Context-Free Fine-Grained Motion Sensing using WiFi

Changlai Du*, Xiaoqun Yuan*†, Wenjing Lou*, Y. Thomas Hou*
*Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
†School of Information Management, Wuhan University, Wuhan, Hubei, China 430070

***Abstract*—WiFi-based motion sensing has received a lot of research attention in recent years. Taking advantage of Channel State Information(CSI) collected from physical layer, previous techniques are able to extract useful information from CSI values to infer human movements. However, these works concentrate either on coarse-grained motion sensing or on fine-grained but context-related motion sensing. In this paper, we propose WiTalk, a new fine-grained human motion sensing technique with the distinct context-free character. To profile human motion using CSI, WiTalk generates CSI spectrograms using signal processing techniques and extracts features by calculating the contours of the CSI spectrograms. We verify the proposed technique in the application scenario of lip reading, where the fine-grained motion is the mouth movements. We implement WiTalk on a commercial laptop. Experiment results show that WiTalk can achieve over 92.3% recognition accuracy to discern a set of 12 syllables and 74.3% accuracy to discern a set of short sentences up to six words.**

I. Introduction

WiFi-based motion sensing has received a lot of research attention in recent years, leveraging the fact that human motion will change the channel states between transceivers. By monitoring these channel state changes, researchers are able to extract useful information to infer human motion. Channel State Information(CSI) is one of the most popular measurements for the purpose of motion sensing because it provides more fine-grained channel information than Received Signal Strength Index(RSSI). CSI is the time series of the channel frequency responses(CFR) which can be collected from physical layer of off-the-shelf WiFi devices thanks to the previously released CSI collecting tools [1], [2].

Previous work on human motion sensing has made great effort focusing on several application scenarios like human localization [9], [10], activity detection and recognition [11], [12], human authentication [13], [14], health care [15], [16] and fine-grained motion sensing [17]–[20], [23]. Using a quadrant classification method, we position these previous research works and the proposed work in this paper in Fig. 1. The four quadrants are determined by whether the motion to be detected is coarse-grained or fine-grained, and by whether the detection is context-free or not. For example, E-eyes [12] can recognize human activities by comparing the testing CSI measurements to a set of CSI profiles. The CSI profiles constructed in time domain are not the same in different contexts such as at different locations. E-eyes needs to set up a profile group for a single human activity. We classify E-eyes into quadrant III as a coarse-grained context-related solution. On the other hand, CARM [11], WifiU [13] extract features from CSI spectrograms in time-frequency domain. CSI spectrogram is proved to be intrinsically correlated to the moving speed of different human body parts but not correlated to contexts. We classify them into quadrant II as coarse-grained context-free solutions. The work proposed by Liu *et al.* [16] makes use of channel information in both time and frequency domain to capture human breathing rate and heart rate. These scalar values of vital signs are estimated in frequency domain and are uncorrelated to contexts. Thus we also classify it into quadrant II.

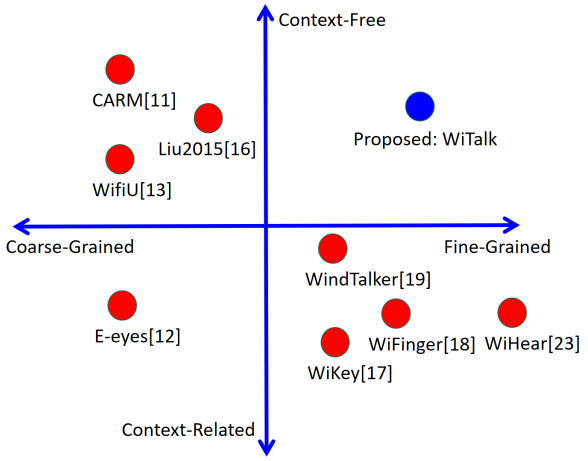


Fig. 1: Research Position: Interpretation of the Four Quadrants

In the right half plane of Fig. 1, existing fine-grained motion sensing solutions construct the CSI profiles in time domain. Time domain CSI profiles are subject to changes of contexts, including changes of locations, users or multipath environments. For example, WiKey [17] recognizes keystrokes based on *CSI-waveform* for each key. WindTalker [19] inferences mobile device keystrokes exploiting the strong correlation between the CSI fluctuation and the keystrokes. WiFinger [18] senses and identifies subtle movements of finger gestures by examining the unique patterns in CSI. WiHear [23] detects and analyzes fine-grained radio reflections from mouth movements by introducing Mouth Motion Profile. These solutions are all context-related. They require the construction and testing of the CSI profiles in the same contexts. For a different user, different location, or a different multipath environment, the profiles need to be reconstructed. We classify them into

quadrant IV as fine-grained context-related solutions.

To design a context-free fine-grained motion sensing solution in quadrant I, some specific *challenges* that differ from previous research settings must be addressed. First, WiFi signal reflections from fine-grained human motion are very tiny, much smaller than those from large-scale human movements. CSI dynamics caused by fine-grained human motion are easily buried in noise and interferences. To ensure the CSI dynamics to be detectable, previous solutions make some assumptions or use special tools. For example WiKey [17] assumes that the tested motion takes place near one end of the transceivers. WiFinger [18] assumes that the tested motion takes place near the line of sight(LOS) between the transceivers. WiHear [23] and WindTalker [19] use special purpose directional antennas to enhance signal-to-noise ratio(SNR) in CSi dynamics. Second, effective denoise methods must be adopted to reduce noises and interferences and obtain a clean CSI waveform that reflects the motion to be detected. Third, to design a context-free solution, intrinsic properties in CSI dynamics that are only correlated to fine-grained motion to be detected must be identified. To effectively detect the fine-grained motion, feasible features must also be carefully identified and selected.

In this paper, we present WiTalk, the first context-free fine-grained motion sensing system using WiFi physical layer channel information. Similar to previous CSI-based motion sensing solutions, WiTalk infers human motion by analyzing the CSI dynamics. To effectively denoise CSI streams, we use principal component analysis(PCA) filtering methods based on the observation that signal fluctuations on all subcarriers are correlated. To address the context-free challenge, we identify CSI spectrogram in time-frequency domain as the stable property in CSI dynamics and extract features from CSI spectrograms by calculating the contours of the spectrograms.

