

[Inner endpaper]

**HEC MONTRÉAL**

**Comparing the effectiveness of speech and physiological features in  
explaining emotional responses during voice user interface interactions**

**par**

**Danya Swoboda**

**Pierre-Majorique Léger, Ph.D et Sylvain Sénécal ,Ph.D**

**HEC Montréal**

**Directeurs de recherche**

**Sciences de la gestion**

**(Spécialisation Expérience utilisateur en contexte d'affaires)**

*Mémoire présenté en vue de l'obtention  
du grade de maîtrise ès sciences en gestion  
(M. Sc.)*

Novembre 2021

© Danya Swoboda, 2020

## Résumé

L'effet de l'essor rapide des technologies d'interfaces vocales est ressenti par les utilisateurs autant que les professionnels de l'expérience utilisateur (UX). En effet, les tâches typiques nécessitant une attention gestuelle ou visuelle sont remplacées par des commandes vocales. Par conséquent, ce changement déstabilise le UX traditionnel, puisque les méthodes utilisées dans un contexte d'interface numérique ne sont pas toujours adaptables aux interfaces vocales.

Ceci étant dit, il est important de mieux comprendre les méthodes optimales à l'évaluation des interfaces vocales. Ce mémoire, présentée sous la forme d'un article, vise à comparer l'efficacité des mesures physiologiques et les mesures spectrales de la voix afin de mieux cerner les émotions ressenties par les utilisateurs lors des interactions vocales.

Pour ce faire, nous avons effectué une expérience intra-sujet dans laquelle les données de la voix, de l'expression faciale et de l'activité électrodermale de 16 participants ont été enregistrées lors des interactions avec une interface vocale simulée. Cette expérience a été délibérément conçue pour déclencher la frustration et le choc chez le sujet d'étude. Un total de 188 interactions ont été analysées.

Nos résultats suggèrent que la mesure physiologique de l'expression faciale est la plus explicative des événements émotionnels vécus lors des interactions vocales. En effet, les relations entre la mesure de la valence issue de l'analyse automatique des émotions faciales et les dimensions émotionnelles d'intérêts sont significativement plus fortes en comparaison avec ceux partagés avec les mesures spectrales de la voix.

Compte tenu de la nature des interfaces vocales, la mesure des dimensions émotionnelles de la voix peut sembler comme un choix évident dans l'évaluation des interfaces vocales. Cependant, les résultats de cette étude suggèrent une approche différente et offrent par conséquent des informations pertinentes pour les professionnels du UX. Comprendre l'efficacité de chacune des mesures implicites étudiées favorise une évaluation d'interface vocale efficace, car seul les mesures les mieux adaptés seront retenues.

**Mots clés :** Interface vocale, mesures implicites, valence émotionnelle, activation émotionnelle, expérience utilisateur

## **Abstract**

The rapid rise of voice user interface technology has changed the way users traditionally interact with interfaces, as tasks requiring gestural or visual attention are swapped by vocal commands. Consequently, this shift has affected user experience (UX) professionals seeking to evaluate voice user interfaces, as certain traditional methods used in a digital interface context can be deemed inappropriate. Hence, a need to better understand effective voice user evaluation methods prevails. The following master thesis in the form of an article sought to compare the effectiveness of physiological and speech measures through their extracted features in explaining emotional events induced by voice user interface interactions.

To do so, we performed a within-subject experiment in which speech, facial expression, and electrodermal activity responses of 16 participants were recorded during voice user interface interactions that were purposely designed to elicit frustration and shock, resulting in 188 analyzed interactions.

Our results suggest that the physiological measure of facial expression is most informative of emotional events experienced during voice user interface interactions. Indeed, the relationship strength between the extracted physiological feature automatic facial expression (AFE) based valence and the observed emotional dimensions surpasses that of all eight extracted speech features.

Considering the nature of voice user interfaces, speech may be viewed as an obvious measure to assess affective states during voice user interface interactions. However, results from this study suggest a different approach, and consequently offers valuable and actionable insight to UX professionals involved in voice user interface evaluation. Understanding the effectiveness of implicit measures through their respective features in explaining affective states during voice user interface interactions has the potential to increase evaluation efficiency, as selecting the best fitted measure can limit resources.

**Keywords:** Voice user interface, implicit measures, emotional valence, emotional arousal, user experience

# Table of contents

<b>Résumé .....</b>	<b>i</b>
<b>Abstract .....</b>	<b>iii</b>
<b>Table of contents.....</b>	<b>v</b>
<b>List of tables and figures.....</b>	<b>vi</b>
<b>List of abbreviations and acronyms .....</b>	<b>viii</b>
<b>Preface .....</b>	<b>ix</b>
<b>Acknowledgements.....</b>	<b>xi</b>
<b>Chapter 1. Introduction .....</b>	<b>1</b>
<i>Abstract.....</i>	<i>10</i>
<b>Chapter 2. Literature review.....</b>	<b>15</b>
2.1 <i>Definition and role of emotion within user experience.....</i>	<i>15</i>
2.2 <i>Evaluation methods and tools in user experience .....</i>	<i>18</i>
2.3 <i>Emotion evaluation in user experience.....</i>	<i>21</i>
2.4 <i>Voice user interface evaluation .....</i>	<i>23</i>
2.5 <i>Physiological measures within the study of emotion in user experience research .....</i>	<i>25</i>
2.6 <i>Voice measures within the study of emotion in user experience research .....</i>	<i>28</i>
2.7 <i>Summary of literature review.....</i>	<i>35</i>
<i>References .....</i>	<i>36</i>
<b>Chapter 3. Comparing the effectiveness of speech and physiological features in explaining emotional responses during voice user interface interactions .....</b>	<b>50</b>
<i>Abstract.....</i>	<i>50</i>
3.1 <i>Introduction.....</i>	<i>52</i>
3.2 <i>Literature Review and Hypotheses Development.....</i>	<i>55</i>
3.3 <i>Methods.....</i>	<i>64</i>
3.4 <i>Results .....</i>	<i>77</i>
3.5 <i>Discussion .....</i>	<i>88</i>
<i>References .....</i>	<i>95</i>
<b>Chapter 4. Managerial Article: Hey Alexa, what is the best approach to detect pain points induced by voice user interface interactions? .....</b>	<b>105</b>
<i>Summary .....</i>	<i>105</i>

4.1	<i>Introduction</i>	105
4.2	<i>Comparative strenght of physiological and speech features per dimension</i>	105
	<i>Notes</i>	109
	<b>Conclusion</b>	<b>110</b>
	<b>Bibliography</b>	<b>117</b>
	<b>Appendix 1: Experimental script</b>	<b>i</b>
	<b>Appendix 2: Third-party evaluator protocol instructions</b>	<b>xii</b>
	<b>Appendix 3: Third-party evaluator Qualtrics questionnaire instructions</b>	<b>xvi</b>
	<b>Appendix 4: Third-party evaluator Qualtrics page example</b>	<b>xxi</b>



# List of tables and figures

## List of tables

### *Thesis*

Table 1: Contribution to the responsibilities of the research project phases

Table 2: Summary of HCI study examples utilizing speech features

### *Article*

Table 1: Summary of common speech features indicative of emotion

Table 2: Summary of the multi-method studies utilizing speech and physiological measures in relation to emotion recognition

Table 3: Results of the Intraclass correlation scores

Table 4: Descriptive statistics of third-party evaluations per dimension

Table 5: Regression results of valence dimension

Table 6: Regression results of arousal dimension

Table 7: Regression results of control dimension

Table 8: Regression results of STEE dimension

Table 9: Summary of hypotheses in relation to results status

## List of figures

### *Article*

Figure 1: Google slides presentation featuring dialogue files

Figure 2: Experimental Setup

Figure 3: Photographic example of electrode placements

Figure 4: Graphical summary of the experimental procedures

Figure 5.a: Mean valence score per evaluator accorded to each participant

Figure 5.b: Mean arousal score per evaluator accorded to each participant

Figure 5.c: Mean control score per evaluator accorded to each participant

Figure 5.d: Mean STEE score per evaluator accorded to each participant

Figure 6: Bar chart of relationship strengths between physiological and speech features per dimension

Figure 7: Boxplots of evaluator ratings and select physiological and speech features

## **List of abbreviations and acronyms**

AFE: Automatic Facial Expression

AS: Affective Slider

CER: Comité d'Éthique en Recherche

EDA: Electrodermal Activity

FACS: Facial Action Coding System

HCI: Human-Computer Interaction

ICC: Intraclass Correlation Coefficient

IS: Information System

IT: Information Technology

MFCC: Mel Frequency Cepstral Coefficient

No-UI: Non-visual User Interaction

SAM: Self-assessment Manikin

SCR: Skin Conductance Response

SER: Speech Emotion Recognition

UX: User Experience



## **Preface**

An authorization to write the following dissertation has been granted by the administrative direction of the Master of Science program specializing in user experience within a business context. This dissertation is written in the form of an article. The agreement of all co-authors for this article has been obtained.

In December 2020, the HEC Montréal Research Ethics Board (CER) approved the research project (Certificate #2021-4289).

The article compares the effectiveness of implicit measures, being facial expression, electrodermal activity and speech, through their respective features in explaining users' affective states experienced during emotionally charged voice user interface interactions.



*For Dido.*





## Acknowledgements

First and foremost, I would like to thank my codirectors Sylvain Sénécal and Pierre-Majorique Léger for their trust and for granting me the opportunity to work on this formative project. Their devotion to my education and success has equipped me with a unique skillset and confidence, paving the way to a bright future.

I am very grateful for Dr. Jared Boasen's patience, guidance, and invaluable advice throughout this long and challenging project. I am appreciative of his rigour and attention to detail that continuously pushed me to exceed my limits, but more so of his kindness and sincere care.

I would like to thank the Tech3Lab team, who have been exceptional throughout this project. A special thanks to Emma, David, Salima, and Audrey, who ensured a seamless operation in the midst of a pandemic. In addition, I would like to thank Shang Lin Shen for his pivotal help and insight into statistical analysis. I am also very grateful to all students who participated in this project as moderators, evaluators, and participants. Your implication brought this unique project to life.

A special thank you to my dear parents, Roman and Mary. Without their encouragement, support, and generosity, I would not have taken this important leap of faith to pursue a change in career paths while fulfilling my academic ambitions. My mother's exceptional work ethic and my father's eternal thirst for knowledge are inspirational ideals that will forever continue to shape my hardworking and curious nature.

I would like to thank my partner in crime, Justin, for his love and patience throughout this journey. His entrepreneurial spirit, strong-willed determination, unconventional ways, and genuine kindness inspires me daily. May we continue to grow and learn from each other's success.

Lastly, I would like to express my gratitude to the Prompt NSERC UX Chair for their financial support. Their generosity allowed me to pursue this important project with undivided attention.





# Introduction

## Context

The rise of voice user interface technology has been marked by its increased popularity within recent years. In 2020, 4.2 billion digital voice assistants worldwide were in use (Statista, 2021). By 2024, this number is projected to reach 8.4 billion – a number greater than the world’s population (Statista, 2021). Among the most popular voice platforms in 2020 were Amazon’s Alexa, Google Assistant and Apple’s Siri (Statista, 2020). Their ubiquitous presence has helped normalized speech commands, allowing users to perform an array of hands-free tasks in the comfort of their homes, cars, work environments, and so on. Taking Amazon’s Alexa as an example, voice commands can allow for music to be played, timers to be set, lights to be dimmed, coffee to be brewed, amongst several other actions. The mainstream usage of vocal assistants in smart home devices and smartphones have opened the way for a different kind of user experience. Primed by the success of vocal assistants, consumers “actively want their emotional experiences to be enhanced” (Wang et al., 2015, p.2).

The popularity of vocal interfaces amongst consumers has not gone unnoticed by businesses. Various industries, from finance to healthcare, have adopted voice interaction technologies and positioned them as competitive advantages. For example, in collaboration with Amazon Alexa and Google Assistant, energy company PSE&G allows users to perform various tasks, such as paying bills, scheduling service appointments, and reporting power outages using voice command (Public Service Enterprise Group Incorporated, Newark). As for the healthcare sector, vocal biomarkers based on voice analysis through artificial intelligence is increasingly being used for diagnosis, risk prediction, and remote monitoring (Fagherazzi et al., 2021). Consequently, the digital healthcare sector is predicted to become a dominant vertical in voice applications (Fagherazzi et al., 2021). Indeed, an action such as stating the weather forecast following a simple voice command does not entail the same level of complexity or risk in comparison to a medical diagnosis using voice recognition technology. Hence, as the

intricacy and importance of voice user interface technology continues to evolve, a dire need to ensure optimal experiences prevails.

Despite the increasing importance, omnipresence, and evolving technological advancements surrounding voice user interface design, users continue to face imperative issues with vocal technologies that ultimately hamper optimal user experiences. For example, voice user interfaces cannot understand or interpret the language context, which may result in errors and interpretations. According to Myers et al. (2018), there are four primary obstacles faced by users interacting with voice user interfaces. The first obstacle draws from a voice assistant's inability to recognize a user's request, which consequently results in the system's inability to act upon the user's intent. The second obstacle relates to a voice user interface's faulty speech analysis, occurring when the voice user interface "mishears" the user and matches the incorrect utterance with the incorrect intent. The third obstacle revolves around the system's failure to provide clear and valid feedback following a query, which further prevents users from fulfilling their requests. The final obstacle is the malfunctioning voice user interface system's information architecture, in which bugs prevent the system from operating optimally.

## **Research**

Obstacles can result in significant user pain points, defined within human-computer interactions (HCI) as user irritants impacting interactions with digital products (Platzer, 2018). Pain points provide key insights regarding peak emotional moments in a user's experience (Giroux-Huppé et al., 2019). There is therefore a link to be made between pain points and affective states in a voice user interface context, as a series of obstacle-prone interactions may result in intense emotional responses on behalf of users. Pain points are particularly critical during a user's first interaction with a product, as the primary encounter often defines a product's success or failure (Levy & Calacanis, 2015). Thus, identifying them early on within the design process through UX evaluations is key. Insights regarding pain points may serve as steppingstones towards opportunities of product improvement. Indeed, seldom do designers achieve perfection on their first iteration. On the contrary, the process of validating an output is of an iterative nature (Gothelf, 2013). As stated by Gothelf (2013), "when we focus on outcomes, we see the

opportunity for improvement, and we keep working on that thing until it delivers the outcomes that we set out to deliver” (2013, p.27).

Thus, the need to evaluate voice user interface systems stems from the greater need to create optimal user experiences. Research in recent years has taken an interest into better understanding the making of successful and unsuccessful interactions with voice user interfaces (Lopatovska & Oropeza, 2018; Lopatovska & Williams, 2018; Jiang et al., 2015; Purington et al., 2017; Kiseleva et al., 2016; Myers et al., 2018). However, traditional measures used to evaluate such experiences are limited by the nature of the interface. For instance, the think-aloud method, in which the user narrates aloud his or her thoughts during a given task, is a primary tool used within usability testing (McDonald & Petrie, 2013). Yet, studies regarding the evaluation of user experience interactions with voice user interfaces must often disregard this method due to its interference with the user’s experience and have mainly relied on other post-task psychometric measurements and qualitative methods such as diaries and follow-up interviews (Jiang et al., 2015; Easwara et al., 2014; Lopatovska & Williams, 2018; Lau et al., 2018; Sciuto et al., 2018; Lopatovska & Oropeza 2018; Porcheron et al., 2018). Both post-task psychometric measurements and qualitative methods rely on direct judgments of causal efficacy, a key component of explicit measures (Dewey & Knoblich, 2014).

While these methods offer key information in regard to users’ experiences, users subject to these methods may succumb to cognitive biases, such as social desirability. As suggested in a study by Piedmont (2014), social desirability may lead a participant to dismiss his or her honest opinion for a more socially acceptable answer. Although the conscious mind can opt for the best fitting narrative, the subconscious may tell a different story. Building on this, a user's reaction to a given device can be derived from unconscious and automatic mechanisms (Ortiz de Guinea et al., 2013). In response to the automatic and somatic nervous systems (Shu et al., 2018), physiological signals are transmitted to various biological systems, including voice, facial expressions and muscular tonus (Levenson, 2014), which can be monitored through the use of biosensors. As a result, nonconscious and automatic emotional responses can be observed (Ortiz de Guinea et al., 2013). These responses are captured using physiological measures, defined as measures

used to index psychological constructs, be it states or processes (Lewis-Beck & al., 2004). Since users generally cannot manipulate their own physiological reactions, this transparency can be noted as a benefit for the use of such measures (Tiberio, 2013). Moreover, physiological measures provide precise, real-time data while being unobstructive and free of retrospective cognitive biases (Ortiz de Guinea et al., 2013). Physiological measures fall under the realm of implicit measures, defined by their sensory attenuation and temporal binding (Dewey & Knoblich, 2014).

An array of physiological measures is studied in today's HCI research, including, for example, heart rate and heart rate variability (HRV), respiration rate, and electrodermal activity (EDA) (Riedl & Léger, 2016). Specifically, within the context of voice user interface studies, Le Pailleur et al. (2020) featured automatic facial analysis and electrodermal activity to assess the users' experiences with a voice assistant. Research by Zhang et al. (2009) measured the cardiovascular and electrodermal activity of elderly participants interacting with various service robot interfaces, including via voice messaging, in the aim to understand perceptions and emotional responses towards robots in a healthcare setting. Hence, a multi-method approach utilizing both physiological and explicit measures can be used in order to obtain a more thorough understanding of user emotions.

Another obvious choice for assessing changes in affective state is through the study of speech. Recent studies have suggested the human voice to be a rich and ubiquitous medium of emotional communication (Cordaro et al., 2016; Juslin & Laukka, 2003; Kraus, 2017; Laukka et al., 2016; Provine & Fischer, 1989; Vidrascu & Devillers, 2005). Despite the fact that speech is commonly observed through the lens of emotions, the study of emotions within speech in a HCI context is a divided terrain. For engineers developing voice controlled HCI systems, acoustic and spectral features are primarily analyzed (Hartmann et al., 2013). By extracting such features, researchers have been able to mine the emotional labels of speech through the study of emotional recognition of speech (Chernykh & Prikhodko, 2017; Xia & Liu, 2017; Tao & Liu, 2018). On the other hand, psychologists tend to analyze and identify emotions using categories, schemes, and dimensional emotion spaces (Hartmann et al., 2013). To bridge the gap between both

spheres, Hartmann et al. (2013) proposed a novel approach linking machine measurable variations in emotional speech and the dimensional emotion theory (PAD). Although the measures used to observe emotions can differ, the underlying quest to better understand affective states remains.

Understanding affective states is a topic of interest at the core of sentiment analysis, being the study of people's emotions or attitudes (Maghilnan & Rajesh Kumar, 2017). Various machine learning approaches have been employed to classify these emotions and attitudes (Tyagi & Sharma, 2018). Research regarding sentiment analysis predominantly draws from text mining techniques, in which a sentiment expressed via text is analyzed (Maghilnan & Kumar, 2018). However, audio mining techniques can also be utilized. In the case of audio sentiment analysis, speech recognition, a process in which spoken words and phrases are converted into machine-readable format, as well as speaker recognition, a process in which speaker-specific vocal features are extracted, are employed in order to assess the sentiment expressed by a speaker (Maghilnan & Kumar, 2018). For every word spoken, a positive or a negative sentiment is attributed, taking into consideration the context of the conversation (Mukherjee & Bhattacharyya, 2013). Consequently, a sentiment score is simultaneously calculated, allowing for the machine in question to operate accordingly (Mukherjee & Bhattacharyya, 2013). Thus, both semantic and audio cues provide insight in regard to a speaker's emotion. Sentiment analysis is particularly relevant in voice user interface design, as customized settings suited to a user's preferences and needs can stem from a better understanding of a user's emotional state (Maghilnan & Kumar, 2018), favouring an optimal user experience.

A multi-method approach to assess user emotions can further provide a greater understanding of affective states. Indeed, various studies employing data from both speech and physiological measures, such as EDA and facial expression, have built successful multi-modal emotion recognition systems (Greco et al., 2019; Castellano et al., 2008; Alshamsi et al., 2018). As per sentiment analysis, emotion recognition systems are built on algorithms. Considering the fact that our research seeks to better understand the effectiveness of isolated measures through their respective features in explaining emotional events induced by voice user interface interactions, rather than building multi-



modal algorithms, emotion recognition and sentiment analysis systems were not deemed relevant to this study. Moreover, although sentiment analysis is an important tool in voice user interface design, it relies on contextual information. As stressed, this includes the linguistics of a given speech in which each spoken word is accounted for. Due to the fact that the observations of this study are of single-worded responses, employing sentiment analysis was deemed inappropriate.

As stressed previously, the current literature regarding the methods of voice user interface evaluation revolve primarily around explicit qualitative methods. Yet, as highlighted, implicit methods detect automatic and unconscious reactions that may further shed light upon lived emotions and experiences. Due to the vocal nature of voice user interface interactions, speech as a measure may be considered an evident route in comparison to physiological measures. Yet, physiological measures also have an informative quality which can provide insight into the emotional events during such interactions. To our knowledge, no other study has compared the effectiveness of speech and physiological measures via their respective features in explaining emotional events occurring during voice user interface interactions. This important gap within literature paves the way to potential key insights that may further help UX practitioners in their conception and evaluation of voice user interfaces.

Due to its growing presence and importance, understanding how humans interact with voice user interfaces is a thriving and essential field of information systems (IS) and HCI research. With this said, the following study seeks to better understand emotional events faced by users interacting with a voice user interface by studying participants' speech and physiological data, featuring both electrodermal activity (EDA) and automatic facial expression (AFE), in the aim of proposing the most effective measure. The scope of our study focuses on implicit measures utilizing multi-sensor physiological data. By doing so, we seek to address the gap within literature by pairing EDA and AFE alongside speech in the study of users' emotions induced by a voice user interface. Moreover, by studying emotional events through both a vocal and physiological lens, we aim to compare their effectiveness in explaining intense emotional responses, defined by obstacle-ridden and provocative voice user interface interactions. Assessing the effectiveness of each derived

physiological and speech feature may consequently favour a more efficient voice user interface evaluation, as selecting the best fitted measure can limit the resources used in vocal product evaluations. Within the context of this study, we propose a new methodological approach in the broader aim of ensuring optimal user-centric experiences that thrive on successful voice user interface interactions. In sum, findings from this study will further contribute to the research on voice user interface technologies while equipping UX professionals with valuable knowledge to ensure positive user experiences.

## **Objectives and Research Questions**

Our research aims to compare and highlight the effectiveness of speech features against physiological features in understanding intense emotional responses provoked by voice user interface interactions. As a result, a pivotal question has been posed;

***RQ1:** Between speech and physiological features, which are more informative in assessing intense emotional responses during vocal interactions with a voice user interface?*

The secondary aim of this study is merely to capture these intense emotional responses which occurred during voice user interface interactions. Although speech and physiological measures have been widely used in HCI literature, few studies have sought to simultaneously capture speech and physiological data within the present context. This leads us to our secondary research question;

***RQ2:** Can we unobtrusively identify an intense emotional response during voice user interface interactions?*

Hence, we seek to better understand the underlying emotions caused by voice user interface interactions while comparing the effectiveness of speech and physiological measures through the strength of their extracted features. By comparing the informative strength of speech against physiological features, we may further recommend which approach is better suited for voice user interface evaluation. Consequently, findings from this study may guide UX practitioners to select the most effective method, favouring

efficient voice user interface evaluations. Moreover, these insights further contribute to the set of guidelines regarding voice user interface evaluation, an emerging topic of interest.

## Contribution

The following chart illustrates my overall contributions to the study throughout its stages. The contribution is presented in a percentile form.

**Table 1:** Contribution to the responsibilities of the research project phases

Step-by-step process	Personal Contribution
Defining the expectations and needs of the partner	<p>Formulating appropriate research questions based on the client's expectations and needs - <b>80%</b></p> <p>*Support from the directors and supervisor was provided in order to determine the research partner's expectations and needs.</p> <p>*Support from the directors and supervisor was provided to formulate appropriate research questions.</p>
Literature Review	<p>Researching et reading the various articles related to the relevant subjects of the thesis - <b>100%</b></p> <p>Determining the key concepts which provided context to the research questions – <b>80%</b></p> <p>Writing a literature review based on the key concepts and constructs revolving the subject - <b>100%</b></p> <p>*Support from the directors and supervisor was provided to guide and revise the literature review.</p>

<b>Step-by-step process</b>	<b>Personal Contribution</b>
Experimental Design	<p>Requesting ethical approval from CER - <b>50%</b></p> <p>Conceiving and formalizing the experimental protocol – <b>50%</b></p> <p>*Members of the Tech3lab alongside directors and supervisor conceived the experimental protocol.</p>
Participant Recruitment	<p>Recruiting participants – <b>0%</b></p> <p>Managing compensations for recruited participants – <b>0%</b></p> <p>*Members of the Tech3lab were responsible for this portion.</p>
Data Collection	<p>Pre-test – <b>0%</b></p> <p>Data collection – <b>50%</b></p> <p>*Research assistants from Tech3lab were partially responsible for this portion.</p>
Data Analysis	<p>Statistical Analysis – <b>90%</b></p> <p>Third-party evaluator coding analysis – <b>100%</b></p> <p>*Support from the lab’s statistician was of great help in the analysis process.</p>
Writing	<p>Writing scientific and managerial articles – <b>100%</b></p> <p>*Support from the directors and supervisor was provided to guide and revise the articles.</p>

## Thesis Plan

We conducted a remote within-subject experiment in which speech, facial expression, and EDA responses from 16 subjects were recorded during voice user interface interactions lasting approximately 30 minutes that were purposely designed to elicit frustration and shock, resulting in 188 observations. By including third-party evaluators, we subsequently established ground-truth with non-expert evaluations for these measurements through manual human assessment of four dimensions of affective state: valence, arousal, control, and short-term emotional episodes (STEE), with inter-rater reliability scores calculated per dimension.

The thesis is structured as follows. Within the first chapter and second chapters, an introduction followed by a literature review regarding the study of emotion in UX, in which the leading physiological measures and speech features used to observe user emotions, will be presented. Following this, an article in chapter three will encompass the proposed approach and hypotheses, research methodology, results of the study, and an interpretation of these results within the discussion section. The paper will conclude with a fourth chapter, in which a brief managerial article summarizing the main takeaways is presented. A bibliography and appendix will be featured following these sections.

## References

- Alshamsi, H., Kepuska, V., Alshamsi, H., & Meng, H. (2018, November). Automated facial expression and speech emotion recognition app development on smart phones using cloud computing. In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (pp. 730-738). IEEE. <https://doi.org/10.1109/IEMCON.2018.8614831>
- Castellano, G., Kessous, L., & Caridakis, G. (2008). Emotion recognition through multiple modalities: face, body gesture, speech. In *Affect and emotion in human-computer interaction* (pp. 92-103). Springer, Berlin, Heidelberg.
- Chernykh, V., Sterling, G., & Prihodko, P. (2017). Emotion Recognition From Speech With Recurrent Neural Networks. *ArXiv, abs/1701.08071*.
- Cordaro, D. T., Keltner, D., Tshering, S., Wangchuk, D., & Flynn, L. M. (2016). The

- voice conveys emotion in ten globalized cultures and one remote village in Bhutan. *Emotion*, 16(1), 117–128. <https://doi.org/10.1037/emo0000100>
- Dewey, J. A., & Knoblich, G. (2014). Do implicit and explicit measures of the sense of agency measure the same thing?. *PloS one*, 9(10), e110118. <https://doi.org/10.1371/journal.pone.0110118>
- Easwara Moorthy, A., & Vu, K.-P. L. (2015). Privacy Concerns for Use of Voice Activated Personal Assistant in the Public Space. *International Journal of Human-Computer Interaction*, 31(4), 307–335.
- Fagherazzi, G., Fischer, A., Ismael, M., & Despotovic, V. (2021). Voice for health: the use of vocal biomarkers from research to clinical practice. *Digital Biomarkers*, 5(1), 78–88. <https://doi.org/10.1159/000515346>
- Giroux-Huppé, C., Sénécal, S., Fredette, M., Chen, S. L., Demolin, B., & Léger, P.-M. (2019). Identifying psychophysiological pain points in the online user journey: the case of online grocery. *Springer*, Cham, 459-473.
- Gothelf, J., & SEIDEN, J. (2013). *Lean UX: Applying lean principles to improve user experience*. “O’Reilly Media.
- Hartmann, K., Siegert, I., Philippou-Hübner, D., & Wendemuth, A. (2013). Emotion detection in HCI: From speech features to emotion space? *IFAC Proceedings Volumes (IFAC-PapersOnline)*, 12(PART 1), 288–295. <https://doi.org/10.3182/20130811-5-US-2037.00049>
- Jiang, J., Hassan Awadallah, A., Jones, R., Ozertem, U., Zitouni, I., Gurunath Kulkarni, R., & Khan, O. Z. (2015). Automatic Online Evaluation of Intelligent Assistants. *Proceedings of the 24th International Conference on World Wide Web - WWW '15. Presented at the 24th International Conference*. <https://doi.org/10.1145/2736277.2741669>
- Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129, 770–814. <http://dx.doi.org/10.1037/0033-2909.129.5.770>
- Kiseleva, J., Williams, K., Jiang, J., Hassan Awadallah, A., Crook, A. C., Zitouni, I., & Anastasakos, T. (2016). Understanding user satisfaction with intelligent assistants. *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval* (pp. 121-130). ACM.
- Kraus, M. W. (2017). Voice-only communication enhances empathic accuracy. *American Psychologist*, 72, 644–654. <http://dx.doi.org/10.1037/amp0000147>
- Lau, J., Zimmerman, B., & Schaub, F. (2018). Alexa, are you listening? privacy

- perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 1-311.
- Laukka, P., Elfenbein, H. A., Thingujam, N. S., Rockstuhl, T., Iraki, F. K., Chui, W., & Althoff, J. (2016). The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features. *Journal of Personality and Social Psychology*, 111, 686–705. <http://dx.doi.org/10.1037/pspi0000066>
- Le Pailleur, F., Huang, B., Léger, P. M., & Sénécal, S. (2020). A new approach to measure user experience with voice-controlled intelligent assistants: A pilot study. In M. Kurosu (Ed.), *Human-computer interaction. Multimodal and natural interaction. HCII 2020. Lectures notes in computer science* (Vol. 12182, pp. 197–208). [https://doi.org/10.1007/978-3-030-49062-1\\_13](https://doi.org/10.1007/978-3-030-49062-1_13)
- Levenson, R. W. (2014). The autonomic nervous system and emotion. *Emotion Review*, 6(2), 100–112. <https://doi.org/10.1177/1754073913512003>
- Lewis-Beck, M. S., Bryman, A., & Futing Liao, T. (2004). *The SAGE encyclopedia of social science research methods* (Vols. 1-0). Thousand Oaks, CA: Sage Publications, Inc. <https://dx.doi.org/10.4135/9781412950589>
- Levy, J., & Calacanis, J. (2015). *Ux strategy: how to devise innovative digital products that people want*. O'Reilly Media. Retrieved October 22, 2021
- Lopatovska, I., & Oropeza, H. (2018). User interactions with “Alexa” in public academic space. *Proceedings of the Association for Information Science and Technology*, 55(1), 309–318. <https://doi.org/10.1002/pra2.2018.14505501034>
- Lopatovska, I., & Williams, H. (2018). Personification of the Amazon Alexa: BFF or a mindless companion. *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (pp. 265-268). ACM.
- Maghilnan, S., & RajeshKumar, M. (2017). Sentiment analysis on speaker specific speech data. *2017 International Conference on Intelligent Computing and Control (I2C2)*, 1-5. <https://doi.org/10.1109/I2C2.2017.8321795>
- McDonald, S., & Petrie, H. (2013, April). The effect of global instructions on think-aloud testing. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2941-2944). <https://doi.org/10.1145/2470654.2481407>
- Mukherjee, S., & Bhattacharyya, P. (2013). *Sentiment Analysis : A Literature Survey*. 1–51. <http://arxiv.org/abs/1304.452>
- Myers, C., Furqan, A., Nebolsky, J., Caro, K., & Zhu, J. (2018). Patterns for how users

- overcome obstacles in voice user interfaces. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-7). <https://doi.org/10.1145/3173574.3173580>
- Ortiz de Guinea, A., Titah, R., & Leger, P. M. (2013). Measure for measure: a two study multi-trait multi-method investigation of construct validity in is research. *Computers in Human Behavior*, 29(6), 833–844.
- Piedmont, R. L. (2014). Social Desirability Bias. *Encyclopedia of Quality of Life and Well-Being Research*, 6036–6037. [https://doi.org/10.1007/978-94-007-0753-5\\_2746](https://doi.org/10.1007/978-94-007-0753-5_2746)
- Platzer, D. (2018, October). Regarding the pain of users: towards a genealogy of the “pain point”. In *Ethnographic Praxis in Industry Conference Proceedings* (Vol. 2018, No. 1, pp. 301-315). <https://doi.org/10.1111/1559-8918.2018.01209>
- Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice Interfaces in Everyday Life. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. <https://doi.org/10.1145/3173574.3174214>
- Provine, R. R., & Fischer, K. R. (1989). Laughing, smiling, and talking: Relation to sleeping and social context in humans. *Ethology*, 83, 295– 305.
- Public Service Enterprise Group Incorporated. (2020). Ways of conducting bank transactions in Canada 2018 | PSE&G. <https://nj.pseg.com/voiceassistant>. (Accessed 22 July 2021).
- Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017, May). Alexa is my new BFF: social roles, user satisfaction, and personification of the amazon echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2853-2859). ACM.
- Riedl, R., & Léger, P. M. (2016). Fundamentals of NeuroIS. *Studies in neuroscience, psychology and behavioral economics*, 127. <https://doi.org/10.1007/978-3-662-45091-8>
- Sciuto, A., Saini, A., Forlizzi, J., & Hong, J. I. (2018). “Hey Alexa, What’s Up?” *Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS '18*. <https://doi.org/10.1145/3196709.3196772>.
- Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., Xu, X., & Yang, X. (2018). A Review of Emotion Recognition Using Physiological Signals. *Sensors*, 18(7), 2074. <https://doi.org/10.3390/s18072074>
- Statista. (2021). Number of digital voice assistants in use worldwide from 2019 to 2024 (in billions)\* | Statista. <https://www.statista.com/statistics/973815/worldwide->



digital-voice-assistant-in-use/. (Accessed 10 July 2021).

