Wearable Social Sensing: Content-Based Processing Methodology and Implementation

Jun Gu, Bin Gao, *Senior Member, IEEE*, Yuanpeng Chen, Long Jiang, Zhao Gao,
Xiaole Ma, Yong Ma, Wai Lok Woo, and Jikun Jin

***Abstract*— Developing wearable activity and speech sensing for assessing human physical and mental health is just as significant as conscious content for determining social behavior. Multiple social relevant sensors, such as microphones and accelerometer, embedded in wearable devices paves the way to provide the opportunity to continuously and non-invasively monitor anxiety and stress in real-life situation. In this paper, we present the design, implementation, and deployment of a wearable computing platform capable of automatically extracting and analyzing social signals. In particular, we benchmarked a set of integrated social features to objectively quantify the level of anxiety using an in-house built wearable device. In addition, in order to protect privacy, we propose a potential method to embed the audio features processing in the hardware to avoid recording their voice directly. In addition, we have implemented the *k*-means classification to determine the level of anxiety of the subjects. The obtained performance has demonstrated that both activity and speech social features have the potential to directly infer anxiety across multiple individuals.**

***Index Terms*— Social signal processing, audio and activity features, wearable device.**

Nomenclature

SSP	Social Signal Processing
AC	Affective Computing
FMRI	Functional Magnetic Resonance Imaging
ADPCM	Adaptive Difference Pulse Code Modulation
PCB	Printed Circuit Board
MCU	Microcontroller Unit
DMP	Digital Motion Processor
STAI	State-Trait Anxiety Inventory
PRCS	Personal Report of Confidence as a Speaker
STE	Short-time Energy
ZCR	Zero-crossing Rate
MP	Maximum Peaks in Short-time
AMP	Autocorrelation Maximum Peaks
APN	Number of Autocorrelation Peaks
MFCC	Mel-Frequency Cepstrum Coefficient
LPCC	Linear Prediction Cepstrum Coefficient
std	Standard Deviation
sp	Sparseness
PPMCC	Pearson Product-moment correlation coefficient
OLED	Organic Light-Emitting Diode
MEMS	Microelectromechanical Systems
DMIPS	Dhrystone Million Instructions executed Per Second
RMS	Root Mean Square
FFT	Fast Fourier Transformation

I. Introduction

SOCIAL Signal Processing (SSP) [1], [2] and Affective Computing (AC) [3] have been widely studied for decades. This topic has attracted researchers various fields such as healthcare [4], [5], psychology [6]–[8], sociology [9], [10], fitness [11], [12] and ambient intelligence [13], [14]. Psychologists have firmly demonstrated that social signals especially the nonlinguistic signals (body language, tone of voice, facial expression) are just as important as conscious context, which is a powerful determinant of human behavior and speculate that they might have evolved as a way to establish hierarchy and group cohesion [15]–[17]. The emergent research fields of SSP and AC have demonstrated that complex human social behavior and psychological states can be inferred from multi-modal data acquisition, fusion and mining by coupling different types of wearable sensors, principally from audiovisual [18] and physiological [19] sensing.

Typically, there exist three standard methods to analyze human social behavior. Firstly, multi-cameras are used to track and recognize the human activities or gesture in an environment while they build the corresponding social network and personality graphs [20]. Secondly, subjects who are doing social interaction are observed and recorded by the independent observers in the specific space. Finally, a form of questionnaire survey is employed for self-reports. To a certain extent, the standard methods can study the characteristics of social behavior, emotional performance and mental health state. However, independent observers trained to do assessment are limited to a small number of people and have inter-observer reliability concerns. Multi-cameras monitoring is expensive and the space that the cameras track and record is limited. All the above suggests that the issue of personal privacy of the subjects cannot be well protected. Studies show that self-report correspond poorly to communication behavior as recorded by independent observers [21], [22] and it suffers from subjectivity and memory efforts.

Manuscript received August 19, 2017; accepted September 11, 2017. Date of publication September 19, 2017; date of current version October 11, 2017. The associate editor coordinating the review of this paper and approving it for publication was Dr. Edward Sazonov. *(Corresponding author: Bin Gao.)*

J. Gu, B. Gao, Y. Chen, L. Jiang, Z. Gao, X. Ma, Y. Ma, and J. Jin are with the School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: bin_gao@uestc.edu.cn).

W. L. Woo is with the School of Electrical and Electronic Engineering, Newcastle University, Newcastle upon Tyne, NE1 7RU, U.K.

Digital Object Identifier 10.1109/JSEN.2017.2754289

In recent years, wearable sensors are widely used to sense and understand human social signals as well as social context in daily life. The studies can dramatically improve collective decision making and help keep remote users in the loop [1]. This innovative approach allows SSP and AC to be automated more naturally. Notwithstanding above, as many of social interactions over the course of weeks and months are repetitive, the method allows us to collect multi-modal data and analyze daily interaction patterns from people over an extended period of time.

Several wearable computing projects have considered the use of wearable devices or smartphones to gather social signals and social context. BodyScope [23] is a wearable acoustic sensor, which can record the sounds produced in the wear's throat area and classify them into activities, such as speaking, laughing, eating, drinking, and speaking. In [4], the smartphone is treated as a multi-sensor device which serves as a continuous monitor capable of informing both clinically-relevant inferences with efficacious and timely patient-feedback. MIT Media Lab has developed several socially aware platforms [24], [25] to measure several aspects of social context, including nonlinguistic social signals measured by analyzing the person's tone of voice, facial movement, and gesture.

Recently, there has been a growing concern about the mental health issues such as anxiety, depression, suicidal ideation, and self-injury on young generation especially college students. Reports show that about 32% of college students suffer from mental health issues [26] and less than 20% of college students with mental health issues received treatment [27]. Anxiety disorders may impose significant functional impairment, especially for juveniles and typically have a chronic course, which can interfere with student's performance at school or social interaction activity [28]. Even in daily life, a large amount of general population fears public speaking with great anxiety emotion and behavior activities, which may induce hands tremble or voice change [32].