We verify the feasibility and performance of WiTalk in the application scenario of lip reading. To ensure that the CSI dynamics generated by mouth movements are detectable, we make similar assumptions as in previous solutions. Specifically, we assume that mouth movements take place near one end of the transceivers similar to [17], based on the observation that people tend to hold the phone close to the cheek while talking over the phone. Directional antennas can also be used to further amplify CSI dynamics and eliminate CSI noises. We leave this as future research work.

The main contributions of WiTalk are summarized as follows:

- To the best of our knowledge, WiTalk is the first feasible system in the context-free fine-grained quadrant of motion sensing solution plane using WiFi CSI dynamics. We show the existence of this quadrant I solution by identifying the CSI spectrograms as the intrinsic stable properties that correlate to fine-grained human motion.
- We identify and extract effective features from CSI spectrograms by calculating the contours of CSI spectrograms. These new discerning features solve the problem of low time-frequency resolution using discrete wavelet transform(DWT).
- We verify the feasibility of WiTalk by applying it to the lip reading scenario. Experiment results show that WiTalk achieves comparable results to previous fine-grained context-related solutions.

We implement WiTalk on a commercial laptop and demonstrate its feasibility through experiments. The performance is evaluated under various contexts with different transceivers distances, different locations and users. The results show that WiTalk can achieve over 92.3% recognition accuracy to discern a set of 12 syllables and 74.3% accuracy to discern a set of short sentences up to six words.

The rest of the paper is organized as follows. We introduce the technical background in Section II. The system design is detailed in Section III. The performance of WiTalk is verified in Section IV under the lip reading scenario. We discuss related work in Section V and conclude the paper in Section VI.

II. Background

CSI-based motion sensing researches rely on the same principle that the change of CSI values has correlation with the motion to be detected. In this section, we briefly introduce the CSI related backgrounds, especially the CSI-speed model proposed in [11].

In WiFi protocols like IEEE 802.11a/n/ac, Orthogonal Frequency Division Multiplexing(OFDM) is adopted as the modulation format. In OFDM, the channel frequency response(CFR) is measured on subcarrier level. Let $X(f,t)$ and $Y(f,t)$ be the transmitted and received signals in frequency domain respectively, then we have $Y(f,t) = H(f,t)X(f,t)$, where $H(f,t)$ is the CFR at frequency f and time t. As the FFT/IFFT operations are integrated in OFDM receivers, the receivers are ready to calculate CFRs. Taking advantage of the released tools [1], [2], CFR values are revealed from NIC firmware to drivers and then to upper layers in the format of CSI. In IEEE 802.11 standards, CFRs on 30 selected subcarriers are reported for every received 802.11 frame. If a WiFi link has N_{tx} and N_{rx} of emitting and receiving antennas respectively, then the reported CFRs form a *CSI matrix* in dimensions of $30 \times N_{tx} \times N_{rx}$. A CSI matrix is instantaneous. For a specific subcarrier on an antenna pair, we name the time-series of CFRs a CSI stream. Then the time-series of the CSI matrix contains $30 \times N_{tx} \times N_{rx}$ CSI streams. We can see that CSI characterizes the frequency response of the wireless channels.

In indoor environments, wireless signals arrive at a receiver antenna through multiple pathes including the LOS path, pathes reflected by static objects like walls and pathes reflected by moving objects like human body. The signals transmitted through these pathes have different amplitudes and phases. The CFR can be modeled as the sum of static and dynamic components [11]:

$$H(f,t) = e^{-j2\pi\Delta f t}(H_s(f) + H_d(f,t)) \quad (1)$$

where Δf is the carrier frequency difference between the sender and the receiver, $H_s(f)$ is the sum of static CFRs and

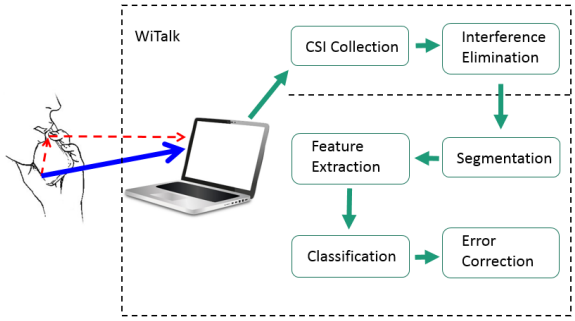
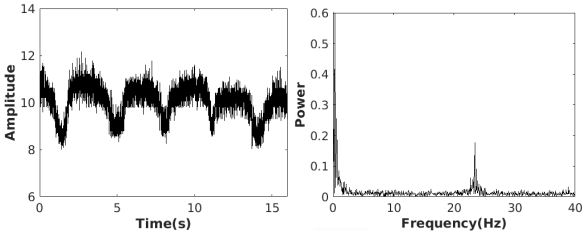


Fig. 2: WiTalk System Design and Workflow



(a) Original CSI Stream (b) Spectrum

Fig. 3: CSI Stream of Breathing and Corresponding Spectrum

$H_d(f,t)$ is the sum of dynamic CFRs:

$$H_d(f,t) = \sum_{k \in \mathscr{P}_d} a_k(f,t) e^{-j2\pi \frac{d_k(t)}{\lambda}} \tag{2}$$

where $\mathscr{P}_d$ is the set of dynamic pathes, a_k is the attenuation and initial phase of k^{th} path and $d_k(t)$ is the length of path k at time t.

