Using Nonverbal Information for Conversation Partners Inference by Wearable Devices

Deeporn Mungtavesinsuk[1], Yan-Ann Chen[2(✉)], Cheng-Wei Wu[1], Ensa Bajo[1], Hsin-Wei Kao[1], and Yu-Chee Tseng[1]

[1] Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan
{deeporn,cww0403}@nctu.edu.tw, bajoensa@gmail.com, scott02308@gmail.com, yctseng@cs.nctu.edu.tw

[2] Department of Computer Science and Engineering, Yuan Ze University, Taoyuan City, Taiwan
chenya@saturn.yzu.edu.tw

Abstract. In this paper, we propose a framework called conversational partner inference using nonverbal information (abbreviated as CFN). We use the wrist-based wearable device that has an accelerometer sensor to detect the user's hand movement. Besides, we propose three different methods, named *leading CFN*, *trainling CFN* and *leading-trailing CFN*, to integrate the detected movement behaviors with the sound data sensed by microphones to effectively infer conservational partners. In experiments, we collect real data to evaluate the proposed framework. The experimental results show that the accuracy of *leading CFN* is better than *trailing CFN* and *leading-trailing CFN*. Moreover, our approach shows higher accuracy than the state-of-the-art approach for conversational partner inference.

Keywords: Conversational partner inference
Nonverbal information · Social interaction analysis · Wearable devices

1 Introduction

Smartphone has become an essential tools for many people's daily life. With rapid advancement of smartphones, more and more types of sensors are embedded in smartphones. Among these sensors, the microphone is a common sensor that can be used to sense the sound around the user. Recently, various audio related applications on smartphones for inferring personal contexts have been proposed [1–4]. They use the sensed sound by smartphones to recognize ambient sounds [3], nearby speakers [1], the stress level of the user [2], and even the user's indoor location [4]. However, these applications mainly focus on individual users' contexts rather than conversation groups' contexts. Here are some of the recent studies that mainly focus on cell phone-based conversation. SocioPhone is a mobile platform that proposed by [5] for conversational monitoring.

Y.-B. Lin et al. (Eds.): IoTaaS 2017, LNICST 246, pp. 187–193, 2018.
https://doi.org/10.1007/978-3-030-00410-1_22

The main idea of their method is to use the volume of the phone to calculate and identify the speaker. However, this method assumes that the conversation group is known and does not propose a method to identify which speakers belong to the same conversation group. Socialweaver [6] uses a clustering-based approach to identify which speakers belong to the same conversation group by using a clustering algorithm. However, to achieve a considerable accuracy, it has to take approximately up to thirty minutes of conversation data, which requires a very high time cost. Afterward, a low time cost approach [7] for conversation partner inference is proposed. This approach uses the smartphone to detect the segments of speech from the speaker (also known as speaker turns [7]), and then uses associations between speaker turns of different speakers to infer conversational partners. The experimental results show that the approach has a good recognition rate. Although the above methods are committed to the study of conversation partner inference, they do not consider the speaker's body language to further enhance the inference performance. In view of this, we address the above research issue by proposing a new framework called *conversational partner inference using nonverbal information* (*CFN*). In the proposed framework, we take the wrist-wearable device into consideration for capturing the acceleration information about the user's hand movements (called moving turns). We further incorporate moving turns and speaker turns into *action turns* for inferring conversational partners in a more effective manner. Based on the way of action turn composition, we propose three methods, namely *leading CFN*, *trailing CFN*, and *leading-trailing CFN*. Extensive experimental results show that the accuracy of *leading CFN* is better than *trailing* and *leading-trailing* ones. Moreover, the results show that the accuracy is approximately 2% to 6% higher than the current best approach [7].

2 Related Work

Several studies [5,8–11] have been conducted on developing novel applications for human's social interactions. Reference [11] designs an interesting social application called E-SmallTalker on mobile platform. It can efficiently compare the common interests and friends of two conversational partners and use such information to recommend users' some chat topics to begin a conversation. Reference [10] designs a wrist-based wearable device to detect the handshake behavior of two users. If a handshake behavior is detected by the developed system, the users' mobile devices will automatically exchange e-mail addresses and social network accounts, which avoids the inconvenience caused by the exchange of traditional business cards. Reference [9] proposes an application called High5, which uses wrist-based wearable devices to detect the clapping behavior between two or more people, which can be used to increase the times of interactions between employees in a company. Reference [12] uses a wearable device attached to the thigh of the user to analyze the movement direction and acceleration of the user, which allows to detect which users belong to which moving groups. Reference [8] analyzes human's nonverbal behaviors to understand the relationships of social

interactions between people. Sociophone [5] is a an interaction monitoring platform, which uses the sound sensed by the microphone of the smartphone to know the identity of the speakers. It can be used to automatically record the user's daily conversation with others.