The traditional method of diagnosing and assessing adolescent mental health state include self-report questionnaires, functional magnetic resonance imaging (FMRI) and through therapist-moderated assessment sessions. They mainly have the following shortcomings. Firstly, patients should periodically fill in the questionnaires or go to hospital which will take them for a long time and uncomfortable way. Secondly, the self-reported data sometimes cannot be accurate, it may suffer from subjectivity and memory efforts. Finally, the mental health state reflected by the data they fill out the surveys is always a reflection of the recent situation and cannot show changes in mental health state for a period of time. In order to overcome the shortcomings, the research studies start to use smartphones as tools to keep track of mental health issues. In [29], it used smartphone-based sensing modalities including phone call duration, speech analysis and movement data can identify manic and depressive states. In [30], it developed a smartphone app for collecting data relevant to behavioral trends of mental illness to provide better disease insights to the patients. In [31], it developed a smartphone app that periodically collects the locations of the users and answers to daily questionnaires that quantify their depressive mood, and demonstrated that there exists a significant correlation between mobility trace characteristics and the depressive mood.

However, the use of sensors in the smart phones has the following shortcomings. Firstly, the smart phone is placed in the pocket while the activity information it captures is inaccurate compared with the smart watch which is worn on the wrist. Secondly, the APP of the smart phone running in the background to record a variety of sensors data will greatly increase the power consumption. Finally, smart phone holds a variety of personal privacy information and the users do not want to leak the important information.

Thus, in this paper, we develop a wearable social sensing watch for unobtrusively and continuously capturing the granular details of behaviors and contexts which might provide important cues about anxiety onset. In particular, we benchmarked a set of integrated social features to objectively recognize the anxiety. Specifically, we test anxiety level immediately after public speaking with English or Chinese language to explore both speech and activity features with index related to anxiety state. In addition, in order to protect privacy, we explored a potential way to embed the audio features processing in the hardware to avoid directly recording their voice and implement the k-means classification to conduct the anxiety level recognition.

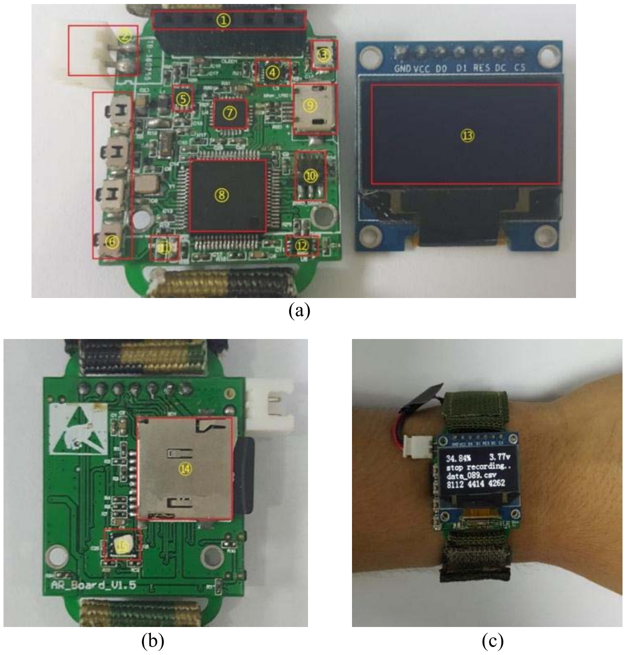
The paper is organized as follows. Section II introduces the proposed social sensing architecture. In Section III, the implementation of experiment set up and feature extraction is derived. Experimental analysis and a series of correlation studies with anxiety onset are presented in Section IV. Several feature extraction algorithms on board and classification analysis based on the most relevant features are presented in Section V. Section VI concludes the paper.

II. Proposed Architecture

The system is constructed based on social awareness to unobtrusively and continuously collect acoustic and physical activity signals in completely natural and unpredictable situation. Briefly, the whole system consists of an acoustic capturing system which hybrids of acoustic sensor and activity capturing system. The activity capturing system consists of an ARM microcontroller, a variety of digital sensors for collecting multi-modal data which can perceive the environment and wear social behavior, a microSD card for extended storage, a power management unit for not only powering up the whole system but also charging the lithium battery, and an OLED screen for showing detail information about the running system.

A. Acoustic Capturing System

Audio signal capturing device with the characteristics of small size, light weight and low power consumption is specified which can be fastened onto the subject' collar to continuously record the speech utterance. The speech utterance captured by the acoustic system is sampled at 48kHz and encoded by adaptive difference pulse code modulation (ADPCM) algorithm which can save memory.



(a)

(b)

(c)

Fig. 1. The activity capturing system prototype. (a) Top side. (b) Bottom side. (c) Wear. 1: OLED connector. 2: aluminum battery socket. 3: manal reset button. 4: charging chip. 5: dc-dc converter. 6: function buttons. 7: accelerometer & gyroscope. 8: MCU. 9: USB port. 10: debug interface. 11: LED indicator. 12: ambient light sensor. 13: OLED 14: microsd card slot.

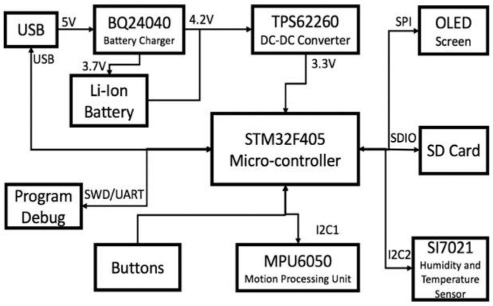


Fig. 2. System block diagram of activity capturing system.

B. Activity Capturing System

Fig. 1 shows the design and assembly of the proposed 2-layer printed circuit board (PCB). The block diagram of system can be shown in Fig. 2. The activity capturing system is based on STM32F405RGT6 which is a programmable system on-chip ARM 32-bit microcontroller unit (MCU) produced by STMicroelectronics. It is designed for consumer applications where the high level of integration and performance (210DMIPS at 168 MHz, DSP instructions and floating point unit) is inside packages as small as 4 × 4.2 mm are required.

The MPU6050 is used due to the low power and high performance requirement of wearable sensors to continuously collect human posture and activity data. The SI7021 is used to measure the temperature and humidity of skin which are significant sign information for assessing the mental state and emotional performance of the subjects. At the top of the PCB, a 0.9-inches OLED with 128 × 64 resolution is connected to the mother board. It is used to indicate the system running information, and display the measurements of various sensors in real time so that we can clearly know whether the system is running normally or not.

In order to meet the different sampling rate of multiple sensors and minimize the complexity of coding, an embedded real-time operating system called RT-Thread and a micro filesystem called FatFs are ported to the system. RT-Thread is an open source real-time operating system for embedded devices and is distributed under the GPLv2 license. FatFs is a generic FAT/exFAT filesystem module for small embedded system and is written in compliance with ANSI C (C89) and completely separated from the disk I/O layer.

