Acoustic-based Alphanumeric Input Interface for Earables

Yilin Wang*
Department of computer science
Florida State University
Tallahassee, USA
ywang40@fsu.edu

Zi Wang*
School of Computer and Cyber Sciences
Augusta University
Augusta, USA
zwang1@augusta.edu

Jie Yang†
Department of computer science
Florida State University
Tallahassee, USA
jie.yang@cs.fsu.edu

***Abstract*—As earables gain popularity, there emerges a need for intuitive user interfaces that adapt to diverse daily scenarios. Traditional methods like touchscreens and voice control often fall short in environments like movie theatres, where silence and darkness are required, or on busy streets where visual distraction introduces extra risk. We propose an innovative earable-based system utilizing unique acoustic friction generated by fingers for alphanumeric input. Our approach digs into the acoustic friction theory, applying this knowledge to better understand the transformation from 2D handwriting into a 1D acoustic time series. This theoretical foundation guides our system design and feature extraction. Specifically, we have redesigned certain characters to enhance their acoustic distinctiveness without compromising the natural handwriting style of users, ensuring the system user-friendly. Our system combines DenseNet and GRU architectures in a multimodal model, refined through transfer learning to adapt to diverse user behaviors. Tested in real-world scenarios with 10 participants, our system achieves a 95% accuracy in recognizing both letters and numbers.**

***Index Terms*—Earable, Face and Ear Interaction, Gestures Recognition, Acoustic Sensing**

I. Introduction

The new generation of in-ear wearables (earables) integrates various sensors in order to provide a better user experience and support a wide range of emerging applications. A total of 34.6% of all wearables shipped were earable devices, which saw the largest growth of 135.1% year over 2022 [1]. A number of research efforts have also been made to leverage earables to provide a better user experience, for instance, AR and VR [2], [3], healthcare monitoring [4], [5], user authentication [6], motion and activity tracking [7]–[9].

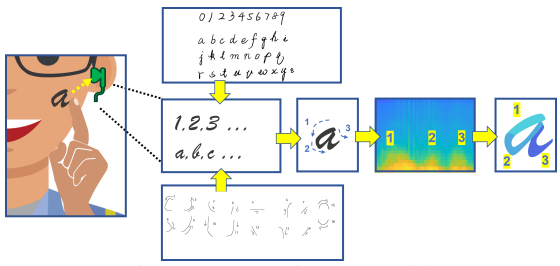


Fig. 1: FaceTyping's core idea.

This work was partially supported by the NSF Grants CNS-2131143.
* These authors contributed equally to this work. † Corresponding Author.

Traditional earables often rely on buttons or touch sensors for user interaction, offering a limited predefined input gestures like single or double touches/presses and sliding. Although Apple Airpods and Samsung Galaxy Buds incorporate these features, the small interaction area and unintentional touch remain issues [3], [10]. Moreover, these sensors occupy crucial hardware space, supporting only a limited number of gestures.

Voice control is another common interaction method in earables, enabling alphanumeric input through a broader range of voice commands. However, speaking in public may be deemed intrusive or inconvenient, particularly in quiet or formal environments such as libraries or banquet. Despite the popularity of voice input, the exploration of alternative interaction methods holds potential value.

A few recent studies have explored earable face-touching interaction methods. For example, using the face as an input pad for touch gestures such as single touch and zoom-in with two fingers [11]. However, similar to buttons and touch-based interaction techniques, these existing solutions are still limited to a small number of predefined gestures. This paper introduces FaceTyping, an interaction system that enables alphanumeric handwriting on the face using passive acoustic sensing in earables. As shown in Fig. 1, the fundamental understanding behind our system is the exploration of domain knowledge regarding the conversion of 2D handwriting patterns into 1D acoustic sequences, improving feature recognition process in our system. Compared to existing methods, FaceTyping offers better security, privacy, and less obtrusion, presenting itself as a viable supplementary interaction technique in specific scenarios.

In particular, by analyzing the impact of acoustic friction and the transformation of 2D handwriting into 1D acoustic signals on earables-based finger-face interfaces and studying the domain knowledge, we found four major factors affecting finger-face handwriting, which are ***the distinct shape of the character***, ***the directional and length confusion***, ***strokes in the character***, as well as ***coupling and decoupling grounding conditions***. Firstly, these insights have driven the redesign of certain characters to ensure clarity and distinction in the 1D acoustic domain and we propose a stroke analysis and redesign to introduce more distinct features. Secondly, the domain

knowledge helped our system in locating specific features that can be representative. Leveraging the domain knowledge of 2D to 1D information conversion, we designed a multimodel machine learning network to better extract features.

In order to evaluate the real-world performance of our system, we conducted an evaluation with 10 participants. The results of the evaluation showed that the FaceTyping system was able to recognize 10 numbers and 26 alphabets with an average accuracy of 96.82% and 95.54% for numbers and characters with only 20 training samples per class, respectively. The main contributions of this paper can be summarized as follows:

- We analyzed the transformation of 2D spatial handwriting into 1D acoustic time sequence information. These insights guide our approach to redesigning the 2D handwriting strokes for less distinguishable characters and more effectively capturing the most representative features from 1D acoustic information.
- We conducted a stroke analysis and character redesign of the handwritten alphabets and numbers to introduce greater distinctiveness for identification while maintaining the user's natural writing habits for better performance. Additionally, we developed multimodal architecture networks to optimally exploit the information provided by both 2D and 1D domains.
- We developed FaceTyping, an alphanumeric interaction system designed for earables. In certain contexts, our system could offers advantages over traditional solutions, including greater security, privacy, and less obtrusiveness. Unlike previous face-touching systems, our approach supports alphanumeric input, which can provide increased efficiency and may serve as a desirable alternative interaction system in situations that require human-earable interaction.
- We conducted a usability study, and evaluated our system in real-world settings. The results of the study indicate that FaceTyping can achieve consistently high accuracy across different usage scenarios and diverse input behaviors exhibited by the participants.

