Cross-Modal Approach for Conversational Well-being Monitoring with Multi-Sensory Earables

Chulhong Min
Nokia Bell Labs
Cambridge, UK
chulhong.min@nokia-bell-labs.com

Seungchul Lee
KAIST and Nokia Bell Labs
Daejeon, South Korea
seungchul@nclab.kaist.ac.kr

Alessandro Montanari
Nokia Bell Labs
Cambridge, UK
alessandro.montanari@nokia-bell-labs.com

Fahim Kawsar
Nokia Bell Labs and TU Delft
Cambridge, UK
fahim.kawsar@nokia-bell-labs.com

Akhil Mathur
Nokia Bell Labs and UCL
Cambridge, UK
akhil.mathur@nokia-bell-labs.com

UbiComp/ISWC'18 Adjunct,, October 8–12, 2018, Singapore, Singapore
ACM 978-1-4503-5966-5/18/10.
https://doi.org/10.1145/3267305.3267695

Abstract
We propose a *cross-modal* approach for conversational well-being monitoring with a *multi-sensory earable*. It consists of motion, audio, and BLE models on earables. Using the IMU sensor, the microphone, and BLE scanning, the models detect speaking activities, stress and emotion, and participants in the conversation, respectively. We discuss the feasibility in qualifying conversations with our purpose-built cross-modal model in an energy-efficient and privacy-preserving way. With the cross-modal model, we develop a mobile application that qualifies on-going conversations and provides personalised feedback on social well-being.

Author Keywords
Earable; well-being, multi-sensory

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

Introduction
Conversations have profound implications for personal-scale social well-being. Mehl et al. argued that more substantive conversations and less small talks are associated with increased happiness of an individual [5]. Also, social connections contribute to improved resilience to stress and increased life satisfaction and happiness [2].

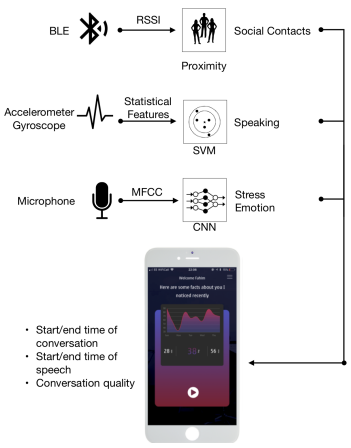

Figure 1: Cross-modal approach.

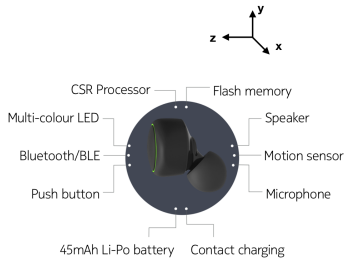

Figure 2: eSense: earbud sensing platform, offering three sensing modalities - audio, motion, and proximity - derived from microphone, accelerometer and gyroscope and BLE, respectively.

Capturing sophisticated social, emotional contexts from conversations has been actively studied in mobile computing, but most work relies on the microphone on smartphones [3, 7]. However, this has two practical limitations. First, these devices have low reliability to monitor conversational interaction, owing to their placement, e.g, the smartphones in a pocket or bag, and hence the quality of the recorded speech signal may be poor in some situations. Second, they require the continuous audio recording during a conversation, which is energy-heavy and privacy-invasive.

Earables present a unique opportunity to address these two challenges. Due to their unique placement on the ear, they have low placement variability and thus a capability of capturing human voice more clearly. Our previous work showed that earbuds show higher SNR of audio data than smartphones and smartwatches in various human contexts [6]. Moreover, they can identify the speaking activity, i.e., whether a user is speaking or not, with the cheap IMU sensor by detecting the movement of the mouth and jaw.

We propose a *cross-modal* approach for conversational well-being monitoring with a *multi-sensory earable* (See Figure 1.) It consists of three sensing models, BLE, motion, and audio models, each of which detects a conversation group, speaking activities, and stress and emotion using BLE advertisements, the IMU sensor, and the microphone, respectively. Our approach saves energy significantly by avoiding continuous audio processing. Also, more importantly, it enables speaker-specific quantification of emotion and stress as the motion model can identify a speaker accurately, i.e., who is speaking and who is not during face-to-face conversation. With the cross-modal model, we develop a mobile application that qualifies on-going conversations and provides personalised feedback on social well-being.

