Situation-Aware Emotion Regulation of Conversational Agents with Kinetic Earables

Katayama, Shin; Mathur, Akhil; Van Den Broeck, Marc; Okoshi, Tadashi; Nakazawa, Jin; Kawsar, Fahim

DOI
10.1109/ACII.2019.8925449

Publication date
2019

Document Version
Final published version

Published in
2019 8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019

Situation-Aware Emotion Regulation of Conversational Agents with Kinetic Earables

Shin Katayama*, Akhil Mathur†, Marc Van den Broeck†, Tadashi Okoshi*, Jin Nakazawa*, Fahim Kawsar†‡
*Keio University, Kanagawa, Japan, †Nokia Bell Labs, Cambridge, UK, ‡TU Delft, Netherlands
*{shinsan, slash, jin}@ht.sfc.keio.ac.jp, †{akhil.mathur, marc.van_den_broeck, fahim.kawsar}@nokia-bell-labs

***Abstract*—Conversational agents are increasingly becoming digital partners of our everyday computing experiences offering a variety of purposeful information and utility services. Although rich on competency, these agents are entirely oblivious to their users' situational and emotional context today and incapable of adjusting their interaction style and tone contextually. To this end, we present a mixed-method study that informs the design of a situation- and emotion-aware conversational agent for kinetic earables. We surveyed 280 users, and qualitatively interviewed 12 users to understand their expectation from a conversational agent in adapting the interaction style. Grounded on our findings, we develop a first-of-its-kind emotion regulator for a conversational agent on kinetic earable that dynamically adjusts its conversation style, tone, volume in response to users emotional, environmental, social and activity context gathered through speech prosody, motion signals and ambient sound. We describe these context models, the end-to-end system including a purpose-built kinetic earable and their real-world assessment. The experimental results demonstrate that our regulation mechanism invariably elicits better and affective user experience in comparison to baseline conditions in different real-world settings.**

***Index Terms*—Conversational Agent, Context Awareness, Emotion Regulation, Earables**

I. Introduction

Conversational agents are now pervasive. Remarkable advancement of machine learning is causing a seismic shift, in that conversational agents are now able to recognise and understand human speech and transform text into speech in a similar way to humans [1]. Naturally, this created interminable possibilities, uncovering novel, productive and useful experiences with conversational agents for accessing and interacting with digital services in many and diverse applications including HCI [2], customer experience [3], conversational commerce [4], medicine [5], entertainment [6] and education [7].

For long affective computing research has focused on bringing emotional awareness to these agents, i.e., by understanding human emotion using machine learning, and more recently representation deep learning techniques [8]. However, unfortunately, the implications of this research in our everyday experience is still limited. For instance, none of the commercial-grade conversational agents (Alexa, Siri, Google, Cortana, etc.) today can react and adapt to users' emotional and situational context. We argue that simple adjustments of the interaction style of the agents' responses can increase users' conversational experience with these agents. Imagine a supportive and discreet response in a slow and mild tone from an agent when a user is upset at work, or a joyous response in an empathic tone to celebrate a user's happy moment.

Although a rich body of literature has looked at understanding user's emotion, and generating an affective voice for agents, there is a striking gap in connecting these two facets. This gap is further intensified considering these agents are entirely oblivious to users' social and situational context. To this end, in this research, we take a user-centred view to design a situation- and emotion-aware conversational agent. We first quantitatively survey 280 users, and interviewed 12 users to understand their expectation concerning an agent's conversation style in a variety of real-life situations (e.g., at home, at work, alone or in a group) and across different emotional contexts (e.g., happy, upset). Grounded on the findings we then move into the development of a situation-aware conversational agent on a kinetic earable for personal-scale conversational experience. Our system operates on two key principle components - i) a *context builder* that uses audio and motion signals from the earable to reliably understand users' emotional, social, environment and activity situations and ii) an *affective adapter* that applies a set of learned rules based on user's context to adapt the affect of agent's response. Early experimental evaluation with 12 users suggests our situation-aware adaptation can increase usability and elicit superior user experience in comparison to baseline conditions (i.e. no adjustment, random adjustment).

In what follows, we first revisit some related past research and then present the overall methodology of this research. We then describe the contextual studies that inform the design of the proposed system. Next, we provide an in-depth technical view of our solution including constituent models, software and hardware artefacts. We then present the evaluation of the system and reflect on a few intriguing issues before concluding the paper.