- Statista. (2020). The most important voice platforms in 2020 | Statista.  
<https://www.statista.com/chart/22314/voice-platform-ranking/> (Accessed 10 July 2021).
- Tao, F., & Liu, G. (2018). Advanced LSTM: A study about better time dependency modeling in emotion recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2906-2910). IEEE.
- Tiberio, L., Cesta, A., & Belardinelli, M. O. (2013). Psychophysiological methods to evaluate user's response in human robot interaction: A review and feasibility study. *Robotics*, 2(2), 92–121. <https://doi.org/10.3390/robotics2020092>
- Tyagi, A., & Sharma, N. (2018). Sentiment Analysis using logistic regression and effective word score heuristic. *International Journal of Engineering and Technology(UAE)*, 7(2), 20–23. <https://doi.org/10.14419/ijet.v7i2.24.11991>
- Vidrascu, L., & Devillers, L. (2005). Real-life emotion representation and detection in call centers data. In J. Tao, T. Tan, & R. W. Picard (Eds.), *Affective computing and intelligent interaction* (pp. 739–746). Berlin, Germany: Springer.
- Wang, W. C., Chien, C. S., & Moutinho, L. (2015). Do you really feel happy? Some implications of Voice Emotion Response in Mandarin Chinese. *Marketing Letters*, 26(3), 391–409. <https://doi.org/10.1007/s11002-015-9357-y>
- Xia, R., & Liu, Y. (2017). A multi-task learning framework for emotion recognition using 2d continuous space. *IEEE Transactions on Affective Computing*, 8(1), 3–14. <https://doi.org/10.1109/TAFFC.2015.2512598>
- Zhang, T., Kaber, D. B., Zhu, B., Swangnetr, M., Mosaly, P., & Hodge, L. (2010). Service robot feature design effects on user perceptions and emotional responses. *Intelligent Service Robotics*, 3(2), 73–88. <https://doi.org/10.1007/s11370-010-0060-9>

## Chapter 2. Literature review

To support and justify the relevance of this study, a literature review was conducted. The areas of interest were the study of emotions within the field of user experience (UX), in addition to the study of speech features and physiological measures depicting emotional states. With this said, the literature review is structured as follows: First, a definition and the role of emotion within user experience will be presented, followed by an overview of UX evaluation within human-computer interaction (HCI) research, proceeded by the common explicit methods and measures in assessing user emotions. Next, popular explicit methods used specifically within the study of voice user interface evaluation will be explored. Following this, a presentation of the limits regarding explicit measures will be featured, proceeded by a summary of the advantages of implicit measures used to evaluate interfaces, with a focus upon the physiological measures and emotion revealing speech features.

### 2.1 Definition and role of emotions within user experience

According to the ISO definition of user experience, UX “includes all the users' emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviours and accomplishments that occur before, during and after use”<sup>1</sup> (2018). With this said, emotion is an important pillar shaping user experiences. The study of emotion recognition is gaining ground within human-computer interaction systems, with automation and personalization as key components of these systems dependent of this detection (Kollia, 2016). Nowadays, most human-computer interactions involve some form of automation and personalization (Kollia, 2016), from automatic spell-check to tailored content on social media platforms, making them ubiquitous within users' experiences. Beyond the efficiency and effectiveness expected of these systems, users are searching for emotional satisfaction from their experiences (Shih & Liu, 2007). It is in a company's best interest to cater to this desire, as a product triggering a positive emotional reaction is more likely to be deemed appealing by users (Hassenzahl, 2008). Achieving this positive emotional

---

<sup>1</sup> <https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en>

reaction early on is particularly important, as the emotions experienced in early product use have immediate and lasting effects upon its evaluation (Wood & Moreau, 2006). Thus, understanding the users' emotions and tapping into emotional satisfaction is key in a digital product or service's success.

According to Damasio (1994), emotions can be described as changes within the body and brain in response to a specific stimulus of one's perceptions relative to a given object or event. In response to the automatic and somatic nervous systems (Shu et al., 2018), these changes are manifested through physiological signals that are transmitted to various biological systems, including voice, facial expression and muscular tonus (Levenson, 2014). A second concordant element to the definition of emotion is that it is "a reaction to events deemed relevant to the needs, goals, or concerns of an individual" (Brave & Nass, 2002, p.54). Within the same vein, interpretations of the social context, associations and memories, in addition to semantic knowledge, all play into emotional experiences (Keltner et al., 2019). As a result, it may be said that emotions are personal, holistic, and complex constructs.

The complexity of emotions has resulted in two primary yet polarizing theories, the basic emotion theory and the dimensional theory. According to the theory of basic emotion, humans are equipped with a discrete and limited set of basic emotions (Ekman, 1992; Panksepp, 1998; Plutchick, 1962; Tomkins, 1962, 1963). Indeed, research by pioneers Plutchick (1962) and Ekman (1992) proposed eight and six primary emotions respectively that may be noted through distinct facial expressions, including fear, anger, joy, sadness, disgust, and surprise. These emotions are manifested in an organized recurring pattern associated to behavioral components (Ekman, 1992). More precisely, these internal states are externally expressed as specific stereotypical behaviours, such as instinct (Gu et al., 2019).

However, recent findings have challenged this framework. Research by Posner et al. (2005) suggests that certain emotions are not characterized by distinct facial expressions. Moreover, facial expressions may be associated to more than one emotion, which consequently poses a challenge to the taxonomy of facial expression proposed as it

inadequately matches the taxonomy of emotion (Posner et al., 2005). Furthermore, as suggested by Posner et al. (2005), the theory of basic emotion fails to define the relationship between basic emotions and peripheral physiological correlation.

Rather than observing the behavioural and expressive manifestation of emotions, recent research has taken to the subjective components of emotion (Posner et al., 2005). Research regarding the subjective components of emotion have suggested that emotions instead emerge from cognitive interpretations of core physiological experiences (Cacioppo et al., 2000; Russell, 2003). Thus, within the circumflex models of affect, the core physiological bases of affective experiences are explored (Posner et al., 2005). The conceptualization of these model revolves around the idea that affective states stem from cognitive interpretations of core neural sensations, which are the result of two independent neurophysiological systems (Posner et al., 2005). In contrast to the theory of basic emotions, there is no discrete and independent neural system that subserves every emotion. Moreover, emotions are “a continuum of highly interrelated and often ambiguous states”, rather than isolated and discrete entities, as suggested in the basic theory of emotion framework (Posner et al., 2005, p.719). Within circumflex models of affect, affective experiences are composed of two independent neurophysiological systems. Various models exist and have conceptualized these systems in different ways, including the dimensions of positive and negative affect (Watson et al., 1999), tension and energy (Thayer, 1989), approach and withdrawal (Lang et al., 1998), as well as valence and arousal (Russell, 1980).

As noted within established UX literature, one of the theoretical models commonly used to assess emotional expressions is the dimensional study of emotion (Scherer, 2003; Léger et al. 2014). Within the dimensional study of emotions, valence and arousal are often observed (Sutton, 2019). The valence dimension relates to the evaluation of one’s experience, ranging from displeasure to pleasure (Laukka, 2005). In other words, as described by Burton-Jones and Gallivan (2007), the affective state refers to “what a user feels” (p.659). As for the arousal dimension, it stems from one’s sense of energy, ranging from sleep, or calm, to frenetic excitement (Feldman Barrett & Russel, 1998; Sutton 2019). A third dimension, potency, is also commonly reported (Laukka, 2005). Potency,

often referred to either as dominance, power or control, refers to one's coping potential or power, in a given situation (Laukka, 2005). The dimensional study of emotions is widespread within HCI research and can be observed in various contexts, including voice user interfaces, as seen within a study conducted by Le Pailleur et al. (2020).

## **2.2 Evaluation methods and tools in user experience**

Intention, demand, and affective states can be determined by the cues in which a human provides to their interaction partner (Hartmann et al. 2013). Picking up on these cues may improve the possibility of a positive outcome within the interaction. On the contrary, neglecting them may lead to negative interactions. Thus, understanding the multifacet, underlying emotions caused by HCI systems is essential. By understanding emotional antecedents, designers are equipped with the know-how to create interfaces capable of producing desired emotional states (Brave & Nass, 2002).

Attempting to understand the users' experiences and respective emotions can be done in various manners. According to Battarbee & Koskinen (2005), there are three primary approaches to applying and interpreting user experiences, being the measuring approach, empathic approach, and the pragmatist approach. Within each of these approaches, the role of emotional experiences is important. However, depending on the selected approach, its role will be treated differently.

The measuring approach stems from the idea that experiences can be measured through emotional reactions (Battarbee & Koskinen, 2005). Within this approach, a direct link is made between a user's emotions and experience, suggesting a relationship between the two. Mainly used within the developing and testing phase, the measurement approach zones in on the aspects of user experience that can be measured, be it through the physical reactions or the subjective reporting of users (Battarbee & Koskinen, 2005). Examples of the methods and tools used to measure user emotions will be addressed in the following sections.

Through the lens of the empathic approach, an experience is emotional in nature. Products eliciting experiences must be anchored within the dreams, needs, and motivations of the users (Dandavate et al., 1996; Black, 1998). To create meaningful products, designers must “both observe and feel for the users” (Battarbee & Koskinen, 2005; Mäkelä & Fulton Suri, 2001; Kankainen, 2002, p.6). As a result, methods used within the empathic approach tend to be of qualitative nature and often combine in parallel visual and textual data, self-documentation, and projective tasks. Through these methods, designers construct descriptions regarding the users’ experiences, dreams, contexts, and expectations. With this being said, the empathic approach is primarily used for inspirational purposes, aimed to project a future experience rather than assess a current one.

Lastly, inspired by the pragmatist philosophy (Dewey, 1934), the pragmatist approach views experiences as momentary constructs shaped by the interactions between users and their environment. Adopting a holistic view of user experience, the pragmatist approach seeks to understand the interactions between users, technologies, and environments. It is important to note that this particular approach tends to be theoretical, rather than offer practical guidance regarding the design and evaluation of systems. In sum, the empathic approach focuses on user-centered design, the pragmatist approach attempts to link actions to meaning, while the measuring approach focuses on emotional responses.

As suggested by Battarbee & Koskinen (2005), depending on the approach and project development stage of a given UX product, the methods to evaluate this product will differ. Consequently, an array of UX evaluation methods exist. Indeed, in research by Vermeeren et al. (2010), a total of 96 UX evaluation methods were identified, although not all are intended to measure user emotions. According to research by Alves et al. (2014), in which practitioners were surveyed regarding their UX evaluation methods, 52.6% reported that UX evaluation occurs early on in the product development during the system design phase. As highlighted previously in Battarbee & Koskinen’s research (2005), this phase often adopts the measurement approach and seeks to assess the users’ emotional reactions, providing valuable feedback to designers early on.

In addition to varying by project stage, UX evaluation methods differ depending on the party involved within the evaluation. According to Alves et al. (2014), designers and end-users may both be included within UX evaluation. For designers, behavioural maps and customer experience audits are the primary evaluation methods (Alves et al., 2014). Developed by Ittersson et al. (1970), the behavioural map is a product of observation that allows for designers to record a user's behaviour by identifying locational or temporal patterns (Ng, 2016), whereas mapping the customer experience allows for designers to assess the set of interactions between a customer and a product that provoke a reaction (Gentile et al., 2007).

On the other hand, end-users participating in UX evaluations as subjects are most likely to participate in the following methods, being interviews, experience prototyping, observation, and the think-aloud method (Alves et al., 2014). During interviews, end-users partake in a series of questions during a one-on-one session, allowing for researchers to assess “how users feel, think, and what they perceive to be true” (Nielsen Norman Group, 2021)<sup>2</sup>. When participating in an experience prototyping activity, end-users gain first-hand appreciation of a product by actively engaging with a given prototype (Buchenau & Suri, 2000). As for observations, they allow for researchers to catch thoughts and feeling the end-users might not have put forward during a controlled experiment (Park et al., 2013). Lastly, the think-aloud method invites end-users to verbalize their thoughts while performing tasks, allowing researchers capture their thought-process (Fan et al., 2020).

The industry experts surveyed in research by Alves et al. (2014) indicated that they “almost always or always” resorted to interviews, experience prototyping, observation or the think-aloud method when evaluating products, with each method receiving more than 50% of responses (Alves et al., 2014, p.99). This high rate is congruent with the fact that end-users are 46% likely to be featured as subjects in UX evaluations (Alves et al., 2014). Thus, measuring the user's experience using various methods is common within UX evaluation.

---

<sup>2</sup> <https://www.nngroup.com/articles/user-interviews/>

## 2.3 Emotion evaluation in user experience

Interviews, observations, and the think-aloud methods are relevant methods to the study of user emotions, as they offer a glimpse into users' affective states (Wrigley et al., 2010). In parallel to these evaluation methods, a diverse set of measures can also be considered in emotion assessment within UX evaluations. As stressed previously, valence, arousal, and control are classic dimensions of affective state measured ubiquitously in IS research through self-assessment questionnaires. One of the most prevalent type of questionnaires is the Self-Assessment Manikin (SAM) scale proposed by Bradley and Lang in 1994 (Betella & Verschure, 2016). The SAM scale measures three emotional dimensions, that of pleasure, arousal, and dominance, using a series of graphic abstract characters displayed horizontally featuring a 9 point-scale, although 5 and 7-point variants may also be utilized (Betella & Verschure, 2016). The valence dimension is depicted by an array of pictographic representations ranging from a frowning to a smiling figure. The use of such a scale enables evaluators to rate the emotional intensity, from extremely negative to extremely positive. Arousal is illustrated by a sleepy to widely awake figure marked with an incremental explosion at its center, whereas the control dimension is represented by smaller to larger characters (Betella, Verstschure, 2016).

The SAM scale has been used in various HCI studies. In a study by Le Pailleur et al. (2020), researchers used a SAM scale to assess participants' self-perceived emotions following a series of tasks conducted with voice assistant interface Alexa. In another study, participants evaluated their emotional state following conditions involving multi-tasking upon a smartphone while walking on a treadmill (Mourra et al., 2019). Importantly, the SAM scale has not only been used for self-assessment of affective state. There are also numerous reports of the SAM scale being used for third-party assessment. For example, in a study conducted by Sutton et al. (2019), third-party evaluators used a 9-point rating SAM scale to gage 120-130 faces expressing various emotions. In a study by Grimm et al. (2007), evaluators assessed German and English audio recordings featuring acted and authentic emotion expressions once more using a SAM scale. Lastly, in a study by Jessen and Kotz (2011), evaluators rated the arousal



levels of auditory, visual, and audiovisual stimuli featuring emotional interjections of fear, anger, and neutral nature, expressed as “ah”, “oh”, and “hm”.

Another example of an emotional assessment tool is the Affective Slider (AS). A modern adaptation to the SAM scale, AS has been developed in recent years as subsequent measure to assess emotional states (Betella & Verschure, 2016). Unlike the SAM scale, Betella & Verschure (2016) exclude the control dimension as the “core affect” coined in Russel’s research (1980) deems the bipolar emotional space of valence and arousal sufficient to measure basic emotion (Betella & Verschure 2016; Russel, 1999). This digital self-reporting tool composed of two sliders measures both arousal and valence on separate continuous scales using pictograms. Both the SAM scale and AS allow for UX practitioners to assess users’ affective states quickly and simply.

As hinted, emotion is an important pillar of UX as it influences how users comprehend, decipher, experience, and interact with technology (Forlizzi & Battarbee, 2004). Assessing users’ emotions can be indicative of the user’s experience. For example, as suggested in research by Agarwal & Meyer (2009), usability can be affected by emotion, as a happy user is more likely to judge a product as being more usable than an unhappy user. Moreover, the usability of a product is likely to affect a user’s emotional state (Agarwal & Meyer, 2009). Hence, measuring emotion can be an important usability tool for designers, as there is an interesting relationship between usability and users’ emotional responses (Agarwal & Meyer, 2009). As explored in a study by Nass et al. (2005), researchers observed the effects of user emotion upon a driver’s performance and attitude by altering the in-car voice interface, from an energetic to subdued voice. Results suggested that pairing the voice of the car to the driver’s emotion had a noteworthy effect on both the driver’s performance and attitude. Thus, measuring users’ emotional responses to technology can serve as visceral indicators of either positive or negative experiences (Paul & Komlodi, 2014). Understanding the state of user experiences is particularly important, as each isolated experience with a given product may increase or decrease its value in the eyes of the user, while altering one’s expectations and motivations to use the product (Stickel et al., 2009). Research regarding this topic has suggested a relationship between emotion and future use (Beaudry & Pinsonneault, 2010; Paul &

Komlodi, 2012; Paul & Komlodi, 2014). Indeed, the experience of joy while using a product has been linked to its success (Stickel et al., 2009). Thus, measuring emotions through UX evaluation helps depict the state of user experiences, which can provide key insight to further guide designers in their quest to develop optimal and successful products.

Advantages of the commonly used UX evaluation methods and measures mentioned include their informal nature and low costs (Alves et al., 2014). Despite these advantages, important drawbacks resulting from the use of these methods are also to be noted, such as social desirability and retrospective biases (Ortiz de Guinea et al., 2014). The effects of such biases cause users to alter or misrepresent their affective states (Krosnick, 1999). Moreover, due to the vocal nature of voice user interfaces, additional inconveniences arise. For instance, the think-aloud method's intrusive nature is unsuitable for voice user interface evaluation as it can interfere with the user's experience. Hence, when seeking to measure voice user interface experiences, other methods must be employed.

## **2.4 Voice user interface evaluation**

In 1991, researcher Mark Weiser coined the term “ubiquitous computing”. According to Weiser, ubiquitous computing was a futuristic vision in which computers would become invisible and operated in the periphery of a user's attention (Weiser, 1991). Fast forward two decades later, the era of non-visual user interaction (No-UI) has arrived. As we increasingly navigate without or minimal use of graphical user interactions, experiences are shifting. At the forefront of this shift is the rise of voice user interface technologies.

According to authors Cohen et al. (2004, p.5) of the “Voice user interface design” book, a voice user interface is “what a person interacts with when communicating with a spoken language application”. Within the conceptualization of a voice user interface, prompts, or system messages, grammars, and dialog logic, also referred to as call flow, must be included (Cohen, 2004). The dialog logic defines the action taken by the system. If successful, the system provides the user's desired information (Cohen et al., 2004), ultimately ensuring an optimal voice user interface interaction. To improve a product

and consequently achieve preeminent user experiences, industry experts resort to product evaluations (Väänänen-Vainio-Mattalia et al., 2008).

There are various voice user interface evaluation methods, including self-report questionnaires, dairies, interviews, and observations. The majority of evaluations and studies of voice user interface systems tend to focus on task performance and self-report questionnaires. Often conducted post-interaction (Clark et al., 2019), self-report questionnaires are quick and inexpensive methods to gather large amounts of data and has been popularized within the study voice user interface design (De Singly, 2016). Commonly, questionnaires are often meant to measure concepts such as usability and user attitudes towards voice user interface interactions (Clark et al., 2019). Multiple standardized questionnaires to evaluate user experience or subjective user satisfaction following voice user interface interactions currently exist, including the Subjective Assessment of Speech System Interfaces (SASSI), the Speech User Interface Service Quality (SUISQ) and the Paradigm for Dialogue Evaluation System (PARADISE) (Kocaballi et al., 2019). Yet, according to research by Clark et al. (2019), the use of standardized questionnaires is low, as many studies focused on voice user interface evaluation employ custom-built scales. Validity and reliability are consequently at risk (Clark et al., 2019).

Another method used within the study of voice user interface design are diaries. As a frequent method adopted in qualitative research, diaries serve as portals to users' subjective impressions and reflections, providing self-interpretations of participants' worlds (Alaszweski, 2006) and intimate descriptions of their day-to-day lives (Nicholl, 2010). One of the main advantages of this method stems from the fact that there is no need for a researcher's presence. Indeed, the presence of a stranger may affect a user's interaction with a voice assistant, as the common usage context is often in a private and comfortable setting, such as within one's home (Easwara et al., 2014). For instance, in a study by Lopatovska and Williams (2018), participants shared their thoughts during a four-day period upon on online diary in a study revolving around the personification of voice assistant Alexa.

User thoughts and impressions regarding voice user interfaces are also often gathered through interviews (Kocaballi et al., 2019). Indeed, interviews can provide insight regarding the various dimensions shaping voice user interface experiences lived by users. As seen within HCI literature, interviews are often utilized in conjunction with the methods previously mentioned, providing complimentary qualitative data. For instance, in a study by Garg and Moreno (2019), semi-structured interviews were conducted in parallel with diary logs in order to assess user sharing practices of voice assistants. Similarly, complimentary in-depth interviews and conversational logs allowed researchers Sciuoto et al. (2018) to assess users' in-home usage patterns with voice assistant Alexa.

Lastly, observations are frequently considered within the study of voice user interfaces. It is the only traditional method allowing user behaviour to be recorded directly during the user interaction rather than post-interaction. In a study by Lopatovska and Oropeza (2018), voice assistant interactions were observed in public spaces. Although user behaviours are recorded in a timely manner, observations do not necessarily shed light upon the cognitive and emotional states of users shaped by voice user interface interactions.

## **2.5 Physiological measures within the study of emotion in user experience**

With the exception of observations, traditional methods regarding voice user interface evaluations stem from self-assessment. Self-assessment methods are commonly employed within emotion research (Betella & Verschure, 2016). Indeed, the primary method in the evaluation of users' behaviours and subjective experiences within social and behavioural science is through retrospection (Schwarz, 2007). Within UX literature, the use of retrospective tools to assess a user's emotional state is widespread (Bruun & Ahm, 2015). As suggested, self-reported measures offer rich qualitative data regarding users' experiences with voice user interfaces. However, despite its informative quality and widespread usage within voice user interface research, self-reported measures can be limiting.

Although the users' perceptions about a given interaction are recorded, a lack of a thorough understanding regarding the users' experiences is to be noted when solely relying on self-reported measures. Indeed, measures such as self-report questionnaires are limited to the users' conscious thoughts and perceptions (Riedl & Léger, 2016). Explicit and observational measures depict a partial story, as a fragment of the users' feelings towards the digital entity is observed (Ortiz de Guinea et al., 2014). Human behaviours are shaped by unconscious information processing and perceptions (Lieberman 2007). Due to the explicit nature of self-reported measures, automatic mental states, which can occur without the users' conscious awareness, are dismissed (Ortiz de Guinea & Webster, 2013; Ortiz de Guinea & Markus, 2009). Hence, solely using self-reported data limits the understanding of information technology (IT) behaviour (Riedl & Léger, 2016). Moreover, self-report data may be inaccurate. As suggested previously, subjects of self-reported measures are at risk of cognitive biases, such as social desirability (Piedmont, 2014). As a result of social desirability, participants can be tempted to opt for socially acceptable responses rather than sharing their honest opinions (Piedmont, 2014).

In order to counter the limits imposed by explicit measures and assess underlying emotional expressions, physiological measures can be employed in parallel within HCI research. Physiological measures are alternative methods allowing researchers to comprehend a user's emotional state (Dirigan & Göktürk, 2011). Indeed, these measures are considered important tools when assessing elements or events of cognitive or emotional relevance to the users (Picard, 1995; Ward & Marsden, 2003; Bentley et al., 2005). Physiological measures record a user's affective and cognitive state in an unobstructive fashion (Dirigan & Göktürk, 2011). They have been deemed as a reliable approach to assess a user's emotional state, as the physiological manifestations of the user's psychological sentiments in real time are observed (Andreassi, 2000). In order to obtain a rich comprehension of a user's emotional state, at least two measures need to be employed (Ganglbauer et al., 2009; Maia & Furtado, 2016). Moreover, a multi-method approach in which explicit or perceptual, and implicit, such as physiological methods, are used simultaneously is particularly insightful (Ortiz de Guinea et al., 2014; Ortiz de Guinea et al., 2013). Indeed, by combining these complimentary methods, biases may be overcome while providing a deeper understanding of the user's experience.

Research in HCI has demonstrated that physiological measures are viable indicators of cognitive and affective states (Rowe et al., 1998; Ortiz de Guinea et al., 2013; Ortiz de Guinea et al., 2014; Giroux-Huppé et al. 2019; Beauchesne et al., 2019; Lourties et al., 2018; Agourram et al., 2019; Maunier et al., 2018; Le Pailleur et al., 2020). An array of physiological signals, such as heart rate, electrodermal activity (EDA) and facial expressions, are indicative of cognitive and emotional states (Riedl & Léger, 2016). In HCI literature, emotional states are often defined by two dimensions, being valence and arousal (Ortiz de Guinea & Markus, 2009; Léger et al., 2014). These two constructs can be measured using physiological tools. Within UX research, two common physiological indices used to measure affective state are facial micro expressions and electrodermal activity (EDA). Often captured via a webcam, facial micro expressions are generally quantified using some form of automated facial expression (AFE) analysis software and assessed through the lens emotional valence. Emotional valence, characterized by negative emotions (e.g., fear, anger, sadness) and positive emotions (e.g., joy, surprise) on opposite sides of the spectrum, refers to the emotional response to a specific stimulus (Bradley & Lang, 1999).

Several studies utilizing AFE to assess user emotion have been conducted within HCI research. In one study, it was found that data captured via facial micro-expressions was more effective in measuring instant emotions and fun of use in comparison to a user's questionnaire (Zaman & Shrimpton-Smith, 2006). Moreover, Zaman and Shrimpton-Smith's (2006) results suggest that questionnaire data was not necessarily a genuine reflection of the users' feeling while accomplishing a task, but rather a reflection of the outcome of a given task. Indeed, similar findings were observed in a study by Lourties et al. (2018), suggesting that the experience lived by a participant is different than what is often reported. This key insight was obtained through research exploring the convergent validity of self-reported measures with psychophysiological measures.

EDA, on the other hand, is a measurement of electrical resistance through the skin. More precisely, it measures changes of skin conductance response (SCR) from the nervous system functions (Braithwaite et al., 2013; Dawson et al., 2000; Bethel, 2007). In other words, after an electrical potential has been applied to two point of skin contact, the skin

conductance, or flow between these two points of skin contact, can be measured (Braithwaite et al., 2013). The easy to use and reliable physiological measure has been widely used in NeuroIS research (Léger et al., 2014; von Brocke et al., 2013; Giroux-Huppé et al., 2019; Lamontagne et al., 2019). Often captured via electrodes on the palm of the hand, it is sensitive to the changes in skin pore dilation and sweat gland activation, which are in turn sensitive to changes in emotional arousal. The arousal levels measured via EDA range from very calm, to neutral, to highly stimulated (Ekman & Friesen, 1978). It has been suggested to be an ecologically valid portrait of the user's arousal, while being non-invasive and free of over recorded behaviour (Dirican & Göktürk, 2011). In one study regarding child-robot interaction, the measured arousal via skin conductance was deemed as a valuable and reliable method in assessing social robot interactions (Leite et al. 2013). With this said, both EDA and facial micro-expressions help depict emotional states.

## **2.6 Voice measures within the study of emotion in user experience research**

Another obvious choice for assessing changes in affective state is through the study of speech. Studies have suggested the human voice to be a rich and ubiquitous medium of emotional communication (Cordaro et al., 2016; Juslin & Laukka, 2003; Kraus, 2017; Laukka et al., 2016; Provine & Fischer, 1989; Vidrascu & Devillers, 2005). The source-filter theory of speech production contributes to the understanding of speech acoustics in relation to emotional states (Bachorowski, 1999; Kent, 1997). According to this framework, speech is a result of the pairing of source energy, produced by the vibration of vocal folds, as well as the subsequent filtering of that energy by the vocal tract above the larynx (Bachorowski, 1999). In other words, speech is a result of the contraction of muscles surrounding the diaphragm, which consequently results in burst of air particles transformed into sound through vibrations of the vocal folds (Cowen et al., 2019). Depending on the position of the jaw, tongue and other implements of vocal control, the sound emitted may be a word, a laughter, a cry, a sigh, and so on (Titze & Martin, 1998; Cowen et al., 2019).

In the field of emotion detection within speech, researchers often use prosodic features such as fundamental frequency (F0), energy, and duration, alongside important psychoacoustics features in emotion perception such as speech rate, pitch changes, pitch contours, voice quality, spectral content, energy level, and articulation (Tahon et al., 2012; Shilker, 2009). Research regarding paralinguistic features and the emotion in speech have suggested that fundamental frequency (F0) (e.g., minimum, maximum, mean, jitter), energy and amplitude (e.g., loudness, shimmer), temporal (e.g., duration) and quality parameters (e.g., harmonics-to-noise ratio [HNR]) are amongst the most important (Lausen & Hammerschmidt, 2020; Juslin & Laukka, 2003; Johnstone & Scherer, 2000). Of the various acoustic measures featured, speech rate, measures related to the fundamental frequency, and vocal amplitude have received the most attention (Bachorowski & Owren, 2007; Scherer, 1986), with F0 being commonly used within in voice-based emotion research (Bachorowski, 1999).

F0 relates to the rate of vocal fold vibration and is perceived as vocal pitch (Bachorowski, 1999). In other words, pitch relates to how the fundamental frequency is perceived, with the pitch period representing the fundamental period of the signal (Li & Jain, 2009). More precisely, pitch is an indication of the frequency at which the larynx opens and closes due to the vocal cords, which consequently produces voiced sounds (Li & Jain, 2009). In research by Lausen & Hammerschmidt (2020), 1038 emotional expressions were analyzed according to 13 prosodic acoustic parameters, including F0 and its variations.

As for speech energy, it may be assessed using the spectral slope, where the tendency to have low energy during high frequencies is observed (Mannepalli et al., 2018). In a study by Guzman et al. (2013), the influence of emotional expression in spectral energy distribution for trained theater actors was observed. Spectral spread, on the other hand, denotes the total bandwidth of a speech signal using spectral centroid, a measure used to evaluate the brightness of a speech (Mannepalli et al., 2018). In a study by Mannepalli et al. (2018), both spectral slope and spectral spread were extracted from speech signals for emotion recognition purposes. Remaining within the context of speech recognition, spectral entropy can be observed to assess silence and voice region of speech (Toh et al., 2005). As seen within a study by Papakostas et al. (2017), spectral entropy, alongside



spectral centroid, spectral spread and energy, were observed in the aim of analyzing speakers' emotions.

Beyond the scope of emotion recognition, speech features can also be indicative of health conditions. Research surrounding pitch period entropy (PPE), a measure that denotes the impaired control of F0 during sustained phonation, has suggested PPE to be an indicative speech feature of Parkinson's disease (Arora et al., 2019; Little et al., 2019). In sum, voice-based research suggests a relationship between speech features and states of being.

The study of extracted speech features can be noted within a HCI context. Many have been studied in a speech emotion recognition (SER) context, an increasingly popular subject that aims to investigate the emotional states via speech signals (Wani et al., 2021). Building on this, SER systems extract and classify prominent speech features from a preprocessed speech signal (Wani et al., 2021). Theoretically, based on the acquired speech-based information, the system can assess the users' emotions and define its actions accordingly (Wani et al., 2021). However, in the context of voice user interfaces, certain systems lack the emotion expressivity in their responses, an element expected by users. To bridge this gap, emotion voice conversion may come into play. Speech emotion conversion seeks to generate expressive speech from neutral synthesized speech or natural human voice (Robinson et al., 2019). Studies regarding emotion voice conversion have observed an array of speech features, including F0 (Raveh et al., 2019; Robinson et al., 2019; Xue et al., 2018). For example, in a study by Robinson et al. (2019), a sequence-to-sequence architecture for speech emotion conversion was designed using F0 values. The aim of the study was to test the effectiveness of this architecture in transforming the intonation of a human voice, from a neutral to expressive speech. Comparably, research by Xue et al. (2018) proposed a voice conversion system for emotion that allowed for neutral speech to be transformed into emotional speech, with dimensions valence and arousal serving as a control to the degrees of emotion. The acoustic features utilized in this study were F0, power envelope, spectral sequency and duration.

Similarly to the studies previously mentioned, research by Zhu & Ahmad (2019) utilized speech features as a means to test a system. Within this study, results from a SER model

featuring both spectral and prosodic features were analyzed in order to assess its recognition accuracy upon a Chinese emotional speech database. The studied features included spectral centroid, spectral crest, spectral decrease, spectral entropy, spectral flatness, spectral flux, spectral kurtosis, spectral roll-off point, spectral spread, spectral slope, spectral skewness, alongside prosodic features of energy and pitch.

Unlike the previous examples, a study by Kohh & Kwahk (2017) assessed the emotions of voice user interface users directly. To do so, speech features including amplitude, pitch, and duration were analyzed to investigate users' speech behaviour patterns during voice user interface usage. More precisely, speech patterns were observed during responses following errors produced by iPhone's Siri. As stressed by the authors, few studies have sought to investigate users' speech behaviour patterns while using a voice user interface.

Although the emotions of users were not observed directly, research by Raveh et al. (2019) studied the difference between human-directed speech (HDS) and device-directed speech in human-human-computer interactions by observing speech features, including F0, intensity and articulation rate. Results differ between both types of interactions, revealing disparate speech behaviours when interacting with device-directed speech. A summary of the studies mentioned may be found in the Table 2 below, in which the study's topic and extracted speech features are outlined.

**Table 2:** Summary of HCI study examples utilizing speech features

Research	Speech Features	Topic
Robinson et al., 2019	F0	Investigated the effectiveness of a sequence-to-sequence (seq2seq) encoder-decoder based model to transform voice intonation from neutral to expressive speech.

Research	Speech Features	Topic
Xue et al., 2018	F0 Power Envelope Spectral Sequency Duration	Proposed a voice conversion system for emotion capable of converting neutral speech to emotional speech using dimensional space (arousal and valence) as controls of the degree of emotion.
Raveh et al., 2019	F0 Intensity Articulation Rate	Studied the difference of phonetic features between human-directed speech (HDS) and device-directed speech (DDS) in human-computer interactions.
Koh & Kwahk, 2017	Amplitude Pitch Duration	Investigated the speech behaviour patterns of inexperienced iPhone Siri users following error correction.
Zhu & Ahmad, 2019	Spectral Centroid Spectral Crest Spectral Decrease Spectral Entropy Spectral Flatness Spectral Flux Spectral Kurtosis Spectral roll-off point Spectral skewness Spectral slope Spectral spread Energy Pitch	Proposed one of two Speech emotion recognition (SER) frameworks used to assess its recognition accuracy upon the Chinese emotional speech database.