It has been validated that the activity capturing system can be able to record three days of data (about 75 hours) as the 3000mAh lithium battery is used.

III. Methodology and Experiment

In this section, we describe the methodology and experimental studies for measuring and assessing adolescent anxiety state using our platform. The methodology can also be applied to other wearable social sensing applications with similar capabilities.

The proposed method hybridizes both speech and activity social features to measure and assess anxiety state. It has several advantages over the existing approaches such as being able to automatically capture the social behavior that allows us to perform fine-grained analysis of human well-being without the need of human observers, embedding the audio features processing in the hardware to avoid directly recording wears' raw voice data to protect privacy.

A. Questionnaire Details

Subjects were asked to complete a series of questionnaires. State-Trait Anxiety Inventory (STAI) [33] to assess their general anxiety and speech state anxiety; Personal Report of Confidence as a Speaker (PRCS) [34] was used to measure their public speaking fear severity. The Fear of Negative Evaluation Scale (FNE) [35] was used to measure a person's apprehension about negative evaluation. They also conducted Five-Factor Inventory (NEO-FFI) [36] and the Rosenberg Self-Esteem Scale (SES) [37] to control possible difference in personal characteristics.

B. Speech Data Processing

Offline social feature process and analysis is implemented to assessing the mental health state of the subjects. Specifically, a series of pre-processing of audio data has been conducted. The raw audio signal is pre-processed to remove the silent part as well as the voice of non-test part of the audio data. After the pre-processing, only the subjects' own voice remains. This process is equivalent to speech segmentation process and they are manually annotated by using the audio editing software "Elan". We analyze blocks of audio data from the

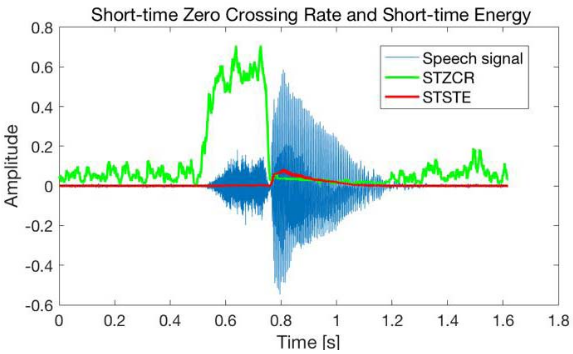


Fig. 3. Example of calculated features.

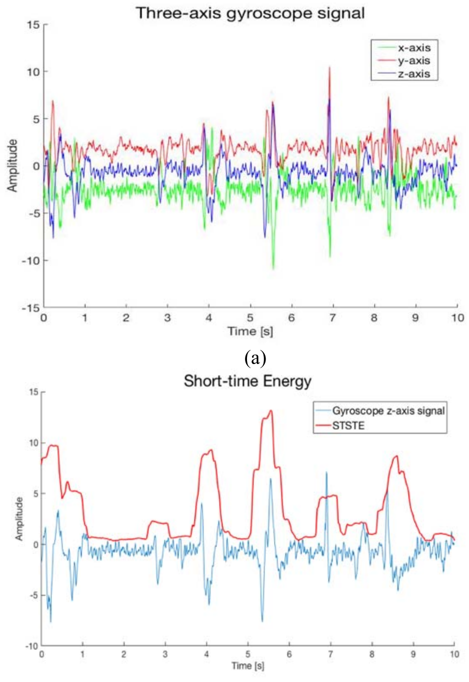
acoustic capturing system (i.e. one block of data consisted of 50 milliseconds). Before calculating features the block audio signals are broken into short-time, 50%-overlapping windows (frames) of 50 milliseconds. For each frame, 16 features are calculated, namely, short-time energy (STE), zero-crossing rate (ZCR), maximum peaks in short-time (MP), autocorrelation maximum peaks (AMP), number of autocorrelation peaks (APN), energy-entropy (Entropy), formant, Mel-Frequency cepstrum coefficient (MFCC), Linear prediction cepstrum coefficient (LPCC), tempo, spectral roll off, spectral brightness, pitch, mode, key, emphasis and time. For most of these features within a frame, three simple statistics are calculated (mean value, standard deviation and sparseness). This step leads to 76 statistic values which are the collected feature values that are employed to characterize the input audio signal. The example of audio raw data and two features can be shown in Fig. 3.

In order to obtain a set of baseline measurements of the "typical" speech social features when subjects under normal environment, a set of standard data have been recorded. We arrange every subject read a well-designed English paragraph and a Chinese paragraph in a completely natural and relaxed state in a closed classroom and record the voice at the same time. Each of 76 statistic feature calculations is performed on all standard data samples to obtain a set of values that can be considered typical of speech and environment sound signal.

C. Activity Data Processing Method

In order to estimate more accurate wrist movements, we combine the data from accelerometer and the data from gyroscope. Studies show that the major energy band for daily activities is 0.3–0.5Hz [38], and 99% of the acceleration power during bare foot walking is contained below 15Hz [32], [33]. In [41], study concluded that in order to assess daily physical activities, accelerometers must be able to measure up to ±4g if they are attached at waist level. Thus, in our experiment the three-axis accelerometer and three-axis gyroscope signal is sampled at 100Hz, which can not only capture the range of human activity but also save the power consumption.

The example of raw data diagram is shown in Fig. 4(a). Before calculating features, the block activity signals are broken into short-time, 50%-overlapping windows (frames) of 600 sample points (6 seconds). For each frame, seven kinds of features are calculated, namely, mean, standard deviation, short-time energy, energy-entropy, correlation coefficient between axis, pitch, roll and peak value in frequency domain. Parts of these features have been successfully used in [42] and [43]. Examples of the calculated feature is shown in Fig. 4(b).



(a)

(b)

Fig. 4. Example of calculated features. (a) Gyroscope raw data. (b) Short-time energy of gyroscope z-axis signal.

D. Correlation Analysis

Once activity and speech social features are extracted, the Pearson product-moment correlation coefficient (PPMCC) between social features (speech and activity features) and the subjects' self-report anxiety levels are calculated. It is a measure of the linear correlation between two variable X and Y. It has a value between $+1$ and -1, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation.

