HearMe: Accurate and Real-time Lip Reading based on Commercial RFID Devices

Shigeng Zhang, Zijing Ma, Kaixuan Lu, Xuan Liu, Jia Liu, Song Guo, *Fellow, IEEE*, Albert Y. Zomaya, *Fellow, IEEE*, Jian Zhang, and Jianxin Wang, *Senior Member, IEEE,*

Abstract—Lip reading can help people with speech disorders to communicate with others and provide them with a new channel to interact with the world. In this paper, we design and implement *HearMe*, an accurate and real-time lip-reading system built on commercial RFID devices. HearMe can be used to accurately recognize different words in a pre-defined vocabulary without limitations in light conditions and can be used in multiple user scenarios by leveraging RFID's ability in identifying different users. We design an effective data collection strategy to well capture the tiny and complex signal patterns caused by mouth motion and propose a set of algorithms to extract signal profiles related to mouth motions and mitigate interference factors like multi-path. A carefully designed set of features, including time-domain statistical features and frequency-domain features, are then extracted from the signal to lift the recognition accuracy at the word level. To reduce training costs when the model is used in a new environment, a transfer-learning-based approach is adopted to enhance the robustness of the model in cross-environment scenarios. Experimental results show that HearMe detects speaking actions of the user with an accuracy higher than 0.95 and recognizes different words in a 20-words vocabulary with an average accuracy higher than 0.88. Moreover, the latency of HearMe ($\sim$150ms) is nearly two orders of magnitude less than traditional approaches, making it applicable to practical scenarios that require real-time lip reading.

Index Terms—RFID sensing, wireless sensing, lip reading, real time

1 Introduction

Automatic lip reading has been studied for a long time to benefit people with difficulties in speaking and hearing [1], [2]. According to the statistics of the World Health Organization (WHO) [3], there are approximately 70 million deaf-mute people in the world. Automatic lip reading can help them enhance their ability in communicating with others and performing daily tasks. It also provides a new channel for those people to better interact with the world [1], [2]. Lip reading also benefits normal people in environments where speaking loudly is inappropriate (*e.g.*, in a meeting room) [2]. Recently, lip reading has also been exploited as a new approach to biometric-based authentication for mobile devices [4], [5].

Traditional approaches to automatic lip reading are generally based on vision analysis [6]–[8]. These approaches are sensitive to light conditions and cannot be used in dark environments. Some other works require the user to attach special sensors [9], [10], which are intrusive and inconvenient to use. Recently, non-intrusive lip reading approaches based on acoustic signals [2], [4], [5], or wireless signals [11] are proposed. The general idea is to capture signal profiles related to different mouth motions and use classification methods to recognize what words the user is *speaking*[1]. These approaches have some limitations that prevent them from being widely adopted in practice. For example, the approaches based on acoustic signals are not convenient to use in daily life as they usually require the user to hold the smartphone in a fixed position [2]. WiFi-based approaches [11] have a relatively large operational range, but they are not suitable in multiple user scenarios because it is difficult to distinguish signals from different users. Moreover, existing approaches usually use dynamic time wrapping (DTW) to match different month motion profiles, which is very time-consuming and thus not applicable to practical scenarios that require real-time services.

In this paper, we design and implement an accurate and real-time lip-reading system based on commercial radio frequency identification (RFID) devices, namely *HearMe*. HearMe recognizes what word the user tries to convey from a predefined vocabulary. HearMe is superior to existing contactless lip-reading approaches by simultaneously providing three merits. First, different from vision-based approaches [6], [7], HearMe uses wireless signals to detect mouth motions and thus is insensitive to light conditions and applicable to more pervasive environments. Second, HearMe has a relatively large operational range of up to several meters and thus is more convenient to use than acoustic-based approaches for which the operational ranges are very

- *Shigeng Zhang, Zijing Ma, Kaixuan Lu, Jian Zhang and Jianxin Wang are with the School of Computer Science and Engineering, Central South University, China. Shigeng Zhang is also with the State Key Laboratory of Information Security (Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093). E-mail: {sgzhang, mazijingcsu, kaixuanlu, jianzhang, jxwang}@csu.edu.cn.*
- *Xuan Liu is with the College of Computer Science and Electronic Engineering, Hunan University, China, 410082. E-mail: xuan_liu@hnu.edu.cn.*
- *Jia Liu is with the Department of Computer Science and Technology, Nanjing University, China. E-mail: jialiu.cs@gmail.com*
- *Song Guo is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. Email: song.guo@polyu.edu.hk.*
- *Albert Y. Zomaya is with the University of Sydney, Australia. E-mail: albert.zomaya@sydney.edu.au.*

Manuscript received January 1st, 2021; revised XX, 2021.

1. For simplicity in presentation, we call the user's action when he tries to convey a word as *speaking*, although the user does not make sound.

limited (∼10 cm). Third, the ability of RFID in identifying different objects enables HearMe to track multiple users and perform lip reading for them simultaneously. In contrast, multiple user tracking is difficult in WiFi-based approaches [11], [12] as it is difficult to map signals to individual users. Moreover, HearMe is implemented on commercial devices. It neither needs any modifications to low-layer drivers such as Wi-Fi-based approaches [11], [13] nor requires specially designed radios like frequency-modulated continuous-wave (FMCW) radar [14].

There are several challenges in implementing HearMe. First, compared to other activities such as gestures or arm activities [10], [15], [16], the mouth motions caused by the user's speaking actions are very tiny, and consequently, the signal fluctuations caused by mouth motions are also weak. How to capture the signal patterns caused by the user's "speaking" actions and correctly separate the signal segments related to mouth motions from the collected data is thus challenging. Second, how to design a set of representative features based on which different words can be accurately recognized is also a challenging problem. Third, when the environment changes, the accuracy of RFID-based activity recognition will be severely affected [15], [17]. We need to enhance the generalization ability of HearMe in different scenarios to make it suitable for daily usage.

We address these challenges with the following novel designs. First, we design an effective data collection strategy with which the tiny and complex signal patterns caused by the user's mouth motion can be well captured. Moreover, a set of algorithms are proposed to extract clear signal profiles related to mouth motions while mitigating interference factors like multi-path. Second, we combine both (coarse-grained) time-domain statistical features and (fine-grained) frequency-domain wavelet transformation features as the feature set to perform word recognition, which well characterizes the signal profile features and achieves high accuracy. Third, we utilize a transfer-learning-based approach to reduce training costs when the user switches between different environments. We implement HearMe with the commercial Impinj R420 reader and passive tags. Experimental results demonstrate that HearMe can detect mouth motions with an average accuracy higher than 0.95. In a predefined vocabulary containing 20 words, HearMe achieves an average recognition accuracy higher than 0.88 with a latency of less than 150 ms. The latency of HearMe is nearly two orders of magnitude less than traditional DTW-based approaches [11], making it applicable to practical scenarios that require real-time lip-reading services.

The rest of the paper is organized as follows. In Section 2 related works are reviewed. We design a data collection strategy that can well capture the tiny signal fluctuation patterns when the user speaks and describe it in detail in Section 3. How to preprocess the data to mitigate the impact of interference is also described in this section. In Section 4 we explain how to segment signal profiles related to mouth motions, followed by feature extraction and classification model selection in Section 5. Section 6 describes how to reduce costs in data collection when the user enters a new environment. Experimental results are reported and analyzed in Section 7. Finally, Section 8 gives the concluding remarks of the paper.

2 Related Work

2.1 Vision-based Approaches

Most lip reading approaches are based on visual analysis. Saitoh *et al.* [6] implemented a lip-reading system that can run on a laptop. The system uses the Viola-Jones algorithm for face detection, the AAM algorithm for lip detection and feature extraction [7], and finally uses the DP matching algorithm to realize mouth motion recognition. It supports multiple postures of sitting, supine, and handheld camera devices. Real-time interaction is realized by registering common sentence databases, inputting instructions to be recognized, and automatic recognition. Kumar *et al.* [8] proposed a vision-based lip-reading system and compared facial movements from a profile and a front view. Zhou *et al.* [18] propose a practical lip-reading system by detecting frames in a video. While achieving fairly high recognition accuracy, however, vision-based approaches are very sensitive to lighting conditions and cannot be used in dark environments, which greatly limits their application scenarios.

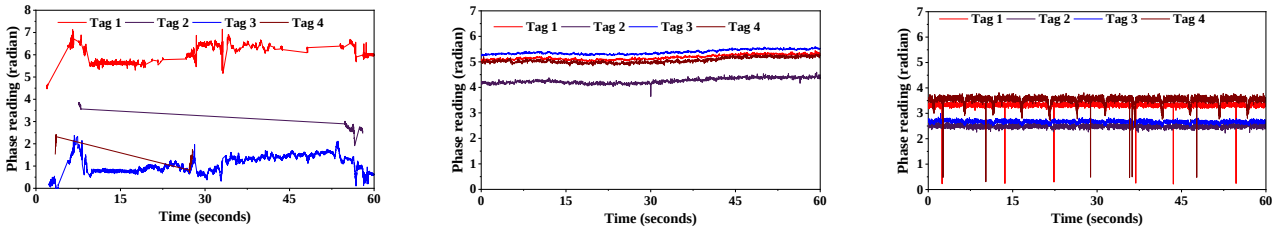
2.2 Lip-Reading based on Wireless Signals

WiHear [11] uses the MIMO technology to extract and receive reflected signals. Due to the negligible Doppler shift and amplitude fluctuation caused by the small motion of the speaking actions, WiHear uses beam-forming technology and wavelet analysis to focus and amplify the characteristics of oral motion, achieving fine-grained activity recognition of lip and tongue movements. Zhao *et al.* [19] demonstrated a Wi-Fi based approach to accurately estimate human postures through walls and occlusions. SilentTalk [2] uses a smartphone to emit ultrasonic signals and capture the Doppler shift in the reflected signals to recognize different mouth motions. Recently, lip reading based on acoustic signals has also been exploited as a new approach to biometric-based authentication for mobile devices [4], [5]. However, lip reading based on RFID is not well explored. The work in [20] leverages customized RFID devices to attach to the user's face to recognize the user's speech, which is uncomfortable and limits its commercial deployment due to customized RFID devices.