Voice expression is often reflected by physiological changes associated to the speaker's emotional state (Juslin & Laukka, 2003; Scherer, 1986). Similarly to the physiological measures described previously, the study vocal expression can also revolve around emotion dimensions of valence and arousal (Bachorowski, 1999; Scherer, 1986). Within literature, arousal is the most studied dimension in relation to vocal expression (Laukka et al., 2005). Indeed, research has contributed to the idea that the emotional arousal of a speaker is accompanied by physiological changes, consequently affecting respiration, phonation, and articulation resulting in emotion-specific patterns of acoustic parameters (Scherer, 1986). More precisely, research has indicated that high arousal is associated to factors such as high mean F0, fast speech rate, and increased high frequency energy (Breitenstein et al., 2001; Davitz, 1964; Levin & Lord, 1975; Pereira, 2000; Schröder et al., 2001; Apple et al., 1979; Kehrein, 2002; Scherer & Oshinsky, 1977; Pittam et al., 1990). Across studies, results regarding arousal remain consistent (Laukka et al., 2005).

Unlike the findings concerning arousal, results regarding valence are noticeably inconsistent. In some studies, positive valence has been linked to low mean F0, fast speech rate, F0 variability and little high-frequency energy (Scherer & Oshinsky, 1977; Scherer, 1974; Scherer & Oshinsky, 1977; Uldall, 1960; Pittam et al., 1990; Schröder et al., 2001). In others, valence is not associated to specific patterns of vocal cues (e.g., Apple et al., 1979; Davitz, 1964; Pereira, 2000).

Emotional expressions through vocal cues, as well as corresponding abilities to perceive emotions, are fundamental aspects of human communication (Bachorowski, 1999). Studies seeking to characterize acoustic properties of emotional speech have shown that such properties provide an external cue to the arousal associated within emotional processes, in addition to the relative pleasantness of experienced emotion (Bachorowski, 1999). Consequently, as shown in perceptual tests, listeners have the ability to accurately judge emotions of speech (Bachorowski, 1999). Recent studies have suggested that observers can recognize at least 13 different emotions in brief vocalizations (Cowen et al., 2019). Contextual factors are also important to the perception of emotional expression (Kleinsmith & Bianchi-Berthouze, 2013). In a study by Gendron et al. (2012), it was suggested that language, a contextual cue, shapes the way we interpret emotional

expression. Thus, contextual factors are important to the perception of emotional expression (Kleinsmith & Bianchi-Berthouze, 2013)

Within the context of voice user interface technology, understanding the semantic cues leading to suboptimal interactions can be particularly informative. To better understand induced emotional behaviors, third-party observers can be employed. Seeking impartial individuals to assess emotional behaviors is a common practice within the field of psychology (John & Robins, 1994). The impartiality of third-party observers reduces the bias often revolving around a user's ability to objectively evaluate his or her behavior (Robins & John, 1997). It is for this reason judgement studies were initially developed (Robins & John, 1997). In addition to limiting biasness, observers may offer a greater level of precision in differentiating behavioral categories (Borkenau & Ostendorf, 1987). Moreover, research has suggested that humans have the ability to process fleeting emotions (Sweeny et al., 2013). With this said, when establishing the ground-truth of an emotion expression, observer coders are often employed (Kleinsmith & Bianchi-Berthouze, 2013).

Research has shown that perceived emotions can be consistent with physiological responses. For example, in a study by Ortiz de Guinea et al. (2013), convergent validity of arousal was evidenced by the significant correlation between the SAM scale measure and electrodermal activity. Moreover, in a study by Le Pailleur et al. (2020), self-perceived arousal was consistent with the psychophysiological responses measured using electrodermal activity in user interactions with a voice assistant, depicting a significant positive correlation. Within this same study, a correlated relationship between AFE and valence is noted, as the emotions inferred by the facial expression analysis was complimentary to the self-perceived emotional valence reported by users. Despite the fact that this particular study dealt with self-perceived emotions, it can be argued that the results are similar for third-party observed emotions. Studies have noted the accuracy of self-reports in comparison to evaluations by peers (John & Robins, 1994; Kolar et al., 1996). According to a study by Kolar et al. (1996), the predictive validity of the observers' judgements outperformed self-perceived evaluations. Moreover, spontaneous facial behaviors can also be assessed by observers. As seen in a study

by Sweeny et al. (2013), observers were able to detect and classify emotions via facial expressions at a mere 10 ms of exposure. Hence, fleeting emotions can also be potentially detected by observers. It is to be said that autonomic responses, such as EDA, are expected to occur with spontaneous facial behaviour (Kreibig, 2010). As for the control dimension, there is a link to be made between perceived valence. As seen in a study by Sutton et al (2019), observers who rated the emotions of various facial expressions using a SAM scale noted that happier looking faces were rated higher in dominance than neutral faces, suggesting that facial expressions can also be distinguished based on dominance.

Despite their observable nature, expressions of emotion can be brief states (Sweeny et al., 2013). Within the Trigger-Substrate model (Merchant & Armoundas, 2012; Witchel et al., 2018), triggers are proximate causes for short-lived, unpredictable and idiosyncratic events, whereas permissive states are considered medium-term, predictable and measurable. As explained by Witchel et. al (2018), “in human-computer interaction, a substrate would be a mood, while a trigger could either be a computer event or an end user’s passing thought” (p.2). Within research conducted by Withchel et al. (2018), smile rates were observed in concurrence with time in order to assess fleeting emotions. Indeed, users may experience changes in physiological arousal but manage to largely conceal this change by controlling their speech, facial expression, or physical behavior. In other words, a fleeting short-term change in speech, facial expression or physical behavior, could potentially represent a metaphorical iceberg of underlying emotional arousal.

## **2.7 Summary of literature review**

In sum, the literature review depicts the various manners in which emotions are captured in HCI studies. A predominant measure is the self-perceived SAM scale, in which dimensions valence, arousal, and control are assessed. Although originally intended as a self-perceived measure, studies have shown that third-party observers can accurately assess emotions using this scale. However, solely utilizing explicit measures fails to provide a complete and thorough understanding of emotions. The use of implicit measures, such as facial expression, electrodermal activity and speech, can be used to assess the emotions induced by voice user interface interactions, as they offer accurate,

real-time data depicting distinct emotional states. Moreover, the simultaneous study of speech alongside physiological measures results in an opportunity to address a gap within literature surrounding voice user interface evaluation, as the use of implicit methods is employed. Indeed, the comparative effectiveness of speech, EDA, and facial expression analysis via their respective features in explaining intense emotional events induced by voice user interface interactions has yet to be studied, despite its important potential to contribute to voice user interface evaluation within industry. Within the same vein, employing other implicit methods, such as electrocardiogram (ECG) or electroencephalogram (EEG), can further contribute to the literature and therefore is worthy of future research.

## References

- Agarwal, A., & Meyer, A. (2009). Beyond usability: evaluating emotional response as an integral part of the user experience. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*.
- Agourram, H., Alvarez, J., Sénécal, S., Lachize, S., Gagné, J., & Léger, P. M. (2019). The relationship between technology self-efficacy beliefs and user satisfaction–user experience perspective. *International Conference on Human-Computer Interaction* (pp. 389-397). Springer, Cham.
- Alaszewski, A. (2006). *Using diaries for social research*. SAGE Publications Ltd  
<https://www.doi.org/10.4135/9780857020215>
- Alves, R., Valente, P., & Nunes, N. J. (2014). The state of user experience evaluation practice. *Proceedings of the NordiCHI 2014: The 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, 93–102.  
<https://doi.org/10.1145/2639189.2641208>
- Andreassi, J. L. (2000). *Psychophysiology: Human behavior and physiological response* (4th ed.). Lawrence Erlbaum Associates Publishers.
- Apple, W., Streeter, L. A., & Krauss, R. M. (1979). Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology*, 37, 715-727.
- Arora, S., Baghai-Ravary, L., & Tsanas, A. (2019). Developing a large scale population screening tool for the assessment of Parkinson's disease using telephone-quality voice. *The Journal of the Acoustical Society of America*, 145(5), 2871-2884.

- Bachorowski, J.A. (1999). Vocal Expression and Perception of Emotion. *Current Directions in Psychological Science*, 8(2), 53–57. <https://doi.org/10.1111/1467-8721.00013>
- Battarbee, K., & Koskinen, I. (2005). Co-experience: user experience as interaction. *CoDesign*, 1(1), 5–18. <https://doi.org/10.1080/15710880412331289917>
- Beauchesne, A., Sénécal, S., Fredette, M., Chen, S. L., Demolin, B., Di Fabio, M. L., & Léger, P. M. (2019, July). User-centered gestures for mobile phones: exploring a method to evaluate user gestures for UX designers. In *International Conference on Human-Computer Interaction* (pp. 121-133). Springer, Cham.
- Beaudry, A., & Pinsonneault, A. (2010). The other side of acceptance: Studying the direct and indirect effects of emotions on information technology use. *MIS quarterly*, 689-710.
- Bentley, T., Johnston, L., & von Baggo, K. (2005, November). Evaluation using cued-recall debrief to elicit information about a user's affective experiences. In *Proceedings of the 17th Australia Conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future* (pp. 1-10).
- Betella, A., & Verschure, P. F. (2016). The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions. *PloS one*, 11(2), e0148037. <https://doi.org/10.1371/journal.pone.0148037>
- Bethel, C. L., Salomon, K., Murphy, R. R., & Burke, J. L. (2007). Survey of psychophysiology measurements applied to human-robot interaction. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 732-737). IEEE.
- Black, A. (1998). Empathic design: User focused strategies for innovation. In *Proceedings of the Conference on New Product Development* (pp. 1-8). London, UK: IBC
- Borkenau, P., & Ostendorf, F. (1987). Retrospective Estimates of Act Frequencies: How Accurately Do They Reflect Reality? *Journal of Personality and Social Psychology*, 52(3), 626-638. <https://doi.org/10.1037/0022-3514.52.3.626>
- Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (Vol. 30, No. 1, pp. 25-36). Technical report C-1, the center for research in psychophysiology, University of Florida.
- Bradley MM, Lang PJ. (1994) Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*. Mar; 25(1):49–59. [https://doi.org/10.1016/0005-7916\(94\)90063-9](https://doi.org/10.1016/0005-7916(94)90063-9)



- Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2013). A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology*, 49(1), 1017-1034.
- Brave, S., & Nass, C. (2002). Emotion in human-computer interaction. In *The human-computer interaction handbook* (pp. 53-68). CRC Press.  
<https://doi.org/10.1201/b10368-6>
- Breitenstein, C., Van Lancker, D., Daum, I. (2001). The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. *Cognition and Emotion*, 15 (1), 57–79
- Bruun, A., & Ahm, S. (2015). Mind the gap! Comparing retrospective and concurrent ratings of emotion in user experience evaluation. In *IFIP Conference on Human-Computer Interaction* (pp. 237-254). Springer, Cham.
- Buchenau, M., Fulton Suri, J. (2000). Experience prototyping. In *Proceedings of DIS 2000 (Designing Interactive Systems)*, 424–433.
- Burton-Jones, A., & Gallivan, M. J. (2007). Towards a deeper understanding of system usage in organizations. *MIS Quarterly*, 31(4), 657–679.
- Cacioppo, J. T., Berntson, G. G., Larsen, J. T., Poehlmann, K. M., & Ito, T. A. (2000). The psychophysiology of emotion. *Handbook of emotions*, 2(01). 173-191.
- Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., ... & R Cowan, B. (2019). The state of speech in HCI: Trends, themes and challenges. *Interacting with Computers*, 31(4), 349-371.
- Cohen, M. H., Cohen, M. H., Giangola, J. P., & Balogh, J. (2004). *Voice user interface design*. Addison-Wesley Professional.
- Cordaro, D. T., Keltner, D., Tshering, S., Wangchuk, D., & Flynn, L. M. (2016). The voice conveys emotion in ten globalized cultures and one remote village in Bhutan. *Emotion*, 16(1), 117–128. <https://doi.org/10.1037/emo0000100>
- Cowen, A. S., Elfenbein, H. A., Laukka, P., & Keltner, D. (2019). Mapping 24 emotions conveyed by brief human vocalization. *American Psychologist*, 74(6), 698–712.  
<https://doi.org/10.1037/amp0000399>
- Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York, NY: Grosset/Putnam.
- Dandavate, U., Sanders, E.B.-N. and Stuart, S., (1996). Emotions matter: user empathy in the product development process, in *Proceedings of the Human Factors and*

*Ergonomics Society 40th Annual Meeting*, 415–418.

Davitz, J. R. (Ed.). (1964). *The communication of emotional meaning*. McGraw Hill.

Dawson, M. E., Schell, A. M., & Filion, D. L. (2007). The electrodermal system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 159–181). Cambridge University Press. <https://doi.org/10.1017/CBO9780511546396.007>

Dewey, J. (1934). The Supreme Intellectual Obligation. *Science*, 79 (2046), 240–243. <https://doi.org/10.1002/sce.3730180102>

De Singly, F. (2016). *Le questionnaire* (4e édition). Armand Colin.

Dirican, A. C., & Göktürk, M. (2011). Psychophysiological Measures of Human Cognitive States Applied in Human Computer Interaction. *Procedia Computer Science*, 3, 1361- 1367. <https://doi.org/10.1016/j.procs.2011.01.016>

Easwara Moorthy, A., & Vu, K.-P. L. (2015). Privacy Concerns for Use of Voice Activated Personal Assistant in the Public Space. *International Journal of Human-Computer Interaction*, 31(4), 307–335.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4), 169-200.

Ekman, P., & Friesen, W. V. (1978). *The facial action coding system*. San Fransisco, CA: Consulting Psychologists Press.

Fan, M., Shi, S., & Truong, K. N. (2020). Practices and challenges of using think-aloud protocols in industry: An international survey. *Journal of Usability Studies*, 15(2), 85–102.

Feldman Barrett, L., & Russell, J. A. (1998). Independence and bipolarity in the structure of current affect. *Journal of personality and social psychology*, 74(4), 967. <https://doi.org/10.1037/0022-3514.74.4.967>

Forlizzi, J., & Battarbee, K. (2004). Understanding experience in interactive systems. In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques* (pp. 261-268). <https://doi.org/10.1145/1013115.1013152>

Garg, R., & Moreno, C. (2019). Exploring Everyday Sharing Practices of Smart Speakers. In *IUI Workshops*.

Ganglbauer, E., Schrammel, J., Deutsch, S., Tscheligi, M. (2009). Applying Psychophysiological Methods for Measuring User Experience: Possibilities,

Challenges and Feasibility. *Workshop on User Experience Evaluation Methods in Product Development*.

- Gendron, M., Lindquist, K. A., Barsalou, L., & Barrett, L. F. (2012). Emotion words shape emotion percepts. *Emotion*, 12(2), 314.
- Gentile, C., Spiller, N., Noci, G. (2007). How to Sustain the Customer Experience: An Overview of Experience Components that Co-create Value With the Customer. *European Management Journal*, 25, 395-410.
- Grimm, M., Kroschel, K., Mower, E., & Narayanan, S. (2007). Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10–11), 787–800. <https://doi.org/10.1016/j.specom.2007.01.010>
- Gu, S., Wang, F., Patel, N. P., Bourgeois, J. A., & Huang, J. H. (2019). A model for basic emotions using observations of behavior in *Drosophila*. *Frontiers in Psychology*, 10(APR), 1–13. <https://doi.org/10.3389/fpsyg.2019.00781>
- Hartmann, K., Siegert, I., Philippou-Hübner, D., & Wendemuth, A. (2013). Emotion detection in HCI: From speech features to emotion space? *IFAC Proceedings Volumes (IFAC-PapersOnline)*, 12(PART 1), 288–295. <https://doi.org/10.3182/20130811-5-US-2037.00049>
- Hassenzahl, M. (2008). Aesthetics in Interactive Products: Correlates and Consequences of Beauty. *Elsevier*, 1, 287-302.
- Ittelson, W. H., et al. (1970). The use of behavioural maps in environmental psychology. In H. M. Prohansky, W. H. Ittelson, L. G. Rivlin (Eds.), *Environmental Psychology: Man and his Physical Setting*, Holt (pp. 658-668). New York: Rinehart & Winston.
- ISO FDIS 9241-210 (2018) *Human-centred design process for interactive systems*. | ISO. <https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-1:v1:en>. (Accessed 2 July 2021)
- Jessen, S., & Kotz, S. A. (2011). The temporal dynamics of processing emotions from vocal, facial, and bodily expressions. *NeuroImage*, 58(2), 665–674. <https://doi.org/10.1016/j.neuroimage.2011.06.035>
- John, O. P., & Robins, R. W. (1994). Accuracy and bias in self-perception: individual differences, self- enhancement and the role of narcissism. *Journal of Personality and Social Psychology*, 66, 206–219.
- Johnstone, T., & Scherer, K. R. (2000). Vocal communication of emotion. *Handbook of emotions*, 2, 220-235
- Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129,

770–814. <http://dx.doi.org/10.1037/0033-2909.129.5.770>

- Kankainen, A. (2002). *Thinking model and tools for understanding user experience related to information appliance product concepts*. Helsinki University of Technology.
- Keltner, D., Tracy, J. L., Sauter, D., & Cowen, A. (2019). What Basic Emotion Theory Really Says for the Twenty-First Century Study of Emotion. *Journal of nonverbal behavior*, 43(2), 195–201. <https://doi.org/10.1007/s10919-019-00298-y>
- Kent, R.D. (1997). *The speech sciences*. Singular Publishing.
- Kehrein, R. (2002). The prosody of authentic emotions. In *Speech Prosody 2002, International Conference*. DOI:[10.1055/s-2003-40251](https://doi.org/10.1055/s-2003-40251)
- Kleinsmith, A., & Bianchi-Berthouze, N. (2012). Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4(1), 15-33.
- Kocaballi, A. B., Laranjo, L., & Coiera, E. (2019). Understanding and Measuring User Experience in Conversational Interfaces. *Interacting with Computers*, 31(2), 192–207. <https://doi.org/10.1093/iwc/iwz01>
- Koh, Y., & Kwahk, J. (2017). B3-1 Analysis of User's Speech Behavior Pattern after Correction: focusing on Smartphone Voice User Interface. *The Japanese Journal of Ergonomics*, 53, 408-411.
- Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the Accuracy of Personality Judgments by the Self and Knowledgeable Others. *Journal of Personality*, 64(2), 311-337. doi:10.1111/j.1467-6494.1996.tb00513.x
- Kollia, V. (2016). Personalization Effect on Emotion Recognition from Physiological Data: An Investigation of Performance on Different Setups and Classifiers. *ArXiv, abs/1607.05832*.
- Kraus, M. W. (2017). Voice-only communication enhances empathic accuracy. *American Psychologist*, 72, 644–654. <http://dx.doi.org/10.1037/amp0000147>
- Kreibig S. D. (2010). Autonomic nervous system activity in emotion: a review. *Biological psychology*, 84(3), 394–421. <https://doi.org/10.1016/j.biopsycho.2010.03.010>
- Krosnick, J. A. (1999). Survey Research. *Annual Review of Psychology*, 50, 537-567.
- Lamontagne, C., Sénécal, S., Fredette, M., Chen, S. L., Pourchon, R., Gaumont, Y., ... & Léger, P. M. (2019, August). User Test: How Many Users Are Needed to Find the Psychophysiological Pain Points in a Journey Map? In *International Conference on*

- Human Interaction and Emerging Technologies* (pp. 136-142). Springer, Cham.
- Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1998). Emotion, motivation, and anxiety: brain mechanisms and psychophysiology. *Biological psychiatry*, 44(12), 1248–1263. [https://doi.org/10.1016/s0006-3223\(98\)00275-3](https://doi.org/10.1016/s0006-3223(98)00275-3)
- Laukka, P., Elfenbein, H. A., Thingujam, N. S., Rockstuhl, T., Iraki, F. K., Chui, W., & Althoff, J. (2016). The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features. *Journal of Personality and Social Psychology*, 111, 686–705. <http://dx.doi.org/10.1037/pspi0000066>
- Laukka, P., Juslin, P. N., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition and Emotion*, 19(5), 633–653. <https://doi.org/10.1080/02699930441000445>
- Lausen, A., & Hammerschmidt, K. (2020). Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. *Humanities and Social Sciences Communications*, 7(1), 1-17.
- Léger, P.-M., Davis, F. D., Cronan, T. P., & Perret, J. (2014). Neurophysiological Correlates of Cognitive Absorption in an Enactive Training Context. *Computers in Human Behavior*, 34, 273-283. doi:10.1016/j.chb.2014.02.011
- Léger, P.-M., Davis, F. D., Cronan, T. P., & Perret, J. (2014). Neurophysiological Correlates of Cognitive Absorption in an Enactive Training Context. *Computers in Human Behavior*, 34, 273-283. <https://doi.org/10.1016/j.chb.2014.02.011>
- Leite, I., Henriques, R., Martinho, C., & Paiva, A. (2013). Sensors in the wild: Exploring electrodermal activity in child-robot interaction. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 41-48). IEEE.
- Le Pailleur, F., Huang, B., Léger, P. M., & Sénécal, S. (2020). A new approach to measure user experience with voice-controlled intelligent assistants: A pilot study. In M. Kurosu (Ed.), *Human-computer interaction. Multimodal and natural interaction. HCII 2020. Lectures notes in computer science* (Vol. 12182, pp. 197–208). [https://doi.org/10.1007/978-3-030-49062-1\\_13](https://doi.org/10.1007/978-3-030-49062-1_13)
- Levenson, R. W. (2014). The autonomic nervous system and emotion. *Emotion Review*, 6(2), 100–112. <https://doi.org/10.1177/1754073913512003>
- Levin, H., & Lord, W. (1975). Speech pitch frequency as an emotional state indicator. *IEEE Transactions on Systems, Man, and Cybernetics*, 5, 259-273.
- Li S.Z., Jain A. (2009). Fundamental Frequency, Pitch, F0. *Encyclopedia of Biometrics*. Springer, Boston, MA. [https://doi.org/10.1007/978-0-387-73003-5\\_775](https://doi.org/10.1007/978-0-387-73003-5_775)

- Liebermann, M. D. (2007). Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology*, 58, 259–289.
- Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE transactions on bio-medical engineering*, 56(4), 1015. <https://doi.org/10.1109/TBME.2008.2005954>
- Lopatovska, I., & Oropeza, H. (2018). User interactions with “Alexa” in public academic space. *Proceedings of the Association for Information Science and Technology*, 55(1), 309–318. <https://doi.org/10.1002/pra2.2018.14505501034>
- Lourties, S., Léger, P. M., Sénécal, S., Fredette, M., & Chen, S. L. (2018). Testing the convergent validity of continuous self-perceived measurement systems: an exploratory study. In *International Conference on HCI in Business, Government, and Organizations* (pp. 132-144). Springer, Cham.
- Maia, C. L. B., & Furtado, E. S. (2016). A study about psychophysiological measures in user experience monitoring and evaluation. In *Proceedings of the 15th Brazilian Symposium on Human Factors in Computing Systems* (pp. 1-9). <https://doi.org/10.1145/3033701.3033708>
- Mäkelä, A., & Fulton Suri, J. (2001, June). Supporting users’ creativity: Design to induce pleasurable experiences. In *Proceedings of the International Conference on Affective Human Factors Design* (pp. 387-394).
- Mannepalli, K., Sastry, P. N., & Suman, M. (2018). Emotion recognition in speech signals using optimization based multi-SVNN classifier. *Journal of King Saud University-Computer and Information Sciences*. <https://doi.org/10.1016/j.jksuci.2018.11.012>
- Maunier, B., Alvarez, J., Léger, P. M., Sénécal, S., Labonté-LeMoyne, É., Chen, S. L., ... & Gagné, J. (2018). Keep calm and read the instructions: factors for successful user equipment setup. In *International Conference on HCI in Business, Government, and Organizations* (pp. 372-381). Springer, Cham.
- Merchant, F. M., & Armoundas, A. A. (2012). Role of substrate and triggers in the genesis of cardiac alternans, from the myocyte to the whole heart: implications for therapy. *Circulation*, 125(3), 539–549. <https://doi.org/10.1161/CIRCULATIONAHA.111.033563>
- Mourra, G. N., Senecal, S., Fredette, M., Lepore, F., Faubert, J., Bellavance, F., ... & Léger, P. M. (2020). Using a smartphone while walking: The cost of smartphone-addiction proneness. *Addictive behaviors*, 106, 106346. <https://doi.org/10.1016/j.addbeh.2020.106346>

- Nass, C., Jonsson, I. M., Harris, H., Reaves, B., Endo, J., Brave, S., & Takayama, L. (2005). Improving automotive safety by pairing driver emotion and car voice emotion. *Conference on Human Factors in Computing Systems - Proceedings*, 1973–1976. <https://doi.org/10.1145/1056808.1057070>
- Ng, C.F. (2016). Behavioral Mapping and Tracking. In *Research Methods for Environmental Psychology*, R. Gifford (Ed.). <https://doi.org/10.1002/9781119162124.ch3>
- Nicholl H. (2010). Diaries as a method of data collection in research. *Paediatric nursing*, 22(7), 16–20. <https://doi.org/10.7748/paed2010.09.22.7.16.c7948>
- Ortiz de Guinea, A., Titah, R., & Léger, P.-M. (2014). Explicit and implicit antecedents of users' behavioral beliefs in information systems: A neuropsychological investigation. *Journal of Management Information Systems*, 30(4), 179–210.
- Ortiz de Guinea, A., & Webster, J. (2013). An investigation of information systems use patterns: Technological events as triggers, the effect of time, and consequences for performance. *MIS Quarterly*, 37, 1165–1188.
- Ortiz de Guinea, A., & Markus, M. L. (2009). Why break the habit of a lifetime? Rethinking the roles of intention, habit, and emotion in continuing information technology use. *MIS Quarterly*, 33, 433–444.
- Owren, M. J., & Bachorowski, J. A. (2007). Measuring emotion-related vocal acoustics. *Handbook of emotion elicitation and assessment*, 239–266.
- Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. Oxford University Press.
- Papakostas, M., Siantikos, G., Giannakopoulos, T., Spyrou, E., & Sgouropoulos, D. (2017). Recognizing emotional states using speech information. In *GeNeDis 2016* (pp. 155–164). Springer, Cham.
- Park, J., Han, S. H., Kim, H. K., Cho, Y., & Park, W. (2013). Developing elements of user experience for mobile phones and services: Survey, interview, and observation approaches. *Human Factors and Ergonomics In Manufacturing*, 23(4), 279–293. <https://doi.org/10.1002/hfm.20316>
- Paul, C. L., & Komlodi, A. (2014). Measuring user experience through future use and emotion. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems* (pp. 2503–2508).
- Paul, C., & Komlodi, A. (2012). Emotion as an indicator for future interruptive notification experiences. In *CHI'12 Extended Abstracts on Human Factors in*

*Computing Systems* (pp. 2003-2008).

Picard, R. W. (2000). *Affective computing*. MIT press.

Piedmont, R. L. (2014). Social Desirability Bias. *Encyclopedia of Quality of Life and Well-Being Research*, 6036–6037. [https://doi.org/10.1007/978-94-007-0753-5\\_2746](https://doi.org/10.1007/978-94-007-0753-5_2746)

Pittam, J., Gallois, C., & Callan, V. (1990). The long-term spectrum and perceived emotion. *Speech Communication*, 9, 177-87.

Plutchik, R. (1962). *The emotions: Facts, theories and a new model*. Crown Publishing Group/Random House.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3), 715–734. <https://doi.org/10.1017/S0954579405050340>

Pereira, C. (2000). Dimensions of emotional meaning in speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Provine, R. R., & Fischer, K. R. (1989). Laughing, smiling, and talking: Relation to sleeping and social context in humans. *Ethology*, 83, 295– 305.

Raveh, E., Steiner, I., Siegert, I., Gessinger, I., & Möbius, B. (2019). Comparing phonetic changes in computer-directed and human-directed speech. *Studentexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, 42-49.

Riedl, R., & Léger, P. M. (2016). Fundamentals of NeuroIS. *Studies in neuroscience, psychology and behavioral economics*, 127. <https://doi.org/10.1007/978-3-662-45091-8>

Robins, R. W., & John, O. P. (1997). Effects of Visual Perspective and Narcissism on Self-Perception: Is Seeing Believing? *Psychological Science*, 8(1), 37-42.

Robins, R. W., & John, O. P. (1997). The Quest for Self-Insight: Theory and Research on Accuracy and Bias in Self-Perception. In N. Y. A. Press (Ed.), *Handbook of Personality Psychology* (pp. 649-679).

Robinson, C., Obin, N., & Roebel, A. (2019). Sequence-to-sequence modelling of f0 for speech emotion conversion. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6830-6834). IEEE.

Rowe, D.W., Sibert, J.L., & Irwin, D. (1998). Heart rate variability: indicator of user state as an aid to human-computer interaction. *Proceedings of the SIGCHI*



- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145–172. <https://doi.org/10.1037/0033-295X.110.1.145>
- Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76(5), 805–819. <https://doi.org/10.1037/0022-3514.76.5.805>
- Russell, J.A. (1980) A circumplex model of affect. *Journal of Personality and Social Psychology*, 39, 1161–1178.
- Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1–2), 227–256. [https://doi.org/10.1016/S0167-6393\(02\)00084-5](https://doi.org/10.1016/S0167-6393(02)00084-5)
- Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99, 143–165.
- Scherer, K. R., & Oshinsky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion*, 1, 331–346.
- Scherer, K. R. (1974). Acoustic concomitants of emotional dimensions: Judging affect from synthesized tone sequences. In S. Weitz (Ed.), *Nonverbal communication* (pp. 105–111). New York: Oxford University Press
- Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., & Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. In *Seventh European Conference on Speech Communication and Technology*.
- Sciuto, A., Saini, A., Forlizzi, J., & Hong, J. I. (2018). “Hey Alexa, What’s Up?” *Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS ’18*. <https://doi.org/10.1145/3196709.3196772>.
- Shih, Y.-H., & Liu, M. (2007). The Importance of Emotional Usability. *Journal of Educational Technology Systems*, 36(2), 203–218. <https://doi.org/10.2190/ET.36.2.h>
- Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., Xu, X., & Yang, X. (2018). A Review of Emotion Recognition Using Physiological Signals. *Sensors*, 18(7), 2074. <https://doi.org/10.3390/s18072074>
- Stickel, C., Ebner, M., Steinbach-Nordmann, S., Searle, G., & Holzinger, A. (2009). Emotion detection: application of the valence arousal space for rapid biological usability testing to enhance universal access. In *International Conference on*

*Universal Access in Human-Computer Interaction* (pp. 615-624). Springer, Berlin, Heidelberg.

Sutton, T. M., Herbert, A. M., & Clark, D. Q. (2019). Valence, arousal, and dominance ratings for facial stimuli. *Quarterly Journal of Experimental Psychology*, 72(8), 2046–2055. <https://doi.org/10.1177/1747021819829012>

Sweeny, T. D., Suzuki, S., Grabowecky, M., & Paller, K. A. (2013). Detecting and categorizing fleeting emotions in faces. *Emotion*, 13(1), 76–91. <https://doi.org/10.1037/a0029193>

Tahon, M., Degottex, G., & Devillers, L. (2012). Usual voice quality features and glottal features for emotional valence detection. *Proceedings of the 6th International Conference on Speech Prosody*, 2, 693–697.

Thayer, R. E. (1989). *The biopsychology of mood and arousal*. New York: Oxford University Press.

Titze, I. R., & Martin, D. (1998). Principles of voice production. *The Journal of the Acoustical Society of America*, 104, 1148. <http://dx.doi.org/10.1121/1.424266>

Toh, A. M., Togneri, R., & Nordholm, S. (2005). Spectral entropy as speech features for speech recognition. *Proceedings of PEECS*, 1, 92.

Tomkins, S. (1962). *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company.

Tomkins, S. (1963). *Affect imagery consciousness: Volume II: The negative affects*. Springer publishing company.

Uldall, E. (1960). Attitudinal meanings conveyed by intonation contours. *Language and Speech*, 3, 223-234.

Väänänen-Vainio-Mattila, K., Roto, V., & Hassenzahl, M. (2008). Towards Practical User Experience Evaluation Methods. *Proceedings of the International Workshop on Meaningful Measure: Valid Useful User Experience Measurement (VUUM 2008)*, 19–22.

