Conversation Monitoring via Low-cost Speaker Diarization using Wearable Wireless Sensors

Muhannad Quwaider
Jordan University of Science and Technology/Department of Computer Engineering, Irbid, Jordan
Email: mqquwaider@just.edu.jo

Subir Biswas and ChaiYong Lim
Michigan State University/Department of Electrical and Computer Engineering, East Lansing, USA
Email: {sbiswas, limchia}@egr.msu.edu

***Abstract*—This paper presents speaker diarization mechanisms using low-cost and low-resolution wearable wireless sensors. Speaker diarization is used for identifying speaking sequence and duration for all individuals engaged in a conversation session. The key advantage of the proposed mechanisms is their ability to monitor human conversation without having to perform energy- and processing-expensive speaker identification algorithms. A prototype system was constructed for experimental acoustic diarization using low-cost and low-resolution wearable sensors. It was experimentally demonstrated that an inexpensive threshold-based diarization mechanism can be used for conversation monitoring with acceptable accuracy. But for more detection accuracy, an acoustic comparator-based diarization is applied. It was shown that comparator-based diarization mechanism is able to consistently deliver significantly better acoustic detection performance than threshold-based mechanism in a more distance and noise independent manner. Controlled experiments using human subjects were carried out for evaluating diarization accuracy and the sensitivity to factors such as sampling rate and inter-speaker distance during conversations.**

***Index Terms*—Wearable Sensors, Coversation Monitoring, Acoustic Diarization, Diagrammatic Speech Diarization, Comparator based Diarization**

I. INTRODUCTION

Wireless Body Area Networks (WBANs) [1–3] consist of a set of wearable or implanted communicating sensors. These devices, communicating through wireless links, can transmit physiological data from the body to an external data aggregator device with Internet connection, from where data can be forwarded to hospitals or clinics in real-time. Such WBANs can enable a wide variety of applications in a number of areas including sports, healthcare, tele-health, and human interaction monitoring. The WBAN technology is still in its primitive stage and is being widely researched in both industry and academia.

Human interaction monitoring using wearable sensors is an interesting WBAN application. Human interaction can be defined as a sequence of brief actions where each action is seen as response to what came before and as a stimulus to what comes afterward [4], [5]. In the absence of visual data, conversation monitoring using acoustic sensors is considered as a feasible technique for human interaction monitoring.

In this paper we develop speaker diarization [6] mechanism using low-cost and low-resolution wearable wireless sensors. Speaker diarization is then used for conversation monitoring by the way of identifying speaking sequence and duration for all individuals engaged in a conversation session. In other words, the objective is to detect who spoke when and for how long, as opposed to detect what was spoken.

Applications of such conversation monitoring include various forms of surveillance, study of behavior and speech development for children with autistic conditions and other speech related disorders. It can also be used for studying team dynamics in controlled and un-controlled settings.

The key advantage of the proposed diarization based conversation monitoring mechanism is its ability to monitor human conversation without having to perform energy- and processing-expensive speaker identification algorithms. In addition to its processing and energy overheads, speaker identification algorithms [7–10] usually require acoustic sampling at higher rates and, subsequently higher wireless bandwidth for collecting data from on-body sensors to out-of-body processing units. In the proposed approachs, on the other hand, low sampling rates (e.g. 5 Hz) lead to lower energy, bandwidth, and processing requirements, while being able to detect conversation dynamics. This suits the proposed mechanisms very well for resource-constrained wearable sensors. The primary reason for this proposed simple approaches to work is that we are mainly interested in detecting who spoke when and for how long, as opposed to specific content of the speech.

Contributions of the paper are as follows. First, the details of a prototype wearable sensor system used for the proposed speaker diarization experiments are presented. Second, the proposed low-complexity speaker diarizations are formally presented. Finally, the algorithms are experimentally validated along with a detailed study of their sensitivity to audio sampling rate and inter-subject distance during a conversation.

doi:10.4304/jetwi.4.4.315-323

II. Related Work

Instrumented detection of human interaction [11] has recently been gaining popularity in both psychology and sensor literature. Example sensing modalities for interaction detection are proximity, relative orientation, and conversation dynamics on an inter-personal basis. The objective of this paper is to address the modality of conversation dynamics [12–15] by the way of speaker diarization.

Speaker diarization in the literature [15] has been used for a wide range of applications including: 1) analyzing broadcast audio news programs in the presence of commercial and other breaks, 2) analyzing recorded meetings with multiple people participants, and 3) phone conversations between two or more people. A large number of diarization approaches in the literature [7–10] use Hidden Markov Model (HMMs) [16] in which each state correspond to an individual speaker. A Gaussian Mixture Model (GMM) [17] is generally used for modeling the conversations.

The other commonly used approach for speaker diarization is Bayesian machine learning [18]. In order to avoid premature hard decisions in the diarization process, the Bayesian mechanisms usually attempt to estimate the complete distribution of the system parameters as opposed to just the averages. Monte Carlo Markov Chains (MCMC) are used in [19] in order to provide a systematic approach to the computation of such distributions via exhaustive sampling. Such sampling, however, is generally slow and expensive when the amount of data is large, and they often require multiple passes as the chains get stuck and not converge in a practical number of iterations. In [20], a variation of Bayesian machine learning algorithm is used to learn a GMM speaker model. In [6], the above is combined successfully with eigenvoice modeling [18] for speaker diarization of telephone conversations. All the above approaches require significant amount of processing overhead that is not usually practical for the on-body diarization problem as targeted in this paper. Our approach addresses such processing complexity issues.