Cross-modal Conversation Monitoring

We used a customised earbud platform, eSense, presented in our previous work [1] (See Figure 2.) It is an aesthetically pleasing, and ergonomically comfortable in-ear high definition wireless stereo wearable instrumented with a microphone, a 6-axis inertial measurement unit and dual mode Bluetooth and Bluetooth Low Energy (BLE) in an open architecture. Each earbud has a dimension of $18mm \times 18mm \times 20mm$ and a 45 mAh battery.

One of the biggest hurdles is the limited battery capacity due to the tiny form factor. eSense has a 45 mAh battery. Apple AirPods and Google Pixel Buds have a 93 mAh and a 120 mAh battery, respectively due to a bigger size. However, the battery on earables is still precious. To address this, we devise an operational flow to run a cross-modal approach in an energy efficient way (see Figure 3). We first leverage the BLE to detect the potential conversation group and participants. Then, we use the IMU on the earable to detect conversation cues (a user's speaking moments) from head and mouth movements. This is followed by capturing a user's speech and its affect using a microphone.

BLE Model for Group Detection

The eSense periodically announces its presence though BLE advertisements which could be received by any device performing a scan operation in the vicinity. Our current model runs on the accompanying phone and performs BLE scanning to discover a potential conversation group and participants. By scanning, it obtains a list of MAC addresses and RSSI values of a user's acquaintances (earbuds). Then, it infers the conversation group using the interaction model proposed in [4]. While its logic is simple, its implementation is not straightforward since the BLE devices are no longer discoverable once they are connected to the other device. We modified the eSense firmware to keep

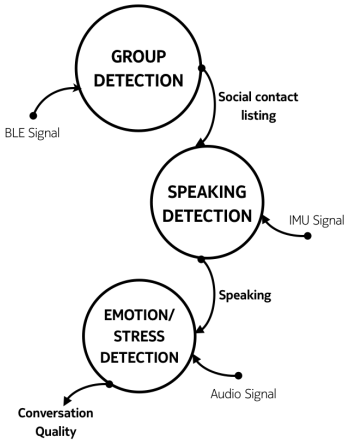

Figure 3: Energy-efficient sensing flow.

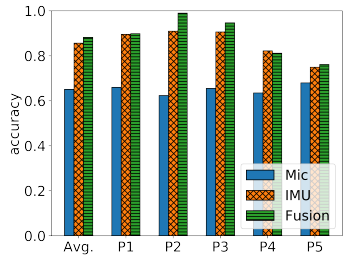

Figure 4: Accuracy for the three models. We report the performance of each model trained with leave-one-user-out method and the average accuracy of the resulting 5 models. P1 - P5 represent the participant's data used to test the model trained on the remaining four participants.

periodically broadcasting *non-connectable* advertisement packets even while being connected. While being disconnected, it periodically broadcasts *connectable* advertisement packets, as other BLE devices do.

Motion model for Speech Activity Detection
We developed a motion model for speaking activity detection, i.e., whether a user is speaking or not. Its key intuition is to capture unique signal patterns on accelerometer and gyroscope which are made by movements of the mouth and jaw while speaking. To explore the feasibility and understand its underlying principle, we used well-established features widely used for inertial sensing. The motion model reads 3-axis accelerometer and 3-axis gyroscope data from eSense at 30 Hz and the sensor streams are segmented into 5-second-long windows with 90% overlap. The model uses a set of time-domain and frequency-domain features (30 features for the accelerometer and 30 features for the gyroscope) and then uses PCA to reduce the dimensionality. 10 features are finally fed into a support vector machine (SVM) to classify the speaking activity.

Audio Model for Conversation Qualification
Upon detecting a potential conversation using IMU sensors, we use the microphone to record and qualify the conversational speech with a subjective label. Currently, we focus on two types of speech qualifications, namely *stress* and *emotion*. Our stress and emotion detection models are based on prior works[3, 7] which show that detecting stress and emotion is feasible using acoustic features such as MFCCs, prosodic features and descriptors of speaking rate and pitch. To this end, we record and aggregate the speech content during a conversation at 16 kHz, (only while a user is speaking) and extract relevant task-specific acoustic features from it, which are then used to train GMM-based models for detecting stress (binary classification) and emotion classes (anger, fear, neutral, sadness, happiness).

Preliminary Evaluation

We provide an initial understanding on the capabilities of eSense in detecting speaking activities using inertial signals. The evaluation of other components (group detection and conversation qualification) can be found in the respective papers [3, 4, 7]. We leave the evaluation of the end-to-end model and its deployment study for future work.