2.3 RFID-based Activity Recognition

Recently there some works on RFID-based activity recognition, but they usually recognize macro and rigid body movements such as gestures and cannot be used in lip reading. Femo [21] recognizes the user's activities during body exercise and assesses the quality of exercise movements. ShopMiner [22] and CBid [23] monitor the customers' behaviors by attaching RFID tags to goods in the supermarket and recognizing different behavior patterns by tracking the motions of tags. In [24], the authors combine Kinect-based activity recognition and RFID-based user identification to improve the quality of augmented reality applications. ID-Sense [25] enables smart interaction between the user and objects by developing activity detection systems based on RFID. Recently, deep learning is also exploited to recognize A user's body activities [26], in which the users need to attach some sensors or RFID tags. Zhang *et al.* propose a read-time RFID-based gesture recognition system [17].

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2022.3208019



(a) DA strategy (b) MT strategy (c) WM strategy

Fig. 2. Signals collected with different strategy: (a) the tags are directly attached to the skin; (b) the tags are attached to a soft mask tightly clung to the skin; (c) the tags are attached to a hard plastic mask and the user wears the mask.

Fig. 1. Framework of HearMe.

3 Data Collection and Preprocessing

The framework of HearMe is shown in Fig. 1. HearMe consists of two parts: a data collection part (the left) and a data processing part (the right). In this section, we first elaborate the motivation of using RFID to perform lip-reading, and then describe the two parts in detail.

3.1 RFID-based Lip-reading: Advantages

Compared with other wireless techniques such as WiFi and Radar, RFID is more suitable for multiple user scenarios due to the following reasons. First, by using tag ID, it is easy to distinguish signals for different users. For example, assume that there are two users in the monitoring region. For the first user, we use tags $t_1, \ldots, t_4$. For the second user, we use tags $t_5, \ldots, t_8$. After receiving the signals backscattered from the tags, we can group signals backscattered from t_1 to t_4 as the signal profile for user 1 and the signals backscattered from t_5 to t_8 as the signal profile for user 2. This can be easily done by filtering signals backscattered from tags according to tag IDs. However, for WiFi-based approaches or Radar-based approaches, separating signals for two different users might be difficult because the signals received at the receiver are a superposition of all the signals reflected by the two users. Second, RFID has a relatively large operational region and the operational region could be further enlarged by using multiple readers/antennas, which makes it more suitable for multiple users scenarios. For example, when there are two users, we can attach different sets of tags to different users and let the user sit still in their positions. When there are more users, we can enlarge the operational region by deploying multiple antennas and letting each antenna cover at most two users. Because different antennas can communicate with tags at different channels (there are 16 available channels for commercial RFID devices), the high accuracy of HearMe can be maintained. We can use RFID to achieve this goal because the RFID reader handles the signal interference problem at the hardware and driver level and thus we need not worry about signal interferences in HearMe. In contrast, although in Wi-Fi-based approaches we can also use multiple channels (at most 3 channels can exist without interference among adjacent channels in WiFi standard), all the signals reflected from different users will mix and it is very difficult to distinguish signals for a specific user. Third, RFID-based approaches are less sensitive to small movements of the user's head and inter-user interferences when the users are separated from each other, while WiFi-based approaches are more sensitive to inter-user interferences and even the user's tiny motion. As reported in previous works [11], even winking can severely impact the signal pattern when Wi-Fi signals are used to perform lip-reading. As a comparison, HearMe can tolerate slight head motions when the user performs speaking actions.

3.2 Data Collection

It is difficult to collect clear signal profiles for lip reading because of two reasons [2], [11]. First, the mouth motions caused by the user's speaking actions are very tiny. Second, the mouth motions are a complex combination of jaw, tongue, and other muscles around the mouth and thus cannot be simply treated as rigid motions as in gesture recognition [15], [17], [27]. As for RFID, it is more challenging because the signals might be absorbed by the skin if we directly attach tags to the user's body as in previous works [27].

To collect clear signal profiles for lip reading, we consider three different data collection strategies described as below.

3.2.1 *Direct Attachment (DA) Strategy*

We first follow existing works on RFID-based activity recognition and attach an array of tags directly to the skin around the mouth. However, we find this strategy has several drawbacks. First, the skin causes significant signal absorption, making it difficult to collect useful signals when the distance between the user (tag) and the antenna is larger than 30 cm. Second, even when the tags are very close to the antenna, the missing reading problem is serious, which makes it difficult to collect signals that can characterize the overall features of mouth motions. Actually, when the tags are placed on conductive materials such as a human being's skin, the reading rate and reading range will both decrease. As pointed out in [28], [29], the skin will cause severe signal absorption which consequently results in the missing

reading problem. For example, Fig. 2(a) plots the phase readings of four tags with the DA strategy. Serious missing readings for tag 2 and tag 4 can be observed, especially for tag 4 whose signals cannot be obtained most of the time. Even for tag 1 and tag 3 whose signals can be collected, the missing readings problem is still serious.

3.2.2 Soft Mask Tightly Clung to Mouth (MT) Strategy

To overcome the missing readings problem caused by signal absorption, we let the user wear a soft mask that is tightly clung to the mouth. With this strategy, the missing readings problem can be well mitigated, as shown in Fig. 2(b). However, because the mouth motions are greatly limited by the tight mask, there are nearly no obvious signal pattern changes when the user performs speaking actions. It can be observed from Fig. 2(b) that the signal patterns during the user's speaking action and during the silent time are similar. This makes it difficult to separate signal profiles related to mouth motions.

3.2.3 Wearable Plastic Mask (WM) Strategy

To overcome the problems in the DA strategy and the MT strategy, we propose a wearable plastic mask strategy. Instead of using a soft mask tightly clung to the mouth, we paste an array of tags to a hard transparent plastic mask and let the user wear the mask. The advantages of the WM strategy are as follows. First, because there is a certain distance between the tags and the mouth, the signal absorption problem and the missing readings problem can be well mitigated. Second, because the distance between the tags and the mouth is very short (only several centimeters), the signals backscattered by the tags can well capture the combined movement of muscles during the user's speaking action. This generates rich patterns for accurate mouth motion detection. Moreover, the tags' signals remain stable when the user is silent, which is helpful for mouth motion segmentation when the user performs speaking actions.

Fig. 2(c) plots the signal collected with the WM strategy when the distance between the antenna and the user is 80 cm. Two observations can be made: 1) The missing readings problem caused by signal absorption is well solved, and 2) the signal patterns when the user performs speaking actions are apparently different from the signal patterns in the silent time. This makes it possible to separate signal segments related to speaking actions from signal segments related to the silent time.

3.3 Data Preprocessing

3.3.1 Phase Ambiguity Mitigation

It is well known that phase readings of RFID tags are usually affected by *phase ambiguities* and *phase wrapping* problems [15], [17]. Phase ambiguity is a phenomenon in which consecutive phase readings might differ by π even when the tag is static, as shown in Fig. 2(c). Phase wrapping means that the phase value reported by the reader wraps when the actual phase approaches 0 or 2π. We adopt the techniques proposed in [30], [31] to mitigate phase ambiguity and phase wrapping problems. Fig. 3(a) shows the signal after resolving phase ambiguities.

3.3.2 Out-band Interference Filtering

The frequency of the signals caused by the mouth motion is around 2-5Hz [11]. To filter out noisy signals that are not caused by the user's speaking action, we use a Savitzky-Golay filter [32] to smooth the signal and filter out-band interferences. The Savitzky-Golay filter performs data filtering based on local polynomial least squares fitting in the time domain. It can preserve the shape and tendency of the signal while filtering out noises.

Consider a set of $2N+1$ data points $x[n]$ centered at $n=0$. We use an J-th order polynomial equation to fit the data

$$p(n) = \sum_{j=0}^{J} a_{\mathrm{j}} n^{\mathrm{j}}, \tag{1}$$

such that the mean-squared approximation error for the $2N+1$ data points is minimized

$$\varepsilon_n = \sum_{n=-N}^{N} (p(n) - x[n])^2. \tag{2}$$

The polynomial coefficients can be calculated as described in [32]. We set $N=3$ to balance the computation complexity and the quality of the filtered data. The signals after filtering are shown in Fig. 3(b). It is apparent that the signals exhibit clear pattern changes when the user performs speaking actions. Moreover, the signals are quite stable when the user is silent.