The rest of this paper is organized as follows. Section II proposes the related works of FaceTyping. Section III presents the preliminary study on acoustic friction and the conversion of 2D spatial handwriting to acoustic 1D time sequence. Section IV introduces the core idea, system flow, detailed design, and implementation of our system. And we present the performance evaluation and experimental results in Section V. Lastly, we conclude our work in Section VI.

II. Related Works

A. Earable Interaction and Sensing

Due to their unique sensing position, advanced sensing capability, and computing capabilities, earables are gaining research interest as a new type of ubiquitous computing platform [12]. One of the major categories is earables-based interaction. Touch sensors are built into some commercial products for interfacing, but these sensors occupy additional hardware space and sometimes suffer from limited interaction areas and unintended touch activation [3], [10]. Another popular category is voice-based controls. For example, Apple AirPods enable users to control their iPhones or iPads with Siri voice assistance [13], [14]. However, that voice-based interaction will become obtrusive in some public scenarios. There are also many studies focusing on gesture-based interaction methods. such as sensing hand movements as input. For instance, mid-ear hand or finger gesture recognition [15], [16], and facial movement [17]. Relevant to our work, on-face finger gestures have been explored in a few previous research projects. However, the existing papers only recognize a few gestures on the face [11]. By supporting alphanumeric input, our system extend the potential of interactions with earables.

B. Acoustics-Based Activity Recognition

Researchers have been analyzing audio signals to determine physical activity and the environment in order to recognize the human physical activity. For instance, Chen et al. [18] tracked finger movements on a wooden tabletop using acoustic signals. There are also studies that utilize smartphone, smart watch and smart glasses to actively track finger drawing on the table or in the air [19], [20]. For the sound-based activities recognition tasks, various classification models have been used, from traditional methods like support vector machines and hidden Markov models [21], [22] to advanced learning techniques such as neural networks [18], [23], [24]. Amoung various traditional feature extraction methods, MFCCs are particularly popular for audio analysis since they distribute spectrogram energy in accordance with human hearing. A proposed system was able to recognize 22 human activities within bathrooms and kitchens by employing a non-Markovian ensemble voting strategy based on MFCC features [25]. More recent studies used pre-trained neural networks and MFCCs to identify 38 environmental events [26], [27].

Acoustic-based activity recognition systems utilizing earables have also been developed. Prior research in this area has involved classifying acoustic signals, such as coughing and chewing recorded by microphones positioned directly on the throat [28], [29]. Closer to our work, EarBuddy [11] presented a real-time system that detected tapping and sliding gestures near the ears using commercial earables' microphones and recognizing facial and ear gestures with deep learning models. However, this approach is limited to recognizing only predefined and coarse-grained gestures, which are inadequate for complex interactive tasks such as text entry.

III. Preliminary

A. Acoustics of friction

Friction acoustic is generated when two sliding surfaces move in relation to each other, converting kinetic energy into thermal energy and scattering some of the energy as sound waves [30]. In our system, friction acoustics are generated by the sliding of the fingertip and face surfaces, producing

intricate sounds. According to our literature study, the characteristics of these friction acoustics are influenced by mainly four factors, the distinct shape of the character, the directional and length confusion, its strokes, as well as coupled and decoupled grounding conditions.

1) Distinct shape of the character: The most straightforward impact of the sound generated on the face when the finger is handwriting is the distinct shape of the character. The distinct shape of the characters refers to the unique silhouette or outline that sets them apart from each other. This is because the distinct shape, i.e., geometric information, of characters affects the resulted 1D acoustic information. The differences in distinct shape lead to differences in frequency distribution and noise level of friction acoustics. Our study found that while the distinctiveness of shapes allows us to distinguish many characters based on both 2D handwriting and 1D acoustic information. For those characters with only minor differences in their distinct shape, it also leads to confusion in 2D handwriting. In order to address this issue, we considered several additional factors affecting the acoustics of friction in our system design.

2) Directional and length confusion: During the conversion of 2D handwriting into 1D acoustic information, the degradation of critical features such as directional information and stroke length occurs will result in difficulties in character differentiation. For example, the letters 'm' and 'w' can be easily distinguished in 2D handwriting but become indistinguishable in the 1D acoustic representation due to the degradation of directional information. Similarly, letters such as 'h' and 'n' which can be differentiated by their height in 2D images become difficult to distinguish in the 1D acoustic representation as a result of the degradation of length information. To address these challenges, it is necessary to consider these information degradation in the design of our system to minimize the impact of directional and length confusion. These issues will be further discussed in the Section of System Design of the paper.

3) Strokes in the character: Characters are composed with strokes. The strokes are different from character to character and generate distinct friction acoustics as well. Strokes can assist in distinguishing characters that are unclear in 2D handwriting by taking advantage of the benefits of 1D acoustics information. For example, the characters '1' and 'l' may look similar in 2D handwriting, but they have distinct strokes. The number '1' has only one vertical stroke, while the letter 'l' has a vertical stroke and a small tail as a second stroke. We observed that even in the most sloppy way of writing 'l', people will still tend to keep the small tail. The presence of different strokes contributes to the acoustic spectrum by creating additional peaks and contributing to the components in the frequency domains of the friction signals. Furthermore, the gaps between the strokes create blank spaces in the acoustic spectrum. And the strokes order also determine the order of peaks in spectrum.