The power of the CFR can be calculated to eliminate Δf:

$$\begin{aligned} |H(f,t)|^2 = &\sum_{k \in \mathscr{P}_d} 2|H_s(f)a_k(f,t)|cos\Big(\frac{2\pi v_k t}{\lambda} + \phi_k(0)\Big) \\ &+ \sum_{\substack{k,l \in \mathscr{P}_d \\ k \neq l}} 2|a_k(f,t)a_l(f,t)|cos\Big(\frac{2\pi(v_k - v_l)t)}{\lambda} + \phi_{k,l}(0)\Big) \\ &+ \sum_{k \in \mathscr{P}_d} |a_k(f,t)|^2 + |H_s(f)|^2 \end{aligned} \tag{3}$$

where $\phi_k(0)$ and $\phi_{k,l}(0)$ are initial phase and phase difference. The total CFR power is the sum of a constant offset and a set of sinusoids. The frequencies are functions of the speeds of path length changes, which is further correlated to human movement speed. This CSI-speed model is the technical principle of using CSI dynamics to detect human motion.

III. System Design

The design of WiTalk is illustrated in Fig. 2. It consists of the following components: CSI data collection and preprocessing, interference elimination, segmentation, feature extraction, classification and error correction. In this paper we mainly focus on three components: interference elimination, feature extraction and classification. We will briefly introduce CSI data collection and preprocessing, segmentation and error correction components because they are not the core contributions of this paper.

A. CSI Data Collection and Preprocessing

WiTalk is implemented on the receiving end of a WiFi link and collects CSI measurements on each received packets. For each pair of sending and receiving antennas, 30 CSI streams are collected.

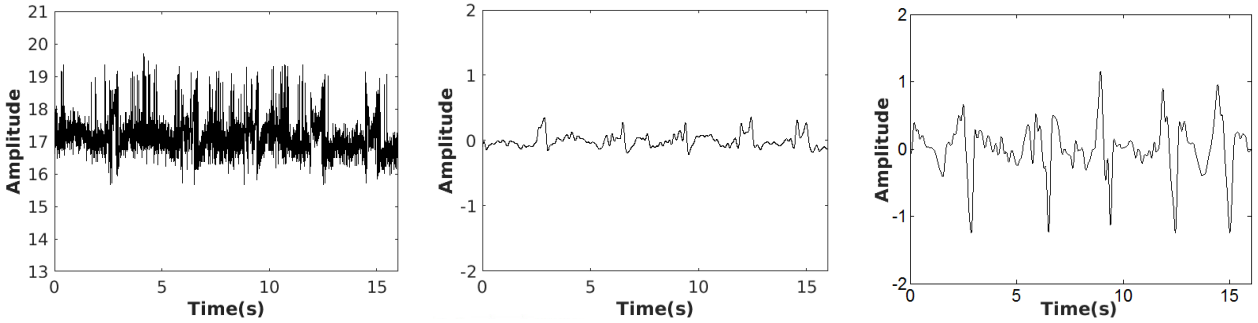
CSI streams are firstly normalized to obtain its z score as $Z = (Y - m)/s$, where m and s are mean and standard deviation vector respectively. After normalization, Z has a mean of 0 and a standard deviation of 1. The reasons why we normalize CSI streams are two-folds. WiTalk uses CSI spectrograms in time-frequency domain, where the amplitude of CSI streams does not affect our analysis. Besides, CSI normalization helps in the PCA based filtering step because after normalization, all CSI streams contribute equally to PCA and none of them will dominate the PCA results.

B. Interference Elimination

CSI streams reported from WiFi NICs are very noisy. Fig. 3a shows one original CSI stream collected at a sampling rate of $250Hz$. The noise sources include environmental noises and that are caused by WiFi NICs internal state transitions. These noises are in the high frequency zone on the spectrum. Besides the high frequency noises, some interferences also exit in CSI streams. Typical interferences include reflected signals from surrounding moving people and the movement of other body parts of the target like chest movements when breathing. In this section, we first use a Butterworth band pass filter to denoise the CSI streams and analyze its parameters. PCA based filter is then applied taking advantage of the correlation among CSI streams.

1) Band Pass Filtering: The key of designing a band pass filter is to determine its cut-off frequencies. Fine-grained human motion has low speed comparing to large-scale movements. For example, previous study on mechanical properties of lip movements [24], [25] shows that average movement speed of human jaw and lips when speaking is between $3 - 6cm/s$, corresponding to $0.5 - 1.1Hz$ dynamics in CSI streams for $5.18GHz$ WiFi signals. Instantaneous speed is higher than the average speed, which means a higher CSI frequency. In WiHear [23] the authors use a frequency range of $2 - 5Hz$. In this paper, We choose a wider frequency range of $1 - 10Hz$ to keep more details. In application scenarios other than lip reading, the cut-off frequencies should be determined by the corresponding applications.

In a static environment, human respiration is the most significant interference existing in CSI streams. Fig. 3a shows one original CSI stream in a static environment. The repeated pattern of breathing can be clearly observed in this figure. Typical respiratory rate for a healthy adult at rest is $12 - 20$ breaths per minute [26], corresponding to $0.2 - 0.33Hz$ dynamics in CSI streams. Fig. 3b shows the spectrum of the CSI stream in Fig. 3a.



(a) Original CSI Stream (b) After Butterworth band-pass filter (c) After PCA based denoising

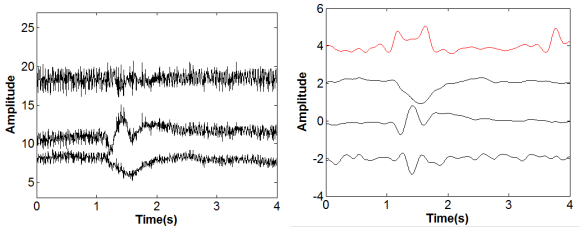
Fig. 4: Denoising the CSI Streams

To eliminate out-band interferences and noises, we use band-pass filter on CSI streams. According to what we discussed above, we set the cutoff frequency of the band-pass filter to be $1 - 10Hz$. We keep the frequency components up to $10Hz$ to get more details of the mouth movements. We choose 3-order Butterworth filter because it has a maximal flat amplitude response in the pass-band. Most of the high frequency burst noises and low frequency interference caused by respiration can be removed by the band-pass filter. Fig. 4b shows the band-pass filter results of the original CSI stream in Fig. 4a.