3.3.3 Multi-path Reflections Removal

Besides the signals reflected by the mouth, the signals collected at the RFID reader might also contain reflection signals from other static objects in the environment, *e.g.*, walls and furniture. Assume that there are P articulators causing reflections during the mouth motion, the total received signals can be represented as

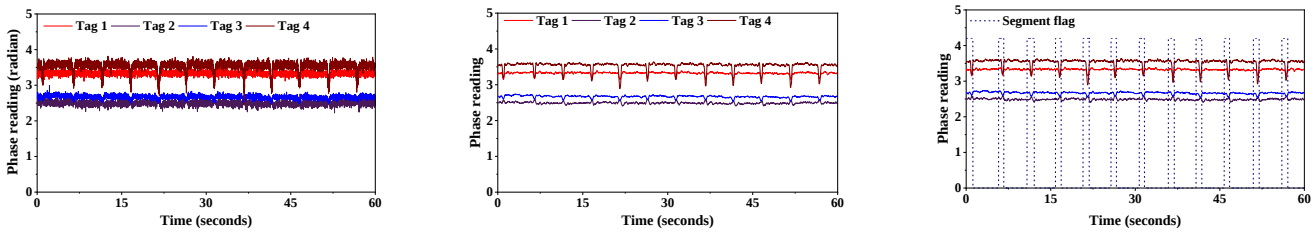
$$d(t) = \sum_{p=1}^{P} a_{\mathrm{p}}(t) sin(2\pi f_{\mathrm{p}} t + \phi_{\mathrm{p}}) + \varphi, \tag{3}$$

where f_{p} is the frequency of the reflected signals from the $p\text{-}th$ articulator, $a_{\mathrm{p}}(t)$ is the reflection coefficient related to the distance from the $p\text{-}th$ articulator to the receiver, ϕ_{p} is the corresponding phase, and φ denotes the reflections caused by static objects in the environment. Because static reflections usually come from a far longer distance than the distance between the tags and the mouth, we can use the method proposed in [2] to remove this interference by setting a threshold on the delay and retain only information related to mouth motions.

4 Mouth Motion Segmentation

4.1 Segmentation Algorithm

It can be observed that the signals are very stable when the user is silent but exhibit significant fluctuations when the user performs speaking actions. Based on these observations, we propose a threshold-based method based on the *Modified Varri Method* [33] to segment signal profiles related to the user's speaking action. The method uses a sliding window that combines a *frequency measure* estimated by the summation of the differences of consecutive signal samples



(a) Removing phase ambiguity (b) Filtering out-band interference (c) Mouth motion segmentation

Fig. 3. Data preprocessing: (a) after mitigating phase ambiguity; (b) after filtering out-band interferences; and (c) after motion segmentation.

and an *amplitude measure* of the signal values in the relevant windows to evaluate changes in the signal. Denote by L be the number of data points in the window and by $x_{\mathrm{i,k}}$ the k-th data point in the i-th window. For the i-th window, the amplitude measure $\mathcal{A}_{\mathrm{i}}$ and the estimated frequency measure $\mathcal{F}_{\mathrm{i}}$ are calculated as

$$\mathcal{A}(i) = \sum_{k=1}^{L} |x_{\mathrm{i,k}}| \text{ and } \mathcal{F}(i) = \sum_{k=1}^{L} |x_{\mathrm{i,k}} - x_{\mathrm{i,k-1}}|, \quad (4)$$

and the measurement difference function $\mathcal{G}$ is defined as

$$\mathcal{G}(i) = \mathcal{C}_{\mathrm{A}}|\mathcal{A}(i+1) - \mathcal{A}(i)| + \mathcal{C}_{\mathrm{F}}|\mathcal{F}(i+1) - \mathcal{F}(i)|, \quad (5)$$

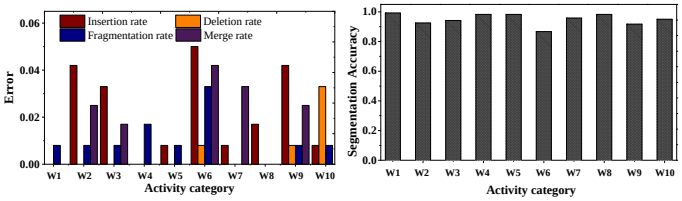
where $\mathcal{C}_{\mathrm{A}}$ and $\mathcal{C}_{\mathrm{F}}$ are two coefficients that change in various applications. We experimentally set their values in our implementation. The value of $\mathcal{G}(i)$ would be small when the user is silent and would be large when the user performs speaking actions, and thus we can use a threshold-based approach to separate the signal profile related to mouth motions.

The detailed segmentation algorithm is as follows. For each data point $x[n]$, a bit $S[n]$ is used to indicate whether it belongs to a mouth motion segment or not: $S[n] = 1$ means YES and $S[n] = 0$ means NO. We first initialize $S[n] = 1$ for all data points and use a two-step algorithm to find mouth motion segments:

- *Determining resting time:* We use a window with length L_{S} to determine the resting time. For the u-th data points, we construct a window consisting of $\{x[u], \ldots, x[u+L_{\mathrm{S}}-1]\}$ and calculate $\mathcal{G}(u)$ according to Eq. (5). If $\mathcal{G}(u)$ is smaller than a predefined threshold, we set $S[n] = 0$ for $u \le n \le u + L_{\mathrm{S}} - 1$. Then we slide the window by one data point and repeat the process. All the data points whose $S[n] = 1$ form a candidate set of potential mouth motion segments.
- *Removing fragmented segments:* Some fragmented motions might be incorrectly identified due to noises. Noting that a speech action usually takes about a relative long time ($\sim 1second$), we remove segments whose length are too short to be a real speech action. To do this, for each identified mouth motion segment, we calculate the time duration of the segment and remove the segment whose duration is shorter than a threshold by setting corresponding $S[n] = 0$.

After we identify the mouth motion segments of all the U tags, we align their boundaries as follows. For each identified motion segment M_{i}, we calculate its left boundary $BL(i,u)$ and right boundary $BR(i,u)$ and set the left boundary and the right boundary for the segment as

$$BL_{\mathrm{i}} = \max\{BL(i,u)\}, BR_{\mathrm{i}} = \min\{BR(i,u)\}, 1 \le u \le U. \quad (6)$$



(a) Motion Segmentation Error (b) Action Detection Accuracy

Fig. 4. Motion segmentation Accuracy.

4.2 Segmentation Accuracy

An example of the segmentation result is shown in Fig. 3(c). It can be observed that most motions can be correctly segmented. We use the following four metrics to evaluate the performance of the proposed mouth motion segmentation algorithm as suggested in [21].

- *Insertion rate* indicates the proportion of pronunciation activities detected during the resting interval. It reflects the algorithm's sensitivity to noise in the resting interval.
- *Deletion rate* indicates the percentage of missed pronunciation activities. It reflects the sensitivity of the algorithm to phase changes caused by different lip patterns.
- *Fragmentation rate* indicates the ratio of dividing a single mouth motion into multiple activities. It evaluates the algorithm's ability to handle complex or incoherent speech activities.
- *Merge rate* indicates the ratio of combining multiple speaking activities into one mouth motion. It evaluates the ability of the algorithm to identify pronunciation activities at relatively high speeds.

Fig. 4(a) plots the four metrics when the user speaks 10 different words. Both fragmentation rate and merge rate are very low for all the words, with an average value of 0.010 and 0.014 respectively, indicating that the proposed algorithm can accurately identify the whole segment related to speaking actions. The deletion rate is very low (0.005), which means that our approach can detect almost all speaking actions. The average insertion rate is 0.021, which means that there is a small probability that a speaking action is detected when the user is silent, *e.g.*, triggering a false alarm. Fig. 4(b) plots the segmentation accuracy of the proposed algorithm when the user speaks 10 different words. The average segmentation accuracy is 0.95 and the highest accuracy is 0.992. The results demonstrate that the

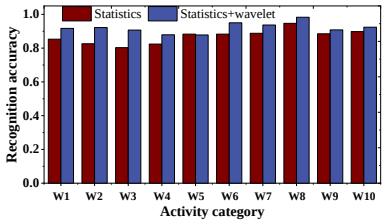


Fig. 5. Comparison of recognition accuracy with/without wavelet features.

proposed algorithm can accurately identify the signal segments related to mouth motions.

5 Feature Extraction and Model Training

We use two types of features to comprehensively capture the characteristics of the signal profiles related to mouth motion: the time-domain statistical features that capture the coarse-grained global tendency of the signal and the frequency-domain wavelet (transformation) coefficient features that capture the fine-grained local features of the signal profile. We then feed the features into a machine-learning-based classifier to recognize what words the user is speaking.

5.1 Statistical Features

The first set of features we use are the statistics of the signal profiles, which can be divided into three categories:

- The statistics that reflect the *central tendency* of the signal profile, including *mode*, *first quartile*, *median*, the *third quartile*, and the *arithmetic mean*.
- The statistics that reflect the *dispersion* property of the signal profile, including *variance*, *coefficient of variation*, *maximum value*, and *minimum value*.
- The statistics that reflect the global shape of the signal profile, including *skewness* and *kurtosis*.

