We Can Hear You with Wi-Fi!

Guanhua Wang, *Student Member, IEEE*, Yongpan Zou, *Student Member, IEEE*,
Zimu Zhou, *Student Member, IEEE*, Kaishun Wu, *Member, IEEE*, and Lionel M. Ni, *Fellow, IEEE*

Abstract—Recent literature advances Wi-Fi signals to "see" people's motions and locations. This paper asks the following question: Can Wi-Fi "hear" our talks? We present WiHear, which enables Wi-Fi signals to "hear" our talks without deploying any devices. To achieve this, WiHear needs to detect and analyze fine-grained radio reflections from mouth movements. WiHear solves this micro-movement detection problem by introducing *Mouth Motion Profile* that leverages partial multipath effects and wavelet packet transformation. Since Wi-Fi signals do not require line-of-sight, WiHear can "hear" people talks within the radio range. Further, WiHear can simultaneously "hear" multiple people's talks leveraging MIMO technology. We implement WiHear on both USRP N210 platform and commercial Wi-Fi infrastructure. Results show that within our pre-defined vocabulary, WiHear can achieve detection accuracy of 91 percent on average for single individual speaking no more than six words and up to 74 percent for no more than three people talking simultaneously. Moreover, the detection accuracy can be further improved by deploying multiple receivers from different angles.

Index Terms—Wi-Fi radar, micro-motion detection, moving pattern recognition, interference cancelation

1 Introduction

Recent research has pushed the limit of ISM (Industrial Scientific and Medical) band radiometric detection to a new level, including motion detection [11], gesture recognition [37], localization [10], and even classification [14]. We can now detect motions through-wall and recognize human gestures, or even detect and locate tumors inside human bodies [14]. By detecting and analyzing signal reflection, they enable Wi-Fi to "SEE" target objects.

Can we use Wi-Fi signals to "HEAR" talks? It is commonsensical to give a negative answer. For many years, the ability of hearing people talks can only be achieved by deploying acoustic sensors closely around the target individuals. It costs a lot and has a limited sensing and communication range. Further, it has detection delay because the sensor must first record the sound and process it, then transmit it to the receiver. In addition, it cannot be decoded when the surrounding is too noisy.

This paper presents WiHear (Wi-Fi Hearing), which explores the potential of using Wi-Fi signals to HEAR people talk and transmit the talking information to the detector at the same time. This may have many potential applications: 1) WiHear introduces a new way to hear people talks without deploying any acoustic sensors. Further, it still works well even when the surrounding is noisy. 2) WiHear will bring a new interactive interface between human and devices, which enables devices to sense and recognize more complicated human behaviors (e.g., mood) with negligible cost. WiHear makes devices "smarter". 3) WiHear can help millions of disabled people to conduct simple commands to devices with only mouth motions instead of complicated and inconvenient body movements.

How can we manage Wi-Fi hearing? It sounds impossible at first glance, as Wi-Fi signals cannot detect or memorize any sound. The key insight is similar to radar systems. WiHear locates the mouth of an individual, and then recognizes his words by monitoring the signal reflections from his mouth. By recognizing mouth moving patterns, WiHear can extract talking information the same way as lip reading. Thus, WiHear introduces a micro-motion detection scheme that most of previous literature can not achieve. And this minor movement detection can also achieve the ability like leap motion [1]. The closest works are WiSee [37] and WiVi [11], which can only detect more notable motions such as moving arms or legs using doppler shifts or ISAR (inverse synthetic aperture radar) techniques.

To transform the above high-level idea into a practical system, we need to address the following challenges:

(1) How to detect and extract tiny signal reflections from the mouth only? Movements of surrounding people, and other facial movement (e.g., wink) from the target user may affect radio reflections more significantly than mouth movements do. It is challenging to cancel these interferences from the received signals while retaining the information from the tiny mouth motions.

To address this issue, WiHear first leverages MIMO beamforming to focus on the target's mouth to reduce irrelevant multipath effects introduced by omnidirectional antennas. Such avoidance of irrelevant multipath will enhance WiHear's detection accuracy, since the impact from other people's movements will not dominate when the radio beam is located on the target individual. Further, since for a specific user, the frequency and pattern of wink is relatively stable, WiHear exploits interference cancelation to remove the periodic fluctuation caused by wink.

- *G. Wang, Y. Zou, Z. Zou and K. Wu are with CSE Department, HKUST and the College of Computer Science and Software Engineering, Shenzhen University, China.*
  E-mail: {gwangab, yzouad, zzhouad, kwinson}@cse.ust.hk.
- *L.M. Ni is with the University of Macau, Macau. E-mail: ni@cse.ust.hk.*

Manuscript received 23 May 2015; revised 6 Jan. 2016; accepted 8 Jan. 2016. Date of publication 18 Jan. 2016; date of current version 28 Sept. 2016.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TMC.2016.2517630

(2) How to analyze the tiny radio reflections without any change on current Wi-Fi signals? Recent advances harness customized modulation like Frequency-Modulated Carrier Waves (FMCW) [10]. Others like [19] use ultra wide-band and large antenna array to achieve precise motion tracking. Moreover, since mouth motions induce negligible doppler shifts, approaches like WiSee [37] are inapplicable.

WiHear can be easily implemented on commercial Wi-Fi devices. We introduce *mouth motion profiles*, which *partially* leverage multipath effects caused by mouth movements. Traditional wireless motion detection focuses on movements of arms or body, which can be simplified as a rigid body. Therefore they remove all the multipath effects. However, mouth movement is a non-rigid motion process. That is, when pronouncing a word, different parts of the mouth (e.g., jaws and tongue) have different moving speeds and directions. We thus cannot regard the mouth movements as a whole. Instead, we need to leverage multipath to capture the movements of different parts of the mouth.

In addition, since naturally only one individual is talking during a conversation, the above difficulties only focus on single individual speaking. How to recognize multiple individuals' talking simultaneously is another big challenge. The reason for this extension is that, in public areas like airports or bus stations, multiple talks happen simultaneously. WiHear enables hear multiple individuals' simultaneously talks using MIMO technology. We let the senders form multiple radio beams to locate on different targets. Thus, we can regard the target group of people as the senders of the reflection signals from their mouths. By implementing a receiver with multiple antennas and enabling MIMO technology, it can decode multiple senders' talks simultaneously.

Summary of results. We implemented WiHear in both USRP N210 [8] and commercial Wi-Fi products. Fig. 1 depicts some syllables (vowels and consonants) that WiHear can recognize[1]. Overall, WiHear can recognize 14 different syllables, 33 trained and tested words. Further, WiHear can correct recognition errors by leveraging related context information. In our experiments, we collect training and testing samples at roughly the same location with the same link pairs. All the experiments are per-person trained and tested. For single user cases, WiHear can achieve an average detection accuracy of 91 percent to correctly recognize sentences made up of no more than six words, and it works in both line-of-sight (LOS) and non-line-of-sight (NLOS) scenarios. With the help of MIMO technology, WiHear can differentiate up to 3 individuals' simultaneously talking with accuracy up to 74 percent. For through-wall detection of single user, the accuracy is up to 26 percent with one link pair, and 32 percent with 3 receivers from different angles. In addition, based on our experimental results, the detection accuracy can be further improved by deploying multiple receivers from different angles.

Contributions. We summarize the main contributions of WiHear as follows:

- WiHear exploits the radiometric characteristics of mouth movements to analyze micro-motion in a non-invasive and device-free manner. To the best of our knowledge, this is the first effort using Wi-Fi signals to hear people talk via PHY layer Channel State Information (CSI) on off-the-shelf WLAN infrastructure.
- WiHear achieves lip reading and speech recognition in LOS, NLOS scenarios. WiHear also has the potential of speech recognition in through-wall scenarios with relatively low accuracy.
- WiHear introduces *mouth motion profile* using partial multipath effect and discrete wavelet packet transformation to achieve lip reading with Wi-Fi.
- We simultaneously differentiate multiple individuals' talks using MIMO technology.

1. Jaws and tongue movement based lip reading can only recognize 30~40 percent of the whole vocabulary of English [24].

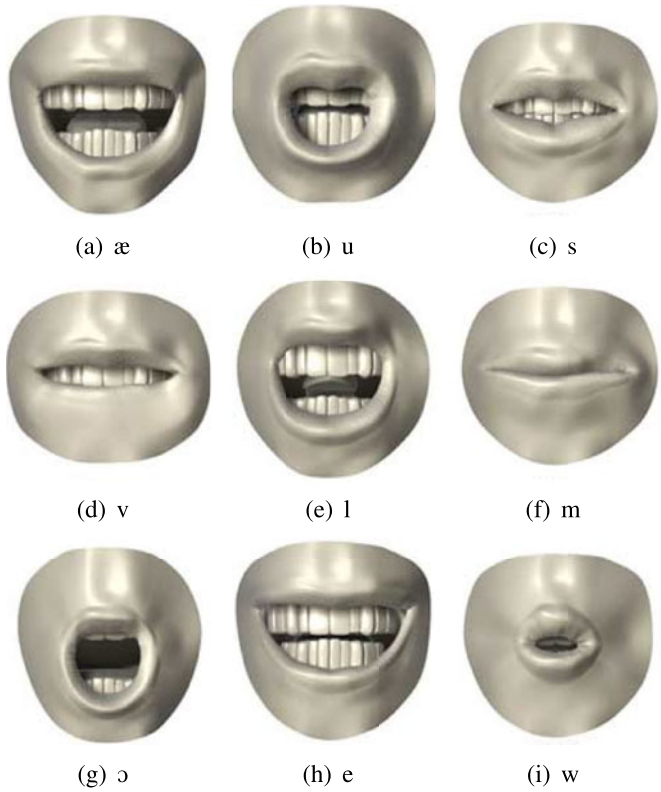


Fig. 1. Illustration of vowels and consonants [36] that WiHear can detect and recognize, © Gary C. Martin.

In the rest of this paper, we first summarize related work in Section 2, followed by an overview in Section 4. Sections 5 and 6 detail the system design. Section 7 extends WiHear to recognize multiple talks. We present the implementation and performance evaluation in Section 8, discuss the limitations in Section 9, and conclude in Section 10.

2 Related Work

The design of WiHear is closely related to the following two categories of research.

