Detecting Verbal and Non-Verbal Gestures Using Earables

Matías Laporte
matias.laporte@usi.ch
Università della Svizzera italiana (USI)
Switzerland

Preety Baglat
preety.baglat@usi.ch
Università della Svizzera italiana (USI)
Lugano, Switzerland

Shkurta Gashi
shkurta.gashi@usi.ch
Università della Svizzera italiana (USI)
Lugano, Switzerland

Martin Gjoreski
martin.gjoreski@usi.ch
Università della Svizzera italiana (USI)
Lugano, Switzerland

Silvia Santini
silvia.santini@usi.ch
Università della Svizzera italiana (USI)
Lugano, Switzerland

Marc Langheinrich
marc.langheinrich@usi.ch
Università della Svizzera italiana (USI)
Lugano, Switzerland

ABSTRACT

Verbal and non-verbal activities convey insightful information about people's affect, empathy, and engagement during social interactions. In this paper, we investigate the usage of inertial sensors to recognize verbal (e.g., *speaking*), non-verbal (e.g., *head nodding*, *shaking*) and other activities (e.g., *eating*, *no movement*). We implement an end-to-end deep neural network to distinguish among these activities. We then explore the generalizability of the approach in three scenarios: (1) using new data to detect a known activity from a known user, (2) detecting a novel activity of a known user and (3) detecting the activity of an unknown user. Results show that using accelerometer and gyroscope sensors, the model achieves a balanced accuracy of 55% when tested on data from a new user, 41% on a new activity of an existing user, and 80% on new data of a known activity from an existing user. The results are between 7-47 percentage points better than baseline classifiers used for comparison.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

KEYWORDS

datasets, earable computing, head gestures recognition, memory recall

ACM Reference Format:
Matías Laporte, Preety Baglat, Shkurta Gashi, Martin Gjoreski, Silvia Santini, and Marc Langheinrich. 2018. Detecting Verbal and Non-Verbal Gestures Using Earables. In *EarComp '21: In Proceedings of the 2nd International Workshop on Earable Computing In Conjunction with UbiComp 2021, September 25, 2021, ,* . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

EarComp '21, September 25, 2021, ,

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Our motivation for detecting verbal and non-verbal activities is rooted in our work on building human memory augmentation systems. Using earable computing, we attempt to recognize different types of human activities, in particular head gestures, with the purpose of detecting when a social interaction is taking place. This is because the presence of others and our interactions with them play important roles in our memories, both during the formation of memory and at retrieval time: moments of social interactions might be easier to remember (formation time), and remembering a particular interaction might also help to remind us of particular details (retrieval time).

The importance of human memory for our daily lives cannot be overstated. It gives us an identity, lets us remember future intentions, carry quotidian tasks, and obtain new knowledge. It also allows us to share experiences and maintain and nurture relationships [19]. Therefore, civilization has applied increasingly complex methods to preserve its memories and overcome their failures. Today, capture technology such as cameras, voice recorders, and fitness trackers are coming close to making total capture (and, consequently, total recall [2]) a possibility, if not already a reality [12]. However, even if every part of our lives is captured and recorded, it is far from trivial to then use this information to aid our memory.

Memory augmentation systems will only succeed as long as they are able to appropriately select the relevant memories for the user [30]. In fact, instead of presenting the user a fully recorded memory, these systems should take advantage of the power of *memory cues* – objects or events that help us remember our original memory or intent. By prompting the user with such a (small) cue, they will be able to recall the original experience in great detail. One key challenge here is to identify appropriate cues among the recorded data that, when played back to the user, will trigger such recall. Social interaction might mark important moments that may make useful memory cues.

Social interaction can easily be detected using audio sensing: detecting a conversation is a sure sign of interpersonal activity. Similarly, closely tracking the movements and orientation of people could allow us to identify social interaction. Alternatively, a wearable camera may pick up faces of others and identify social interaction. All three options rely on highly sensitive personal data. Instead, we seek to identify head gestures to detect both verbal and nonverbal social interaction.

To summarize, this paper presents the following contributions:

- We present a new dataset comprised of accelerometer and gyroscope data collected using ear-worn devices. The dataset was collected from 10 participants while performing 5 activities: *nodding*, *speaking*, *eating*, *staying*, and *head shaking*.
- We investigate the feasibility of using a deep Convolutional Neural Network (CNN) to recognize activities related to social interactions such as verbal (e.g., *speaking*) and non-verbal (e.g., *nodding*) gestures, as well as gestures unrelated to interactions (e.g., *eating*).

