Wearable Audio Monitoring: Content-Based Processing Methodology and Implementation

Bin Gao, *Member, IEEE*, and Wai Lok Woo, *Senior Member, IEEE*

***Abstract*—Developing audio processing tools for extracting social-audio features are just as important as conscious content for determining human behavior. Psychologists speculate these features may have evolved as a way to establish hierarchy and group cohesion because they function as a subconscious discussion about relationships, resources, risks, and rewards. In this paper, we present the design, implementation, and deployment of a wearable computing platform capable of automatically extracting and analyzing social-audio signals. Unlike conventional research that concentrates on data which have been recorded under constrained conditions, our data were recorded in completely natural and unpredictable situations. In particular, we benchmarked a set of integrated algorithms (sound speech detection and classification, sound level meter calculation, voice and nonvoice segmentation, speaker segmentation, and prediction) to obtain speech and environmental sound social-audio signals using an in-house built wearable device. In addition, we derive a novel method that incorporates the recently published audio feature extraction technique based on power normalized cepstral coefficient and gap statistics for speaker segmentation and prediction. The performance of the proposed integrated platform is robust to natural and unpredictable situations. Experiments show that the method has successfully segmented natural speech with 89.6% accuracy.**

***Index Terms*—Audio detection and classification, social signal analysis, speaker segmentation, wearable device.**

I. Introduction

SOCIAL signal processing (SSP) [1], [2] is a hot research field, where intelligent devices sense and understand human social behavior. SSP has already attracted researchers in areas such as psychology, ambient intelligence, and healthcare. Standard methods to measure and evaluate SSP have issues. Monitoring humans is very expensive, is limited to a small number of people per observer, and may have interobserver reliability concerns. Using cameras is also expensive and the range of measurement is limited. Surveys suffer from subjectivity and memory effects. Thus, portable computing system-based intelligent devices capable of automatically capturing social signals in a persuasive manner offer an alternative.

Manuscript received May 24, 2013; revised December 12, 2013; accepted January 2, 2014. Date of publication February 5, 2014; date of current version March 12, 2014. This paper was recommended by Associate Editor L. Rothrock.

B. Gao is with the School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 610051, China (e-mail: bin_gao@uestc.edu.cn).

W. L. Woo is with the School of Electrical and Electronic Engineering, Newcastle University, Tyne and Wear NE1 7RU, U.K. (e-mail: w.l.woo@ncl.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2014.2300698

The idea of integrating wearable computing (WC) with the analysis of human social signals to generate more natural, flexible computing technology was introduced in [2]. This allows the SSP to be automated. In addition, the automatic discovery and characterization of human communication and social interaction would allow us to gather interaction data from large groups of people. Specially, the audio signal plays an import role in studying WC and can support automatic assessment of human social behavior. It can support multimodal tools and enhance researcher productivity.

Several WC projects have considered the use of speech and audio in the interface. Ubiquitous Talker [3] is a camera-enabled system that provides information related to recognized physical objects using a display and synthesized voice. A prototype-augmented audio tour guide [4] presented digital audio recordings indexed by the spatial location of visitors in a museum. SpeechWear [5] enabled users to perform data entry and retrieval using speech recognition and synthesis.

Human speech plays a significant role in social signal learning [3]. However, speech research in modeling conversations generally considers limited situation such as in meeting room scenarios [4], [5] or with acted speech [6], which is known to poorly reflect natural emotion [7]. The datasets that do capture real emotion [8]–[10] are generally limited to a handful of observations of each subject and cannot be used to compare one person's speech across different situations over time. Most are also recorded in relatively unnatural settings (such as in the case of television shows or interviews) that are not representative of everyday human communication.

There is a little research in naturalistic settings. In [10], this study only recorded short-time frames of a single participant in isolation. Some prior work has difficulties in disambiguating speakers. In particular, the speech signals from specific individuals are likely to be significantly attenuated relative to the ambient noise [11].

Thus, existing techniques for automatically learning social audio signals are limited and fall short of the success enjoyed in the other areas of SSP. In this paper, a novel system for extracting social-audio signals is proposed. The contributions are summarized as follows.

1) Developing a fully automatic, computationally efficient method of extracting and analyzing social-audio features from a target person using the wearable acoustic monitor (WAM) device.
2) Analyzing and assessing the performance of the novel development of speech and sound detection and classification algorithms based on block audio features. Seven well-established audio features were integrated to improve

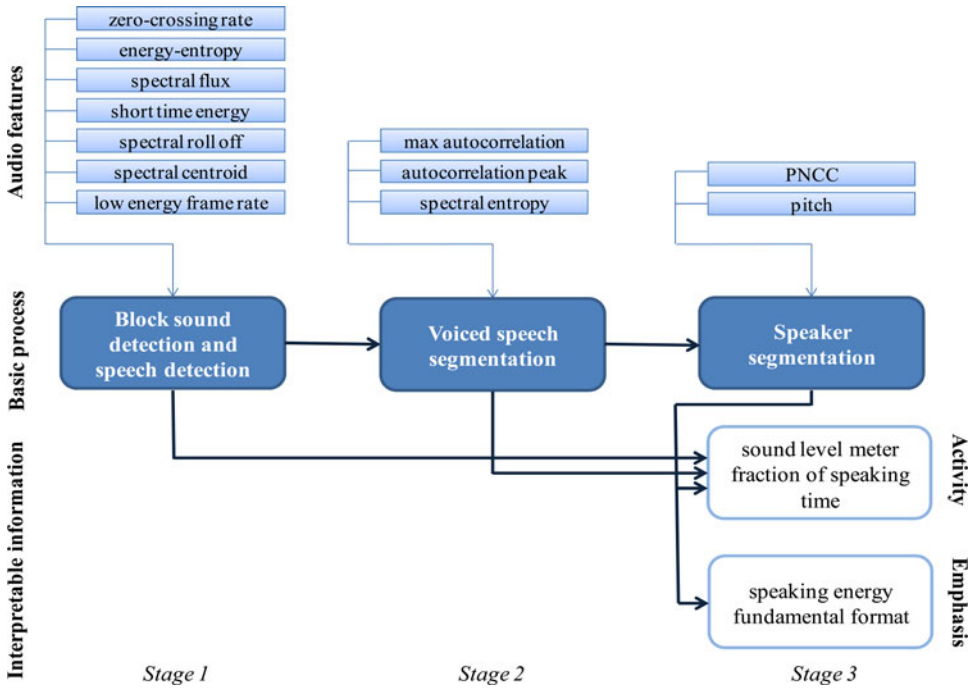

Fig. 1. Proposed social-audio signal extraction architecture.

the robustness of the detection system and to address any ambiguity in any single feature set.

3) Developing a novel speaker segmentation method based on a power normalized cepstral coefficient (PNCC) [12]. This novel method is highly desirable since it enables the noise speech signal to be segmented with significantly higher accuracy, when contrasted with other tradition features such as linear prediction cepstral coefficients (LPCC), mel-frequency cepstral coefficients (MFCC) [13], which only work well with clean speech signals. In addition, a gap statistic method [14] is incorporated as part of the overall system to predict the number of unknown speakers within the conversational data.

The paper is organized as follows. Section II introduces the proposed social audio signal extraction (SASE) architecture. In Section III, the "new speaker segmentation algorithm" as well as the "speaker number prediction technique" is derived. Experimental results and a series of performance comparison with alternative methods are presented in Section IV. Section V concludes the paper.

II. Proposed Social-Audio Signal Extraction Architecture

We design the WAM to continuously collect audio signals in completely natural and unpredictable situations. The proposed SASE architecture can be divided into three stages.

Stage 1: Block detection of sound and speech and classification of both environmental and speech sounds.

Stage 2: Voiced and nonvoiced speech segmentation and sound level meter calculation.

Stage 3: Individual speaker segmentation, clustering, unknown number of speaker prediction, and social signal calculation.

These stages are explained below, and the proposed architecture is shown in Fig. 1.

