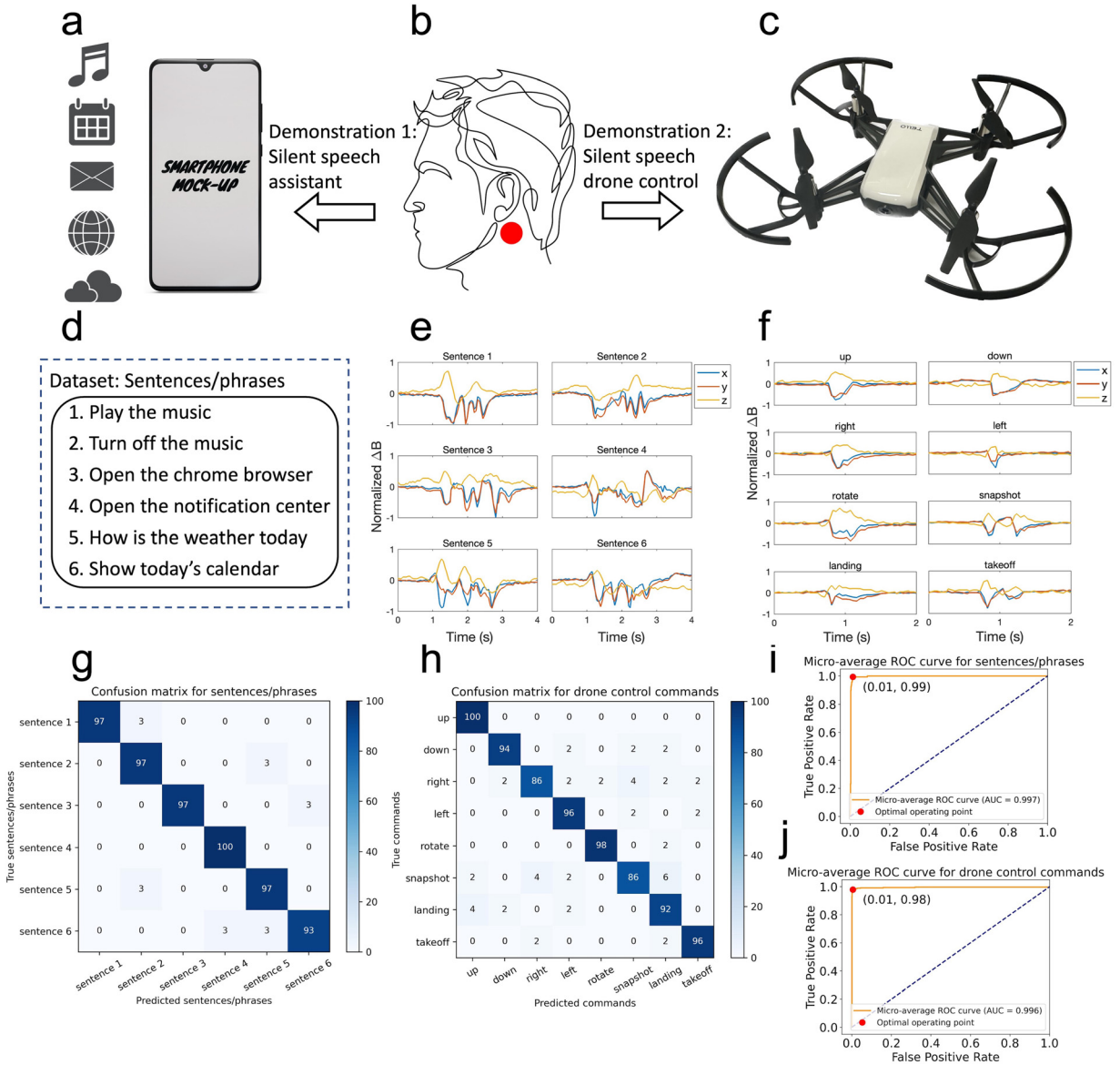

Fig. 7 Two demonstrations based on magnetic skin-enabled silent speech interfaces. (a) Demonstration 1: silent speech assistant for smartphone control. (b) Schematic of a user wearing the silent speech interface. (c) Demonstration 2: silent speech interaction for drone control. (d) Selected sentences/phrases for demonstration 1. (e) Time series signals of the selected sentences/phrases for demonstration 1. (f) Time series signals of the selected commands for demonstration 2. (g) Confusion matrix of selected sentences/phrases for demonstration 1. (h) Confusion matrix of selected commands for demonstration 2. (i) Micro-average ROC curve corresponding to (g). (j) Micro-average ROC curve corresponding to (h).