4) Coupling and decoupling grounding conditions: One of the major impactors of acoustic friction is coupling and decoupling grounding conditions between the finger and face, which affect friction and the generated sound. Depending on the coupling and decoupling grounding conditions at the interface where the fingertip contacts the face, the sounds generated by a particular face and finger pair differ largely. In general, as a result of weak grounding, or decoupling, the finger and face respond at their own natural frequency, almost independently of each other. Strong grounding, or coupling, occurs when the finger and face are in strong contact conditions. The grounding force can cause the finger and face to become a coupled system and create a more complicated and often nonlinear response. For instance, for the pair (0, o), in consistent hand-writing, the larger the radius of trajectory, then the lower drawing speed, which makes the grounding condition more coupling, thus the number '0' has more straight edges on left and right, which is faster and more decoupling than the letter 'o'. The number '0' tends to scatter more energy at low frequencies and produces less friction noise.

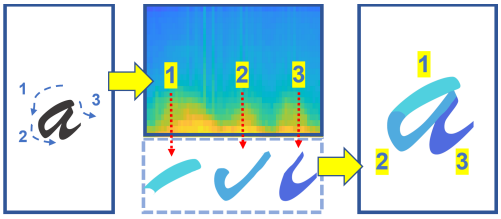


Fig. 2: 2D geometric handwriting to 1D acoustic time sequence information.

B. 2D Handwriting to 1D Acoustic Time-Spatial Information

Previous face-touching interaction system didn't study the conversion from 2D spacial handwriting to 1D acoustic sequence, limiting the recognition ability of systems. As a result, the system was able to support only limited number of predefined gestures.

In this study, we study the conversion from 2D handwriting into a 1D sequence of acoustic friction to differentiate a large number of inputs. The 1D sequence of acoustic friction contains the acoustic connections, time sequence, and frequency information of each handwritten character, as depicted in Fig. 2.

As we have previously discussed in Section III.A, the conversion of 2D handwriting into 1D acoustic information results in the degraded geometric information of relationships between strokes, including directional information and stroke length. This degradation of information can lead to confusion between characters that were distinct in 2D. As seen in top 4 rows of Fig. 3, the category of "Distinct in 2D but Confused in 1D" includes pairs of characters such as (x, t) and (w, m), where confusion arises from the degradation of directional information during the transformation from 2D to 1D. Meanwhile, the pair of (2, z) is confused due to the fact that they merely have minor differences in distinct shape. Pair of (h, n) is also confused in 1D domain due to the degradation of stroke length information. In order to improve the discrimination of character pairs in handwriting recognition, it

is necessary to address the information degradation that occurs when transforming a 2D visual representation of handwriting to a 1D acoustic sequence. As a solution, we considered to incorporate additional features into the acoustic sequences.

Additionally, the acoustic friction introduced in Sections III.A.2 and III.A.3 allows us to distinguish some characters that were previously confused in 2D, but become distinct in 1D acoustic information, as shown in the second category of Fig. 3. As a result, the pair (1, l) are sometimes confused in 2D writting but become distinct because the tail of the letter 'l' generates an additional peak in the acoustic spectrum. Furthermore, the decoupling effect in the 1D acoustic representation of the character pair of (0, o) results in a weaker acoustic friction for the number '0' compared to the letter 'o'. This unique 1D acoustic feature enables differentiation between these characters based solely on 2D information. Moreover, there are other categories of letters are both undistinguishable in 2D handwriting and 1D acoustic domain due to the face they have almost the same distinct shape, as illustrated in the third category of Fig. 3. For instance, the letter pairs of 'u' and 'v' which are very similar in handwriting, given they are usually scrawled, thus are difficult to distinguish in both 2D and 1D domain.

The conversion of 2D handwriting into 1D acoustic information presents significant challenges to our recognition system. The first challenge is compensating for the degradation of directional and length information during the transformation by incorporating additional features. The second challenge is how to effectively utilize the unique 1D acoustic features to improve recognition accuracy. To address these challenges, guided by the domain knowledge of 2D to 1D information conversion, we propose a redesign of the writing strokes of those less indistinguishable characters to incorporate more features while maintaining their writing style natural to users, which is detailed in Section IV.B. Also, we leverage the friction domain knowledge to locate most representative features, incorporated with the MFCC spectrogram, to better utilize the time-frequency domain information, and our system employs a combination of DenseNet and GRU networks to effectively capture both the 1D time-spatial and frequency information, as well as the 2D geometry information that remains preserved, which is described in Section IV.D.

IV. System Design

A. System Overview

Our system's core concept leverages an in-depth understanding of coverting 2D handwriting into 1D acoustic signals. Through exploring this knowledge, we come up with several principles and redesigned less distinguishable characters for more accurate feature identification. Taking the advantages, our system can recognize 26 letters and 10 numbers written on the face with a fingertip. It offers an alternative to voice control, enhacing security and privacy in settings where traditional methods might be obtrusive, such as during a movie.

As illustrated in Fig. 4, the system captures the finger-face friction acoustics via earables' inward-facing microphones, utilizing occlusion effect from face-bone conducted signals. To improve handwriting character distinctiveness, we analyze and suggest redesigns for handwriting patterns, enhancing their recognizability. The system transforms 2D geometric handwriting into 1D acoustic data. We developed a multimodal model to capture both 2D geometric and 1D acoustic features effectively. As a user writes on their face, the generated friction acoustic signals are transmitted through the face-ear channel, captured by the microphone, converting some 2D information into 1D, including character shapes and stroke order. While directional and length information being degraded, this conversion impacts the spectrogram's overall envelope and the spectrum peak order. We leverage remaining 2D shape and stroke features, along with 1D acoustic features, for character identification. Our multimodal model combines DenseNet, for 2D shape and stroke feature, with GRU, for sequential 1D acoustic features extraction. Utilizing transfer learning addresses user behavior variability, allowing for effective character classification with less user-specific data enrollment.