Vision/sensor based motion sensing. The flourish of smart devices has spurred an urge for new human-device interaction interfaces. Vision and sensors are among prevalent ways to detect and recognize motions.

Popular vision-based approaches include Xbox Kinect [2] and Leap Motion [1], which use RGB hybrid cameras and depth sensing for gesture recognition. A slightly different approach which has been grounded into commercial

products called Vicon systems [3]. These systems can achieve precise motion tracking using cameras by detecting and analysing markers placed on human body, which needs both instrumentation to environments and target human body. Yet they are limited to the field of view and are sensitive to lighting conditions. Thermal imaging [34] acts as an enhancement in dim lighting conditions and non-line-of-sight scenarios at the cost of extra infrastructure. Vision has also been employed for lip reading. [26] and [25] present a combination of acoustic speech and mouth movement image to achieve higher accuracy of automatic speech recognition in noisy environment. [33] presents a vision-based lip reading system and compares viewing a person's facial motion from profile and front view. [23] shows the possibility of sound recovery from the silent video.

Another thread exploits various wearable sensors or handhold devices. Skinput [29] uses acoustic sensors to detect on-body tapping locations. Agrawal et al. [12] enable writing in the air by holding a smartphone with embedded sensors. TEXIVE [15] leverages smartphone sensors to detect driving and texting simultaneously.

WiHear is motivated by these precise motion detection systems, yet aims to harness the ubiquitously deployed Wi-Fi infrastructure, and works non-intrusively (without on-body sensors) and through-wall.

Wireless-based motion detection and tracking. WiHear builds upon recent research that leverages radio reflections from human bodies to detect, track, and recognize motions [41]. WiVi [11] initializes through-wall motion imaging using MIMO nulling [35]. WiTrack [10] implemented an FMCW (Frequency Modulated Carrier Wave) 3D motion tracking system at the granularity of 10 cm. WiSee [37] recognizes gestures via Doppler shifts. AllSee [32] achieves low-power gesture recognition on customized RFID tags.

Device-free human localization systems locate a person by analyzing his impact on wireless signals received by pre-deployed monitors, while the person carries no wireless enabled devices [51]. The underlying wireless infrastructure varies, including RFID [52], Wi-Fi [51], ZigBee [47], and the signal metrics range from coarse signal strength [51], [47] to finer-grained PHY layer features [49], [50].

Adopting a similar principle, WiHear extracts and interprets reflected signals, yet differs in that WiHear targets at finer-grained motions from lips and tongue. Since the micro motions of the mouth produce negligible Doppler shifts and amplitude fluctuations, WiHear exploits beamforming techniques and wavelet analysis to focus on and zoom in the characteristics of mouth motions only. Also, WiHear is tailored for off-the-shelf WLAN infrastructure and is compatible with the current Wi-Fi standards. We envision WiHear as an initial step towards centimetre-order motion detection (e.g., finger tapping) and higher-level human perception (e.g., inferring mood from speech pacing).

3 Background on Channel State Information

In typical cluttered indoor environments, signals often propagate to the receiver via multiple paths. Such multipath effect creates varying path loss across frequencies, known as *frequency diversity* [38]. Frequency diversity depicts the small-scale spectral structure of wireless channels, and has been adopted for fine-grained location distinction [53], motion detection [48] and localization [43].

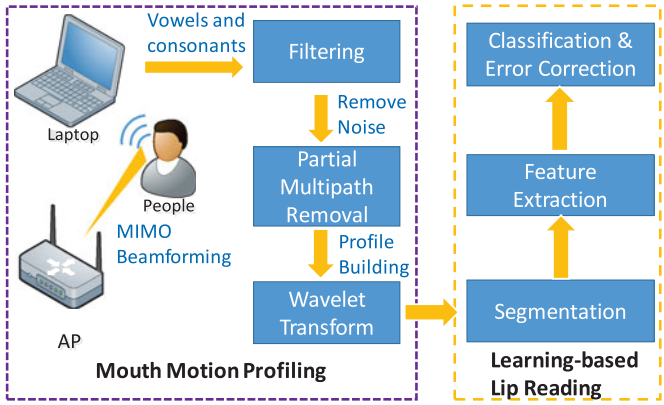


Fig. 2. Framework of WiHear.

Conventional MAC layer RSSI provides only a single-valued signal strength indicator. Model multi-carrier radio such as OFDM measures frequency diversity at the granularity of subcarrier, and stores the information in the form of Channel State Information. Each CSI depicts the amplitude and phase of a subcarrier:

$$H(f_k) = ||H(f_k)||e^{j\sin(\angle H)}, \tag{1}$$

where $H(f_k)$ is the CSI at the subcarrier with central frequency of f_k, and $\angle H$ denotes its phase. Leveraging the off-the-shelf Intel 5300 network card with a publicly available driver [28], a group of CSIs $H(f)$ of $K = 30$ subcarriers are exported to upper layers,

$$H(f) = [H(f_1), H(f_2), \ldots, H(f_K)]. \tag{2}$$

Recent WLAN standards (e.g., 802.11n/ac) also exploit MIMO techniques to boost capacity via *spatial diversity*. We thus involve spatial diversity to further enrich channel measurements. Given M receiver antennas and N transmitter antennas, we obtain an $M \times N$ matrix of CSIs $\{H_{mn}(f)\}_{M\times N}$, where each element $H_{mn}(f)$ is defined as Equation (2).

In a nutshell, PHY layer CSI portrays finer-grained spectral structure of wireless channels. Spatial diversity provided by multiple antennas further expands the dimensions of channel measurements. While RSSI based device-free human detection systems mostly make binary decisions whether a person is present along the link [47] or resort to multiple APs to fingerprint a location [51], TagFree utilizes the rich feature space of CSI to identify different objects with only a single AP.

4 WiHear Overview

WiHear is a wireless system that enables commercial Wi-Fi devices to hear people talks using OFDM (Orthogonal Frequency Division Multiplexing) Wi-Fi devices. Fig. 2 illustrates the framework of WiHear. It consists of a transmitter and a receiver for single user lip reading. The transmitter can be configured with either two (or more) omnidirectional antennas on current mobile devices or one directional antenna (easily changeable) on current APs (access points). The receiver

only needs one antenna to capture radio reflections. WiHear can be extended to multiple APs or mobile devices to support multiple simultaneous users.

WiHear transmitter sends Wi-Fi signals towards the mouth of a user using beamforming. WiHear receiver extracts and analyzes reflections from mouth motions. It interprets mouth motions in two steps:

1) *Wavelet-based mouth motion profiling.* WiHear sanitizes received signals by filtering out-band interference and partially eliminating multipath. It then constructs mouth motion profiles via discrete wavelet packet decomposition.
2) *Learning-based lip reading.* Once WiHear extracts mouth motion profiles, it applies machine learning to recognize pronunciations, and translates them via classification and context-based error correction.

At the current stage, WiHear can only detect and recognize human talks if the user performs no other movements during speaking. We envision the combination of device-free localization [49] and WiHear may achieve continuous Wi-Fi hearing for mobile users. For irrelevant human interference or ISM band interference, WiHear can tolerant irrelevant human motions 3 m away from the link pair without dramatic performance degradation.

5 Mouth Motion Profiling

The first step of WiHear is to construct *Mouth Motion Profile* from received signals.

5.1 Locating on Mouth

Due to the small size of the mouth and the weak extent of its movements, it is crucial to concentrate maximum signal power towards the direction of the mouth. In WiHear, we exploit MIMO beamforming techniques to locate and focus on the mouth, thus both introducing less irrelevant multipath propagation and magnifying signal changes induced by mouth motions [20]. We assume the target user does not move when he speaks.

The locating process works in two steps:

1) The transmitter sweeps its beam for multiple rounds while the user repeats a predefined gesture (e.g., pronouncing [æ] once per second). The beam sweeping is achieved via a simple rotator made by stepper motors similar in [54]. We adjust the beam directions in both azimuth and elevation as in [55]. Meanwhile, the receiver searches for the time when the gesture pattern is most notable during each round of sweeping. With trained samples (e.g., waveform of [æ] for the target user), the receiver can compare the collected signals with trained samples. And it chooses the time stamp in which the collected signals share highest similarity with trained samples.
2) The receiver sends the selected time stamp back to the transmitter and the transmitter then adjusts and fixes its beam accordingly. After each round of sweeping, the transmitter will get the time stamp feedback to adjust the emitted angle of the radio beam. The receiver may also further feedback to the transmitter during the analyzing process to refine the direction of the beam. As the example shown in Fig. 3, after the transmitter sweeping the beam for several rounds, the receiver sends back time slot 3 to the transmitter.

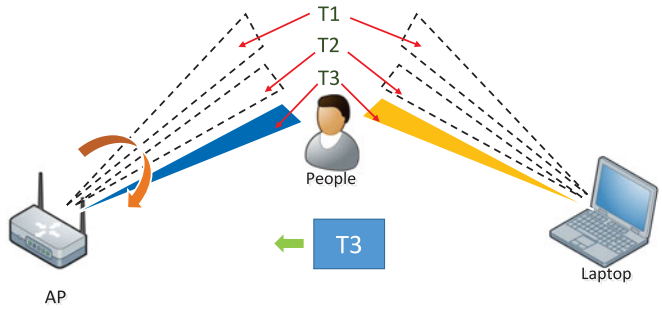


Fig. 3. Illustration of locating process.

Based on our experimental results, the whole locating process usually costs around 5-7 seconds, which is acceptable in real-world implementation. we can And we define correctly locating as the mouth is within the beam's coverage. More precisely, since the horizontal angle of our radio beam is roughly 120 degree, our directional antenna rotates 120 degree per second. Thus basically we sweep our radio beam for around two rounds. And then it can locate to the correct direction. For single user scenarios, we tested 20 times with three times failure, and thus the accuracy is around 85 percent. For multiple user scenarios, we define the correct locating as all users' mouths are within the radio beams. We tested with three people for 10 times with two times failure, and thus the accuracy is around 80 percent.

5.2 Filtering Out-Band Interference

As the speed of human speaking is low, signal changes caused by mouth motion in the temporal domain are often within 2-5 Hz [46]. Therefore, we apply band-pass filtering on the received samples to eliminate out-band interference.