A. Implementation of Stage 1

1) Sound and Speech Detection: Sound detection [15] is a useful preprocessing tool which periodically explores the threshold detector output to judge the presence of a strong audio signal. We analyze blocks of audio data from the WAM recorder (i.e., one block of data consisted of 30 s of microphone data). If a strong audio signal is detected (the energy of the signal beyond threshold) more than 50 out of 1000 times within a 0.5-s period, the sound detection stage assumed the presence of a strong audio signal, and thus preserved the whole block of data.

Simultaneously, the proposed platform starts the speech detection process to detect the presence of voice. In the speech detection process, a large number of signal features are employed to discriminate between different environmental sounds and speech. Once the sound signal has been detected within the block data, the block signals are then broken into short-term, nonoverlapping windows (frames) of 50 ms. For each frame, six features [16]–[18] are calculated, namely, zero-crossing rate, energy-entropy, spectral flux, short-time energy, spectral roll off, and spectral centroid. An additional feature, low-energy frame rate, has also been incorporated (calculated by every window of 64 frames). For each of the first six feature sequences within a block, a simple statistic is calculated (the standard deviation divided by the mean value). This step leads to six single statistic values. These seven values are the final feature values that characterize the input audio signal. In the proposed speech detection

process, all features are combined to classify the speech and environmental sound to improve the robustness of the classification system and address any ambiguities in any single feature set.

2) Sound and Speech Classification: In order to obtain a set of baseline measurements for the above features for "typical" speech and environment sound, a set of training data is used. Each of the seven feature calculations is performed on all training data samples, to obtain a set of values that can be considered typical of speech and environmental sound signals. In our training phase, the training speech data consisted of five types of speech: two sets of single male speech, two sets of single female speech, one set of male and male conversation, one set of male and female conversation, and one set of female and female conversation. Each set of training data for the speech sets were 5 min long. The training datasets for environmental sounds consisted of walking sounds, mouse click sounds, keyboard typing sounds, scratching sounds in the pocket, water sounds in the bathroom, noise from the underground train, street noise, hoover sounds, music, sound from kitchen, sound from within church. All together, the ambient sound dataset contained approximately 100 min of environmental sound.

Once a set of baseline feature measurements is established, a system is needed to classify test data as having features more like speech or other sounds. We tested several classifying algorithms before choosing the most suitable classifier. These results will be discussed in Session IV. Based on the results, a k-nearest neighbor (KNN) algorithm [19] is chosen for simplicity and efficiency. Also, the range covered by each of the seven features is normalized to a unit norm in order to weight each feature equally in the distance calculations. This ensures that larger-valued features will not dominate over smaller ones.

B. Implementation of Stage 2

In order to measure the sound level of ambient environment, for blocks containing nonspeech signals, an A-weighted sound level meter [20] is designed. The fast Fourier transform (FFT) algorithm is used to estimate the frequency spectrum of a windowed set of samples. The frequency spectrum is then weighted using a closed-form expression for the A-weighting filter, and the average signal energy is then estimated in the frequency domain using the Parseval relation.

Human speech can be divided into voiced speech and nonvoiced speech. Voiced speech [21] is defined as speech generated from the vibrations of the vocal chords; it includes all vowel sounds and some consonant sounds. In contrast to the nonspeech signal blocks, if a speech block is detected, the first stage then activates the voice and nonvoiced speech segmentation system. The role of this stage is to pick high-quality speech frames from the block speech signal and to discard low-quality speech frames as well as frames of silence, which occur naturally from brief pauses (i.e., any period of silence less than 0.2 sec. was considered a pause) in human speech.

1) Features and Segmentation: We use three features [22] for the voice and nonvoiced segmentation conducted on the blocks of speech data, namely, 1) noninitial maximum of the normalized noisy autocorrelation, 2) number of autocorrelation peaks, and 3) normalized spectral entropy. These are computed on a per-frame basis. In the parameters setting, we work with a frame size of 32 ms and an overlap of 16 ms between frames. Once all features have been calculated, we then use a two-level hidden Markov model (HMM) [36] to segment the speech block data into voiced and nonvoiced segments.

C. Implementation of Stage 3

The idea of speaker segmentation can be considered as a partition of speech signal into subsets and a judgment whether one person is present or belong to others [23]. To achieve this, a common way is to search for a change point that represents that a possible change of speaker may have occurred. Once all change points have been found, the speaker clustering algorithm can classify the subsets based on who is now speaking. Generally, the statistical language modeling methods such as MFCC for speech feature calculation has greatly improved the performance of speaker recognition systems in 'clean' environments [12]. Nevertheless, the accuracy still degrades significantly in noisy environments especially in cases when the audio data are recorded in completely natural and unpredictable situations.

In this study, we deployed a more sophisticated audio feature, PNCC as our data are recorded in unpredictable situations. The overall speaker segmentation system consisted of several steps: 1) calculate speech PNCC features, as well as pitch information [25], to improve robustness and segmentation efficiency; 2) employ a change point detection algorithm [23] to find the possible speaker change point given in the feature space; and 3) classify WAM speaker (i.e., the wearer of the WAM device) and other unknown speakers, based on a Gaussian mixture model (GMM) classifier [24], and predict the number of unknown speakers based on the gap statistic method. All details of the proposed segmentation system are described in Section III. Once the above three stages have been implemented, the sound level meter and the classified speaker information can be further analyzed so that they could possibly serve as predictive social-audio signals. Two well-known social signals, namely, activity and emphasis [26], could be computed from the captured audio to infer the social events.

1) Activity: Activity is defined as the fraction of time a person is speaking. The percentage of speaking time is known to be correlated with interest level [27] and extraversion [28]. In the domain of negotiation, the authors in [29] found a trend whereby extraversion correlated positively with individual outcomes in an integrative bargaining task, similar to the one used in the present study. In our case, we are targeting the social signal of the fraction of speaking time which could be directly calculated by using the segmented WAM speech frames. In addition, we are also able to predict the number of different places that the WAM speaker engaged in conversational activity according to the sound level meter. In general, the sound level can be divided into three phases based on Monte Carlo realizations. The results are tabulated in Table I.

The results in Table I are calculated by averaging the outcomes of chunk data (each chunk consists of 3-min audio data).

TABLE I
SOUND LEVEL METER CLASSIFICATION

Environment	Noisy	Normal	Quiet
E_{dB} (dB)	$E_{dB} > 82$	$60 < E_{dB} \leq 82$	$E_{dB} \leq 60$

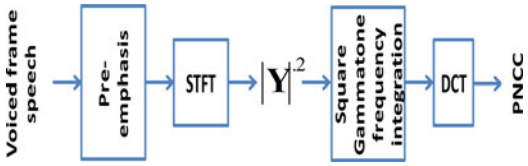

Fig. 2. Computation of PNCC. "**Y**" is the spectrum of STFT of voiced frame speech.

2) Emphasis: Emphasis is measured by variations in speech prosody—specifically, variation in pitch and volume. Prosody refers to speech features that are longer than one phonetic segment and are perceived as stress, intonation, or rhythm [31]. To measure emphasis, we began by extracting the speaking energy and the fundamental format for voiced segments within the speech block data. We then calculate the standard deviation of the energy and frequency measures, each scaled by their respective means.

III. SPEAKER SEGMENTATION

A. PNCC Features

The most widely used audio feature extraction algorithms are MFCC and LPCC [13]. Even though many speech recognition systems [32] have obtained satisfactory performance based on the above two features, they are dependent on clean test environments. Recognition accuracy significantly degrades if the test environment is different from the training environment. These environmental differences might be due to additive noise, channel distortion, and acoustical differences between different speakers. The recently proposed audio feature PNCC has been developed to enhance the environmental robustness of speech recognition systems. The major innovations of this feature can be summarized as follows.