2) PCA Based Filtering: To further denoise the CSI streams and strengthen the effective CSI dynamics, we use principal components analysis(PCA) to track the correlation introduced in CSI streams by human motion. We get $30 \times 3 \times 1 = 90$ CSI streams in total when a WiTalk device has three antennas and the other end of the WiFi link has one antenna.

Among all the 90 CSI streams, we observe that not all of them show strong correlation. Specifically, the correlation is time varying and antenna related. Subcarriers from different antennas tends to be uncorrelated. For example, Fig. 5a shows three CSI streams that are collected from the three different antennas. Though the bottom two streams has some observable correlations, the top CSI stream from the third antenna does not seem to be correlated with them. CSI streams of the third antenna are more "noisy". If we use PCA on all 90 CSI streams from 3 antennas, noises from the third antenna will impinge the performance of PCA. Thus before PCA, we calculate the mean values of the correlation coefficients of the 30 CSI streams from an antenna, and setup a threshold to filter out the "noisy" antenna(s). We choose the threshold value 0.95. If none of the three antennas has a correlation coefficients higher than this threshold, we simply choose the antenna with the highest mean coefficient.

There are four main steps of applying PCA to CSI streams. The first step is data preprocessing. Data of CSI streams are segmented into small chunks to form a data matrix $\mathbf{H}$. Next, we calculate the correlation matrix of the data matrix as $\mathbf{H}^T \times \mathbf{H}$. The dimension of the correlation matrix is $N \times N$, where $N = 90$ is the number of CSI streams. The third step is to perform eigen-decomposition of the correlation matrix. In the last step, the principal components are reconstructed as $\mathbf{h}_i = \mathbf{H} \times \mathbf{q}_i$,



(a) Original CSI Stream (b) Filtered CSI Stream

Fig. 5: Different CSI Waveform for the Same Syllable

where $\mathbf{q}_i$ is i^{th} eigenvector.

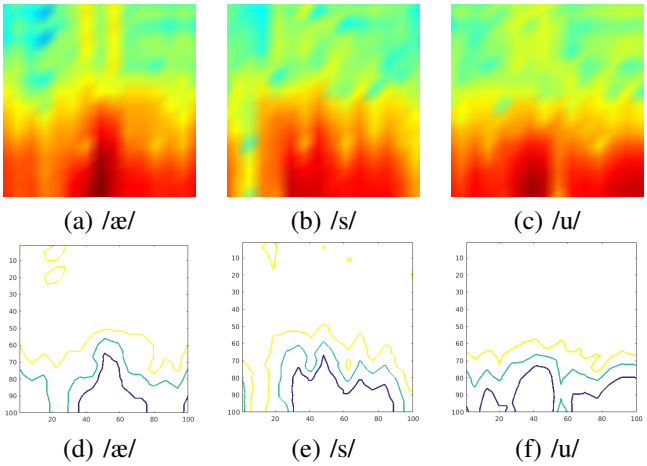
Only the first several principal components of PCA results with highest variance are valuable to our analysis. Due to correlated nature of CSI streams, all principal PCA components contain the same information. We discard the first principal component because noises caused by internal state changes are highly correlated and are captured in the first principal component [11]. We chose the second principal component as the input of the feature extraction step. Fig. 4c shows the second principle component of the PCA results. Compared to the band-pass filtered result of the same CSI stream in Fig. 4b, this PCA component contains more details of the CSI stream with higher strength.

C. Segmentation

Segmentation is an important preprocessing step to determine the start and end points of human motion. A good quality of segmentation will improve the performance of fine-grained motion sensing. WiTalk simply requires a short static interval between the movements to be detected. The static interval serves as the sentinel signal, helping WiTalk to segment the movements. We make this requirement because segmentation is not the research focus of this paper. In lip reading application, more details about syllable and word segmentation can be found in [23].

D. Feature Extraction

The design of a context-free fine-grained motion sensing system requires us to find the intrinsic properties in CSI streams that are stable and correlated to human motion only.



(a) /æ/ (b) /s/ (c) /u/

(d) /æ/ (e) /s/ (f) /u/

Fig. 6: Spectrogram and its Contour of Different Syllables

1) spectrogram Construction: Previous fine-grained motion sensing systems [17]–[19], [23] use time domain CSI profiles to extract features for subsequent classification step. However, as time domain CSI waveforms change with different contexts, it is infeasible to use CSI waveforms as the discerning profiles in this paper. We verify this claim by performing experiments on mouth movement sensing. Fig. 5a shows the original CSI waveforms of /æ/ form three subcarriers on three different antennas. Fig. 5b shows the corresponding band-pass filtered waveform. The top red line of Fig. 5b is a CSI waveform of /æ/ collected at a different location. Please be noted that there are translations of the lines in Fig. 5b to see them clearly. From Fig. 5b we can see that the waveforms from different contexts are significantly different. This verify our claim that it's infeasible to use CSI waveforms as the profiles in context-free settings.

As discussed in Section II, the CSi-Speed model proves that CSI spectrogram in time-frequency domain is a stable property of CSI streams that are highly correlated to human movement speeds. The movement speeds of different human body parts are correlated to a specific human activity like pronouncing a syllable. In the lip reading scenario, when pronouncing a specific syllable, there exists a specific movement pattern of all mouth parts [34]. A movement pattern includes the speed,direction and duration of the movements of every evolving mouth parts. WiTalk identifies CSI spectrogram in time-frequency domain as the stable property in CSI dynamics and extract features from these CSI spectrograms.

We take the second principle component of PCA based filtering process as the input to construct the spectrogram following these steps:

(1) Divide the input into equal-length segments. The segments must be short enough that the frequency content of the signal does not change appreciably within a segment. The segments may or may not overlap. We choose the segment size to be 128, corresponding to about 0.5 second of samples, so that the time is smaller than pronouncing a syllable and at the same time the number of samples is large enough to calculate the short-time Fourier transform. The overlap is set to be 126. Large overlap produces more spectrum lines and therefore a smoother spectrogram.