The goal of this paper is to propose our wearable sensor network that can be used for human interaction exposure through detecting human speech activity. Specific contributions of the paper are as follows. First, propose our prototype as one modern framework that can be used for human interaction and speaker diarization. Second, develop an analytical framework for determining the current speaker, for a given collecting data, representing the current speaker. Third, validate the proposed prototype and the analytical framework by comparing the experimental speaker diarization data and the acoustic recording data. Finally, Study the impact of the sample rate and signal compression data on the speaker diarization accuracy in the present of different parameters, like, human subject's pair-wise distance and orientation and environment noise.

III. Wearable Sensors

This section presents the experimental setting of wearable sensors that was used for the proposed light-weight speaker diarization mechanism.

A. Experiment Settings

The wearable node was constructed using a 900MHz Mica2Dot MOTE [21] (running TinyOS operating system), with Chipcon's SmartRF CC1000 radio chip (chipcon.com), and the sensor card MTS510 from Crossbow Inc. (xbow.com). The Mica2Dot node runs from a 570mAH button cell with a total sensor weight of approximately 10 grams. The default CSMA MAC protocol was used with a data rate of 19.2kbps. In our experiments, each sensor was worn as a badge fastened at a subject's chests so that the sensor does not move during an experiment.

A wireless link is established from the wearable sensor to an external processing server to transport raw data or the results from on-body diarization depending on the chosen processing mode. A Mica2Dot radio node with custom-built serial interface, running RS232 protocol, was used as a base station for collecting data from on-body sensors and for sending to a Windows PC processing server through its serial port.

B. Acoustic Sensing

Within the MTS510 sensor card, acoustic data from a microphone is fed into an ADC through an amplifier. The ADC values are read by the Mica2Dot's microprocessor and formatted as packets before sending it out to processing server PC through 900MHz radio links. The speaker diarization was then executed within the processing server PC. The sampling rate of the acoustic data was varied from 5Hz to 40Hz during our experiments.

As a control, for each scripted conversation session the entire conversation was also recorded using a separate high fidelity microphone so that the actual conversation dynamics can be post-coded by listening to that recording and can be compared with the output of the diarization algorithms for their accuracy. The high fidelity microphone is connected to the same processing server PC so that it can time-stamp both the externally recorded audio data and the samples received from on-body sensors using the same clock, thus synchronizing the two data streams.

C. Polling Based Channel Access for Collision Control

Since each person is required to wear a separate sensor badge, multiple sensors can be simultaneously active and share the radio for sending data to the processor server. In order to avoid the CSMA MAC collisions due to such multiple accesses, we implemented a higher layer polling based TDMA access strategy that is managed by the base station (BS) connected to the processing server. In addition to avoiding collisions, TDMA operation also enables the system to run in a more energy-efficient manner by turning the wireless interface of a sensor node off during the TDMA slots of other sensor nodes.

The BS polls the on-body sensors in a round-robin fashion. A node forwards its packet only when it is polled by the BS for giving access to the channel. With *n-node* network (see Figure 1), a polling time frame of $100 \times n$ *msec* is used which is divided into $100 \times n$ msec time slots, one for each on-body node. If a node misses a polling packet from the BS, it simply misses one transmission opportunity. Each slot is further divided into two 50 msec sub-slots. The first sub-slot is used for polling packets from the BS to an on-body sensor node, while the second sub-slot is used for the data packet from the on-body node to the BS.

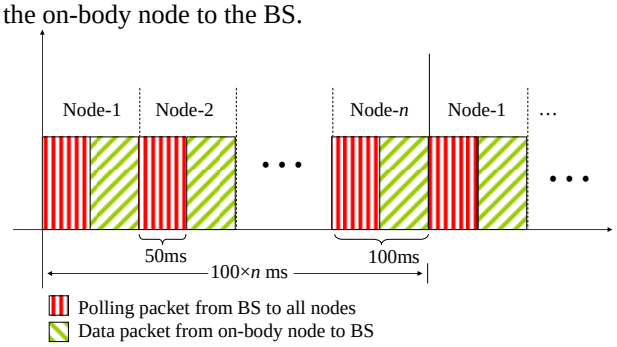

Figure 1. Collision-free MAC access via polling

IV. SPEAKER DIARIZATION

The goal of this section is to develop a speaker diarization mechanism for the down sampled signal, by which a current speaker can be decided.

A. Speech Signal Variation

The captured acoustic signal from the on-body sensor of the i^{th} subject during an *n*-subject conversation can be represented as:

$$X_i(t) = \sum_{j=1}^{n} A_{ij} S_j(t) + N_i(t), \ for\ all\ i \in n \tag{1}$$

where S_j denotes the acoustic signal from the j^{th} speaker and N denotes the ambient noise. A_{ij} denotes the coefficient of the acoustic signal between the j^{th} speaker and node i in the network, which depends on the distance between the speaker and the node. Therefore A_{ij} is maximize when $i=j$. Means, when the speech signal is captured by the speaker node itself, because that is the shortest distance between the speaker and the corresponding node.