2 RELATED WORK

The continuous development of unobtrusive wearable sensors has made possible the recording of new types of data in uncontrolled settings. Of particular interest to our work is the use of earable sensors, i.e., head-mounted in-ear/behind-the-ear sensors, to detect speech and head gestures, as cues for human interaction. As previously mentioned, other approaches (e.g., cameras and microphones) require an involved setup with additional privacy issues to consider.

Current earable devices can accommodate several sensors (e.g. accelerometers, gyroscopes, microphones or biometric sensors) and actuators (e.g., speaker) in a comfortable size with decent battery autonomy, allowing not only sound and head movement measurements, but also of head rotation and bio-signals, among others.

2.1 Earable Systems

Earable sensors have been proposed as a tool with "enormous potential in accelerating our understanding of a wide range of human activities in a nonintrusive manner", with applications ranging from "health tracking" to "contextual notification management" , including "cognitive assistance" and "lifelogging" [21].

Among the applications to deepen our understanding of human behaviour, Frohn et al. [10] have used an earable sensor to characterize the emotional intent of study participants performing a series of scripted scenes. Although the results were limited due to the reduced sample size and the use of non-actors, they showed that participants act more energetically and in sync when the scenes have a positive intent, than otherwise.

Röddiger et al. [29] instead used in-ear accelerometer and gyroscope sensors for health tracking, by measuring the respiration rate of the participants.

Other applications include: EarDynamic [32], a biometric-based authentication method which models users' ear canal deformation through the emission of inaudible audio signals and their reflections; and EarBuddy [33], a gesture recognition system which uses the microphones on the earbuds to detect different types of finger touches in the face.

Although these approaches have done novel applications with the available earable technology, none of them have focused on the detection of human behavior.

2.2 Human Behaviour Detection

Earable sensors benefit particularly from the proximity and contact with the face to be able to distinguish the movements of the jawbone and the activation of the different muscles.

EarBit [1], for example, was a prototype with multimodal (acoustic, motion) sensors to detect chewing episodes. It used an optical proximity sensor to measure the deformation of the ear canal produced by the movement activity of the mandibular bone, and a 9-axis Inertial Measurement Unit (IMU) to capture the movement of the temporalis muscle, used when chewing. Earbit also included a microphone located around the neck to detect swallowing events. A chest-mounted GoPro was used as ground truth collector. Auracle [3] is another example of eating detection, but with the use of a contact microphone instead, and an unobtrusive ground truth collector embedded into a cap.

In STEAR [28], ear mounted IMU sensors have been proposed as a new approach for step counting, with the benefit of not being affected by random motions of leg and hand, like it would happen with a smartphone or a smartwatch, respectively.

There also exists previous research on the recognition of head gestures and human activities, even in social interaction settings, with the use of earables.

Gjoreski et al. [15] used a 9-axis IMU to detect 8 individual daily life activities from a dataset of 4 subjects. Ferlini et al. [9] used an ear-worn device to track head rotations while performing activities like chewing and speaking. Min et al. [26] used an IMU sensor and a microphone for monitoring conversational well-being, using models that recognize speaking activities, altogether with stress and emotion detection. Tan et al. [31] used earable devices to detect the head orientation of interacting groups and used it as a cue for directed social attention. Lee et al. [23] focused instead on the recognition of smile and frowns gestures, while Islam et al. [20] proposed an activity recognition framework differentiating between head and mouth related activities (e.g. head shaking, nodding, eating and speaking), and normal activities (e.g. staying, walking and speaking while walking).

Our work further expands on Islam et al.'s by considering the detection of verbal and non-verbal gestures in the context of social interactions, with the intent of marking part of those moments as important for the use of memory cues.

2.3 Human Memory Augmentation

The idea of a system that stores one's digital records (e.g., documents, images, multimedia etc.) for a lifetime goes back to the 1945 vision of the Memex by Vannevar Bush [6]. While Bush did not detail the exact technology for implementing his vision, he predicted an era when storage will be virtually unlimited. Some 60 years later, the MyLifeBits project attempted to fulfill the promise of Bush's vision [13]. MyLifeBits started as a platform that could log all personal information generated and accessed on a PC, but its memory enhancing aspects quickly emerged [12]. More recently, Davies et al. [8] described the vision and core architectural building blocks of a future pervasive memory augmentation ecosystem, while Harvey et al. describe the role of lifelogging technology in this vision [18].