B. Redesigning principles

Leveraging the knowledge of transforming 2D writing patterns into 1D acoustics, we identified three key factors influencing friction acoustics. Correspondingly, we came up with three redesign principles: adding extra stroke component, adjusting the stroke order and emphasizing specific stroke, as shown in the Fig. 5.

- **Adding extra stroke component.** Adding extra stroke as a starting or tail contributes to the unique to the shape of characters. In 1D acoustics, it generates a short and sharp peak in the corresponding part in the signal to enhance the distinct shape and strokes in the character.
- **Adjusting the order of strokes.** The different strokes can be more similar in 1D acoustics due to the degradation of directional and length information. Such as letter 'x' and 't' shown in Fig. 7. We come up with an adjusted order of strokes for such characters' handwritting which makes significant difference on 1D acoustics, reintroducing the directional information while keeping them within user writing habits. During our experiments, participants get used to the change in a few tries.
- **Emphasizing the key stroke.** The acoustics of some characters such as 'f' and 's' are easier to affected by coupled condition due to the multiple curves within their strokes. We emphasize the key stroke in the redesigning to make their coupled condition more consistent thus more features remain in the 1D acoustics consistently and identifiable.

C. Strokes Analysis and Handwriting Redesign

1) Alphabets categories and strokes analysis: Transforming 2D handwriting into 1D acoustics introduces challenges for handwriting recognition accuracy. Our preliminary study indicates that this conversion can degrade certain information, like stroke direction and length, but also introduces new data through 1D time-spatial characteristics. Consequently, while

Categories	Pairs	Shape Related	Stroke Related		Coupling and Decoupling Related		Information Lost During 2D to 1D	
		Shape	Extra stroke generate peak	Gap between stroke generate blank	Drawing Speed	Stroke Radius	Directional Confusion	Length Confusion
Distinct in 2D but confused in 1D	(x,t)						1D Confused	
	(m,w)		Redesign				1D Confused	
	(h,n)	1D Confused	Redesign					1D Confused
	(2,z)	1D Confused			Redesign	Redesign		
Confused in 2D but distinct in 1D	(0,o)	2D Confused			1D Distinct	1D Distinct		
	(6,b)	2D Confused			1D Distinct	1D Distinct		
	(1,l)	2D Confused	1D Distinct					
Both confused	(9,q)	Both Confused	Redesign					
	(u,v)	Both Confused	Redesign					
Both distinct	Others	Both Distinct	Both Distinct	Both Distinct				

Fig. 3: Different categories of handwriting pairs, based on the analysis of the transformation from 2D special handwriting to 1D acoustic time sequence, we categorize the 26 alphabets and 10 numbers into 4 groups, i.e., 1. ***Distinct in 2D but Confused in 1D***, 2. ***Confused in 2D but Distinct in 1D***, 3. ***Both Confused***, and 4. ***Both Distinct***.

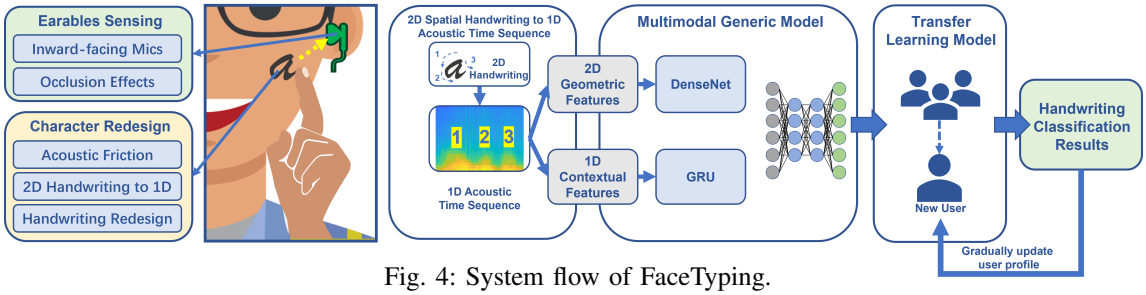


Fig. 4: System flow of FaceTyping.

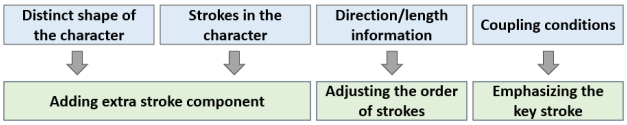


Fig. 5: The key redesign methods guided by domain knowledge of conversion from 2D writing pattern to 1D acoustics.

some characters are easily identifiable in 2D, they may become less distinguishable by 1D acoustics, and vice versa.

To address these challenges, we study the character pairs that cause confusion and divide them into four groups, namely ***Distinct in 2D but Confused in 1D***, ***Confused in 2D but Distinct in 1D***, ***Both Confused***, and ***Both Distinct***. In order to improve the accuracy of recognition for the categories ***Distinct in 2D but Confused in 1D*** and ***Both Confused***, we utilize the three methods to redesign the characters, enhancing the features of characters. Meanwhile, we take advantage of the 1D acoustic information for characters in the category ***Confused in 2D but Distinct in 1D***, as illustrated in Fig. 6.

2) Categories of Handwriting and Redesign: To enhance system performance, we redesigned certain handwriting characters to improve distinguishability, incorporating richer features while permitting user-preferred writing styles.