Sigma-Aldrich. The silicone adhesive (Liveo™ MG 7-9900) was provided by Knowde. The silicone elastomer (Ecoflex™ 00-30) and release agent (Easy Release™ 200) were obtained from Smooth-On. The clown white and custom body paint were obtained from Mehron and TCP Global, respectively. All materials were used as received.

Fabrication of the magnetic skin

Fig. 2 depicts the fabrication process of the magnetic films. 0.20 g silicone adhesive part A, 0.20 g silicone adhesive part B, 0.05 g silicone elastomer part A, 0.05 g silicone elastomer part B, and 1.5 g Toluene were first mixed using the mixer (AR-100, Thinky) at 1000 rpm for 30 s. Subsequently, 3.0 g NdFeB micro magnetic particles were introduced to the mixture and mixed at 1000 rpm for an additional 30 s. The release agent was then uniformly sprayed onto a glass substrate (5.08 cm by 7.62 cm) followed by a drying period of 5 minutes at the ambient temperature. Next, a 3 g portion of the mixture was poured onto the prepared glass substrate and allowed to dry naturally for 30 minutes to evaporate the toluene. After the toluene had completely evaporated, the sample was cured at 80 °C for 1 hour. Afterward, three pieces of the sample with a length of 6 mm and a width of 5 cm were cut off. The three sample pieces were then placed into two distinct 3D-printed molds (Fig. S17, ESI†) with different orientations, after which they were magnetized along the x, y, and z directions using an

impulse magnetizer (IM-10-30, ASC Scientific). The impulse magnetizer employed a 1.25″ coil that yielded a resulting magnetic flux density of 2.6 T. Finally, the three sample pieces were arranged on the glass substrate and bonded together using silicone adhesive.

Measurement of the strain–stress curve

Measurement of the strain–stress curve was conducted using a material testing system (858 Mini Bionix II, MTS) operating at a constant speed of 10 mm min^{-1}. The load cell embedded within the system offers a resolution of 0.001 N. The magnetic skin was applied with 50% strain. Simultaneous recording of distance and tensile force was performed at a frequency of 100 Hz. The stress values were obtained by dividing the tensile force by the cross-sectional area of the magnetic skin. The strain values were calculated by dividing the distance by the length of the magnetic skin.

Comparisons of magnetic flux densities for samples with varying material ratios

A series of samples with varying weight ratios (1 : 1, 1 : 2, 1 : 3, 1 : 4, 1 : 5, 1 : 6, 1 : 7, and 1 : 8) between the silicone mixture and magnetic particles were prepared. From each sample, pieces measuring 5 mm × 5 mm × 0.9 mm were cut. Subsequently, magnetic flux densities of all sample pieces were measured using a magnetometer (Metamotions, Mbientlab). The size of the magnetic skin is larger than the magnetometer chip. Therefore, multiple locations were measured, including the edges and central points of the magnetic skin. The magnetic flux density at the central point is larger than that at the edges. The results shown in Fig. 3(b) indicate the mean value and the error bar.