II. Related Work

Over the past decade, a rich body of work has looked at affect embodiment in everyday computational experiences. We reflect on some of these works that shape our work. On **Emotion Understanding**, audio signal has been extensively studied for detecting and regulating human emotions and more recently through a deep learning lens. In [9], Bertero et al. proposed a CNN model to classify human speech into eight emotions from raw speech. Latif et al. [10] proposed the use of parallel convolutional layers with an LSTM network for emotion

978-1-7281-3888-6/19/$31.00

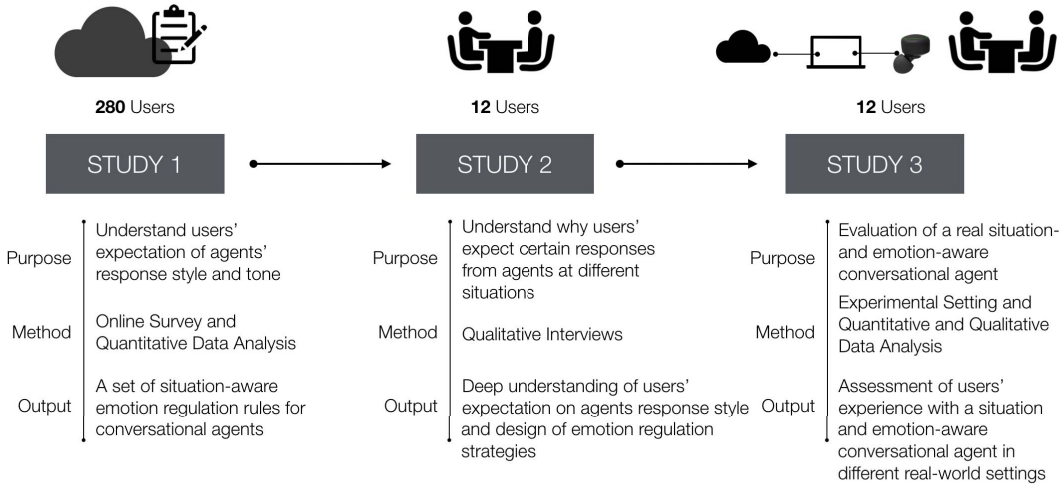

Fig. 1. Overall mixed-method methodology followed in this research including quantitative data analysis, qualitative study, real-world evaluation and qualitative interviews

recognition. On **Speech Synthesis**, a collection of work, such as Wavenet [11], Tacotron [12], and Deep Voice [13] looked at synthesising voice using reference acoustic representation for the desired prosody. Our work is built on these established research, and essentially connects these two threads of research with an adaptation strategy, i.e., transforming the response of the agent in correspondence to a user's emotion and situational context. On this latter **Context Understanding**, audio has as long been used as an honest signal carrying meaningful information to explain surrounding environments. In [14] Lu et al. presented SoundSense, an audio event classification system specifically designed for mobile phones. Stork et al. [15] used Mel-frequency cepstral coefficients (MFCC) with non-Markovian ensemble voting to discriminate among 22 indoor activities. Laput et al. [16] proposed Ubicoustics, which uses state-of-the acoustic classification model trained with special effect dataset to recognise unconstrained human activities. We built on these works to develop a custom acoustic recognition pipeline that can explain users situation faithfully. On **Emotion Regulation**, Schrder et al. [17] proposed a substantial effort to build a real-time interactive multimodal dialogue system with a focus on emotional and nonverbal interaction capabilities. Dongkeon Lee et al. [18] introduced conversational mental healthcare service based on emotion recognition leveraging advanced natural language-based technique. Siddique [19] proposed an interactive system that generates an empathic response by learning users personality. All of these work influenced our research; however, we take a more general approach in designing situation specific responses.

III. Study Methodology

We begin by briefly offering an overview of the methodology followed in this research. The objective of this work is to devise an emotion regulation strategy for conversational agents in correspondence to a user's situational context. We have approached this research in three stages applying mixed-method study methodology and small-scale system deployment. The overall method is reflected in Figure 1. We begin with a quantitative data analysis of 280 users' survey responses to understand their expectation of an agent's interaction style in a variety of real-world situations. Then we performed a qualitative study with twelve users to understand these facets in depth. Grounded on the findings of these analyses, we developed a system with purpose-built components for emotion regulation in agent's response. We evaluated them in a small-scale real-world deployment with twelve users followed by qualitative interviews with them to understand their experience and impression of such system and engagement. In the rest of this paper, we describe each of these studies in depth.