In WiHear, considering the trade-off between computational complexity and functionality, we adopt a 3-order Butterworth IIR band-pass filter [21], of which the frequency response is defined by equation (3). Butterworth filter is designed to have maximum flat frequency response in the pass band and roll off towards zero in the stop band, which ensures the fidelity of signals in target frequency range while removing out-band noises greatly. The gain of an n-order Butterworth filter is:

$$G^2(w) = |H(jw)|^2 = \frac{G_0^2}{1 + \left(\frac{w}{w_c}\right)^{2n}}, \tag{3}$$

where $G(w)$ is the gain of Butterworth filter; w represents the angular frequency; w_c is the cutoff frequency; n is the order of filter, in our case, n = 3; G_0 is the DC gain.

Specifically, since normal speaking frequency is 150-300 syllables/minute [46], we set the cutoff frequency to be (60/60-300/60) Hz to cancel the DC component (corresponding to static reflections) and high frequency interference. In practice, as the radio beam may not be narrow enough, a common low-frequency interference is caused by winking. As shown in Fig. 4, however, the frequency of winking is

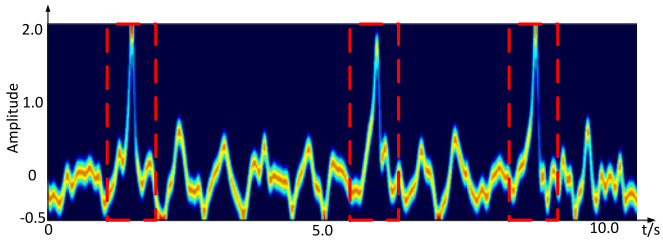


Fig. 4. The impact of wink (as denoted in the dashed red box).

smaller than 1 Hz (0.25 Hz on average). Thus, most of reflections from winking are also eliminated by filtering.

5.3 Partial Multipath Removal

Unlike previous work (e.g., [10]), where multipath reflections are eliminated thoroughly, WiHear performs *partial* multipath removal. The rationale is that mouth motions are non-rigid compared with arm or leg movements. It is common for the tongue, lips, and jaws to move in different patterns and deform in shape sometimes. Consequently, a group of multipath reflections with similar delays may all convey information about the movements of different parts of the mouth. Therefore, we need to remove reflections with long delays (often due to reflections from surroundings), and retain those within a delay threshold (corresponding to non-rigid movements of the mouth).

WiHear exploits CSI of commercial OFDM based Wi-Fi devices to conduct partial multipath removal. CSI represents a sampled version of the channel frequency response at the granularity of subcarrier. An IFFT (Inverse Fast Fourier Transformation) is first operated on the collected CSI to approximate the power delay profile in the time domain [42]. We then empirically remove multipath components with delay over 500 ns [30], and convert the remaining power delay profile back to the frequency domain CSI via an FFT (Fast Fourier Transformation). Since for typical indoor channel, the maximum excess delay is usually less than 500 ns [30], we set it as the initial value. The maximum excess delay of power delay profile is defined to be the temporal extent of the multipath that above a particular threshold. The delay threshold is empirically selected and adjusted based on the training and classification process (Section 6). More precisely, if we cannot get well-trained waveform (i.e., easy to be classified as a group) of one specific word/syllable, we empirically adjust the multipath threshold value.

5.4 Mouth Motion Profile Construction

After filtering and partial multipath removal, we obtain a sequence of cleaned CSI. Each CSI represents the phases and amplitudes on a group of 30 OFDM subcarriers. To reduce computational complexity with keeping the temporal-spectral characteristics, we explore to select a single representative value for each time slot.

We apply identical and synchronous sliding windows on all subcarriers and compute a coefficient C for each of them in each time slot. The coefficient C is defined as the peak to peak value on each subcarrier within a sliding window. Since we have filtered the high frequency components, there would be little dramatic fluctuation caused by interference or noise [21]. Thus the peak-to-peak value can represent human talking behaviors. We also compute another metric, the mean of signal strength in each time slot for each subcarrier. The mean values of all subcarriers facilitate us to pick the several subcarriers (in our case, we choose ten such subcarriers) which represent the most centralized ones, by analyzing the distribution of mean values in each time slot. Among the chosen subcarriers, based on C calculated within each time slot, we pick the waveform of the subcarrier which has the maximum coefficient C. By sliding the window on each subcarrier synchronously, we can pick a series of waveform segments from different subcarriers and assemble them into a single one by arranging them one by one. We define the assembled CSIs as a *Mouth Motion Profile*.

Some may argue that this peak-to-peak value may be dominated by environment changes. However, first of all, we have filtered the high frequency components. In addition, as mentioned in introduction and previous sessions, during our experiment, we keep the surrounding environment static. Thus it is unlikely to introduce irrelevant signal fluctuation caused by the environment. Furthermore, the sliding window we use is 200 ms (we can change the duration of sliding window according to different people's speaking patterns). These three reasons may ensure that, for most scenarios, our peak-to-peak value is dominated by mouth movements. Further, we use all the 30 subcarriers information to remove irrelevant multipath and keep partial multipath in Section 4.3. Thus we do not waste any information collected from PHY layer.

5.5 Discrete Wavelet Packet Decomposition

WiHear performs discrete wavelet packet decomposition on the obtained *Mouth Motion Profiles* as input for the learning based lip reading.

The advantages of wavelet analysis are two-folds: 1) It facilitates signal analysis on both time and frequency domain. This attribute can be leveraged in WiHear for analysing the motion of different parts on mouth (e.g., jaws and tongue) in varied frequency domains. It is because each part of mouth moves at different pace. It can also help WiHear locate the time periods for different parts of mouth motion when one specific pronouncing happens. 2) It achieves fine-grained multi-scale analysis. In WiHear, the motion of mouth when pronouncing some syllables shares a lot in common (e.g., [e], [i]), which makes them difficult to be distinguished. By applying discrete wavelet packet transform to the original signals, we can figure out the tiny difference which is beneficial for our classification process.

Here we first introduce discrete wavelet transform (DWT). As with the Fourier transform, where the signal is decomposed into linear combination of the basis if the signal is in the space spanned by the basis, wavelet decomposition also decomposes a signal to a combination of a series of expansion functions. It is given by equation (4):

$$f(t) = \sum_k a_k \phi_k(t), \tag{4}$$

where: k is an integer index of the finite or infinite sum, the a_k are the expansion coefficients, and the $\phi_k(t)$ are expansion functions, or the basis. If the basis chosen appropriately, there exists another set of basis $\widetilde{\phi}_k(t)$ which is orthogonal to $\phi_k(t)$.

The inner product of these two functions is ginven by equation (5):

$$<\phi_i(t), \widetilde{\phi}_j(t)> = \int \phi_i(t)\widetilde{\phi}_j^*(t)dt = \sigma_{ij}. \tag{5}$$

With the orthonormal property, it is easy to find the coefficients by equation (6):

$$\begin{aligned} <f(t), \widetilde{\phi}_k(t)> &= \int f(t)\widetilde{\phi}_k^*(t)dt \\ &= \int \left(\sum_{k'} a_{k'}\phi_{k'}(t)\right)\widetilde{\phi}_k^*(t)dt \\ &= \sum_{k'} a_{k'}\sigma_{k'k} \\ &= a_k. \end{aligned} \tag{6}$$

We can rewrite it as follows equation (7)

$$a_k = <f(t), \widetilde{\phi}_k(t)> = \int f(t)\widetilde{\phi}_k^*(t)dt. \tag{7}$$

For the signal we want to deal with, apply a particular basis satisfying the orthogonal property on that signal. It is easy to find the expansion coefficients a_k. Fortunately, the coefficients concentrate on some critical values, while others are close to zero.

Discrete wavelet packet decomposition is based on the well-known discrete wavelet transform, where a discrete signal $f[n]$ is approximated by a combination of expansion functions (the basis).

$$\begin{aligned} f[n] =& \frac{1}{\sqrt{M}}\sum_k W_\phi[j_0, k]\phi_{j_0,k}[n] \\ &+ \frac{1}{\sqrt{M}}\sum_{j=j_0}^{\infty}\sum_k W_\psi[j, k]\psi_{j,k}[n], \end{aligned} \tag{8}$$

where $f[n]$ represents the original discrete signal, which is defined in $[0, M-1]$, including totally M points. $\phi_{j_0,k}[n]$ and $\psi_{j,k}[n]$ are both discrete functions defined in $[0, M-1]$, called wavelet basis. Usually, the basis sets $\phi_{j_0,k}[n]_{k\in Z}$ and $\psi_{j,k}[n]_{(j,k)\in Z^2, j\geq j_0}$ are chosen to be orthogonal to each other in order for the convenience of obtaining the wavelet coefficients in the decomposition process, which means:

$$<\phi_{j_0,k}[n], \psi_{j,m}[n]> = \delta_{j_0,j}\delta_{k,m}. \tag{9}$$

In discrete wavelet decomposition, during the decomposition procedure, the initial step splits the original signal into two parts, approximation coefficients (i.e., $W_\phi[j_0, k]$) and detail coefficients (i.e., $W_\psi[j, k]$). After that, the following steps consist of recursively decomposing the approximation coefficients and detail coefficients into two new parts, respectively, using the same strategy as in initial step. This offers the richest analysis: the complete binary tree in the decomposition producer is produced as shown in Fig. 5:

The wavelet packet coefficients in each level can be computed using the following equations as:

$$W_\phi[j_0, k] = \frac{1}{\sqrt{M}}\sum_n f[n]\phi_{j_0,k}[n] \tag{10}$$

$$W_\psi[j, k] = \frac{1}{\sqrt{M}}\sum_n f[n]\psi_{j,k}[n], \quad j \geq j_0, \tag{11}$$

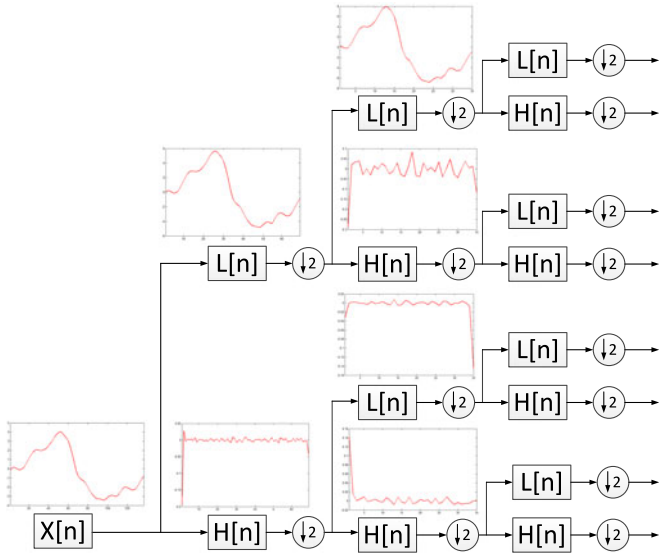


Fig. 5. Discrete wavelet packet transformation.

where $W_\phi[j_0, k]$ refers to the approximation coefficients while $W_\psi[j, k]$ represents the detailed coefficients respectively.