3 DATA COLLECTION

We provide below details about the participants, the type of data we collected, the tools used to do it and the data collection procedure.

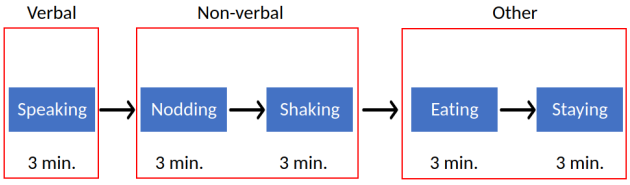

Figure 1: Experimental Protocol. Participants performed each activity during 3 minutes, with no particular order.

3.1 Collected Data and Tools

For each participant, we recorded data from the accelerometer and gyroscope sensors. To collect sensor data, we used the eSense earbuds developed by Kawsar et al. [22] at Nokia Bell Labs. The eSense earbuds are equipped with 6-axis Inertial Measurement Unit (IMU) sensors, comprised of 3-axis accelerometer and 3-axis gyroscope sensors [5]. Being worn on the ear, the eSense is suited for gathering sensor data for detecting human gestures in an unobtrusive and continuous manner.

The acccelerometer sensor measures the acceleration of the device in G-force [5]. The gyroscope sensor measures the rotation of the device in degrees per second (deg/sec) [5]. Acceleration and gyroscope data measured from ear-worn devices have been shown to reflect the movements of the head and facial muscles [1, 21, 23]. Thereby, they seem suitable to detect whether a person is interacting with another individual or not. The eSense device also contains a microphone sensor, which could be used to detect verbal activities during social interactions such as e.g., speaking. However, microphone use raises privacy concerns for users [24, 26] and is not suitable to detect non-verbal types of interactions (e.g., nodding).

To collect the sensor data, we use the eSense app[1]. The eSense app is a smartphone application developed for the Android operating system. The application was initially implemented by Islam et al. [20] and then extended by Frohn et al. [10]. The app connects via Bluetooth Low Energy (BLE) to the eSense earbuds and obtains the sensor data. We set the sampling rate of the sensors to 25 Hz.

3.2 Participants and Procedure

We recruited 10 participants (6 females and 4 males). The majority of the participants were between 18 - 34 years old, and one participant was above 55 years old. The participants had different occupations such as e.g., worker (3), postdoctoral researcher (1), Ph.D. student (1), and University student (5).

Before meeting each participant, we charged the eSense and the mobile phone. Previous to the experiment, the researcher responsible for running the data collection, explained the study goal and the data collection procedure to the participant. All participants signed an informed consent form. The experimenter provided the left earbud to the participant and instructed how to wear it. The left eSense earbud was then connected to the Android application. The participant was then instructed to first select the activity they wished to perform, and then to select the start button on the app to record the sensor data. At the end of recording an activity, the participants stopped the data recording and repeated the same procedure for another activity.

The participants performed five activities, namely, *nodding*, *speaking*, *eating*, *standing still*, and *head shaking*. We choose these activities to investigate whether verbal and non-verbal interaction activities (e.g., *speaking* and *nodding*) are distinguishable from other head and mouth-related activities (e.g., *eating*, *head shaking*) as well as no activity at all (e.g., *standing still*). The participants performed each activity for 3 minutes, one after the other, and they were free to pick the order in which the activities were performed. A simple diagram of this procedure can be seen in Figure 1.

4 DATA ANALYSIS

The main goal of our work is to develop a method to recognize human verbal, non-verbal interactions or no interactions using inertial signals. In this section, we describe the end-to-end deep learning pipeline we developed as well as the evaluation procedures, metrics and baselines used.

4.1 Data Pre-processing

To pre-process the signals, we follow common pre-processing steps used in the literature for human activity recognition from inertial signals [5]. In particular, after dividing the dataset into train and test splits, we segment the sensor data for each split into 4 seconds windows with 75% overlap. After the segmentation, our final dataset contains 1210 *speaking* samples, 1162 *nodding*, 1272 *eating*, 1179 *head shaking* and 1127 *standing still*. The measurement unit of acceleration data is converted to ± 4g and gyroscope data to ± 500 deg/s directly in the application used to collect the data.