1) The use of a well-motivated power function that replaces the log function, and the use of a novel approach to the blind removal of background excitation based on medium duration power estimation. This normalization makes use of the ratio of the arithmetic mean to the geometric mean, which has proved to be a useful measure in determining the extent to which speech is corrupted by noise.
2) PNCC uses frequency weighting based on the gammatone filter shape [33] rather than the triangular frequency weighting or the trapezoidal frequency weighting associated with the MFCC and LPCC computation.
3) The Power-Bias Subtraction algorithm has proved to be a useful and easily computed way to characterize the data, and the power flooring technique can reduce spectral distortion between training and test sets for these regions [12].

The structure of the PNCC feature is shown in Fig. 2. A preemphasis filter of the form $H(z) = 1 - 0.97z^{-1}$ is applied first. The short-time Fourier transform (STFT) analysis is performed using Hamming windows of duration 2 s, with 10 ms for the time step for a sampling frequency of 16 kHz (where broader windows and longer time steps to smooth out the data), 40 gammatone channels. After passing through the gammatone channel, the power is normalized using peak power (i.e., the 95% of short-time power).

B. Change Detection Algorithm

The change detection algorithm can be summarized as follows. Once the block of the PNCC feature set is generated, the algorithm detects whether the points ahead are different from those we have already seen. If this is true, the current point is considered as a possible change of the speaker. Otherwise, the algorithm keeps going ahead. In order to quantify the difference between two points, the Kullback–Liebler divergence has been used achieve a distance measure [34]. Given two vector sequences $\mathbf{F}^a$ and $\mathbf{F}^b$, the symmetric Kullback–Liebler distance (KLd) between the two is

$$\mathrm{KLd}\left(\mathbf{F}^a, \mathbf{F}^b\right) = \int_x \left[p_{\mathbf{a}}(x) - p_{\mathbf{b}}(x)\right] \log \frac{p_{\mathbf{a}}}{p_{\mathbf{b}}}. \tag{1}$$

We assume $p_a \sim N(\mathbf{u}_a, \mathbf{\Sigma}_a)$ and $p_b \sim N(\mathbf{u}_b, \mathbf{\Sigma}_b)$ are T-variate Gaussian distribution which can be expressed as

$$p(x) = \frac{1}{(2\pi)^{T/2}|\mathbf{\Sigma}|} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{u})^{\mathbf{T}} \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{u})\right\} \tag{2}$$

where "**u**" is a mean vector, "**Σ**" is a covariance matrix and superscript "**T**" denotes transpose. Thus, (1) can be expressed as

$$\begin{aligned}\mathrm{KLd}\left(\mathbf{F}^a, \mathbf{F}^b\right) &= \frac{1}{2} Tr\left[\left(\mathbf{\Sigma}_{\mathbf{a}} - \mathbf{\Sigma}_{\mathbf{b}}\right)\left(\mathbf{\Sigma}_{\mathbf{a}}^{-1} - \mathbf{\Sigma}_{\mathbf{b}}^{-1}\right)\right] + \frac{1}{2} Tr \\ &\quad \times \left[\left(\mathbf{\Sigma}_{\mathbf{b}}^{-1} - \mathbf{\Sigma}_{\mathbf{a}}^{-1}\right)\left(\mathbf{u}_{\mathbf{a}} - \mathbf{u}_{\mathbf{b}}\right)\left(\mathbf{u}_{\mathbf{a}} - \mathbf{u}_{\mathbf{b}}\right)^{\mathbf{T}}\right]\end{aligned} \tag{3}$$

where "Tr" denotes trace operation. $\mathbf{F}^a$ represents the sequence of feature vectors extracted from the analysis window of size 2 samples that begins at the given point, and $\mathbf{F}^b$ denotes the feature vectors extracted from the analysis window that follows directly after the first, as the same size samples. If the distance between either of these is above a certain threshold, the current point is considered to be a change point. In our experiments, two thresholds, ε_A and ε_B are employed. The first is calculated as the mean of a window around the given point, multiplied by a constant, i.e.,

$$\varepsilon_A = \alpha \frac{1}{2N_1} \sum \mathbf{F} \tag{4}$$

where "**F**" represents the feature vector and summation symbol represent sum across all the elements in **F**, "N_1" is the size of the window in one direction, and "α" is a constant used to tune the recall of the threshold. The experiments section will detail the performance of this variable. However, in order to be taken as a true change point, the given value must also be greater than

ε_B, which is calculated by

$$\varepsilon_B = \sigma_F + \beta \frac{1}{2N_2} \sum \mathbf{F} \tag{5}$$

where "σ_F" is the standard deviation over the windowed area, "N_2" is the size of the window, and "β" is another constant. The first threshold ensures that the given value is greater than the surrounding area, calculated over a small window. The second threshold is calculated over a larger window, and ensures that the change takes into account the general trend of the data's changes. Current values for α and β are 1.4 and 1, respectively. The window sizes N_1 and N_2 are currently set to 3 and 4 s, respectively. These setting are based on the Monte Carlo experiment of 100 independent realizations and better segmentation results are obtained from these settings.

C. Clustering

After the segmentation process, we assume that each of the recovered segments represents a potential speaker; therefore, we create models for each speaker. Traditionally, speaker verification models are created using GMMs [36]. GMMs are a hybrid between parametric and nonparametric statistical methods, which have been used with considerable success in areas where there is little information available for model creation and one cannot assume a fixed distribution generating the data. The EM method is used here for training purposes. In speech verification, the average number of mixtures considered to be sufficient to model a speaker ranges from 32 to 256 [15], [35]. We train the GMM for WAM speaker models and use a universal background GMM to represent all unknown speakers [24]. The features that were used for training included the PNCC mentioned in the previous paragraph, along with the fundamental frequency (pitch). For every potential speaker, the WAM speaker and unknown speakers were identified, as the one talking during the segment is the one whose speaker model gives the highest sum of log likelihood across the segment.

D. Predicting Number of Unknown Speakers

Given the ambiguity in the "unknown speaker" clustered segments, we need to predict how many unknown speakers are involved in the conversation. In our case, the standard k-means algorithm [36] and the gap-statistic technique are used to achieve this purpose. The issue of cluster validation is an open problem in the area. One method of estimating the number of clusters is to plot the within-cluster sum of squares graph for each number of possible clusters (i.e., from 1 to the total number of segments). The idea here is that the sum of squares can only decrease as the number of clusters increases; after a certain point, the sum of squares should decrease more slowly than for previous clusters. This point is called the "elbow," and is determined to be the optimal number of clusters. The overall steps of the prediction method are summarized as follows.

1) Given all GMM models for each separate segment of the unknown speaker PNCC feature, use k-means algorithm for clustering. Varying the total number of clusters $k = 1, \ldots, K$ and calculating the within-dispersion measures W_k $k = 1, \ldots, K$.
2) Generate D number of references sets using the uniform prescription, and cluster each one giving within-dispersion measures W^*_{kd}, $d = 1, \ldots, D$, $k = 1, \ldots, K$. The gap statistic are estimated as

$$\text{Gap}\,(k) = \frac{1}{D} \sum_d \log\left(W^*_{kd}\right) - \log\left(W_k\right). \tag{6}$$

3) Let $l = \frac{1}{D} \sum_d \log\left(W^*_{kd}\right)$ and the standard deviation can be expressed as

$$sd_k = \left[\frac{1}{D} \sum_d \left\{ \log\left(W^*_{kd}\right) - l \right\}^2 \right]^{\frac{1}{2}} \tag{7}$$

and $m_k = sd_k \sqrt{(1 + 1/D)}$ to account for the modeling error in l. Finally, the estimated number of clusters can be obtained:

$$\hat{k} = \arg\min_k \left[\text{Gap}(k) \geq \text{Gap}(k+1) - m_{k+1} \right]. \tag{8}$$

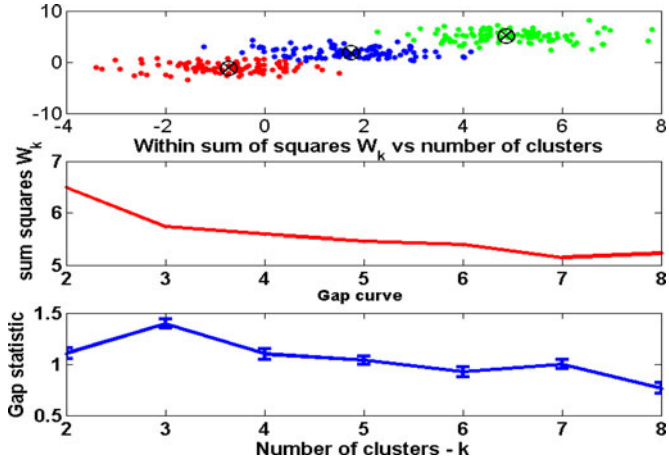

Fig. 3. Example of gap statistic prediction.