Measurement of changes in magnetic flux density under strains

The Dragon Skin 20 elastomer (Smooth-On) was selected to fabricate the skin replica due to its similar elastic modulus compared to the human skin.[62] Parts A and B were mixed in a 1 : 1 ratio using a mixer (AR-100, Thinky) operating at 1000 rpm for 30 s followed by curing at 80 °C for 1 hour. The resulting cured Dragon Skin elastomer was then cut into a rectangular piece measuring 90 mm × 25 mm. During the investigation of magnetic flux changes, the magnetic skin was affixed onto the Dragon Skin elastomer, which was subjected to stretching using a customized tensile stage (Fig. 3(l) and Fig. S2a, ESI†) at a controlled speed of 0.75 mm s^{-1} until reaching a strain of 10%. Meanwhile, a magnetometer (Metamotions, Mbientlab) was used for measuring the magnetic flux change (Fig. S2b, ESI†). To reveal the effect of stretchability on the magnetic flux changes, after testing the as-prepared stretchable magnetic skins, tapes (Transpore™ 3M) were attached to the magnetic skin and placed on the skin replica to limit its stretchability. The skin replica was once again stretched to achieve a strain of 10%, while the magnetometer simultaneously recorded the magnetic flux changes. Another experiment was conducted to assess the repeatability of the magnetic skin with an optimized ratio of 6 : 1, under a 10% strain. Over 1100 cycles of stretching/releasing were conducted using the same setup as shown in Fig. 3(l).

DIC analysis of the movement patterns of the temporomandibular joint region

The 3D-DIC system (Trilion) was used to obtain the motion of the temporomandibular joint region. The hardware setup of the 3D-DIC system is illustrated in Fig. S18a, ESI.† Two cameras were positioned at an approximate angle of 30° before calibration. Thirteen pairs of photographs of the calibration pad (Fig. S18b, ESI†), taken at various angles, were acquired using the cameras. The distance between the DIC system and the calibration pad was approximately 1 m. These photographs were subsequently imported into the commercial software GOM Correlate for calibration.

To capture images of the subject's skin surface of interest (temporomandibular joint region), the area was cleaned with water and gently dried with paper towels. Fig. S19 and S20, ESI† show the skin painting process. Clown white makeup (Mehron) was applied to the target skin area using a paintbrush (Amazon Basics), as demonstrated in Fig. S20a, ESI.† Afterward, a mask (Fig. S19, ESI†) was cut by a mechanical plotter (Cameo 4, Silhouette). The mask pattern, generated using a MATLAB toolbox,[63] consisted of 40 × 40 holes with random shapes, spanning a 10 cm square. The random shapes could help the DIC system track skin motion more effectively. The mask was then affixed to the skin area of interest and secured using tapes (Fig. S20b, ESI†). Next, black custom body paint makeup (TCP Global) was sprayed onto the skin surface with a white background using an airbrush (Model G222, Master Airbrush). Following the removal of the mask, a marker pen (Sharpie) was used to add dots and fill any remaining blank areas (Fig. S20c, ESI†) to get the final appearance (Fig. S20d, ESI†). All makeup products employed on the skin were FDA-approved, biocompatible, and easy to clean, ensuring their safety and compatibility with human skin.[62] During the experimental setup, the subject was seated in a chair while maintaining the head position aligned with the previously placed calibration pad. The subject silently spoke various phonemes, words, and sentences, while the DIC system captured photos at a frequency of 4 Hz. Following the photo collection, all images were imported into the GOM Correlate software to calculate the displacement and strain data for the skin area of interest.

Signal acquisition and processing of silent speech

All signal processing was performed using the Python programming language. The collected signals were transmitted wirelessly to a laptop *via* Bluetooth. The signal processing procedure is outlined as follows: (1) zero baseline: the change of the magnetic flux density is the key to recognition. Hence, zero-baseline manipulations were first conducted by subtracting the average of the first ten data points to get the time series data ΔB. (2) Blip glitch removal: the blip glitch is a common noise problem of time series data. The signals were first traversed to find each blip glitch. One blip glitch was identified by comparing the change between two data points. When detected, a blip glitch