Specifically, the category ***Confused in 2D but Distinct in 1D*** encompasses letter pairs such as (0, o), (6, b), and (1, l), as demonstrated in Fig. 8, which has clearer differentiation in 1D acoustics. We explored why their 1D acoustics stand out compared to 2D handwritings, guiding our redesign. In pair (0, o), the difference on drawing speed and radius of stroke resulted in differences in the high frequency component of the spectrogram, as illustrated in Fig. 8a. The number '0' has straight edges on both the left and right, and is drawn at a faster speed compared to the letter 'o', which results in more energy being scattered at low-frequency and less friction noise. Similarly, in pair (6, b), the different connecting part between strokes leaded the difference. The number '6' is written in a single continuous stroke, while the letter 'b' requires the writer to slowdown at the end of the first vertical stroke, change direction, and then complete the character. This difference results in a slower drawing speed and generates greater coupling friction, leading to increased friction and noise, as shown in Fig. 8b. In pair (1, l), a notable difference can be seen in the spectrogram of the letter 'l' due to the shape and trajectory of the drawing generates an extra stroke, i.e., the tail, which creates an additional component in the 1D time-spatial spectrogram and distinguishes these two characters, as shown in Fig. 8c.

The category ***Distinct in 2D but Confused in 1D*** includes letter pairs such as (x, t), (m, w), (h, n), and (2, z), where distinction becomes less significant from 2D handwriting to 1D acoustics, as illustrated in Fig. 9. To adress this, we adjusted stroke orders for (x, t). As shown in Fig. 7, the letter 't' was similar to the letter 'x' when rotated. After the stroke adjustment, the degradation of directional information would not confuse those letters any more, as shown in Fig. 9a. Similarly, for (m, w), an extra stroke on 'm' differentiate its acoustic profile, as shown in Fig. 9b, an extra peak shows the additional information. For (h, n), emphasizing the first stroke of h and adding a swash to 'n' creates distinct peaks in their spectrogram, as shown in Fig. 9c. These minor 2D changes significantly impact 1D acoustic distinction.

The ***Both Confused*** category includes pairs like (9, q) and (u, v), which are challenging to differentiate in both 2D handwriting and 1D acoustics. To improve recognition, we

Categories	Pairs	2D Handwriting to 1D Acoustic Transformation		Handwriting Stroke Redesign
Distinct in 2D but confused in 1D	(x,t)	1D Confused:	Directional Confusion	Different stroke order for 't'
	(m,w)	1D Confused:	Directional Confusion	Extra stroke added to 'm'
	(h,n)	1D Confused:	Length Confusion	Extra stroke added to 'n'
	(2,z)	1D Confused:	Shape	Emphasis stroke radius of '2'
Confused in 2D but distinct in 1D	(0,o)	1D Distinct:	Coupling and Decoupling	No Redesign Needed
	(6,b)	1D Distinct:	Coupling and Decoupling	No Redesign Needed
	(1,l)	1D Distinct:	Extra stroke	No Redesign Needed
Both confused	(9,q)	Both Confused:	Shape	Extra stroke added to 'q'
	(u,v)	Both Confused:	Shape	Extra stroke added to 'u'
Both distinct	Others	Both Distinct		No Redesign Needed

Fig. 6: Handwriting redesign.

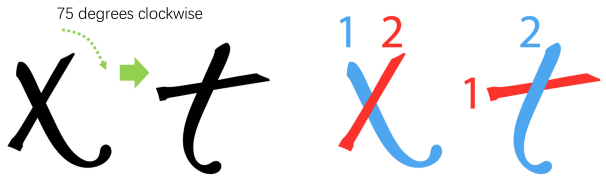


Fig. 7: Strokes analysis example: x and t.

added a tail to 'u', introducing an extra spectral component that enhances distinguishability, as demonstrated in Fig. 10, which shows an extra peak after redesign.

Finally, the last category, referred to as ***Both Distinct***, encompasses characters that are easily distinguishable in both the 2D handwriting and 1D acoustic domains, most characters are originally in this category.

Besides the redesign following the principles above, we have developed a switching method for different keyboards detailed in the subsequent section that assigns numbers and letters to distinct keyboards and eliminates the need for further redesigns, such as the case of (2, z) and (9, q). We utilize gestures from orthogonal domains as control signals for the keyboard switching with an extremely high level of accuracy close to 100%. The user can accurately use these gestures switch between keyboards. We used two-finger taps, palm touches on the face, and three-finger sliding as the switching gestures.

D. Multimodal Generic Model Design

Fingertips drawing on faces generates two pieces of information in both 2D and 1D domain. As we discussed, the remaining 2D handwriting including the distinct shape of each character and the strokes order within the character. The 1D domain features including the time sequence information, the mel-specturm and its frequency component. . To make the most of both information and extract the right features from each domain, we have designed a multimodal generic model.

In particular, the remaining 2D handwriting features are extracted by a DenseNet model. It captures features from the spectrogram of handwriting's geometry information that remains such as distinct shape and stroke order. Meanwhile, the domain knowledge of acoustic frictions also helps us on locating the most representative features of 1D acoustics of characters. We use the domain knowledge of the four main factors affecting the acoustics of friction as a guide to choose the features that can best represent the characteristics of characters. In particular, we consider spectral flux to reflect the transitions between the acoustic signal, representing the switching between the strokes, which could further help identify the distinct shape. Also, we utilize spectral spread and spectral skewness to provide information about the timbre related to the state of the finger-face contact surface, as a representation of coupled conditions. Moreover, we use spectral roll-off to capture sudden changes in the signal, which can be the symbol of decoupling effects at the turning point inside the writing characters. The 1D acoustic friction time sequence features, on the other hand, are extracted using a GRU (Gated Recurrent Unit) model. In our work, GRU module is utilized to analyze the time series of acoustic friction signals generated by the friction of each stroke during handwriting. It takes the acoustic friction sequence as input and processes it through recurrent layers, and produces a compact representation of the friction patterns in handwriting.