4.2 Convolutional Neural Network (CNN)

We developed an end-to-end CNN, which takes as input the 4-second windows of raw accelerometer and gyroscope signals (see Figure 2). The accelerometer and gyroscope sensor data is first processed by three convolutional layers, each with a kernel size of 7, 128 feature maps and ReLU activation function. These layers learn feature representation from the raw sensor data. The output of the last convolutional layer is then flattened and provided as input to a max pooling layer. To avoid over-fitting, we employed dropout regularization with dropout rate of 0.5. The output of the last layer of the model is provided as input to a `sigmoid` function, which returns a k dimensional output with estimated probability between 0 and 1, where k is the number of activity classes, which is 3 (*non-verbal*, *verbal*, or *other*).

4.3 Evaluation Procedures

To evaluate the performance of the CNN classifier, we follow common procedures in machine learning [14, 27]. In particular, we investigate three validation procedures described as following.
Leave-one-part-out (LOPO) validation procedure uses the data of all participants, except one, and the first 80% of the data for each activity from the left-out user in the training set. The remaining 20% of each activity of the left-out user is used for the test set. The procedure is repeated for all participants and the results are reported as average of all iterations. This approach verifies the ability of the model to generalize to unseen data of a known user. In

[1]https://github.com/SabrinaFrohn/Esense

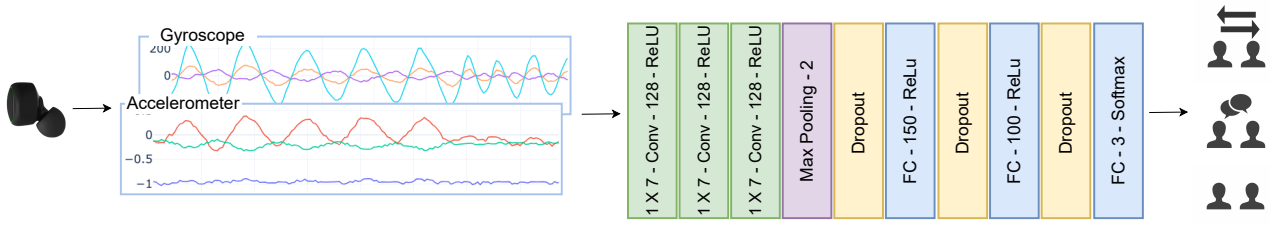

Figure 2: Overview of the end-to-end deep learning pipeline. The raw acceleremeter and gyroscope data collected from eSense earbuds is provided as input the CNN. The model classifies each data sample as a verbal, non-verbal or other activity.

addition, it avoids the *temporal leak* issue, discussed in [7], which refers to situations when a model is trained on data from the future. With this approach, we ensure that the test set is posterior to the data in the training set.

Leave-one-activity-out (LOAO) evaluation approach uses all the data of all users, except one activity of one user, in the training set. The left-out activity of the user is kept as the test set. The same procedure is repeated for all activities of the left-out user and for all the users. We report the classification results as the average of all iterations. The main goal of this technique is to avoid having segments from a same trace (e.g., activity) collected from one user in both training and testing set. This is because adjacent segments are not statistically independent as discussed in [17]. This approach verifies the ability of the model to recognize new activities from a known user by learning the presence of a particular activity in the training set from other users.

Leave-one-subject-out (LOSO) validation scheme uses the data of all users except one in the training set and the left-out user as the test set, as used in [23, 24]. This procedure is repeated for all the users in the dataset. We report the classification results as the mean metrics for all users. This validation procedure ensures that the activities performed by the same user are not present simultaneously in the train and test sets. With this technique, we aim to investigate the generalization of the model to new users.

4.4 Evaluation Metrics

To evaluate the performance of the model, we use *accuracy*, *balanced accuracy* and *F1*. Accuracy quantifies the number of samples correctly classified by the model [27]. Balanced accuracy score is defined as the average of recall score obtained in each class [4]. This score is suitable to compare the performance of imbalanced datasets because it also takes into consideration the minority class. To further explore the performance of the classifier in all the classes, we report also the F1 metric. The F1 score is the harmonic mean of precision and recall [27].

4.5 Baseline Classifiers

We compare the performance of the CNN with Random Guess (RG) and Biased Random Guess (BRG) baselines. RG provides a classification uniformly at random. BRG takes into consideration the distribution of the classes in the training set and generates a biased prediction. In particular, BRG always predicts the most frequent label in the training set, as used in [11].