3 Conversation Partners Analysis

Figure 1 shows the workflow of our conversation inference mechanism. The smartphones of users form a proximity group via short-range communication. They continuously collect verbal information by recognizing their owners' speaker turns and emotions and nonverbal information from wearable sensors by understanding users' hand moving periods. After sharing information among these phones, they analyze self-conversational relationship with the others by investigating the fused data of verbal and nonverbal information. Figure 2 presents the proposed system architecture of this wearable sensing system for conversation partner inference. We then introduce components of the system architecture as follows.

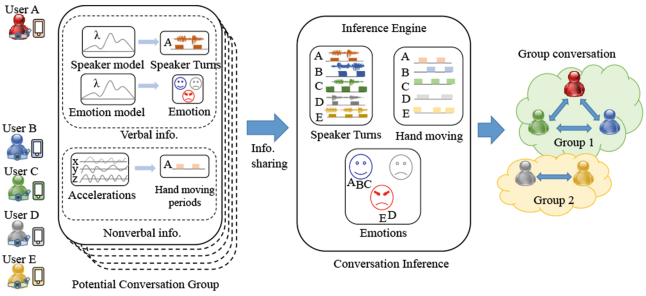


Fig. 1. The workflow of our analysis model.

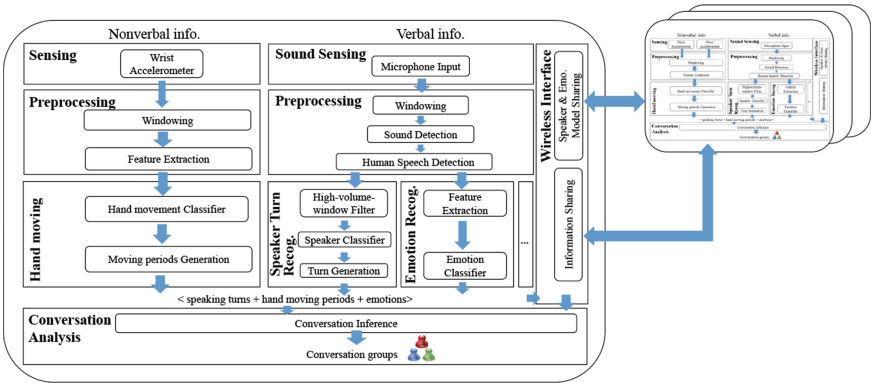


Fig. 2. The system architecture

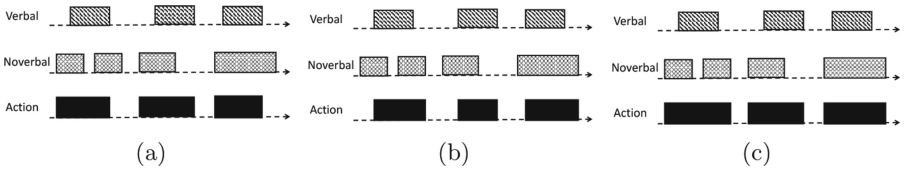
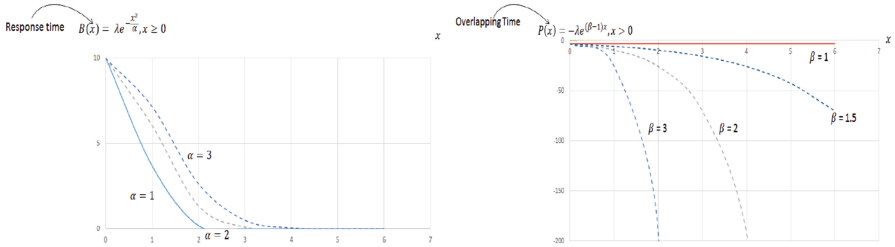


Fig. 3. Examples of (a) leading, (b) trailing, and (c) leading-trailing action turn composition.

Sensing, Processing, and Recognition Classifier. Here, we explain how to extract verbal and nonverbal information by this wearable sensing system. For verbal information, we follow the procedures of sensing, processing, and recognition in reference [7]. We then can obtain speaker turns and emotion of a user by analyzing user's voice in a conversation. For nonverbal information, we investigate user's hand movement during a conversation by analyzing accelerations from the sensor worn on the user's wrist. In the sensing component, thus, we simply measure wrist acceleration by a 3-axis accelerometer. Then, in the preprocessing component, we compute the total acceleration of 3-axis accelerations and extract 4 statistical features, *mean crossing rate*, *variance*, *average absolute deviation*, and *kurtosis*, within a period of window time. Next, the hand movement classifier utilizes these 4 features to build a recognition model to distinguish the hand movement which is caused by hand-gestures during a conversation. Finally, the component of conversation analysis will acquire a sequence of hand movement and speaking periods where we call moving turns and speaker turns, respectively.