As shown in the Fig. 11, we embedded two GRU layers between the Dense blocks and FC layers to extract and leverage both the 2D geometry and 1D time sequence information. Our analysis shows that by incorporating both modalities, our multimodal generic model is able to improve the overall accuracy of handwriting identification compared to using only DenseNet or GRU.

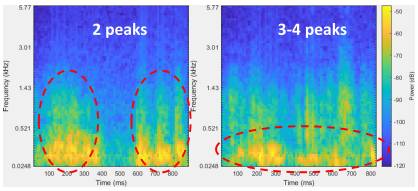
E. User Behaviour Mitigation

The variability in user behavior, particularly with complex gestures such as handwriting, significantly affects classification accuracy. To address this challenge, we introduced transfer learning method to mitigate the accuracy drop. During the enrolling process, all parameters in the global model are fixed as a feature extractor, except for the closest fully connected layer to the output end. The model is then retrained using only the data from the newly added user.

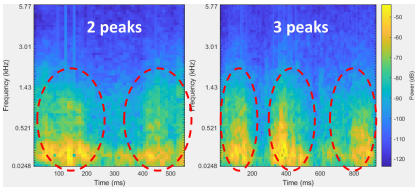
V. Evaluation

A. Experiment Setup

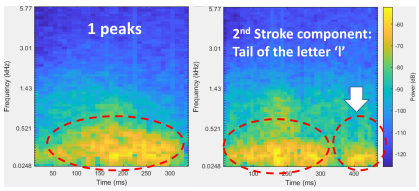
Environment Our system is applicable in a variety of everyday scenarios, including situations where people interact with earables or use them to interact with other devices. During our experiments, we evaluated the performance of the system



(a) 0 and o

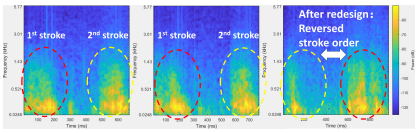


(b) 6 and b

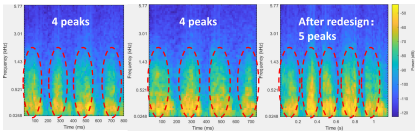


(c) 1 and l

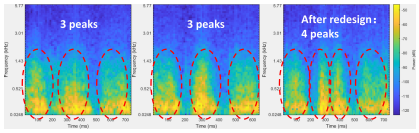
Fig. 8: Examples letter pairs in category ***Confused in 2D but Distinct in 1D***.



(a) x, t and redesign of t

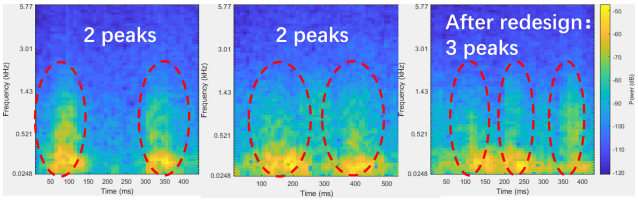


(b) m, w and redesign of m



(c) h, n and redesign of n

Fig. 9: Examples letter pairs in category ***Distinct in 2D but Confused in 1D***.



(a) v, u and redesign of u

Fig. 10: Examples letter pairs in category ***Both Confused***.

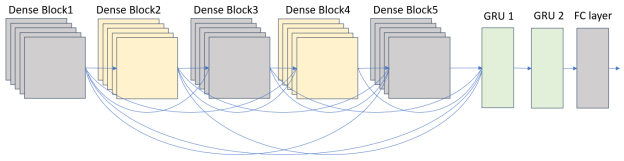


Fig. 11: Multimodal generic model design of the proposed system.

in four distinct environmental settings with varying levels of ambient noise. These settings included the office, living room, car, and street environments, as illustrated in Fig. 12 and described in detail below.

Participants and Data Collection. We conducted an experiment to assess the performance of our finger-face handwriting input system. We recruited 10 participants with ages ranging from 23 to 35 years. All participants were right-handed. Each participant was instructed to wear the prototype at their habitual position and asked to perform a complete set of numbers (0-9) and letters (a-z) using finger-face handwriting.

Hardware. Due to firmware restrictions that limit access to raw data, instead of existing commercial earables, we developed a low-cost earables prototype using readily available hardware components, as illustrated in Fig. 12b. Our system incorporates a single microphone chip with sensitivity of -28 dB and a 3.5 mm audio jack. The cost of the equipment used in our prototype is less than 10 dollars, making it accessible to a broader customer base compared to commercial products. The sampling rate used was 192kHz. The evaluation experiments were performed on a computer system equipped with a GPU RTX 3090, CPU i7-11700KF, and 32GB of RAM.

B. Overall Performance

We first evaluate the overall performance of our system. We trained our generic model with 9 users and evaluate the generic model without keyboard-switching mechanism on testing data from the same users, the average accuracy of our generic model reached 96.18%. We then implement transfer learning on the remaining user with only taking a few samples. As shown in Fig. 13, results also showed that with a relatively small training sample of only 8 samples per class, the accuracy of both numbers and letters was over 90%. Furthermore, when the number of training samples is 20, the average accuracy of the number identification 96.82%. And the accuracy of letters identification is 95.54%. These results demonstrate the high effectiveness of the system in classifying different characters with minimal training data.

We also conducted an experiment to evaluate the performance of our system and compare it with previous face-touching apporaches without the use of a multimodal generic model and character redesign as employed in our work. The results of the experiment are presented in Fig. 14, which includes the confused pairs disscussed above. Our analysis revealed that after the integration of our domain knowledge, the error rate for each of these character pairs significantly decreased, leading to a substantial improvement in the recognition accuracy of our handwriting system.