5 RESULTS

In what follows, we present and discuss the evaluation results. We first report the performance of the CNN using different evaluation procedures and baseline classifiers described in Section 4. We then investigate the performance of each sensor separately (`unimodal`) and their combination (`multimodal`).

5.1 Evaluation Procedures Comparison

We first compare the performance of the CNN model using `LOPO`, `LOAO` and `LOSO` validation techniques. Figure 3 shows the balanced accuracy of CNN and baseline classifiers for each evaluation technique. These results imply that it is feasible to use ear-worn devices to distinguish between verbal, non-verbal and other activities performed during social interactions. Overall the classification results using `LOPO` are significantly higher than using `LOAO` or `LOSO`. In particular, the CNN has a balanced accuracy of 80%, which is 25 and 39 percentage points increment from `LOSO` and `LOAO` validation techniques. As expected, the presence of annotated data from the test user allows the model to achieve a higher performance. Therefore, future systems that aim to distinguish between verbal, nonverbal and other activities using earbuds, should first train the model with data from the user to avoid the cold start problem. The performance drop of the CNN when using the `LOSO` or `LOAO` techniques, suggests that such systems are difficult to generalize to the data of a new user or new activity of a user. Given that `LOPO` validation procedure provides the best results, in the next experiments we present more detailed results for this validation procedure.

5.2 Comparison to Baseline Classifiers

Figure 4 shows different classification metrics for RG, BRG and CNN classifiers, using the best validation procedure explored in this work, `LOPO`. In particular, balanced accuracy for the CNN is 80%, 36% for the RG and 33% for the BRG. Our model shows 44 and 47 percentage points increment compared to RG and BRG classifiers.

Figure 5 shows the balanced accuracy for each participant using the LOPO validation approach and the CNN, as the best model available among those tested. We observe that the performance of the CNN for the majority of the users is higher than 60%, with the exception being users P01 and P03.

5.3 Unimodal vs Multimodal

In this set of results, we investigate the classification performance of training with single (`unimodal`) and multiple sensors (`multimodal`).

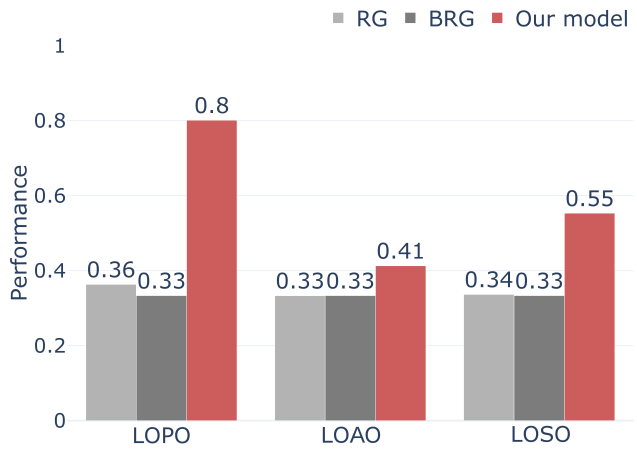

Figure 3: Classification results of CNN model using leave-one-part-out (LOPO), leave-one-session-out (LOSO) and leave-one-user-out (LOUO) validation techniques.

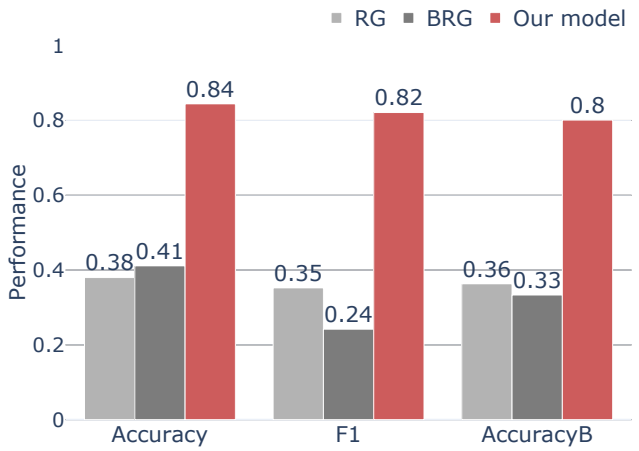

Figure 4: Accuracy, F1, and balanced accuracy (AccuracyB) for CNN, RG, and BRG classifiers using accelerometer and gyroscope data as input and LOPO validation accuracy.