data point was replaced with the value of the preceding data point. (3) Average filtering: an average filter that averaged the adjacent three data points was applied to the signal. This process served to smooth out the signal and reduce noise. (4) Interference removal: interferences include the geomagnetic field and motion artifacts. The geomagnetic field was assumed evenly distributed around the subject. A magnetometer placed behind the subject's right ear captured interference signals, while another magnetometer behind the subject's left ear detected both magnetic skin signals and interferences. As shown in Fig. S1, ESI,† the coordinate of the working (left) magnetometer was denoted as X, Y, and Z, while the coordinate of the reference (right) magnetometer was denoted as x, y, and z. The rotation matrix between the two coordinates can be determined by eqs (S1)–(S5) (ESI). Hence the noise caused by interferences can be cancelled (Fig. S21, ESI†) by using Equation S6. A detailed description can be found in the 'Method for eliminating signal interferences' section of the ESI.† (5) Differentiation: differentiation was performed by calculating the rate of change of the signal over time, resulting in another set of data for subsequent analysis. The average filter can help to reduce the noise. (6) Signal segment: each phoneme, word, and sentence was repeated 50 times so the segment program helped extract the data of each trial. (7) Feature extraction: thirteen features in the time domain and eight features in the frequency domain, as listed in Table S5, ESI,† were calculated. (8) Classifier: the LDA algorithm was applied to classify silent speech. Five-fold validation was employed for the evaluation of the classifier. The calculated feature data and label data were input into the algorithm to obtain the results.

Demonstration of phone assistant

After training the LDA classification model, the model was utilized for real-time identification of the collected signals. A phone assistant using the silent speech interface was developed. To establish a connection between an Android phone (Moto G) and a Windows laptop (Dell Latitude 7410), both devices were connected to the same WIFI network. Then the software Android Studio was used to enable the wireless data transfer between the phone and the laptop. The Android debug bridge (ADB)[64] command-line tool was integrated into the Python program running on the laptop. After the collected signals were converted into sentences or phrases by the pre-strained ML model, the ADB tool would transmit the corresponding command line to the phone. The phone would then perform the corresponding task, as requested by the command line. Supporting Video 1, ESI† presents a demonstration of the phone assistant enabled by silent speech recognition.

Demonstration of drone control

Similar to the phone assistant application, the captured signals were decoded in real-time by a pre-trained ML model for drone control. The drone (Ryze Tech Tello) was connected to a Windows laptop (Dell Latitude 7410) *via* WIFI service. The Python package DJITelloPy[65] was employed for the program running on the laptop. When the model classified the specific command words from the acquired magnetic skin signals, these words were transmitted from the laptop to the drone using a designated command line enabled by DJITelloPy. The drone would subsequently execute the corresponding movement as controlled by the command line. Supporting Video 2 (ESI†) presents a demonstration of the drone control application.

Statistical analysis

All characterization measurements were performed five times for each sample piece, and the average value was selected to represent the final result. For silent speech recognition, magnetic signals were collected from five subjects, comprising 3 male and 2 female subjects, with ages ranging from 20 to 30 years old. The authors have complied with all relevant ethical regulations. Study procedures were conducted in accordance with the guidelines provided by Stony Brook University. Prior to participation, informed consent was obtained from all subjects. Fig. 5(c)–(f) and Fig. S13–S16, ESI† provide a comprehensive summary of the confusion matrices and ROC curves for five individual subjects. The recognition accuracy (Table S3, ESI†) for the nine phonemes was measured to be 93.2% $\pm$ 2.62%. For a dictionary containing word pairs from the same viseme group, it was determined to be 87.3% $\pm$ 2.14%. These recognition accuracies are presented as the mean $\pm$ standard deviation. The statistical analysis was conducted using the MATLAB software.

Data availability

All data needed to evaluate the conclusions in the paper are presented in the manuscript and the ESI.† Additional data related to this paper may be provided on request.

Author contributions

Conceptualization: Penghao Dong and Shanshan Yao. Methodology: Penghao Dong, Shanshan Yao, Yizong Li, and Si Chen. Investigation: Penghao Dong and Shanshan Yao. Validation: Penghao Dong and Shanshan Yao. Operation: Penghao Dong. Data collection and processing: Penghao Dong, Justin T. Grafstein, and Irfaan Khan. Visualization: Penghao Dong and Shanshan Yao. Supervision: Shanshan Yao. Writing – original draft: Penghao Dong. Writing – review & editing: Shanshan Yao.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors express sincere gratitude to Professor Petar Djuric from the Department of Electrical and Computer Engineering and Professor Shubham Jain from the Department of Computer Science at Stony Brook University for helpful discussions. We also would like to thank Professors Paolo Celli and

Rigoberto Burgueno from the Department of Civil Engineering at Stony Brook University for providing support in Digital Image Correlation (DIC). This material is based upon work supported by the National Science Foundation under award no. ECCS-2129673 and 2238363.