For a system with *n* speakers, each node can potentially receive acoustic signal from all the speakers in the system. Eqn. 1 can be extended to include *n* such signals captured at any node in the network as follows:

$$\begin{bmatrix} X_1(t) \\ \vdots \\ X_i(t) \\ \vdots \\ X_n(t) \end{bmatrix} = \begin{bmatrix} A_{11} & \dots & A_{1i} & \dots & A_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ A_{i1} & \dots & A_{ii} & \dots & A_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ A_{n1} & \dots & A_{ni} & \dots & A_{nn} \end{bmatrix} \cdot \begin{bmatrix} S_1(t) \\ \vdots \\ S_i(t) \\ \vdots \\ S_n(t) \end{bmatrix} + \begin{bmatrix} N_1(t) \\ \vdots \\ N_i(t) \\ \vdots \\ N_n(t) \end{bmatrix}, \ for\ all\ i \in n \tag{2}$$

B. Diagrammatic Speaker Diarization (DSD)

In this section we propose diagrammatic based speaker *diarization* mechanism. The key idea behind the proposed power based diagrammatic speaker diarization (DSD) is that when a sensor captures signal from its wearer, the variability in the captured samples (i.e. the power) is higher than when signal is captured from other sensors.

The power of the acoustic activity at node i during a discrete time window of T samples can be computed as

$$P_i = \frac{1}{T} \sum_{t=1}^{T} |X_i(t)|^2 \tag{3}$$

where $X_i(t)$ is the discrete acoustic signal captured at node i as defined in Eqn. 1. Statistically, Eqn. 3 can be written in terms of the variance of the signals within a window of T samples as:

$$\sigma_i^2 = P_i - \mu_i^2 \tag{4}$$

where σ_i^2 is the variance and μ_i is the mean of the acoustic signals captured at node i during a window of T samples. As define in Eqn. 4, the acoustic power is directly proportion to the diversity of the data samples. In other words, the power of a signal can be estimated by observing the variance of samples within pre-specified time windows.

In Eqn. 3, the number of samples T depends on the time window W in seconds, during which the acoustic signals are captured at a sampling rate of f_s Hertz.

$$T = W \cdot f_s \tag{5}$$

In all the presented experiments, the window size W is kept fixed to be 300ms, and the sampling frequency f_s is changed from 5 to 40Hz.

After the acoustic power is computed, a set of acoustic signals can be classified to be a speech or not by computing the histogram of the collected data power. The power is classified to be, either high, if it corresponds to a speech time frame, or low, if it corresponds to a silent or a low speech time frame, as defined in Eqn. 3. With DSD, we propose a diagrammatic process to classify the power and then to decide the current speaker. DSD mechanism is summarized by computing the histogram of P_i at node i, and then classifying the speech and the silent time frames.

Ideally, the histogram of P_i should classify the acoustic power into two sets or in a form of a bimodal distribution. One set is in the left with low power, which represents the silent power, and the other set is to the right with high power, which represents the speech power. The goal of DSD is to assign a threshold value by which the power can be classified into silent or speech, as shown in

Figure 2.

In this figure, the discrete acoustic signals are fist computed as described in Eqn. 1. Then, the speech power

of the collected signals is computed at each node according to Eqn. 3. Finally, the computed powers of the all signals are compared with DSD threshold value at each user to decide the current speaker.

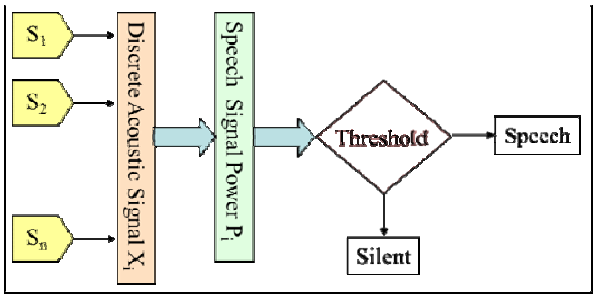

Figure 2. Capturing speech using diagrammatic speech diarization

Ideally, DSD at any node should be in the middle of the acoustic power distribution, and in this case it will be computed as

$$DSD_{mid} = \frac{1}{T}\sum_{j=1}^{T}\sigma_j^2 \quad (6)$$

where T is the total number of the sampled data within the window W and σ_i^2 is the variance of the acoustic signals captured at the node, which corresponds to the speech activity power. On the other hand, due to the diversity of the expected acoustic power, the bimodal distribution of P_i is expected to shift to the left or to the right, depending on the speech conversation, subjects, noise orientation and distance. Therefore, DSD threshold values are chosen to be proportional to DSD_{mid} and the relative frequency of the distribution. More details about DSD threshold values are presented in *Section V*.

C. Acoustic Comparator based Diarization (ACD)

With DSD, the down sampled signal $X_i(t)$ will still have some other speakers' acoustics, as shown also in Eqn. 2. Therefore, DSD accuracy will be affected during the diarization, and it will not perform well, as we will sdiscuss in *Section V*. In this section we propose *Acoustic Comparator based Diarization* or *ACD* for speech detection.

For a given i^{th} node, let us define $X_i(1), X_i(2), \ldots, X_i(T)$ be T acoustic samples that are collected during T discrete time at node i. In order to measure the variation of the acoustic samples, let us define Acoustic Variation Coefficient (*AVC*) as:

$$AVC_i = \frac{AVar_i}{\mu_i} \quad (7)$$

where μ_i is the mean of T acoustic samples, and it can be defined by:

$$\mu_i = \frac{1}{T}\sum_{t=1}^{T} X_i(t) \quad (8)$$

$X_i(t)$ is the acoustic signal computed in node i and at time t, as described in Eqn. 1. $AVar_i$ in Eqn. 7 is the variance of the T acoustic samples, and can be computed as:

$$AVar_i = \frac{1}{T}\sum_{t=1}^{T}(X_i(t) - \mu_i)^2 \quad (9)$$

After the *AVC* is computed in every node in the network, *ACD* is applied among the all nodes at the same time. Consequently, the sampling and processing times are synchronized among all the nodes as reported in *Section III*. For given N nodes in the network and $AVC_1(t), AVC_2(t), \cdots, AVC_N(t)$ acoustic variation coefficients reported from all nodes, the current SPeaker (*SP*) at time t corresponds to a node with the following argument among the all nodes in the network.