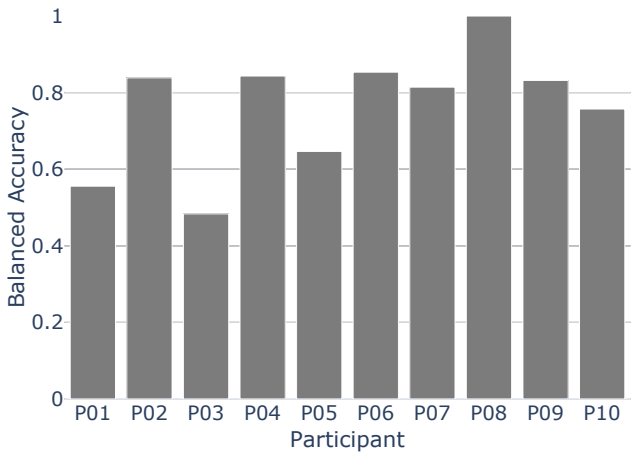

Figure 5: Balanced accuracy of CNN model for each participant using accelerometer and gyroscope signals as input and LOPO validation procedure.

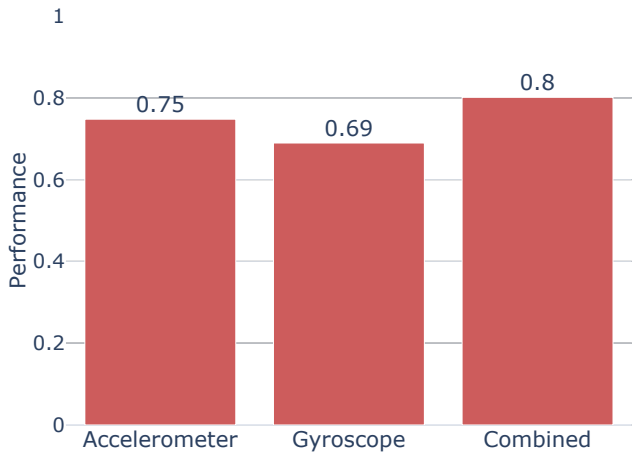

Figure 6: Comparison of the performance of the CNN model using only accelerometer, only gyroscope or both as input.

Unimodal refers to experiments where only one sensor's data (e.g., accelerometer) is used as input to the CNN model. Multimodal refers to experiments using both accelerometer and gyroscope data. Figure 6 presents the balanced accuracy scores obtained for unimodal and multimodal approaches using the LOPO validation procedure. The balanced accuracy for the accelerometer data is 75%, for the gyroscope data is 69%, and for their combination (accelerometer and gyroscope) is 80%. We observe that the performance of the multimodal classifier is higher than the performance of unimodal classifiers by 5 and 11 percentage points respectively. These results imply that combining data from accelerometer and gyroscope sensors allows recognizing user's interactions better than using only one of these sensors. This outcome highlights the importance of considering not only the movements but also the rotation angle of the device during these activities, which is in line with other end-to-end deep learning studies on activity recognition [16].

6 LIMITATIONS AND FUTURE WORK

An important limitation of our study is that it relies on data collected in a controlled laboratory setting. This might not reflect the challenges of collecting such data in real-world scenarios. In future work, we plan to run a larger study in naturalistic settings and verify the generalizability of our approach to new settings where users' movements are not constrained. In addition, we plan to investigate the relationship between the frequency of occurrence of such activities to participants' memory recall.

We segmented accelerometer and gyroscope sensor data using a sliding window of 4 seconds with 75% overlap. Other studies, like

[20], used different window and step sizes. We did not investigate the effect of these variables on our results. In future work, we plan to experiment with different segmentation strategies (e.g., overlapping, non-overlapping) and window sizes used in [24, 25].

7 CONCLUSIONS

Social interaction presents an interesting feature for identifying moments that lend themselves to memory cue generation. Instead of using video, audio, or location tracking technology, we envision the use of unobtrusive inertial sensors to identify moments of social interaction – both verbal and non-verbal.

While end-to-end deep learning offers the possibility to build activity recognition models without feature engineering, it may also require larger datasets to train those model. Since the dataset used in this study is relatively small, in the future we plan to implement shallow, feature-based classifiers. Another option would be to increase the size of the dataset and to implement other end-to-end deep learning architectures.