IV. Contextual Study

The objective of this study is to understand users' expectation from a conversational agent in adjusting their interaction style in accordance to different real-life situations. To this end, at this phase of our research we conducted an online survey and a contextual study to qualify these aspects. In particular, we defined the following context dimensions[1] and emotion states to assess how an agent should respond to these settings.

- **Location:** This context includes two attributes, *Home and Public* settings.
- **Sociality:** This context contains two attributes, *Alone and Group*.
- **Activity:** This context covers two possible activity situations *Walking and Driving*.
- **Emotion:** Four emotional states are considered including *Neutral, Upset, Happy, Angry*.

[1]Please note that we do not claim these dimensions and their constituent attributes are complete, however, this offers a first-hand view of the implicit effect of contextualising conversational agents.

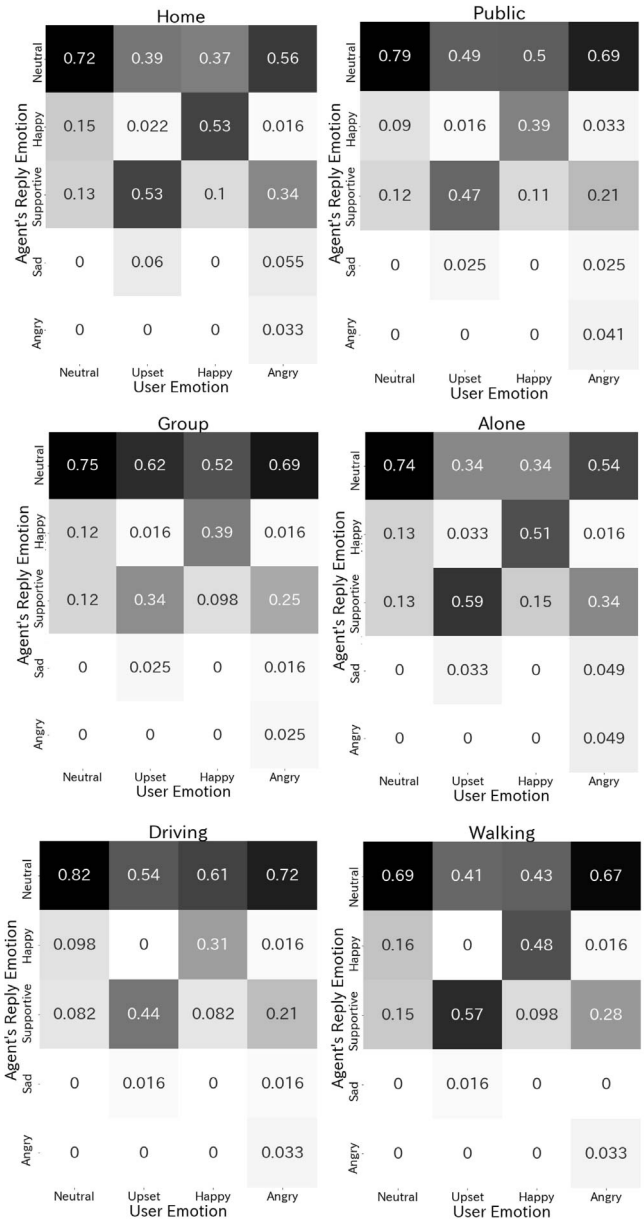

Fig. 2. Heat maps illustrating participants' expectations of the response style (emotion) of a conversational agent in a variety of contextual settings.

We have designed a set of survey questions ($n = 24$), mainly, combining each context attribute with an emotional state and ask the participants to express their desired interaction style of the agents. An example survey question is: *When you are at **home** and feeling **upset**, how would you like a conversational agent to respond to your questions?*. The possible responses to each survey question are *Neutral, Supportive, Happy, Sad, Angry*, which reflect the different ways a conversational agent can regulate the affect of its voice.

Survey: We gathered 280 responses (6,720 answers) using Amazon Mechanical Turk. We followed guidelines from established literature, e.g., interaction time, lockout to control the quality of the responses [20], [21].

Qualitative Interviews: To further understand and qualitatively assess the survey responses, we conducted semi-structured interviews with 12 participants (3 female, recruited through snowball sampling), all of whom are active users of conversational agents. We asked the same 24 questions as the survey, following an interview technique called laddering[2] to uncover the core reasons behind users reactions. We analysed the interview data by coding the individual responses using affinity diagramming to derive final rules.