$$SP(t) = \arg\max_{1 \le i \le N}[AVC_i(t)] \quad (10)$$

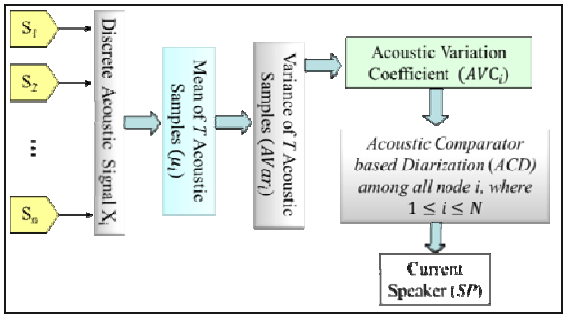

Figure 3. Capturing Speech using acoustic comparator based diarization

Figure 3 shows ACD steps for speaker diarization. As shown in the figure, AVC is computed at each node and the maximum AVC value will correspond to the current speaker, as defined in Eqn. 10. With two subjects conversation, the current speaker can be detected by simply comparing AVC_1 with AVC_2, or $AVC_1 - AVC_2$. In this case the output will show either, positive (subject-1 speak), negative (subject-2 speak) or silent. More details about ACD results are presented in *Section V*.

V. EXPERIMENTAL PERFORMANCE

In this section we study the performance of *ACD* speech diarization and its performance in comparison with *DSD* based approaches that are described in *Section IV*.

Controlled experiments are designed in which two human subjects, each with an acoustic data collection sensor node, are given pre-determined sequences of conversation. The two subjects are conversing in unknown noisy environment and within a pre-defend distance as shown in. Speaker identified using acoustic

detection mechanisms, are then temporally correlated with the actual conversation given to the subjects for evaluating the identification accuracy.

A. Acoustic Power Dissemination

In order to study the acoustic power in a conversation, the acoustic power distribution is computed using Eqn. 3 with a window size of 300ms, as discussed in *Section IV*. Figure 4 shows four experiments results that were conducted in same environment with sampling frequency f_s of 5, 10, 20 and 40Hz. With each frequency, the figure shows each subject's acoustic power. The following conclusions can be made from Figure 4.

First, the figure shows how the acoustic power constructs a very close to a bimodal distribution for each subject's speech. The form of the distribution is caused due to the silent and the speech scenario of the two subjects speeches during the conversation. Second, even with different sampling frequencies, the acoustic powers construct very similar distributions. Finally, the speech and the silent can be detected by choosing a threshold power value between the two peaks of the acoustic powers. This threshold value can be easily determined due to the wide range between the silent and the speech power.

B. Acoustic Diarization using Diagrammatic

In order to determine the performance of DSD, different *DSD* threshold values are applied on the conducted experiments results that are shown in Figure 4. The threshold values are chosen from the acoustic power distributions that are described in *Section IV.B*. The exact threshold values that are used for speaker diarization are presented in TABLE I. The performance of DSD is then computed by comparing the DSD output with the recorded acoustic data. Table I includes also a benchmark scenario, at which, DSD represents the best case of maximized percentage match with the recorded acoustic data.

TABLE I:

DSD THRESHOLD SELECTED VALUES

DSD Threshold Value	DSD Chosen Description
DSD_1	A value below the 10% of the relative frequency
DSD_2	A value above the 10% of the relative frequency
DSD_3	10% of DSD_{mid}
DSD_4	50% of DSD_{mid}
Bmark	Best-case DSD assignment scenario

Figure 5 shows the performance results correspond to DSD threshold values presented in TABLE I. As shown in the figure, for a given DSD threshold value, it was able to identify the current speaker. Similarly, different threshold values perform well and almost close to each other. That is because of the noise, distance and sampling rate as discussed in *Section IV.B*. On the other hand, even with different sampling rate, the best case benchmark performance is very close to the other DSD thresholds performance.