Fig. 5 plots the recognition accuracy when using only statistical features on a vocabulary containing 10 different English words (the detailed word list is given in Section 7). For nine out of the ten words, the recognition accuracy is lower than 0.9, and the average accuracy is 0.87. We find that the statistics can well capture global coarse-grained features of the signal profile, but they cannot reflect fine-grained local features of signal profiles related to different speaking actions. For example, Fig. 6(a) and Fig. 6(c) show the signal profiles of two words "`you`" and "`go`" respectively. While the shapes of the two profiles are different, their time-domain statistics are very similar and thus are difficult to distinguish by using only the aforementioned statistical features. We need to extract more fine-grained features of the signal profile to improve the recognition accuracy.

5.2 Wavelet Coefficients Features

To capture the fine-grained local features of the signal profile, we exploit the wavelet transform of the signals and use the wavelet transform coefficients as additional features.

5.2.1 Wavelet Transform of Signal Profiles

Compared with the statistical features, the wavelet transformation features provide additional information from two aspects. First, the wavelet transform captures the features of the signal profile from both time domain and frequency domain, while the statistical features consider only time domain. Second, the wavelet features provide multi-scale feature of the signal profile, which is helpful to distinguish between words that have common pronunciation actions (*e.g.*, "`you`" and "`go`").

We use the discrete wavelet transformation (DWT) [11] to obtain the wavelet features. The discrete signal $x[n]$ can be represented by a combination of wavelet basis functions

$$x[n] = \frac{1}{\sqrt{M}}(\sum_{k} W_{\phi}[j_0,k]\phi_{\mathrm{j_0,k}}[n] + \sum_{j=j_0}^{\infty}\sum_{k} W_{\psi}[j,k]\psi_{\mathrm{j,k}}[n]), \tag{7}$$

where $x[n]$ represents the original discrete signal defined in $[0, M-1]$, $\phi_{\mathrm{j_0,k}}[n]$ and $\psi_{\mathrm{j,k}}[n]$ are discrete wavelet basis functions defined in $[0, M-1]$. In order to obtain the wavelet coefficients, we select a set of basis functions $\phi_{\mathrm{j_0,k}}[n], k \in Z$ and $\psi_{\mathrm{j,k}}[n], (j,k) \in Z^2, j \geq j_0$ that are orthogonal in the decomposition process, *e.g.*,

$$< \phi_{\mathrm{j_0,k}}[n], \psi_{\mathrm{j,k}}[n] >= \delta_{\mathrm{j_0,j}}\delta_{\mathrm{k,m}}. \tag{8}$$

In discrete wavelet decomposition, the signals are iteratively decomposed into two parts: the approximate coefficients and the detailed coefficients. The following equations are used to calculate the wavelet packet coefficients at each level:

$$W_{\phi}[j_0,k] = \frac{1}{\sqrt{M}}\sum_{n} f[n]\phi_{\mathrm{j_0,k}}[n], \tag{9}$$

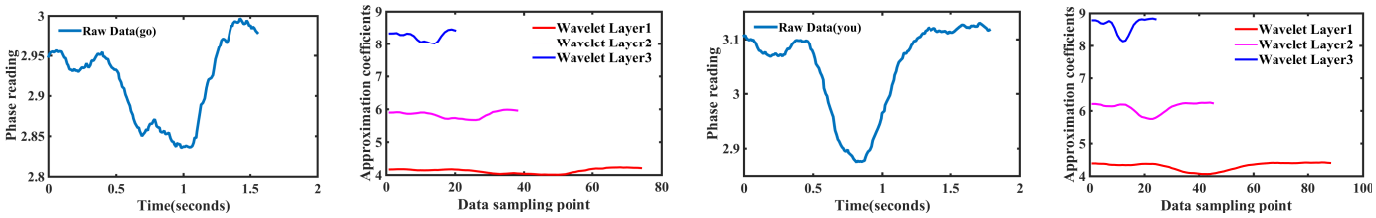
$$W_{\psi}[j,k] = \frac{1}{\sqrt{M}}\sum_{n} f[n]\psi_{\mathrm{j,k}}[n], j \geq j_0, \tag{10}$$

where $W_{\phi}[j_0,k]$ and $W_{\psi}[j,k]$ represent the approximation coefficients and the detailed coefficients, respectively.

5.2.2 Data Alignment

To apply wavelet decomposition to the signal profile, the data points for each tag should be of the same length. However, the channel access mechanism used in RFID such as ALOHA protocol is inherently a time division random access protocol, and thus the data points for each tag are usually different. Furthermore, the data are collected at different time, making them not aligned with each other. Thus, before applying the wavelet transformation on the signal profile, we should align the data by interpolation. We use a cubic Hermite polynomial to interpolate the data points [17]. Each speaking action takes about 1 second. Considering that the tag identification rate is about 400 readings per second, we interpolate 400 data points for each tag.

After the data are aligned, we apply the wavelet decomposition to the interpolated data for each tag and obtain a set of wavelet coefficients. We choose the *Daubechies* wavelet bases [34] and evaluate the recognition accuracy with different decomposition levels. A RandomForest classifier is used to evaluate the recognition accuracy. The results are listed in TABLE 1. It can be observed that the highest recognition accuracy is achieved when the signals are decomposed at level 4.



(a) "Signal profile ('go')" (b) "Wavelet coefficients ('go')" (c) "Signal profile ('you')" (d) "Wavelet coefficients ('you')"

Fig. 6. Comparison of wavelet decomposition of "you" and "go" with similar pronunciation.

TABLE 1
Recognition accuracy vs. decomposition levels.

Decomposition level	2	3	4	5	6
Recognition Accuracy	0.879	0.884	**0.891**	0.869	0.843

TABLE 2
Recognition accuracy with different classifiers

Classifier model	Recognition accuracy
RandomForest	**0.927**
J48	0.803
RandomTree	0.732
DecisionTable	0.606
Logistic	0.850
BayesNet	0.740
LogitBoost	0.811

We combine both the wavelet coefficient features and the statistic features obtained in Section 5.1 to classify different speaking actions and plot the results in Fig. 5. It can be observed for 9 out of the 10 words, the recognition accuracy is improved when the wavelet coefficient features are used. The most significant improvement is for the third word, for which the recognition accuracy is improved by nearly 10 percent. The only word for which the recognition accuracy decreases is the fifth word, for which the accuracy is slightly decreased from 0.883 to 0.878. By using the wavelet coefficient features, the average recognition accuracy over all words is improved from 0.87 to 0.92.

5.3 Classifier Model Selection

We test the recognition accuracy of different classifier models and list the results in TABLE 2. The *RandomForest* model performs significantly better than all the other classifiers, which is consistent with previous studies on RFID-based activity recognition. Thus in all the following experiments we use RandomForest as the default classifier.

6 Cross Environment Retraining based on Transfer Learning

One inherent problem of wireless sensing is environment dependence. When the trained classification model is deployed to a new environment different from the training environment, its performance usually significantly degrades. To maintain a high accuracy of the model in the new environment, the user needs to collect samples in the new environment and retrain the model. As model training usually requires a large number of samples to achieve a high accuracy, collecting data and retraining the model incur high costs. To reduce the cost of collecting new samples, we use an approach based on the mapping model and transfer learning to synthesize samples in the new environment with the samples collected in the training environment, similar to [35]. The flowchart is shown in Fig. 7, and it can be divided into two parts: synthesizing samples with the mapping model, and transferring the mapping model to a new environment.

The goal of synthesizing samples with the mapping model is to generate synthetic data of the new environment to decrease the collecting costs in the new environment. Assume that there are enough samples collected in the training environment and a few samples (*e.g.*, 5 samples of each class) are available in the new environment. To synthesize samples of the new environment, we need to learn the mapping relationship between the samples collected in the training environment and the samples collected in the new environment. After obtaining the relationship, we can map the samples in the training environment to the new environment, which can be used to train a new classification model in the new environment. Specifically, we first build a neural network model as the mapping model. Denote by S_t the set of samples collected in the training environment and by S_n the set of samples collected in the new environment. We first divide S_t into two subsets S_t^1 and S_t^2, where the samples in S_t^1 are used to train the mapping model and the samples in S_t^2 are used to generate synthetic samples.

When training the mapping model, for each sample $st \in S_t^1$, we select a sample $sn \in S_n$ with the distance between sn and st being the shortest among all the samples in S_n as its paring sample. (Note that st and sn should be in the same class.) We then use all the st as input and all the sn as output to train the mapping model. In our implementation, we train a network containing 7 fully-connected layers as the mapping model. After the mapping model is built, we feed the samples in S_t^2 into the model and take the output of the model as the synthesized samples in the new environment. These synthetic samples are generated by the model based on the samples of the training environment, which do not exist in the real world but contain features of real world data because of the mapping relationship. With the method, we only need to collect a few samples in the new environment to build the mapping model, and then generate synthetic samples to train the new classification model, which avoids the collection of a large number of samples in the new environment and thus decreases the cost of sample collection in the new environment.