(2) Window each segment using a Hamming window and compute its spectrum using short-time Fourier transform.

(3) Display segment-by-segment the power of each spectrum in decibels and depict the magnitudes side-by-side as an image with magnitude-dependent colormap.

(4) Segment the CSI spectrogram using the static intervals to get the spectrograms for different syllables.

Fig. 6 shows the spectrograms for three different syllables. The spectrograms show how the energy of each frequency component evolves with time, where high-energy components are colored in red. We can see that there exists distinguishable patterns in the spectrograms, though not very clearly. As an example, the energy of spectrogram of /æ/ concentrate in the center. It is because when pronouncing /æ/, the jaw moves at a relatively higher speed in a short time. On the contrary, the spectrogram of /s/ spreads wider than /æ/, because when we pronounce /s/, the lips move slower and last longer time.

2) Feature Extraction: Though the spectrogram patterns are human distinguishable, we need to further extract features from CSI spectrograms for classification of the fine-grained motion.

We find that discrete wavelet transform (DWT) on CSI spectrograms is not suitable for fine-grained motion sensing, which is used in coarse-grained solutions in quadrant II such as CARM [11]. Fine-grained motion like mouth movements have lower speed than large-scale human activities like walking or falling, which results in low frequency components in CSI spectrograms. Without enough frequency resolution, it is infeasible to extract frequencies at multiple resolutions on multiple time scales using DWT.

In WiTalk, we propose to first calculate the contours of the spectrogram images to extract features. The contours represent the edges of different energy levels of the spectrograms, and depicts the unique patterns of the spectrograms of fine-grained motion. Fig. 6 shows three contour lines for the corresponding upper spectrograms. The top yellow contours mark the lines of signal energy and noise. The bottom blue contours enclose the major part of signal energy.

Directly using the contour lines as the classification features leads to high computational costs for classification. Therefore, we use the most relevant signal processing tool DWT on the contour lines to compress their length by extracting approximate sequences. In WiTalk we choose Daubechies wavelet filter of order 4 because it has the best classification performance.

E. Classification

People may perform the same micro motion at different speeds, and even for the same person, the motion speeds may vary from time to time. Dynamic time warping(DTW) can be used to measure similarity between two temporal sequences which may vary in speed. DTW calculates the optimal match between the two time sequences and wraps the sequences to

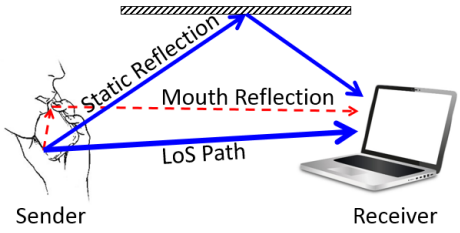


Fig. 7: System Scenario of WiTalk

measure their similarity. The output of DTW is the distance between the two series. Low distance means high similarity between the two sequences. We build a classifier using the DWT compressed contours of the CSI spectrograms as features. The classifier calculate the DTW distances between the input and all the contours in the dataset. The one with the shortest distance is identified as the recognized motion.

F. Error Correction

The performance of fine-grained motion sensing can be further improved using application specific context information. For example, in the lip reading application, such information includes constraints that reject sentences that are not following these constraints. For example the sentence "The apple is red" will be accepted but "The apple is angry" will be rejected [36]. In WiTalk, we implement context-based error correction using a simple Bayesian method similar to [34].

IV. Performance Evaluation

We implement WiTalk on a commercial laptop, and evaluate its performance in a typical lab environment. The system scenario of WiTalk is illustrated in Fig. 7. WiTalk is implemented on the receiving end of a WiFi link. The transmitter continually sends packets to WiTalk at a speed of 250 packets/second. WiTalk collects CSI data and use the proposed algorithms to infer the fine-grained motion from the hidden patterns of CSI streams. We design experiments to detect a set of pronounced syllables, which is the bases for lip reading applications. We select lip reading as our example application scenario because it is the most fine-grained motion sensing in the literature that is previously reported using WiFi CSI dynamics [23].

A. System Setup

WiTalk is implemented on a commercial Thinkpad X301 laptop. The laptop is equipped with an Intel Core 2 U9600 processor, 4GB memory and an Intel 5300 NIC with 3 omni-directional antennas. The operating system running on the laptop is Ubuntu 14.04 LTS. We install and configure Linux 802.11n CSI Tool as described in [1]. The laptop works as the receiver and collects CSI streams using the CSI tool. Collected data are processed using Matlab scripts for signal processing and classification. Matlab version is R2016a. We do the experiments on channel 36 at $5.180GHz$.

We test WiTalk using two model of smartphones: a LG Nexus 5 with Android 6.0.1 and a Samsung Note 5 with

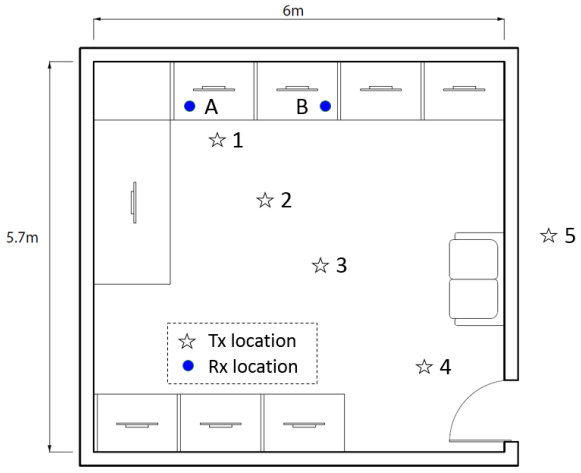


Fig. 8: WiTalk Testbed

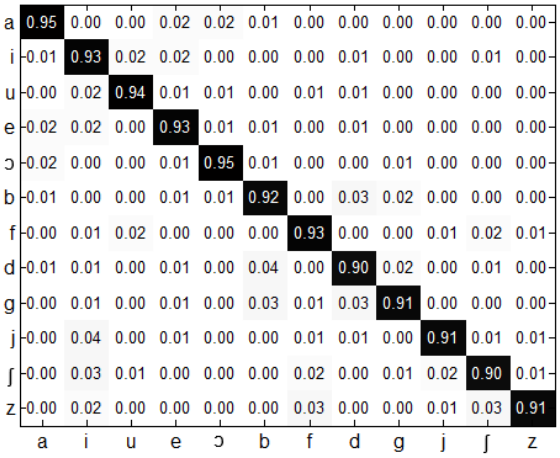


Fig. 9: Confusion Matrix of 12 Syllables: Same Context

Android 5.1.1. The smartphones work as the transmitters. The transmitter continually sends packets to WiTalk at a rate of 250 packets/second during the experiments. The CSI streams are collected and stored for later processing. WiTalk can also work in realtime currently. By writing the CSI dynamics to a named pipe, Matlab script can read from the pipe and process the data simultaneously.