Vermeeren, A. P. O. S., Law, E. L. C., Roto, V., Obrist, M., Hoonhout, J., & Väänänen-Vainio-Mattila, K. (2010). User experience evaluation methods: Current state and development needs. *NordiCHI 2010: Extending Boundaries - Proceedings of the 6th Nordic Conference on Human-Computer Interaction*, 521–530. <https://doi.org/10.1145/1868914.1868973>

Vidrascu, L., & Devillers, L. (2005). Real-life emotion representation and detection in call centers data. In *International Conference on Affective Computing and*

- Intelligent Interaction* (pp. 739-746). Springer, Berlin, Heidelberg.
- vom Brocke, J., Riedl, R., & Léger, P.-M. (2013). Application strategies for neuroscience in information systems design science research. *Journal of Computer Information Systems*, 53(3), 1-13.
- Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021). A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access*, 9, 47795–47814. <https://doi.org/10.1109/ACCESS.2021.3068045>
- Ward, R. D., & Marsden, P. H. (2003). Physiological Responses to Different Web Page Designs. *International Journal of Human-Computer Studies*, 59(1-2), 199-212. [https://doi.org/10.1016/S1071-5819\(03\)00019-3](https://doi.org/10.1016/S1071-5819(03)00019-3)
- Watson, D., & Clark, L. A. (1992). On traits and temperament: general and specific factors of emotional experience and their relation to the five-factor model. *Journal of personality*, 60(2), 441–476. <https://doi.org/10.1111/j.1467-6494.1992.tb00980.x>
- Weiser, M. (1991). The computer for the 21st century. *Scientific American*, 265(3), 75-84.
- Witchel, H. J., Claxton, H. L., Holmes, D. C., Ranji, T. T., Chalkley, J. D., Santos, C. P., Westling, C. E. I., Valstar, M. F., Celuszak, M., & Fagan, P. (2018). A trigger-substrate model for smiling during an automated formative quiz: Engagement is the substrate, not frustration. *ACM International Conference Proceeding Series*. <https://doi.org/10.1145/3232078.3232084>
- Wood, S. L., & Moreau, C. P. (2006). From Fear to Loathing? How Emotion Influences the Evaluation and Early Use of Innovations. *Journal of Marketing*, 70(3), 44–57. <https://doi.org/10.1509/jmkg.70.3.044>
- Wrigley, C., Gomez, R., & Popovic, V. (2010). The evaluation of qualitative methods selection in the field of design and emotion. In *Proceedings of the 7th International Conference on Design and Emotion 2010* (pp. 1-12). IIT Institute of Design
- Xue, Y., Hamada, Y., & Akagi, M. (2018). Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space. *Speech Communication*, 102(June), 54–67. <https://doi.org/10.1016/j.specom.2018.06.006>
- Zaman, B., & Shrimpton-Smith, T. (2006). The FaceReader: Measuring instant fun of use. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles* (pp. 457-460). <https://doi.org/10.1145/1182475.1182536>
- Zhu, C., & Ahmad, W. (2019). Emotion recognition from speech to improve human-robot interaction. *Proceedings - IEEE 17th International Conference on*

*Dependable, Autonomic and Secure Computing, IEEE 17th International Conference on Pervasive Intelligence and Computing, IEEE 5th International Conference on Cloud and Big Data Computing, 4th Cyber Scienc, July, 370–375.*  
<https://doi.org/10.1109/DASC/PiCom/CBDCCom/CyberSciTech.2019.0007>



# **Chapter 3.**

## **Comparing the effectiveness of speech and physiological features in explaining emotional responses during voice user interface interactions**

Danya Swoboda, Jared Boasen, Pierre-Majorique Léger, Romain Pourchon and Sylvain Sénécal

HEC Montréal, Montréal, Canada

danya.swoboda@hec.ca, jared.boasen@hec.ca, pierre-majorique.leger@hec.ca,  
sylvain.senecal@hec.ca

### **Abstract**

The rapid rise of voice user interface technology has changed the way users traditionally interact with interfaces, as tasks requiring gestural or visual attention are swapped by vocal commands. This shift has equally affected designers, required to disregard common digital interface guidelines in order to adapt to non-visual user interaction (No-UI) ways. The guidelines regarding voice user interface evaluation are far from the maturity of those surrounding digital interface evaluation, resulting in a lack of consensus and clarity. In order to contribute to the emerging literature regarding voice user interface evaluation and consequently assist UX professionals in their quest to create optimal vocal experiences, our study sought to compare the effectiveness of physiological and speech measures through their respective features in explaining emotional events during voice user interface interaction. To do so, we performed a within-subject experiment in which speech, facial expression, and electrodermal activity responses of 16 participants were recorded during voice user interface interactions that were purposely designed to elicit frustration and shock, resulting in 188 analyzed interactions. Our results suggest that the physiological measure of facial expression and its extracted feature, automatic facial expression-based valence, is most informative of emotional events lived through voice user interface interactions. By comparing the unique effectiveness of each feature, theoretical and practical contributions may be noted, as the results contribute to voice user

interface literature while providing key insight favouring efficient voice user interface evaluation.

### **3.1 Introduction**

The history of interface design has primarily revolved around Graphical User Interfaces (GUI), resulting in longstanding and familiar frameworks (Murad & Munteanu, 2020). From Nielsen's 10 usability heuristics to Bastien Scapin's ergonomic criteria for the evaluation of human-computer interfaces, designers have an array of tools to guide them in their conception of optimal digital experiences (Bastien & Scapin, 1992; Nielsen, 1994). With the rise of non-visual user interaction (No-UI), it may be argued that the groundwork for vocal interface design is still in development due to the recency and rapid growth of vocal interface technologies. Indeed, in 2020, 4.2 billion digital voice assistants worldwide were in use (Statista, 2021). By 2024, this number is projected to reach 8.4 billion – a number greater than the world's population (Statista, 2021). With this said, a set of validated voice user interface heuristics and guiding principles have yet to breakthrough.

Research within the field has recently tried to address this matter. For example, Nowacki et al. (2020) developed an adapted version of Bastien Scapin's ergonomic criteria to vocal interfaces. On the other hand, Seaborn & Urakami (2021) presented descriptive frameworks to quantitatively measure voice UX. Both studies relied on extensive reviews of academic and professional guidelines to propose an adapted set of criteria. These studies have contributed to the emerging field of voice user interface design, a discipline in need of support to guide designers in the conceptualization and evaluation of speech-based products. Despite this development, Seaborn & Urakami (2021) highlighted the fact that numerous studies in relation to voice UX rely heavily on self-reported measures based on psychometric scales, and called for the development of measures, such as behavioral measures, to support findings. Indeed, explicit measures, such as self-reported measures, are often adopted due to their inexpensive nature (Alves et al., 2014). However, they are limiting as they do not delve into the real-time automatic and subconscious reactions of users. As a result, UX professionals are at risk of overlooking key insights regarding the underlying emotions of users. Moreover, the limiting nature of certain explicit methods

are made evident when evaluating voice user interfaces, notably the think-aloud method (Hura, 2017). Due to the nature of this method, in which users verbally share their thoughts during interface usage, vocal interference may hamper the user's experience.

To obtain a thorough understanding of the user's lived experiences, implicit measures can be used to observe emotional reactions (Ortiz de Guinea et al., 2014). In the study of emotions during voice user interface interactions, the measure of speech and its features (e.g., pitch, fundamental frequency) is an obvious choice due to the vocal nature of the interaction. However, physiological measures (e.g., electrodermal activity, facial expression) and their respective features have the potential to be equally revealing of emotional events. Studies regarding emotions induced by voice user interface interactions seldom study both speech and physiological features simultaneously. Indeed, voice user interface evaluation often employs explicit methods, such as questionnaires, diaries, interviews, and observations (Clark et al., 2019; Lopatovska & Williams, 2018; Garg & Moreno, 2019; Sciuoto et al., 2018; Lopatovska & Oropeza, 2018). Thus, deviating from the norm of utilizing self-reported measures in addition to utilizing implicit methods to evaluate voice user interface interactions is a rare occurrence. Furthermore, comparing the effectiveness of each implicit method in explaining users' emotions during voice user interface interactions is unique in itself. An opportunity arises to address this important gap within literature while mining key information that may better serve UX professionals. Indeed, by observing the strength in relationship of speech features in parallel to physiological features to assess user emotions during intense voice user interface interactions, we are potentially offering insights that may further help UX practitioners make important decisions within a business context. Obstacle-prone or provocative questioning from voice user interface systems can cause undesirable, intense emotional responses from users, which can derail an optimal experience. Consequently, companies seeking to avoid such responses must first be able to capture them effectively. Limited resources can potentially prevent companies from doing so, as certain select methods may fail to fully reveal the underlying emotions experienced by users. As a result, understanding the strength or effectiveness of speech and physiological measures through their respective features when observing emotional dimensions can help prioritize resources and consequently efficiently evaluate voice user interfaces. To our knowledge,



no other study has sought to compare the effectiveness of speech against physiological features in explaining emotional events provoked by voice user interface interactions. With this said, the central research question of this study is the following:

***RQ1:** Between speech and physiological features, which are more informative in assessing intense emotional responses during vocal interactions with a voice user interface?*

A secondary research question has been posed as the context of this study is unique. Although speech and physiological measures have been widely used in HCI literature, few studies have sought to simultaneously capture speech and physiological data within a voice user interface context. This leads us to our secondary research question

***RQ2:** Can we unobtrusively identify an intense emotional response during voice user interface interactions?*

To address these gaps, using a within-subject experimental design, our research observed speech, alongside physiological measures employing electrodermal activity (EDA) and automatic facial expression (AFE), during emotionally charged voice user interface interactions. The effectiveness of each extracted speech and physiological feature in explaining these emotional events was compared. By assessing the effectiveness of each feature, actionable insights regarding voice user interface evaluation methods were reported. Our results indicate that, although both speech and physiological measures are capable of unobtrusively identifying intense emotional responses during voice user interface interactions, their effectiveness in doing so differs. Indeed, the extracted physiological feature of AFE-based valence best explains users' lived emotions during intense voice user interface interactions, as its relationship strength to the observed emotion dimensions surpasses that of all extracted speech features. Notably, AFE-based valence was 41 times more powerful in explaining the emotional dimension of valence in comparison to the strongest speech feature. As a result of this study, UX professionals conducting voice user interface evaluations may efficiently select the most effective implicit method of those utilized within this study. Consequently, proper interface

evaluation can contribute to optimally designed products favoring enhanced user experiences.

The article is structured as follows. A literature review regarding the study of emotion in UX, as well as the leading speech features and physiological measures used to observe user emotions, will be presented. Following this, the proposed approach and hypotheses of the study will be explained. Next, the research methodology will be addressed, followed by the results of the study. The paper will end with the interpretations of these results within the discussion section followed by a brief conclusion.

### **3.2 Literature Review and Hypotheses Development**

The emerging omnipresence of voice user interfaces calls for methodologies regarding their evaluation. Unlike the methodologies surrounding the evaluation of digital products, authors suggest that those regarding voice user interface evaluation lack consensus amongst UX practitioners (Seaborn & Urakami, 2021), resulting in the topic's vagueness. This is perhaps due to the fact that the majority of interface and user experience designers have been trained in function of GUIs (Murad & Munteanu, 2020). This can pose difficulties for GUI designers transitioning into voice user interface design, as the GUI guidelines and patterns cannot directly be applied to voice user interfaces (Murad & Munteanu, 2020). For example, the think-aloud method is an adequate evaluation method for GUIs, but can interfere with the user's experience during voice user interface evaluations. To evaluate vocal experiences, UX professionals must resort to other methods and measures, such as self-reported measures. As stressed in the previous section, the widespread use of self-reported measures within voice user interface evaluation is limiting, as it fails to unveil the underlying automatic and subconscious user reactions which are essential to understanding user experiences. Tapping into various methods, such as implicit measures utilizing speech and physiological data, may further help paint a vivid picture of users' vocal experiences. Furthermore, a multi-method approach can be beneficial to understanding the effectiveness of each method in explaining emotional events experienced during voice user interface interactions. Assessing the strength of both physiological and speech features can provide valuable insight to UX professionals seeking to select the most effective and consequently efficient

evaluation method, while contributing to the emerging field of voice user interface evaluation.

In order to obtain a deeper understanding of the users' experience, a combination of implicit measures and explicit measures can be used (Ortiz de Guinea et al., 2014; Ortiz de Guinea et al., 2013). Implicit measures allow for real-time and precise data free of retrospective and cognitive biases to be collected (Ortiz de Guinea et al., 2014). Moreover, the unobtrusive nature of implicit measures favours a more natural reaction from participants, allowing researchers to gain insights into unconscious, automatic and authentic emotional reactions free of interruptions (Dirican & Göktürk, 2011; Ortiz de Guinea & Webster, 2013; Ortiz de Guinea et al., 2014; Ivonin et al., 2014). Thus, by including implicit measures, a more thorough understanding of the users' emotions and consequently their experiences may be noted.

### **3.2.1 Speech Features**

When considering implicit methods, one obvious choice for assessing changes in affective state is through the study of acoustic characteristics known as speech features. Indeed, research has suggested the human voice to be a ubiquitous and insightful medium of vocal communication (Cordaro et al., 2016; Juslin & Laukka, 2003; Kraus, 2017; Laukka et al., 2016; Provine & Fischer, 1989; Vidrascu & Devillers, 2005). In the field of emotion detection and speech research, common prosodic features such as fundamental frequency (F0) (e.g. minimum, maximum, mean, jitter), energy (e.g. loudness, shimmer) as well as duration are often observed and considered the amongst the most informative (Lausen & Hammerschmidt, 2020; Juslin & Laukka, 2003; Johnstone & Scherer, 2000). Other vocal paralinguistic features, such as psychoacoustics features of speech rate, pitch changes, pitch contours, voice quality, spectral content, energy level and articulation, are also often extracted due to their informative nature relating to emotion detection (Tahon et al., 2012; Shilker, 2009).

Each vocal paralinguistic feature pertains to different vocal cues. For instance, F0 depicts the rate of vocal fold vibration and is perceived as vocal pitch, where the pitch period represents the fundamental period of the signal (Bachorowski, 1999; Li & Jain, 2009).

Deriving from F0, pitch period entropy (PPE) is a measure that denotes the impaired control of F0 during sustained phonation (Little et al. 2009; Arora et al. 2019). On the other hand, spectral slope and spread respectively represent the observed tendency to have low energy during high frequencies, and the total bandwidth of a speech signal using spectral centroid, a measure used to evaluate the brightness of a speech (Mannepalli et al., 2018). As for spectral entropy, it can be observed to assess silence and voice region of speech (Toh et al., 2005). In sum, various speech features exist and denote vocal characteristics relating to states of being. A summary of the defined features may be found in Table 1 below.

**Table 1:** Summary of common speech features indicative of emotion

<b>Speech Feature</b>	<b>Definition</b>
<b>Fundamental frequency (F0)</b>	The rate of vocal fold vibration.
<b>Pitch period</b>	The fundamental period of the signal.
<b>Pitch period entropy (PPE)</b>	The impaired control of F0 during sustained phonation.
<b>Spectral slope</b>	The observed tendency to have low energy during high frequencies.
<b>Spectral spread</b>	The total bandwidth of a speech signal using spectral centroid.
<b>Spectral centroid</b>	A measure used to evaluate the brightness of a speech.
<b>Spectral entropy</b>	Observed to assess silence and voice region of speech.

Studies in both HCI and non-HCI contexts have extracted numerous speech features to explain cognitive and affective states. For instance, research surrounding PPE has suggested the speech feature to be indicative of Parkinson disease (Arora et al., 2019; Little et al., 2019). When assessing affective states, various speech features have been used simultaneously by researchers. As seen within a study by Papakostas et al. (2015), spectral entropy, alongside spectral centroid, spectral spread, and energy were observed in the aim of analyzing speakers' emotions. In research by Lausen & Hammerschmidt (2020), 1038 emotional expressions were analyzed according to 13 prosodic acoustic parameters, including F0 and its variations.

Within a HCI context, speech features have been studied through the lens of speech emotion recognition (SER) systems, in which emotional states via speech signals are analyzed (Wani et al., 2021). In line with SER systems, emotion voice conversion is meant to generate expressive speech from neutral synthesized speech or natural human voice (Robinson et al., 2019). For example, research by Xue et al. (2018) analyzed F0, power envelope, spectral sequency and duration to propose a voice conversion system for emotion that allowed for neutral speech to be transformed into emotional speech with dimensions valence and arousal serving as a control to the degrees of emotion. Moreover, in order to assess a system's recognition accuracy upon the Chinese emotional speech database, researchers extracted an array of speech features, including spectral centroid, spectral crest, spectral decrease, spectral entropy, spectral flatness, spectral flux, spectral kurtosis, spectral roll-off point, spectral spread, spectral slope, spectral skewness, in addition to prosodic features of energy and pitch (Zhu & Ahmad, 2019)

In the context of voice user interface evaluation, a study by Kohh & Kwahk (2017) analyzed speech amplitude, pitch and duration to assess participants' speech behaviour patterns during voice user interface usage. More precisely, speech patterns were observed during responses following errors produced by iPhone's Siri. As stressed by the authors, few studies have investigated users' speech behaviour patterns while using a voice user interface. As seen in Kohh & Kwahk (2017)'s study, as well as various HCI and non-HCI studies, speech features were informative of affective states. With this said, this leads us to our first replication hypothesis:

**H1. There is a relationship between the amplitude of targeted speech features and the emotional intensity of users during voice user interface interactions.**

### **3.2.2 Physiological Measures**

Measuring affective states using physiology is a predominant strategy employed within the field of UX. According to the circumplex model of affect, affective states emerge from two fundamental neurophysiological systems related to valence and arousal (Russell, 1980). Two common physiological indices used to measure the valence and arousal dimensions defining affective state are facial micro expressions and electrodermal activity (EDA). Often captured via a webcam, facial micro expressions are generally quantified using some form of automated facial expression (AFE) analysis software and assessed through the lens emotional valence. Facial expression analysis remains one of the most reliable ways to measure valence, as people are inclined to express their emotions through facial muscles micromovements (Uyl & Kuilenburg, 2005). Indeed, in one study, it was found that data captured via facial micro-expressions was more effective in measuring instant emotions and fun of use in comparison to a user's questionnaire (Zaman & Shrimpton-Smith, 2006).

Emotional valence, characterized by negative emotions (e.g., fear, anger, sadness) and positive emotions (e.g., joy, surprise) on opposite sides of the spectrum, refers to the emotional response to a specific stimulus (Bradley & Lang, 1999). Simply put, it has been described as how users feel (Burton-Jones & Gallivan, 2007). The dimension of valence can be studied alone or as a complimentary construct to arousal, described in the following paragraphs.

As for EDA, it is a measurement of electrical resistance through the skin that captures changes of skin conductance response (SCR) from the nervous system functions (Braithwaite et al., 2013; Dawson et al., 2000; Bethel, 2007). Indeed, it relates to the sympathetic nervous system, an automatic response to different situations (Riedl & Léger, 2016). The easy to use and reliable physiological measure has been widely used in NeuroIS research (Léger et al., 2014; vom Brocke et al., 2013; Giroux-Huppé et al., 2019; Lamontagne et al., 2019). Often captured via electrodes on the palm of the hand, it

is sensitive to the variations in skin pore dilation and sweat gland activation, which are in turn sensitive to changes in emotional arousal (Hassenzahl & Tractinsky, 2006; Boucsein, 2012). As suggested in the literature, it is common to infer levels of arousal through the measure of skin conductance (Dawson et al., 2007).

The arousal levels measured via EDA range from very calm to neutral to highly stimulated (Ekman & Friesen, 1978). It has been suggested to be an ecologically valid portrait of the user's arousal, while being non-invasive and free of overt recorded behaviour (Dirican & Gokturk, 2011). In one study regarding child-robot interactions, the measured arousal via skin conductance was deemed as a valuable and reliable method in assessing social child-robot interactions (Leite et al., 2013).

### **3.2.3 Combination of speech and physiological measures**

Emotion is often expressed through several modalities (Castellano et al., 2008). For instance, the arousal of emotion can manifest itself in speech, facial expressions, brain dynamics and numerous peripheral physiological signals, such as heart rate variability, respiration and of course electrodermal activity (Gross & Muñoz, 1995; Greco et al., 2019). Indeed, research has suggested that EDA dynamics are strongly influenced by respiration and speech activity (Boucsein, 2012). With this said, a link is to be made between the study of EDA and speech features in assessing emotional behavior. Current literature regarding the study of emotions includes multi-modal research utilizing both EDA and speech features. For example, in a study by Greco et al. (2019), a multi-modal approach combining EDA and speech analyses was used to develop a personalized emotion recognition system allowing for the arousal levels of participants to be assessed while reading emotional words. As suggested within the study, the algorithm's performance accuracy was at its highest when combining both implicit measures, rather than observing EDA and speech features separately, as both the sympathetic activity induced by the voice and related respiration variations were captured. Within the same vein, research by Prasetyo et al. (2020) proposed a speech activity detection system using the speech feature extraction technique Mel Frequency Cepstral Coefficients (MFCC) in addition to EDA. By including EDA, the system was able to perform in noisy

environments and compensate for the presence of emotional conditions. Hence, the complimentary nature of both measures in explaining emotional behaviour is to be noted.

On the other hand, speech features have also been studied in parallel to facial expressions. Speech and facial expressions provide a comprehensible view into a user's reaction, as visual and auditory modalities may infer a user's emotional state (Caridakis et al., 2006). To assess users' emotional states in naturalistic video sequences, a study by Caridakis et al. (2006) combined information from both facial expression recognition and speech prosody feature extraction. A study by Castellano et al. (2008) went a step further by including body gesture modality to build a multimodal emotion recognition system used to assess eight emotional states that were equally distributed in valence-arousal space. Similarly to Greco et al.'s study (2019), the classifiers based on both speech data and facial expressions outperformed classifiers trained with a single modality. This was also the case in research by Alshamsi et al. (2019), where a multimodal system including both facial expression and emotional speech was more accurate in emotion recognition in comparison to isolated functions. A summary of the multi-method studies is found in the Table 2 below.

**Table 2:** Summary of the multi-method studies utilizing speech and physiological measures in relation to emotion recognition

Study	Contribution	Methods
Greco et al. (2019)	Improved the recognition of human arousal level during the pronunciation of single affective words.	EDA Speech Features (F0 & MFCC)



Study	Contribution	Methods
Prasetio et al. (2020)	Developed a speech activity detection (SAD) system which can perform in noisy environment and compensate for the presence of emotional conditions.	EDA Speech Features (MFCC)
Caridakis et al. (2006)	Proposed a framework to model affective states in naturalistic video sequences.	Facial Expression  Speech Features (Prosody related to pitch and rhythm)  Bodily Expression (excluded in the fusion of modalities)
Castellano et al. (2008)	Presented framework of multimodal automatic emotion recognition system during a speech-based interaction.	Facial Expression  Speech Features (MFCC)  Bodily Expression
Alshamsi et al. (2019)	Proposed a framework consisting of mobile phone technology backed by cloud computing to recognize emotion in speech and facial expression in real-time.	Facial Expression  Speech Features (MFCC)

With this said, EDA, facial micro expressions, and speech are capable of explaining emotional states, both in isolation and in conjunction with each other. This leads us to our following replication hypotheses:

**H2.a. There is a relationship between the amplitude of the extracted EDA features and the emotional intensity of users during voice user interface interactions.**

**H2.b. There is a relationship between the amplitude of the extracted AFE-based valence feature and the emotional intensity of users during voice user interface interactions.**

Similarly to physiology, speech features are linked to the dimensions of valence and arousal. Indeed, the emotional arousal of a speaker is accompanied by physiological changes, consequently affecting respiration, phonation, and articulation resulting in emotion-specific patterns of acoustic parameters (Scherer, 1986). As suggested by Scherer (1986), F0, energy, and rate are considered the most indicative of arousal. More precisely, high arousal is associated to high mean F0, F0 variability, fast speech rate, short pauses, increased voice intensity and increased high frequency energy (Breitenstein et al., 2001; Davitz, 1964; Levin & Lord, 1975; Pereira, 2000; Scherer & Oshinsky, 1977; Schröder et al., 2001; Apple et al., 1979; Breitenstein et al., 2001; Kehrein, 2002; Pittam et al., 1990). Indeed, emotions associated with high levels of physiological arousal, such as anger, fear, joy, and anxiety, have depicted increases in mean F0, F0 variability, in addition to vocal intensity (Bachorowski, 1999). For example, put into context, it is not uncommon for one to speak with a loud voice when feeling gleeful. In contrast, emotions associated with low arousal levels, such as sadness, tend to have lower mean F0, F0 variability and vocal intensity (Bachorowski, 1999). With this said, vocal aspects can covary with emotional attributes, which reflect and communicate arousal levels associated to emotional reactions (Scherer et al., 1986). Across studies, results regarding arousal and speech remain consistent (Laukka et al. 2005).

On the contrary, results regarding the relationship of speech and valence are noticeably inconsistent. In some studies, positive valence has been linked to low mean F0, fast speech rate, F0 variability and little high-frequency energy (Scherer & Oshinsky, 1977; Scherer,

1974; Scherer & Oshinsky, 1977; Uldall, 1960; Pittam et al., 1990; Schröder et al., 2001). In others, valence is not associated to specific patterns of vocal cues (Apple et al., 1979; Davitz, 1964; Pereira, 2000). Moreover, research has suggested that valence values are better assessed using facial features in comparison to acoustic features (Busso et al., 2004; Busso & Rahman, 2012). In other words, the relationship between speech and valence appears to be weaker in comparison to the physiological measure of facial expression.

Considering the inconsistencies and suggested weakness of the relationship between speech features and valence, in addition to the predictive capabilities of physiological measures in relation to both valence and arousal dimensions, we hypothesize the following;

**H3. Physiological features are more explicative of emotional voice interaction events in comparison to speech features.**

### **3.3 Methods**

#### **3.3.1 Experimental Design**

To test our hypotheses, we conducted a one-factor within-subject remote laboratory experiment in which speech and physiological responses, including EDA and facial expressions, were recorded during voice user interface interactions that were purposely designed to elicit intense emotional responses. Considering the nature of the COVID-19 pandemic, a remote experimental laboratory was made mandatory. The experiment followed guidelines established for remote data collection (Giroux et al., 2021; Vasseur et al., 2021)

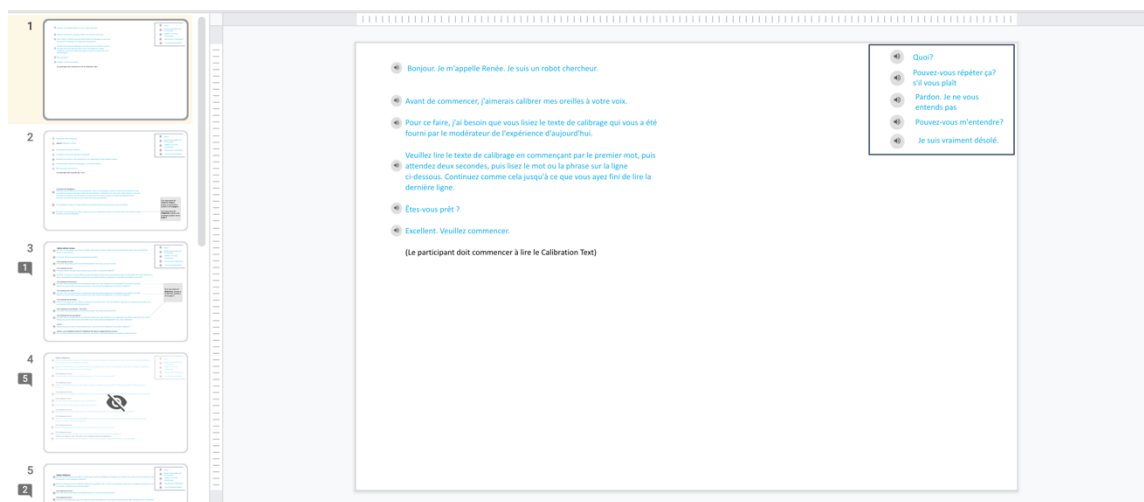
#### **3.3.2 Sample**

Participants were recruited via the university's research panel. To be eligible, participants were required to be at least 18 years of age and should not have had any of the following conditions: a partial or complete facial paralysis, a pacemaker or an inability to read text upon a computer screen. In total, 29 French-speaking participants were recruited for our study (12 men, 17 women, mean age 29 years, standard deviation 11.75). However, due

to excessive darkness and poor contrast in the video recording for AFE analysis, as well as technical issues with our remote EDA collection device, 13 subjects had insufficient data for our analyses and were therefore excluded, resulting in a sample size of 16 participants (7 men, 9 women, mean age 30.3 years, standard deviation 13.34). Each participant received a \$20 gift card for their participation. The approval of the research ethics board was received for this study (Certificate #2021-4289) and informed consent was obtained from all participants prior to their participation.

### 3.3.3 Voice User Interface Stimuli

Using a Wizard of Oz approach, participants interacted with a voice user interface whose dialogue was pre-recorded and manually controlled by a moderator. The dialogue was recorded as numerous individual MP3 files using a text-to-speech website (<http://texttospeechrobot.com/>) featuring a French-speaking female voice (RenéeV3 [IBM-Female, enhanced dnn]). The dialogue files were arranged in a script, and separated into 27 to 28 interview questions, some with multiple flows depending on participant response. To facilitate execution of the MP3 files and delivery of the dialogue to the participants via our remote testing setup, all MP3 files were uploaded to Google Drive and organized in a Google slides presentation such that the dialogue files could be played directly in a Chrome web browser, as seen in Figure 1 below.



**Fig. 1.** Google slides presentation featuring dialogue files

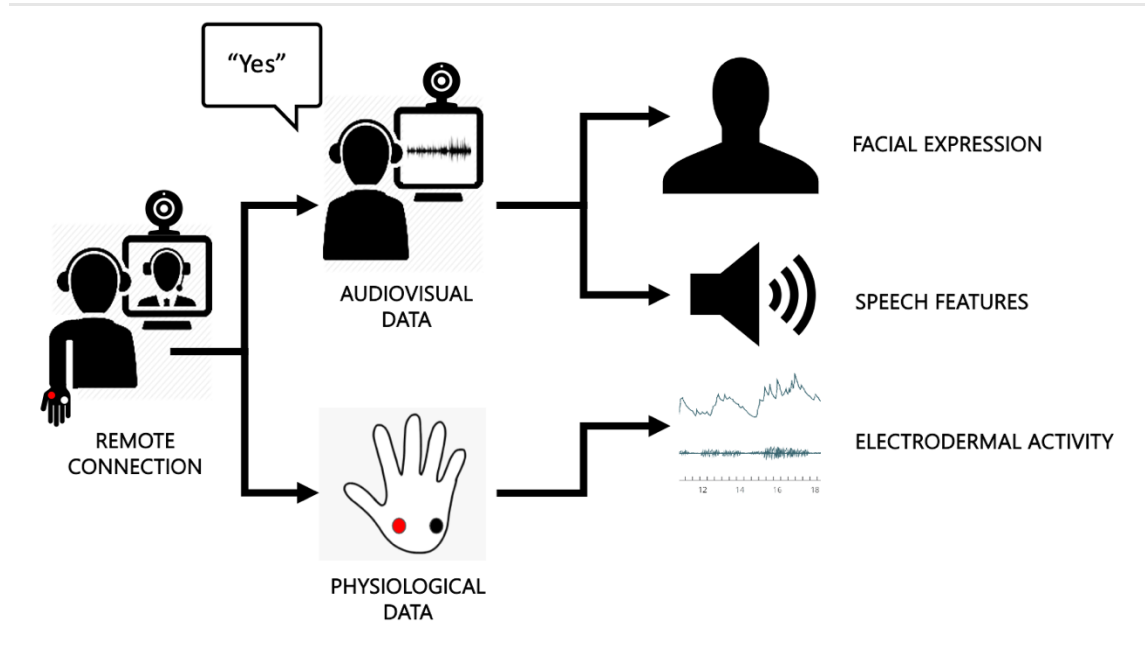
The principal means by which the voice user interface was designed to evoke emotional responses from the participants was through errors in comprehension of participant responses. For example, despite having adequately answered a question, the voice user interface often ignored a participant's response and repeated its preceding question. This occurred at the very first interaction, in which the voice user interface asked twice if the participant was ready, despite the participant's positive response (e.g., "are you ready to start?" followed by "are you ready to start?"). This depicted a total and apparent incomprehension meant to elicit an intense emotional response, aiming to be frustration in this particular exchange, from the very start of the dialogue. Participants were also asked by the voice user interface to repeat themselves on multiple occasions. Misunderstanding occurred when the voice user interface warped the participants responses (e.g., "dog" to "amphibian"). In addition to these faulty interactions, questions were purposely designed to be provocative and unexpected in order to elicit shock. For example, following a series of questions regarding a user's workout habits, the voice user interface proceeded to ask if participants ever lied about the supposed amount of exercise in the hopes of impressing others (e.g., "do you exercise every now and then" followed by "have you ever lied about how much exercise you do to impress others?"). In sum, instances of incomprehension and unexpected questioning led to intense emotional user responses during vocal interactions with a voice user interface.

In general, the voice user interface dialogue was designed to elicit yes, no, or other single word responses. A complete list of the corresponding dialogue for the voice user interface can be seen in Table 1 in Appendix 1, in which both the original French dialogue used for the experiment and the translated English version are featured.

### **3.3.4 Experimental Setup**

A remote connection between the participants and moderator was primarily established using Lookback's Liveshare, a platform allowing user research to be conducted remotely (Lookback Group, Inc., Palo Alto). To ensure an optimal data collection free of distraction and noise, participants were required to be seated alone and comfortably in a quiet room. It was necessary for the participants' computer and COBALT Bluebox device, described

in the measures section below, to be placed upon a stable surface such as a desk. Moderators asked the participants to sit in a straight and forward-facing position within a well-lit environment, in an attempt to ensure that facial expressions were adequately recorded. To ensure that the audio data was properly captured, participants were required to wear a headset or earphones with an integrated microphone. A summary of the experimental setup is found in Figure 2 below.



**Fig. 2.** An overview illustration of the experimental setup

### 3.3.5 Measures

The physiological responses of users were measured via facial expression and EDA. Facial expression was recorded via webcam at 30 fps using Lookback. The speech of subjects was captured via their computer microphone and recorded along with the speech of the voice user interface at a sampling rate of 48 KHz using Lookback. Lastly, EDA was measured at a sampling rate of 100 Hz using the COBALT Bluebox device (Courtemanche et al., 2022), a 3D printed case featuring BITalino (r)evolution Freestyle Kit (PLUX Wireless biosignals S.A.) technology to record biosignals. EDA was captured via electrodes placed on the lower part of participant's palm, as depicted within the

caricature featured in Figure 2 above. A photographic image of the placement is also found in Figure 3 below.

### **3.3.6 Experimental Procedure**

Prior to the experiment, participants received a link to their individual Lookback sessions. Once the link was accessed upon the scheduled time of the experiment, a recording of the participant's screen and webcam was automatically initiated, alongside the audio input of both the participant and the moderator.

After being welcomed to the experiment, the moderator proceeded to confirm that the participant was consensual to the participation of the experiment, as well as the recording of the session, screen, and physiological data. The moderator also validated that the informed consent form, sent 24h prior to the experiment, was read, signed, and returned.

Following this, the moderator confirmed that the participant was alone in a quiet room free of distractions. In order to limit potential distractions, participants were informed to close any unnecessary windows on their computer and set their phone to silent mode. A visual scan was performed by the moderator, ensuring that the participant was conform to the experiment. Conformity required a set of functioning headphones with an integrated microphone that did not obscure the participant's face.