The goal of the second part, transferring the recognition model to a new environment, is to transfer a trained recognition model to another new environment by transfer learning to decrease the number of training samples since

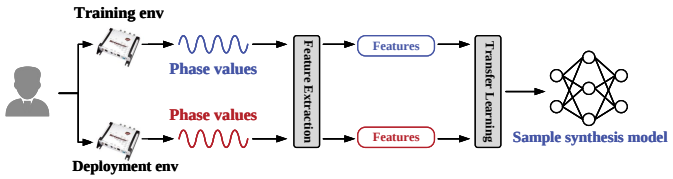


Fig. 7. Synthesizing samples in a new environment by transferring samples collected in the training environment.

we do not have to train the network from scratch. Specifically, when a trained recognition model is deployed to a new environment A, we freeze the parameters of the first several layers and only fine-tune the parameters of the last fully connected layer with samples from A, either collected samples or synthesized samples. The rationale for freezing the parameters and fine-tuning the last layer lies in transfer learning. Based on the theory the transfer learning [36], the shallow layers of the neural network extract general features, which are the same across similar tasks, while the deep layers of neural network extract task-specific features among similar tasks. Hence, we can transfer the trained shallow layers to a new recognition model to avoid training the model from scratch.

As shown in Fig. 7, this synthesis process still requires collecting some samples in the new environment. However, the number of required samples are much smaller than the number of samples when all the samples are newly collected, and thus the cost in collecting samples and training model can be significantly reduced.

7 Performance Evaluation

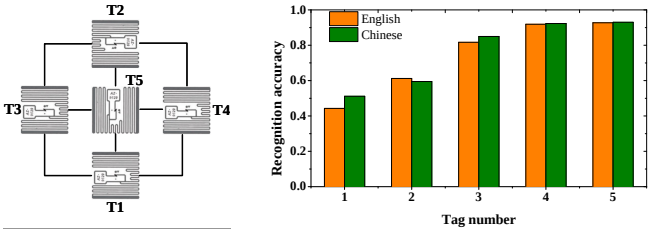
7.1 Experiment Setup

We implement HearMe with the commercial Impinj R420 Speedway reader and a circularly polarized Laird S9028PCR antenna. We attach an array of passive tags (Monza AZ-9629) to a transparent plastic mask and let the user wear the plastic mask and speak different words (without making sound). To minimize the impact of mutual coupling between tags [37], we follow the deployment of tags in the work [16]. In specific, we deploy the nearby tags perpendicular to each other such that the interference between tags is low enough. The reader uses the maximum throughput mode to continuously interrogate data from the tags, with the tag reading rate of around 400 readings per second. The default distance between the antenna and the user is set at 80 cm. The low level reader protocol (LLRP) [38] is used to transmit the data from the reader to a laptop for data processing, which is equipped with a 2.6GHz Intel(R) Core(TM) i5 CPU and 8GB RAM memory. The data processing software is implemented in Java.

7.1.1 Vocabulary

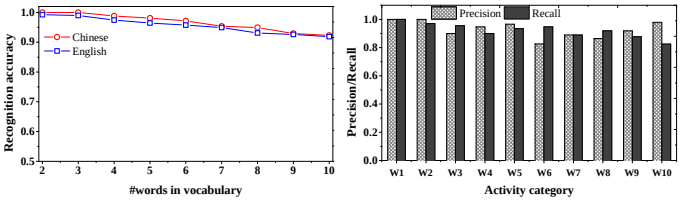
We consider an English vocabulary and a Chinese vocabulary, each containing 10 frequently used words.

- English vocabulary [11]: `how`, `are`, `you`, `good`, `like`, `go`, `play`, `any`, `watch`, `dog`.
- Chinese vocabulary: 中, 爱, 是, 华, 国, 吃饭, 睡觉, 学习, 走路, 上课.



(a) Tag Deployment (b) Recognition Accuracy

Fig. 8. Recognition accuracy with different number of tags.



(a) Recognition accuracy (b) Precision/Rrecall

Fig. 9. Impact of vocabulary size: (a) recognition accuracy; (b) Precision/recall for English words.

7.1.2 Data Collection

We invite 7 volunteers and let each volunteer perform the speaking actions for about 100 minutes. In each speaking action, the volunteer randomly select one word in the vocabulary and speaks the word without making sound. There is a short silence time (1-2 seconds) between two consecutive speaking actions. The collected data are divided into two parts: 80% of the data are used as training data to train a classifier model, 10% of the data are used as validation data to validate classifier, and the rest 10% data are used as testing data to evaluate the performance. The default classifier model is RandomForest.

7.2 Accuracy of HearMe under Different Settings

7.2.1 Impact of Tag Array Deployment

We first investigate how the number of tags and different deployment strategy of tags affect the recognition accuracy. Due to the limited space of the mask, we can attach at most five tags to the mask without causing signal coupling between tags. The positions of tags are illustrated in Fig. 8(a). When k tags are used, these tags are pasted to positions $T1, \ldots, Tk$. For example, when four tags are used, the four tags are pasted to $T1, T2, T3, T4$, respectively.

The recognition accuracy with different number of tags is plotted in Fig. 8(b). It can be observed the recognition accuracy is low (≤ 0.6) when only one or two tags are used. The reason is that the mouth motions are complex combinations of different components including jaw, tongue and other muscles around the mouth, and thus two tags are not enough to capture the comprehensive signal changes of such complex motions. The recognition accuracy is improved to a high level (0.919 for English words and 0.923 for Chinese words) when four tags are used. However, using more than four tags does not further improve the recognition accuracy. Moreover, more tags would cause more signal collisions and decrease the quality of the obtained signal profiles. Thus, in the following experiments, we use the four tag deployment strategy in default.

TABLE 3
Accuracy when #word in [11, 20].

#Word	11	12	13	14	15	16	17	18	19	20
Accuracy	.925	.923	.920	.919	.909	.907	.898	.897	.887	.881

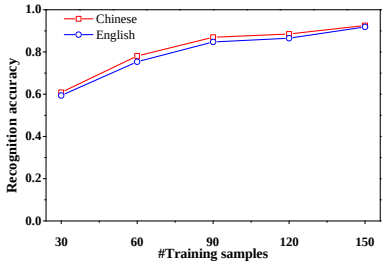


Fig. 10. The recognition accuracy vs. number of training samples of each word.

7.2.2 Impact of Vocabulary Size

We investigate the recognition accuracy with different number of words in the vocabulary and plot the results in Fig. 9(a). For both English and Chinese, the average recognition is higher than 0.95 when there are less than 7 words in the vocabulary. The recognition accuracy gradually decreases when the vocabulary length increases, but remains higher than 0.92 even when the vocabulary contains as much as 10 words.

Fig. 9(b) plots the precision and recall for each word in the English vocabulary. There are 7 words for which the precision is higher than 0.90 and there is only 1 word for which the precision is lower than 0.85. The average precision is 0.93, showing the high recognition precision of HearMe. The average recall over the 10 words is as high as 0.92, showing that HearMe can correctly segment speaking actions and recognize related words. We further merged the English and the Chinese vocabulary to investigate the accuracy of HearMe when the number of words exceeds 10. As shown in TABLE 3, the recognition accuracy remains higher than 0.88 even when there are 20 words in the vocabulary.

7.2.3 Impact of Training Sample Numbers

The number of training samples will impact the recognition accuracy of HearMe. We plot the accuracy of HearMe with respect to the number of training samples in Fig. 10. It can be observed that with only 60 training samples for each word, the classification accuracy for both Chinese words and English words are close to 0.8. The recognition accuracy improves when more samples are used for training and tends stable after the number of samples is larger than 150, at which point the classification accuracy is 0.93 for Chinese and 0.92 for English, respectively.

7.2.4 Impact of Distance

The phase readings in RFID system are affected by the distance between the antenna and tags. To investigate how the distance impacts the recognition accuracy of HearMe, we change the distance between the user and the antenna by $-\lambda/4$ and $\lambda/4$ respectively[2]. It can be observed that for both Chinese and English the recognition accuracy decreases when the testing distance and training distance are different, but the differences in all cases are smaller than 0.1, which means HearMe can resist to the deformation in signal profile caused by distance changes. In detail, the accuracy decreases for Chinese words is smaller than 0.085 and the accuracy decreases for English words is smaller than 0.092. Such a decrease in accuracy is acceptable and can be further reduced by considering context of different words. Moreover, we can further reduce the decrease in accuracy by building multiple classifier models at different distances and select the best classifier adaptively.

TABLE 4
Recognition accuracy at different distances.

Distance	40cm	80cm	120cm
Chinese	0.857	0.923	0.838
English	0.830	0.918	0.824

7.2.5 Resistance to Environmental Interferences

The recognition accuracy of HearMe might be affected by environmental interferences, such as other moving objects in the same space. To investigate HearMe's performance in noisy environments, we consider three cases. 1) *Interference-free scenario*, in which the user performs speaking actions with the distance between the user and the antenna set at 80cm and there are no moving objects in the environment. The data collected in this scenario are used to train the classifier model. 2) *Slight interference scenario*, in which there is one volunteer randomly moving around the user but the volunteer always keeps 3 meters away from the user. The volunteers can jump or using mobile phones. 3) *Serve interference scenario*, in which there are two volunteers randomly moving around the user who performs the speaking actions. The volunteers can do the same actions as in the second scenario, but the distance between the two volunteers and the user is less than 1 meter.

We use the data collected in the interference-free scenario to train the classifier and use the data collected in all the three scenarios as testing data. The recognition accuracy in different scenarios are given in the Fig. 12. It can be observed HearMe still performs very well in the slight interference scenario, with only 1 to 2 percent decrease in recognition accuracy. However, in the serve interference scenario, the recognition accuracy for both Chinese words and English words decreases significantly. The drop in accuracy is 0.22 for Chinese and 0.25 for English. In such cases, we should develop effective methods to remove interference signals caused by the moving objects.