Accuracy of Motion Model for Speaking Activity Detection
Experimental setup: we recruited 5 participants and collected the data under 4 scenarios for each participant; speaking (4 minutes), sitting (2 minutes), watching TV (2 minutes), and nearby person's speaking (2 minutes). In all scenarios, the participants did not show significant head movement. For the evaluation, we conducted 10-fold cross-validation and leave-one-user-out. Due to space limitations, we report only the results for leave-one-user-out. To compare its performance, we further developed two methods, *microphone-based detection* and *fusion-based detection*.

Microphone-based detection: the model reads the audio stream at 16 kHz and segments it into 5-second-long windows. Then, it further segments the window into 25-ms-long frames with 15 ms overlap, and extracts 24 MFCCs. It aggregates all features from 498 frames and applies PCA to reduce the dimensionality to 10 features.

Fusion-based detection: here, we applied early-fusion: we concatenate all features from audio (30 features after applying PCA to reduce the dimensionality) and IMU sensors (30 features for accelerometer and 30 for gyroscope) into a 1D-array, we then apply PCA to reduce the dimensionality to 10 and then employ a single SVM classifier.

Results and implications: Figure 4 shows the monitoring

Configuration	Avg. Power (mW)	Battery Life (hours)
No sampling	24.27	18.75
IMU @ 50Hz	42.92	3.65
Mic @ 16kHz	50.33	2.36

Table 1: Power consumed by eSense in various configurations. For all cases, eSense is connected to a phone and transmits BLE advertisement packets with an interval between 625 and 750ms. The model we use to approximate the battery life is $Battery\ Life = Capacity/Avg.\ Current * 0.7$ [a]

[a]The factor 0.7 takes into consideration external factors which can affect battery life, e.g. the battery cannot be completely discharged, the capacity reported is not perfect.

Figure 5: Application prototype.

accuracy of speaking activities. The results show that our motion model (86%) outperforms the mic-based method (65%). This is mainly because the mic-based model is vulnerable to the ambient noise, so it sometimes incorrectly identifies a nearby person's speaking as a user's speaking activity. Surprisingly, the motion model shows comparable performance to the fusion-based detection (88%), while it is much more energy-efficient than the fusion method.

Power Consumption

We profiled the power consumption of eSense using a Monsoon Power Monitor. Table 1 shows that sampling the microphone consumes on average more power than sampling the IMU sensor and reduced the expected battery life of more than 1 hour. This supports our proposed approach where least expensive sensors are used to trigger more expensive ones only when needed.

Outlook

We prototyped a mobile app that provides personalised feedback on conversational wellbeing (Figure 5). For each conversation session, it provides details including nearby people, speech sessions, and emotion/stress. We also provide human-understandable information: *affect score*. We map the labels of emotion and stress to the specific score, e.g., 1 for stress and 0 for non-stress, and simply compute the affect score as a weighted sum of emotion and stress.

Wearing earbuds during a conversation may not be preferred as earbuds are usually worn in a way that blocks the ear and ambient sound as well. Also, it could be considered socially inappropriate by conversation partners. Thus, we assume the situation where a user wears an earbud in one ear and keeps the other ear open. This is reasonable considering how people recently use real-time translation on earbuds. Also, tiny earables designed for hearing aid are already prevalent and worn in everyday life.

We proposed a *cross-modal* approach for conversational wellbeing monitoring with a *multi-sensory earable*. We presented the details of our implementation and showed the preliminary evaluation that shows the feasibility of energy efficient IMU sensing for speaking activity detection.

REFERENCES

1. Kawsar et al. 2018. Earables for Personal-scale Behaviour Analytics. *IEEE Pervasive Computing* 17, 3 (2018).
2. Helliwell et al. 2004. The social context of well-being. *Philosophical Transactions of the Royal Society B: Biological Sciences* 359, 1449 (2004).
3. Lu et al. 2012. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*.
4. Mashhadi et al. 2016. Understanding the impact of personal feedback on face-to-face interactions in the workplace. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*.
5. Mehl et al. 2010. Eavesdropping on happiness: Well-being is related to having less small talk and more substantive conversations. *Psychological science* 21, 4 (2010), 539–541.
6. Min et al. 2018. Exploring Audio and Kinetic Sensing on Earable Devices. In *Proceedings of the 2018 Workshop on Wearable Systems and Applications*.
7. Rachuri et al. 2010. EmotionSense: a mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*.