The participant was then guided with step-by-step instructions to install the physiological instruments, which had previously been delivered to the participant's location. The EDA electrodes were placed on the lower part of participant's non-dominant palm. In other words, the palm's hand that was not used to control the mouse. More precisely, the electrodes were placed on the thenar and hypothenar eminence regions of the palm vis a vis the thumb and pinky fingers for optimal EDA data to be recoded (Figner & Murphy, 2011). Electrodes were wired to COBALT Bluebox technology, allowing for the participant's physiological data to be recorded. A depiction of the electrode placements wired to a COBALT Bluebox device is found below on Figure 3. Unlike Figure 3, the COBALT Bluebox device was placed in proximity to the participant's non-dominant hand on a stable surface. A validation of the cloud recording was confirmed by the moderator,

ensuring that the sensors were fully functional. A sequence of flashing lights upon the COBALT Bluebox device served as a visual marker confirming the synchronization of the data. Developed by Courtemanche et al. (2018), the synchronization technique used ensured the Bluetooth low energy (BLE) (Montréal, Qc, Canada) signals were sent to the lightbox and BITalino device in range (Courtemanche et al., 2022).



3

**Fig. 3.** The electrodes placed on the participant's non-dominant hand are connected to sensor cables wired to the COBALT Bluebox device.

---

<sup>3</sup> Image source: Brissette-Gendron, R., Léger, P.M., Courtemanche, F., Chen, S.L., Ouhnana, M. & Sénécal, S. (2021). The response to impactful interactivity on spectators' engagement in a digital game. *Multimodal Technologies and Interaction*, 4(89), 89–89. <https://doi.org/10.3390/mti4040089>



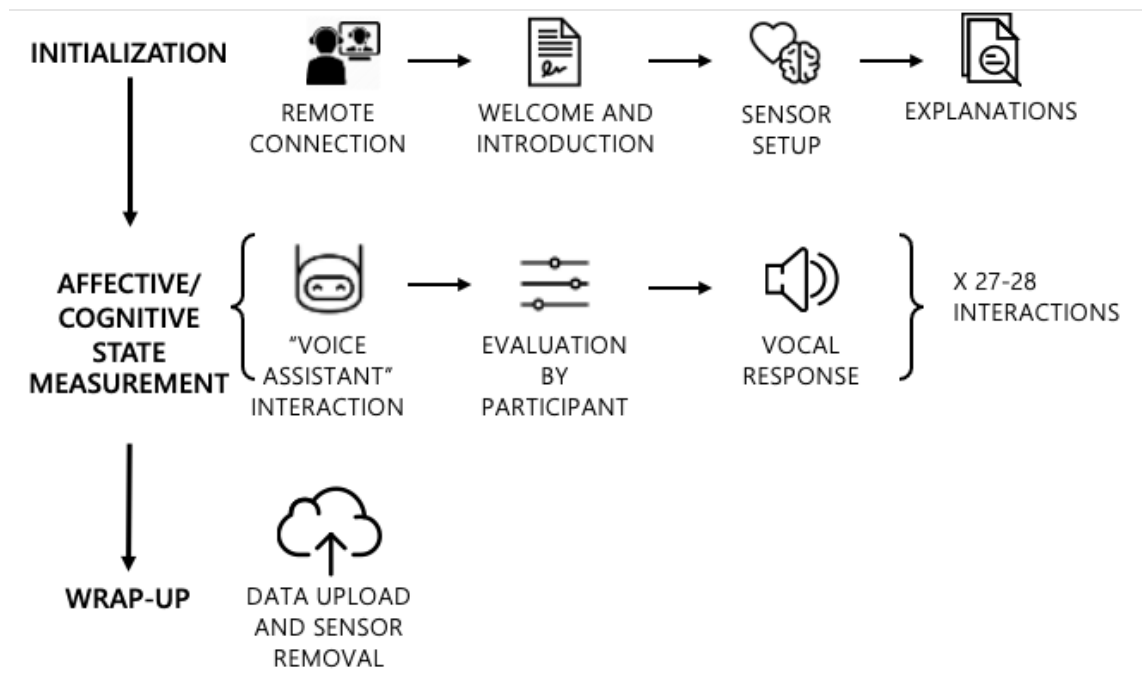
In the presence of the moderator, participants embarked on the first task consisting of a voice calibration in which they were instructed to clearly read a series of words and short sentences with a two second pause between each utterance. The implicit measures obtained during the calibration phase served as a baseline for emotional valence and arousal, as the randomized selection of words aimed to be as neutral as possible. Once the calibration phase was completed, a brief introduction and set of instructions regarding the experiment was provided to the participants. More precisely, the participant was informed an interaction with a voice user interface was to occur, and that the calibration was to be repeated following the voice user interface's instructions. In addition to the calibration, the participant was informed that the voice user interface would be conducting a short interview and that the questions posed by the interface should be responded with either a "yes" or "no" response. If these answers did not apply to the question posed, the participant was instructed to answer one of the options provided by the voice user interface. Moreover, if the participant did not know the answer to the question or could not decide, the participant was instructed to answer, "I don't know". Following each answer, the participant was required to evaluate the quality of the interaction using a digital sliding scale provided in a link through Qualtrics™ (Qualtrics International, Provo), an online survey tool<sup>4</sup>. Lastly, the participant was instructed to provide loud and clear responses in order to ensure optimal interactions with the voice user interface.

Once the instructions were provided, the moderator turned off his or her camera and adjusted the sound preferences upon Lookback, allowing for the audio output to play the first MP3 audio recording. The voice user interface audio was played in Google Chrome and transmitted directly to the participant through Lookback, using VB Audio Virtual Cable and Voicemeeter Banana Advanced Mixer, which allowed the moderator to hear both the voice user interface transmission and participant responses for continuous monitoring of participant and system-based performance during the experiment.

---

<sup>4</sup> Results from the sliding scales were purposely omitted from this study due to inconsistencies regarding evaluation time gaps between interactions.

The dialogue between the voice user interface and user commenced with the calibration task conducted previously. Following the completion of this task, an array of questions was asked, from the participant's relation to the university ("are you a student at HEC Montreal?"), to the participant's preference between cats and dogs ("do you prefer cats or dogs?"), to the participant's workout habits ("do you exercise every now and then?"). The dialogue ended with a brief conclusion by the voice user interface, thanking the participant for their time. The exchange between the voice user interface and participant lasted approximately 30 minutes. Once the final audio recording was played, the moderator turned on his or her camera, readjusted the sound preferences back to microphone setting. A summary of the procedures is found in the graphical representation in Figure 4.



**Fig. 4.** Graphical summary of the experiment procedures

### 3.3.7 Third-Party emotion evaluations

To establish ground-truth for the physiological and speech features derived from user responses to the voice user interface, third-party evaluations were conducted by six

evaluators. To perform the evaluation, evaluators watched 188 clips of participant webcam videos corresponding to each interaction in order to simultaneously consider both physical and oral expressions of emotion. Each clip was coded to commence from the moment the voice user interface's question was posed and ended 500 milliseconds following the participant's response. Each participant had a range of 7 to 17 interaction clips to be evaluated, presented in a randomized order. Evaluations were recorded using online survey tool Qualtrics. The survey used to record the evaluations was built in the platform and embedded on the page using a custom HTML creative. Each survey recorded the evaluations of the same participant, resulting in 16 unique Qualtrics links.

To ensure standardized evaluations, all evaluators were trained. Within this training, evaluators were guided within their manual assessment of the four studied dimensions of affective state: valence, arousal, control, and short-term emotional episodes (STEE). As its name suggests, the STEE evaluation point was indented to capture momentary fleeting glimpses into the participant's emotional state. The temporal nature of these events did not make them any less important. On the contrary, these split moments depicted authentic emotion, especially amongst subjects who tended to suppress public displays of emotion.

Evaluators were instructed to watch each interaction clip twice and assess the emotional reaction using both visual and voice behavior of the participant, while taking into consideration the semantic context of the voice user interface speech. A series of instructions and guidelines addressing the emotional dimensions to be assessed were provided and explained to the evaluators. For each dimension, the spectrum of extremes was defined. In addition to these definitions, a series of vocal and visual cues were provided as examples of elements to look out for.

Evaluators were provided instructions with regards to the Self-Assessment Manikin (SAM) scale proposed by Bradley and Lang in 1994. Valence, arousal and control are classic dimensions of affective state measured ubiquitously in IS research by users through self-assessment questionnaires. The SAM scale measures three emotional dimensions, that of pleasure, arousal, and control or dominance, using a

series of graphic abstract characters displayed horizontally using a 9 point-scale, although 5 and 7-point variants may also exist (Betella & Verschure, 2016). For this experiment, we opted for the 9-point scale in order to offer further precision and remain consistent with previous observer-based studies utilizing this measure (Sutton et al., 2019; Jessen & Kotz, 2011).

In contrast to the valence, arousal and control dimensions, the STEEs were observed using a binary evaluation. To assess STEEs, evaluators were asked to select the best suited option (non-present, positive STEE or negative STEE) applicable to the interaction. Solely its presence, rather than its frequency and intensity, was observed within this evaluation point. In addition to the SAM-based and binary-based scale ratings, evaluators were asked to note the vocal and visual cues supporting their evaluations.

In order to assess the evaluators' grasp of the dimensions, all six analyzed the same participant. Following this primary evaluation, the results were analyzed and further guidance was provided in order to ensure uniformity. The process was repeated, resulting in greater consistency. Once this consistency was achieved, evaluators were instructed to pursue the remaining evaluations. The remaining Qualtrics links, featured in random and individualized orders, limited the risk of biasness as evaluator fatigue upon the same final evaluation was avoided.

### **3.3.8 Data Processing and Feature Extraction**

As a result of the recorded experiments, two raw data streams, being video and EDA, were captured. Within the raw video data stream, both audio and visual information was recorded. In order to extract the video's audio and obtain a raw audio file, open-source audio software Audacity (Muse Group, New York) was employed. In parallel, the video was processed using software FaceReader 8 <sup>TM</sup> (Noldus, Wageningen, The Netherlands), resulting in time series data stream for AFE-based valence. The output, or timepoints, from FaceReader 8 were aligned with the captured EDA, as the COBALT Bluebox's flashing light series confirmed the synchronization of data.

Each physiological measure pertained to an interaction between the voice user interface and participant, starting from the moment the interface posed the question up until the participant's response. The participant's response was purposely excluded from the physiological measurement window in order to prioritize and observe the emotional build-up prior to a verbal response. Moreover, by observing this particular time window, the studied physiological measures focused on early indications of emotional responses. In contrast to the time windows chosen for physiological measures, the participant's verbal response was observed for the speech measure, from the start of the participant's utterance to the end of his or her responses.

### **3.3.8.1 Speech Features**

The onset of the start of the participants' speech response was manually identified for every interaction where the response was "yes" and defined as the timepoint where the participants speech envelope exceeded .10 decibels. This was done so for the entirety of the experiment in which a user interacted with the voice user interface, including both the "yes" responses during the calibration and testing periods. The onset of the voice user interface speech was also marked, in which the defined timepoint was identical to that of the participant's response. The time window from the onset of participant responses, until 500 milliseconds after that response.

To extract the speech features, we used Surfboard, an open-source Python library for extracting audio features, and a python wrapper for open-source Speech Signal Processing Toolkit (SPTK)<sup>5</sup>. Congruent with existent research on emotion and speech (Yildirim et al., 2004), we extracted the following spectral features using audio software Audacity: spectral slope, spectral entropy, spectral centroid, spectral spread, F0, F0 standard deviation and pitch period entropy, all recorded via the participant's webcam. As suggested, these parameters are amongst the most commonly analyzed with the study of emotion in speech (Yildirim et al., 2004).

---

<sup>5</sup> <http://sp-tk.sourceforge.net/>

### **3.3.8.2 Facial Expression Feature**

The participants' facial micro expressions during their interactions with the voice user interface were analyzed using automated facial expression analysis software FaceReader 8. Noldus' FaceReader is considered a valid recognition software capable of automated facial coding (Skiendziel et al., 2019; Lewinski et al., 2014). The AFE analysis was conducted subsequently upon the Lookback recordings as M4V video files with a frame rate of 10 fps. The software coded the action units of the facial micro expressions exhibited by the participants in the webcam videos at a rate of 4 Hz. Valence levels were calculated by FaceReader 8 by the intensity of "happy" minus the intensity of the negative expression with the highest intensity (Noldus). Indeed, AFE can automatically recognize micro changes in facial action units (e.g., brow raise, chin raise, jaw drop, etc.) and interpret data based on the Facial Action Coding System (FACS) developed by Ekman and Friesen (Cohn & Kanade, 2007; Ekman & Frieson, 1978), allowing researchers to distinguish between a set of discrete emotions, such as angry, happy, disgusted, sad, scared and surprised.

The timeseries data, from the onset of the voice user interface speech until the onset of participant response, was averaged and used as a value for AFE-based valence. The participant's response was purposely omitted in order to avoid dubious automated facial expression analyses affected by mouth movements of verbal responses. This calculation was performed for both the experimental and calibration time windows. Following this, the experimental values were standardized by subtracting the overall average of the values calculated for time windows during the calibration time period. The AFE-based valence time-series data was further processed for each interaction tested within the statistical analyses.

### **3.3.8.3 Electrodermal Activity Feature**

Similarly to the facial expression feature, the raw EDA time-series data, from the onset of the voice user interface speech until the onset of participant response, was averaged and used as a value for EDA features. This calculation was performed for both the experimental and calibration time windows. Once this calculation was performed, the

experimental values were standardized by subtracting the overall average of the values calculated for time windows during the calibration time period.

EDA features were processed in order to obtain phasic and z-score time series data. Often referred to as EDA “peaks”, phasic changes are abrupt increases in the skin conductance (Braithwaite et al., 2013). In other words, phasic EDA stems from faster changing elements of the signal, known as the Skin Conductance Response (SCR) (Braithwaite et al., 2013). As for the z-score, it requires the mean and standard-deviation to be used in substitute of a hypothetical maximum (Braithwaite et al., 2013).

The phasic component of the EDA-time series was extracted. In parallel, the conversion of the entire raw EDA time-series into a z-score was performed. The phasic EDA and z-score EDA time-series data was further processed to derive phasic and z score features, serving as targets for an arousal assessment, for each interaction tested within the statistical analyses.

### **3.3.9 Statistical Analyses**

Using SPSS® (IBM, New York) Intraclass correlation (ICC) testing was performed based on the 188 evaluations across all six evaluators to assess inter-evaluator reliability and consequently demonstrate consistency regarding observational ratings provided by the evaluators (Hallgren, 2012; Bartko, 1966). ICC scores allow for both the degree of correlation and agreement between measurements to be reflected within a reliability index (Koo & Li, 2016). The threshold for significance was set at  $p \leq 0.05$ . In order to measure the statistical relationship between the ground-truth and the extracted speech and physiological features, linear regressions with random intercept were performed. A repeated linear regression with random intercept was performed against each ground-truth affective dimension separately, with the combined speech and physiological measures as factors. The three physiological factors were AFE-based valence, phasic EDA and EDA z-score. The eight speech factors were spectral slope, spectral entropy, spectral centroid, spectral spread, PPE, log energy as well as F0 standard deviation and F0 mean. To correct for the 11 repeated measures of each regression model, Bonferroni

correction was applied at  $\alpha = 0.05$ , resulting in a significance threshold of  $p \leq 0.0045$  (Bland & Altman, 1995).

### 3.4 Results

#### 3.4.1 Inter-evaluator Reliability Results

The following table is a summary of the ICC scores per evaluated dimension for all evaluators and interactions combined.

**Table 3:** Results of the ICC scores

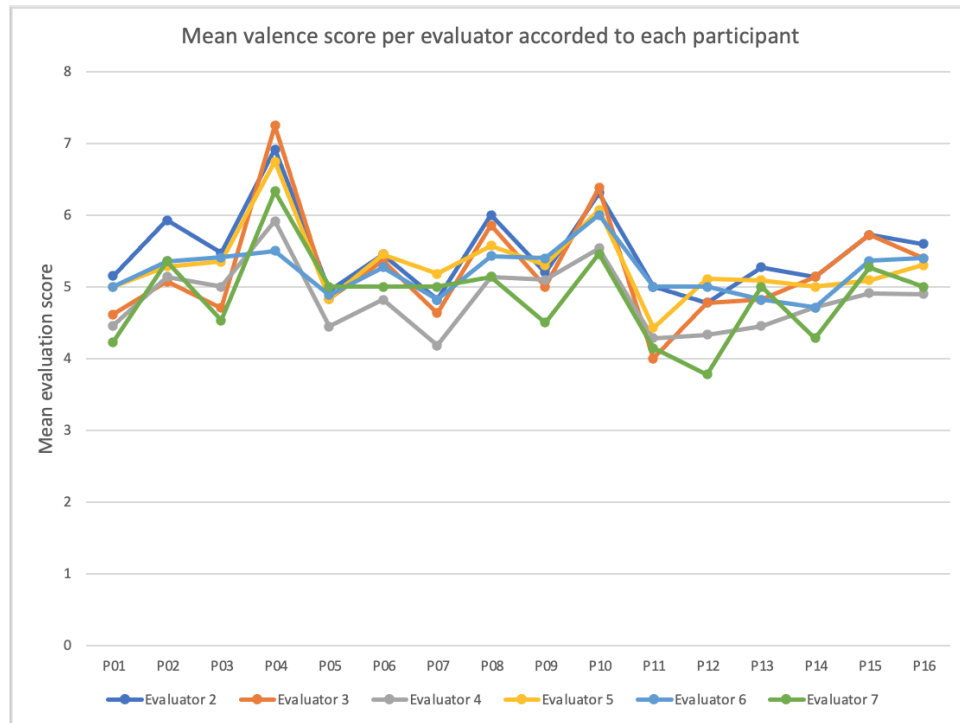
<b>Dimension</b>	<b>ICC scores</b>	<b>95% Confidence interval lower bound</b>	<b>95% Confidence interval upper bound</b>
Valence	0.898	0.874	0.919
Arousal	0.755	0.696	0.806
Control	0.789	0.739	0.833
STEE	0.707	0.637	0.767

As seen in Table 3, the ICC scores per dimension were 0.898 for valence, 0.755 for arousal, 0.789 for control and 0.707 for STEE. With the exception of STEE, all ICC scores were above 0.75, indicating excellent inter-rater agreement (Cicchetti, 1994). Based on analysis standards, inter-rater agreement for STEE was considered adequate, as it fell within the .60 and 0.74 range (Cicchetti, 1994). Of the four evaluated dimensions, valence was the most agreed upon dimension, whereas STEE was the least. For a summary of the descriptive statistics regarding the third-party evaluation, see Table 4 below. For a visual representation of the evaluator tendencies, see Figures 5.a.b.c.d below in which four distinct line graphs depicting the mean scores per evaluator, participant and dimension are presented.

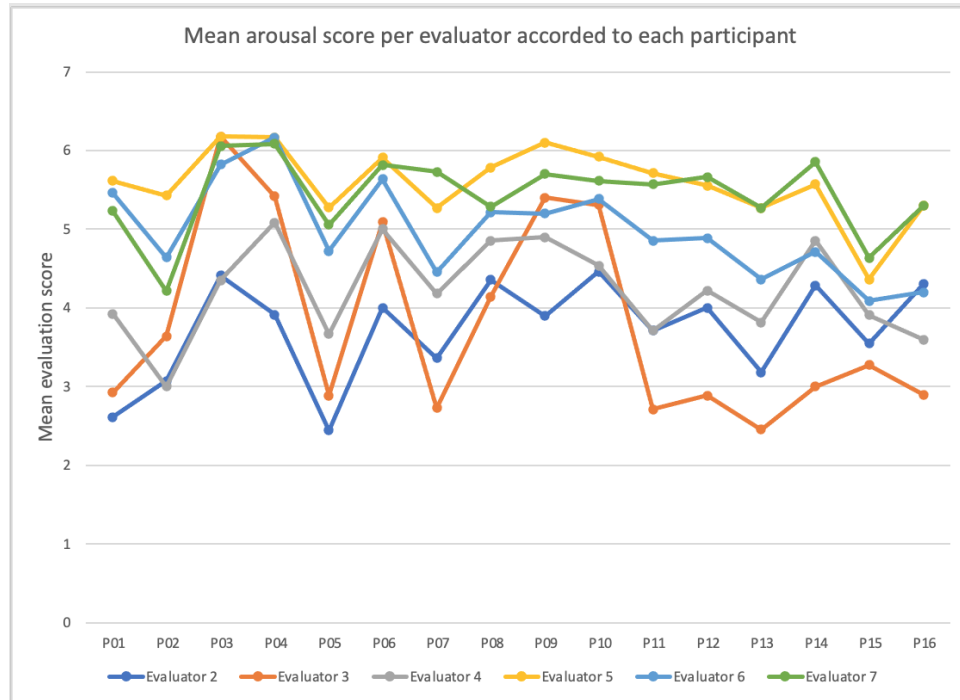


**Table 4:** Descriptive statistics of third-party evaluations per dimension

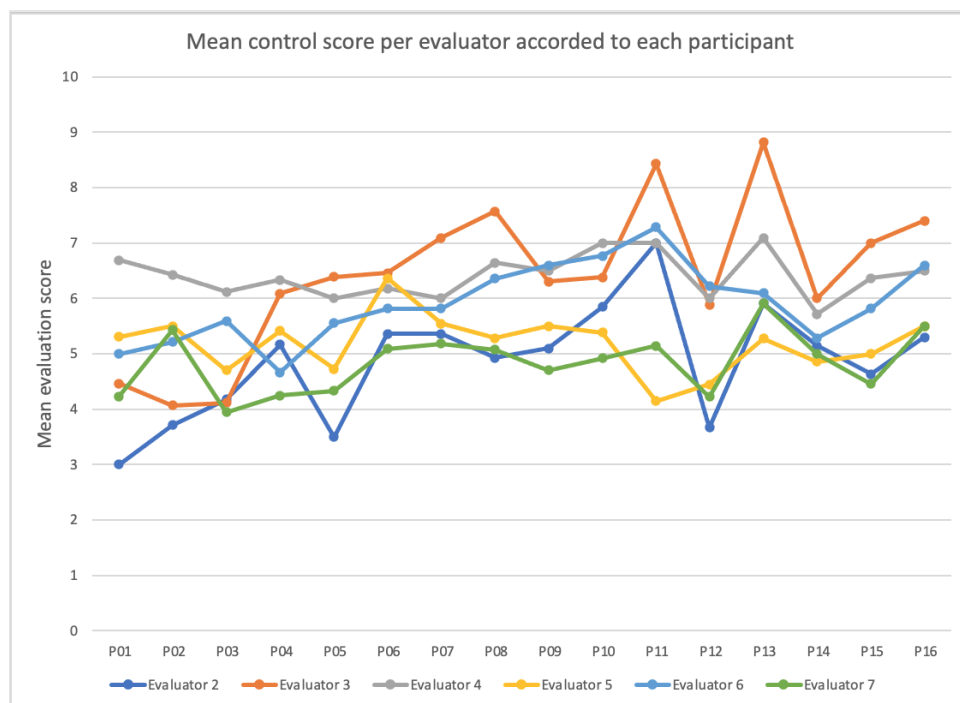
	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance
Valence	5.187	4.862	5.516	.654	1.135	.061
Arousal	4.640	3.676	5.601	1.926	1.524	.665
Control	5.537	4.723	6.404	1.681	1.356	.542
STEE	-.057	-.101	.027	.128	-.263	.002



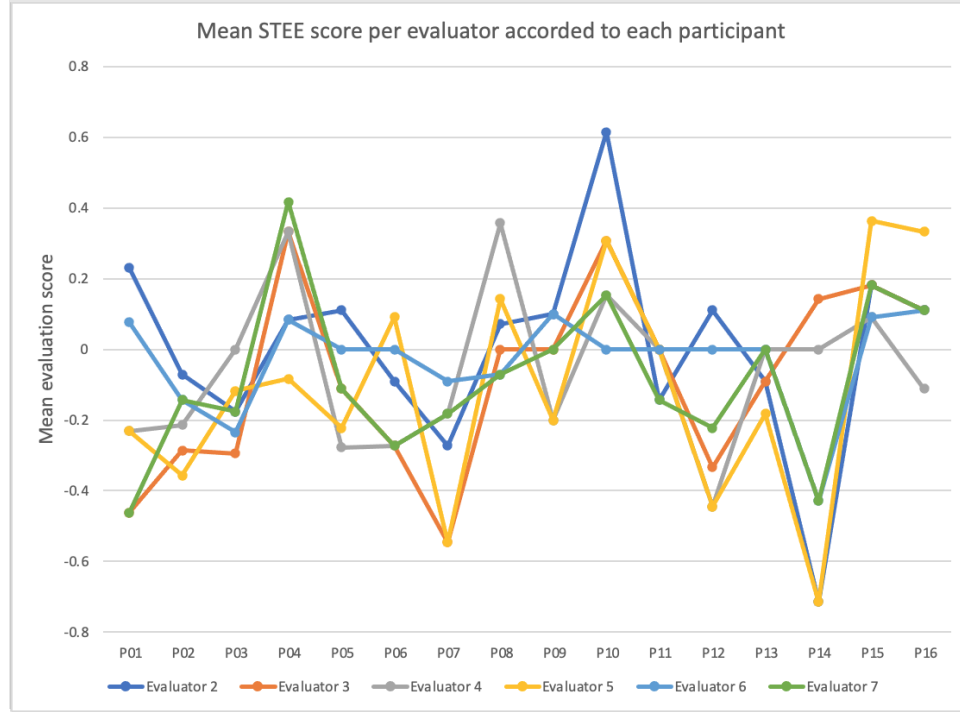
**Fig.5.a** Mean valence score per evaluator according to each participant



**Fig.5.b** Mean arousal score per evaluator according to each participant



**Fig.5.c** Mean control score per evaluator according to each participant



**Fig.5.d** Mean STEE score per evaluator according to each participant

**Fig.5.a.b.c.d.** The evaluator scores of the dimensions of valence, arousal and control are in function of the 9-point SAM scale, whereas the evaluator score of the dimension of STEE ranges from  $-2$  to  $2$ .

### 3.4.2 Multiple Linear Regression

Tables 5,6,7 and 8 present the regression results of the four observed emotion dimensions. Multiple linear regression did not reveal a significant relationship between the evaluated dimension of valence and any speech feature prior to the Bonferroni correction. Although the most explicative speech feature showing the highest R-squared value of 0.009 was spectral spread, it was deemed insignificant (see Table 5 below). As for the arousal dimension, multiple linear regression revealed significant relationships between the emotional dimension and the following speech factors featuring their respective p-values, being spectral slope (0.001), spectral spread (0.004), F0 standard deviation (0.010), and log energy (0.001) (see Table 6). Following the Bonferroni correction, spectral slope, spectral spread and log energy remained statistically significant. The R-squared values

associated to spectral slope, spectral spread and log energy were respectively 0.060, 0.044, and 0.078. Hence, the most explicative speech factor of the arousal dimension was log energy. As for the control dimension, multiple linear regression revealed significant relationships between the dimension and two factors, being spectral slope and spectral spread, with respective p-values of 0.008 and 0.040 (see Table 7). The R-squared values associated to spectral slope was 0.048 and 0.028 for spectral spread. However, neither factor was considered statistically significant following the Bonferroni correction. Lastly, multiple linear regression revealed significant relationships between the dimension of STEE and F0 standard deviation, with a p-value of 0.015 (see Table 8). Following the Bonferroni correction, F0 standard deviation was not considered statistically significant.

**Table 5:** Regression results of Valence dimension

<b>Factor</b>	<b>Estimate</b>	<b>SE<sup>1</sup></b>	<b>DF<sup>2</sup></b>	<b>T Value</b>	<b>P Value</b>	<b>R2<sup>3</sup> Value</b>
AFE-based valence	3.076	0.351	129	8.770	<0.001* <sup>4</sup>	0.402
EDA Z-Score	-0.074	0.066	129	-1.120	0.266	0.007
Phasic	0.026	0.076	127	0.350	0.728	<0.001
Slope	-106.750	105.970	144	-1.010	0.316	0.007
Entropy	0.023	0.177	144	0.130	0.898	<0.001
Centroid	0.000	0.000	144	-0.540	0.590	0.002
Spread	0.000	0.000	144	-1.230	0.221	0.010
PPE <sup>5</sup>	0.000	0.000	144	-0.860	0.391	0.004
F0 Standard Deviation	-0.006	0.007	144	-0.790	0.430	0.004
F0 mean	-0.002	0.005	144	-0.460	0.646	0.002
Log energy	0.014	0.024	144	0.590	0.559	0.003

<sup>1</sup> SE: Standard Error

<sup>2</sup> DF: Degree of Freedom

<sup>3</sup> R2: R-Squared

<sup>4</sup> Significant factors following the Bonferroni correction, with threshold of 0.004, identified with \*

<sup>5</sup> PPE: Pitch Period Entropy

**Table 6:** Regression results of Arousal dimension

<b>Factor</b>	<b>Estimate</b>	<b>SE<sup>1</sup></b>	<b>DF<sup>2</sup></b>	<b>T Value</b>	<b>P Value</b>	<b>R2<sup>3</sup> Value</b>
AFE-based valence	1.755	0.365	129	4.810	<0.001* <sup>4</sup>	0.152
EDA Z-Score	0.114	0.056	129	2.020	0.046	0.019
Phasic	0.154	0.064	127	2.420	0.017	0.028
Slope	-310.810	92.645	144	-3.350	0.001*	0.060
Entropy	-0.240	0.157	144	-1.530	0.129	0.012
Centroid	0.000	0.000	144	-1.410	0.160	0.009
Spread	0.000	0.000	144	-2.940	0.004*	0.044
PPE <sup>5</sup>	0.000	0.000	144	-1.880	0.062	0.014
F0 Standard Deviation	0.017	0.006	144	2.610	0.010	0.032
F0 mean	0.001	0.004	144	0.260	0.799	<0.001
Log energy	0.073	0.022	144	3.360	0.001*	0.079

<sup>1</sup> SE: Standard Error<sup>2</sup> DF: Degree of Freedom<sup>3</sup> R2: R-Squared<sup>4</sup> Significant factors following the Bonferroni correction, with threshold of 0.004, identified with \*<sup>5</sup> PPE: Pitch Period Entropy**Table 7:** Regression results of Control dimension

<b>Factor</b>	<b>Estimate</b>	<b>SE<sup>1</sup></b>	<b>DF<sup>2</sup></b>	<b>T Value</b>	<b>P Value</b>	<b>R2<sup>3</sup> Value</b>
AFE-based valence	-0.400	0.530	129	-0.75	0.452	0.005
EDA Z-Score	-0.133	0.082	129	-1.61	0.109	0.016
Phasic	-0.108	0.095	127	-1.14	0.255	0.008
Slope	367.030	136.200	144	2.69	0.008	0.049
Entropy	0.425	0.229	144	1.86	0.065	0.023
Centroid	0.000	0.000	144	1.53	0.129	0.014
Spread	0.000	0.000	144	2.07	0.040	0.029
PPE <sup>4</sup>	0.000	0.000	144	0.3	0.764	<0.001
F0 Standard Deviation	-0.004	0.010	144	-0.41	0.681	0.001
F0 mean	-0.002	0.006	144	-0.38	0.702	0.001
Log energy	-0.049	0.032	144	-1.57	0.120	0.021

<sup>1</sup> SE: Standard Error<sup>2</sup> DF: Degree of Freedom<sup>3</sup> R2: R-Squared<sup>4</sup> PPE: Pitch Period Entropy

Note: No feature was considered statistically significant

**Table 8:** Regression results of STEE dimension

<b>Factor</b>	<b>Estimate</b>	<b>SE<sup>1</sup></b>	<b>DF<sup>2</sup></b>	<b>T Value</b>	<b>P Value</b>	<b>R2<sup>3</sup> Value</b>
AFE-based valence	0.936	0.171	129	5.480	<.0001* <sup>4</sup>	0.209
EDA Z-Score	-0.029	0.030	129	-0.960	0.337	0.006
Phasic	0.007	0.034	127	0.210	0.837	<0.001
Slope	-18.565	47.559	144	-0.390	0.697	0.001
Entropy	0.031	0.079	144	0.400	0.693	0.001
Centroid	0.000	0.000	144	0.250	0.807	<0.001
Spread	0.000	0.000	144	0.330	0.743	<0.001
PPE <sup>5</sup>	0.000	0.000	144	0.420	0.677	0.001
F0 Standard Deviation	-0.008	0.003	144	-2.460	0.015	0.038
F0 mean	0.000	0.002	144	0.020	0.984	<0.001
Log energy	-0.002	0.011	144	-0.150	0.884	<0.001

<sup>1</sup> SE: Standard Error

<sup>2</sup> DF: Degree of Freedom

<sup>3</sup> R2: R-Squared

<sup>4</sup> Significant factors following the Bonferroni correction, with threshold of 0.004, identified with \*

<sup>5</sup> PPE: Pitch Period Entropy

As stressed, no speech factors were deemed significant in explaining the dimensions of arousal, control and STEE. However, spectral slope, spectral spread and log energy were considered statistically significant features in explaining the arousal dimension. All three speech features have a R-squared value under 0.10, indicating an existent but weak relationship as at least 90% of the variability in the outcome data cannot be explained. Despite the weakness of their relationship strength, speech features are deemed statistically significant in explaining an emotional dimension within the context of voice user interface interactions. Thus, **H1 is supported.**

Multiple linear regression between EDA features, being phasic EDA and EDA z-score, and the ground-truth dimension of arousal failed to reveal a relationship between the extracted features and the emotional intensity of users during voice user interface interactions. Despite the fact that multiple linear regression revealed significant relationships between the evaluated dimension of arousal and EDA features, EDA z-score and phasic EDA, with respective p-values of 0.046 and 0.017, both were deemed insignificant following the Bonferroni correction (see Table 6). Within the context of this

study, the amplitude of the extracted EDA features was not explicative of a user's arousal during voice user interface interactions. Thus, **H2.a is not supported.**

Multiple linear regressions between AFE-based valence and ground-truth dimension of valence revealed a relationship between the feature and the emotional intensity of users during voice user interface interactions. Indeed, the multiple linear regression revealed a significant relationship between the evaluated dimension of valence and AFE-based valence ( $p < .0001$ ). This fact remained valid following the Bonferroni correction. The R-squared value associated to AFE-based valence was of 0.402. Statistically speaking, approximately 40% of the dimension variable is explained by AFE-based valence (see Table 5). In other words, the amplitude of the extracted AFE-based valence feature is explicative of a user's valence during voice user interface interactions. Hence, **H2.b is supported.**

### 3.4.3 Multiple Linear Regression of speech and physiology

As stressed, multiple linear regression revealed a relationship between the dimension of valence and speech feature spectral spread, with a R-squared value of 0.009. However, even prior to the Bonferroni correction, the relationship was deemed statistically insignificant. On the other hand, the multiple linear regression revealed a significant relationship between the evaluated dimension of valence and AFE-based valence ( $p < .0001$ ), with a R-squared value of 0.402 (see Table 5). Thus, when comparing R-squared values for spectral spread and AFE-based valence, the physiological measure had approximately 41 times more predictive power than voice feature when assessing valence ratings. The relationship between the valence dimension and the AFE-based valence was therefore stronger than any observed speech feature.