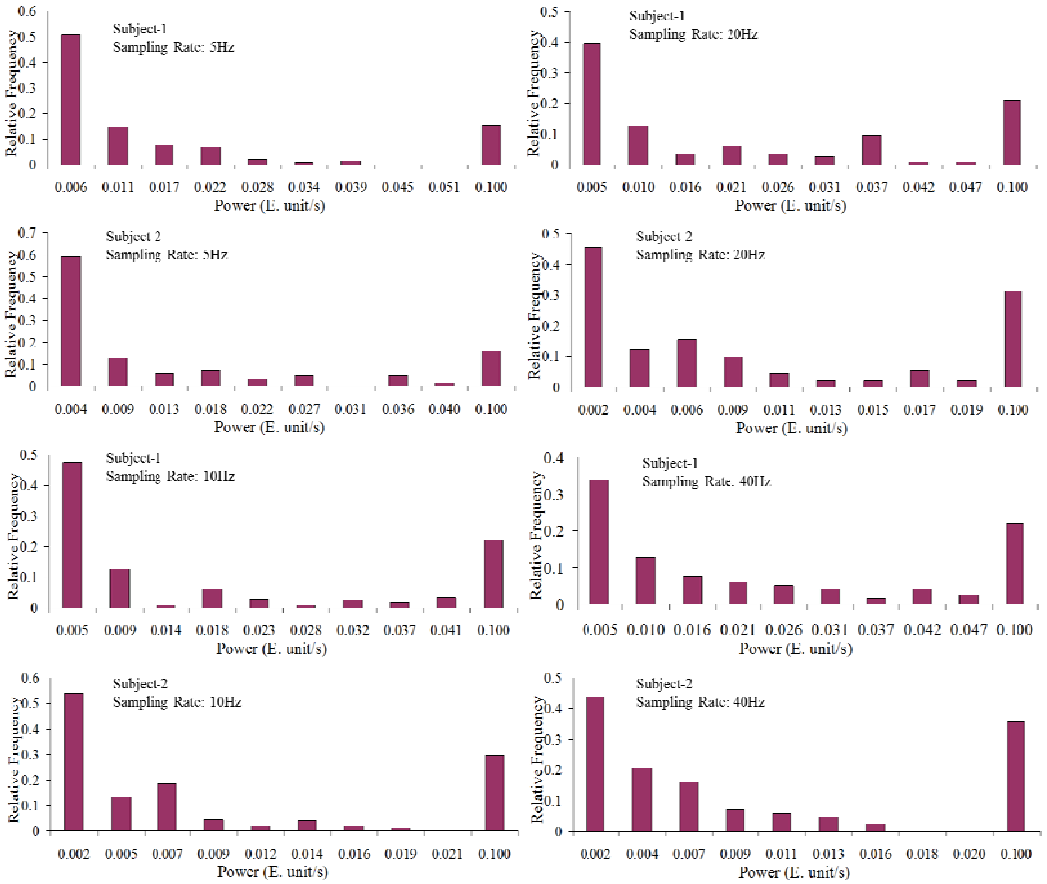

Figure 4. Acoustic power Distribution

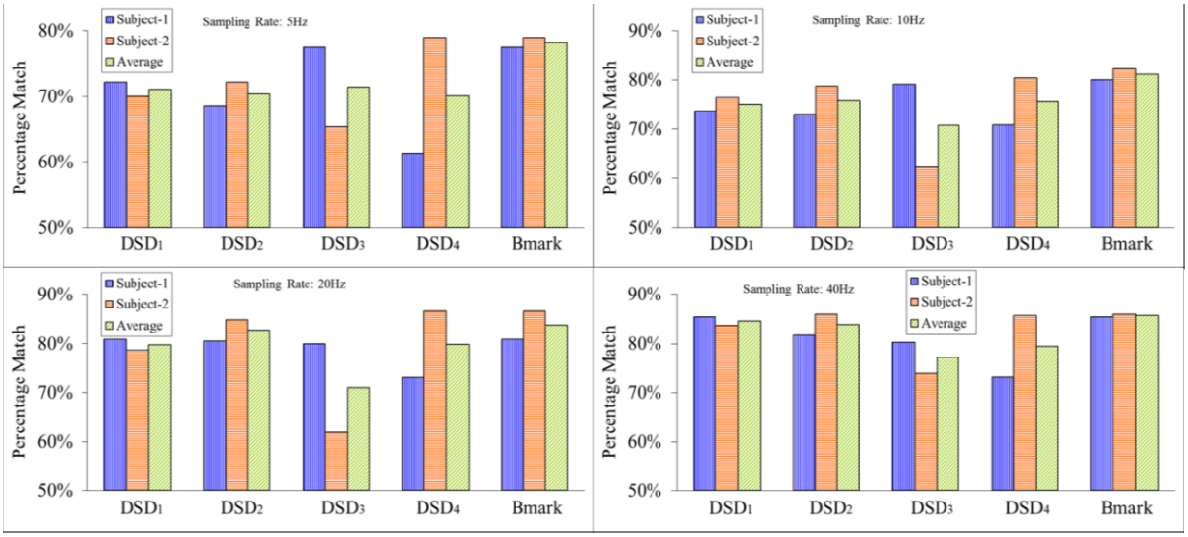

Figure 5. DSD acoustic diarization performance

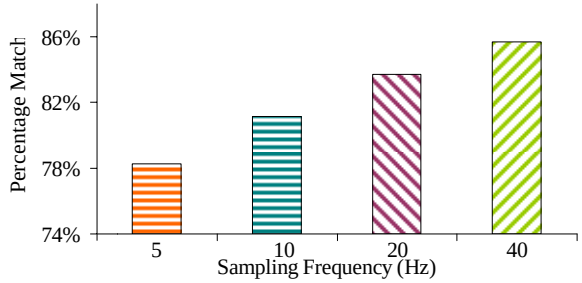

Figure 6. The impact of the sampling frequency on the performance

C. Sampling Frequency

Figure 6 shows the impact of the sampling frequency on the performance. Where sampling frequency of 5, 10, 20 and 40Hz are applied. As shown the figure, as the rate of collecting acoustic data increased, the percentage match of acoustic diarization is increased. The reason is due to of missing of some acoustic data by using low sampling frequencies, which causes missing some matching with the actual speech.

D. Conversational Distance

To see the impact of the subjects distance on the diagrammatic acoustic diarization, same conversation experiment with sampling frequency of 40Hz was repeated many times with a distance between the two subjects of 0.5, 1, 2 and 3m. Same subjects, environment and dialog were used in these experiments. The idea of doing these experiments is to see the impact of the distance on the collected acoustic data on each subject's sensor node.