The purpose of the paper is to determine both the most relevant and the least relevant sets of features to objectively identify anxiety. In order to further explore language difference (English vs. Chinese) and identify the most relevant social features, correlation analysis between social signal features and the self-report anxiety grades was conducted in different categories.

E. Experiment Setup

We have collected a corpus containing sensors-derived measurements of speech and activity within 94 subjects. They are all sophomores at the University of Electronic Science and Technology of China (UESTC) and we record the data from a English lesson class. At the beginning of each class, they

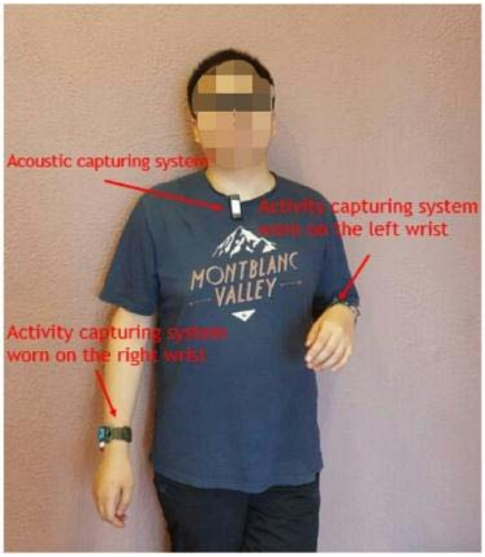

Acoustic capturing system
Activity capturing system worn on the left wrist
Activity capturing system worn on the right wrist

Fig. 5. Test subject.

are asked to make presentations in groups for about five minutes. Each group is required to perform in both English and Chinese. During the presentation period, each member of the group is instructed to wear the acoustic capturing system to continuously record the audio signal and the activity capturing system containing multiple sensors for detecting activities, postures and environmental context. Fig. 5 shows how a subject wears the audio capturing system and the activity capturing system. After the presentations, every subject is asked to complete a survey at the end of each data collection episode, reporting on their immediate feeling when they are presenting in front of the class. In total, we collected 10hrs of audio and activity signals and 912 self-reports. Other questionnaires for personal characteristic and general state were carried out at the end of the semester. The data is collected during presentation process for one day per week over a semester. As some of subjects' sample data is not complete or do not satisfy the instruction (e.g. some subjects only do the English presentation but no Chinese presentation or missed filling in the questionnaire) this brings us to a total of 22 subjects. We evaluated these subjects' data and focused on analyzing the correlation between self-report grades and the features derived from activity and speech. The objective of the experiment is to use the proposed platform to correlate the anxiety state with social sensing features for individual subjects. The self-report questionnaires provide us with a detailed picture of the subjects' recent anxiety level during presentation.

IV. Evaluation and Analysis

In this section, we will specify the progress of social feature analysis

A. Speech Analysis

In order to do comparative analysis, we have selected those data that contain both Chinese and English presentations from the collection of data set. For each of the collected speech data, the following features are calculated: STE, ZCR, MP, AMP, APN, Formant, Entropy, MFCC, LPCC, Tempo, Rolloff, Brightness, Pitch, Mode, Key, Emphasis and time of presentation. Since most features are vectors, the length of the vector represents the number of frames of each segment of speech. We calculate three statistics to characterize these features i.e., mean, standard deviation and sparseness. These statistics will be used to correlate with the degree of anxiety of the subjects.

TABLE I
Analysis of Overall Presentations

PPMCC	SAI		PRCS	
	SAICN	SAIEN	PRCSCN	PRCSEN
>0.6 or <-0.6	2	0	0	0
>0.5 or <-0.5	11	2	4	1

1) General Analysis: Based on different PPMCC, the most relevant features of speech are statistically evaluated. The detailed information is shown in Table1.

The SAI questionnaire includes 20 items (total scores from 20 to 80) to assess the feeling of the subjects 'at the moment' which is collected after each of experiment and the higher the score, the higher the level of anxiety the subjects are. The PRCS questionnaire include 30 items (total scores from 0 to 30) to assess the feeling of subjects when they are performing presentations and the higher the score, the higher the level of anxiety the subjects are.

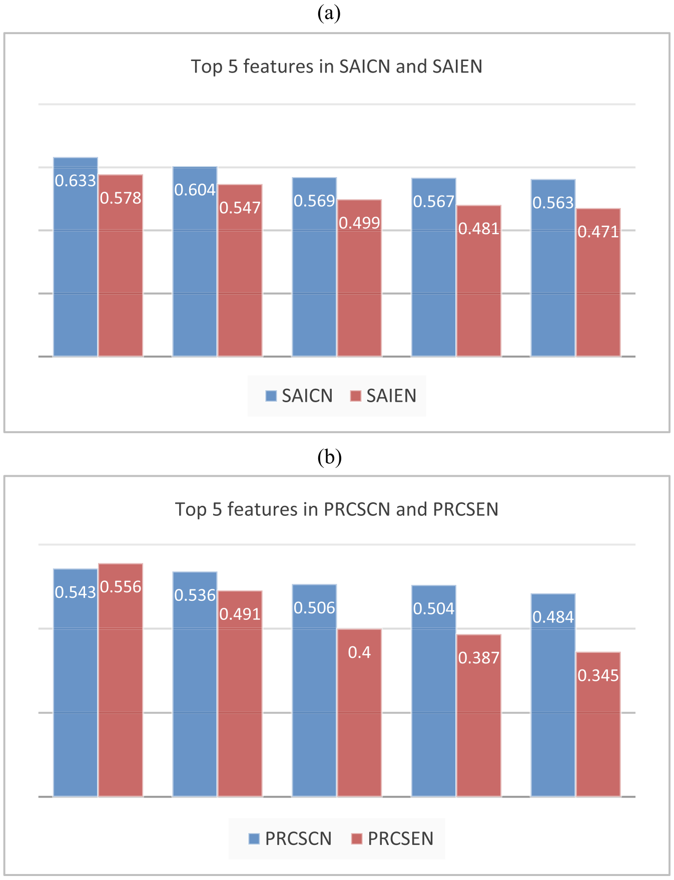
SAICN represents the SAI questionnaire completed after the subjects' Chinese presentation, and SAIEN represents the SAI questionnaire completed after the subjects' English presentation. This naming rule is equally applicable to PRCSCN and PRCSEN.