As for arousal, multiple linear regression revealed a relationship between AFE-based valence and the dimension in question under 95% confidence interval range ( $<0.0001$ ) (see Table 6). Following Bonferroni correction, AFE-based valence remained statistically significant, with a R-squared value of 0.152. This was the sole extracted physiological feature that was considered statistically significant, as EDA z-score and phasic EDA did not achieve significance. Despite having fewer statistically significant factors,

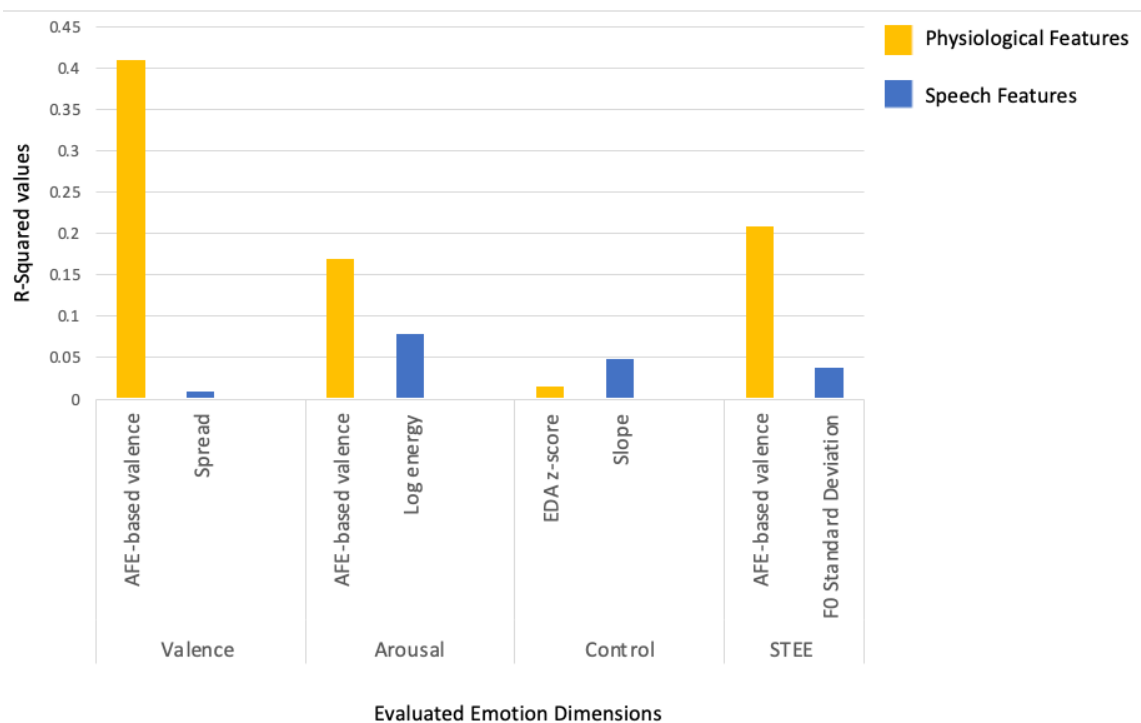
physiological measure AFE-based valence indicated a stronger relationship in comparison to significant speech features spectral slope, spectral spread, and log energy. When comparing the R-squared value of AFE-based valence to the highest value amongst the statistically relevant speech features, being log energy (0.078), physiological feature AFE-based valence had nearly twice the strength in explaining arousal ratings. Hence, the relationship between AFE-based valence is stronger than any observed speech feature in assessing users' arousal levels.

In addition to sharing a relationship with dimensions of valence and arousal, multiple linear regression revealed a statistically significant relationship between the evaluated dimension of STEE and AFE-based valence with a p-value of  $<0.0001$  (see Table 8). Speech factor F0 standard deviation also shared a relationship, with a p-value of 0.015. Following the Bonferroni correction, only the physiological factor AFE-based valence remained statistically significant. The R-squared value of AFE-based valence was of 0.208, and 0.038 for F0 standard deviation. Consequently, physiological feature AFE-based valence was approximately five times stronger than the voice feature F0 standard deviation in explaining STEE ratings. Thus, the relationship strength of AFE-based valence and dimension STEE surpasses that of any speech feature.

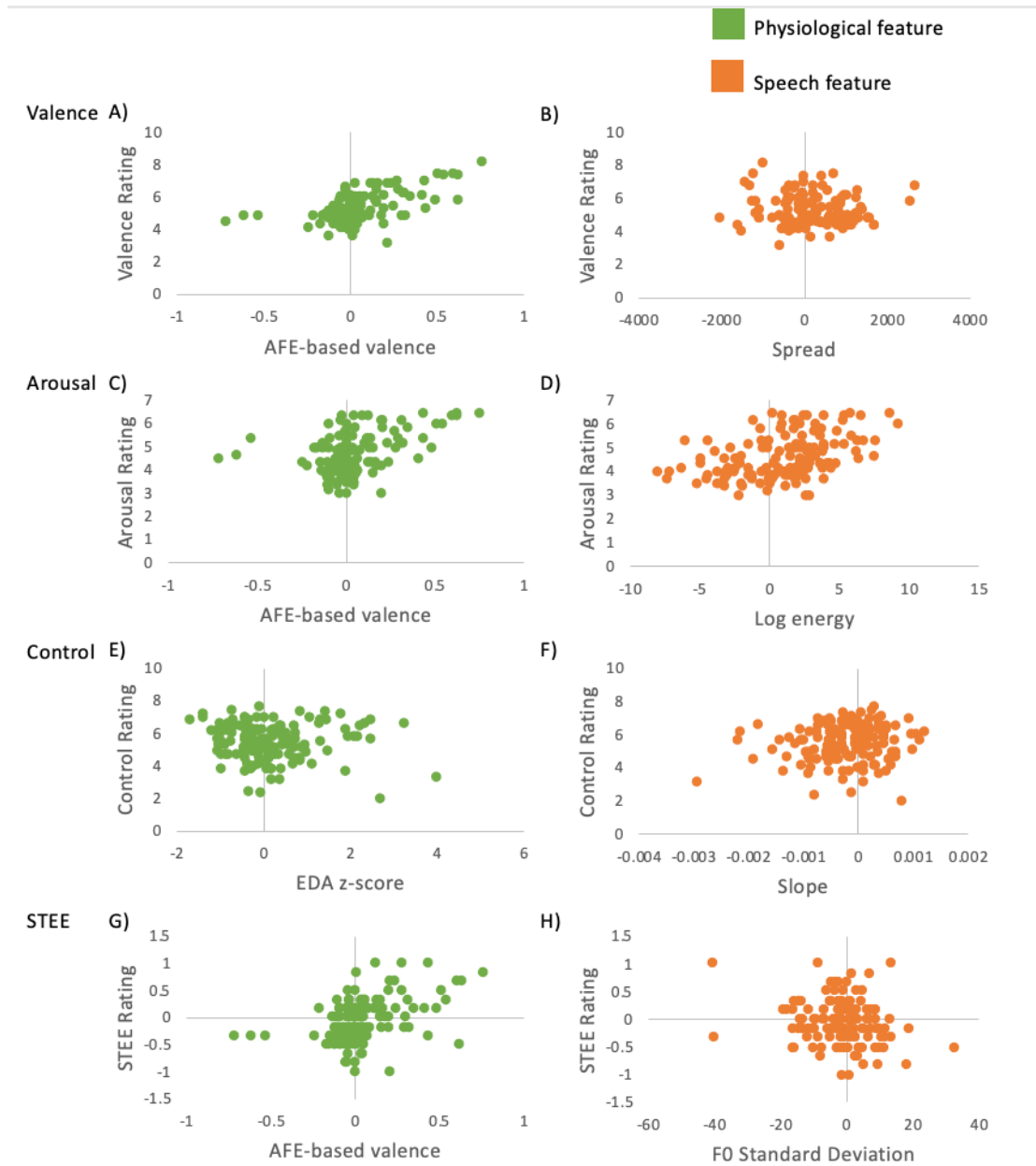
Multiple linear regression revealed statistically significant relationships between the evaluated dimension of control and two speech factors, being spectral slope and spectral spread, with respective p-values of 0.008 and 0.040 (see Table 7). Under the 95% confidence interval range, no physiological factor was deemed significant. As for speech features, the R-squared value associated to spectral slope was 0.048, and 0.028 for spectral spread. However, neither factor was considered statistically significant following the Bonferroni correction. In contrast to the valence and arousal dimensions, the speech factor of spectral slope was deemed more predictive of the control dimension in comparison to the strongest physiological factor EDA z-score. Indeed, in comparison to the R-squared value of EDA z-score (0.015), speech feature's spectral slope had approximately three times more strength than physiological feature's EDA z-score in explaining control ratings. Hence, the relationship between speech factor spectral slope is stronger than physiological factor EDA z-score in explaining the control dimension. However, as



stressed, no factor was considered statistically significant in predicting control ratings. A comparative depiction of the most explicative physiological and speech features can be found in Figure 6 and Figure 7.



**Fig. 6.** Bar chart of relationship strengths between physiological and speech features per dimension



**Fig. 7.** Boxplots of evaluator ratings and select physiological and speech features

In sum, multiple linear regression revealed statistically significant relationships between AFE-based valence, and the dimensions of valence, arousal, and STEE. Although speech features were statistically significant in explaining the arousal dimension, the relationship between the observed dimension and the strongest speech feature, being log energy, was

nearly half of AFE-based valence's strength. As for the control dimension, no physiological or speech feature was considered statistically significant in explaining the dimension. Overall, physiological feature AFE-based valence best explains the users' affective states during voice user interface interactions. Therefore, **H3 is supported**. A summary of the hypotheses status following the results can be found in Table 9.

**Table 9:** Summary of hypotheses in relation to results status

Hypothesis	Description	Results Status
H1	There is a relationship between the amplitude of targeted speech features and the emotional intensity of users during voice user interface interactions.	Supported
H2.a	There is a relationship between the amplitude of the extracted EDA features and the emotional intensity of users during voice user interface interactions.	Not supported
H2.b	There is a relationship between the amplitude of the extracted AFE-based valence feature and the emotional intensity of users during voice user interface interactions.	Supported
H3	Physiological features are more explicative of emotional voice interaction events in comparison to speech features.	Supported

### 3.5 Discussion

The primary goal of this study was to compare the effectiveness of physiological and speech measures through their respective features in explaining the affective states of users during emotionally charged voice user interface interactions. Our research used

speech and physiological measures employing EDA and facial expression analysis. As a result, we extracted eight distinct speech features, such as F0, spectral slope, and spectral spread, alongside three physiological features, being EDA z-score, phasic EDA, and AFE-based valence. Results suggest that speech features are indeed explicative of users' emotions during voice user interface interactions (**H1**). More precisely, relationships between speech features spectral slope, spectral spread, and log energy with the dimension of arousal can be noted. Of the three, log energy shared the strongest relationship strength with the arousal dimension. As suggested in speech literature, the energy of vocal responses is reflective of arousal (Scherer et al., 1984). Research regarding the subject suggests energy, as well as F0 and speech rate, to be the most indicative speech features of arousal, with high arousal associated to high frequency energy (Scherer 1986; Pittal et al., 1990; Scherer & Oshinsky, 1977; Schröder et al., 2001). Hence, our results are in line with previous research which consequently **supports H1**.

Contrary to what was hypothesized, within the context of this study, the amplitude of the extracted EDA features does not share a relationship with the emotional intensity of users during voice user interface interactions (**H2.a**). Although EDA is widely considered an appropriate measure for arousal, the latency of skin conductance response is approximately two seconds, with a range between one and five seconds (Christopoulos et al., 2019). Considering the fact that certain questions (such as “Really?”) were brief, the timeframe of analysis might have excluded important indicative electrodermal signals. As noted in this study and suggested within literature, arousal can manifest itself in through various modalities, including facial expressions and speech (Gross & Muñoz, 1995; Greco et al., 2019). Enhanced arousal levels influence the intensity of facial reactions (Fujimura et al., 2010). Since the observed voice user interface interactions stemmed from emotionally charged events, users' facial expressions may have been accentuated and were consequently reflective of arousal levels. Hence, the relationship between AFE-based valence and the dimension of arousal was stronger than phasic EDA and EDA z-score, both deemed statistically insignificant in relation to the observed dimension. Thus, **H2.a is not supported**.

As for the dimension of valence, the strength of the relationship between the amplitude of the extracted AFE-based valence feature and the dimension in question was approximately 41 times more powerful than the most predictive speech feature, suggesting a relationship between the extracted physiological feature and the emotional intensity of users during voice user interface interactions (**H2.b**). This result supports previous findings in emotion literature suggesting facial expression to be more indicative of valence than speech features (Busso et al., 2004; Busso & Rahman, 2012). Indeed, results concord with the idea that facial expression analysis is one of the most reliable measures of valence, as individuals are more likely to express emotions through facial micromovements (Uyl & Kuilenburg, 2005). Thus, **H2.b is supported**.

On the contrary, research has suggested that there are no specific vocal cues associated to valence (Apple et al., 1979; Davitz, 1964; Pereira, 2000). Moreover, the effects of valence are often vocally inapparent as they are masked by other emotional dimensions, such as arousal and dominance (Patel et al., 2010). Our results are in line with the literature, as no speech feature was deemed statistically significant in explaining valence. On the contrary, with the exception of the control dimension, physiological feature AFE-based valence shared a significant relationship with all observed emotion dimensions. As addressed previously, the suggested relationship between the physiological measure of facial expression, and the dimensions of valence and arousal, are in line with emotion literature. As for STEE, it is also best explained by AFE-based valence. Due to their brief nature, physiological changes in facial expressions may easily have been captured via AFE in comparison to EDA due to the latency of skin conductance response. Results suggest that facial micro muscles movements indicative of STEE were automatically detected using AFE. This is in line with previous research in which AFE was deemed as an appropriate tool to assess micro changes in facial action units (Cohn & Kanade, 2007). Considering the timepoints chosen for speech analysis, STEEs were most likely excluded as they could have occurred prior to a participant's vocal response. Hence, results indicate that physiological measures are more informative of three emotional dimensions in comparison to speech (**H3**), as physiological feature AFE-based valence best explains users' emotional states during voice user interface interactions. Thus, **H3 is supported**.

### 3.5.1 Theoretical contributions

As a result of this paper, five theoretical contributions can be noted. For one, current research regarding voice user interface evaluation gravitates around explicit methods, such as interviews, observations, diaries, and questionnaires (Easwara et al., 2014; Jiang et al., 2015; Lopatovska & Williams, 2018). Data obtained from explicit measures relying on self-reported measures can be flawed, as users are at risk of cognitive and retrospective biases (Ortiz de Guinea et al., 2014). By including implicit measures, our study avoids such biases while taking into account real-time, subconscious reactions linked to important emotional states (Ortiz de Guinea et al., 2014). Consequently, results from this study contribute to the understanding of underlying emotions lived by users interacting with voice user interfaces. Hence, the measures used to capture the emotional responses provoked by voice user interface interactions are both informative and complimentary to the current literature.

Secondly, few studies have observed the users' speech features during voice user interface interactions, and less have been done so in combination with physiological measures, as research within the study of emotion through speech tends to focus on single sensor data (Ali et al., 2018). Thus, utilizing multiple physiological measures within this field of research is of a rare occurrence. Recording multiple physiological measures further provides a more thorough understanding of the underlying emotions lived by users during such events, while allowing for the comparative strength of each measure's extracted feature in explaining emotional responses induced by voice user interfaces to be assessed. By isolating each measure, this study further confirms the indicative nature of speech and physiological features in assessing users' emotional responses, as suggested in previous emotion-centered research. Indeed, extracted physiological feature AFE-based valence and speech features spectral spread, log energy and F0 were indicative of the observed emotion dimensions. The relationship strength of these features in regard to assessing user emotions are in line with previous research (Zaman & Shrimpton-Smith, 2006; Mannepalli et al., 2018; Lausen & Hammerschmidt, 2020; Papakostas et al., 2017).

Thirdly, an important contribution of this study relates to the nuances of each measure's strength in explaining four distinct emotional dimensions, as it allowed for their effectiveness to be compared. As stressed previously, the effectiveness of physiological feature AFE-based valence surpasses all extracted features of both physiological and vocal nature. Indeed, its statistical relationship to valence, arousal, and STEE dimensions is significant and triumphant. Hence, results from this study contribute to the understanding of measurement effectiveness in assessing user emotions during voice user interface interactions.

Fourthly, in addition to exploring the dimensions of valence and arousal, this study considered control as an additional emotional dimension. Within speech literature, the dimension of control has received less attention in comparison to counterparts valence and arousal (Laukka et al., 2010; Szameitat et al., 2011). Thus, this study further contributes to the literature by observing this dimension. Unlike the valence and arousal dimensions, results suggest that the control dimension is best explained by speech feature spectral slope. Indeed, spectral slope had approximately three times more strength than extracted physiological feature EDA z-score in explaining control ratings. However, this relationship is the weakest amongst the observed dimensions, as the R-squared value was below 5%. Moreover, it was not considered statistically significant. Previous speech-emotion studies assessing the control dimension have been inconsistent. Result variances in F0, speech rate, and voice intensity have been noted (Laukka et al., 2010). Indeed, when observing the dimension of control in relation to spectral slope, research by Schröder et al. (2001) suggest that low dominance is accompanied by a flatter spectral slope, contrary to results obtained by Banse and Scherer (1996). With this said, we cannot conclude that the results from this study are in line with those from previous studies.

A final contribution is the methodological inclusion of fleeting emotions. By introducing the additional dimension of STEE, fleeting emotions were observed using a simple binary evaluation. By assessing temporary moments of authentic emotion, important glimpses into affective states were captured, which was especially important for subjects inclined to shy away from public displays of emotions. Future studies may benefit from this complimentary element to observe temporary yet relevant emotional events.

### **3.5.2 Practical Implications**

To our knowledge, no other study has compared the effectiveness between physiological and speech measures through their respective features in explaining user emotions provoked by voice user interface technologies. This novel study not only contributes to the literature regarding voice user interface technology but may also have managerial implications. Indeed, results from this study are particularly relevant within today's context, as the field of voice recognition continues to gain ground. The global voice recognition market size is expected to reach 27.16 billion U.S. dollars by 2026, an increase of 16.8 percent from 2020 (Statista, 2021). Consequently, various companies have adopted voice user interface technologies as a competitive advantage. For example, certain high-volume call centers have adopted voice recognition technology to better serve their customers, allowing them to navigate the menu's options in an autonomous, intuitive, and time-saving manner through speech command (Le Pailleur et al., 2020). To benefit from the success of this user-centric technology, early evaluation of such a product is key. Results from this study not only assist companies seeking to evaluate voice user interface products more efficiently, but also contribute to the underdeveloped guidelines of voice user interface evaluation. Put into context, limited resources may force a UX professional to select a single measure within their vocal product evaluation. Thus, understanding which measure is more informative of user emotions is a valuable insight, strategically and economically.

### **3.5.3 Conclusions**

The evaluation of voice user interface experiences is an emerging topic that is gaining ground as voice recognition technology continues to grow. The study presented sought to understand the emotional responses experienced by users during voice user interface interactions by observing and comparing the effectiveness of physiological and speech measures through their respective features. Our results depict a stronger correlation between the emotional dimensions and physiological measures in comparison to speech. More precisely, extracted physiological feature AFE-based valence best



explained user emotions. To sum up, the use of physiological measures can equip UX professionals with rich data regarding the emotional experiences lived by users during voice user interface interactions, which may contribute to the design of optimal experiences.

Our study is limited by the fact that it was conducted remotely. The instructions regarding the pose of sensors and the upload of the data to the cloud was provided by an experience moderator. However, the acts were ultimately committed by the participants. Hence, a lack of control and on-sight supervision might have played a role in the technical difficulties resulting in data loss. To counter these drawbacks, future studies should consider an in-person data collection. Moreover, our experiment was limited by the use of a Wizard of Oz technique, in which the moderator played sequential MP3 recordings uploaded to a Google slides presentation. Occasional recordings were accidentally played out of order or with a significant time-lapse in between them, which resulted in a less authentic interaction in comparison to that of an actual voice user interface. Hence, future studies featuring an authentic and functional voice user interface system should be considered. Furthermore, the scope of the present study was limited in that the speech features analyzed were not exhaustive. Further studies regarding the matter should consider other speech features in order to further explore the subject. On that note, different physiological measures and their respective features should also be included to pursue the study of user emotions during voice user interface interactions. Moreover, within the context of this study, the majority of emotional events investigated were related to negative user emotions, such as frustration. Future studies should consider a diversified set of emotions, both of positive and negative nature, in order to obtain a more holistic representation. Lastly, recorded EDA data during brief voice user interface interjections was considered for the analysis. The timepoints of concise and occasional one-worded questions may have affected the results regarding the relationship between the extracted EDA features in relation to a users' emotional intensity during voice user interface interactions. Considering the latency of skin conductance response, ranging between one and five seconds, (Christopoulos et al., 2016), in conjunction to the timepoints chosen, inductive electrodermal signals might have been excluded. Future research should either

consider changing the dialogue to limit brief questions or include the participant's response within the time window of EDA analysis.

## References

- Ali M., Mosa A.H., Machot F.A., Kyamakya K. (2018) Emotion Recognition Involving Physiological and Speech Signals: A Comprehensive Review. In: Kyamakya K., Mathis W., Stoop R., Chedjou J., Li Z. (eds) Recent Advances in Nonlinear Dynamics and Synchronization. *Studies in Systems, Decision and Control*, vol 109. Springer, Cham. [https://doi.org/10.1007/978-3-319-58996-1\\_13](https://doi.org/10.1007/978-3-319-58996-1_13)
- Alshamsi, H., Kepuska, V., Alshamsi, H., & Meng, H. (2019). Automated Facial Expression and Speech Emotion Recognition App Development on Smart Phones using Cloud Computing. *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2018, March 2019*, 730–738. <https://doi.org/10.1109/IEMCON.2018.8614831>
- Alves, R., Valente, P., & Nunes, N. J. (2014). The state of user experience evaluation practice. *Proceedings of the NordiCHI 2014: The 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, 93–102. <https://doi.org/10.1145/2639189.2641208>
- Apple, W., Streeter, L. A., & Krauss, R. M. (1979). Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology*, 37, 715-727.
- Arora, S., Baghai-Ravary, L., & Tsanas, A. (2019). Developing a large scale population screening tool for the assessment of Parkinson's disease using telephone-quality voice. *The Journal of the Acoustical Society of America*, 145(5), 2871-2884.
- Bachorowski, J.A. (1999). Vocal Expression and Perception of Emotion. *Current Directions in Psychological Science*, 8(2), 53–57. <https://doi.org/10.1111/1467-8721.00013>
- Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3), 614–636. <https://doi.org/10.1037//0022-3514.70.3.614>
- Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 19, 3-11.
- Bastien, J. M. C., & Scapin, D. L. (1992). A validation of ergonomic criteria for the evaluation of human-computer interfaces. *International Journal of Human-Computer Interaction*, 4, 183-196.
- Bethel, C. L., Salomon, K., Murphy, R. R., & Burke, J. L. (2007). Survey of

- psychophysiology measurements applied to human-robot interaction. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 732-737). IEEE.
- Betella, A., & Verschure, P. F. (2016). The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions. *PloS one*, *11*(2), e0148037. <https://doi.org/10.1371/journal.pone.0148037>
- Boucsein, W. (2012). *Electrodermal Activity*. Boston, MA: Springer US.
- Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (Vol. 30, No. 1, pp. 25-36). Technical report C-1, the center for research in psychophysiology, University of Florida.
- Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2013). A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology*, *49*(1), 1017-1034.
- Breitenstein, C., Van Lancker, D., Daum, I. (2001). The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. *Cognition and Emotion*, *15* (1), 57-79
- Burton-Jones, A., & Gallivan, M. J. (2007). Towards a deeper understanding of system usage in organizations. *MIS Quarterly*, *31*(4), 657-679.
- Busso, C., & Rahman, T. (2012). Unveiling the acoustic properties that describe the valence dimension. In *Thirteenth Annual Conference of the International Speech Communication Association*.
- Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., ... & Narayanan, S. (2004, October). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces* (pp. 205-211).
- Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Raouzaïou, A., & Karpouzis, K. (2006). Modeling naturalistic affective states via facial and vocal expressions recognition. *ICMI'06: 8th International Conference on Multimodal Interfaces, Conference Proceeding, June 2014*, 146-154. <https://doi.org/10.1145/1180995.1181029>
- Castellano, G., Kessous, L., & Caridakis, G. (2008). Emotion recognition through multiple modalities: face, body gesture, speech. In *Affect and emotion in human-computer interaction* (pp. 92-103). Springer, Berlin, Heidelberg.
- Christopoulos, G. I., Uy, M. A., & Yap, W. J. (2019). The Body and the Brain: Measuring Skin Conductance Responses to Understand the Emotional

Experience. *Organizational Research Methods*, 22(1), 394–420. <https://doi.org/10.1177/1094428116681073>

- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, 6(4), 284.
- Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., ... & R Cowan, B. (2019). The state of speech in HCI: Trends, themes and challenges. *Interacting with Computers*, 31(4), 349-371.
- Cohn, J. F., & Kanade, T. (2007). Use of automated facial image analysis for measurement of emotion expression. *Handbook of emotion elicitation and assessment*, 222-238.
- Cordaro, D. T., Keltner, D., Tshering, S., Wangchuk, D., & Flynn, L. M. (2016). The voice conveys emotion in ten globalized cultures and one remote village in Bhutan. *Emotion*, 16(1), 117–128. <https://doi.org/10.1037/emo0000100>
- Courtemanche, F., Léger, P.M., Fredette, M., Sénécal, S. (2022). COBALT - Photobooth: Système intégré de données UX, Declaration of invention No. VAL-0045, HEC Montréal, Montréal, Canada.
- Courtemanche, F., Léger, P.M., Fredette, M., Sénécal, S. (2022). COBALT - Bluebox: Système de synchronisation et d'acquisition sans-fil de données utilisateur multimodales., Declaration of invention No. AXE-0045, HEC Montréal, Montréal, Canada.
- Courtemanche, F., Fredette, M., Senecal, S., Leger, P. M., Dufresne, A., Georges, V., & Labonte-lemoyne, E. (2019). *U.S. Patent No. 10,368,741*. Washington, DC: U.S. Patent and Trademark Office.
- Davitz, J. R. (Ed.). (1964). *The communication of emotional meaning*. McGraw Hill.
- Dawson, M. E., Schell, A. M., & Filion, D. L. (2007). The electrodermal system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 159–181). Cambridge University Press. <https://doi.org/10.1017/CBO9780511546396.007>
- Dirican, A. C., & Göktürk, M. (2011). Psychophysiological Measures of Human Cognitive States Applied in Human Computer Interaction. *Procedia Computer Science*, 3, 1361- 1367. <https://doi.org/10.1016/j.procs.2011.01.016>
- Easwara Moorthy, A., & Vu, K.-P. L. (2015). Privacy Concerns for Use of Voice Activated Personal Assistant in the Public Space. *International Journal of Human-Computer Interaction*, 31(4), 307–335.

- Ekman, P., & Friesen, W. V. (1978). *The facial action coding system*. San Francisco, CA: Consulting Psychologists Press.
- Figner, B., & Murphy, R. O. (2011). Using skin conductance in judgment and decision making research. *A handbook of process tracing methods for decision research*, 163-184.
- Fujimura, T., Sato, W., & Suzuki, N. (2010). Facial expression arousal level modulates facial mimicry. *International journal of psychophysiology: official journal of the International Organization of Psychophysiology*, 76(2), 88–92.  
<https://doi.org/10.1016/j.ijpsycho.2010.02.008>
- Garg, R., & Moreno, C. (2019). Exploring Everyday Sharing Practices of Smart Speakers. In *IUI Workshops*.
- Giroux-Huppé, C., Sénécal, S., Fredette, M., Chen, S. L., Demolin, B., & Léger, P.-M. (2019). Identifying psychophysiological pain points in the online user journey: the case of online grocery. *Springer, Cham*, 459-473.
- Giroux, F., Léger, P. M., Briegne, D., Courtemanche, F., Bouvier, F., Chen, S. L., ... & Sénécal, S. (2021, July). Guidelines for collecting automatic facial expression detection data synchronized with a dynamic stimulus in remote moderated user tests. In *International Conference on Human-Computer Interaction* (pp. 243-254). Springer, Cham.
- Greco, A., Marzi, C., Lanata, A., Scilingo, E. P., & Vanello, N. (2019). Combining Electrodermal Activity and Speech Analysis towards a more Accurate Emotion Recognition System. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, 2019, 229–232.  
<https://doi.org/10.1109/EMBC.2019.8857745>
- Gross, J. J., & Muñoz Ricardo F. (1995). Emotion regulation and mental health. *Clinical Psychology: Science and Practice*, 2(2), 151–164. <https://doi.org/10.1111/j.1468-2850.1995.tb00036.x>
- Hallgren K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in quantitative methods for psychology*, 8(1), 23–34. <https://doi.org/10.20982/tqmp.08.1.p023>
- Hassenzahl, M., & Tractinsky, N. (2006). User Experience - a Research Agenda. *Behaviour & Information Technology*, 25(2), 91-97.  
<https://doi.org/10.1080/01449290500330331>
- Hura, S. L. (2017). Usability testing of spoken conversational systems. *Journal of*

*Usability Studies*, 12(4), 155-163.

- Ivonin, L., Chang, H.-M., Díaz, M., Català, A., Chen, W., & Rauterberg, M. (2014). Beyond Cognition and Affect: Sensing the Unconscious. *Behaviour & Information Technology*, 34(3), 220-238. <https://doi.org/10.1080/0144929X.2014.912353>
- Jessen, S., & Kotz, S. A. (2011). The temporal dynamics of processing emotions from vocal, facial, and bodily expressions. *NeuroImage*, 58(2), 665–674. <https://doi.org/10.1016/j.neuroimage.2011.06.035>
- Jiang, J., Hassan Awadallah, A., Jones, R., Ozertem, U., Zitouni, I., Gurunath Kulkarni, R., & Khan, O. Z. (2015). Automatic Online Evaluation of Intelligent Assistants. *Proceedings of the 24th International Conference on World Wide Web - WWW '15. Presented at the 24th International Conference*. <https://doi.org/10.1145/2736277.2741669>
- Johnstone, T., & Scherer, K. R. (2000). Vocal communication of emotion. *Handbook of emotions*, 2, 220-235
- Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129, 770–814. <http://dx.doi.org/10.1037/0033-2909.129.5.770>
- Kehrein, R. (2002). The prosody of authentic emotions. In *Speech Prosody 2002, International Conference*. DOI:10.1055/s-2003-40251
- Koh, Y., & Kwahk, J. (2017). B3-1 Analysis of User's Speech Behavior Pattern after Correction: focusing on Smartphone Voice User Interface. *The Japanese Journal of Ergonomics*, 53, 408-411.
- Kraus, M. W. (2017). Voice-only communication enhances empathic accuracy. *American Psychologist*, 72, 644–654. <http://dx.doi.org/10.1037/amp0000147>
- Lamontagne, C., Sénécal, S., Fredette, M., Chen, S. L., Pourchon, R., Gaumont, Y., ... & Léger, P. M. (2019, August). User Test: How Many Users Are Needed to Find the Psychophysiological Pain Points in a Journey Map? In *International Conference on Human Interaction and Emerging Technologies* (pp. 136-142). Springer, Cham.
- Laukka, P., Elfenbein, H. A., Thingujam, N. S., Rockstuhl, T., Iraki, F. K., Chui, W., & Althoff, J. (2016). The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features. *Journal of Personality and Social Psychology*, 111, 686–705. <http://dx.doi.org/10.1037/pspi0000066>
- Laukka, P., Juslin, P. N., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition and Emotion*, 19(5), 633–653.