For the reference sets, the simplest choice is to generate each reference feature uniformly over the range of the observed values for that feature. The specific calculations of W_k and W^*_{kd} are referred to in [30]. Fig. 3 shows an example of using the gap statistic method to predict an optimal number of clusters. In Fig. 3, we have generated three normally distributed datasets, where green, blue, red dots represent the different means. Each dataset is a 100×2 matrix from a normal distribution with mean $\mathbf{u}_1(5, 5)$, mean $\mathbf{u}_2(-1, -1)$, and $\mathbf{u}_3(2, 2)$, respectively. All datasets have unique standard deviation with $\sigma(1, 1)$. The data (top panel) fall in three distinct clusters. We are attempting to predict the number of clusters given in these datasets. The middle panel shows step one of the gap statistic method which calculates the within-dispersion measures W_k $k = 1, \ldots, K$ (here, the number of clusters vary from $k = 2, \ldots, 8$). The bottom panel shows step two of the gap statistic method, whereby the gap value (with ± 1 standard error bars) can be calculated by using (6). The final predicted number of clusters can be obtained by using (6) where the optimized number of clusters is $\hat{k} = 3$.

IV. Evaluation and Analysis

A. Prototype of the Wearable Acoustic Monitor

To prototype the WAM, a custom four-layer printed circuit board (PCB) has been developed. The PCB measured 50 × 50 mm and featured a PIC32MX695F512 H 32 bit microcontroller from Microchip Inc.

The prototype also featured a MEMS microphone (WM1720 from Wolfson Microelectronics) which is interfaced to a voice band audio CODEC chip (AIC111 from Texas Instruments). The CODEC chip is also interfaced to an auxiliary microphone which was used for experimenting with microphones of different frequency responses and also optimizing the microphone position in the final design. Audio can be sampled from the microphone in mono up to 40 kHz in 16 bit resolution; however, we find 8 kHz and 16 kHz to be the most appropriate compromise between low power and performance. Once the samples are gathered from the microphone, they are written to a high-capacity secure digital (SDHC) card of up to 32-GB capacity.

The prototype also features an OLED display and an RGB LED for communicating a state during development. To collect naturalistic data for the algorithm development, a small, low power, wearable device is necessary. At the design stage, any component preventing us from collecting live data has been excluded from the design.

Once the prototype PCB is evaluated, a second version of the PCB is designed. This PCB was designed to be suitable for a wrist worn enclosure which was deemed to be a mounting point of high social acceptability for a wide range of people. The same components that are chosen for the prototype PCB are placed on this second PCB (with the exception of the auxiliary microphone input) and a resizable silicate wrist band was designed to house the electronics. A rechargeable lithium-polymer battery is chosen for this circuit.

B. Experimental Setup

The proposed SASE system is tested on recorded audio signals. For the training phase of stage one and stage three, we collect clean voice data from 30 speakers (15 male and 15 female). The age distribution of our participants is between 20 and 70 years. For each speaker, we have collected approximately 5 min. of "clean" speech.

The formats consisted of 8-kHz sampling rate, 16-bit mono and 16-kHz sampling rate, 16-bit mono. In order to compare the efficiency of using PNCC features with other well-known audio features such as MFCC and LPCC, we use a training set from the established English Language Speech Database for Speaker Recognition (ELSDS) [37] database. This database consists of 22 speakers: 10 female, and 12 male, with an age span of 24 to 63 years. The training set consists of seven reading paragraphs, which include 11 sentences and a set of 44 random sentences. These paragraphs have been developed to ensure that they capture all of the possible speech sounds (phonemes) that are used within the English language (these include vowels, consonants and diphthongs). Altogether there are 154 (7 × 22) utterances in the training set. On average, the duration it takes to read the training data was 78.6 s for male, 88.3 s for female, 83 s for all.

For the testing phase of stage I, 35 chunks (each chunk being 3 minutes long) of completely natural and unpredictable conversational interaction, between two people or more, or single speech, and another 35 chunks of different ambient noise are employed. All test speech samples are collected either from noisy environments or quiet environments and the WAM recording device is positioned on the main speaker's wrist (also referred to as the WAM speaker).

Several well-established classifiers are compared in Stage 1 to obtain the best accuracy results. These classifiers [39] include linear discriminant analysis (LDA), NaiveBayes classifier, Decision Tree (DT), K-nearest neighbor (KNN), and support vector machine (SVM).

In the testing phase of Stage 2, the algorithms are evaluated on: 1) natural and unpredictable recorded speech signals; and 2) clean speech which is corrupted by artificial noise. The artificial noise is chosen from the AURORA database [38].

In Stage 3 testing, two evaluation processes are considered: 1) testing and obtaining the suitable audio features using the speech recognition system. In this experiment, the ELSDSR database provided 44 (2 × 22) utterances. The duration for reading of the test data, on average, is: 16.1 s (male); 19.6 s (female); 17.6 s (for all). 2) Testing the proposed speaker segmentation on completely natural and unpredictable conversation. In order to obtain ground truth results, the test speech chunks needed to undergo human annotation. We implement this by using ELAN [40], software designed for the manual annotation of audio and video data, and three separate annotators per audio, to control for any inter-annotator reliability issues. For the evaluation process, we consider a measured binary classification problem for evaluating the results. We define speech classification as positive and non-speech classification as negative. The evaluation of each of the category combination involves computing the resulting sensitivity, specificity, and accuracy as follows:

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$
$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{9}$$

where TP = true positive, FP = false positive, TN = true negative, and FN = false negative.

C. Stage 1 Evaluation

Table II shows the comparison of our proposed speech and sound classification method based on the sampling frequency 8 and 16 kHz, under various classifiers. The classification results for all classifier types based on the 8 kHz sampling frequency gives an average accuracy of 89.72%, while a higher performance is attained by the 16 kHz sampling frequency with an average accuracy of 92.8%. This leads to an improvement of 2.28%. However, the best results for both sampling frequency are obtained by using a KNN classifier with an accuracy of 92.8% and 94.3%, respectively.

TABLE II
EVALUATION RESULTS OF STAGE 1 PROCESS

Methods	8 kHz sampling frequency		
	Sensitivity	Specificity	Accuracy
LDA	90.1%	94.6%	92.3%
NaiveBayes	88.6%	92.3%	90.5%
DT	82.1%	83.1%	82.8%
KNN	92.8%	100%	96.4%
SVM	88.6%	95.5%	92%

Methods	16 kHz sampling frequency		
	Sensitivity	Specificity	Accuracy
LDA	92.8%	94.5%	93.6%
NaiveBayes	90.2%	95.4%	92.8%
DT	90.1%	88.3%	89.2%
KNN	94.3%	100%	97.1%
SVM	92.8%	97.2%	95%

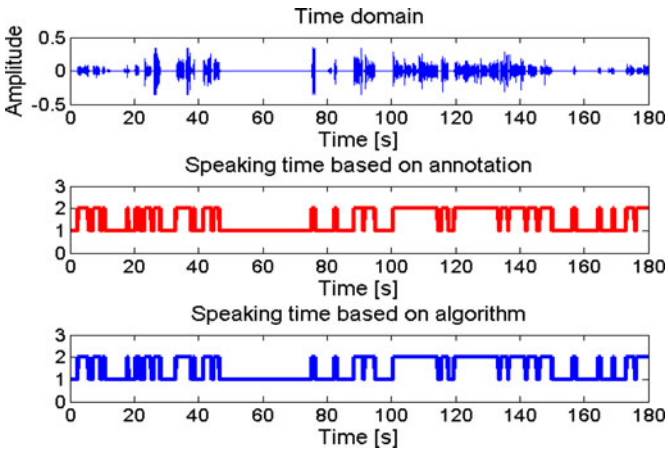

Fig. 4. Segmentation results (clean speech).