错误！未找到引用源。 reports experimentally obtained average percentage match with different DSD thresholds compared with the actual recorded speech. As seen in the figure, the percentage match does not change with a distance of 1m or more, while with small distance the percentage match is reduced significantly. The fall of accuracy with small distance is caused due to the starting of acoustic interfacing from each subject to the other subject's sensor node. In other words, by reducing the distance, the microphone of each sensor node will start collecting the other subject's speech. More conclusion results will be presented in *Section F* in order to study the impact of inter-speaker distance using acoustic variation coefficient.

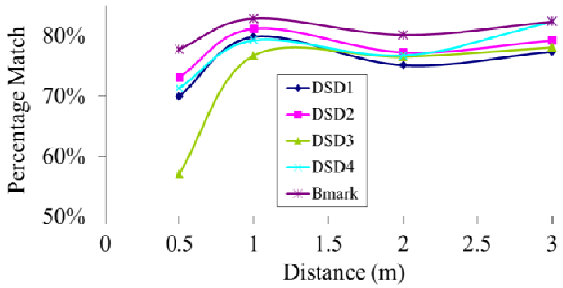

Figure 7. The impact of distance on the diarization performance

E. Comparator based Diarization

In order to study the performance of the comparator based diarization, new experiments were conducted. Same subjects and dialog were used in these experiments with a distance of 1 and 2m between the two subjects. In these experiments, a source of noise is also introduced, to see the impact of the noise on the performance. The sampled data was then processed, and AVC values at each node were computed. Then, the current speaker was decided using Eqn. 10.

Figure 8, 9, 10 and 11 show the performance results of two subject's conversations under different scenarios. The first two plots of each figure show subject-1 and subject-2 AVC output compared with the actual speech. While the third plot of each figure shows SP values in terms of $AVC_1 - AVC_2$, of subject-1 and subject-2 compared with the actual speech. On other words, the peak of AVC cross the x-axis should belong to the current speaker speech. As seen in these figures, the AVC outputs belong to the current speaker, which means the current speaker can be directly decided as discussed in *Section IV* and Eqn. 10.

The corresponding percentage match of the results that are shown in Figure 8, 9, 10 and 11 are shown in

Figure 12. Where each percentage match is computed by comparing the ACD outputs with the actual speech. As seen in Figure 8, even with different distances and noise environment, the ACD approach delivers very high state match rates (84% and above) diarization in all scenarios. This is because with ACD mechanism it was able to capture the speech activity, as shown in Eqns. 7-9, and removing the impact of the distance and noise at each node, as computed in Eqn. 10.

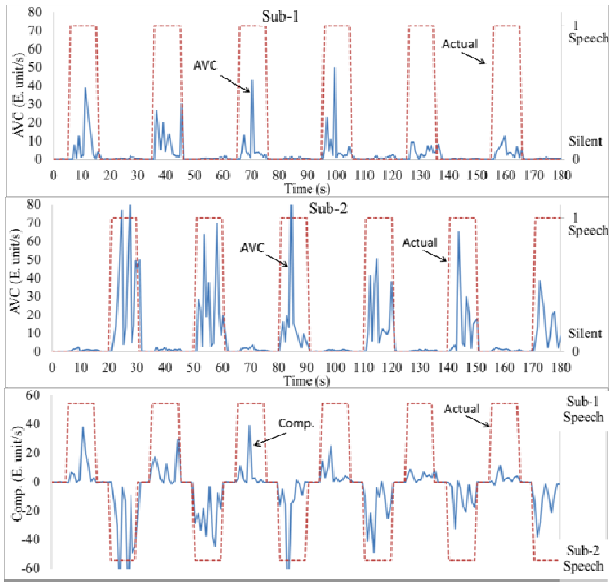

Figure 8. ACD with 1m distance and no noise

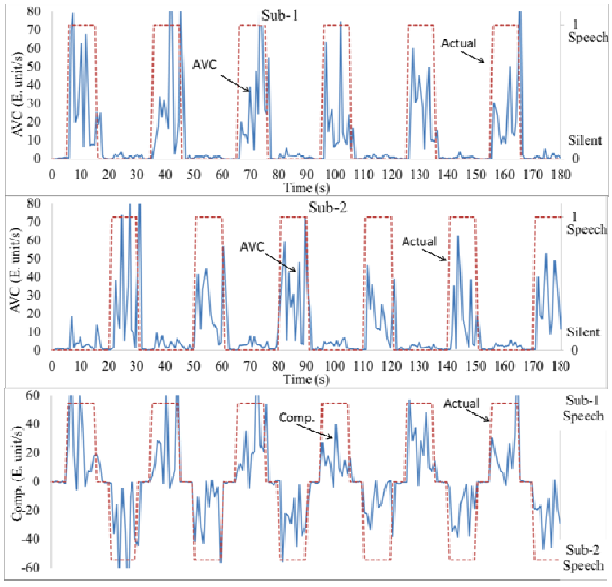

Figure 9. ACD with 1m distance and noise

F. *Inter-speaker Distance*

In order to study the impact of the distance between the source of the acoustic and the sensor node on the acoustic diarization, more experiments were conducted. Here, we are considering the small distances (less than 1m). In this new set of experiments, Acoustic Variation Coefficient *AVC* is computed, and as described in *Section IV*. In this sets of experiments, same source of acoustic is used with same dialog and environment. In each set of experiments, a distance of 10cm is changed, where a distance of 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100cm in each set of experiments is applied.