In Table 1, each row represents the number of relevant speech features that reach a certain level of PPMCC (0.5 or 0.6) which is in the first column of each table and each column of the table represents a specific questionnaire which reflects the anxiety state of the speaker at different times. For example, the number 2 of the third row and second column indicates there are 2 features (actually the Brightness_sp and MFCC5_sp) whose absolute correlation with SAICN questionnaire score exceeded 0.6 which means these two speech features have a high correlation with the degree of anxiety.

2) Chinese vs. English: Through the comparative analysis of Table 1, we can clearly find that the number of relevant features in Chinese presentation is always more than that in English presentation. In this case, there are 11 speech features (i.e., STE_std, STE_sp, MP_mean, MP_std, MP_sp, AMP_std, AMP_sp, Formant_mean, MFCC5_sp, Brightness_sp, Emphasis) whose absolute value of the correlation coefficient exceeds 0.5. However, when it comes to English, there are only two features (MFCC9_mean and Brightness_sp).

Tables 2a-b respectively show the most five relevant speech features associated with SAI and PRCS questionnaire scores and their correlation coefficients. In Table 2a and 2b, the red histograms represent the five speech features that are

TABLE II
(a) Analysis of Overall Presentations.
(b) Analysis of Overall Presentations


(a)

(b)

TABLE III
Number of Common Features in Top10

SAI	PRCS
5	3

most relevant to SAICN questionnaire scores and are arranged in descending order of correlation coefficients while the blue ones describe the five speech features that are most relevant to SAIEN questionnaire scores. Table 2a-2b show that of the five most relevant features, almost all of Chinese speech features have a higher correlation coefficient than English ones. Specifically, in these features, the average correlation coefficient for each feature of Chinese is 0.72(SAI) and 0.79(PRCS) higher than that in English. This means the speech features of Chinese presentation show a better consistency with the level of anxiety. Since Chinese is the native language of all subjects, this further implies that comparing with foreign language, when the subjects peak in native language, the speech features have higher potential to better reflect the degree of inner anxiety.

In order to further study the differences between Chinese and English presentations, Table 3 shows the number of common features in the most relevant 10 features both in Chinese and English presentations respectively.

TABLE IV
Number of Common Features in Top10

SAI		PRCS	
SAICN	SAIEN	PRCSCN	PRCSEN
0	2	0	2

Table 3 shows the number of common most 10 relevant features both in Chinese and English presentation. In particular, in PRCS questionnaire there are only 3 features in the top 10 most relevant features that have relatively high correlation with psychological anxiety in both Chinese and English speeches. This shows that when the subjects speak in different languages, the features of the voice that indicate the degree of inner anxiety are not the same. The reason for this is attributed to the different language used while they behave with the different verbal skill and the occurrence of different habits. This also shows that when the subjects use phonetic features to characterize the degree of psychological anxiety, different language reflects the different voice features.

3) Male vs. Female: Table 4 show the number of common most 10 relevant features both in male and female presentation. Through Table 4 we find that the 10 most relevant features of male and female are almost different in both Chinese and English. For example, in Chinese presentation, there is no common feature in the most 10 features in the SAI questionnaire, and in SAIEN the number of features that both in top 10 features of male and female is only 2. This shows that the speech features which can indicate men's anxiety are not the same as women's. The reason for this phenomenon can be attributed to male and female speaking with different voice frequencies. This means that we should distinguish between men and women when charactering anxiety with speech features.

B. Activity Analysis

In order to perform comparative analysis, we have selected data that contain both right hand and left hand data set form the whole collected data set.

For each of the collected samples, 8 types of features are calculated i.e., mean, standard deviation, short-time energy, entropy, correlation between different axis, pitch, roll and peak value in frequency domain. These features will be used to correlate with the degree of anxiety of the subjects.

1) General Analysis: In this section, based on different correlation coefficients, the most relevant features of activity are statistically evaluated. The detailed information has been summarized in Table 5a-b. The specifications of the form are identical to those of the speech analysis form. For the activity capturing system, it is found that the devices worn on the left hand and right hand have generated different data, especially the axis of the accelerometer and gyroscope. Thus, the later analysis of the activity data set will be distinguished whether the data is collected by the left hand or right hand.

2) Chinese vs English: Based on different PPMCC, the most relevant features of physical activity are statistically evaluated. The detailed information is shown in Table 5a-b.

TABLE V

(a) ANALYSIS OF OVERALL CHINESE PRESENTATIONS.
(b) ANALYSIS OF OVERALL ENGLISH PRESENTATIONS

(a)

PPMC	SAICN		PRCSCN	
	Left hand	Right hand	Left hand	Right hand
>0.5 or <-0.5	0	0	0	1
>0.4 or <-0.4	2	3	1	3

(b)

PPMC	SAIEN		PRCSEN	
	Left hand	Right hand	Left hand	Right hand
>0.5 or <-0.5	0	2	4	4
>0.4 or <-0.4	0	5	6	7

TABLE VI

COMPARISON OF CALCULATION PERFORMANCE

Computing Time (s) Speech Signal (s)	Matlab 2016a (s)				Wearable device (s)
	Energy	Entropy	Formant	Total	Total
60	0.045	0.058	0.367	0.470	34
230	0.134	0.225	1.207	1.566	126
300	0.169	0.294	1.503	1.966	162
600	0.331	0.579	2.926	3.836	336

In Table 5a-b, each row represents the number of relevant activity features that reach a certain level of PPMCC (0.4 or 0.5) which is in the first column of each table and each column of the table represents a specific questionnaire which reflects the anxiety state of the speaker at different times. For example, the number 2 of the fourth row and second column in Table 5a indicates there are 2 features (actually the g_y_std and Peak value) whose absolute correlation with SAICN questionnaire score exceeded 0.4 which means these two speech features have a high correlation with the degree of anxiety.

Through the comparative analysis of Table 5a-b, we can clearly find that the activity features associated with the state of anxiety in English presentations are always more than those in Chinese presentations. This phenomenon occurs simultaneously in the left and right hand scenes. When the subjects make Chinese presentations, there is only one activity feature (a_{yz}_cov) where absolute value of the correlation coefficient exceeds 0.5 in PRCS questionnaire. However, when it comes to English presentations, the number significantly increases to four (i.e., a_{RMS}_mean, a_{RMS}_power, g_x_std, peak number). One reason for this is that when the subjects speak their mother tongue, behavioral action will be more natural, which will enable them to better control their body language. On the contrary, when the subjects speak in non-native language, their behavioral action changes and they use the body language as much as possible to complement their presentations. Thus, the body language reflects the degree of the subjects' inner anxiety.