The efficacy of wavelet transform relies on choosing proper wavelet basis. One approach that aims at maximizing the discriminating ability of the discrete wavelet packet decomposition is applied, in which a class separability function is adopted [39]. We applied this method for all possible wavelets in the following families: Daubechies, Coiflets, Symlets, and got their class separability respectively. Based on their classification performance, a Symlet wavelet filter of order 4 is selected.

6 Lip Reading

The next step of WiHear is to recognize and translate the extracted signal features into words. To this end, WiHear detects the changes of pronouncing adjacent vowels and consonants by machine learning, and maps the patterns to words using automatic speech recognition. That is, WiHear builds a wireless-based provocation dictionary for automatic speech recognition system [22]. To make WiHear an automatic and real-time system, we need to address the following issues: segmentation, feature extraction and classification.

6.1 Segmentation

The segmentation process includes inner word segmentation and inter word segmentation.

For inner word segmentation, each word is divided into multiple phonetic events [31]. And WiHear then uses the training samples of pronouncing each syllable (e.g., sibilants and plosive sounds) to match the parts of the word and then using the syllables' combination to recognize the word.

For inter word segmentation, since there is usually a short interval (e.g., 300 ms) between pronouncing two successive words, WiHear detects the silent interval to separate words apart. Specifically, we first compute the finite difference (i.e., sample-to-sample difference) of the signal we obtained, which is referred as S_{dif}. Next we apply a sliding window to S_{dif} signal. Within each time slot, we compute the absolute mean value of signals in that window to determine whether this window is active or not, w.r.t, by comparing with a

dynamically computed threshold, we can determine whether the user is speaking within time period that the sliding window covers. In our experiments, the threshold is set to be 0.75 times the standard deviation of the differential signal across the whole process of pronouncing a certain word. This metric identifies the time slot when signal changes rapidly, indicating the process of pronouncing a word.

6.2 Feature Extraction

After signal segmentation, we can obtain wavelet profiles for different pronunciations, each with 16 fourth-order sub-waveforms from high frequency to low frequency components. To avoid the well-known "dimensionality curse" [27], We apply a Multi-Cluster/Class Feature Selection (MCFS) scheme [18] to extract representative features from wavelet profiles to reduce the quantity of sub-waveforms. Compared with other feature selection methods like SVM, MCFS can produce an optimal feature subset by considering possible correlations between different features, which better conforms to the characteristics of the dataset. We run the same dataset on SVM, which cost 3-5 minutes. The same dataset processing using MCFS only takes around 5 seconds. MCFS works as follows.

First, a m-nearest neighbor graph is constructed from the original dataset P. For each p_i, once its m nearest neighbors are determined, weighted edges are assigned between p_i and each of its neighbors, respectively. We define the weight matrix W for the edge connecting node i and node j as:

$$W_{i,j} = e^{-\frac{\|p_i - p_j\|^2}{\varepsilon}}. \tag{12}$$

Second, MCFS solves the following generalized eigen-problem:

$$Lv = \lambda Av, \tag{13}$$

where A is a diagonal matrix and $A_{ii} = \sum_j W_{ij}$. The graph Laplacian L is defined as $L = A - W$. And V is defined as $V = [v_1, v_2, ...v_K]$, in which all the v_k are the eigenvectors of Equation (13) with respect to the smallest eigenvalue.

Third, given v_k, a column of V, MCFS searches for a relevant feature subset by minimizing fitting errors:

$$\min_{\alpha_k} \|v_k - P^T \alpha_k\|^2 + \gamma|\alpha_k|, \tag{14}$$

where α_k is a M-dimensional vector and $|\alpha_k| = \sum_{j=1}^{M} |\alpha_{k,j}|$ represents the L_1-norm of α_k.

Finally, for every feature j, MCFS defines the MCFS score for the feature as:

$$Score(j) = \max_k |\alpha_{k,j}|, \tag{15}$$

where $\alpha_{k,j}$ is the jth element of vector α_k and all the features are sorted by their MCFS scores in descending order.

Fig. 6 shows the features selected by MCFS w.r.t. the mouth motion reflections in Fig. 1, which differ in each pronunciation.

6.3 Classification

For a specific individual, his speed and rhythm of speaking each word share similar patterns. We can thus directly compare the similarity of the current signals and previously sampled ones by generalized least squares [16].

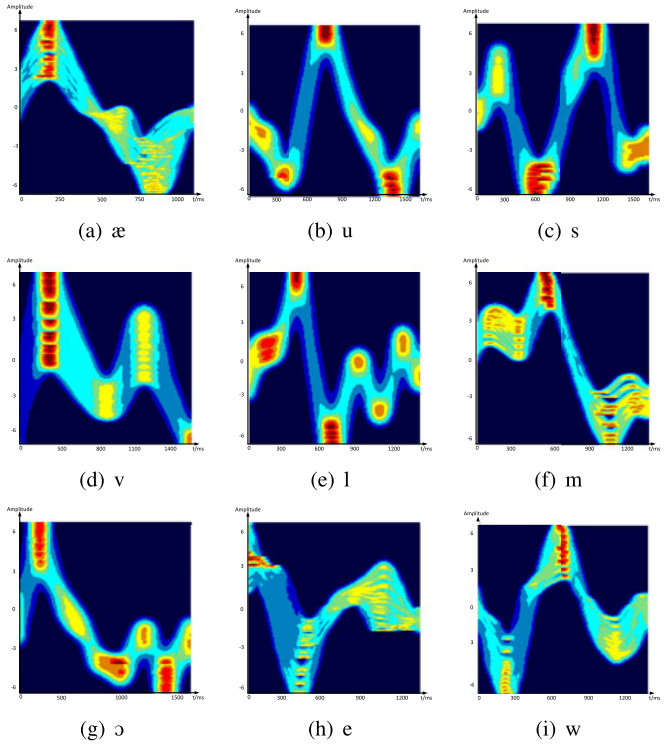


Fig. 6. Extracted features of pronouncing different vowels and consonants.

For scenarios where the user speaks at different speeds in a specific place, we can use dynamic time warping (DTW) [45] to classify the same word spoken at different speeds into the same group. DTW overcomes the local or global time series' shifts in time domain. It calculates intuitive distance between two time series waveforms. For more information, we recommend [40] which describes it in detail. Further, for people that share similar speaking patterns, we can also use DTW to enable word recognition with only one training individual.

For other unknown scenarios (e.g., different environments, etc.), due to the fine-grained analysis of wavelet transform, any small changes in the environment will lead to a single classifier with very poor performance [44]. Therefore instead of using a single classifier, we explore a more advanced machine learning scheme: Spectral Regression Discriminant Analysis (SRDA) [17]. SRDA is based on the popular Linear Discriminant Analysis (LDA) yet mitigates computational redundancy. We use this scheme to classify the test signals in order to recognize and match them into the corresponding mouth motions.

6.4 Context-Based Error Correction

So far we only explore direct word recognition with mouth motions. However, since the pronunciations spoken are correlated, we can leverage context-aware approaches widely used in automatic speech recognition [13] to improve recognition accuracy. As a toy example, when WiHear detects "you" and "ride", if the next word is "horse", WiHear can automatically distinguish and recognize "horse" instead of "house". Thus we can easily reduce the mistakes in recognizing words with similar mouth motion pattern, and further

improve recognition accuracy. Therefore, after applying machine learning for classification of signal reflections and mapping to their corresponding mouth motions, we use context-based error correction to further enhance our lip reading recognition.

7 Extending to Multiple Targets

For one conversation, it is common that only one person is talking at one time. Therefore it seems sufficient for WiHear to track one individual each time. To support debate and discussion, however, WiHear needs to be extended to track multiple talks simultaneously.

A natural approach is to leverage MIMO techniques. As shown in previous work [37], we can use spatial diversity to recognize multiple talks (often from different directions) at the receiver with multiple antennas. Here we also assume that people stay still while talking. To simultaneously track multiple users, we can first let each of them perform a unique pre-defined gesture (e.g., Person A repeatedly speaks [æ], Person B repeatedly speaks [h], etc.). Then we try to locate radio beams on them. The detailed beam locating process is illustrated in Section 4.1. After locating, WiHear's multi-antenna receiver can detect their talks simultaneously by leveraging spatial diversity in MIMO system.

However, due to additional power consumption of multiple RF links [32] and physical sizes of multiple antennas, we explore an alternative approach called *ZigZag cancelation* to support multiple talks with only one receiving antenna. The key insight is that, for most of the circumstances, multiple people do not begin pronouncing each word exactly at the same time. Therefore we can use *ZigZag cancelation*. After we recognize the first word of a user, we can predictably recognize the word he would like to say. Then in the middle of the first person speaking the first word, the second person speaks his first word. We can rely on the previous part of the first person part of first word, and use this information to predict the following part of his first word, and we can cancel the following part of first person speaking the first word and recognize the second person speaking. And we repeat the process back and forth. Thus we can achieve multiple hearing without deploying additional devices.