Conversation Analysis. Through the wireless interface, one smartphone can acquire speaker turns and moving turns of nearby users. The conversation analysis component exploits these verbal and nonverbal information to infer conversation groups. The problem is how to do the data fusion with verbal and nonverbal information. Here, we define 3 methods, namely *leading CFN*, *trailing CFN*, and *leading-trailing CFN*, to compose the fused data called *action turns*. An action turn represents the period of time where a user dominates a conversation for interacting with others. The *leading CFN* is to model a situation where a speaker's hand-gesture leads a speaking sentence during a conversation. Thus, it composes an action turn by merging a speaker turn with the moving turn acted before it as shown in Fig. 3(a). On the other hand, the *trailing CFN* is to model a situation where a speaker's speaking sentence leads a hand-gesture. It composes an action turn by merging a speaker turn with the moving turn acted after it as shown in Fig. 3(b). Whereas, *leading-trailing CFN* is model both situations by merging a speaker turn with the moving turn acted around it as shown in Fig. 3(c).

To infer the group conversations in an environment, we analyze the relations among action turns of all the users. We adopt pairwise conversation possibility *Dialog Confidence(DC)* in [7] where we have a higher confidence that two people



(a) (b)

Fig. 4. (a) Bonus function (b) Penalty function.

have a conversation if their action turns are nearly close and hardly overlapping. We thus define two functions to evaluate the gap and overlapping of two adjacent turns. Given a gap x between one action turn of one user and that of the other user, the bonus function shown in Fig. 4(a) is defined as

$$f_b(x) = \lambda_1 e^{-\frac{x^2}{\alpha}} \ , \tag{1}$$

where λ_1 and α are predefined constants to control the amplitude and the slope of the curve. The value $f_b(x)$ is inversely proportional to the gap x. On the contrary, given two turns with an overlapping period x, the penalty function shown in Fig. 4(b) is defined as

$$f_p(x) = -\lambda_2 e^{(\beta-1)x} \ , \tag{2}$$

where λ_2 and β are predefined constants to control the amplitude and the slope of the curve.

Assume that there are users i and user j. we compute the DC by the following equation:

$$D_{i,j} = \frac{\sum\limits_{x_b \in B^{i,j}} f_b(x_b) + \sum\limits_{x_p \in P^{i,j}} f_p(x_p)}{|B^{i,j}| + |P^{i,j}|} \ , \tag{3}$$

where $B^{i,j}$ and $P^{i,j}$ are sets of bonus and penalty cases between user i and j's actions turns. If the dialog confidence $D_{i,j}$ is above a threshold Δ_D, we determine that users i and j have a conversation in this interval. Therefore, we can infer that user i and j is in the same conversation group. Once we apply this DC computation for all 2-combinations of users in this environment, we will know the conversation partners of each user.

4 Performance Evaluation

In this section, we evaluate the performance of the conversation inference considering nonverbal information. We conduct experiments of having a conversation with 2 or 3 speakers while recording each speaker's voice and wrist acceleration.

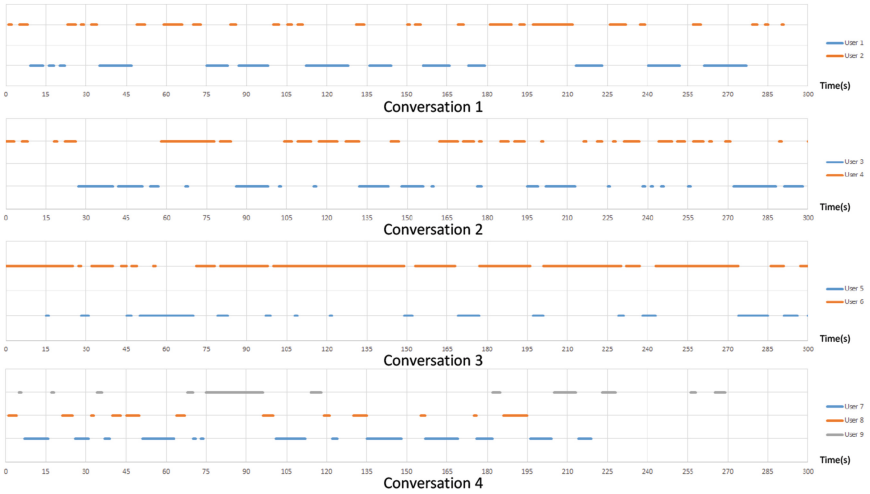


Fig. 5. Experiments of real-life conversations.