D. Stage 2 Evaluation

In this stage, three features and two states of HMM are used to segment voice and nonvoice frames. The three features are the noninitial maximum of the normalized noisy autocorrelation, number of autocorrelation peaks, and normalized spectral entropy. All three have proven successful for robustly detecting voiced speech under varying noise conditions. We obtain segmentation results on two types of speech data: 1) clean speech corrupted by different artificial noise; and 2) testing on completely natural and unpredictable speech.

1) Clean Speech Corrupted by Different Artificial Noise: Figs. 4 and 5 show the test examples of segmenting clean and noisy speech using the proposed features.

In Fig. 4, the top panel denotes the clean female speech; the middle and bottom panel denote the annotation and algorithm segmentation results. Upon comparing the algorithm segmentation with ground truth annotation results, voice segmentation has been well obtained, when using clean speech as the audio sample.

Fig. 5 demonstrates an example of segmentation results when clean speech is corrupted by strong noise, e.g., signal-to-noise ratio (SNR = 0 dB white noise). Fig. 5 shows the segmentation results given when working with a noisy speech signal. Despite failing to detect a few voiced segments, the overall segmentation

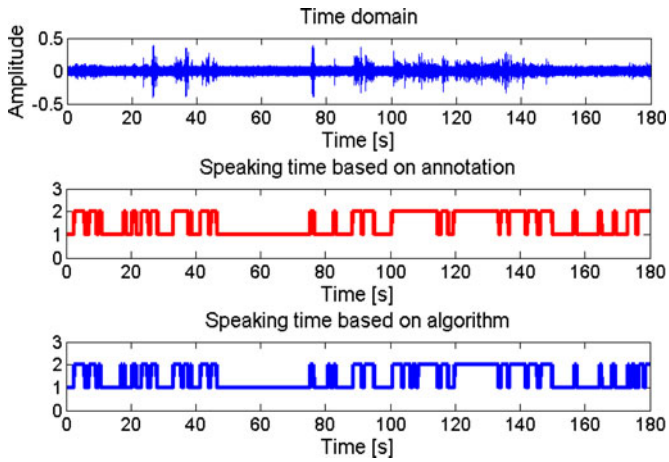

Fig. 5. Segmentation results (noisy speech).

results based on the performance of the algorithm are very accurate when compare with the ground truth. In total, four types of corrupting noise were tested: 1) white noise, 2) color noise, 3) street noise, and 4) restaurant noise. For each type of noise, the level of SNRs is 0, 5, and 10 dB. In general, the results show that the method employed with clean speech samples still can segment voice and nonvoice segments when they are corrupted with noise. Tables III and IV show the segmentation results based on sensitivity, specificity, and accuracy.

Tables III and IV show the segmentation results for 8 and 16 kHz sampling frequencies of recorded speech. For all types of noise, the sensitivity, specificity, and accuracy increase coherently when SNR increases. Overall, the results based on 16 kHz sampling frequency are always better than those based on 8 kHz sampling frequency with an average improvement of 2.5% accuracy. In Table IV, the segmentation performance of speech contaminated with color noise is found to be better than those contaminated with other noise, with an average of 95.6% accuracy. A high segmentation performance has been observed, even when the speech is corrupted with high power noise at SNR = 0 dB. The segmentation performance of speech contaminated with other types of noise still maintain significantly high accuracy, with an average accuracy of 93.2% for white noise, 94.1% for street noise and 93.2% for restaurant noise. The results of Tables III and IV are average performance. In our experiments, the noise is randomly selected from the AURORA database [41] and the segmentation performance for each type of noise is obtained in Tables III and IV by averaging over 20 realizations.

2) Natural and Unpredictable Speech: In this experiment, all speech samples are collected either from noisy or quiet environments and the WAM device was positioned on the wearer's wrist (WAM speaker). We test 35 speech chunks from various conversational situations (e.g., 3 people sit in the home having a discussion, with variable television volume in the background; 4 people have a conversation in a car; 2 people have a conversation in a quiet room; 3 people have a conversation in a noisy restaurant). Fig. 6 shows an example of the speech segmentation results from the conversation occurring with background TV sound. We tested speech chunks for both 8 and 16 kHz

TABLE III
EVALUATION OF SEGMENTATION RESULTS UNDER VARYING NOISE CONDITIONS (8 KHZ SAMPLING FREQUENCY)

Noise type	Noise level (signal to noise ratio)								
	0 dB			5 dB			10 dB		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
White noise	82.3%	98.5%	90.4%	84.6%	98.1%	91.2%	86.2%	100%	93.3%
Color noise	85.4%	97.2%	91.6%	90.7%	98.2%	94.3%	92.4%	98.1%	95.5%
Street noise	81.2%	95.4%	88.2%	87.2%	97.8%	92.6%	91.1%	98.3%	94.3%
Restaurant noise	71.7%	96.1%	84.3%	83.5%	97.3%	90.1%	90.6%	98.7%	94.2%

TABLE IV
EVALUATION OF SEGMENTATION RESULTS UNDER VARYING NOISE CONDITIONS (16 KHZ SAMPLING FREQUENCY)

Noise type	Noise level (signal to noise ratio)								
	0 dB			5 dB			10 dB		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
White noise	85.4%	100%	93.3%	87.2%	100%	94.5%	89.6%	100%	94.9%
Color noise	90.5%	98.1%	94.4%	93.3%	98.2%	96.2%	93.3%	98.2%	95.7%
Street noise	85.2%	99.3%	92.1%	93.6%	98.6%	95.1%	96.4%	98.1%	97.1%
Restaurant noise	73.7%	100%	87.6%	93.1%	98.4%	95.7%	96.3%	98.6%	97.1%

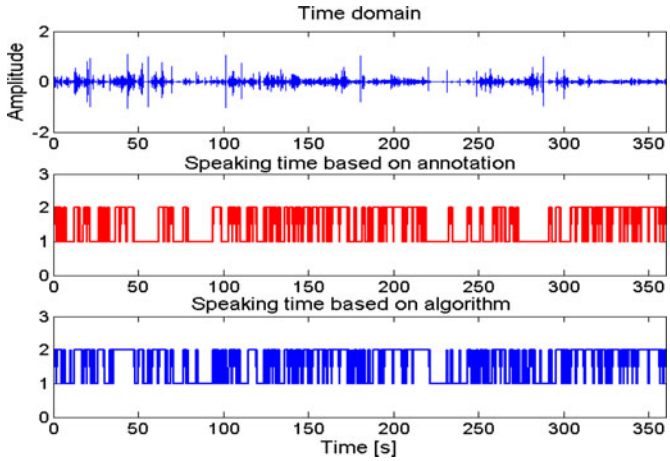

Fig. 6. Segmentation results.

TABLE V
EVALUATION OF SEGMENTATION RESULTS FOR THE WAM SPEAKER

WAM test speech (8kHz sampling frequency)					
Noiseless			Noise		
Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
66.6%	77.7%	71.5%	58.9%	73.5%	66.8%
WAM test speech (16 kHz sampling frequency)					
Noiseless			Noise		
Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
71.6%	92.5%	84.7%	68.5%	90.7%	81.6%

sampling rates. The averaged results are shown in Table V for the two cases of unpredictable speech in noiseless environment and ambient noise environment.