References

1 Q. Yang, W. Jin, Q. Zhang, Y. Wei, Z. Guo, X. Li, Y. Yang, Q. Luo, H. Tian and T.-L. Ren, *Nat. Mach. Intell.*, 2023, **5**, 169–180.
2 S. Brown, A. R. Laird, P. Q. Pfordresher, S. M. Thelen, P. Turkeltaub and M. Liotti, *Brain Cogn.*, 2009, **70**, 31–41.
3 P. Khanna, T. Srivastava, S. Pan, S. Jain and P. Nguyen, Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications, Virtual, United Kingdom, 2021.
4 J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. Martin Donas, J. L. Perez-Cordoba and A. M. Gomez, *IEEE Access*, 2020, **8**, 177995.
5 P. Dong, Y. Song, S. Yu, Z. Zhang, S. K. Mallipattu, P. M. Djuric and S. Yao, *Small*, 2023, **19**, e2205058.
6 B. J. Betts, K. Binsted and C. Jorgensen, *Interact. Comput.*, 2006, **18**, 1242–1259.
7 Wikipedia, Silent Speech Interface, **https://en.wikipedia.org/wiki/Silent_speech_interface**, (accessed May, 2023).
8 W. Lee, J. J. Seong, B. Ozlu, B. S. Shim, A. Marakhimov and S. Lee, *Sensors*, 2021, **21**, 22.
9 N. Kimura, K. Hayashi and J. Rekimoto, Proceedings of the International Conference on Advanced Visual Interfaces, Salerno, Italy, 2020.
10 L. Pandey and A. S. Arif, Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 2021.
11 A. Fernandez-Lopez and F. M. Sukno, *Image Vis.*, 2018, **78**, 53–72.
12 K. Sun, C. Yu, W. Shi, L. Liu and Y. Shi, Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology, Berlin, Germany, 2018.
13 T. Afouras, J. S. Chung, A. Senior, O. Vinyals and A. Zisserman, *IEEE Trans. Pattern Anal. Mach. Intell*, 2018, **44**, 11.
14 Y. Mroueh, E. Marcheret and V. Goel, Proceedings of 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing South Brisbane, Australia, 2015.
15 S. Petridis, T. Stafylakis, P. C. Ma, G. Tzimiropoulos and M. Pantic, Proceedings of 2018 IEEE Workshop on Spoken Language Technology, Athens, Greece, 2018.
16 T. G. Csapo, C. Zainko, L. Toth, G. Gosztolya and A. Marko, Proceedings of Interspeech 2020, Shanghai, China, 2020.
17 N. Kimura, M. Kono and J. Rekimoto, Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 2019.
18 Y. Jin, Y. Gao, X. Xu, S. Choi, J. Li, F. Liu, Z. Li and Z. Jin, Proc. ACM interact. mob. wearable ubiquitous technol., 2022, 6, 1-28.
19 J. Tan, C.-T. Nguyen and X. Wang, Proceedings of 2017 IEEE Conference on Computer Communications, Atlanta, USA, 2017.
20 Y. Gao, Y. Jin, J. Li, S. Choi and Z. Jin, Proc. ACM interact. mob. wearable ubiquitous technol., 2020, 4, 1-27.