7.2.6 Impact of Data Volume

Due to the throughput limitation of RFID systems, when the number of users increases, the effective number of samples for each user(tag) will decrease, which will consequently affect the accuracy of HearMe. For example, when there are n users in the monitoring region, on average the collected

2. Because $\lambda/4$ ≈8cm is a short distance difficult to control, we actually change the distance to $-5\lambda/4$ and $5\lambda/4$ respectively as shown in TABLE 4.

Actual Word	Predicted Word									
	CW1	CW2	CW3	CW4	CW5	CW6	CW7	CW8	CW9	CW10
CW1	1.00	0	0	0	0	0	0	0	0	0
CW2	0	1.00	0	0	0	0	0	0	0	0
CW3	0	0.01	0.90	0	0.01	0	0.01	0	0.06	0
CW4	0	0	0	0.94	0.02	0	0.02	0.02	0	0
CW5	0	0	0	0	0.97	0	0	0	0	0.03
CW6	0	0.01	0.02	0	0	0.83	0.06	0.03	0.04	0.01
CW7	0	0	0	0	0.02	0.03	0.89	0	0.02	0.04
CW8	0	0	0	0.08	0	0	0	0.86	0	0.06
CW9	0	0	0.02	0.02	0	0.03	0	0.01	0.92	0
CW10	0	0	0	0	0.02	0	0	0	0	0.98

(a) HearMe (Chinese)

Actual Word	Predicted Word									
	EW1	EW2	EW3	EW4	EW5	EW6	EW7	EW8	EW9	EW10
EW1	0.92	0.05	0	0	0	0.03	0	0	0	0
EW2	0.03	0.92	0	0	0	0.01	0.02	0.02	0	0
EW3	0	0.01	0.91	0.04	0.01	0	0	0	0.03	0
EW4	0.03	0	0.02	0.88	0.03	0	0.04	0	0	0
EW5	0	0.01	0.04	0.03	0.88	0	0	0	0.03	0.01
EW6	0.03	0	0	0	0.02	0.95	0	0	0	0
EW7	0.01	0	0	0.03	0	0	0.94	0.02	0	0
EW8	0	0	0.02	0	0	0	0	0.98	0	0
EW9	0	0	0	0.01	0.04	0	0	0.01	0.91	0.03
EW10	0	0	0	0.01	0.03	0	0.01	0.03	0	0.92

(b) HearMe (English)

Actual Word	Predicted Word									
	CW1	CW2	CW3	CW4	CW5	CW6	CW7	CW8	CW9	CW10
CW1	0.83	0	0.08	0.08	0	0	0	0	0	0
CW2	0	0.90	0	0	0	0	0	0	0.10	0
CW3	0	0.01	0.80	0	0.10	0	0	0	0	0.10
CW4	0	0	0	0.88	0	0	0	0.12	0	0
CW5	0	0	0	0	1.00	0	0	0	0	0
CW6	0	0.10	0	0	0	0.90	0	0	0	0
CW7	0	0	0	0.17	0	0	0.83	0	0	0
CW8	0	0	0.11		0	0	0	0.89	0	0
CW9	0	0	0	0	0.10	0	0	0	0.90	0
CW10	0	0	0	0	0.09	0	0	0.09	0	0.82

(c) DTW (Chinese)

Actual Word	Predicted Word									
	EW1	EW2	EW3	EW4	EW5	EW6	EW7	EW8	EW9	EW10
EW1	0.80	0	0	0	0	0	0	0	0.10	0.10
EW2	0	0.92	0	0	0	0.08	0	0	0	0
EW3	0	0	0.90	0	0	0	0	0	0	0.10
EW4	0	0	0	0.78	0	0.22	0	0	0	0
EW5	0.09	0	0	0	0.91	0	0	0	0	0
EW6	0	0	0.10	0	0	0.80	0	0.10	0	0
EW7	0	0	0	0.1	0	0	0.90	0	0	0
EW8	0.12	0	0	0	0	0	0	0.88	0	0
EW9	0	0	0	0	0	0	0	0	1.00	0
EW10	0	0	0	0.08	0	0	0.08	0.08	0	0.75

(d) DTW (English)

Fig. 11. Confusion matrix of HearMe and the DTW based approach: (a) HearMe on Chinese words; (b) HearMe on English words; (c) DTW-based approach on Chinese words; and (d) DTW-based approach on English words. The average accuracy of HearMe is 0.923 and 0.919 for Chinese words and English words, respectively. The average accuracy of DTW-based approach is 0.87 and 0.85, respectively.

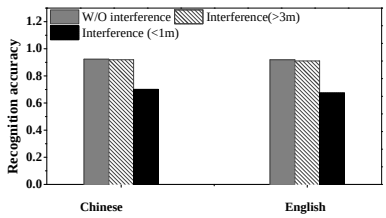


Fig. 12. Recognition accuracy vs. interferences.

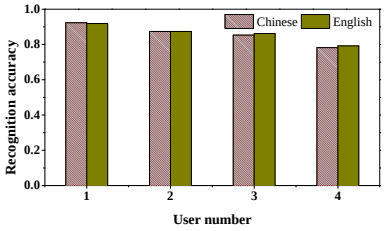


Fig. 13. Recognition accuracy vs. user number.

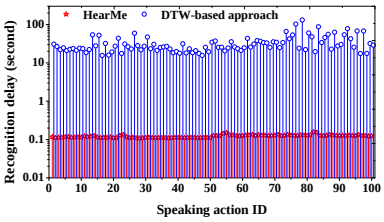


Fig. 14. Recognition delay of different approaches.

data for each tag would drop to $1/n$ of when there is only one user. We evaluated the performance of HearMe when the data rate decreases to mimic a multiple user scenario. Please note that compared with approaches based on other wireless techniques such as WiFi, HearMe can easily distinguish different signals for different users in a multiple user scenario by taking advantage of tag ID. In RFID system, each tag has a unique ID. Thus, by grouping signals according to tag ID, we can easily separate signal profiles for different users in RFID-based lip-reading system.

We plot the recognition accuracy of HearMe with different data volume in Fig. 13. The recognition accuracy decreases only slightly when the data volume fraction decreases from 1 to 1/3 (which mimics a three still user scenario), from 0.923 to 0.853 for Chinese and from 0.919 to 0.862, respectively. However, when there are more than four users, the recognition accuracy sharply drops to below 0.8 (0.782 for Chinese and 0.792 for English). The reason is that when there are too many users, the data collected for each user are very sparse, which makes our mouth motion segmentation algorithm fail to correctly segment the signal profile corresponding to the speaking actions. This in turn decreases the recognition accuracy in word recognition. More robust motion segmentation algorithms need to be developed for such cases.

7.2.7 Impact of Different Users

We also investigate the accuracy of HearMe for different individual users. We collect data for 7 different volunteers and we test HearMe's accuracy in a person-specific manner. The results are shown in Fig. 15. It can be observed that for the different users, the recognition accuracy varies. For four volunteers, the accuracy is higher than 0.9, and the highest accuracy is 0.95. The lowest accuracy is 0.84 (user 4). We speculate that the performance fluctuation might be caused by the different speaking habits of different users. For example, some users speak with smaller mouth motions

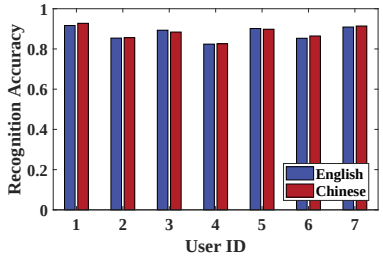


Fig. 15. The recognition accuracy of HearMe for 7 different users.

TABLE 5
Recognition accuracy at different angles.

Angle	-60°	0°	60°
40cm	0.634	0.830	0.793
80cm	0.795	0.918	0.796

and at a faster speed than others, which might lead to low accuracy for these users.

7.2.8 Impact of Different Angles

We evaluated the performance of HearMe when the user faces the antenna from different angles and give the results in TABLE 5. We test HearMe at three different angles: 60° from lest, 0°, and 60° from right. It can be observed that the performance is best when the user is exactly at the front of the antenna, and the performance slightly degrades when the user are at other angles. However, the recognition accuracy is still close to 0.8 even when the angle between the user and the antenna is as large as 60° when the distance is 80 cm, showing that the operational range of HearMe is relatively large.

7.3 Impact of Head Movement

We also investigated how slight head movements affect the performance of HearMe. Three different scenarios are considered: 1) the normal scenario, in which the user can

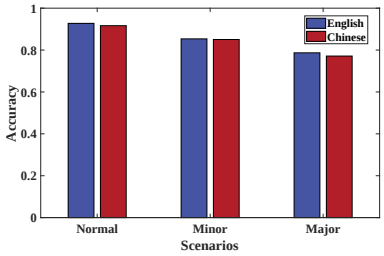


Fig. 16. Accuracy of HearMe when the user moves his/her head to a different extent during the speaking action.