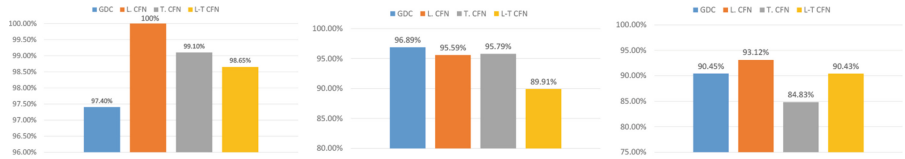


Fig. 6. Performance evaluation of scenario (a) 1, (b) 2, and (c) 3.

Figure 5 shows the domination of each speaker in the conversations where these turns are manually labelled. We define 3 scenarios of simulating concurrent group conversations in an environment from these conversation records. The scenario 1 simulates 2 concurrent group conversations by mixing the conversation 1 and 2; The scenario 2 simulates 3 concurrent group conversations by mixing the conversations 1, 2, and 3, each group has 2 speakers; Then, the scenario 3 simulates concurrent 2-speaker and 3-speaker conversation groups. Figure 6 shows performance of conversation inference while considering hand-gesture information. The performance matric is the accuracy of conversation inference which is computed by $\frac{TP+TN}{TP+FP+FN+TN}$. The TP, TN, FP, and FN represents the numbers of true positive, true negative, false positive, and false negative cases of conversation inference, respectively. The *leading CFN* can achieve better performance than the others. The reason is that the speaker may have some gestures before speaking and these gestures may enhance the continuity of action turns. However, the gesture behind a sentence may decrease the performance by overlapping with others' gesture or speaking.

5 Conclusion and Future Works

In this work, we propose a mechanism to infer conversation partner using nonverbal information with the assistance of wearable devices. We utilize wrist accelerometer to detect the movement of hand-gestures in a conversation. We observe the correlation of action turns, which are composed by verbal and nonverbal information, among speakers to enhance the performance of *Dialog Confidence* Finally, we show that the leading method of action turn composition can have 2% to 6% improvement of conversation inference.

References

1. Lu, H., Bernheim Brush, A.J., Priyantha, B., Karlson, A.K., Liu, J.: SpeakerSense: energy efficient unobtrusive speaker identification on mobile phones. In: Lyons, K., Hightower, J., Huang, E.M. (eds.) Pervasive 2011. LNCS, vol. 6696, pp. 188–205. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21726-5_12
2. Lu, H., et al.: Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In: Proceedings of ACM International Conference on Ubiquitous Computing (UbiComp) (2012)
3. Rossi, M., Feese, S., Amft, O., Braune, N., Martis, S., Troster, G.: Ambientsense: a real-time ambient sound recognition system for smartphones (2013)
4. Tarzia, S.P., Dinda, P.A., Dick, R.P., Memik, G.: Indoor localization without infrastructure using the acoustic background spectrum. In: Proceedings of International Conference on Mobile Systems, Applications, and Services (MobiSys) (2011)
5. Lee, Y., et al.: Sociophone: everyday face-to-face interaction monitoring platform using multi-phone sensor fusion. In: Proceedings of International Conference on Mobile Systems, Applications, and Services (MobiSys) (2013)
6. Luo, C., Chan, M.C.: Socialweaver: collaborative inference of human conversation networks using smartphones. In: Proceedings of ACM Conference on Embedded Networked Sensor Systems (SenSys) (2013)
7. Chen, Y.A., Chen, J., Tseng, Y.C.: Inference of conversation partners by cooperative acoustic sensing in smartphone networks. IEEE Trans. Mob. Comput. **15**(6), 1387–1400 (2016)
8. Basu, S.: Social signal processing: understanding social interactions through nonverbal behavior analysis. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (2009)
9. Kim, Y., et al.: High5: promoting interpersonal hand-to-hand touch for vibrant workplace with electrodermal sensor watches. In: Proceedings of ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) (2014)
10. Wu, F.J., Chu, F.I., Tseng, Y.C.: Cyber-physical handshake. In: Proceedings of ACM Special Interest Group on Data Communication (SIGCOMM) (2011)
11. Yang, Z., Zhang, B., Dai, J., Champion, A.C., Xuan, D., Li, D.: E-SmallTalker: a distributed mobile system for social networking in physical proximity. In: Proceedings of International Conference on Distributed Computing Systems (ICDCS) (2010)
12. Gordon, D., Wirz, M., Roggen, D., Tröster, G., Beigl, M.: Group affiliation detection using model divergence for wearable devices. In: Proceedings of International Symposium on Wearable Computers (ISWC) (2014)