C. Usability Study

The usability study of our system was evaluated with 20 participants. After interacting with the system, the participants were asked to fill out a questionnaire to rate the design of each character on a seven-point scale, from very comfortable to very uncomfortable. The results of this evaluation are presented and

(a) Experiment environments: living room, car, office, and street.

(b) Prototype and wearing.

Fig. 12: Experiment environment and prototype.

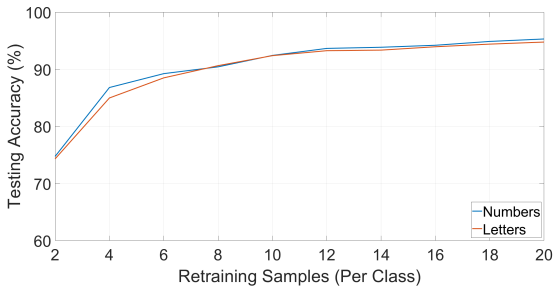


Fig. 13: Overall performance.

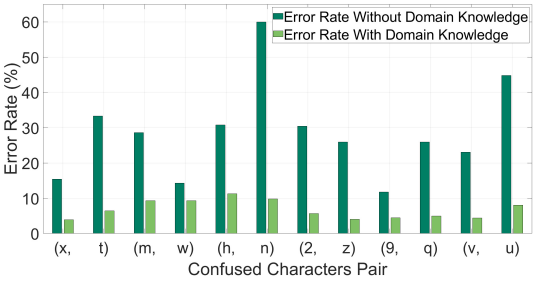


Fig. 14: Error rate compared with traditional approaches.

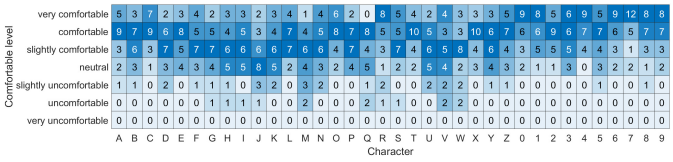


Fig. 15: Usability study.

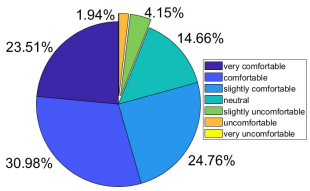


Fig. 16: Distribution of user comfort level.

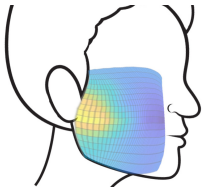


Fig. 17: Impacts of different touching areas.

indicate the overall comfort level of the system as perceived by the participants.

As shown in Fig. 15, the redesigned character pattern received positive feedback from the 20 participants who voted on all the letters and numbers. On average, as shown in Fig. 16, 93.91% of the votes were above neutral. Among them, about 85% of the votes were rated as "very comfortable" or "comfortable", demonstrating that the interaction method in the system is well received in terms of user experience.

For characters that have been redesigned, the ratings slightly decreased, however, over 85% of the votes were still positive or neutral, suggesting that the redesign an acceptable trade-off between user experience and system performance.

D. Robustness Analysis

1) Impacts of Body Movements: The system's performance was tested during various movements, including head rotation and walking in a living room environment. We sampled 1440 instances for each of the body movement. As shown in Fig. 18, head rotation has little impact with the average accuracy 94.69%. Meanwhile, walking has a slight impact on the average accuracy as 90.53%.

2) Impacts of Different Touching Areas: We also carried out experiments to evaluate the sound intensity produced by friction on various facial regions. We analyzed the results to construct a heat map of the optimal interaction positions on the face. The yellow areas indicate strong recorded handwriting sound, while the blue areas indicate weaker sound. As shown in Fig. 17, our results reveal that the most desirable location for finger-face handwriting is near the tragus of the ear.

3) Impacts of Different Background Environment: We assessed our system's performance across four real-life background environments: living room, car with its engine on, office, and a street, to simulate potential usage scenarios. These settings were chosen to evaluate the system's handwriting identification accuracy under various ambient noises. As shown in Fig. 19, the accuracy in these environments were 93.47%, 91.60%, 92.22%, and 88.68%, respectively. The result shows the system's high accuracy and robustness in recognizing handwriting under diverse environmental sounds.

VI. Conclusion

In this paper, we propose a face-handwriting interaction system that addresses the need for seamless user interfaces on earable devices. We studied domain knowledge of acoustic friction and conversion from 2D to 1D domains. On the basis of this domain knowledge, we proposed FaceTyping, a finger-face acoustic friction system that enables alphanumeric text input on earables without additional devices and has the potential to act as complementary input in the scenario

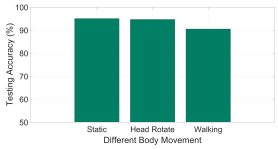


Fig. 18: Impacts of body movements.

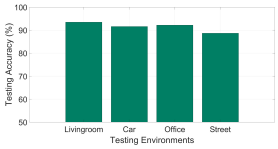


Fig. 19: Impacts of different background environment.

where traditional input methods become less feasible. Guided by domain knowledge, we redesigned the characters' handwriting as recommended writing patterns to incorporate more geometric features. This was done without changing users' natural writing habits. This domain knowledge of 2D to 1D also helped us locate and select the most representative special features. We developed a multimodal generic model using DenseNet and GRU to locate, capture, and leverage the most representative features for handwriting identification. We also incorporated transfer learning to mitigate user behavior variances and gradually updated user profiles over time. Our experiments involved 10 participants in real-world settings, and FaceTyping achieved high accuracy in identifying both numbers and characters.