We test WiTalk in a normal lab environment depicted in Fig. 8. The WiTalk device is tested at two positions marked as blue dots. We collect data with three volunteers(all males). The volunteers are asked to stand still at the positions marked as stars. They hold the phone steadily in normal phone call position the same way as depicted in Fig. 7. Instead of making a real phone call, they are asked to read a set of syllables and a set of short sentences no more than six words. Static intervals are inserted intentionally between syllables and words to facilitate segmentation. The set of syllables includes 12 elements: /a/ /i/ /u/ /e/ /ɔ/ /b/ /f/ /d/ /g/ /j/ /ʃ/ /z/. Each volunteer reads the set of syllables 10 times at each test location, and the set of short sentences 5 times at each location.

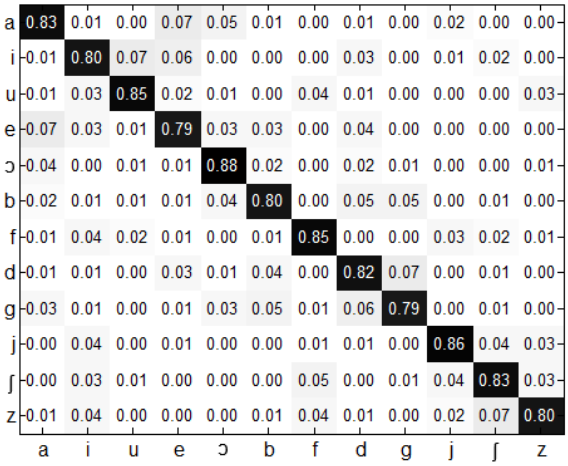


Fig. 10: Confusion Matrix of 12 Syllables: Mixed Context

B. Syllables Classification Accuracy

The recognition results for the syllables set are depicted in Fig. 9 and Fig. 10. We do the experiments in two steps to verify the syllable detection performance. We first train and test the classifier in the same context with the same volunteer at the same location. The confusion matrix is reported in Fig. 9 with an average detection accuracy of 92.3%. Next we mix the data collected from different users, different transmitters at different locations for both training and testing. The confusion matrix is reported in Fig. 10. In the mixed contexts situation, the average detection accuracy is 82.5%. The reasons for the lower accuracy in mix contexts situation are two-folds: 1) There are slight differences for different people to pronounce the same syllable. For example, in our experiments one of the volunteers tends to pronounce /u/ more lightly than others. His lips pucker very little when he makes this pronunciation, which impinges the detection accuracy of this syllable. 2) Signal reflection multipathes are significantly changed at different locations. This will introduce noises to CSI streams which cannot be removed completely. This impinges the classification performance compared to the same-context situation.

C. Sentence Recognition Accuracy

We evaluate the performance of sentence recognition with and without context-based error correction. The set of short sentences include sentences from 1 word to 6 words. As shown in Fig. 11, with the increase of the number of words, the accuracy drops significantly. For 6 words situation, without context based correction, the recognition accuracy is only about 43%. With context based correction, the accuracy is improved by 16%. This is because the longer the sentences, the more they context information can be applied. The drop of the performance is mainly because the difficulty of in-word syllables segmentation. To solve this problem, continuous lip reading model could be adopted as in [35].

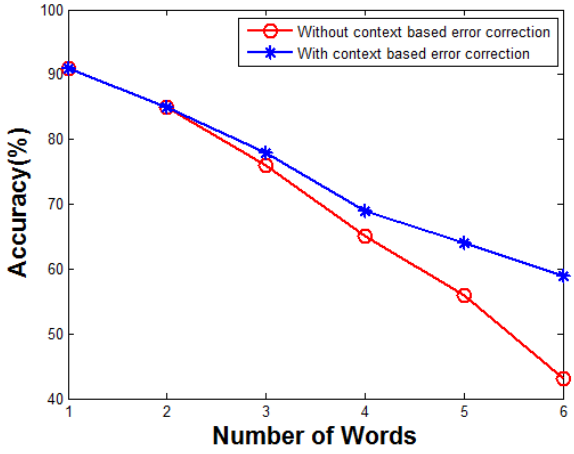


Fig. 11: Sentence Recognition Accuracy

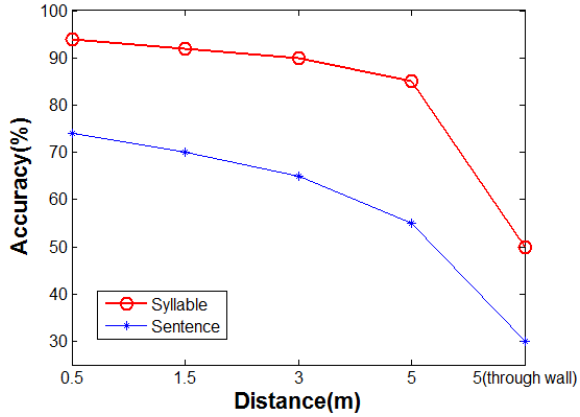


Fig. 12: Performance with Distance

D. Performance with Distance

Fig. 12 depicts the performance with increase of distance between the transceivers. For the receiver location marked as B, we only did the experiments at transmitter location 1, 2 and 5. The distance of $B1$,$B2$ and $B5$ are $1.5m$, $1.5m$ and $3m$ respectively. As we can see from Fig. 12, the syllable discerning accuracy drops by 9% when the distance increase from $0.5m$ to $5m$ in the same room with LOS available. And the accuracy of sentence detection drops by 19%. However in through-the-wall scenario as in location 5, both syllable and sentence detection rate drop significantly by over 25% compared to location 4 at the same distance. To improve the performance and increase the working range of WiTalk, using a directional antenna will be a promising choice.