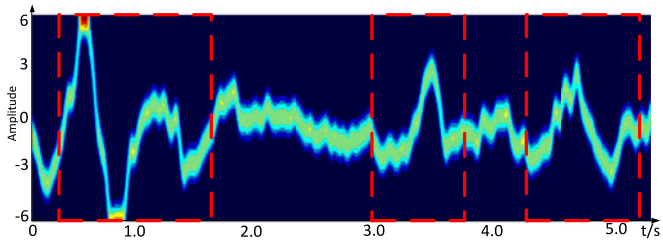
Figs. 7a and 7b depict the speaking of two users, respectively. After segmentation and classification, we can see each word as encompassed in the dashed red box. As is shown, three words from user1 have different starting and ending time compared with those of user2. Take the first word of the two users as an example, we can first recognize the beginning part of user1 speaking word1, and then use the predicted ending part of user1's word1 to cancel in the combined signals of user1 and user2's word1. Thus we use one antenna to simultaneously decode two users' words.

8 Implementation and Evaluation

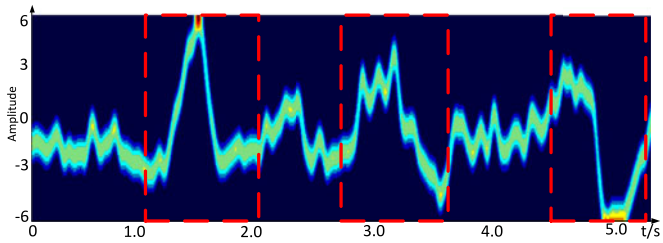
We implement WiHear on both commercial Wi-Fi infrastructure and USRP N210 [8], and evaluate its performance in typical indoor scenarios.

8.1 Hardware Testbed

We use a TP-LINK N750 serial, TL-WDR4300 type wireless router as the transmitter, and use a 3.20 GHz Intel(R) Pentium 4 CPU 2 GB RAM desktop equipped with Intel 5300



(a) Representive signals of user1 speaking



(b) Representive signals of user2 speaking

Fig. 7. Feature extraction of multiple human talks with ZigZag decoding on a single Rx antenna.

NIC (Network Interface Controller) as the receiver. As shown in the Fig. 10, the transmitter possesses directional antennas TL-ANT2406A [4] (beam width: Horizontal 120 degree, Vertical 90 degree) and operates in IEEE 802.11n AP mode at working at 2.4 GHz band. The receiver has three working antennas and the firmware is modified as in [28] to report original CSI to upper layers.

During the measurement campaign, the receiver continuously pings packets from the AP at the rate of 100 packets per second and we collect CSIs for 1 minute during each measurement. The collected CSIs are then stored and processed at the receiver.

For USRP implementation, we use GNURadio software platform [5], and implement WiHear into a 2×2 MU-MIMO system with 4 USRP N210 [8] boards and XCVR2450 daughterboards, which operate in the 2.4 GHz range. We use IEEE 802.11 OFDM standard [9], which has 64 sub-carriers (48 for data). We connect USRP N210 nodes via Gigabit Ethernet to our laboratory PCs, which are all equipped with a qual-core 3.2 GHz processor, 3.3 GB memory and running Ubuntu 10.04 with GNURadio software platform [5]. Since USRP N210 boards cannot support multiple daughter boards, we combine two USRP N210 nodes with an external clock [7] to build a two-antenna MIMO node. We use the other two USRP N210 nodes as clients.

8.2 Experimental Scenarios

We conduct the measurement campaign in a typical office environment and run our experiments with four people (one female and three males). We conduct measurements in a relatively open lobby area covering $9\,m \times 16\,m$ as Fig. 8. During our experiments, we always keep the distance between the radio and the user within roughly 2 m. To evaluate WiHear's ability to achieve LOS, NLOS and through-wall speech recognition, we extensively evaluate WiHear's performance in the following six scenarios (shown in Fig. 9).

- *Line of sight*. The target person is on the line of sight range between the transmitter and the receiver.

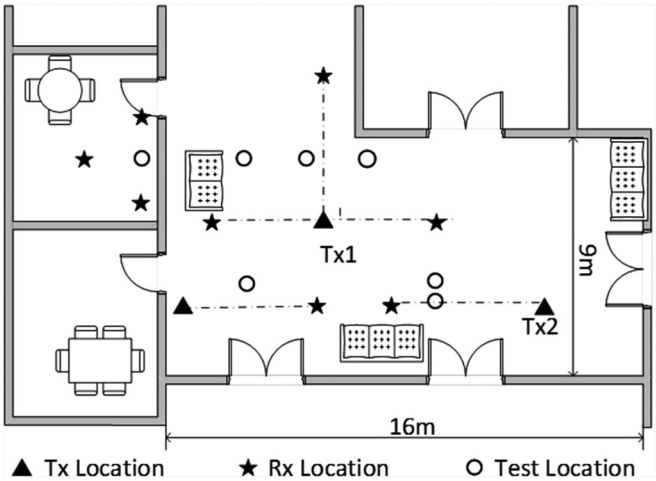


Fig. 8. Floor plan of the testing environment.

- *None line of sight.* The target person is not on the line of sight places, but within the radio range between the transmitter and the receiver.
- *Through wall Tx side.* The receiver and the transmitter are separated by a wall (roughly 6 inches). The target person is on the same side as the transmitter.
- *Through wall Rx side.* The receiver and the transmitter are separated by a wall (roughly 6 inches). The target person is on the same side as the receiver.
- *Multiple Rx.* One transmitter and multiple receivers are on the same side of a wall. The target person is within the range of these devices.
- *Multiple link pairs.* Multiple link pairs work simultaneously on multiple individuals.

Due to the high detection complexity of analyzing mouth motions, for practical issues, the following experiments are per-person trained and tested. Further, we tested two different types of directional antennas, namely, TL-ANT2406A and TENDA-D2407. With roughly the same location of users and link pairs, we found that WiHear does not need training per commercial Wi-Fi device. However, for devices that have huge differences like USRPs and commercial Wi-Fi devices, we recommend per device training and testing.

8.3 Lip Reading Vocabulary

As previously mentioned, lip reading can only recognize a subset of vocabulary [24]. WiHear can correctly classify and

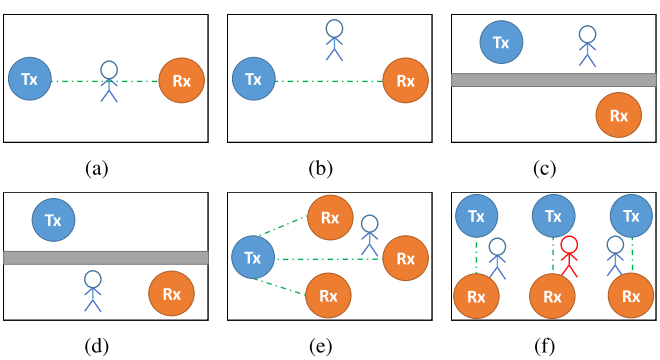


Fig. 9. Experimental scenarios layouts. (a) line-of-sight, (b) non-line-of-sight, (c) through wall Tx side, (d) through wall Rx side, (e) multiple Rx, and (f) multiple link pairs.

Fig. 10. The commercial hardware testbed.

recognize following syllables (vowels and consonants) and words.

Syllables. [æ], [e], [i], [u], [s], [l], [m], [h], [v], [ɔ], [w], [b], [j], [ʃ].

Words. see, good, how, are, you, fine, look, open, is, the, door, thank, boy, any, show, dog, bird, cat, zoo, yes, meet, some, watch, horse, sing, play, dance, lady, ride, today, like, he, she.

We note that it is unlikely any words or syllables can be recognized by WiHear. However, we believe the vocabulary of the above words and syllables are sufficient for simple commands and conversations. To further improve the recognition accuracy and extend the vocabulary, one can leverage techniques like Hidden Markov Models and Linear Predictive Coding [16], which is beyond the scope of this paper.

8.4 Automatic Segmentation Accuracy

We mainly focus on two aspects of segmentation accuracy in LOS and NLOS scenarios like Figs. 9a and 9b: inter word and inner word. Our tests consist of speaking sentences with varied quantity of words ranging from 3 to 6. For inner word segmentation, due to its higher complexity, we try to speak 4-9 syllables in one sentence. We test on both USRP N210 and commercial Wi-Fi devices. Based on our experimental results, we found that the performance for LOS (i.e., Fig. 9a) and NLOS (i.e., Fig. 9b) achieve similar accuracy. Given this, we average both LOS and NLOS performance as the final results. And Sections 7.5, 7.6, 7.7 follow the same rule.

Fig. 11 shows the inner-word and inter-word segmentation accuracy. The correct rate of inter-word segmentation is higher than that of inner-word segmentation. The main reason is that for inner-word segmentation, we directly use the waveform of each vowel or consonant to match the test waveform. Since different segmentation will lead to different combinations of vowels and consonants, even some of the combinations do not exist. In contrast, inter-word segmentation is relatively easy since it has a silent interval between two adjacent words.

When comparing between commercial devices and USRPs, we find the overall segmentation performance of commercial devices is a little better than USRPs. The key reason may be the number of antennas on the receiver. The receiver NIC card of commercial devices has 3 antennas whereas MIMO-based USRP N210 receiver only has two receiving antennas. Thus the commercial receiver may have richer information and spacial diversity than USRP N210's receiver.

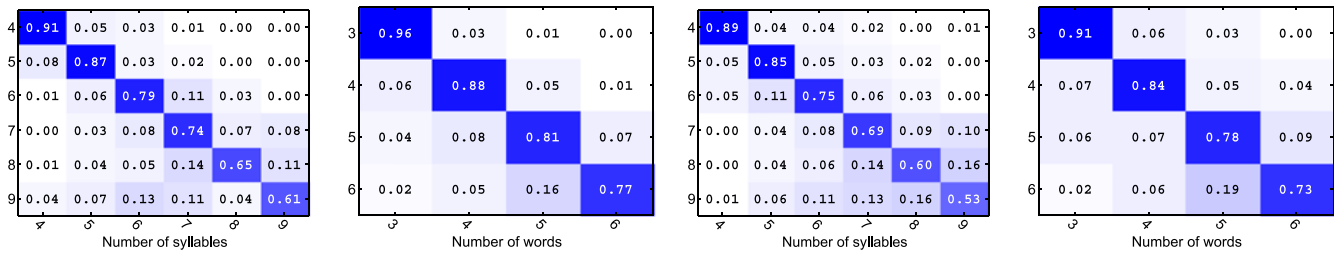


Fig. 11. Automatic segmentation accuracy for (a) inner-word segmentation on commercial devices, (b) inter-word segmentation on commercial devices, (c) inner-word segmentation on USRP, and (d) inter-word segmentation on USRP.