Results and Design Implications: Figure 2 shows participants' expectations of the response style, i.e., emotional reply of a conversational agent in a variety of contextual settings. We observe that participants' expect different response style in different situations. For example, when the user feels happy at home, 53% of the participants want the agent to reply in a happy tone, but when the user feels happy at a public place, 50% people want the agent to reply in a neutral tone. Although, we have observed the desire for such adjustments, in the majority of the situations, however, participants prefer a neutral tone. This is particularly the case when the participant in a public place or a group. Our qualitative interviews shed some light on this, as multiple ($n = 9$) participants remarked that they do not want the agent to show any affect in public settings. We also observed that participants prefer a supportive reply when they are upset or angry and in a private setting, e.g., at home, or alone. There are cases, for instance, when an agent helps a user during a physical workout, or other coaching sessions; participants ($n = 8$) desired a mirrored reply, i.e., if a user is angry, the response can be in an angry tone too. Based on these insights we put forth three design goals for our solution:

1) The system should recognise situational context leveraging sensory perceptions.
2) Based on the situation, the system should determine the target emotion of the agent's response. Through our survey and interviews, we have developed a set of adaptation rules that determine the target emotion. In cases where multiple target emotions are possible, the system takes activity as the critical context to drive the interaction. These strategies are illustrated in Table 1.
3) Besides prosody transformation to reflect emotion, the solution should also be able to change the speed and volume of the response in correspondence to the situation.

In the next section, we describe our situation-aware conversational agents built on these principles.

V. System

In this section, we present a first-of-its-kind situation-aware conversational agent that dynamically adjusts its conversation style, tone, volume in response to users emotional, environmental, social and activity context gathered through speech prosody, ambient sound and motion signatures. We use an off-the-shelf ear-worn device named eSense [22] as the source of

[2]http://www.uxmatters.com/mt/archives/2009/07/laddering-a-research-interview-technique-for-uncovering-core-values.php

TABLE I
RULES DERIVED FROM THE SURVEY AND INTERVIEWS FOR EMOTION REGULATION BASED ON SITUATIONAL CONTEXT.

User Emotion	User Context	Reply Emotion
Neutral	Home, Public, Group, Alone, Driving, Walking	Neutral
Upset	Public, Group, Driving	Neutral
	Home, Alone, Walking	Supportive
Happy	Public, Group, Driving	Neutral
	Home, Alone,Walking	Happy
Angry	Public, Group	Neutral
	Alone, Driving, Walking, Home	Supportive

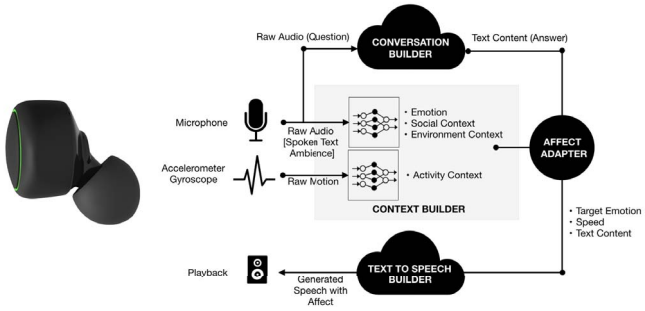

Fig. 3. The end-to-end architecture of the situation-aware conversational agent with a kinetic earable that adjusts its interaction style contextually.

sensory signals for our system. The kinetic earable, eSense, is an in-ear high definition wireless stereo wearable instrumented with a microphone, a 6-axis inertial measurement unit (IMU) and dual model Bluetooth and Bluetooth Low Energy (BLE). These embodiments collectively enable eSense to offer three sensing modalities - audio, motion, and proximity that we have used for our sensory inferences.

Our system, built on top of eSense, is shown in Figure 3 and is composed of the following components:

Context Builder. This component consists of a number of sensing modules which infer a user's momentary context by analyzing the various modalities obtained from eSense earable. More specifically, we use motion and audio as the two main sensing modalities to detect four types of user contexts, namely physical activity, emotional state, social context, and environmental context.

Firstly, accelerometer and gyroscope data obtained from the IMU is used to infer the *physical context* of the users – for this work, we restrict the detection of physical context to two classes: walking and driving, in line with our survey and qualitative studies. We use a state-of-the-art deep neural network architecture proposed by [23] to build the physical activity models. This architecture consists of a CNN-based feature extractor with 4 residual blocks containing 2 convolutional layers each. This is followed by two fully-connected layers respectively with 1024 and 128 units, and then with a output layer of 4 units corresponding to our locomotion target classes. The model is trained on two publicly available HAR datasets, namely the *Opportunity dataset* [24] and *RealWorld HAR dataset* [25].