V. Related Work

Motion sensing. Motion sensing based human localization, human tracking, gesture and activity recognition have been studied a lot in the research community. Existing work on motion sensing can be divided in three categories: vision-based, sensor-based and RF-based.

The most popular approaches for video gaming and virtual reality platforms are vision-based. Such systems include

Microsoft Xbox Kinect [3], Leap Motion [4], and Sony PlayStation Camera [5]. They use color and infrared cameras to do body-depth perception, motion tracking and gesture recognition. The main problem of vision-based motion sensing is that its performance is highly influenced by the condition of lighting. These systems also require line-of-sight(LOS) for proper operation.

Wearable device based methods like RF-IDraw [6] traces trajectory of fingers and hands by attaching RFID to the fingers. Xu et. [7] uses smartwatch to identify 37 gestures with an accuracy of 98%. TypingRing [8] asks the users to wear a ring for text inputting with the capability of detecting and sending key events in real-time with an average accuracy of 98.67%.

Earlier work on RF-based motion sensing rely on specialized hardware. WiTrack [27] tracks 3D human body motion using an FMCW(Frequency Modulated Carrier Wave) radar at the granularity of 10cm. WiSee [28] works by looking at the minute Doppler shifts and multi-path distortions for gesture recognition. Google Project Soli [29] uses on-chip 60GHz radar to detect fine-grained motion. However, the short effective range limit its application in long distance scenarios.

CSI-based method like CARM [11] builds a CSI-speed model and a CSI-activity model, which depicts the relationship between CSI value dynamics and human body parts movement speeds, and the relationship between the body movement speeds and specific human activities. CARM is coarse-grained as it discerns human activities like walking, falling and sitting down. Different from CARM, WiTalk is fine-grained and reads the motion of human mouth. CARM also requires a sampling rate as high as 2500 samples per second. It is very difficult to reach such high sampling rate in WiTalk scenario. WiKey [17] uses CSI waveform shape as the features and can recognize keystrokes in a continuously typed sentence with an accuracy of 93.5%. WiKey works well only in controlled environments and specific devices positioning. WiFinger [18] also uses CSI waveform shape as the features and can discern 8 finger gestures with 93% recognition accuracy. WiFinger also requires static transceivers and finger motion must be near the LOS line of the transceivers. Different from WiKey and WiFinger, WiTalk removes the limitation of static transceivers and works on mobile devices. WindTalker [19] allows an attacker to infer the sensitive keystrokes on a mobile device using CSI. However, WindTalker requires that the mobile device being placed in a stable environment. Wi-Wri [21] uses WiFi signals to recognize written letters. WiDraw [22] leverages WiFi signals from commodity mobile devices to enable hands-free drawing in the air. These two projects focus on the users hand trajectory tracking, which is not WiTalk's research target. The most related work to our work is WiHear [23]. Our work is inspired by WiHear, but we must point out that the system setting and techniques used are significantly different. WiHear uses specialized directional antennas to obtain usable CSI variations. It takes 5-7 seconds for stepper motors to adjust the emitted angle of the radio beam to locate the target's mouth, which is not acceptable for a real-time eavesdropping system in our setting. Furthermore, WiHear does not have enough noise filter mechanisms. It still needs the training process per location per user. On the other hand, WiTalk can be implemented on commercial WiFi devices and has the one-time training feature.

Lip reading. [31] present a combination of acoustic speech and mouth movement image to achieve higher accuracy of automatic speech recognition in noisy environment. [32] presents a vision-based lip reading system and compares viewing a person's facial motion from profile and front view. [33] shows the possibility of sound recovery from the silent video. SilentTalk [34] generates ultrosonic signals from mobile phone and analyzes the frequency-shift caused by mouth movements from the reflections. It can identify 12 basic mouth motion up to 95.4% accuracy. Chung et. [35] presents their recent results on lip reading which performs better than hum pros. The system was trained using 5000 hours of videos including 118,000 sentences.

VI. Conclusion and Future Work

WiTalk is the first fine-grained motion sensing system using CSI dynamics of WiFi. WiTalk can be implemented on a single WiFi device. We analyse and verify the feasibility of WiTalk in the application of CSI-based lip reading on smartphones. We propose to denoise CSI streams using band-pass filtering and PCA based filtering. We identify the spectrograms of CSI dynamics as the intrinsic features that correlate to human fine-grained motion only, and extract features from the contours of CSI spectrograms. WiTalk needs only one-time training and works for different environments. Experiment results show that WiTalk can discern a set of 12 syllables with an accuracy of 92.3% and short sentences up to six words with an accuracy of 74.3%.

The current implementation of WiTalk has some limitations. First, the performance of WiTalk degrades with the increase of the distance between the two transceivers and the distance between the transceiver and moving human body parts. The main reason is that the reflected signal power from fine-grained motion is too small and is easily buried in noises and interferences. Using directional antennas is one possible solution, which we leave as future work.

Second, WiTalk requires the user staying relatively still for the fine-grained motion sensing. If the user talks over the phone while walking around, WiTalk will fail because the useful CSI dynamics will be buried in the CSI changes caused by the moving user body and legs. Even if the user stays still at one location, if he/she makes some hand or body gestures, the extraction of useful CSI dynamics will also become harder. How to eliminate these interferences will be our future research target.

Acknowledgment

This work was supported in part by US National Science Foundation under grants CNS-1446478 and CNS-1443889.

REFERENCES

[1] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11 n traces with channel state information," in ACM SIGCOMM CCR, 2011.