Table V shows the overall segmentation results given when working with a natural speech signal. Speech signals recorded from noiseless environment have been recovered successfully for both sampling frequencies with an average of 85.4% and 87.3%. The segmentation performance deteriorates when speech signals are recorded from unpredictable noisy environment because more interference exists when segmenting every voiced frame and there is higher probability of incurring an error. However, the segmentation results are still acceptable with an average accuracy of 77.4% for 8 kHz and 82.2% for 16 kHz sampling frequency. Comparing the results in the table, lower sampling frequency results in poorer performance than a higher sampled one. One reason could be that the seemingly lower sampling frequency, i.e., 8 kHz might satisfy the speech signal. However, some kinds of noise such as street noise or background music with high-frequency properties will lead to aliasing when the sampling frequency is low, which will impact the segmentation results thus leading to less efficiency in the segmentation of voiced frames.

E. Stage 3 Evaluation

1) Feature Selection Test: As described in Section III, robust features highly impact the accuracy of speaker segmentation results, especially when clean speech is corrupted by unpredictable noise. In this experiment, the speaker recognition system was first employed to test the efficiency of the different features when the speech sample is corrupted by different types of noise. The training and testing speech samples are obtained from the ELSDSR database.

Speaker recognition is the task of comparing an unknown speaker with a set of known speakers in a database to find the

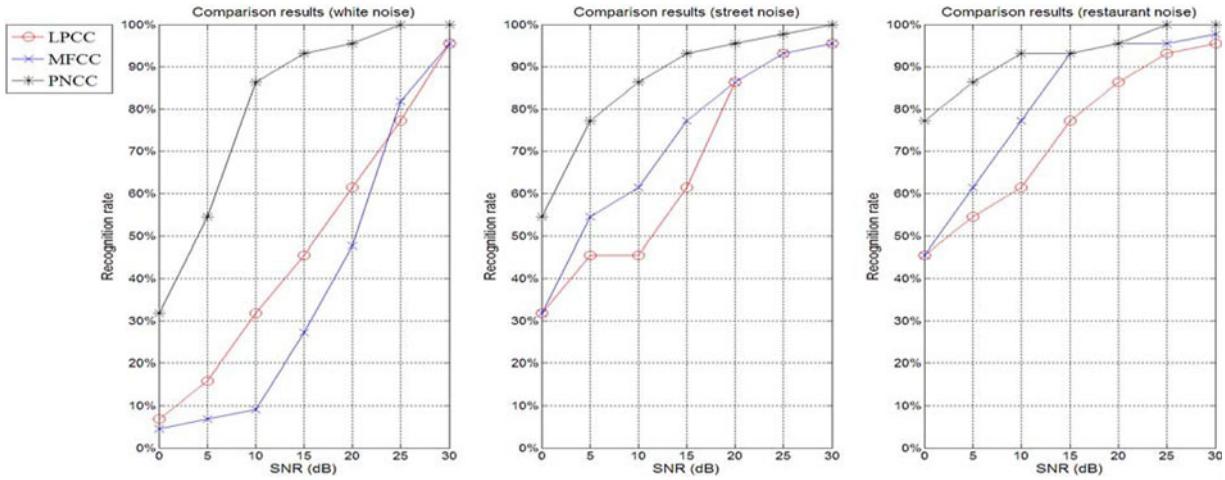

Fig. 7. Speaker recognition accuracy obtained using test speech corrupted with different type of noise.

best matching speaker. In this experiment, we consider the use of stochastic models, where pattern matching is probabilistic and results in a measure of the likelihood, or conditional probability, of the observation occurring within a model. Here, a certain type of distribution is fitted to the training data by searching the parameters of the distribution that maximize some criterion. Stochastic models provide more flexibility and better results. They include GMM, HMM, and ANN, as well as linear classifiers [36]. In our experiment, we chose to use GMM [15] with 32 components and 12 coefficients for every test results based on different types of additive noise under various SNR levels. For all types of features and noise, the recognition rate increased coherently when SNR increased.

Overall, from Fig. 7, PNCC leads to the highest recognition rate for all types of noise. For strong noisy environments (i.e., SNR between 0 and 15 dB), PNCC provides a superior average of 50.3% accuracy for white noise, 20.2% accuracy for street noise, and 16% accuracy for restaurant noise when compared to MFCC features, and 40.4% accuracy for white noise, 27.1% accuracy for street noise, and 32.3% accuracy for restaurant noise when compared with LPCC features. For low-level noise environments (i.e., SNR between 20 and 30 dB), PNCC provided an average of 28.2% accuracy for white noise, 7.5% accuracy for street noise and 0.4% accuracy for restaurant noise, when compared with MFCC features, and 30.1% accuracy for white noise, 7.6% accuracy for street noise and 0.8% accuracy for restaurant noise, when compared with LPCC features.

2) Segmenting the Wearable Acoustic Monitor Speaker from Natural and Unpredictable Conversation: In this test, we target the segmentation of the WAM speaker as well as predicting the number of unknown speakers within the unpredictable conversation. We trained the GMM WAM speaker PNCC models individually and used a universal background GMM to represent all unknown speakers. This can be implemented by training the GMMs on different kinds of speakers. In our experiment, we train 40 different people, ranging in age from 20 to 70 years, in order to generate our background GMM PNCC models. For parameters settings, 180 s (3 min) is deemed a good minimum amount of training data for the WAM speaker

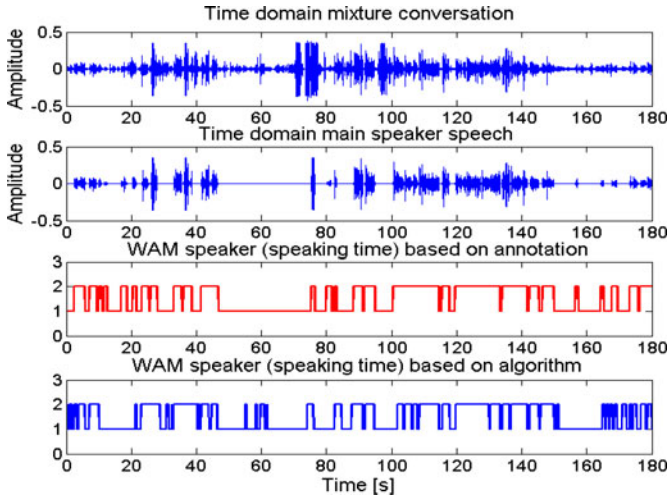

Fig. 8. Segmentation results (WAM speaker).

and 32 GMM components is chosen to balance accuracy and efficiency [15]. In addition, the fundamental frequency (pitch) of each WAM speaker is found useful in discriminating, at the very least, between male and female speakers, so, we have included it as another feature in the model training.

Fig. 8 shows an example of the segmentation results. The top panel shows the natural conversation, where the WAM speaker talks with her female friend in the kitchen. Later, an additional two females and two males join the conversation. We have annotated the ground truth of the WAM speaker, which can be viewed in the middle two panels and then compared to the algorithm's ability to segment the WAM speaker (bottom panel), which presented a successful result of 82.4% accuracy. In this case, TP denotes that the WAM speaker is correctly identified as the WAM speaker, FP denotes nonspeech segments and other speakers who are incorrectly identified as the WAM speaker, TN denotes nonspeech segments and other speakers who are correctly identified as nonspeech and other speakers, and finally, FN denotes that the WAM speaker is incorrectly identified as nonspeech or other speakers. Fig. 8 shows the results obtained

TABLE VI
SOCIAL-AUDIO SIGNALS OF EACH SUBJECT

Subject	Total conversation time (minutes)	Percentage of speaking time		Sound level meter	Estimated speaking energy (std/mean)	Estimated first format frequency (std/mean)
		Actual	Estimate			
# 1	22.30	55.1%	52.2%	76.3dB	1.86	0.36
# 2	22.16	22.8%	21.4%	62.9dB	2.96	0.52
# 3	15.07	32.2%	34.1%	70.5dB	1.90	0.49
# 4	23.21	18.0%	20.0%	83.2dB	1.74	0.39
# 5	15.13	51.4%	48.5%	74.2dB	1.36	0.34
# 6	10.08	16.4%	14.7%	72.5dB	1.36	0.34
# 7	10.15	19.5%	16.8%	72.1dB	3.92	0.36
# 8	15.29	32.6%	34.1%	76.7dB	1.57	0.47
# 9	15.11	30.8%	34.6%	72.9dB	2.00	0.49
# 10	15.04	48.1%	44.9%	73.7dB	1.36	0.35

TABLE VII
GAP STATISTIC PREDICTION

Unpredictable test speech (8kHz sampling frequency)	
Noiseless	Noise
Correct %	Correct %
75.2%	61.1%
Unpredictable test speech (16 kHz sampling frequency)	
Noiseless	Noise
Correct %	Correct %
79.3%	67.1%

for sensitivity 79.5%, specificity 84.2% and accuracy 82.3%. In addition, Table V shows the overall segmentation results for all test speech chunks based on sensitivity, specificity, and accuracy.