8.5 Classification Accuracy

Fig. 12 depicts the recognition accuracy on both USRP N210s and commercial Wi-Fi infrastructure in LOS (i.e., Fig. 9a) and NLOS (i.e., Fig. 9b). We also average the performance of LOS and NLOS for each kind devices. All the correctly segmented words are used for classification. We define the correct detection as correctly recognizing the whole sentence and we do not use context-based error correction here. As is shown in Fig. 12, the accuracy performance of commercial Wi-Fi infrastructure system achieves 91 percent on average with no more than 6-word sentences. In addition, with multiple receivers deployed, WiHear can achieve 91 percent on average with fewer than 10-word sentences, which is further discussed in Section 7.8.

Results show that the accuracy of commercial Wi-Fi infrastructure with directional antenna is much higher than that of USRP devices. The overall USRP accuracy performance for 6-word sentences is around 82 percent. The key reasons are two-folds: 1) the USRP N210 uses omni-directional antennas which may introduce more irrelevant multipath. 2) the receiver of commercial Wi-Fi product has one more antenna, which gives one more dimension of spacial diversity.

Since overall commercial Wi-Fi devices perform better than USRP N210, we mainly focus on commercial Wi-Fi devices in the following evaluations.

8.6 Training Overhead

WiHear requires a training process before recognizing human talks. We evaluate the training process in LOS and NLOS scenarios in Figs. 9a and 9b, and then average the performance. Fig. 13 shows the training overhead of WiHear. For each word or syllable, we present the quantity of training set and its corresponding recognition accuracy. As a whole, we can see that for each word or syllable, the accuracy of word-based is higher than syllable-based scheme. Given this result, empirically we choose the quantity of training sample ranging from 50 to 100, which has good recognition accuracy with acceptable training overhead.

However, the training overhead of word-based scheme is much larger than syllable-based one. Note that the amount of syllables in a language is limited, but the quantity of words is huge. We should make a trade off between syllable-based recognition and word-based recognition.

8.7 Impact of Context-Based Error Correction

We evaluate the importance of context-based error correction in LOS and NLOS scenarios as in Figs. 9a, 9b, and then average the performance. We compare WiHear's recognition accuracy with and without context-based error correction. We divide the quantity of words into three groups, namely fewer than three words (i.e., <3), 4 to 6 words (i.e., 4-6), more than six words but fewer than 10 words (i.e., $6<$). By testing different quantity of words in each group, we average the performance as the group's recognition accuracy. The following sections follow the same rule.

As shown in Fig. 14, without context-based error correction, the performance drops dramatically. Especially in the scenario of more than six words, context-based error correction achieves 13 percent performance gain than without it. This is because the longer the sentence, the more context information can be exploited for error correction.

Even with context-based error correction, the detection accuracy still tends to drop for longer sentences. The main problem is segmentation. For syllable-based technique, it is

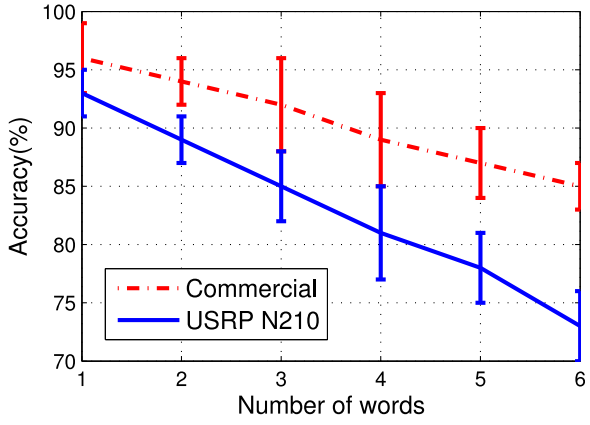


Fig. 12. Classification performance.

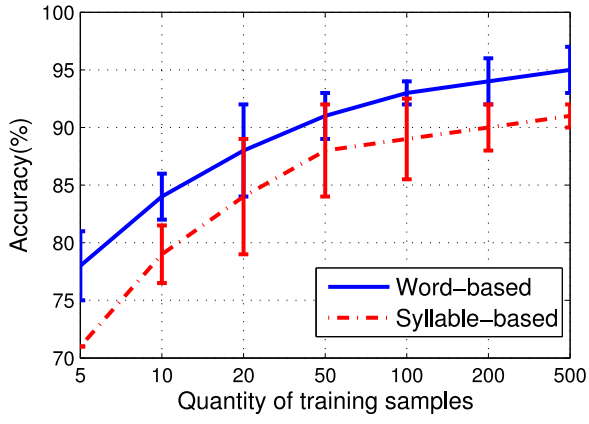


Fig. 13. Training overhead.

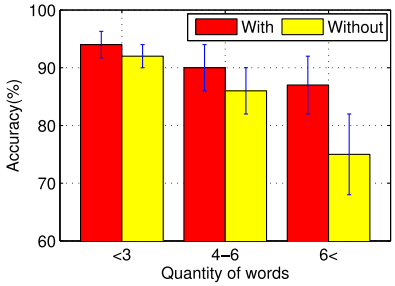


Fig. 14. With/without context-based error correction.

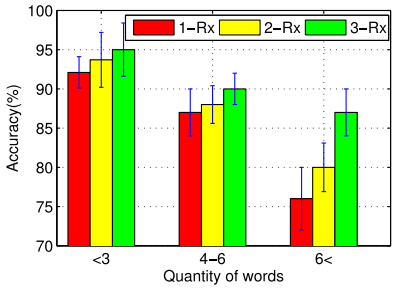


Fig. 15. Performance with Multiple Rx.

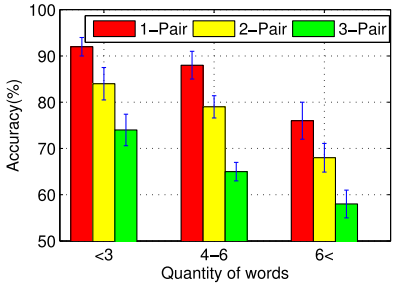


Fig. 16. Performance of multiple users with multiple link pairs.

obviously hard to segment the waveforms. For word-based technique, even though a short interval often exists between two successive words, the magnitudes of waveforms during these silent intervals are not strictly 0. Thus some of them may be regarded as part of the waveforms of some words. This may cause wrong segmentation of the words and decrease the detection accuracy. Thus the detection accuracy is dependent on the number of words. The performance in the following parts suffers from the same issue.

8.8 Performance with Multiple Receivers

Here we analyze radiometric impacts of human talks from different perspectives (i.e., scenarios like Fig. 9e). Specifically, to enhance recognition accuracy, we collect CSI from different receivers in multiple angle of views.

Based on our experiments, even though each NIC receiver has three antennas, the spatial diversity is not significant enough. In other words, the mouth motion's impacts on different links in one NIC are quite similar. This may be because the antennas are closely placed to each other. Thus we propose to use multiple receivers for better spatial diversity. As shown in Fig. 20, the same person pronouncing the word "GOOD" has different radiometric impacts on the received signals from different perspectives (from the angles of 0, 90 and 180 degree).

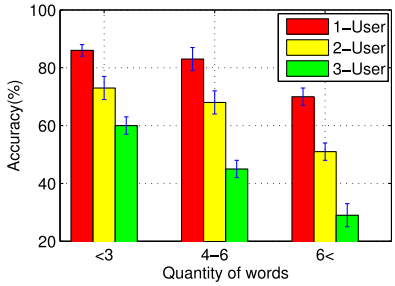


Fig. 17. Performance of zigzag decoding for multiple users.

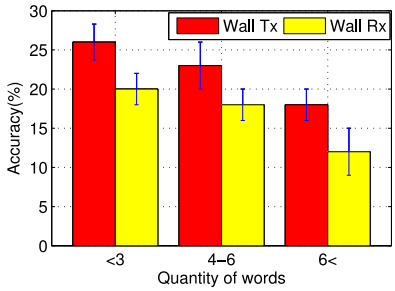


Fig. 18. Performance of two through wall scenarios.

With WiHear receiving signals in different perspectives, we can build up *Mouth Motion Profile* with these dimensions of different receivers. Thus it will enhance the performance and improve recognition accuracy. As depicted in Fig. 15, with multiple (3 in our case) dimensional training data, WiHear can achieve 87 percent accuracy even when the user speaks more than six words. It ensures the overall accuracy to be 91 percent in all three words' group scenarios. Given this, if it is needed for high accuracy of Wi-Fi hearing, we recommend to deploy more receivers from different views.

8.9 Performance for Multiple Targets

Here we present WiHear's performance for multiple targets. We use two and three pairs of transceivers to simultaneously target on two and three individuals, respectively (i.e., scenarios like Fig. 9f). As shown in Fig. 16, compared with a single target, the overall performance decreases with the number of targets increasing. Further, the performance drops dramatically when each user speaks more than six words. However, the overall performance is acceptable. The highest accuracy of three users' simultaneously talking less than three words is 74 percent. The worst situation can achieve nearly 60 percent accuracy with thre users speaking more than six words at the same time.

For *ZigZag cancelation* decoding, since NIC card [28] has three antennas, we enable only one antenna for our measurement. As depicted in Fig. 17, the performance drops more severely than that of multiple link pairs. The worst case (i.e., 3 users, 6 < words) only achieves less than 30 percent recognition accuracy. Thus we recommend to use *ZigZag cancelation* scheme with no more than two users who speak fewer than six words. Otherwise, we increase link pairs to ensure the overall performance.

8.10 Through Wall Performance

We tested two through wall scenarios, target on the Tx side (Fig. 9c) and on the Rx side (Fig. 9d). As shown in Fig. 18,

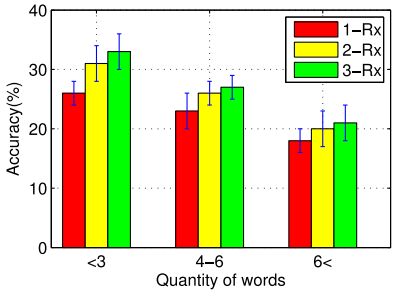


Fig. 19. Performance of through wall with multiple Rx.