To infer the *emotion state* of the user, we employ a speech processing model which analyses the prosody of the speech to infer one of the four target emotion classes as discussed in section IV. We use a CNN-based architecture comprising of 2 convolutional layers and 1 fully-connected layer proposed in [26] to build our emotion detection model. The model is trained on a publicly available speech emotion dataset called RAVDESS [27] which is a collection of 1,440 English-language speech segments spoken by actors while expressing a range of emotions.

Finally, to infer the *social and environmental context* of the user, we use the pre-trained acoustic environment detection model provided by the authors of Ubicoustics [16] to classify a given audio segment into two environment contexts: *Home and Public*, and two social contexts: *Alone and Group*. Note that the Ubicoustics model performs acoustic classification at a finer granularity and we semantically group the output classes of Ubicoustics to generate classes of interest for our work.

Conversation Builder. This is a question-and-answer component which enables a user to interact with the agent using a predefined dialogue base. In our current implementation, we have used DialogFlow [28] populated with a set of situation-specific questions that the users can ask the agent (e.g., how is the weather today in New York). DialogFlow receives a speech segment recorded from eSense as input and answers it with a textual response.

Affect Adapter. This component is responsible for guiding the adaptation strategy for the agent's response corresponding to the user's context, taking into account the output of the Context Builder and a data-driven rule engine. More specifically, this component takes the current context and current emotion of the user as input (from the Context Builder) and outputs a target emotion for agent's response based on the rules learned from our survey (as shown in Table I). For example, if the Emotion Detection model outputs the user emotion as *Upset* and the Environment Context models infer the user's location as *Home*, then the Affect Adapter would output the target emotion of agent's response as *Supportive*. Although our current affect adaptation is based on the rules extracted from the survey responses, in future work it can be extended to a learning-based approach where the rules are automatically learned over time using machine learning techniques.

Text-to-Speech Builder. Once the emotion of the agent is determined, the final step in the pipeline is to play the answer to user's query (from DialogFlow) with the correct emotion adaptation. To this end, we use IBM Cloud Voice service [3] to synthesise the agent's response in a way that accurately reflects the emotion adaptation. IBM Voice Service provides a Speech Synthesis Markup Language (SSML) to adapt various parameters of the speech such as pitch, pitch range, rate, glottal tension, breathiness, when performing a Text-to-Speech synthesis. Based on established literature in the

[3]https://cloud.ibm.com/docs/services/text-to-speech

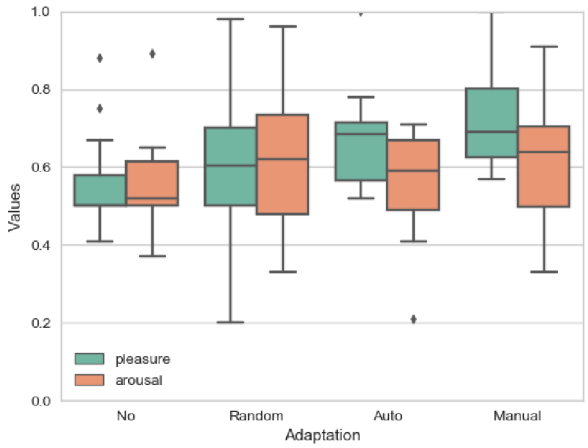

Fig. 4. Pleasure and Arousal scores using the Affective Slider

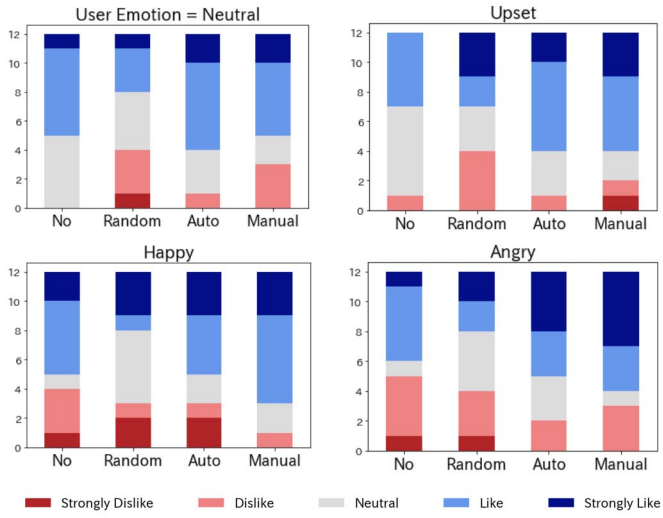

Fig. 5. Subjective Likert-scale preferences of the users towards different adaptation strategies.

speech synthesis community [29], we chose appropriate values for each of these parameters to simulate various emotions in the agent's response. For example, to synthesize 'Angry' speech, we use {pitch=50%, pitch_range=80%, rate=40%, glottal_tension=100%, breathiness=60%}.