Table V shows the overall segmentation results given when working with naturalistic test speech chunks. The WAM speaker segments have been extracted successfully, when conversations happen in a noiseless environment, for both sampling frequencies, with an average of 71.5% and 84.4%. The segmentation performance deteriorated when conversation was recorded from within unpredictable, noisy environments. This is due to fact that more interference exists in the change point detection algorithm and model match calculation and hence results in a higher probability of incurring an error. However, the segmentation results are still acceptable, with an average accuracy of 66.5% for 8 kHz and 81.3% for 16 kHz sampling frequency. Once the WAM speaker has been segmented out of conversation, we expected to be able to predict the number of other speakers within the conversation. This prediction was achieved through the use of gap statistic techniques and the results are tabulated in Table VII.

Table VII summarizes the overall prediction results when working with naturalistic speech chunks. It can be seen that the obtained results have been very successful when conversation occurs in the noiseless environment for both sampling frequencies, with an average of 75.2% and 79.3%, respectively. The prediction results are still acceptable with an average accuracy of 61.1% for 8 kHz and 67.1% for 16 kHz sampling frequency.

Once all of the above stages have been implemented, the social-audio signals can then be calculated. The sound level meter is used to analyze the nonspeech sounds occurring in the background, to measure the number of places that conversations are happening. The social signals *activity* and *emphasis* can be predicted by using the WAM speaker segmentation results.

Overall, we calculated the social signals for all WAM subjects. Ten out of thirty results have been tabulated in Table VI. These features can be applied to enable the performance of automated social analysis of conversational data to infer the relationship between speakers. Fig. 9 shows an example of calculating the social signal of the WAM speaker based on the proposed work. The top panel shows the conversation about holiday discussion between the WAM speaker and other speakers where the talk happens inside a house. The second panel shows the WAM speaker segmentation results (89.3% sensitivity, 88.5% specificity, and 89.6% accuracy compared with annotated ground truth) based on the proposed algorithm. The third panel estimates sound level meter for every 1 min block conversation. It indicates that the environment of this conversation happens in a normal place due to the classification range summarized in Table I of E_{dB}. The fourth panel estimates the percentage of WAM speaking time. This indicates how activity of the WAM speaker is within the conversation. This shows that with percentage of WAM speaking time, we can readily infer the activity level of the speaker. Fig. 9 shows that the WAM speaker is more active in the first 3 min conversation, where his percentage of speaking time is larger than 50%, and then slowly decreases in the latter 2 min to approximately 40%. The activity suddenly decreases to the bottom (10%) in the conversation between $t = 6$ and 7 min. The WAM speaker starts to be active again in the conversation at $t = 8$ mins and decreases to 10% of speaking time at $t = 13$ mins and finally returns around 30% in the last 2 min. In line with above, the bottom panel estimates the emphasis of the WAM speaker during conversation based

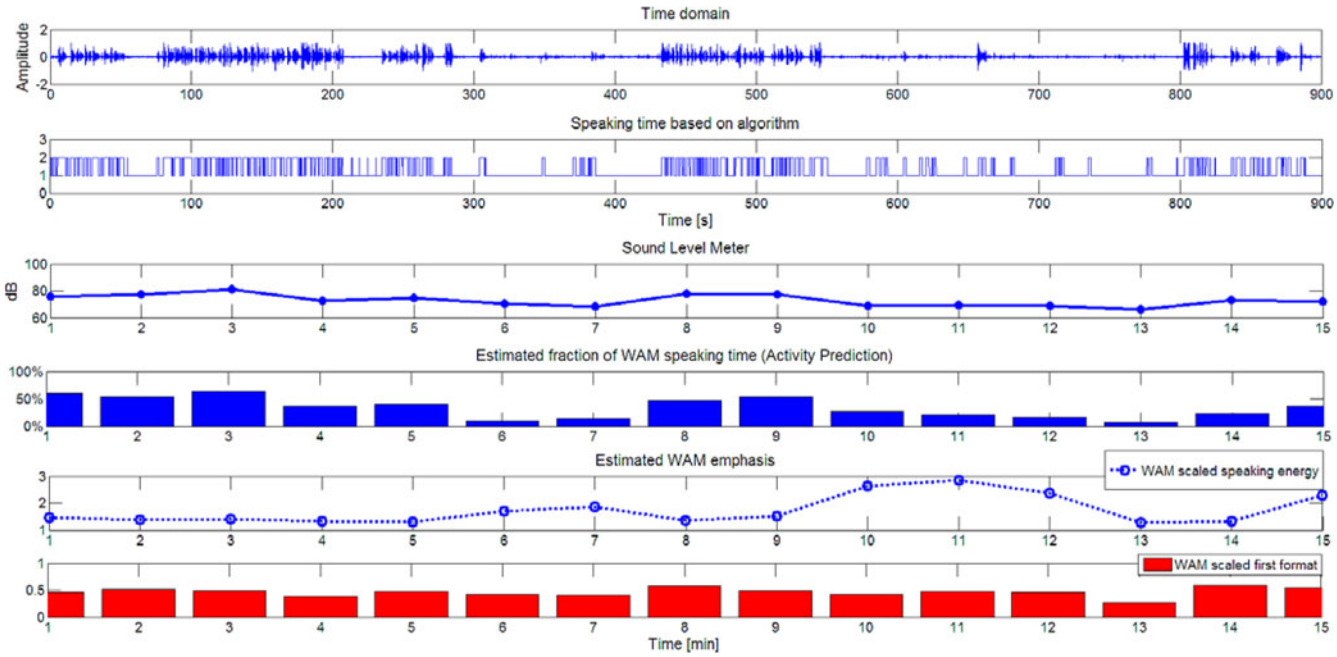

Fig. 9. Social signal prediction.

on the estimated scaled speaking energy and first format. The first format of the WAM speaker does not change prominently across the whole conversation except at $t = 13$ min and 14 min. In addition, the emphasis of the WAM speaker in the first 5 min retains a relatively stable level based on the scaled speaking energy and starts to fluctuate between conversation from $t = 6$ min and 15 min where the emphasis approaches the highest level at the $t = 11$ min.

V. Conclusion

In this paper, a novel platform for the analysis of longitudinal and unpredictable social-audio signals has been proposed. The proposed method enjoys at least three significant advantages. First, an efficient architecture has been developed to enable continuous audio sensing and scalable four-stage methods to gather social-audio signal. Second, an integrated system of speech and sound detection and classification to reliably analyze longitudinal audio signals has been introduced. This is used to capture the changeable/unstable characteristics of the longitudinal and unpredictable audio signals. Finally, the analysis of the audio data captured by the wearable device has yielded significantly high performance for social-audio signal learning using the proposed SASE architecture.