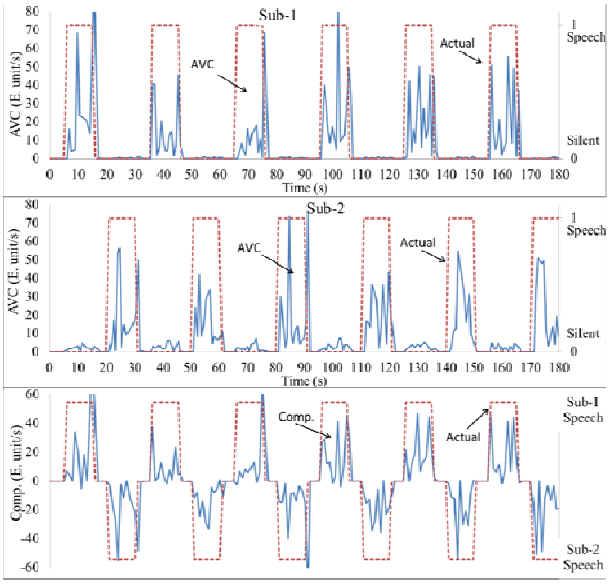

Figure 10. ACD with 2m distance and no noise

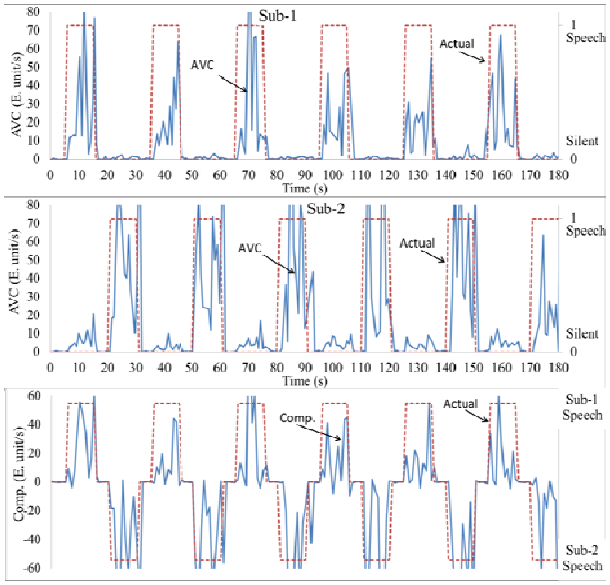

Figure 11. ACD with 2m distance and noise

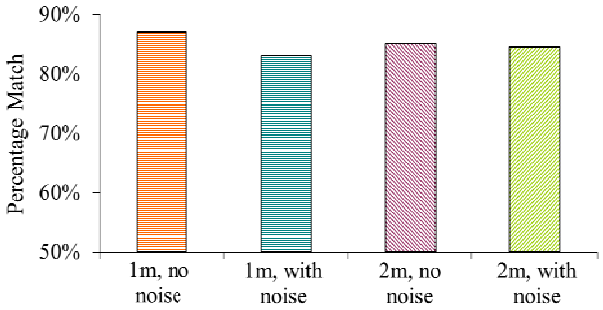

Figure 12. ACD performance results

Figure 13 shows the results, where each point in the figure represents an average of five AVC values collected form five experiments. As shown in the figure, by increasing the distance from the source of the acoustic to the sensor node, AVC goes down significantly. That

means, by increasing the distance, the sensitivity of the sensor node goes down. This is because by increasing the distance, the ADC converter output of the sensor was unable to deliver the acoustic data. Then, the acoustic data will not be collected very well at the sensor to see an impact on the computed AVC variation, as shown in Figure 13. The results in this figure clarify also why ACD performs well with different subject's distances, as shown in Figure 8, 9, 10, 11 and 12.

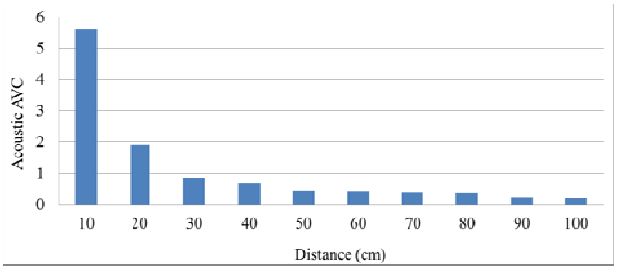

Figure 13. Acoustic variation coefficient vs. distance

VI. Conclusion And Future Work

We present an experimental framework for a wearable sensor network that can be used for networked human acoustic diarization. Diagrammatic Speech Diarization (DSD), coupled with Acoustic Comparator based Diarization (*ACD*), have been used for detecting the current speaker. It was first demonstrated that the acoustic power is shown in a form of a bimodal distribution, corresponds to the silent and speech acoustic powers. Then, DSD was proposed to decide some threshold power, by which, the current speaker can be decided. Although DSD threshold based mechanism can be used for reasonable acoustic diarization performance, the intrinsic noise and unpredictability subject's distance require a delicate dimensioning of the used threshold values for consistent acoustic diarization performance across various distance and noise. To avoid this, an ACD based detection process is applied. It was shown that the ACD method is able to consistently deliver significantly better acoustic detection performance than DSD threshold based mechanism in a more distance and noise independent manner. Ongoing work on this topic includes, adjusting the DSD, ACD and processing mechanism to adapt for many subjects and different scenarios of subjects orientation and environment noise.

References

[1] E. Jovanov, A. Milenkovic, C. Otto, P. De Groen, B. Johnson, S. Warren, and G. Taibi, "A WBAN System for Ambulatory Monitoring of Physical Activity and Health Status: Applications and Challenges," in *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, 2005, pp. 3810–3813.