VI. Evaluation

In this section, we report the evaluation of our solution containing three parts. First we mention the performance of our different system components, then we discuss the real-world experiment of the system with twelve participants, and finally, we briefly mention the qualitative feedback we have received from these participants concerning the usability of the system during the experiment.

A. Performance of System Components

First we evaluate the performance of the different sensing models in our system. To this end, we ran the various sensory models on a laptop (Apple MacBook Pro, 3 GHz Intel Core i7, 16 GB 1600 MHz DDR3) and evaluated their performance with 10-fold cross validation. For the physical context and emotion state detection, our trained models yielded a F_1-score of 0.97 and 0.91 respectively. Similarly, the pre-trained UbiAcoustics models had a test F_1-score of 0.906 on the Environment Context detection task, and 0.943 on the Social Context detection task. As for the latency of our system, the average latency for the conversation builder was 1.28 seconds, the text to speech was 1.2 seconds, and the end-to-end delay was for 2.48 seconds.

B. Real-World Experimental Evaluation

In this phase of our evaluation, we recruited twelve participants through snowball sampling. Each of these participants took part in two sessions.

Methodology: In the first session, participants were exposed to four experimental conditions of adaptation while interacting with a conversational agent in a variety of real-life situations in a simulated environment. These experimental conditions were: **No Adaptation, Random Adaptation, Auto Adaptation, Manual Adaptation**. No adaptation represents the state-of-the-art baseline where an agent reply in a neutral voice in all situations. In random adaptation condition, agents pick randomly one of the five emotions (*Neutral, Happy, Supportive, Angry, Sad*) while replying. Auto Adaptation responds using our system of situation-aware adaptation, and finally, In Manual Adaptation, participants can manually select from five emotions for agent's reply.

We used a role-playing methodology for the user study, wherein we asked the participants to imaging that they are in a certain situation (e.g., at home, at a public cafe, alone, in a group). In each context, we asked the participants to interact with the agent in four different emotions (*Neutral, Upset, Happy, Angry*, which were simulated using pre-populated questions. For example, to simulate the Angry emotion, the users could ask the agent *"Why can you not understand my voice? It's just English."*. The agent would then respond in one of the four ways of adaptation as discussed previously.

In total, each participant had 16 interactions with the agent; the order of these conditions was counter-balanced using Latin square. After each experimental condition, participants were requested to finish a questionnaire that includes an Affective Slider [30] to assess the overall user experience of the condition. Affective Slider is a digital self-reporting tool composed of two slider controls for the quick assessment of Pleasure and Arousal. After the Affective Slider, participants evaluated the condition of adaptation with 5-point Likert scale questions. Example questions include *How did you find the agent's response to your angry voice at home?* Finally, each participant was interviewed to qualitatively assess their experience with the system.

Results. Figure 4 illustrates the pleasure and arousal of each scenario in the numerical value of the Affective Slider. A pair-wise comparison using Wilcoxon rank sum tests showed significant differences in pleasure between No Adaptation

and Auto Adaptation (p-value: 0.02), No Adaptation and Manual Adaptation (p-value: 0.003), and Random Adaptation and Manual Adaptation (p-value: 0.04). In arousal, between Auto Adaptation and Manual Adaptation (p-value: 0.02) was only significantly different. These results suggest that Auto Adaptation of the agent's interaction style depending on the user's situation and emotion is better in terms of pleasure than No Adaptation, and showed the superiority of our system.

Figure 5 shows the participants subjective experience across different situations gathered through Likert scale.

Converting the Likert scale values of 1 and 2 into negative values and 4 and 5 into positive values, the positive rate for the Angry emotion increased to 58% in Auto Adaptation compared to baseline conditions(No Adaptation:50%, Random Adaptation:33%), and the negative rate decreased to 16% in Auto Adaptation compared to baseline conditions (No:41%, Random:33%). In other words, when the users were Angry, they preferred our proposed Auto Adaptation scenario from a user-experience perspective over the baseline scenarios. Similarly, in the Upset scenario, the positive rate rose to 66% in Auto Adaptation compared to baseline conditions(No:41%, Random:41%). While the Auto Adaptation setting was preferred over No and Random Adaptation, we also observe that Manual Adaptation baseline was the most preferred option in non-neutral scenarios.