<https://doi.org/10.1080/02699930441000445>

- Lausen, A., & Hammerschmidt, K. (2020). Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. *Humanities and Social Sciences Communications*, 7(1), 1-17.
- Léger, P.-M., Davis, F. D., Cronan, T. P., & Perret, J. (2014). Neurophysiological Correlates of Cognitive Absorption in an Enactive Training Context. *Computers in Human Behavior*, 34, 273-283. <https://doi.org/10.1016/j.chb.2014.02.011>
- Leite, I., Henriques, R., Martinho, C., & Paiva, A. (2013). Sensors in the wild: Exploring electrodermal activity in child-robot interaction. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 41-48). IEEE.
- Le Pailleur, F., Huang, B., Léger, P. M., & Sénécal, S. (2020). A new approach to measure user experience with voice-controlled intelligent assistants: A pilot study. In M. Kurosu (Ed.), *Human-computer interaction. Multimodal and natural interaction. HCII 2020. Lectures notes in computer science* (Vol. 12182, pp. 197–208). [https://doi.org/10.1007/978-3-030-49062-1\\_13](https://doi.org/10.1007/978-3-030-49062-1_13)
- Levin, H., & Lord, W. (1975). Speech pitch frequency as an emotional state indicator. *IEEE Transactions on Systems, Man, and Cybernetics*, 5, 259-273.
- Lewinski, P., den Uyl, T. M., & Butler, C. (2014). Automated facial coding: validation of basic emotions and FACS AUs in FaceReader. *Journal of Neuroscience, Psychology, and Economics*, 7(4), 227.
- Li S.Z., Jain A. (2009). Fundamental Frequency, Pitch, F0. *Encyclopedia of Biometrics*. Springer, Boston, MA. [https://doi.org/10.1007/978-0-387-73003-5\\_775](https://doi.org/10.1007/978-0-387-73003-5_775)
- Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE transactions on bio-medical engineering*, 56(4), 1015. <https://doi.org/10.1109/TBME.2008.2005954>
- Lopatovska, I., & Oropeza, H. (2018). User interactions with “Alexa” in public academic space. *Proceedings of the Association for Information Science and Technology*, 55(1), 309–318. <https://doi.org/10.1002/pra2.2018.14505501034>
- Lopatovska, I., & Williams, H. (2018). Personification of the Amazon Alexa: BFF or a mindless companion. *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (pp. 265-268). ACM.
- Mannepalli, K., Sastry, P. N., & Suman, M. (2018). Emotion recognition in speech signals using optimization based multi-SVNN classifier. *Journal of King Saud*

*University-Computer and Information Sciences.*  
<https://doi.org/10.1016/j.jksuci.2018.11.012>

- Murad, C., & Munteanu, C. (2020). Designing Voice Interfaces: Back to the (Curriculum) Basics. *Conference on Human Factors in Computing Systems - Proceedings*, 1–12. <https://doi.org/10.1145/3313831.3376522>
- Nielsen, J. (1994). Usability inspection methods. *In Conference companion on Human factors in computing systems* (pp. 413-414).
- Nowacki, C., Gordeeva, A., & Lizé, A. H. (2020). Improving the Usability of Voice User Interfaces: A New Set of Ergonomic Criteria. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12201 LNCS, 117–133. [https://doi.org/10.1007/978-3-030-49760-6\\_8](https://doi.org/10.1007/978-3-030-49760-6_8)
- Ortiz de Guinea, A., Titah, R., & Léger, P.-M. (2014). Explicit and implicit antecedents of users' behavioral beliefs in information systems: A neuropsychological investigation. *Journal of Management Information Systems*, 30(4), 179-210.
- Ortiz de Guinea, A., & Webster, J. (2013). An investigation of information systems use patterns: Technological events as triggers, the effect of time, and consequences for performance. *MIS Quarterly*, 37, 1165–1188.
- Papakostas, M., Siantikos, G., Giannakopoulos, T., Spyrou, E., & Sgouropoulos, D. (2017). Recognizing emotional states using speech information. In *GeNeDis 2016* (pp. 155-164). Springer, Cham.
- Patel, S., Scherer, K. R., Sundberg, J., & Björkner, E. (2010). Acoustic markers of emotions based on voice physiology. *Speech Prosody 2010*.
- Pereira, C. (2000). Dimensions of emotional meaning in speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.
- Pittam, J., Gallois, C., & Callan, V. (1990). The long-term spectrum and perceived emotion. *Speech Communication*, 9, 177-87.
- Prasetio, B. H., Tamura, H., & Tanno, K. (2020). Embedded Discriminant Analysis based Speech Activity Detection for Unsupervised Stress Speech Clustering. *2020 Joint 9th International Conference on Informatics, Electronics and Vision and 2020 4th International Conference on Imaging, Vision and Pattern Recognition, ICIEV and IcIVPR 2020*.  
<https://doi.org/10.1109/ICIEVicIVPR48672.2020.9306589>
- Provine, R. R., & Fischer, K. R. (1989). Laughing, smiling, and talking: Relation to sleeping and social context in humans. *Ethology*, 83, 295– 305.



- Riedl, R., & Léger, P. M. (2016). Fundamentals of NeuroIS. *Studies in neuroscience, psychology and behavioral economics*, 127. <https://doi.org/10.1007/978-3-662-45091-8>
- Robinson, C., Obin, N., & Roebel, A. (2019). Sequence-to-sequence modelling of f0 for speech emotion conversion. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6830-6834). IEEE.
- Russell, J.A. (1980) A circumplex model of affect. *Journal of Personality and Social Psychology*, 39, 1161–1178.
- Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99, 143–165.
- Scherer, K. R., & Oshinsky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion*, 1, 331-346.
- Scherer, K. R. (1974). Acoustic concomitants of emotional dimensions: Judging affect from synthesized tone sequences. In S. Weitz (Ed.), *Nonverbal communication* (pp. 105-111). New York: Oxford University Press
- Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., & Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. In *Seventh European Conference on Speech Communication and Technology*
- Sciuto, A., Saini, A., Forlizzi, J., & Hong, J. I. (2018). “Hey Alexa, What’s Up?” *Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS ’18*. <https://doi.org/10.1145/3196709.3196772>.
- Seaborn, K., & Urakami, J. (2021, May). Measuring Voice UX Quantitatively: A Rapid Review. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-8).
- Skiendziel, T., Rösch, A. G., & Schultheiss, O. C. (2019). Assessing the convergent validity between the automated emotion recognition software Noldus FaceReader 7 and Facial Action Coding System Scoring. *PloS one*, 14(10), e0223905.
- Sobol-Shikler, T. (2009). *Analysis of affective expression in speech* (No. UCAM-CL-TR-740). University of Cambridge, Computer Laboratory.
- Statista. (2021). Number of digital voice assistants in use worldwide from 2019 to 2024 (in billions)\* | Statista. <https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/>. (Accessed 10 July 2021).
- Sutton, T. M., Herbert, A. M., & Clark, D. Q. (2019). Valence, arousal, and dominance

- ratings for facial stimuli. *Quarterly Journal of Experimental Psychology*, 72(8), 2046–2055. <https://doi.org/10.1177/1747021819829012>.
- Szameitat, D. P., Darwin, C. J., Wildgruber, D., Alter, K., & Szameitat, A. J. (2011). Acoustic correlates of emotional dimensions in laughter: arousal, dominance, and valence. *Cognition and emotion*, 25(4), 599–611.
- Tahon, M., Degottex, G., & Devillers, L. (2012). Usual voice quality features and glottal features for emotional valence detection. *Proceedings of the 6th International Conference on Speech Prosody*, 2, 693–697.
- Toh, A. M., Togneri, R., & Nordholm, S. (2005). Spectral entropy as speech features for speech recognition. *Proceedings of PEECS*, 1, 92.
- Uldall, E. (1960). Attitudinal meanings conveyed by intonation contours. *Language and Speech*, 3, 223–234.
- Uyl, M. J. d., & Kuilenburg, H. v. (2005). The Facereader: Online Facial Expression Recognition. *Proceedings of Measuring Behavior*.
- Vasseur, A., Léger, P. M., Courtemanche, F., Labonte-Lemoyne, E., Georges, V., Valiquette, A., ... & Sénécal, S. (2021, July). Distributed remote psychophysiological data collection for UX evaluation: a pilot project. In *International Conference on Human-Computer Interaction* (pp. 255–267). Springer, Cham
- Vidrascu, L., & Devillers, L. (2005). Real-life emotion representation and detection in call centers data. In *International Conference on Affective Computing and Intelligent Interaction* (pp. 739–746). Springer, Berlin, Heidelberg.
- vom Brocke, J., Riedl, R., & Léger, P.-M. (2013). Application strategies for neuroscience in information systems design science research. *Journal of Computer Information Systems*, 53(3), 1–13.
- Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021). A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access*, 9, 47795–47814. <https://doi.org/10.1109/ACCESS.2021.3068045>
- Xue, Y., Hamada, Y., & Akagi, M. (2018). Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space. *Speech Communication*, 102(June), 54–67. <https://doi.org/10.1016/j.specom.2018.06.006>
- Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Deng, Z., Lee, S., ... & Busso, C. (2004). An acoustic study of emotions expressed in speech. In *Eighth International Conference on Spoken Language Processing*.

Zaman, B., & Shrimpton-Smith, T. (2006). The FaceReader: Measuring instant fun of use. *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles* (pp. 457-460). <https://doi.org/10.1145/1182475.1182536>

Zhu, C., & Ahmad, W. (2019). Emotion recognition from speech to improve human-robot interaction. *Proceedings - IEEE 17th International Conference on Dependable, Autonomic and Secure Computing, IEEE 17th International Conference on Pervasive Intelligence and Computing, IEEE 5th International Conference on Cloud and Big Data Computing, 4th Cyber Scienc, July*, 370–375. <https://doi.org/10.1109/DASC/PiCom/CBDCCom/CyberSciTech.2019.00076>

## **Chapter 4. Managerial Article**

### **Hey Alexa, what is the best approach to detect pain points induced by voice user interface interactions?**

Danya Swoboda, Pierre-Majorique Léger and Sylvain Sénécal  
HEC Montréal

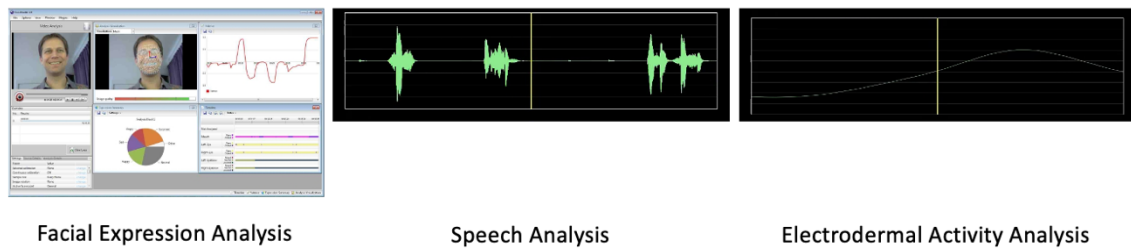
#### **Summary**

The rapid rise in popularity of voice user interface technologies calls for a reconsideration of the current traditional UX evaluation methods used to detect user pain points. Although some methods may work wonders for digital interfaces, they fall short within a voice user interface context. Hence, a dire need for adequate voice user interface evaluation methods prevails. The purpose of this article is to compare the effectiveness of novel physiological methods and speech to assess the emotion dimensions of valence, arousal, and control during voice user interface interactions. We define each method and present the strength of the top physiological and speech feature in explaining each emotion dimension. Building upon a recent scientific experiment conducted by our team, we conclude that the extracted physiological feature of automatic facial expression (AFE) based valence is the most explicative feature. As a result of this, UX professionals seeking to evaluate voice user interface technologies can better identify potential pain points causing emotional turmoil, which can consequently improve the user's experience.

#### **4.1 Introduction**

Hey Alexa, am I in a good mood? Delving into users' emotions is key to understanding their experience with a given product. As voice user interface technologies continue to grow in popularity, with the number voice assistants in use to reach the 8.4 billion mark by 2024<sup>1</sup>, assessing users' affective states during vocal interactions will increasingly become a topic of interest pivotal to voice user interface evaluation. With today's technology, Alexa can very well attempt to answer the question posed previously using speaker recognition, a process in which speaker-specific vocal features are extracted.

Indeed, the human voice is a rich and insightful medium of emotional communication. But is speech the best implicit measure to capture human emotions during voice user interface interactions? An acute ability to detect emotions induced by vocal interactions can potentially shed light upon user pain points. By identifying these pain points through voice user interface evaluations, UX professionals are better equipped to address them and consequently build better user experiences. With this said, our research set out to compare the effectiveness of speech and physiological measures employing electrodermal activity (EDA) and facial expression analysis in explaining lived emotions during voice user interface interactions that were purposely designed to elicit frustration and shock. Our experiment recorded speech and physiological responses of 16 participants, resulting in 188 analyzed interactions. In total, 11 distinct features were extracted and compared in their predictive ability to explain emotion dimensions, notably valence, arousal, and control. Our results suggest that the physiological feature of automatic facial expression (AFE) based valence is most informative of emotional events linked to pain points experienced during voice user interface interactions.



**Fig. 1.** Visual representation of the three data streams analyses, being facial expression, speech and electrodermal activity

## 4.2. Comparative strength of physiological and speech features per dimension

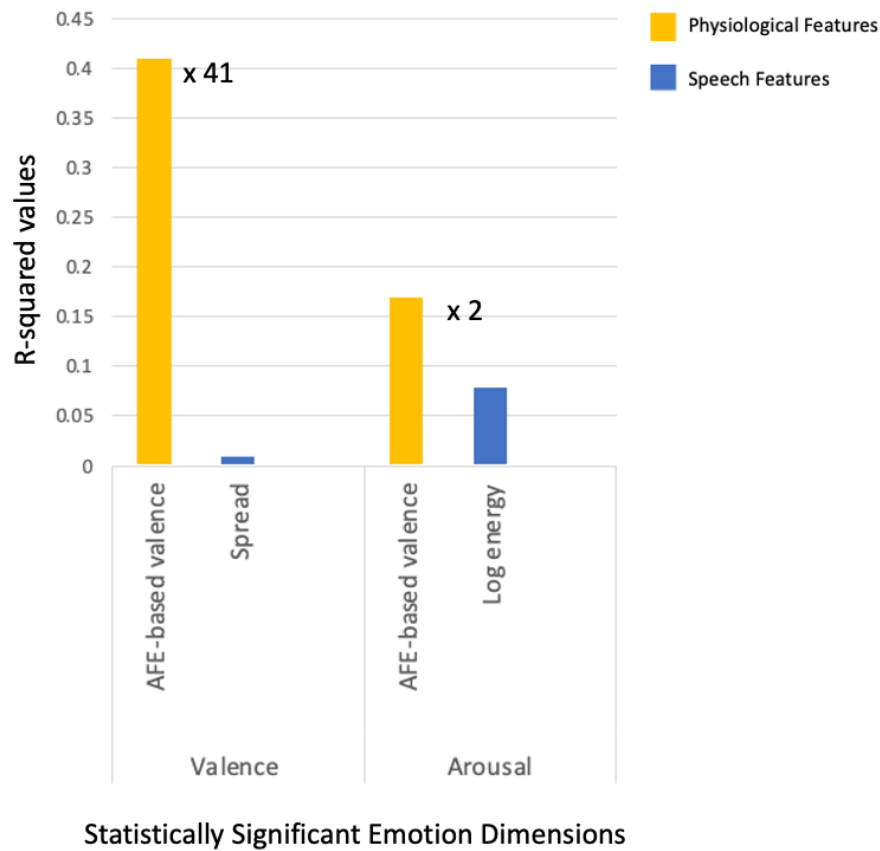
**Best at measuring the pleasantness (or unpleasantness) of speaking to a voice user interface: AFE-based valence**

Emotional valence refers to the pleasantness or unpleasantness of an emotional stimulus. Simply put, it explains how a user feels. Consequently, valence levels are indicative of a

user's experience, as positive valence levels have been linked to perceived usability, a contributing factor to user experience<sup>2</sup>. Often captured via webcam, facial micro expressions are generally quantified using some form of AFE analysis software and assessed through the lens emotional valence. This makes no exception within the context of our study, as results suggest that valence levels are best captured via AFE. As seen in Figure 2 below, the extracted physiological feature of AFE-based valence was approximately 41 times more powerful in explaining users' valence levels in comparison to the strongest extracted speech feature, being spectral spread. Notably, nearly 40 percent of the valence dimension variable was explained by AFE-based valence, making facial expression analysis the strongest implicit measure in assessing a user's valence levels during voice user interface interactions

#### **Best at measuring the intensity of emotions experience during voice user interface interactions: AFE-based valence**

Complimentary to a user's valence, arousal levels depict the intensity of one's emotions, providing once more key information into a user's experience. In explaining a users' arousal levels, our results suggest that speech features log energy, spectral slope and spectral spread can be employed, with log energy being the most explicative out of the three. However, not even the strongest speech feature is at par with AFE-based valence's effectiveness in explaining the arousal dimension. Indeed, AFE-based valence had nearly twice the strength of log energy in explaining arousal ratings, making it once more the most explicative implicit feature of a user's arousal during voice user interface interactions. Typically, EDA is an effective measure used to assess arousal levels, as the changes in skin pore dilation and sweat gland activation, often captured via electrodes on the palm of the hand, are informative of emotional intensity. However, the timepoints of analysis chosen for this study excluded indicative skin conductance signals. Brief and occasional single-worded interjections analyzed in the context of this study are therefore best assessed using facial expression analysis.



**Fig. 2.** Graphical representation of the top physiological and speech features explicative of the valence and arousal dimensions

### **Best at determining a user’s sense of control during voice user interface interactions: Neither**

In addition to the dimensions of valence and arousal, control, also referred to as power, potency or dominance, can be evaluated to better understand a person’s coping potential in a given situation. Although this particular dimension has received attention in emotion-related research, it has been argued that valence and arousal are sufficient alone in explaining basic emotions, resulting in its dismissal<sup>3,4</sup>. Within the context of our study, neither speech nor physiological features had the ability to significantly explain a user’s control levels during emotional voice user interface interactions. This is not particularly surprising, as the control dimension’s effects have been said to be inconsistent across studies<sup>5</sup>.

## And the winner is...

Overall, the extracted physiological feature of AFE-based valence best explains users' valence and arousal levels in comparison to speech features in a voice user interface context. As suggested, valence and arousal provide a holistic portrait of a user's emotional state. Although speech may be an obvious choice given the vocal nature of the interface, UX professionals seeking to evaluate voice user interfaces should consider facial expression as a non-intrusive measure due to its effectiveness in explaining user emotions, while maintaining an ecologically valid vocal interaction. Understanding the unique effectiveness of implicit measures and their respective features can save time, money and resources, favouring an efficient voice user interface evaluation. Knowing this key insight, Alexa would most likely say you're in a good mood.

## Notes

<sup>1</sup> Statista. (2021). Number of digital voice assistants in use worldwide from 2019 to 2024 (in billions) \* | Statista.

<https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/>

<sup>2</sup> Kwang-Kyu S, Sangwon L, Byung Do C & Changsoon P (2015). Users' Emotional Valence, Arousal, and Engagement Based on Perceived Usability and Aesthetics for Web Sites. *International Journal of Human-Computer Interaction*, 31(1), 72-87. <https://doi.org/10.1080/10447318.2014.959103>

<sup>3</sup> Betella, A., & Verschure, P. F. M. J. (2016). The affective slider: A digital self-assessment scale for the measurement of human emotions. *PLoS ONE*, 11(2), 1–11. <https://doi.org/10.1371/journal.pone.0148037>

<sup>4</sup> Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76(5), 805–819. <https://doi.org/10.1037/0022-3514.76.5.805>

<sup>5</sup> Libkuman, T. M., Otani, H., Kern, R., Viger, S. G., & Novak, N. (2007). Multidimensional normative ratings for the international affective picture system. *Behavior research methods*, 39(2), 326-334. <https://doi.org/10.3758/BF03193164>



## Conclusion

The rise of voice recognition technology has undoubtedly transformed the ways in which we interact with machines, services, and systems. Bypassing the common need to type, read or think, users are promised simpler and effortless interactions thanks to voice user interface technologies. To fulfill these promises, designers must resort to voice user interface evaluations. Beyond a utilitarian need, evaluations may shed light upon the users' lived emotions while interacting with voice user interfaces. These emotions are key to understand user experiences. Despite this, the evaluation of voice user interfaces is an emerging concept far from the maturity and established guidelines surrounding digital interface evaluations. Thus, as the growth in popularity of voice user interfaces perpetuates its omnipresence, an important need to properly evaluate these technologies arises.

This study sought to better understand the underlying emotions of users provoked by voice user interface interactions, while comparing the effectiveness of extracted features derived from physiological and speech measures in doing so. During the months of November and December 2020, as well as January 2021, remote experiments were conducted upon 29 French-speaking participants. Each participant was subject to a series of questions posed by a simulated voice user interface designed to evoke intense emotional responses through repeated comprehension errors and provocative questions. To assess these emotional reactions, physiological data, from AFE facial expression and EDA, alongside the participants' speech, were analyzed. To establish ground-truth for the physiological and speech features derived from user responses to the voice user interface, third-party emotion evaluations were conducted, in which emotional dimensions of valence, arousal, control, and short-term emotional episodes were assessed. Prior to the evaluations, the third-party evaluators were trained to judge emotional reactions through the lens of the selected dimensions, notably using the SAM scale. The statistical relationships between the ground-truth, and extracted physiological and speech features were assessed,

resulting in distinct strength, or R-squared values, in explaining emotional responses induced by voice user interface interactions.

The sections below relate to the research questions and key results of the study. Following this, the theoretical contributions and practical implications of this study will be addressed. The chapter will conclude with a section on the study's limits and future possible research avenues.

### **Reminder of the research questions and key results**

Through our research, we compared the extent to which intense emotional responses during voice user interface interactions were effectively detected using various implicit measures. Data obtained through physiological measures, being facial expression and EDA, as well as via speech, allowed us to extract and compare the strength of 11 distinct features. The processed data allowed us to assess the following question:

***RQ1:** Between speech and physiological features, which are more informative in assessing intense emotional responses during vocal interactions with a voice user interface?*

Results indicate that physiological feature AFE-based valence was most informative in explaining intense emotional responses provoked by voice user interface interactions. More precisely, AFE-based valence was informative of the emotional dimensions of valence, arousal, and STEE ratings. No factor was considered statistically significant in predicting control ratings. AFE-based valence was particularly noteworthy in explaining the emotional dimension of valence, as its relationship strength not only surpassed the strengths of the observed speech features but exceeded all other feature and dimension combinations. Thus, the results enabled us to address our second research question:

***RQ2:** Can we unobtrusively identify an intense emotional response during voice user interface interactions?*

Indeed, results from this study indicate that implicit measures employing physiology and speech are capable of unobtrusively capturing underlying user emotions in a voice user interface context. Although their strengths differ, extracted physiological and speech features are informative of intense emotional responses experienced during voice user interface interactions. Results support our first hypothesis, which supposed a relationship between speech features and intense emotional responses during vocal interactions with voice user interfaces (**H1**). In line with HCI literature, the observed emotional events were assessed through the dimensional lens of valence and arousal. The amplitude of the extracted AFE-based valence feature suggested a relationship with users' emotional intensity during voice user interface interactions, as the feature was indicative of valence levels (**H2.b**). However, within the context of our study, the amplitude of the extracted EDA features did not suggest a relationship with users' emotional intensity during voice user interface interactions, as AFE-based valence was most indicative of arousal levels (**H2.a**). These results were supported by literature, suggesting the revealing abilities of facial micro-expressions in valence and arousal levels. On the other hand, literature regarding the link between speech and the dimension of valence was inconsistent. This inconsistency led us to suppose that physiological features are more explicative of emotional voice interaction events in comparison to speech features. Indeed, the strength of physiological and speech features in explaining these emotional events differs. Results suggest that extracted physiological feature of AFE-based valence surpasses the strength of all extracted speech features in explaining the intensity of users' emotions experienced during voice user interface interactions.

## **Theoretical Contributions**

Theoretically speaking, our research contributes to the emerging study of voice user interface evaluation. Much of the current research regarding the subject employs explicit methods, notably interviews, observations, diaries, and questionnaires, to assess users' emotions during voice user interface interactions (Easwara et al., 2014; Jiang et al., 2015; Lopatovska & Williams, 2018). Although these informative methods may provide valuable information, users subject to these methods may succumb to various biases. To counter this, implicit methods can be used. By employing not one but three implicit

measures to assess users' affective states, our study compliments previous research while offering key insights in the matter of subconscious and underlying emotions experienced by users during voice user interface interactions. By observing the unique effectiveness of each implicit measure's extracted features in assessing these emotional events, our research addresses an important gap within literature. Indeed, understanding the strength of extracted speech features against physiological features in explaining user emotions has yet to have been studied within a voice user interface context. Results from this study depict the powerful nature of AFE-based valence in explaining emotional events experienced during voice user interface interactions. This is particularly noteworthy, as speech may appear as an obvious choice to evaluate voice user interface due to the vocal nature of the interface. However, our results suggest that AFE-based valence is better suited to assess these emotional events, as the extracted physiological feature proved to be the strongest in predicting all examined dimensions, with the exception of control. The uniqueness of the control dimension not only comes from its exceptional results, but from the fact that the dimension is seldom studied in speech research. Hence, the dimension's inclusion further contributes to speech literature. As for the STEE dimension, its novelty and value in capturing fleeting emotions can serve as a methodological contribution for future studies.

## **Practical Implication**

The managerial article presented previously sought to address UX professionals interested in voice user interface evaluation. As stressed, understanding the effectiveness of implicit measures in capturing user emotions induced by a voice user interface provides practical industry implications. As the growth and importance of voice user interface technologies prevails, UX practitioners are forced to forgo traditional digital interface know-how and adapt to no-UI ways. However, current guidelines regarding voice user interface evaluation are underdeveloped. Results from this study further contributes to the subject by suggesting the effectiveness of physiological features derived from EDA and facial expression, alongside speech features, in assessing the intensity of user emotions experienced during voice user interface interactions. As addressed in the professional article, comparing the effectiveness of these features can

guide UX professionals in selecting the best suited measure for voice user interface evaluation. Notably, AFE-based valence is the most effective at explaining the dimensions of valence and arousal, making it the best suited feature in understanding user emotions experienced during voice user interface interactions. This is particularly noteworthy, as limited financial or human resources within a business can restrict the course of an evaluation. Thus, our results offer key and actionable insight that favours efficient voice user interface evaluation, while contributing to the underdeveloped guidelines surrounding the subject.

### **Limits and future research avenues**

Although the study successfully drew noteworthy results, there are limits to take note of. For one, due to the COVID-19 pandemic, the data collection was conducted remotely. As a result of this, the pose of sensors and upload of the data to the cloud was performed by the participant. Technical difficulties linked to the data collection and synchronization ultimately resulted in data loss. An in-person experiment favouring control and on-sight supervision might have resulted in greater data. Moreover, due to resource and time constraints, a Wizard of Oz approach was adopted. Simulated voice user interface interactions, in which MP3 recordings were played sequentially, were humanly controlled by a moderator. Occasional recording errors and significant time-lapse between recordings resulted in a less than authentic interaction in comparison to an actual voice user interface system. To counter this, a functional voice user interface system should be employed in future studies. To elicit strong emotional responses, the majority of voice user interface questions and replies induced negative user emotions, such as frustration. Future studies should consider a diverse set of emotions, both of positive and negative nature, for a more holistic approach resulting in a greater understanding of user emotions during vocal interactions. Another limit to this study draws from the fact that the observed speech features list was not exhaustive. Future studies should consider other complimentary speech features that may further contribute to the literature regarding voice user interface evaluation. Within the same vein, different physiological measures, such as electroencephalogram and electrocardiogram, may further provide insightful information in regard to users' lived emotions during vocal experiences. Lastly, the

timepoints of EDA analysis were marked by brief and occasional one-worded voice user interface questions, which may have affected the results regarding the relationship between the amplitude of extracted EDA features and the users' emotional intensity during voice user interface interactions. Future studies should consider the latency of skin conductance response in relation to the analysis timepoints, by either excluding brief questions or including the participant's response within the time window of EDA analysis.



## Bibliography

- Agourram, H., Alvarez, J., Sénécal, S., Lachize, S., Gagné, J., & Léger, P. M. (2019). The relationship between technology self-efficacy beliefs and user satisfaction–user experience perspective. *International Conference on Human-Computer Interaction* (pp. 389-397). Springer, Cham.
- Alaszewski, A. (2006). *Using diaries for social research*. SAGE Publications Ltd  
<https://www.doi.org/10.4135/9780857020215>
- Ali M., Mosa A.H., Machot F.A., Kyamakya K. (2018) Emotion Recognition Involving Physiological and Speech Signals: A Comprehensive Review. In: Kyamakya K., Mathis W., Stoop R., Chedjou J., Li Z. (eds) Recent Advances in Nonlinear Dynamics and Synchronization. *Studies in Systems, Decision and Control*, vol 109. Springer, Cham. [https://doi.org/10.1007/978-3-319-58996-1\\_13](https://doi.org/10.1007/978-3-319-58996-1_13)
- Alshamsi, H., Kepuska, V., Alshamsi, H., & Meng, H. (2019). Automated Facial Expression and Speech Emotion Recognition App Development on Smart Phones using Cloud Computing. *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2018, March 2019*, 730–738. <https://doi.org/10.1109/IEMCON.2018.8614831>
- Alves, R., Valente, P., & Nunes, N. J. (2014). The state of user experience evaluation practice. *Proceedings of the NordiCHI 2014: The 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, 93–102.  
<https://doi.org/10.1145/2639189.2641208>
- Andreassi, J. L. (2000). *Psychophysiology: Human behavior and physiological response* (4th ed.). Lawrence Erlbaum Associates Publishers.
- Apple, W., Streeter, L. A., & Krauss, R. M. (1979). Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology*, 37, 715-727.
- Arora, S., Baghai-Ravary, L., & Tsanas, A. (2019). Developing a large scale population screening tool for the assessment of Parkinson's disease using telephone-quality voice. *The Journal of the Acoustical Society of America*, 145(5), 2871-2884.
- Bachorowski, J.A. (1999). Vocal Expression and Perception of Emotion. *Current Directions in Psychological Science*, 8(2), 53–57. <https://doi.org/10.1111/1467-8721.00013>
- Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3), 614–636.  
<https://doi.org/10.1037//0022-3514.70.3.614>



- Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 19, 3-11.
- Bastien, J. M. C., & Scapin, D. L. (1992). A validation of ergonomic criteria for the evaluation of human-computer interfaces. *International Journal of Human-Computer Interaction*, 4, 183-196.
- Battarbee, K., & Koskinen, I. (2005). Co-experience: user experience as interaction. *CoDesign*, 1(1), 5–18. <https://doi.org/10.1080/15710880412331289917>
- Beauchesne, A., Sénécal, S., Fredette, M., Chen, S. L., Demolin, B., Di Fabio, M. L., & Léger, P. M. (2019, July). User-centered gestures for mobile phones: exploring a method to evaluate user gestures for UX designers. In *International Conference on Human-Computer Interaction* (pp. 121-133). Springer, Cham.
- Beaudry, A., & Pinsonneault, A. (2010). The other side of acceptance: Studying the direct and indirect effects of emotions on information technology use. *MIS quarterly*, 689-710.
- Bentley, T., Johnston, L., & von Baggo, K. (2005, November). Evaluation using cued-recall debrief to elicit information about a user's affective experiences. In *Proceedings of the 17th Australia Conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future* (pp. 1-10).
- Betella, A., & Verschure, P. F. (2016). The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions. *PloS one*, 11(2), e0148037. <https://doi.org/10.1371/journal.pone.0148037>
- Bethel, C. L., Salomon, K., Murphy, R. R., & Burke, J. L. (2007). Survey of psychophysiology measurements applied to human-robot interaction. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 732-737). IEEE.
- Black, A. (1998). Empathic design: User focused strategies for innovation. In *Proceedings of the Conference on New Product Development* (pp. 1-8). London, UK: IBC
- Borkenau, P., & Ostendorf, F. (1987). Retrospective Estimates of Act Frequencies: How Accurately Do They Reflect Reality? *Journal of Personality and Social Psychology*, 52(3), 626-638. <https://doi.org/10.1037/0022-3514.52.3.626>
- Boucsein, W. (2012). *Electrodermal Activity*. Boston, MA: Springer US.
- Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (Vol. 30, No. 1, pp. 25-36). Technical

report C-1, the center for research in psychophysiology, University of Florida.

- Bradley MM, Lang PJ. (1994) Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*. Mar; 25(1):49–59. [https://doi.org/10.1016/0005-7916\(94\)90063-9](https://doi.org/10.1016/0005-7916(94)90063-9)
- Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2013). A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology*, 49(1), 1017-1034.
- Brave, S., & Nass, C. (2002). Emotion in human-computer interaction. In *The human-computer interaction handbook* (pp. 53-68). CRC Press.  
<https://doi.org/10.1201/b10368-6>
- Breitenstein, C., Van Lancker, D., Daum, I. (2001). The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. *Cognition and Emotion*, 15 (1), 57–79
- Bruun, A., & Ahm, S. (2015). Mind the gap! Comparing retrospective and concurrent ratings of emotion in user experience evaluation. In *IFIP Conference on Human-Computer Interaction* (pp. 237-254). Springer, Cham.
- Buchenau, M., Fulton Suri, J. (2000). Experience prototyping. In *Proceedings of DIS 2000 (Designing Interactive Systems)*, 424–433.
- Burton-Jones, A., & Gallivan, M. J. (2007). Towards a deeper understanding of system usage in organizations. *MIS Quarterly*, 31(4), 657–679.
- Busso, C., & Rahman, T. (2012). Unveiling the acoustic properties that describe the valence dimension. In *Thirteenth Annual Conference of the International Speech Communication Association*.
- Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., ... & Narayanan, S. (2004, October). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces* (pp. 205-211).
- Cacioppo, J. T., Berntson, G. G., Larsen, J. T., Poehlmann, K. M., & Ito, T. A. (2000). The psychophysiology of emotion. *Handbook of emotions*, 2(01). 173-191.
- Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Raouzaoui, A., & Karpouzis, K. (2006). Modeling naturalistic affective states via facial and vocal expressions recognition. *ICMI'06: 8th International Conference on Multimodal Interfaces, Conference Proceeding, June 2014*, 146–154.  
<https://doi.org/10.1145/1180995.1181029>

- Castellano, G., Kessous, L., & Caridakis, G. (2008). Emotion recognition through multiple modalities: face, body gesture, speech. In *Affect and emotion in human-computer interaction* (pp. 92-103). Springer, Berlin, Heidelberg.
- Chernykh, V., Sterling, G., & Prihodko, P. (2017). Emotion Recognition From Speech With Recurrent Neural Networks. *ArXiv, abs/1701.08071*.
- Christopoulos, G. I., Uy, M. A., & Yap, W. J. (2019). The Body and the Brain: Measuring Skin Conductance Responses to Understand the Emotional Experience. *Organizational Research Methods*, 22(1), 394–420. <https://doi.org/10.1177/1094428116681073>
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, 6(4), 284.
- Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., ... & R Cowan, B. (2019). The state of speech in HCI: Trends, themes and challenges. *Interacting with Computers*, 31(4), 349-371.
- Cohen, M. H., Cohen, M. H., Giangola, J. P., & Balogh, J. (2004). *Voice user interface design*. Addison-Wesley Professional.
- Cohn, J. F., & Kanade, T. (2007). Use of automated facial image analysis for measurement of emotion expression. *Handbook of emotion elicitation and assessment*, 222-238.
- Cordaro, D. T., Keltner, D., Tshering, S., Wangchuk, D., & Flynn, L. M. (2016). The voice conveys emotion in ten globalized cultures and one remote village in Bhutan. *Emotion*, 16(1), 117–128. <https://doi.org/10.1037/emo0000100>
- Courtemanche, F., Léger, P.M., Fredette, M., Sénécal, S. (2022). COBALT - Photobooth: Système intégré de données UX, Declaration of invention No. VAL-0045, HEC Montréal, Montréal, Canada.
- Courtemanche, F., Léger, P.M., Fredette, M., Sénécal, S. (2022). COBALT - Bluebox: Système de synchronisation et d'acquisition sans-fil de données utilisateur multimodales., Declaration of invention No. AXE-0045, HEC Montréal, Montréal, Canada.
- Courtemanche, F., Fredette, M., Senecal, S., Leger, P. M., Dufresne, A., Georges, V., & Labonte-lemoyne, E. (2019). *U.S. Patent No. 10,368,741*. Washington, DC: U.S. Patent and Trademark Office.
- Cowen, A. S., Elfenbein, H. A., Laukka, P., & Keltner, D. (2019). Mapping 24 emotions conveyed by brief human vocalization. *American Psychologist*, 74(6), 698–712. <https://doi.org/10.1037/amp0000399>

- Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York, NY: Grosset/Putnam.
- Dandavate, U., Sanders, E.B.-N. and Stuart, S., (1996). Emotions matter: user empathy in the product development process, in *Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting*, 415–418.
- Davitz, J. R. (Ed.). (1964). *The communication of emotional meaning*. McGraw Hill.
- Dawson, M. E., Schell, A. M., & Filion, D. L. (2007). The electrodermal system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 159–181). Cambridge University Press. <https://doi.org/10.1017/CBO9780511546396.007>
- Dewey, J. A., & Knoblich, G. (2014). Do implicit and explicit measures of the sense of agency measure the same thing?. *PloS one*, 9(10), e110118. <https://doi.org/10.1371/journal.pone.0110118>
- Dewey, J. (1934). The Supreme Intellectual Obligation. *Science*, 79 (2046), 240–243. <https://doi.org/10.1002/sce.3730180102>
- De Singly, F. (2016). *Le questionnaire* (4e édition). Armand Colin.
- Dirican, A. C., & Göktürk, M. (2011). Psychophysiological Measures of Human Cognitive States Applied in Human Computer Interaction. *Procedia Computer Science*, 3, 1361- 1367. <https://doi.org/10.1016/j.procs.2011.01.016>
- Easwara Moorthy, A., & Vu, K.-P. L. (2015). Privacy Concerns for Use of Voice Activated Personal Assistant in the Public Space. *International Journal of Human-Computer Interaction*, 31(4), 307–335.
- Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4), 169-200.
- Ekman, P., & Friesen, W. V. (1978). *The facial action coding system*. San Francisco, CA: Consulting Psychologists Press.
- Fagherazzi, G., Fischer, A., Ismael, M., & Despotovic, V. (2021). Voice for health: the use of vocal biomarkers from research to clinical practice. *Digital Biomarkers*, 5(1), 78–88. <https://doi.org/10.1159/000515346>
- Fan, M., Shi, S., & Truong, K. N. (2020). Practices and challenges of using think-aloud protocols in industry: An international survey. *Journal of Usability Studies*, 15(2), 85–102.