ACKNOWLEDGMENTS

We thank Nokia Bell Labs for donating the eSense devices to us and study participants for volunteering to be part of the experiment.

REFERENCES

[1] Abdelkareem Bedri, Richard Li, Malcolm Haynes, Raj Prateek Kosaraju, Ishaan Grover, Temiloluwa Prioleau, Min Yan Beh, Mayank Goel, Thad Starner, and Gregory Abowd. 2017. EarBit: Using Wearable Sensors to Detect Eating Episodes in Unconstrained Environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)* 1, 3 (2017), 1–20.

[2] C.G. Bell, J.G.G. Bell, J. Gemmell, and B. Gates. [n.d.]. *Total Recall: How the e-Memory Revolution Will Change Everything*. Dutton. https://books.google.it/books?id=AoBdPgAACAAJ

[3] Shengjie Bi, Tao Wang, Nicole Tobias, Josephine Nordrum, Shang Wang, George Halvorsen, Sougata Sen, Ronald Peterson, Kofi Odame, Kelly Caine, et al. 2018. Auracle: Detecting Eating Episodes With an Ear-mounted Sensor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)* 2, 3 (2018), 1–27.

[4] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. 2010. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*. IEEE, 3121–3124.

[5] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors. *ACM Computing Surveys (CSUR)* 46, 3 (2014), 1–33.

[6] Vannevar Bush et al. 1945. As we may think. *The atlantic monthly* 176, 1 (1945), 101–108.

[7] Francois Chollet. 2017. *Deep Learning with Python*. Manning Publications Company.

[8] Nigel Davies, Adrian Friday, Sarah Clinch, Corina Sas, Marc Langheinrich, Geoff Ward, and Albrecht Schmidt. [n.d.]. Security and Privacy Implications of Pervasive Memory Augmentation. 14, 1 ([n. d.]), 44–53. https://doi.org/10.1109/MPRV.2015.13

[9] Andrea Ferlini, Alessandro Montanari, Cecilia Mascolo, and Robert Harle. 2019. Head Motion Tracking Through in-Ear Wearables. In *Proceedings of the 1st International Workshop on Earable Computing*. 8–13.

[10] Sabrina AL Frohn, Jeevan S Matharu, and Jamie A Ward. 2020. Towards a Characterisation of Emotional Intent During Scripted Scenes Using In-ear Movement Sensors. In *Proceedings of the 2020 International Symposium on Wearable Computers*. 37–39.

[11] Shkurta Gashi, Elena Di Lascio, Bianca Stancu, Vedant Das Swain, Varun Mishra, Martin Gjoreski, and Silvia Santini. 2020. Detection of Artifacts in Ambulatory Electrodermal Activity Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–31.

[12] Jim Gemmell, Gordon Bell, and Roger Lueder. [n.d.]. MyLifeBits: A Personal Database for Everything. 49, 1 ([n. d.]), 88–95. https://doi.org/10.1145/1107458.1107460

[13] Jim Gemmell, Gordon Bell, Roger Lueder, Steven Drucker, and Curtis Wong. 2002. MyLifeBits: Fulfilling the Memex Vision. In *Proceedings of the Tenth ACM International Conference on Multimedia* (Juan-les-Pins, France) *(MULTIMEDIA '02)*. Association for Computing Machinery, New York, NY, USA, 235–238. https://doi.org/10.1145/641007.641053

[14] Aurélien Géron. 2019. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.

[15] Hristijan Gjoreski, Ivana Kiprijanovska, Simon Stankoski, Stefan Kalabakov, John Broulidakis, Charles Nduka, and Martin Gjoreski. [n.d.]. Head-Ar: Human Activity Recognition with Head-Mounted IMU Using Weighted Ensemble Learning. In *Activity and Behavior Computing*, Md Atiqur Rahman Ahad, Sozo Inoue, Daniel Roggen, and Kaori Fujinami (Eds.). Springer Singapore, 153–167.

[16] Martin Gjoreski, Vito Janko, Gašper Slapničar, Miha Mlakar, Nina Reščič, Jani Bizjak, Vid Drobnič, Matej Marinko, Nejc Mlakar, Mitja Luštrek, et al. 2020. Classical and deep learning methods for recognizing human activities and modes of transportation with smartphone sensors. *Information Fusion* 62 (2020), 47–62.