This has two implications: a) adaptation of agents in general is preferred by the users, b) automatic adaptation strategies like ours still need to be improved to reach the level of personalisation desired by the users (as reflected in manual adaptation). In future work, we will explore ways of personalising our adaptation rules to each user in order to improve the overall user experience.

Our qualitative interviews with the participants also throw some light on these results. Essentially, all the participants evaluated the adaptation positively and found subtle changes in the agent's interaction style pleasurable and natural. One remark was (P3), *At the office, I expect nothing but professionalism. There, I need the voice assistant to behave, at home or in a cafe, it can be as crazy as me. So I welcome these interaction styles, its refreshing than the same monotonous voice all the time...*

Another interesting comments was (P7): *When I'm at home and sad, I want to stay in that state, but when I'm in a restaurant surrounded by people, maybe the machine can bring me back to more happier levels. So, I like such adjustments.*

We have received similar comments from other participants which we consider a further validation of our approach. However, these interviews also revealed some interesting dynamics concerning such situation- and emotion-aware conversational agents. We discuss these issues in the next section.

VII. Outlook

In this section we reflect on a few insights from our study that we found the most compelling.

Utility and Companionship: Conversational agents are now pervasive and capable of automating various facets of our life. Although, in this work we have demonstrated an emotion-regulation strategy of these agents, our study identified one interesting dynamics towards such regulation. Essentially, emotion regulation adds minimal value to agents that are designed as utility providers (e.g., home automator, news reader). However, agents that aim to become digital companions, e.g., guiding different decision making with just-in-time information or offering support in stressful situations require effective emotion regulation. As such, our work calls attention to designers of conversational agents to understand the purpose and role of the agents before considering regulation strategies.

Culture: Another important aspect is the impact of demography on the ultimate conversational experience. In our study, we have observed a striking difference in people's expectation of agents interaction style across generations, and culture. For example, many of the Asian participants preferred the agent to mirror their emotions, whereas Western participants opted for more subtle regulation. Besides, during our qualitative studies, participants (mostly western origin) suggested they do not want the agent to become humane in its response, given that it is a machine after all. These findings suggest that future agents need to consider both personality [19] and the cultural origin of the user to offer the most personal experience.

Privacy: Not surprisingly, privacy has been a constant concern for our participants. While they recognise the benefit of the contextualisation, they were concerned that their emotional state wouldn't remain private any longer with such technology. These are legitimate concerns and demands the context processing to be done entirely locally in a privacy-preserving way. Research in embedded machine learning space are addressing local compute issues, and we are hopeful that soon we will be able to run such agents in a local device preserving user's privacy.

Limitations: This study was conducted in a relatively small and constrained environment. Indeed, the results presented here must be interpreted in the context of the experimental setting, and must not be generalised without further longitudinal studies. Besides, our selection of context dimensions are certainly limited, and more fine-grained context can be further included for more accurate situational awareness. We acknowledge these limitations and consider these as future avenues of this research.

To conclude, we reported the design and development of a situation-aware conversational agent purposefully built for a kinetic earable. The design was grounded on a mixed-method study that informed the regulation strategy of the agent. We describe the study, technical details of the system, and a small-scale evaluation. We consider, our research uncovers exciting opportunities for building next-generation conversational agents transforming them truly into our digital partners.

Acknowledgement

This work was conducted at Nokia Bell Labs, Cambridge through a research collaboration with Keio University supported (in part) by National Institute of Information and Communications Technology (NICT), Japan

References

[1] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural Responding Machine for Short-Text Conversation. *arXiv:1503.02364 [cs]*, March 2015. arXiv: 1503.02364.

[2] Paul Luff, David Frohlich, and Nigel G Gilbert. *Computers and conversation*. Elsevier, 2014.

[3] Salvatore Parise, Patricia J Guinan, and Ron Kafka. Solving the crisis of immediacy: How digital technology can transform the customer experience. *Business Horizons*, 59(4):411–420, 2016.

[4] Nishant Piyush, Tanupriya Choudhury, and Praveen Kumar. Conversational commerce a new era of e-business. In *2016 International Conference System Modeling & Advancement in Research Trends (SMART)*, pages 322–327. IEEE, 2016.