3) Left Hand vs Right Hand: When the subjects give Chinese presentation, it is observed that the number of hand activity features is less than that in English. In Table 5a-b, we can clearly find that compared to the left hand, right hand action is more able to reflect a male's anxiety state. There are 6 activity features of right hand (a_{RMS}_power, a_{RMS}_std, a_y_std, a_z_std, a_zg_z_cov, peak value) where the absolute value of the correlation coefficients exceeds 0.5 in English presentation. However, when it comes to left hand, it is found that there is no activity feature where the absolute value of the correlation coefficients even exceeds 0.4. This indicates that, compared to the left hand, the inner anxiety state is often reflected by the action of the right hand which is the dominant hand of most subjects.

V. DISCUSSION

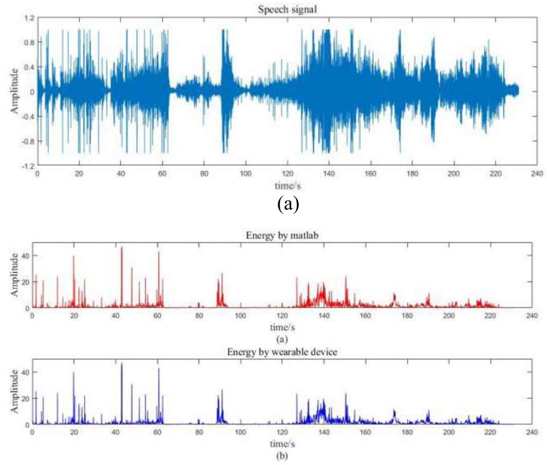
A. Integrated Feature Extraction Algorithm on Board

In this project, we use Matlab software to process the collected data off-line. On the contrary, we have fully considered that the collected data processed by the smart device. The Table 6 describes the comparison of computing time between microcontroller and Matlab.

We run Matlab 2016a on the PC with i7-6700 CPU@3.4GHz and the wearable device run at 168MHz. Several examples of feature calculation have been compared, which is Energy, Entropy and Formant of speech data. As is shown in Table 6, although the computing time of the wearable device is longer than those in Matlab, we can infer the possibility of performing features calculation while collecting data because the computing time is much shorter than the collecting time. The wearable device run an embedded real-time operating system where the tasks are separately processed and these include data collection, a write data to files through 'FatFs' as it writes the sensor data file every T second. Once a file is closed, the data processing task starts reading the file to RAM and processing the data. Because it is a multi-tasking operating system, the old data processing and the acquisition of the new data can be processed in 'parallel'. Although this is pseudo-parallel, the delay between new data collection and old data processing is T. When T is small enough, we can actually think this is parallel.

Fig. 6. shows the comparison of energy calculated by Matlab and the wearable device. Through the comparative analysis in Fig. 6(b), we can clearly find that the feature energy computed by the wearable device are similar to Matlab.

In order to further verify the correctness of the feature extraction algorithm of the hardware system, we have conducted the error analysis by comparing features calculated by Matlab and that by the wearable device. Fig. 7 shows that the error of energy ranges from 0 to 4×10^{-4}, and it is below 1×10^{-4} for most of the time. This proves that our algorithm on hardware system is almost the same as it in Matlab.



(a)

(b)

Fig. 6. Comparison of Features Between Matlab and Wearable Device. (a) Speech signal. (b) Energy calculated by Matlab and wearable device.

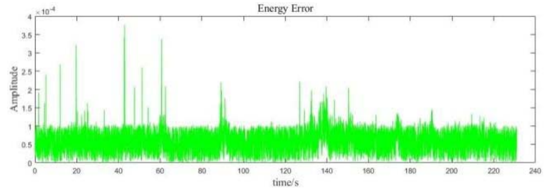


Fig. 7. Energy error analysis between Matlab and wearable device.

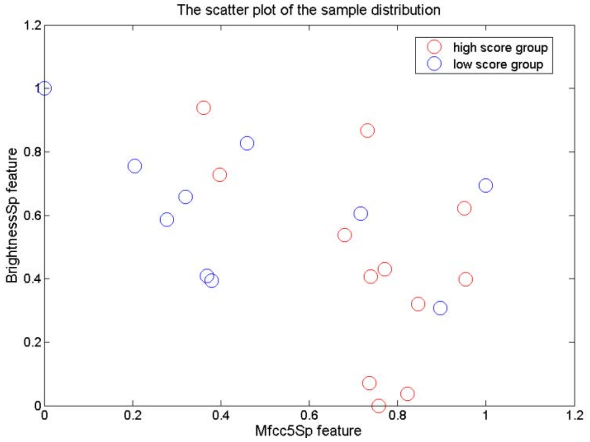


Fig. 8. The scatter plot the sample distribution.

B. Classification Analysis

We used K-means algorithm to do the classification analysis. In particularly, we selected the analysis results in Table 2 in which the most two features (MFCC5_sp and Brightness_sp) associated with SAICN are selected. We take the median value of the SAICN scores as the criterion of classification. Fig. 8. shows that the distribution of the samples with speech features of Brightness_sp and MFCC5_sp. The red circles represent the subjects who get high score which indicates a high level of anxiety and the blue circle represent the subjects who get low score which indicates a low level of anxiety.

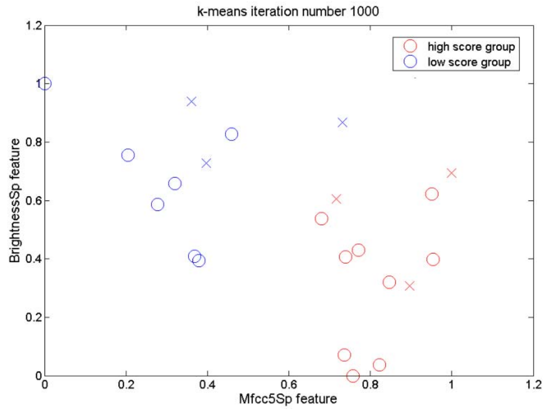


Fig. 9. Classification result of K-Means algorithm.

We initialize the K-means as iteration number of 1000. After converge, it automatically separate two groups with low and high score which is shown in Fig. 9.