[17] Nils Y Hammerla and Thomas Plötz. 2015. Let's (not) Stick Together: Pairwise Similarity Biases Cross-validation in Activity Recognition. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 1041–1051.

[18] Morgan Harvey, Marc Langheinrich, and Geoff Ward. [n.d.]. Remembering through Lifelogging: A Survey of Human Memory Augmentation. 27 ([n. d.]), 14–26. https://doi.org/10.1016/j.pmcj.2015.12.002

[19] Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, and Ken Wood. [n.d.]. SenseCam: A Retrospective Memory Aid. In *UbiComp 2006: Ubiquitous Computing*, Paul Dourish and Adrian Friday (Eds.). Vol. 4206. Springer Berlin Heidelberg, 177–193. https://doi.org/10.1007/11853565_11

[20] Md Shafiqul Islam, Tahera Hossain, Md Atiqur Rahman Ahad, and Sozo Inoue. 2021. Exploring Human Activities Using eSense Earable Device. In *Activity and Behavior Computing (ABC)*. Springer, 169–185.

[21] Fahim Kawsar, Chulhong Min, Akhil Mathur, and Allesandro Montanari. 2018. Earables For Personal-Scale Behavior Analytics. *IEEE Pervasive Computing* 17, 3 (2018), 83–89.

[22] Fahim Kawsar, Chulhong Min, Akhil Mathur, Alessandro Montanari, Utku Günay Acer, and Marc Van den Broeck. 2018. eSense: Open Earable Platform for Human Sensing. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. 371–372.

[23] Seungchul Lee, Chulhong Min, Alessandro Montanari, Akhil Mathur, Youngjae Chang, Junehwa Song, and Fahim Kawsar. 2019. Automatic Smile and Frown Recognition with Kinetic Earables. In *Proceedings of the 10th Augmented Human International Conference 2019*. 1–4.

[24] Roya Lotfi, George Tzanetakis, Rasit Eskicioglu, and Pourang Irani. 2020. A Comparison Between Audio and IMU Data to Detect Chewing Events Based on an Earable Device. In *Proceedings of the 11th Augmented Human International Conference (AH)*. 1–8.

[25] Chulhong Min, Akhil Mathur, and Fahim Kawsar. 2018. Exploring Audio and Kinetic Sensing on Earable Devices. In *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications (WearSys)*. 5–10.

[26] Chulhong Min, Alessandro Montanari, Akhil Mathur, Seungchul Lee, and Fahim Kawsar. 2018. Cross-modal Approach for Conversational Well-being Monitoring With Multi-sensory Earables. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp)*. 706–709.

[27] Andreas C Müller and Sarah Guido. 2016. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. " O'Reilly Media, Inc.".

[28] Jay Prakash, Zhijian Yang, Yu-Lin Wei, and Romit Roy Choudhury. 2019. STEAR: Robust Step Counting from Earables. In *Proceedings of the 1st International Workshop on Earable Computing*. 36–41.

[29] Tobias Röddiger, Daniel Wolffram, David Laubenstein, Matthias Budde, and Michael Beigl. 2019. Towards Respiration Rate Monitoring Using an In-Ear Headphone Inertial Measurement Unit. In *Proceedings of the 1st International Workshop on Earable Computing*. 48–53.

[30] Abigail J. Sellen and Steve Whittaker. [n.d.]. Beyond Total Capture: A Constructive Critique of Lifelogging. 53, 5 ([n. d.]), 70. https://doi.org/10.1145/1735223.1735243

[31] Stephanie Tan, David MJ Tax, and Hayley Hung. 2021. Multimodal Joint Head Orientation Estimation in Interacting Groups via Proxemics and Interaction Dynamics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)* 5, 1 (2021), 1–22.

[32] Zi Wang, Sheng Tan, Linghan Zhang, Yili Ren, Zhi Wang, and Jie Yang. 2021. EarDynamic: An Ear Canal Deformation Based Continuous User Authentication Using In-Ear Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–27.

[33] Xuhai Xu, Haitian Shi, Xin Yi, Wenjia Liu, Yukang Yan, Yuanchun Shi, Alex Mariakakis, Jennifer Mankoff, and Anind K Dey. 2020. EarBuddy: Enabling On-Face Interaction via Wireless Earbuds. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI)*. 1–14.