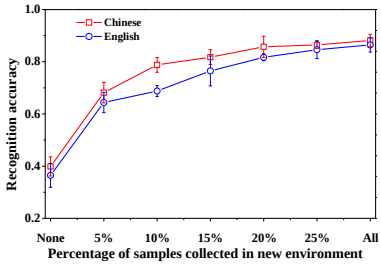


Fig. 17. Accuracy in new environment with synthesized samples.

only slightly shake his head with an angle up to 5 degrees; 2) the minor movement scenario, in which the user can shake his head up to about 20 degrees; and 3) the major movement scenario, in which the user can shake his head up to about 45 degrees. The results are shown in Fig. 16. It can be observed that when the user only slightly moves his head during the speaking action (i.e., in the normal scenario), the accuracy of lip reading recognition is about 0.93. In the minor movement scenario, the accuracy slightly drops to 0.86. When the user shakes his head violently (i.e., in the major movement scenario), the accuracy will drop to about 0.78. Based on these results, we conclude that HearMe can be used in scenarios in which the user slightly moves his head, e.g., in scenarios in which the movement angle is within 20 degrees. However, we also point out that if the users move their heads violently when speaking, HearMe cannot work well and the accuracy drops to lower than 0.8. The reason is that when the user shakes his head violently, the RF signals will be significantly changed by the head movement.

7.4 Comparison with DTW-based Approach

Dynamic time wrapping (DTW) has been used in RFID systems to classify different gestures [15], [16]. We also implement a DTW-based approach to recognizing different speaking motions. For each word we select 10 templates, and when calculating the distances between a signal segment and a template we use the multi-dimensional DTW (MDTW) as in [15]. The confusion matrix on both the Chinese vocabulary and the English vocabulary for HearMe and the DTW-based approach are given in Fig. 11. Compared with the DTW-based approach, HearMe improves recognition accuracy by 5 percent for Chinese words and by 7 percent for English words.

The recognition delay for each speaking action is plotted in Fig. 14. The recognition delay includes all the time needed to process the data, but does not include the time spent in collecting data from tags, which is limited by the throughput of the RFID system and on the order of several seconds. The average recognition delay for HearMe is less than 150 ms. In contrast, DTW-based approaches are much slower because they need to find the optimal match in a large number of templates [15], [17]. The average recognition delay for DTW-based approaches is longer than 20 seconds, which is two orders of magnitude higher than HearMe.

7.5 Cross-Environment Training Performance

Fig. 17 plots the recognition accuracy when cross-environment synthesized samples are used. We first train the classification model in environment A. When the model is directly applied to a new environment B, the recognition accuracy is lower than 0.4 for both Chinese and English. When we use new samples collected in B (about 200 samples each word) to train a new model, the accuracy for Chinese and English are 0.88 and 0.86, respectively. We then use a part of the new samples to learn the relationship between the samples collected from the two environments and transfer samples collected in environment A to environment B with the learned relationship. Fig. 17 shows the accuracy with different number of new samples. It can be observed that with about 25% new samples (50 samples each word), the classifier trained with the synthesized samples performs nearly the same as using all the new samples. This significantly reduces the sample collection cost by 75%.

8 Conclusion

This paper presents the design and implementation of HearMe, an RFID-based contactless lip reading system that can work in a long operational range and can simultaneously track multiple users by leveraging the ability of RFID tags to uniquely identify an object. HearMe achieves an average recognition accuracy of higher than 0.93 for a vocabulary containing 10 words. Moreover, HearMe is very fast, with a recognition latency less than 150 ms. This means that HearMe can support real-time communications for those people having difficulties in speaking. A transfer learning based approach is also proposed to effectively reduce sample collection cost when the user switches to a new environment. HearMe still has much room to be improved. For instance, it can be enhanced to recognize short sentences by using the hidden Markov model (HMM) to exploit the context of the words. Currently, lip-reading systems based on wireless signals requires the user remain still during the speaking actions. So another improve direction is to design new signal processing algorithms to handle the impact of user mobility on the performance of such systems

Acknowledgement

This work is partially supported by the National Natural Science Foundation of China (Grant Nos. 61772559, 62177047, 62172154, 61872310, 62072231), the Hunan Provincial Natural Science Foundation of China under grant No. 2020JJ3016, Shenzhen Science and Technology Innovation

Commission (JCYJ20200109142008673), and the Collaborative Innovation Center of Novel Software Technology and Industrialization. Prof. Jian Zhang is the corresponding author of this paper. Professor Zomaya would like to acknowledge the support of the Australian Research Council Discovery Project (DP200103494).

REFERENCES

[1] Fakhteh Soltani, Fatemeh Eskandari, and Shadan Golestan. Developing a gesture-based game for deaf/mute people using microsoft kinect. In *2012 Sixth International Conference on Complex, Intelligent, and Software Intensive Systems*, pages 491–495. IEEE, 2012.

[2] Jiayao Tan, Cam-Tu Nguyen, and Xiaoliang Wang. Silenttalk: Lip reading through ultrasonic sensing on mobile phones. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, pages 1–9. IEEE, 2017.

[3] Mohamed Aktham Ahmed, Bilal Bahaa Zaidan, Aws Alaa Zaidan, Mahmood Maher Salih, and Muhammad Modi bin Lakulu. A review on systems-based sensory gloves for sign language recognition state of the art between 2007 and 2017. *Sensors*, 18(7):2208, 2018.

[4] Li Lu, Jiadi Yu, Yingying Chen, Hongbo Liu, Yanmin Zhu, Linghe Kong, and Minglu Li. Lip reading-based user authentication through acoustic sensing on smartphones. *IEEE/ACM Transactions on Networking*, 27(1):447–460, 2019.

[5] Jiayao Tan, Xiaoliang Wang, Cam-Tu Nguyen, and Yu Shi. Silentkey: A new authentication framework through ultrasonic-based lip reading. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1):36:1–36:18, 2018.

[6] Takeshi Saitoh. Development of communication support system using lip reading. *IEEJ Transactions on Electrical and Electronic Engineering*, 8(6):574–579, 2013.

[7] KBRK Ramesha, KB Raja, KR Venugopal, and LM Patnaik. Feature extraction based face recognition, gender and age classification, 2010.

[8] Kshitiz Kumar, Tsuhan Chen, and Richard M Stern. Profile view lip reading. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–429. IEEE, 2007.

[9] Wenqiang Chen, Maoning Guan, Yandao Huang, Lu Wang, Rukhsana Ruby, Wen Hu, and Kaishun Wu. Vitype: A cost efficient on-body typing system through vibration. In *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 1–9. IEEE, 2018.

[10] Liang Wang, Tao Gu, Xianping Tao, and Jian Lu. Toward a wearable rfid system for real-time activity recognition using radio patterns. *IEEE Transactions on Mobile Computing*, 16(1):228–242, 2017.

[11] Guanhua Wang, Yongpan Zou, Zimu Zhou, Kaishun Wu, and Lionel M Ni. We can hear you with wi-fi! *IEEE Transactions on Mobile Computing*, 15(11):2907–2920, 2016.

[12] Feng Lin, Chen Song, Yan Zhuang, Wenyao Xu, Changzhi Li, and Kui Ren. Cardiac scan: A non-contact and continuous heart-based user authentication system. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking (Mobicom)*, pages 315–328, 2017.

[13] Jian Liu, Yan Wang, Yingying Chen, Jie Yang, Xu Chen, and Jerry Cheng. Tracking vital signs during sleep leveraging off-the-shelf wifi. In *Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pages 267–276. ACM, 2015.

[14] Mingmin Zhao, Fadel Adib, and Dina Katabi. Emotion recognition using wireless signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking (Mobicom)*, pages 95–108, 2016.

[15] Yongpan Zou, Jiang Xiao, Jinsong Han, Kaishun Wu, Yun Li, and Lionel M Ni. Grfid: A device-free rfid-based gesture recognition system. *IEEE Transactions on Mobile Computing*, 16(2):381–393, 2017.

[16] Chuyu Wang, Jian Liu, Yingying Chen, Hongbo Liu, Lei Xie, Wei Wang, Bingbing He, and Sanglu Lu. Multi-touch in the air: Device-free finger tracking and gesture recognition via cots rfid. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pages 1691–1699. IEEE, 2018.

[17] Shigeng Zhang, Chengwei Yang, Xiaoyan Kui, Jianxin Wang, Xuan Liu, and Song Guo. Reactor: Real-time and accurate contactless gesture recognition with rfid. In *Proceedings of 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 212–220. IEEE, 2019.

[18] Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen. Towards a practical lipreading system. In *CVPR 2011*, pages 137–144. IEEE, 2011.

[19] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7356–7365, 2018.

[20] Jingxian Wang, Chengfeng Pan, Haojian Jin, Vaibhav Singh, Yash Jain, Jason I Hong, Carmel Majidi, and Swarun Kumar. RFID tattoo: A wireless platform for speech recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(4):1–24, 2019.

[21] Han Ding, Jinsong Han, Longfei Shangguan, Wei Xi, Zhiping Jiang, Zheng Yang, Zimu Zhou, Panlong Yang, and Jizhong Zhao. A platform for free-weight exercise monitoring with passive tags. *IEEE Transactions on Mobile Computing*, 16(12):3279–3293, 2017.

[22] Zimu Zhou, Longfei Shangguan, Xiaolong Zheng, Lei Yang, and Yunhao Liu. Design and implementation of an rfid-based customer shopping behavior mining system. *IEEE/ACM Transactions on Networking*, 25(4):2405–2418, 2017.