In Fig. 9, we clearly see the samples classified by the K-Means algorithm into two groups, and the samples of misclassification is represented by forks. As shown in Fig. 9, the total number of samples is 22, and number of samples correctly classified is 16. Thus, the correct rate of K-means algorithm classification based on the speech features of Brightness_sp and MFCC5_sp is 72.73% which means the two features have a certain reference value in assessing the degree of anxiety.

VI. Conclusion

In this paper, a novel platform and approach for the analysis of human physical and vocal activities and their associations with anxiety has been proposed. The proposed method enjoys at least three significant advantages. Firstly, several features of activity and audio data that most relate to mental health especially anxiety state have been found. Secondly, a novel activity capturing system has been proposed to capture the unobtrusively and continuously collect acoustic and physical activity signals in completely natural and unpredictable situation. Finally, the speech features extraction algorithm is embedded in the wearable device for protecting privacy so that there is no need to record the wear's raw speech data.

References

[1] A. Pentland, "Socially aware, computation and communication," *Computer*, vol. 38, no. 3, pp. 33–40, Mar. 2005.

[2] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1743–1759, 2009.

[3] R. W. Picard, "Affective computing: Challenges," *Int. J. Hum.-Comput. Stud.*, vol. 59, nos. 1–2, pp. 55–64, 2003.

[4] M. S. H. Aung *et al.*, "Leveraging multi-modal sensing for mobile health: A case review in chronic pain," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 5, pp. 962–974, Aug. 2016.

[5] P. Gope and T. Hwang, "BSN-care: A secure IoT-based modern healthcare system using body sensor network," *IEEE Sensors J.*, vol. 16, no. 5, pp. 1368–1376, Mar. 2015.

[6] R. Wang, "CrossCheck: Toward passive sensing and detection of mental health changes in people with schizophrenia," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 886–897.

[7] J. Costa, A. T. Adams, M. F. Jung, F. Guimbetiere, and T. Choudhury, "EmotionCheck: Leveraging bodily signals and false feedback to regulate our emotions," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. (UbiComp)*, 2016, pp. 758–769.

[8] V. W. S. Tseng, M. Merrill, F. Wittleder, S. Abdullah, M. H. Aung, and T. Choudhury, "Assessing mental health issues on college campuses: Preliminary findings from a pilot study," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 1200–1208.

[9] L. Sun, K. W. Axhausen, D.-H. Lee, and X. Huang, "Understanding metropolitan patterns of daily encounters," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 34, pp. 13774–13779, 2013.

[10] Y.-A. de Montjoye, L. Radaelli, and V. K. Singh, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, no. 6221, pp. 536–539, 2015.

[11] F. Buttussi and L. Chittaro, "MOPET: A context-aware and user-adaptive wearable system for fitness training," *Artif. Intell. Med.*, vol. 42, no. 2, pp. 153–163, 2008.

[12] F. Gu *et al.*, "RunnerPal: A runner monitoring and advisory system based on smart devices," *IEEE Trans. Serv. Comput.*, to be published.

[13] G. Mois, T. Sanislav, and S. C. Folea, "A cyber-physical system for environmental monitoring," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 6, pp. 1463–1471, Jun. 2016.

[14] A. Gómez-Goiri, P. Orduña, J. Diego, and D. López-de-Ipiña, "Otsopack: Lightweight semantic framework for interoperable ambient intelligence applications," *Comput. Hum. Behav.*, vol. 30, no. 30, pp. 460–467, 2014.

[15] C. Nass and S. Brave, *Voice Activated: How People are Wired for Speech and how Computers will Speak With Us*. Cambridge, MA, USA: MIT Press, 2004.

[16] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," *Psychol. Bull.*, vol. 111, no. 2, pp. 256–274, 1992.

[17] A. Pentland, "Social dynamics: Signals and behavior," in *Proc. Int. Conf. Develop. Learn.*, 2004, pp. 263–267.

[18] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[19] E. Vyzas and R. W. Picard, "Affective pattern classification," in *Proc. AAAI Fall Symp. Ser. Emotional Intell. Tangled Cognit.*, 1998, pp. 176–182.

[20] I.-C. Chang, J.-H. Yang, and Y.-H. Liao, "Multi-camera based social network analysis," in *Proc. IEEE 8th Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.*, Jul. 2012, pp. 174–177.

[21] H. R. Bernard, P. Killworth, D. Kronenfeld, and L. Sailer, "The problem of informant accuracy: The validity of retrospective data," *Annu. Rev. Anthropol.*, vol. 13, no. 1, pp. 495–517, 1984.

[22] S. R. Corman and C. R. Scott, "A synchronous digital signal processing method for detecting face-to-face organizational communication behavior?" *Soc. Netw.*, vol. 16, no. 2, pp. 163–179, 1994.

[23] K. Yatani and K. N. Truong, "BodyScope: A wearable acoustic sensor for activity recognition," in *Proc. ACM Conf. Ubiquitous Comput.*, 2012, pp. 341–350.

[24] D. Wyatt, J. Bilmes, T. Choudhury, and J. A. Kitts, "Towards the automated social analysis of situated speech data," in *Proc. Int. Conf. Ubiquitous Comput.*, 2008, pp. 168–171.

[25] T. Choudhury *et al.*, "The mobile sensing platform: An embedded activity recognition system," *IEEE Pervasive Comput.*, vol. 7, no. 2, pp. 32–41, Apr./Jun. 2008.

[26] D. Eisenberg, J. Hunt, and N. Speer, "Mental health in American colleges and universities: Variation across student subgroups and across campuses," *J. Nervous Mental Disease*, vol. 201, no. 1, pp. 60–67, 2013.

[27] C. Blanco, *et al.*, "Mental health of college students and their non–college-attending peersresults from the national epidemiologic study on alcohol and related conditions," *Arch. Gen. Psychiatry*, vol. 65, no. 12, pp. 1429–1437, 2008.

[28] L. N. Ravindran and M. B. Stein, "The pharmacologic treatment of anxiety disorders: A review of progress," *J. Clin. Psychiatry*, vol. 71, no. 7, pp. 839–854, 2010.

[29] A. Grünerbl *et al.*, "Smartphone-based recognition of states and state changes in bipolar disorder patients," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 1, pp. 140–148, Jan. 2014.

[30] M. Frost, A. Doryab, M. Faurholt-Jepsen, L. V. Kessing, and J. E. Bardram, "Supporting disease insight through data analysis: Refinements of the monarca self-assessment system," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2013, pp. 133–142.