21 Q. Zhang, D. Wang, R. Zhao and Y. Yu, Proc. ACM interact. mob. wearable ubiquitous technol., 2021, 5, 1-28.
22 Y. Zhang, W.-H. Huang, C.-Y. Yang, W.-P. Wang, Y.-C. Chen, C.-W. You, D.-Y. Huang, G. Xue and J. Yu, Proc. ACM interact. mob. wearable ubiquitous technol., 2020, 4, 1-26.
23 Y. Zhang, Y.-C. Chen, H. Wang and X. Jin, Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers, Virtual USA, 2021.
24 B. Denby and M. Stone, Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Canada, 2004.
25 A. Jaumard-Hakoun, K. Xu, C. Leboullenger, P. Roussel-Ragot and B. Denby, Proceedings of Interspeech 2016, San Francisco, USA, 2016.
26 A. Brownlee and L. Bruening, Living with ALS: Changes in Speech and Communication Solutions, **https://arc.phhp.ufl.edu/wordpress/files/2017/06/Living-with-ALS-_-Speech.pdf**, (accessed Jun, 2022).
27 K. Brigham and B. V. Kumar, Proceedings of the 2010 4th International Conference on Bioinformatics and Biomedical Engineering, Chengdu, China, 2010.
28 P. Suppes, Z. L. Lu and B. Han, *Proc. Natl. Acad. Sci. U. S. A.*, 1997, **94**, 14965–14969.
29 T. Srivastava, P. Khanna, S. Pan, P. Nguyen and S. Jain, Proc. ACM interact. mob. wearable ubiquitous technol., 2022, 6, 1-26.
30 J. Rekimoto and Y. Nishimura, Proceedings of Augmented Humans Conference 2021, Rovaniemi, Finland 2021.
31 H. Sahni, A. Bedri, G. Reyes, P. Thukral, Z. Guo, T. Starner and M. Ghovanloo, Proceedings of the 2014 ACM International Symposium on Wearable Computers, Seattle, USA, 2014.
32 J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore and E. Holdsworth, *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2017, **25**, 2362–2374.
33 B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert and J. S. Brumberg, *Speech Commun.*, 2010, **52**, 270–287.
34 R. Hofe, S. R. Ell, M. J. Fagan, J. M. Gilbert, P. D. Green, R. K. Moore and S. I. Rybchenko, *Speech Commun.*, 2013, **55**, 22–32.
35 N. Kimura, T. Gemicioglu, J. Womack, R. Li, Y. Zhao, A. Bedri, Z. Su, A. Olwal, J. Rekimoto and T. Starner, Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, USA, 2022.
36 Y. H. Wang, T. Y. Tang, Y. Xu, Y. Z. Bai, L. Yin, G. Li, H. M. Zhang, H. C. Liu and Y. A. Huang, *npj Flexible Electron.*, 2021, **5**, 1–9.
37 H. Liu, W. Dong, Y. Li, F. Li, J. Geng, M. Zhu, T. Chen, H. Zhang, L. Sun and C. Lee, *Microsyst. Nanoeng.*, 2020, **6**, 16.