REFERENCES

[1] M. NEEDHAM, "Wearable devices market share," Sep 2022. [Online]. Available: https://www.idc.com/promo/wearablevendor

[2] Z. Yang, Y.-L. Wei, S. Shen, and R. R. Choudhury, "Ear-ar: indoor acoustic augmented reality on earphones," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020, pp. 1–14.

[3] R. R. Choudhury, "Earable computing: A new area to think about," in *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications*, 2021, pp. 147–153.

[4] N. Pham, T. Dinh, Z. Raghebi, T. Kim, N. Bui, P. Nguyen, H. Truong, F. Banaei-Kashani, A. Halbower, T. Dinh *et al.*, "Wake: a behind-the-ear wearable system for microsleep detection," in *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, 2020, pp. 404–418.

[5] N. Bui, N. Pham, J. J. Barnitz, Z. Zou, P. Nguyen, H. Truong, T. Kim, N. Farrow, A. Nguyen, J. Xiao *et al.*, "ebp: A wearable system for frequent and comfortable blood pressure monitoring from user's ear," in *The 25th annual international conference on mobile computing and networking*, 2019, pp. 1–17.

[6] Z. Wang, S. Tan, L. Zhang, Y. Ren, Z. Wang, and J. Yang, "Eardynamic: An ear canal deformation based continuous user authentication using in-ear wearables," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–27, 2021.

[7] D. Kim, C. Min, and S. Kang, "Towards automatic recognition of perceived level of understanding on online lectures using earables," in *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*, 2021, pp. 158–164.

[8] J. Prakash, Z. Yang, Y.-L. Wei, and R. R. Choudhury, "Stear: Robust step counting from earables," in *Proceedings of the 1st International Workshop on Earable Computing*, 2019, pp. 36–41.

[9] C. Min, A. Montanari, A. Mathur, S. Lee, and F. Kawsar, "Cross-modal approach for conversational well-being monitoring with multi-sensory earables," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 2018, pp. 706–709.

[10] F. Kawsar, C. Min, A. Mathur, A. Montanari, U. G. Acer, and M. Van den Broeck, "esense: Open earable platform for human sensing," in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, 2018, pp. 371–372.

[11] X. Xu, H. Shi, X. Yi, W. Liu, Y. Yan, Y. Shi, A. Mariakakis, J. Mankoff, and A. K. Dey, "Earbuddy: Enabling on-face interaction via wireless earbuds," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[12] T. Röddiger, C. Clarke, P. Breitling, T. Schneegans, H. Zhao, H. Gellersen, and M. Beigl, "Sensing with earables: A systematic literature review and taxonomy of phenomena," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–57, 2022.

[13] A. Zeidan, H. T. Abdelgelil, E. Edwin, and D. Alqarni, "Apple siri as communication conduit during covid-19: between inside and outside the or," *BMJ Simulation & Technology Enhanced Learning*, vol. 7, no. 4, pp. 274–275, 2021.

[14] Apple. (2022) Airpods. https://www.apple.com/airpods/.

[15] Y.-C. Chen, C.-Y. Liao, S.-w. Hsu, D.-Y. Huang, and B.-Y. Chen, "Exploring user defined gestures for ear-based interactions," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. ISS, pp. 1–20, 2020.

[16] C. Metzger, M. Anderson, and T. Starner, "Freedigiter: A contact-free device for gesture control," in *Eighth International Symposium on Wearable Computers*, vol. 1. IEEE, 2004, pp. 18–21.

[17] K. Li, R. Zhang, B. Liang, F. Guimbretière, and C. Zhang, "Eario: A low-power acoustic sensing earable for continuously tracking detailed facial movements," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 2, pp. 1–24, 2022.

[18] M. Chen, P. Yang, J. Xiong, M. Zhang, Y. Lee, C. Xiang, and C. Tian, "Your table can be an input panel: Acoustic-based device-free interaction recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 1, pp. 1–21, 2019.

[19] Y. Zou, Z. Xiao, S. Hong, Z. Guo, and K. Wu, "Echowrite 2.0: A lightweight zero-shot text-entry system based on acoustics," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 6, pp. 1313–1326, 2022.

[20] K. Wu, Q. Yang, B. Yuan, Y. Zou, R. Ruby, and M. Li, "Echowrite: An acoustic-based finger input system without training," *IEEE Transactions on Mobile Computing*, vol. 20, no. 5, pp. 1789–1803, 2020.

[21] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Reliable detection of audio events in highly noisy environments," *Pattern Recognition Letters*, vol. 65, pp. 22–28, 2015.

[22] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, 2005.

[23] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[24] N. D. Lane, P. Georgiev, and L. Qendro, "Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning," in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 2015, pp. 283–294.

[25] J. A. Stork, L. Spinello, J. Silva, and K. O. Arras, "Audio-based human activity recognition using non-markovian ensemble voting," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2012, pp. 509–514.

[26] G. Laput, K. Ahuja, M. Goel, and C. Harrison, "Ubicoustics: Plug-and-play acoustic activity recognition," in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, 2018, pp. 213–224.

[27] G. Laput, Y. Zhang, and C. Harrison, "Synthetic sensors: Towards general-purpose sensing," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 3986–3999.

[28] K. Yatani and K. N. Truong, "Bodyscope: a wearable acoustic sensor for activity recognition," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 2012, pp. 341–350.

[29] T. Rahman, A. T. Adams, M. Zhang, E. Cherry, B. Zhou, H. Peng, and T. Choudhury, "Bodybeat: A mobile system for sensing non-speech body sounds," in *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, 2014, pp. 2–13.

[30] A. Akay, "Acoustics of friction," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1525–1548, 2002.