[23] Jinsong Han, Han Ding, Chen Qian, Wei Xi, Zhi Wang, Zhiping Jiang, Longfei Shangguan, and Jizhong Zhao. Cbid: A customer behavior identification system using passive tags. *IEEE/ACM Transactions on Networking*, 24(5):2885–2898, 2016.

[24] Lei Xie, Jianqiang Sun, Qingliang Cai, Chuyu Wang, Jie Wu, and Sanglu Lu. Tell me what i see: Recognize rfid tagged objects in augmented reality systems. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 916–927. ACM, 2016.

[25] Hanchuan Li, Can Ye, and Alanson P Sample. Idsense: A human object interaction detection system based on passive uhf rfid. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2555–2564. ACM, 2015.

[26] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 2018.

[27] Yuxiao Hou, Yanwen Wang, and Yuanqing Zheng. Tagbreathe: Monitor breathing with commodity RFID systems. In *Proceedings of the 37th IEEE International Conference on Distributed Computing Systems (ICDCS)*, pages 404–413. IEEE, 2017.

[28] Abdelmoula Bekkali, Sicheng Zou, Abdullah Kadri, Michael Crisp, and Richard V Penty. Performance analysis of passive uhf rfid systems under cascaded fading channels and interference effects. *IEEE Transactions on Wireless Communications*, 14(3):1421–1433, 2015.

[29] Pavel V Nikitin and KVS Rao. Performance limitations of passive uhf rfid systems. In *2006 IEEE Antennas and Propagation Society International Symposium*, pages 1011–1014. IEEE, 2006.

[30] Jia Liu, Feng Zhu, Yanyan Wang, Xia Wang, Qing-feng Pan, and Lijun Chen. Rf-scanner: Shelf scanning with robot-assisted RFID systems. In *Proceedings of INFOCOM*, 2017.

[31] Shigeng Zhang, Chengwei Yang, Danming Jiang, Xiaoyan Kui, Song Guo, Albert Y. Zomaya, and Jianxin Wang. Nothing blocks me: Precise and real-time LOS/NLOS path recognition in RFID systems. *IEEE Internet of Things Journal*, 6(3):5814–5824, 2019.

[32] Ronald W Schafer et al. What is a savitzky-golay filter. *IEEE Signal processing magazine*, 28(4):111–117, 2011.

[33] Hamed Azami, Karim Mohammadi, and Behzad Bozorgtabar. An improved signal segmentation using moving average and savitzky-golay filter. *Journal of Signal and Information Processing*, 3(01):39, 2012.

[34] Hafeez Ullah Amin, Aamir Saeed Malik, Rana Fayyaz Ahmad, Nasreen Badruddin, Nidal Kamel, Muhammad Hussain, and Weng-Tink Chooi. Feature extraction and classification for eeg signals using wavelet transform and machine learning techniques. *Australasian physical & engineering sciences in medicine*, 38(1):139–149, 2015.

[35] Jie Zhang, Zhanyong Tang, Meng Li, Dingyi Fang, Petteri Nurmi, and Zheng Wang. Crosssense: Towards cross-site and large-scale wifi sensing. In Rajeev Shorey, Rohan Murty, Yingying (Jennifer) Chen, and Kyle Jamieson, editors, *Proceedings of the 24th Annual*

International Conference on Mobile Computing and Networking (MobiCom), pages 305–320, 2018.

[36] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[37] Zhong-qin Wang, J. Andrew Zhang, Fu Xiao, and Min Xu. Accurate aoa estimation for RFID tag array with mutual coupling. *IEEE Internet of Things Journal*, 9(15):12954–12972, 2022.

[38] https://www.gs1.org/standards/epc-rfid/llrp/1-1-0.

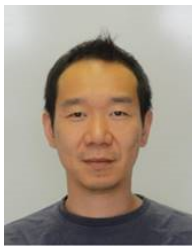
Shigeng Zhang Shigeng Zhang received the BSc, MSc, and DEng degrees, all in Computer Science, from Nanjing University, China, in 2004, 2007, and 2010, respectively. He is currently a Professor in School of Computer Science and Engineering at Central South University, China. His research interests include Internet of Things, mobile computing, RFID systems, and IoT security. He has published more than 70 technique papers in top international journals and conferences including Ubicomp, Infocom, Mobihoc, ICNP, TMC, TC, TPDS, TOSN, and JSAC. He is on the editorial board of International Journal of Distributed Sensor Networks, and was a program committee member of many international conferences including ICC, ICPADS, MASS, UIC and ISPA. He is a member of IEEE and ACM.

Zijing Ma Zijing Ma received the BSc degree in computer science and technology from South China Agricultural University in 2020. He is currently working towards his MSc degree in computer science and technology from Central South University, China. His research interests include wireless sensing, RFID, and the Internet of Things.

Kaixuan Lu Kaixuan Lu received the B.S. degree in network engineering from Zhongyuan University of Technology in 2017. He then graduated from Central South University for the MS degree in computer technology. His research interests focus on wireless sensing and the Internet of Things.

Xuan Liu Xuan Liu is currently a Professor in the College of Computer Science and Electronic Engineering at Hunan University. She received the BSc degree in information and computing mathematics from XiangTan University in 2005, the MSc degree in Computer Science from National University of Defense Technology in 2008, and the PhD degree in the Hong Kong Polytechnic University in 2015, respectively. Her research interests include Multi-agent reinforcement learning, RFID systems and Internet of Things. She has published more than 40 technique papers in top international journals including JSAC/ToN/TMC/TC/TPDS and top conferences including Infocom/ICNP/Mobihoc/Ubicomp.

Jia Liu Jia Liu is an associate professor with the Department of Computer Science and Technology at Nanjing University, Nanjing, China. Before that, he received the B.E. degree in software engineering from Xidian University, Xi'an, China, in 2010. He received the Ph.D. degree in computer science and technology from Nanjing University, Nanjing, China, in 2016. His research mainly focuses on RFID systems. He is a member of the IEEE and ACM.



Song Guo Song Guo is a Full Professor at Department of Computing, The Hong Kong Polytechnic University. He also holds a Changjiang Chair Professorship awarded by the Ministry of Education of China. Prof. Guo is a Fellow of the Canadian Academy of Engineering, Member of Academia Europaea, and Fellow of the IEEE (Computer Society). His research interests are mainly in federated learning, edge AI, mobile computing, and distributed systems. He published many papers in top venues with wide impact in these areas and was recognized as a Highly Cited Researcher (Clarivate Web of Science). He is the recipient of over a dozen Best Paper Awards from IEEE/ACM conferences, journals, and technical committees. Prof. Guo is the Editor-in-Chief of IEEE Open Journal of the Computer Society. He was an IEEE ComSoc Distinguished Lecturer and a member of IEEE ComSoc Board of Governors. He has served for IEEE Computer Society on Fellow Evaluation Committee, Transactions Operations Committee, Steering Committee of IEEE Transactions on Cloud Computing, Editor-in-Chief Search Committee, and been named on editorial board of a number of prestigious international journals like IEEE TC, IEEE TPDS, IEEE TCC, IEEE TETC, ACM CSUR, etc. He has also served as chairs of organizing and technical committees of many international conferences.

Albert Y. Zomaya Albert Y. ZOMAYA is Peter Nicol Russell Chair Professor of Computer Science and Director of the Centre for Distributed and High-Performance Computing at the University of Sydney. To date, he has published ¿ 700 scientific papers and articles and is (co-)author/editor of ¿30 books. A sought-after speaker, he has delivered ¿250 keynote addresses, invited seminars, and media briefings. He is currently the Editor in Chief of the ACM Computing Surveys and served in the past as Editor in Chief of the IEEE Transactions on Computers (2010-2014) and the IEEE Transactions on Sustainable Computing (2016-2020).

Professor Zomaya is a decorated scholar with numerous accolades including Fellowship of the IEEE, the American Association for the Advancement of Science, and the Institution of Engineering and Technology. He is a Fellow of the Australian Academy of Science, Royal Society of New South Wales, Foreign Member of Academia Europaea, and Member of the European Academy of Sciences and Arts. Some of Professor Zomaya recent awards include the New South Wales Premier's Prize of Excellence in Engineering and Information and Communications Technology (2019) and the Research Innovation Award, IEEE Technical Committee on Cloud Computing (2021). His research interests lie in parallel and distributed computing, networking, and complex systems.

Jian Zhang JIAN ZHANG received the B.Eng. degree in computer science from the National University of Defense Technology, in 1998, and the M.Eng. and Ph.D. degree in computer science from Central South University, in 2002 and 2007, respectively, where he is currently an Associate Professor with the School of Computer Science and Engineering. His research interests include optimization theory, cyberspace security, cloud computing, and cognitive radio technology.

Jianxin Wang Jianxin Wang received the BEng and MEng degrees in computer engineering from Central South University, China, in 1992 and 1996, respectively, and the PhD degree in computer science from Central South University, China, in2001. He is the Dean of and a professor in School of Computer Science and Engineering, Central South University, Changsha, Hunan, P.R. China. His current research interests include algorithm analysis and optimization, parameterized algorithm, Bioinformatics and computer network. He is a senior member of the IEEE.