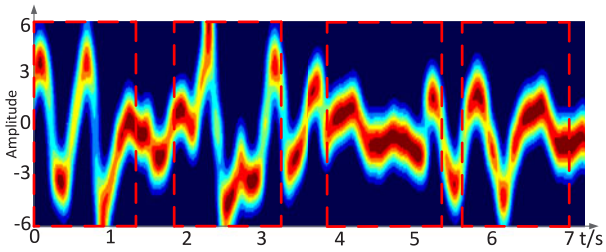
although recognition accuracy is pretty low (around 18 percent on average), compared with the probability of random guess (i.e., $1/33 = 3$ percent), the recognition accuracy is acceptable. Performance with target on the Tx side is better.

We believe by implementing interference nulling as in [11] can improve the performance, which unfortunately cannot be achieved with commercial Wi-Fi products. However, an alternative approach is to leverage spatial diversity with multiple receivers. As shown in Fig. 19, with two and three receivers, we can analyze signals from different perspectives with the target on the Tx side. Especially with three receivers, the maximum accuracy gain is 7 percent. With trained samples from different views, multiple receivers can enhance through wall performance.

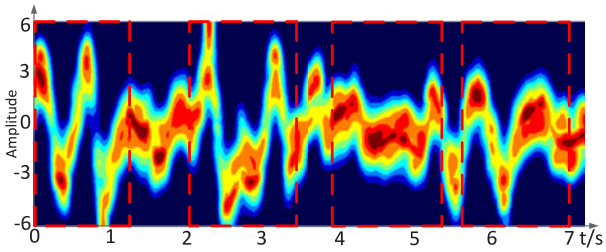
8.11 Resistance to Environmental Dynamics

We evaluate the influence of other ISM-band interference and irrelevant human movements on the detection accuracy of WiHear. We test these two kinds of interference in both LOS and NLOS scenarios as depicted in Figs. 9a and 9b. The resistance results of these two scenarios also share high similarity. Thus here we depict environmental effects on NLOS scenarios in Fig. 21.

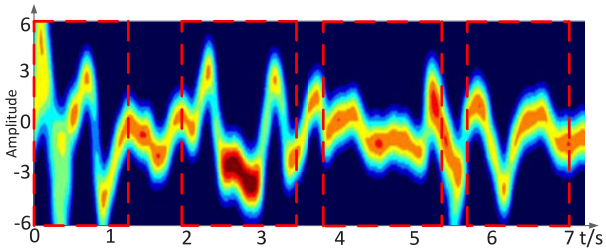
As shown in Fig. 21, one user repeatedly speaks a 4-word sentence. For each of the following three cases, we collect the radio sequences of speaking the repeated 4-word sentence for 30 times and draw the combined waveform in Fig. 21. For the first case, we remain the surroundings stable. With pretrained waveform of each word that the user speaks, as shown in Fig. 21a, we can easily recognize four words that user speaks. For the second case, we let three men randomly stroll in the room but always keep 3 m away from the WiHear's link pair. As shown in Fig. 21b, the words can still be correctly detected even though the waveform is loose compared with that in Fig. 21a. This loose character may be



(a) Waveform of a 4-word sentence without interference of ISM band signals or irrelevant human motions



(b) Impact of irrelevant human movements interference

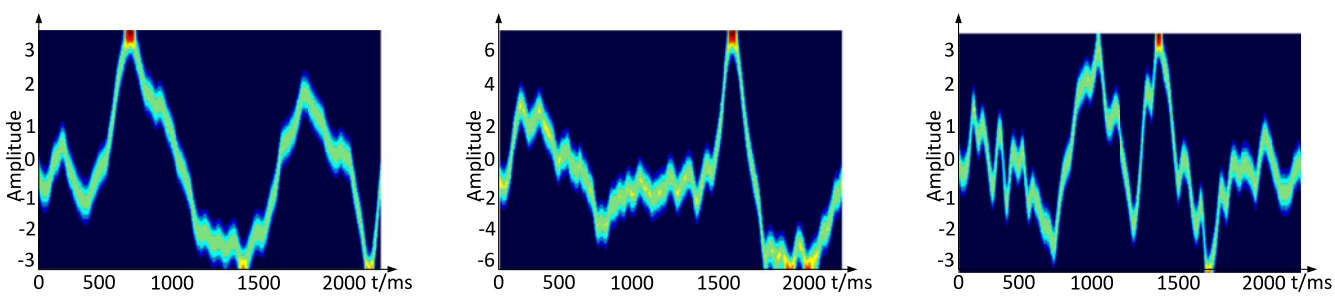


(c) Impact of ISM band interference

Fig. 21. Illustration of WiHear's resistance to environmental dynamics.

the effect of irrelevant human motions. For the third case, we use a mobile phone to communicate with an AP (e.g., surfing online) and keep them 3 m away from WiHear's link pair. As shown in Fig. 21c, the generated waveform fluctuates a little compared with that in Fig. 21a. This fluctuation may be the effect of ISM band interference.

Based on above results, we can conclude that WiHear can be resistant to ISM band interference and irrelevant human motions 3 m away without significant recognition performance degradation. For interference within 3 m around transceivers, the interference sometimes dominates the WiHear signal fluctuation. Therefore, the performance is unacceptable (usually around 5 percent accuracy) and we leave it as one of our future works.



(a) GOOD 0° (b) GOOD 90° (c) GOOD 180°

Fig. 20. Example of different views for pronouncing words.

9 Discussion

So far we assume people do not move when they speak. It is possible that a person talks while walking. We believe the combination of device-free localization techniques [49] and WiHear would enable real-time tracking and continuous hearing. We leave it as a future direction.

Generally, people share similar mouth movements when pronouncing the same syllables or words. Given this, we may achieve Wi-Fi hearing via DTW (details in Section 5.3) with training data from one person, and testing on another individual. We leave it as part of the future work.

Due to the longer distance between the target person and the directional antenna, the larger noise and interference occurs. For long range Wi-Fi hearing, we recommend grid parabolic antennas like TL-ANT2424B [6] to accurately locate the target for better performance.

To support real-time processing, we can only use CSI on one subchannel to reduce the computational complexity. Since we found the radiometric impact of mouth motions is similar across subchannels, we may safely select one representative subchannel without sacrificing much performance. However, the full potential of the whole CSI information is still under-explored.

10 Concluding Remarks

This paper presents WiHear, a novel system that enables Wi-Fi signals to hear talks. WiHear is compatible with existing Wi-Fi standards and can be extended easily to commercial Wi-Fi products. To achieve lip reading, WiHear introduces a novel system for sensing and recognizing micro-motions (e.g., mouth movements). WiHear consists of two key components, *mouth motion profile* for extracting features, and learning-based signal analysis for lip reading. Further, *Mouth motion profile* is the first effort that leverage *partial* multipath effects to get the whole mouth motions' impacts on radio. Extensive experiments demonstrate that, with correct segmentation, WiHear can achieve recognition accuracy of 91 percent for single user speaking no more than 6 words and up to 74 percent for hearing no more than three users simultaneously.

WiHear may have many application scenarios. Since Wi-Fi signals do not require LOS, even though experimental results are not promising, we believe WiHear has the potential to "hear" people talks through walls and doors within the radio range. In addition, WiHear can "understand" people talking, which can get more complicated information from talks than gesture-based interfaces like Xbox Kinect [2] (e.g., mood). Further, WiHear can also help disabled people to conduct simple commands to devices with mouth movements instead of inconvenient body gestures. We can also extend WiHear for motion detection on hands. Since WiHear can be easily extended into commercial products, we envision it as a practical solution for Wi-Fi hearing in real-world deployment.

Acknowledgments

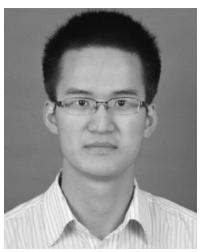
This research is supported in part by Guangdong Natural Science Funds for Distinguished Young Scholar (No. S20120011468), the Shenzhen Science and Technology Foundation (No. JCYJ20140509172719309, KQCX20150324 160536457), China NSFC Grant 61472259, Guangdong Young Talent Project 2014TQ01X238, Hong Kong RGC Grant HKUST16207714, and GDUPS (2015). Kaishun Wu is the corresponding author.