[2] E. Jovanov, A. Milenkovic, C. Otto, and P. C. de Groen, "A Wireless Body Area Network of Intelligent Motion Sensors for Computer Assisted Physical Rehabilitation," *Journal NeuroEng. and Rehab*, vol. 2, no. 11, p. 6, Mar. 2005.

[3] E. Jovanov, A. Milenkovic, and C. Otto, "Wireless sensor networks for personal health monitoring: Issues and an implementation," 2006.

[4] D. Gibson, "Taking Turns and Talking Ties: Networks and Conversational Interaction," *The American Journal of Sociology*, vol. 110, no. 6, pp. 1561–1597, 2005.

[5] H. Sacks, "Lectures on Conversation." Cambridge, Mass.: Basil Blackwell, 1995.

[6] D. Reynolds, P. Kenny, and F. Castaldo, "A study of new approaches to speaker diarization," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[7] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker Diarization: A Review of Recent Research," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 356–370, 2012.

[8] M. Kotti, E. Benetos, and C. Kotropoulos, "Computationally efficient and robust BIC-based speaker segmentation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 920–933, 2008.

[9] X. Zhu, C. Barras, L. Lamel, and J.-L. Gauvain, "Multi-stage Speaker Diarization for Conference and Lecture Meetings," in *Multimodal Technologies for Perception of Humans*, vol. 4625, R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 533–542.

[10] S. Jothilakshmi, V. Ramalingam, and S. Palanivel, "Speaker diarization using autoassociative neural networks," *Engineering Applications of Artificial Intelligence*, vol. 22, no. 4–5, pp. 667–675, 2009.

[11] R. Sinha, S. E. Tranter, M. J. F. Gales, and P. C. Woodland, "The Cambridge university March 2005 speaker diarisation system," *IN PROC. INTERSPEECH*, vol. 2437, p. 2005, 2005.

[12] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic Analysis of Multimodal Group Actions in Meetings," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, pp. 305–317, Mar. 2005.

[13] O. Brdiczka, D. Vaufreydaz, J. Maisonnasse, and P. Reignier, "Unsupervised segmentation of meeting configurations and activities using speech activity detection," *Artificial Intelligence Applications and Innovations*, pp. 195–203, 2006.

[14] T. Choudhury and A. Pentland, "Characterizing Social Interactions Using the Sociometer," *PROCEEDINGS OF NAACOS 2004*, 2004.

[15] D. A. Reynolds and P. Torres-Carrasquillo, "The MIT Lincoln Laboratory RT-04F Diarization Systems: Applications to Broadcast Audio and Telephone Conversations," Nov. 2004.

[16] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[17] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1–3, pp. 19–41, 2000.

[18] P. Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors," *CRIM, Montreal, Technical Report*, 2008.

[19] S. N. Maceachern, "Estimating normal means with a conjugate style dirichlet process prior," *Communications in Statistics - Simulation and Computation*, vol. 23, no. 3, pp. 727–741, 1994.

[20] F. Valente and C. Wellekens, "Variational Bayesian methods for audio indexing," *Machine Learning for Multimodal Interaction*, pp. 307–319, 2006.

[21] "Crossbow Technology, Inc.," *http://www.xbow.com*. [Online]. Available: http://www.xbow.com.

Muhannad Quwaider is an Assistant Professor at Jordan University of Science and Technology (*JUST*). Dr. Quwaider earned his Ph.D. and M.S. at Michigan State University in East Lansing, USA, and his B.S. at Jordan University of Science and Technology in Irbid, Jordan. Prior to joining JUST in 2010, Dr. Quwaider was senior researcher in Networked Embedded and Wireless Systems (*NeEWS*) laboratory at the Electrical and Computer Engineering (ECE) Department of Michigan State University (MSU). Among other distinction, paper from his publication won best paper award in ICICS 2011. He served as TPC chair of ICICS2012 and guest editor of Elsevier Journal of Ad Hoc Networks. Dr. Quwaider is a member of IEEE and a researcher consultant in NeEWS in ECE Department of MSU. His current research interests include the broad area of wireless data networking, low-power network protocols, application-specific sensor networks, wireless network security, mobile ad hoc networks, and body area network.

Subir Biswas is a Professor and the director of the Networked Embedded and Wireless Systems laboratory at Michigan State University. He received his Ph.D. from University of Cambridge and held various research positions in NEC Research Institute, Princeton, AT&T Laboratories, Cambridge, and Tellium Optical Systems, NJ. He published over 140 peer-reviewed articles in the area of network protocols, and co-invented 6 (awarded and pending) US patents. His current research includes Pricing Calculus in Social Wireless Networks, Capacity Scavenging in Cognitive Networks, UWB Switching in Sensor Networks, Safety and Content based Applications in Vehicular Networks, Anonymous Protocols in Body Area Networks, Wearable Sensing for Health Applications, and Group Communication in DTN Networks. He is a senior member of IEEE and a fellow of Cambridge Philosophical Society.

ChaiYong Lim received the B.S degree in electrical engineering at Michigan State University (MSU), East Lansing, in 2012. He is currently pursuing on his Ph.D. at Arizona State University (ASU), Tempe. From May 2010 to May 2011, he was an undergraduate research assistant at NeEWS lab, worked on the project of wearable wireless sensors for group activity monitoring. From May 2011 to July 2012, he was an undergraduate research assistant at SML lab, worked on the dynamic modeling for ionic polymer-metal composite (IPMC) sensor. His current research interests include biological signal processing, random process modeling, system on chip, wireless neural-prosthesis system.