References

[1] A. Pentland, "Socially aware computation and communication," *Computer*, vol. 38, pp. 33–40, 2005.
[2] A. Pentland, "Social signal processing," *IEEE Signal Process. Mag.*, vol. 24, no. 4, pp. 108–111, Jul. 2007.
[3] R. Jun and K. Nagao, "The world through the computer: Computer augmented interaction with real world environments," in *Proc. User Interface Softw. and Technol.*, Nov. 14–17, 1995, pp. 29–38.
[4] B. Benjamin, "Audio augmented reality: A prototype automated tour guide," in *Proc. Conf. Human Factors Comput.*, May 1996, vol. 95, pp. 210–211.
[5] R. Alexander, S. Reed, and E. Thayer, "Speech wear: A mobile speech system," in *Proc. Int. Conf. Spoken Language Process.*, 1996, vol. 96.
[6] D. Wyatt, T. Choudhury, J. Bilmes, and J. Kitts, "Towards the automated social analysis of situated speech data," in *Proc Int. Conf. Ubiquitous Comput.*, 2008, pp. 168–171.
[7] "NIST rich transcription evaluations," 2006. [Online]. Available: http://www.nist.gov/speech/tests/rt/rt2006/spring/
[8] T. L. Nwe, H. Sun, B. Ma, and H. Li, "Speaker clustering and cluster purification methods for RT07 and RT09 evaluation meeting data," *IEEE Trans. Audio, Speech Language Process.*, vol. 20, no. 2, pp. 461–473, Feb. 2012.
[9] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Commun.*, vol. 40, pp. 33–60, 2003.
[10] A. Batliner, K. Fisher, R. Huber, J. Spilker, and E. Nöth, "Desperately seeking emotions or: actors, wizards and human beings," in *Proc Int. Symp. Circuits Syst., Tutorial Res. Workshop Speech Emotion*, 2000, pp. 195–200.
[11] P. Greasley, J. Setter, M. Waterman, C. Sherrard, P. Roach, S. Arnfield, and D. Horton, "Representation of prosodic and emotional features in a spoken language database," in *Proc. 13th Int. Congr. of Phonetic Sci.*, 1995, pp. 242–245.
[12] E. Douglas-Cowie, R. Cowie, and M. Schroeder, "A new emotion database: Considerations, sources and scope," in *Proc. Int. Symp. Circuits Syst., Tutorial Res. Workshop Speech Emotion*, 2000, pp. 39–44.
[13] M. Zelenak, C. Segura, J. Luque, and J. Hernando, "Simultaneous speech detection with spatial features for speaker diarization," *IEEE Trans. Audio, Speech, Signal Process.*, vol. 20, no. 2, pp. 436–446, Feb. 2012.
[14] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez, "Estimating dominance in multi-party meetings using speaker diarization," *IEEE Trans. Audio, Speech Language Process.*, vol. 19, no. 2, pp. 847–860, Feb. 2012.
[15] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2010, pp. 4574–4577.
[16] J. P. Campbell, Jr., "Speaker recognition: A tutorial," in *Proc. IEEE*, Sep. 1997, vol. 85, no. 9, pp. 1437–1462.
[17] R. J. Tibshirani, G. Walther, and T. J. Hastie, "Estimating the number of clusters in a dataset via the gap statistic," *J. Royal Stat. Soc.: Ser. B. (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.
[18] H. Lu, A. J. Brush, B. Priyantha, A. Karlson, and J. Liu, "Energy efficient unobtrusive speaker identification on mobile phones," in *Proc. 9th Int. Conf. Pervasive Comput.*, San Francisco, CA, USA, 2011, pp. 188–205.

[19] J. Saunders, "Real time discrimination of broadcast speech/music," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1996, pp. 993–996.
[20] I. M. Cowan, D. Dean, M. McLaren, R. Vogt, and S. Sridharan, "The delta-phase spectrum with application to voice activity detection and speaker recognition," *IEEE Trans Audio, Speech Language Process.*, vol. 19, no. 7, pp. 2026–2038, Sep. 2011.
[21] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 3, pp. 600–613, Sep. 2011.
[22] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.
[23] A. D. Wallis, "From mahogany to computers proceedings euronoise," London, U.K., Plenary Paper., Sep. 1992.
[24] J. Saunders, "Real time discrimination of broadcast speech/music," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1996, pp. 993–996.
[25] D. O. Olguín, B. Waber, T. Kim, A. Mohan, K. Ara, and A. Pentland, "Sensible organizations: Technology and methodology for automatically measuring organizational behavior," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 1, pp. 43–55, Feb. 2009.
[26] P. Davalos and H. A. Kingravi, "Unsupervised speaker segmentation and clustering," Tech. Rep., 2007.
[27] M. Huijbregts and D. A. van Leeuwen, "Large-scale speaker diarization for long recordings and small collections," *IEEE Trans. Audio, Speech Language Process.*, vol. 20, no. 2, pp. 403–413, Feb. 2012.
[28] L. R. Rabiner, N. J. Murray Hill, M. Cheng, A. E. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. SSP- 24, no. 5, pp. 399–418, Oct. 1976.
[29] J. R. Curhan and A. Pentland, "Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes," *J. Appl. Psychol.*, vol. 92, no. 3, pp. 802–811, 2007.
[30] R. Dunbar, *Grooming, Gossip, And The Evolution Of Language*. Cambridge, MA, USA: Harvard Univ. Press, 1998.
[31] X. L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 4, pp. 697–710, Apr. 2013.
[32] B. Barry and R. A. Friedman, "Bargainer characteristics in distributive and integrative negotiation," *J. Personality Social Psychol.*, vol. 74, pp. 345–359, 1998.
[33] R. J. Tibshirani, G. Walther, and T. J. Hastie, "Estimating the number of clusters in a dataset via the gap statistic," Department of Statistics, Stanford University, Stanford, CA, USA, Tech. Rep., 2000.
[34] S. Werner and E. Keller, *Prosodic Aspects of Speech. In E. Keller (Ed.), Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges*. Chichester, U.K.: Wiley, 1994, pp. 23–40.
[35] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2009, pp. 28–31.
[36] P. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. H. Allerhand, "Complex sounds and auditory images," in *Auditory and Perception. Oxford*, Y. Cazals, L. Demany, and K. Horner, Eds. Oxford, U.K.: Pergamon Press, 1992, pp. 429–446.
[37] K. W. Jorgensen, L. L. Molgaard, and L. K. Hansen, "Unsupervised speaker change detection for broadcast news segmentation," in *Proc. Eur. Signal Process. Conf.*, 2006.
[38] V. V. Digalakis, M. Ostendorf, and J. R. Rohlicek, "Fast algorithms for phone classification and recognition using segment-based models," *IEEE Trans Signal Process.*, vol. 40, no. 12, pp. 2885–2896, Aug. 1992.
[39] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Chichester, U.K.: Wiley-Interscience, 2000.
[40] L. Feng, "Speaker recognition," Master's thesis, Tech. Univ. Denmark, Informatics and Mathematical Modelling, 2004.
[41] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. Automat. Speech Recog.: Challenges New Millennium*, Paris, France, 2000, pp. 29–32.

Bin Gao (M'12) received the B.S. degree in communications and signal processing from China, the M.Sc. and the Ph.D. degrees (with Dist.) in communications and signal processing from Newcastle University, Tyne and Wear, U.K, from Newcastle University.

Upon graduation, he worked as a Research Associate with Newcastle University on wearable acoustic sensor technology. He is currently an Associate Professor with the School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include sensor signal processing, machine learning, structural health monitoring, and nondestructive testing and evaluation.

Wai Lok Woo (SM'12) received the B.Eng. degree (1st Class Hons.) in electrical and electronics engineering and the Ph.D. degree from Newcastle University, Tyne and Wear, U.K.

He is currently a Senior Lecturer and Director of Operations with the School of Electrical and Electronic Engineering. His major research spans the area of 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, signal/image deconvolution, and social signal processing.

Dr Woo is a Member of the Institution Engineering Technology.