References

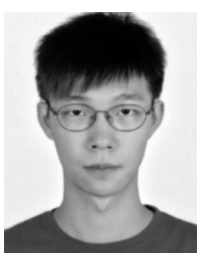
[1] (2012). Leap Motion [Online]. Available: https://www.leapmotion.com/
[2] (2013). Xbox Kinect [Online]. Available: http://www.xbox.com/en-US/kinect
[3] (2013). Vicon [Online]. Available: http://www.vicon.com
[4] (2014). TP-LINK 2.4 GHz 6dBi Indoor Directional Antenna [Online]. Available: http://www.tp-link.com/en/products/details/?categoryid=2473&model=TL-ANT2406A#over
[5] (2012). GNU software defined radio [Online]. Available: http://www.gnu.org/software/gnuradio
[6] (2014). TP-LINK 2.4 GHz 24dBi Grid Parabolic Antenna [Online]. Available: http://www.tp-link.com/en/products/details/?categoryid=2474&model=TL-ANT2424B#over
[7] (2012). *Oscilloquartz SA, OSA 5200B GPS Clock* [Online]. Available: http: //www.oscilloquartz.com
[8] (2012). *Universal Software Radio Peripheral*, Ettus Research LLC [Online]. Available: http://www.ettus.com
[9] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Std 802.11, 2012.
[10] F. Adib, Z. Kabelac, D. Katabi, and R. C. Miller, "3d tracking via body radio reflections," in *Proc. 11th USENIX Conf. Netw. Syst. Des. Implementation*, 2014, pp. 317–329.
[11] F. Adib and D. Katabi, "See through walls with Wi-Fi!," in *Proc. ACM SIGCOMM*, 2013, pp. 75–86.
[12] S. Agrawal, I. Constandache, S. Gaonkar, R. R. Choudhury, K. Cave, and F. DeRuyter, "Using mobile phones to write in air," in *Proc. 9th Int. Conf. Mobile Syst., Appl. Serv.*, 2011, pp. 15–28.
[13] M. Aynehband, A. M. Rahmani, and S. Setayeshi, "Coast: Context-aware pervasive speech recognition system," in *Proc. IEEE Int. Symp. Wireless Pervasive Comput.*, 2011, pp. 1–4.
[14] D. Bharadia, K. R. Joshi, and S. Katti, "Full duplex backscatter," in *Proc. 12th ACM Workshop Hot Topics Netw.*, 2013, pp. 4:1– 4:7.
[15] C. Bo, X. Jian, X.-Y. Li, X. Mao, Y. Wang, and F. Li, "You're driving and texting: Detecting drivers using personal smart phones by leveraging inertial sensors," in *Proc. 19th Annu. Int. Conf. Mobile Comput. Netw.*, 2013, pp. 199–202.
[16] J. Bradbury, "Linear predictive coding," *Mc G. Hill*, 2000.
[17] D. Cai, X. He, and J. Han, "SRDA: An efficient algorithm for large-scale discriminant analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 1, pp. 1–12, Jan. 2008.
[18] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 333–342.
[19] G. L. Charvat, L. C. Kempel, E. J. Rothwell, C. M. Coleman, and E. L. Mokole, "A through-dielectric radar imaging system," *IEEE Trans. Antennas Propagation*, vol. 58, no. 8, pp. 2594–2603, Aug. 2010.
[20] L. J. Chu, "Physical limitations of omni-directional antennas," *J. Appl. Phys.*, vol. 19, pp. 1163–1175, 1948.
[21] G. Cohn, D. Morris, S. N. Patel, and D. S. Tan, "Humantenna: Using the body as an antenna for real-time whole-body interaction," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2012, pp. 1901–1910.
[22] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, pp. 267–285, 2001.
[23] A. Davis, M. Rubinstein, N. Wadhwa, G. Mysore, F. Durand, and W. T. Freeman, "The visual microphone: Passive recovery of sound from video," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 79:1–79:10, 2014.
[24] B. Dodd and R. Campbell, *Hearing by Eye: The Psychology of Lip-Reading*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 1987.
[25] P. Duchnowski, M. Hunke, D. Busching, U. Meier, and A. Waibel, "Toward movement-invariant automatic lip-reading and speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1995, pp. 109–112.
[26] P. Duchnowski, U. Meier, and A. Waibel, "See me, hear me: Integrating automatic speech recognition and lip-reading," in *Proc. Int. Conf. Spoken Lang. Process*, 1994, pp. 547–550.

[27] I. A. Gheyas and L. S. Smith, "Feature subset selection in large dimensionality domains," *Pattern Recog.*, vol. 43, pp. 5–13, 2010.
[28] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Predictable 802.11 packet delivery from wireless channel measurements," in *Proc. ACM SIGCOMM Conf.*, 2010, pp. 159–170.
[29] C. Harrison, D. Tan, and D. Morris, "Skinput: Appropriating the body as an input surface," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2010, pp. 453–462.
[30] Y. Jin, W. Seng Soh, and W. Choong Wong, "Indoor localization with channel impulse response based fingerprint and nonparametric regression," *IEEE Trans. Wireless Commun.*, vol. 9, no. 3, pp. 1120–1127, Mar. 2010.
[31] H. Junker, P. Lukowicz, and G. Troster, "On the automatic segmentation of speech signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1987, pp. 77–80.
[32] B. Kellogg, V. Talla, and S. Gollakota, "Bringing gesture recognition to all devices," in *Proc. 11th USENIX Conf. Netw. Syst. Des. Implementation*, 2014, pp. 303–316.
[33] K. Kumar, T. Chen, and R. M. Sternl, "Profile view lip reading," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2007, pp. 429–432.
[34] E. Larson, G. Cohn, S. Gupta, X. Ren, B. Harrison, D. Fox, and S. N. Patel, "Heatwave: Thermal imaging for surface user interaction," in *Proc. Conf. Human Factors Comput. Syst.*, 2011, pp. 2565–2574.
[35] K. Ching-Ju Lin, S. Gollakota, and D. Katabi, "Random access heterogeneous MIMO networks," in *Proc. ACM SIGCOMM*, 2011, pp. 146–157.
[36] G. C. Martin. (2014). Preston blair phoneme series [Online]. Available: http://www.garycmartin.com/mouth_shapes.html
[37] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in *Proc. 19th Annu. Int. Conf. Mobile Comput. Netw.*, 2013, pp. 27–38.
[38] Theodore Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 2001.
[39] N. Saito and R. Coifman, "Local discriminant bases and their applications," *J. Math. Imaging Vis.*, vol. 5, no. 4, pp. 337–358, 1995.
[40] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intell. Data Anal.*, vol. 11, no. 5, Oct. 2007.
[41] M. Scholz, S. Sigg, H. R. Schmidtke, and M. Beigl, "Challenges for device-free radio-based activity recognition," in *Proc. Workshop Context Syst, Des., Eval. Optim.*, 2011, pp. 3:1–3:12.
[42] S. Sen, J. Lee, K.-H. Kim, and P. Congdon, "Avoiding multipath to revive inbuilding Wi-Fi localization," in *Proc. 11th Annu. Int. Conf. Mobile Syst., Appl. Serv.*, 2013, pp. 249–262.
[43] S. Sen, B. Radunovic, R. R. Choudhury, and T. Minka, "You are Facing the Mona Lisa: Spot localization using PHY layer information," in *Proc. ACM Int. Conf. Mobile Syst., Appl. Serv.*, 2012, pp. 183–196.
[44] M. Skurichina and R. P. W. Duin, "Bagging, boosting and the random subspace method for linear classifiers," *Pattern Anal. Appl.*, vol. 5, no. 2, pp. 121–135, 2002.
[45] J. Wang and D. Katabi, "Dude, where's my card? RFID positioning that works with multipath and non-line of sight," in *Proc. ACM SIGCOMM*, 2013, pp. 51–62.
[46] J. R. Williams, "Guidelines for the use of multimedia in instruction," in *Proc. 42nd Annu. Meeting Human Factors Ergonomics Soc.*, 1998, pp. 1447–1451.
[47] J. Wilson and N. Patwari, "Radio tomographic imaging with wireless networks," *IEEE Trans. Mobile Comput.*, vol. 9, no. 5, pp. 621–632, May 2010.
[48] J. Xiao, K. Wu, Y. Yi, L. Wang, and L. M. Ni, "FIMD: Fine-grained device-free motion detection," in *Proc. IEEE Int. Conf. Parallel Distrib. Syst.*, 2012, pp. 229–235.
[49] J. Xiao, K. Wu, Y. Yi, L. Wang, and L. M. Ni, "Pilot: Passive device-free indoor localization using channel state information," in *Proc. IEEE 33rd Int. Conf. Distrib. Comput. Syst.*, 2013, pp. 236–245.
[50] Z. Yang, Z. Zhou, and Y. Liu, "From RSSI to CSI: Indoor localization via channel response," *ACM Comput. Surveys*, vol. 46, no. 2, pp. 25:1–25:32, 2013.
[51] M. Youssef, M. Mah, and A. Agrawala, ," Challenges: Device-free passive localization for wireless environments," in *Proc. 13th Annu. ACM Int. Conf. Mobile Comput. Netw.*, 2007, pp. 222–229.
[52] D. Zhang, J. Zhou, M. Guo, J. Cao, and T. Li, "TASA: Tag-free activity sensing using RFID tag arrays," *IEEE Trans. Parallel Distrib. Syst.*, vol. 22, no. 4, pp. 558–570, Apr. 2011.
[53] J. Zhang, M. H. Firooz, N. Patwari, and S. K. Kasera, "Advancing wireless link signatures for location distinction," in *Proc. ACM Int. Conf. Mobile Comput. Netw.*, 2008, pp. 26–37.
[54] W. Zhang, X. Zhou, L. Yang, Z. Zhang, B. Y. Zhao, and H. Zheng, "3D beamforming for wireless data centers," in *Proc. ACM 10th ACM Workshop Hot Topics Netw.*, 2011, pp. 4:1–4:6.
[55] X. Zhou, Z. Zhang, Y. Zhu, Y. Li, S. Kumar, A. Vahdat, B. Zhao, and H. Zheng, "Mirror mirror on the ceiling: Flexible wireless links for data centers," in *Proc. ACM SIGCOMM*, 2012, pp. 443–454.

Guanhua Wang received the BEng degree in computer science from Southeast University, China, in 2012, the MPhil degree in computer science and engineering, from the Hong Kong University of Science and Technology in 2015, advised by Prof. Lionel M. Ni, and is a first year computer science PhD student in the AMPLab, at UC Berkeley, advised by Prof. Ion Stoica.. His main research interests include big data and networking. He is a student member of the IEEE.



Yongpan Zou received the BEng degree of chemical machinery from Xi'an Jiaotong University, Xi'an, China. Since 2013, he is working toward the PhD degree in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology (HKUST). His current research interests mainly include: wearable/mobile computing and wireless communication. He is a student member of the IEEE.



Zimu Zhou received the BE degree in 2011 from the Department of Electronic Engineering at Tsinghua University, Beijing, China. He is currently working toward the PhD degree in the Department of Computer Science and Engineering, Hong Kong University of Science and Technology. His main research interests include wireless networks and mobile computing. He is a student member of the IEEE and ACM.

Kaishun Wu received the PhD degree in computer science and engineering from HKUST in 2011. He is currently a distinguish professor at Shenzhen University. Previously, he was a research assistant professor in the Fok Ying Tung Graduate School at the Hong Kong University of Science and Technology (HKUST). He received the Hong Kong Young Scientist Award in 2012. His research interests include wireless communication, mobile computing, wireless sensor networks, and data center networks. He is a member of the IEEE.

Lionel M. Ni received the PhD degree in electrical and computer engineering from Purdue University in 1980. He is a chair professor in the Department of Computer and Information Science and a vice rector of academic affairs at the University of Macau. Previously, he was a chair professor of computer science and engineering at the Hong Kong University of Science and Technology. He has chaired more than 30 professional conferences and has received eight awards for authoring outstanding papers. He is a fellow of the IEEE and Hong Kong Academy of Engineering Science.