38 A. Kapur, S. Kapur and P. Maes, Proceedings of 23rd International Conference on Intelligent User Interfaces, Tokyo, Japan, 2018.
39 J. M. Vojtech, M. D. Chan, B. Shiwani, S. H. Roy, J. T. Heaton, G. S. Meltzner, P. Contessa, G. De Luca, R. Patel and J. C. Kline, *J. Speech Lang. Hear. Res.*, 2021, **64**, 2134–2153.
40 G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy and J. C. Kline, *J. Neural. Eng.*, 2018, **15**, 046031.
41 D. Gaddy and D. Klein, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Virtual, USA, 2020.
42 S. Yao, W. Zhou, R. Hinson, P. Dong, S. Wu, J. Ives, X. Hu, H. Huang and Y. Zhu, *Adv. Mater. Technol.*, 2022, **7**, 2101637.
43 W. Zhou, S. Yao, H. Wang, Q. Du, Y. Ma and Y. Zhu, *ACS Nano*, 2020, **14**, 5798–5805.
44 S. Xu, J. X. Yu, H. Guo, S. Tian, Y. Long, J. Yang and L. Zhang, *Nat. Commun.*, 2023, **14**, 219.
45 Z. Zhou, K. Chen, X. Li, S. Zhang, Y. Wu, Y. Zhou, K. Meng, C. Sun, Q. He, W. Fan, E. Fan, Z. Lin, X. Tan, W. Deng, J. Yang and J. Chen, *Nat. Electron.*, 2020, **3**, 571–578.
46 T. Kim, Y. Shin, K. Kang, K. Kim, G. Kim, Y. Byeon, H. Kim, Y. Gao, J. R. Lee, G. Son, T. Kim, Y. Jun, J. Kim, J. Lee, S. Um, Y. Kwon, B. G. Son, M. Cho, M. Sang, J. Shin, K. Kim, J. Suh, H. Choi, S. Hong, H. Cheng, H. G. Kang, D. Hwang and K. J. Yu, *Nat. Commun.*, 2022, **13**, 5815.
47 J. Wang, C. Pan, H. Jin, V. Singh, Y. Jain, J. I. Hong, C. Majidi and S. Kumar, Proc. ACM interact. mob. wearable ubiquitous technol., 2019, 3, 1-24.
48 H. Yoo, E. Kim, J. W. Chung, H. Cho, S. Jeong, H. Kim, D. Jang, H. Kim, J. Yoon, G. H. Lee, H. Kang, J. Y. Kim, Y. Yun, S. Yoon and Y. Hong, *ACS Appl. Mater. Interfaces*, 2022, **14**, 54157–54169.
49 L. Cheng, D. Q. Ruan, Y. W. He, J. Y. Yang, W. Qian, L. W. Zhu, P. D. Zhu, H. P. Wu and A. P. Liu, *J. Mater. Chem. C*, 2023, **1**, 1.
50 Y. Li, Y. Liu, S. R. A. Bhuiyan, Y. Zhu and S. Yao, *Small Struct.*, 2021, **3**, 2100131.
51 Y. Lu, H. Tian, J. Cheng, F. Zhu, B. Liu, S. Wei, L. Ji and Z. L. Wang, *Nat. Commun.*, 2022, **13**, 1401.
52 Y. Kunimi, M. Ogata, H. Hiraki, M. Itagaki, S. Kanazawa and M. Mochimaru, Proceedings of Augmented Humans 2022, Kashiwa, Japan, 2022.
53 Y. Alapan, A. C. Karacakol, S. N. Guzelhan, I. Isik and M. Sitti, *Sci. Adv.*, 2020, **6**, eabc6414.
54 H. Song, H. Lee, J. Lee, J. K. Choe, S. Lee, J. Y. Yi, S. Park, J. W. Yoo, M. S. Kwon and J. Kim, *Nano Lett.*, 2020, **20**, 5185–5192.
55 J. Tian, X. Zhao, X. D. Gu and S. Chen, Proceedings of 2020 IEEE International Conference on Robotics and Automation, Paris, France, 2020.
56 J. Tian, M. Li, Z. Han, Y. Chen, X. D. Gu, Q. J. Ge and S. Chen, *Comput. Methods Appl. Mech. Eng.*, 2022, **389**, 114394.
57 C. Li, G. Guan, R. Reif, Z. Huang and R. K. Wang, *J. R. Soc., Interface*, 2012, **9**, 831–841.
58 P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar and J. Vepa, Proceedings of the 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2018.
59 D. M. W. Powers, *J. Mach. Learn. Technol.*, 2011, **2**, 37–63.
60 Wikipedia, Kabsch algorithm, **https://en.wikipedia.org/wiki/Kabsch_algorithm**, (accessed May, 2023).
61 Y. Li, A. Parsan, B. Wang, P. Dong, S. Yao and R. Qin, *Eng. Appl. Artif. Intell.*, 2023, **117**, 105597.
62 T. Sun, F. Tasnim, R. T. McIntosh, N. Amiri, D. Solav, M. T. Anbarani, D. Sadat, L. Zhang, Y. Gu, M. A. Karami and C. Dagdeviren, *Nat. Biomed. Eng.*, 2020, **4**, 954–972.
63 G. Kwiatek, DXFLib, **https://www.mathworks.com/matlabcentral/fileexchange/33884-dxflib**, (accessed May, 2023).
64 Google, Android Debug Bridge (adb), **https://developer.android.com/tools/adb**, (accessed May, 2023).
65 DJI, DJITelloPy, **https://github.com/damiafuentes/DJITelloPy/blob/master/README.md**, (accessed May, 2023).