[2] Y. Xie, Z. Li, and M. Li, "Precise power delay profiling with commodity wifi," in ACM MobiCom, 2015.

[3] Kinect. https://dev.windows.com/en-us/kinect.

[4] Leap Motion. https://www.leapmotion.com/.

[5] PlayStation Camera. https://www.playstation.com/en-us/explore/accessories/playstation-camera-ps4/

[6] J. Wang, D. Vasisht, and D. Katabi, "Rf-idraw: virtual touch screen in the air using rf signals," In ACM SIGCOMM, 2014.

[7] C. Xu, P. H. Pathak, and P. Mohapatra, "Finger-writing with smartwatch: A case for finger and hand gesture recognition using smartwatch," In ACM HotMobile, 2015.

[8] S. Nirjon, J. Gummeson, D. Gelb, and K.-H. Kim, "Typingring: A wearable ring platform for text input," In ACM MobiSys, 2015.

[9] Z. Yang, Z. Zhou, and Y. Liu, "From RSSI to CSI: Indoor localization via channel response," ACM Computing Surveys, 46(2): 25, 2013.

[10] K. Wu, J. Xiao, Y. Yi, D. Chen, X. Luo, L. M. Ni, "Csi-based indoor localization", IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 7, pp. 1300-1309, Jul. 2013.

[11] W. Wang , A. Liu , M. Shahzad , K. Ling , S. Lu, "Understanding and Modeling of WiFi Signal Based Human Activity Recognition," In Proceedings of the 21st Annual International Conference on Mobile Computing and Networking, September, 2015, Paris, France

[12] Y. Wang, J. Liu, Y. Chen, M. Gruteser, J. Yang, and H. Liu, "E-eyes: device-free location-oriented activity identification using fine-grained wifi signatures," In ACM MobiCom, 2014.

[13] W. Wang, A.X. Liu, M. Shahzad, "Gait Recognition Using WiFi Signals", Proc. Int'l Conf. Pervasive and Ubiquitous Computing, 2016.

[14] Y. Zeng, P. H Pathak, and P. Mohapatra. "WiWho: WiFi-based Person Identification in Smart Spaces," In Proc. IEEE/ACM IPSN, 2016.

[15] H. Wang, D. Zhang, J. Ma, Y. Wang, Y. Wang, D. Wu, T. Gu, and B. Xie, "Human respiration detection with commodity wifi devices: do user location and body orientation matter?" In Proceedings of UbiComp '16. ACM, New York, NY, USA, 25-36.

[16] J. Liu, Y. Wang, Y. Chen, J. Yang, X. Chen, and J. Cheng. "Tracking Vital Signs During Sleep Leveraging Off-the-shelf WiFi," In Proceedings of MobiHoc 2015.

[17] Kamran Ali , Alex X. Liu , Wei Wang , Muhammad Shahzad, "Keystroke Recognition Using WiFi Signals," Proceedings of the 21st Annual International Conference on Mobile Computing and Networking, September 07-11, 2015, Paris, France

[18] S. Tan and J. Yang, "WiFinger: Leveraging Commodity WiFi for Fine-grained Finger Gesture Recognition," Proceedings of the Seventeenth International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2016), Paderborn, Germany, July, 2016.

[19] Mengyuan Li, Yan Meng, Junyi Liu, Haojin Zhu, Xiaohui Liang, Yao Liu, and Na Ruan, "When CSI Meets Public WiFi: Inferring Your Mobile Phone Password via WiFi Signals," In Proc. ACM SIGSAC CCS, 2016.

[20] Ouyang Zhang and Kannan Srinivasan, "Mudra: User-friendly Fine-grained Gesture Recognition using WiFi signals," Proc. of ACM CoNEXT16, Dec 12-15, 2016.

[21] X. Cao, B. Chen and Y. Zhao, "Wi-Wri: Fine-Grained Writing Recognition Using Wi-Fi Signals," 2016 IEEE Trustcom/BigDataSE/ISPA, Tianjin, 2016, pp. 1366-1373.

[22] L. Sun, S. Sen, D. Koutsonikolas, and K.-H. Kim, "Widraw: Enabling hands-free drawing in the air on commodity wifi devices," In ACM MobiCom, 2015.

[23] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We can hear you with Wi-Fi!," In Proceedings of ACM MobiCom, 2014.

[24] S. Maeda and M. Toda, "Mechanical properties of lip movements: How to characterize different speaking styles?" Proceedings of the XVth ICPhS, Barcelona: 189-192, 2003.

[25] D.J. Ostry, J.R. Flanagan "Human jaw movement in mastication and speech," Archives of Oral Biology, 34 (1989), pp. 685-693

[26] K. Barrett, S. Barman, S. Boitano, H. Brooks "Ganong's Review of Medical Physiology (24 ed.)," McGraw-Hill Education/Medical, 2012, p.619, ISBN 0071780033.

[27] F. Adib, Z. Kabelac, D. Katabi, and R. C. Miller, "3d tracking via body radio reflections," In NSDI, 2014.

[28] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," In ACM MobiCom, 2013.

[29] Google Project Soli, https://www.google.com/atap/project-soli/.

[30] J. Dennis, H. D. Tran and H. Li, "Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions," in IEEE Signal Processing Letters, vol. 18, no. 2, pp. 130-133, Feb. 2011.

[31] 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. 109C112.

[32] K. Kumar, T. Chen, and R. M. Sternl, "Profile view lip reading," in Proc. Int. Conf. Acoust., Speech Signal Process., 2007, pp. 429C432.

[33] 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. 79:1C79:10, 2014.

[34] J. Tan, C. Nguyen, X. Wang, "SilentTalk Lip Reading through Ultrasonic Sensing on Mobile Phones," In INFOCOM. IEEE, March 2017, Atlanta, US.

[35] J.S. Chung, A. Senior, O. Vinyals, A. Zisserman, "Lip reading sentences in the wild," In: Proc. CVPR (2017)

[36] Wikipedia, speech recognition, https://en.wikipedia.org/wiki/Speech_recognition.