[31] L. Canzian and M. Musolesi, "Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2015, pp. 1293–1304.

[32] A. Heeren, G. Ceschi, D. P. Valentiner, V. Dethier, and P. Philippot, "Assessing public speaking fear with the short form of the personal report of confidence as a speaker scale: Confirmatory factor analyses among a French-speaking community sample," *Neuropsychiatric Disease Treat.*, vol. 9, pp. 609–618, May 2013.

[33] C. D. Spielberger, R. L. Gorsuch, and R. E. Lushene, *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA, USA: Consulting Psychologists Press, 1970.

[34] R. Hunter, "Insight vs. desensitization in psychotherapy," *Proc. Roy. Soc. Med.*, vol. 59, p. 1167, Nov. 1966.

[35] D. Watson and R. Friend, "Measurement of social-evaluative anxiety," *J. Consulting Clin. Psychol.*, vol. 33, no. 4, pp. 448–457, 1969.

[36] M. C. Melchers, M. Li, B. W. Haas, M. Reuter, L. Bischoff, and C. Montag, "Similar personality patterns are associated with empathy in four different countries," *Frontiers Psychol.*, vol. 7, p. 290, Mar. 2016.

[37] M. Rosenberg, *Society and the Adolescent Self–Image*, vol. 11. Princeton, NJ, USA: Princeton Univ. Press, 1965.

[38] M. Sun and J. O. Hill, "A method for measuring mechanical work and work efficiency during human activities," *J. Biomech.*, vol. 26, no. 3, pp. 229–241, 1993.

[39] E. K. Antonsson and R. W. Mann, "The frequency content of gait," *J. Biomech.*, vol. 18, no. 1, pp. 39–47, 1985.

[40] K. Aminian, P. Robert, E. Jéquier, and Y. Schutz, "Incline, speed, and distance assessment during unconstrained walking," *Med. Sci. Sports Exerc.*, vol. 27, no. 2, pp. 226–234, 1995.

[41] C. V. C. Bouten, K. T. M. Koekkoek, M. Verduin, R. Kodde, and J. D. Janssen, "A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity," *IEEE Trans. Biomed. Eng.*, vol. 44, no. 3, pp. 136–147, Mar. 1997.

[42] M. L. Blum, "Real-time context recognition," M.S. thesis, Dept. Inf. Technol., ETH Zurich, 2005.

[43] N. Kern and B. Schiele, "Context-aware notification for wearable computing," in *Proc. 7th IEEE Int. Symp. Wearable Comput.*, 2003, pp. 223–230.

Jun Gu received the B.Sc. degree from the School of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing, China, in 2013. He is currently pursuing the M.Sc. degree in wearable device research with the University of Electronic Science and Technology of China, Chengdu, China. His research interests include embedded software development and digital signal processing.

Bin Gao (M'12–SM'14) received the B.Sc. degree in communications and signal processing from Southwest Jiaotong University, China, in 2005, and the M.Sc. (Hons.) degree in communications and signal processing and the Ph.D. degree from Newcastle University, U.K., in 2006 and 2011, respectively. He was a Research Associate with Newcastle University from 2011 to 2013, where he worked on wearable acoustic sensor technology. He is currently a Professor with the School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China. He has coordinated several research projects from the National Natural Science Foundation of China. His research interests include sensor signal processing, machine learning, social signal processing, and nondestructive testing and evaluation, where he actively publishes in these areas. He is a very active reviewer for many international journals and long standing conferences.

Yuanpeng Chen received the B.Sc. degree from the School of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Lanzhou, China, in 2016. He is currently pursuing the M.Sc. degree in speech scene recognition and machine translation with the University of Electronic Science and Technology of China, Chengdu, China. His research interests include machine translation, image identification, and speech scene recognition machine learning.

Long Jiang received the B.Sc. degree from the School of Information Engineering, Southwest University of Science and Technology, Mianyang, China, in 2012. He is currently pursuing the M.Sc. degree in control engineering and science with the University of Electronic Science and Technology of China, Chengdu, China. His research interests include intelligent hardware and wearable sensor.

Zhao Gao was born in Sichuan, China, in 1974. She is currently pursuing the Ph.D. degree in biomedical engineering with the School of Life Science and Technology. From 2012 to 2014, she was an Academic Visitor at the Cognition and Brain Science Unit, MRC, Cambridge, U.K., the Mind Research Network, and the University of New Mexico, Albuquerque, USA. She is also an Associate Professor with the School of Foreign Languages, University of Electronic Science and Technology of China. Her research interests are the neuroendocrinological mechanism of metaphor bias in social and emotional contexts, the foreign language classroom anxiety, and English for academic purpose reading pedagogy.

Xiaole Ma received the B.Sc. degree in applied psychology from Taiyuan Normal university, China, in 2012. She is currently pursuing the Ph.D. degree in biomedical engineering with the University of Electronic Science and Technology of China, Chengdu, China. Her research interests include cognitive and affective neural science and data analysis.

Yong Ma received the B.Sc. degree from the School of Mechatronics and Control Engineering, Hubei Normal University, Huangshi, China, in 2013. He is currently pursuing the M.Sc. degree in speech segmentation method based on wearable social perceptual system with the University of Electronic Science and Technology of China, Chengdu, China. His research interests include speech signal processing, feature extraction, and speech segmentation.

Wai Lok Woo was born in Malaysia. He received the B.Eng. (Hons.) degree in electrical and electronics engineering and the Ph.D. degree from Newcastle University, U.K. He is currently a Senior Lecturer and the Director of Operations with the School of Electrical and Electronic Engineering, Newcastle University. His major research is in the mathematical theory and algorithms for nonlinear signal and image processing. This includes areas of machine learning for signal processing, blind source separation, multidimensional signal processing, and signal/image deconvolution and restoration. He has an extensive portfolio of relevant research supported by a variety of funding agencies. He has published over 250 papers on these topics on various journals and international conference proceedings. He received the IEE Prize and the British Scholarship to continue his research work. He has served as a lead-editor of journals' special issues. He is currently an associate editor of several international journals.

Jikun Jin received the B.Sc. degree in electronic information engineering from the University of Electronic Science and Technology of China in 2017, where he is currently pursuing the M.Sc. degree. His research focuses on motions detection and recognition algorithm of the wearable device.