[5] Myrthe L. Tielman, Mark A. Neerincx, Rafael Bidarra, Ben Kybartas, and Willem-Paul Brinkman. A Therapy System for Post-Traumatic Stress Disorder Using a Virtual Agent and Virtual Storytelling to Reconstruct Traumatic Memories. *Journal of Medical Systems*, 41(8):125, August 2017.

[6] Michael J Kuhn. Virtual game assistant based on artificial intelligence, December 1 2015. US Patent 9,202,171.

[7] Lindsay C Page and Hunter Gehlbach. How an Artificially Intelligent Virtual Assistant Helps Students Navigate the Road to College. page 12.

[8] Kun Han, Dong Yu, and Ivan Tashev. Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. page 5.

[9] Dario Bertero, Farhad Bin Siddique, Chien-Sheng Wu, Yan Wan, Ricky Ho Yin Chan, and Pascale Fung. Real-Time Speech Emotion and Sentiment Recognition for Interactive Dialogue Systems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1042–1047, Austin, Texas, 2016. Association for Computational Linguistics.

[10] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, and Julien Epps. Direct Modelling of Speech Emotion from Raw Speech. *arXiv:1904.03833 [cs, eess]*, April 2019. arXiv: 1904.03833.

[11] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. *arXiv:1609.03499 [cs]*, September 2016. arXiv: 1609.03499.

[12] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards End-to-End Speech Synthesis. *arXiv:1703.10135 [cs]*, March 2017. arXiv: 1703.10135.

[13] Sercan O. Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, and Mohammad Shoeybi. Deep Voice: Real-time Neural Text-to-Speech. *arXiv:1702.07825 [cs]*, February 2017. arXiv: 1702.07825.

[14] Hong Lu, Wei Pan, Nicholas D. Lane, Tanzeem Choudhury, and Andrew T. Campbell. SoundSense: scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the 7th international conference on Mobile systems, applications, and services - Mobisys '09*, page 165, Wroclaw, Poland, 2009. ACM Press.

[15] Johannes A. Stork, Luciano Spinello, Jens Silva, and Kai O. Arras. Audio-based human activity recognition using Non-Markovian Ensemble Voting. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 509–514, Paris, France, September 2012. IEEE.

[16] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. Ubicoustics: Plug-and-Play Acoustic Activity Recognition. In *The 31st Annual ACM Symposium on User Interface Software and Technology - UIST '18*, pages 213–224, Berlin, Germany, 2018. ACM Press.

[17] Marc Schroder, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark Ter Maat, Gary McKeown, Sathish Pammi, Maja Pantic, et al. Building autonomous sensitive artificial listeners. *IEEE transactions on affective computing*, 3(2):165–183, 2011.

[18] Dongkeon Lee, Kyo-Joong Oh, and Ho-Jin Choi. The chatbot feels you - a counseling service using emotional response generation. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 437–440, Jeju Island, South Korea, February 2017. IEEE.

[19] Farhad Bin Siddique, Onno Kampman, Yang Yang, Anik Dey, and Pascale Fung. Zara Returns: Improved Personality Induction and Adaptation by an Empathetic Virtual Agent. In *Proceedings of ACL 2017, System Demonstrations*, pages 121–126, Vancouver, Canada, 2017. Association for Computational Linguistics.

[20] Gabriele Paolacci. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5):9, 2010.

[21] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1):3–5, January 2011.

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

[23] Liangying Peng, Ling Chen, Zhenan Ye, and Yi Zhang. Aroma: A deep multi-task learning based simple and complex human activity recognition method using wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):74, 2018.

[24] Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, et al. Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh international conference on networked sensing systems (INSS)*, pages 233–240. IEEE, 2010.

[25] Timo Sztyler and Heiner Stuckenschmidt. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–9. IEEE, 2016.

[26] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *Platform Technology and Service (PlatCon), 2017 International Conference on*, pages 1–5. IEEE, 2017.

[27] Steven R. Livingstone and Frank A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, 13(5):e0196391, May 2018.

[28] Google LLC. Dialogflow, 2019. Accessed: 2019-04-23.

[29] Marc Schröder. Emotional speech synthesis: A review. In *Seventh European Conference on Speech Communication and Technology*, 2001.

[30] Alberto Betella and Paul F. M. J. Verschure. The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions. *PLOS ONE*, 11(2):e0148037, February 2016.