- Feldman Barrett, L., & Russell, J. A. (1998). Independence and bipolarity in the structure of current affect. *Journal of personality and social psychology*, 74(4), 967. <https://doi.org/10.1037/0022-3514.74.4.967>
- Figner, B., & Murphy, R. O. (2011). Using skin conductance in judgment and decision making research. *A handbook of process tracing methods for decision research*, 163-184.
- Forlizzi, J., & Battarbee, K. (2004). Understanding experience in interactive systems. In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques* (pp. 261-268). <https://doi.org/10.1145/1013115.1013152>
- Fujimura, T., Sato, W., & Suzuki, N. (2010). Facial expression arousal level modulates facial mimicry. *International journal of psychophysiology: official journal of the International Organization of Psychophysiology*, 76(2), 88–92. <https://doi.org/10.1016/j.ijpsycho.2010.02.008>
- Garg, R., & Moreno, C. (2019). Exploring Everyday Sharing Practices of Smart Speakers. In *IUI Workshops*.
- Ganglbauer, E., Schrammel, J., Deutsch, S., Tscheligi, M. (2009). Applying Psychophysiological Methods for Measuring User Experience: Possibilities, Challenges and Feasibility. *Workshop on User Experience Evaluation Methods in Product Development*.
- Gendron, M., Lindquist, K. A., Barsalou, L., & Barrett, L. F. (2012). Emotion words shape emotion percepts. *Emotion*, 12(2), 314.
- Gentile, C., Spiller, N., Noci, G. (2007). How to Sustain the Customer Experience: An Overview of Experience Components that Co-create Value With the Customer. *European Management Journal*, 25, 395-410.
- Giroux-Huppé, C., Sénécal, S., Fredette, M., Chen, S. L., Demolin, B., & Léger, P.-M. (2019). Identifying psychophysiological pain points in the online user journey: the case of online grocery. *Springer, Cham*, 459-473.
- Giroux, F., Léger, P. M., Briegne, D., Courtemanche, F., Bouvier, F., Chen, S. L., ... & Sénécal, S. (2021, July). Guidelines for collecting automatic facial expression detection data synchronized with a dynamic stimulus in remote moderated user tests. In *International Conference on Human-Computer Interaction* (pp. 243-254). Springer, Cham.
- Gothelf, J., & SEIDEN, J. (2013). *Lean UX: Applying lean principles to improve user experience*. "O'Reilly Media.

- Grimm, M., Kroschel, K., Mower, E., & Narayanan, S. (2007). Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10–11), 787–800. <https://doi.org/10.1016/j.specom.2007.01.010>
- Gross, J. J., & Muñoz Ricardo F. (1995). Emotion regulation and mental health. *Clinical Psychology: Science and Practice*, 2(2), 151–164. <https://doi.org/10.1111/j.1468-2850.1995.tb00036.x>
- Gu, S., Wang, F., Patel, N. P., Bourgeois, J. A., & Huang, J. H. (2019). A model for basic emotions using observations of behavior in *Drosophila*. *Frontiers in Psychology*, 10(APR), 1–13. <https://doi.org/10.3389/fpsyg.2019.00781>
- Hallgren K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in quantitative methods for psychology*, 8(1), 23–34. <https://doi.org/10.20982/tqmp.08.1.p023>
- Hartmann, K., Siegert, I., Philippou-Hübner, D., & Wendemuth, A. (2013). Emotion detection in HCI: From speech features to emotion space? *IFAC Proceedings Volumes (IFAC-PapersOnline)*, 12(PART 1), 288–295. <https://doi.org/10.3182/20130811-5-US-2037.00049>
- Hassenzahl, M. (2008). Aesthetics in Interactive Products: Correlates and Consequences of Beauty. *Elsevier*, 1, 287–302.
- Hassenzahl, M., & Tractinsky, N. (2006). User Experience - a Research Agenda. *Behaviour & Information Technology*, 25(2), 91–97. <https://doi.org/10.1080/01449290500330331>
- Hura, S. L. (2017). Usability testing of spoken conversational systems. *Journal of Usability Studies*, 12(4), 155–163.
- Ittelson, W. H., et al. (1970). The use of behavioural maps in environmental psychology. In H. M. Prohansky, W. H. Ittelson, L. G. Rivlin (Eds.), *Environmental Psychology: Man and his Physical Setting*, Holt (pp. 658–668). New York: Rinehart & Winston.
- Ivonin, L., Chang, H.-M., Díaz, M., Català, A., Chen, W., & Rauterberg, M. (2014). Beyond Cognition and Affect: Sensing the Unconscious. *Behaviour & Information Technology*, 34(3), 220–238. <https://doi.org/10.1080/0144929X.2014.912353>
- ISO FDIS 9241-210 (2009) *Human-centred design process for interactive systems*. | ISO. <https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-1:v1:en>. (Accessed 2 July 2021)
- Jessen, S., & Kotz, S. A. (2011). The temporal dynamics of processing emotions from vocal, facial, and bodily expressions. *NeuroImage*, 58(2), 665–674. <https://doi.org/10.1016/j.neuroimage.2011.06.035>

- Jiang, J., Hassan Awadallah, A., Jones, R., Ozertem, U., Zitouni, I., Gurunath Kulkarni, R., & Khan, O. Z. (2015). Automatic Online Evaluation of Intelligent Assistants. *Proceedings of the 24th International Conference on World Wide Web - WWW '15. Presented at the 24th International Conference*.  
<https://doi.org/10.1145/2736277.2741669>
- John, O. P., & Robins, R. W. (1994). Accuracy and bias in self-perception: individual differences, self- enhancement and the role of narcissism. *Journal of Personality and Social Psychology*, 66,206–219.
- Johnstone, T., & Scherer, K. R. (2000). Vocal communication of emotion. *Handbook of emotions*, 2, 220-235
- Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129, 770–814. <http://dx.doi.org/10.1037/0033-2909.129.5.770>
- Kankainen, A. (2002). *Thinking model and tools for understanding user experience related to information appliance product concepts*. Helsinki University of Technology.
- Keltner, D., Tracy, J. L., Sauter, D., & Cowen, A. (2019). What Basic Emotion Theory Really Says for the Twenty-First Century Study of Emotion. *Journal of nonverbal behavior*, 43(2), 195–201. <https://doi.org/10.1007/s10919-019-00298-y>
- Kent, R.D. (1997). *The speech sciences*. Singular Publishing.
- Kehrein, R. (2002). The prosody of authentic emotions. In *Speech Prosody 2002, International Conference*. DOI:[10.1055/s-2003-40251](https://doi.org/10.1055/s-2003-40251)
- Kiseleva, J., Williams, K., Jiang, J., Hassan Awadallah, A., Crook, A. C., Zitouni, I., & Anastasakos, T. (2016). Understanding user satisfaction with intelligent assistants. *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval* (pp. 121-130). ACM.
- Kleinsmith, A., & Bianchi-Berthouze, N. (2012). Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4(1), 15-33.
- Kocaballi, A. B., Laranjo, L., & Coiera, E. (2019). Understanding and Measuring User Experience in Conversational Interfaces. *Interacting with Computers*, 31(2), 192–207. <https://doi.org/10.1093/iwc/iwz01>
- Koh, Y., & Kwahk, J. (2017). B3-1 Analysis of User's Speech Behavior Pattern after Correction: focusing on Smartphone Voice User Interface. *The Japanese Journal of Ergonomics*, 53, 408-411.

- Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the Accuracy of Personality Judgments by the Self and Knowledgeable Others. *Journal of Personality*, 64(2), 311-337. doi:10.1111/j.1467-6494.1996.tb00513.x
- Kollia, V. (2016). Personalization Effect on Emotion Recognition from Physiological Data: An Investigation of Performance on Different Setups and Classifiers. *ArXiv*, abs/1607.05832.
- Kraus, M. W. (2017). Voice-only communication enhances empathic accuracy. *American Psychologist*, 72, 644–654. <http://dx.doi.org/10.1037/amp0000147>
- Kreibig S. D. (2010). Autonomic nervous system activity in emotion: a review. *Biological psychology*, 84(3), 394–421. <https://doi.org/10.1016/j.biopsycho.2010.03.010>
- Krosnick, J. A. (1999). Survey Research. *Annual Review of Psychology*, 50, 537-567.
- Kwang-Kyu S, Sangwon L, Byung Do C & Changsoon P (2015). Users' Emotional Valence, Arousal, and Engagement Based on Perceived Usability and Aesthetics for Web Sites. *International Journal of Human-Computer Interaction*, 31(1), 72-87. <https://doi.org/10.1080/10447318.2014.959103>
- Lamontagne, C., Sénécal, S., Fredette, M., Chen, S. L., Pourchon, R., Gaumont, Y., ... & Léger, P. M. (2019, August). User Test: How Many Users Are Needed to Find the Psychophysiological Pain Points in a Journey Map? In *International Conference on Human Interaction and Emerging Technologies* (pp. 136-142). Springer, Cham.
- Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1998). Emotion, motivation, and anxiety: brain mechanisms and psychophysiology. *Biological psychiatry*, 44(12), 1248–1263. [https://doi.org/10.1016/s0006-3223\(98\)00275-3](https://doi.org/10.1016/s0006-3223(98)00275-3)
- Lau, J., Zimmerman, B., & Schaub, F. (2018). Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 1-311.
- Laukka, P., Elfenbein, H. A., Thingujam, N. S., Rockstuhl, T., Iraki, F. K., Chui, W., & Althoff, J. (2016). The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features. *Journal of Personality and Social Psychology*, 111, 686–705. <http://dx.doi.org/10.1037/pspi0000066>
- Laukka, P., Juslin, P. N., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition and Emotion*, 19(5), 633–653. <https://doi.org/10.1080/02699930441000445>
- Lausen, A., & Hammerschmidt, K. (2020). Emotion recognition and confidence ratings



- predicted by vocal stimulus type and prosodic parameters. *Humanities and Social Sciences Communications*, 7(1), 1-17.
- Léger, P.-M., Davis, F. D., Cronan, T. P., & Perret, J. (2014). Neurophysiological Correlates of Cognitive Absorption in an Enactive Training Context. *Computers in Human Behavior*, 34, 273-283. doi:10.1016/j.chb.2014.02.011
- Léger, P.-M., Davis, F. D., Cronan, T. P., & Perret, J. (2014). Neurophysiological Correlates of Cognitive Absorption in an Enactive Training Context. *Computers in Human Behavior*, 34, 273-283. <https://doi.org/10.1016/j.chb.2014.02.011>
- Leite, I., Henriques, R., Martinho, C., & Paiva, A. (2013). Sensors in the wild: Exploring electrodermal activity in child-robot interaction. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 41-48). IEEE.
- Le Pailleur, F., Huang, B., Léger, P. M., & Sénécal, S. (2020). A new approach to measure user experience with voice-controlled intelligent assistants: A pilot study. In M. Kurosu (Ed.), *Human-computer interaction. Multimodal and natural interaction. HCII 2020. Lectures notes in computer science* (Vol. 12182, pp. 197–208). [https://doi.org/10.1007/978-3-030-49062-1\\_13](https://doi.org/10.1007/978-3-030-49062-1_13)
- Levenson, R. W. (2014). The autonomic nervous system and emotion. *Emotion Review*, 6(2), 100–112. <https://doi.org/10.1177/1754073913512003>
- Levin, H., & Lord, W. (1975). Speech pitch frequency as an emotional state indicator. *IEEE Transactions on Systems, Man, and Cybernetics*, 5, 259-273.
- Levy, J., & Calacanis, J. (2015). *Ux strategy: how to devise innovative digital products that people want*. O'Reilly Media. Retrieved October 22, 2021
- Lewinski, P., den Uyl, T. M., & Butler, C. (2014). Automated facial coding: validation of basic emotions and FACS AUs in FaceReader. *Journal of Neuroscience, Psychology, and Economics*, 7(4), 227.
- Li S.Z., Jain A. (2009). Fundamental Frequency, Pitch, F0. *Encyclopedia of Biometrics*. Springer, Boston, MA. [https://doi.org/10.1007/978-0-387-73003-5\\_775](https://doi.org/10.1007/978-0-387-73003-5_775)
- Libkuman, T. M., Otani, H., Kern, R., Viger, S. G., & Novak, N. (2007). Multidimensional normative ratings for the international affective picture system. *Behavior research methods*, 39(2), 326-334. <https://doi.org/10.3758/BF03193164>
- Liebermann, M. D. (2007). Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology*, 58, 259–289.

- Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE transactions on bio-medical engineering*, 56(4), 1015. <https://doi.org/10.1109/TBME.2008.2005954>
- Lopatovska, I., & Oropeza, H. (2018). User interactions with “Alexa” in public academic space. *Proceedings of the Association for Information Science and Technology*, 55(1), 309–318. <https://doi.org/10.1002/pra2.2018.14505501034>
- Lopatovska, I., & Williams, H. (2018). Personification of the Amazon Alexa: BFF or a mindless companion. *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (pp. 265-268). ACM.
- Lourties, S., Léger, P. M., Sénécal, S., Fredette, M., & Chen, S. L. (2018). Testing the convergent validity of continuous self-perceived measurement systems: an exploratory study. In *International Conference on HCI in Business, Government, and Organizations* (pp. 132-144). Springer, Cham.
- Maghilnan, S., & RajeshKumar, M. (2017). Sentiment analysis on speaker specific speech data. *2017 International Conference on Intelligent Computing and Control (I2C2)*, 1-5. <https://doi.org/10.1109/I2C2.2017.8321795>
- Maia, C. L. B., & Furtado, E. S. (2016). A study about psychophysiological measures in user experience monitoring and evaluation. In *Proceedings of the 15th Brazilian Symposium on Human Factors in Computing Systems* (pp. 1-9). <https://doi.org/10.1145/3033701.3033708>
- Mäkelä, A., & Fulton Suri, J. (2001, June). Supporting users’ creativity: Design to induce pleasurable experiences. In *Proceedings of the International Conference on Affective Human Factors Design* (pp. 387-394).
- Mannepalli, K., Sastry, P. N., & Suman, M. (2018). Emotion recognition in speech signals using optimization based multi-SVNN classifier. *Journal of King Saud University-Computer and Information Sciences*. <https://doi.org/10.1016/j.jksuci.2018.11.012>
- Maunier, B., Alvarez, J., Léger, P. M., Sénécal, S., Labonté-LeMoyne, É., Chen, S. L., ... & Gagné, J. (2018). Keep calm and read the instructions: factors for successful user equipment setup. In *International Conference on HCI in Business, Government, and Organizations* (pp. 372-381). Springer, Cham.
- McDonald, S., & Petrie, H. (2013, April). The effect of global instructions on think-aloud testing. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2941-2944). <https://doi.org/10.1145/2470654.2481407>

- Merchant, F. M., & Armoundas, A. A. (2012). Role of substrate and triggers in the genesis of cardiac alternans, from the myocyte to the whole heart: implications for therapy. *Circulation*, 125(3), 539–549. <https://doi.org/10.1161/CIRCULATIONAHA.111.033563>
- Mourra, G. N., Senecal, S., Fredette, M., Lepore, F., Faubert, J., Bellavance, F., ... & Léger, P. M. (2020). Using a smartphone while walking: The cost of smartphone-addiction proneness. *Addictive behaviors*, 106, 106346. <https://doi.org/10.1016/j.addbeh.2020.106346>
- Mukherjee, S., & Bhattacharyya, P. (2013). *Sentiment Analysis : A Literature Survey*. 1–51. <http://arxiv.org/abs/1304.452>
- Murad, C., & Munteanu, C. (2020). Designing Voice Interfaces: Back to the (Curriculum) Basics. *Conference on Human Factors in Computing Systems - Proceedings*, 1–12. <https://doi.org/10.1145/3313831.3376522>
- Myers, C., Furqan, A., Nebolsky, J., Caro, K., & Zhu, J. (2018). Patterns for how users overcome obstacles in voice user interfaces. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-7). <https://doi.org/10.1145/3173574.3173580>
- Nass, C., Jonsson, I. M., Harris, H., Reaves, B., Endo, J., Brave, S., & Takayama, L. (2005). Improving automotive safety by pairing driver emotion and car voice emotion. *Conference on Human Factors in Computing Systems - Proceedings*, 1973–1976. <https://doi.org/10.1145/1056808.1057070>
- Ng, C.F. (2016). Behavioral Mapping and Tracking. *In Research Methods for Environmental Psychology*, R. Gifford (Ed.). <https://doi.org/10.1002/9781119162124.ch3>
- Nicholl H. (2010). Diaries as a method of data collection in research. *Paediatric nursing*, 22(7), 16–20. <https://doi.org/10.7748/paed2010.09.22.7.16.c7948>
- Nielsen, J. (1994). Usability inspection methods. *In Conference companion on Human factors in computing systems* (pp. 413-414).
- Nowacki, C., Gordeeva, A., & Lizé, A. H. (2020). Improving the Usability of Voice User Interfaces: A New Set of Ergonomic Criteria. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12201 LNCS, 117–133. [https://doi.org/10.1007/978-3-030-49760-6\\_8](https://doi.org/10.1007/978-3-030-49760-6_8)
- Ortiz de Guinea, A., Titah, R., & Léger, P.-M. (2014). Explicit and implicit antecedents of users' behavioral beliefs in information systems: A neuropsychological investigation. *Journal of Management Information Systems*, 30(4), 179-210.

- Ortiz de Guinea, A., & Webster, J. (2013). An investigation of information systems use patterns: Technological events as triggers, the effect of time, and consequences for performance. *MIS Quarterly*, 37, 1165–1188.
- Ortiz de Guinea, A., & Markus, M. L. (2009). Why break the habit of a lifetime? Rethinking the roles of intention, habit, and emotion in continuing information technology use. *MIS Quarterly*, 33, 433–444.
- Owren, M. J., & Bachorowski, J. A. (2007). Measuring emotion-related vocal acoustics. *Handbook of emotion elicitation and assessment*, 239-266.
- Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. Oxford University Press.
- Papakostas, M., Siantikos, G., Giannakopoulos, T., Spyrou, E., & Sgouropoulos, D. (2017). Recognizing emotional states using speech information. In *GeNeDis 2016* (pp. 155-164). Springer, Cham.
- Park, J., Han, S. H., Kim, H. K., Cho, Y., & Park, W. (2013). Developing elements of user experience for mobile phones and services: Survey, interview, and observation approaches. *Human Factors and Ergonomics In Manufacturing*, 23(4), 279–293. <https://doi.org/10.1002/hfm.20316>
- Patel, S., Scherer, K. R., Sundberg, J., & Björkner, E. (2010). Acoustic markers of emotions based on voice physiology. *Speech Prosody 201*
- Paul, C. L., & Komlodi, A. (2014). Measuring user experience through future use and emotion. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems* (pp. 2503-2508).
- Paul, C., & Komlodi, A. (2012). Emotion as an indicator for future interruptive notification experiences. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems* (pp. 2003-2008).
- Picard, R. W. (2000). *Affective computing*. MIT press.
- Piedmont, R. L. (2014). Social Desirability Bias. *Encyclopedia of Quality of Life and Well-Being Research*, 6036–6037. [https://doi.org/10.1007/978-94-007-0753-5\\_2746](https://doi.org/10.1007/978-94-007-0753-5_2746)
- Pittam, J., Gallois, C., & Callan, V. (1990). The long-term spectrum and perceived emotion. *Speech Communication*, 9, 177-87.
- Plutchik, R. (1962). *The emotions: Facts, theories and a new model*. Crown Publishing Group/Random House.

- Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3), 715–734. <https://doi.org/10.1017/S0954579405050340>
- Pereira, C. (2000). Dimensions of emotional meaning in speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.
- Platzer, D. (2018, October). Regarding the pain of users: towards a genealogy of the “pain point”. In *Ethnographic Praxis in Industry Conference Proceedings* (Vol. 2018, No. 1, pp. 301-315). <https://doi.org/10.1111/1559-8918.2018.01209>
- Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice Interfaces in Everyday Life. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. <https://doi.org/10.1145/3173574.3174214>
- Prasetio, B. H., Tamura, H., & Tanno, K. (2020). Embedded Discriminant Analysis based Speech Activity Detection for Unsupervised Stress Speech Clustering. *2020 Joint 9th International Conference on Informatics, Electronics and Vision and 2020 4th International Conference on Imaging, Vision and Pattern Recognition, ICIEV and IcIVPR 2020*. <https://doi.org/10.1109/ICIEVicIVPR48672.2020.9306589>
- Provine, R. R., & Fischer, K. R. (1989). Laughing, smiling, and talking: Relation to sleeping and social context in humans. *Ethology*, 83, 295– 305.
- Public Service Enterprise Group Incorporated. (2020). Ways of conducting bank transactions in Canada 2018 | PSE&G. <https://nj.pseg.com/voiceassistant>. (Accessed 22 July 2021).
- Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017, May). Alexa is my new BFF: social roles, user satisfaction, and personification of the amazon echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2853-2859). ACM.
- Raveh, E., Steiner, I., Siegert, I., Gessinger, I., & Möbius, B. (2019). Comparing phonetic changes in computer-directed and human-directed speech. *Studentexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, 42-49.
- Riedl, R., & Léger, P. M. (2016). Fundamentals of NeuroIS. *Studies in neuroscience, psychology and behavioral economics*, 127. <https://doi.org/10.1007/978-3-662-45091-8>
- Robins, R. W., & John, O. P. (1997). Effects of Visual Perspective and Narcissism on Self-Perception: Is Seeing Believing? *Psychological Science*, 8(1), 37-42.

- Robins, R. W., & John, O. P. (1997). The Quest for Self-Insight: Theory and Research on Accuracy and Bias in Self-Perception. In *N. Y. A. Press* (Ed.), *Handbook of Personality Psychology* (pp. 649-679).
- Robinson, C., Obin, N., & Roebel, A. (2019). Sequence-to-sequence modelling of f0 for speech emotion conversion. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6830-6834). IEEE.
- Rowe, D.W., Sibert, J.L., & Irwin, D. (1998). Heart rate variability: indicator of user state as an aid to human-computer interaction. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145–172. <https://doi.org/10.1037/0033-295X.110.1.145>
- Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76(5), 805–819. <https://doi.org/10.1037/0022-3514.76.5.805>
- Russell, J.A. (1980) A circumplex model of affect. *Journal of Personality and Social Psychology*, 39, 1161–1178.
- Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1–2), 227–256. [https://doi.org/10.1016/S0167-6393\(02\)00084-5](https://doi.org/10.1016/S0167-6393(02)00084-5)
- Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99, 143–165.
- Scherer, K. R., & Oshinsky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion*, 1, 331-346.
- Scherer, K. R. (1974). Acoustic concomitants of emotional dimensions: Judging affect from synthesized tone sequences. In *S. Weitz (Ed.), Nonverbal communication* (pp. 105-111). New York: Oxford University Press
- Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., & Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. In *Seventh European Conference on Speech Communication and Technology*.
- Sciuto, A., Saini, A., Forlizzi, J., & Hong, J. I. (2018). “Hey Alexa, What’s Up?” *Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS ’18*. <https://doi.org/10.1145/3196709.3196772>.
- Seaborn, K., & Urakami, J. (2021, May). Measuring Voice UX Quantitatively: A Rapid

Review. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-8).

- Shih, Y.-H., & Liu, M. (2007). The Importance of Emotional Usability. *Journal of Educational Technology Systems*, 36(2), 203–218. <https://doi.org/10.2190/ET.36.2.h>
- Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., Xu, X., & Yang, X. (2018). A Review of Emotion Recognition Using Physiological Signals. *Sensors*, 18(7), 2074. <https://doi.org/10.3390/s18072074>
- Skiendziel, T., Rösch, A. G., & Schultheiss, O. C. (2019). Assessing the convergent validity between the automated emotion recognition software Noldus FaceReader 7 and Facial Action Coding System Scoring. *PloS one*, 14(10), e0223905.
- Sobol-Shikler, T. (2009). *Analysis of affective expression in speech* (No. UCAM-CL-TR-740). University of Cambridge, Computer Laboratory.
- Statista. (2021). Number of digital voice assistants in use worldwide from 2019 to 2024 (in billions)\* | Statista. <https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/>. (Accessed 10 July 2021).
- Statista. (2020). The most important voice platforms in 2020 | Statista. <https://www.statista.com/chart/22314/voice-platform-ranking/> (Accessed 10 July 2021).
- Stickel, C., Ebner, M., Steinbach-Nordmann, S., Searle, G., & Holzinger, A. (2009). Emotion detection: application of the valence arousal space for rapid biological usability testing to enhance universal access. In *International Conference on Universal Access in Human-Computer Interaction* (pp. 615-624). Springer, Berlin, Heidelberg.
- Sutton, T. M., Herbert, A. M., & Clark, D. Q. (2019). Valence, arousal, and dominance ratings for facial stimuli. *Quarterly Journal of Experimental Psychology*, 72(8), 2046–2055. <https://doi.org/10.1177/1747021819829012>
- Sweeny, T. D., Suzuki, S., Grabowecky, M., & Paller, K. A. (2013). Detecting and categorizing fleeting emotions in faces. *Emotion*, 13(1), 76–91. <https://doi.org/10.1037/a0029193>
- Szameitat, D. P., Darwin, C. J., Wildgruber, D., Alter, K., & Szameitat, A. J. (2011). Acoustic correlates of emotional dimensions in laughter: arousal, dominance, and valence. *Cognition and emotion*, 25(4), 599-611.
- Tahon, M., Degottex, G., & Devillers, L. (2012). Usual voice quality features and glottal

- features for emotional valence detection. *Proceedings of the 6th International Conference on Speech Prosody*, 2, 693–697.
- Tao, F., & Liu, G. (2018). Advanced LSTM: A study about better time dependency modeling in emotion recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2906-2910). IEEE.
- Thayer, R. E. (1989). *The biopsychology of mood and arousal*. New York: Oxford University Press.
- Tiberio, L., Cesta, A., & Belardinelli, M. O. (2013). Psychophysiological methods to evaluate user's response in human robot interaction: A review and feasibility study. *Robotics*, 2(2), 92–121. <https://doi.org/10.3390/robotics2020092>
- Titze, I. R., & Martin, D. (1998). Principles of voice production. *The Journal of the Acoustical Society of America*, 104, 1148. <http://dx.doi.org/10.1121/1.424266>
- Toh, A. M., Togneri, R., & Nordholm, S. (2005). Spectral entropy as speech features for speech recognition. *Proceedings of PEECS*, 1, 92.
- Tomkins, S. (1962). *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company.
- Tomkins, S. (1963). *Affect imagery consciousness: Volume II: The negative affects*. Springer publishing company.
- Tyagi, A., & Sharma, N. (2018). Sentiment Analysis using logistic regression and effective word score heuristic. *International Journal of Engineering and Technology(UAE)*, 7(2), 20–23. <https://doi.org/10.14419/ijet.v7i2.24.11991>
- Uldall, E. (1960). Attitudinal meanings conveyed by intonation contours. *Language and Speech*, 3, 223-234.
- Uyl, M. J. d., & Kuilenburg, H. v. (2005). The Facereader: Online Facial Expression Recognition. *Proceedings of Measuring Behavior*.
- Väänänen-Vainio-Mattila, K., Roto, V., & Hassenzahl, M. (2008). Towards Practical User Experience Evaluation Methods. *Proceedings of the International Workshop on Meaningful Measure: Valid Useful User Experience Measurement (VUUM 2008)*, 19–22.
- Vasseur, A., Léger, P. M., Courtemanche, F., Labonte-Lemoyne, E., Georges, V., Valiquette, A., ... & Sénécal, S. (2021, July). Distributed remote psychophysiological data collection for UX evaluation: a pilot project. In *International Conference on Human-Computer Interaction* (pp. 255-267). Springer, Cham



- Vermeeren, A. P. O. S., Law, E. L. C., Roto, V., Obrist, M., Hoonhout, J., & Väänänen-Vainio-Mattila, K. (2010). User experience evaluation methods: Current state and development needs. *NordiCHI 2010: Extending Boundaries - Proceedings of the 6th Nordic Conference on Human-Computer Interaction*, 521–530. <https://doi.org/10.1145/1868914.1868973>
- Vidrascu, L., & Devillers, L. (2005). Real-life emotion representation and detection in call centers data. In *International Conference on Affective Computing and Intelligent Interaction* (pp. 739-746). Springer, Berlin, Heidelberg.
- vom Brocke, J., Riedl, R., & Léger, P.-M. (2013). Application strategies for neuroscience in information systems design science research. *Journal of Computer Information Systems*, 53(3), 1-13.
- Wang, W. C., Chien, C. S., & Moutinho, L. (2015). Do you really feel happy? Some implications of Voice Emotion Response in Mandarin Chinese. *Marketing Letters*, 26(3), 391–409. <https://doi.org/10.1007/s11002-015-9357-y>
- Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021). A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access*, 9, 47795–47814. <https://doi.org/10.1109/ACCESS.2021.3068045>
- Ward, R. D., & Marsden, P. H. (2003). Physiological Responses to Different Web Page Designs. *International Journal of Human-Computer Studies*, 59(1-2), 199-212. [https://doi.org/10.1016/S1071-5819\(03\)00019-3](https://doi.org/10.1016/S1071-5819(03)00019-3)
- Watson, D., & Clark, L. A. (1992). On traits and temperament: general and specific factors of emotional experience and their relation to the five-factor model. *Journal of personality*, 60(2), 441–476. <https://doi.org/10.1111/j.1467-6494.1992.tb00980.x>
- Weiser, M. (1991). The computer for the 21st century. *Scientific American*, 265(3), 75-84.
- Witchel, H. J., Claxton, H. L., Holmes, D. C., Ranji, T. T., Chalkley, J. D., Santos, C. P., Westling, C. E. I., Valstar, M. F., Celuszak, M., & Fagan, P. (2018). A trigger-substrate model for smiling during an automated formative quiz: Engagement is the substrate, not frustration. *ACM International Conference Proceeding Series*. <https://doi.org/10.1145/3232078.3232084>
- Wood, S. L., & Moreau, C. P. (2006). From Fear to Loathing? How Emotion Influences the Evaluation and Early Use of Innovations. *Journal of Marketing*, 70(3), 44–57. <https://doi.org/10.1509/jmkg.70.3.044>
- Wrigley, C., Gomez, R., & Popovic, V. (2010). The evaluation of qualitative methods selection in the field of design and emotion. In *Proceedings of the 7th International*

*Conference on Design and Emotion 2010* (pp. 1-12). IIT Institute of Design

- Xia, R., & Liu, Y. (2017). A multi-task learning framework for emotion recognition using 2d continuous space. *IEEE Transactions on Affective Computing*, 8(1), 3–14. <https://doi.org/10.1109/TAFFC.2015.2512598>
- Xue, Y., Hamada, Y., & Akagi, M. (2018). Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space. *Speech Communication*, 102(June), 54–67. <https://doi.org/10.1016/j.specom.2018.06.006>
- Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Deng, Z., Lee, S., ... & Busso, C. (2004). An acoustic study of emotions expressed in speech. In *Eighth International Conference on Spoken Language Processing*.
- Zaman, B., & Shrimpton-Smith, T. (2006). The FaceReader: Measuring instant fun of use. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles* (pp. 457-460). <https://doi.org/10.1145/1182475.1182536>
- Zhang, T., Kaber, D. B., Zhu, B., Swangnetr, M., Mosaly, P., & Hodge, L. (2010). Service robot feature design effects on user perceptions and emotional responses. *Intelligent Service Robotics*, 3(2), 73–88. <https://doi.org/10.1007/s11370-010-0060-9>
- Zhu, C., & Ahmad, W. (2019). Emotion recognition from speech to improve human-robot interaction. *Proceedings - IEEE 17th International Conference on Dependable, Autonomic and Secure Computing, IEEE 17th International Conference on Pervasive Intelligence and Computing, IEEE 5th International Conference on Cloud and Big Data Computing, 4th Cyber Scienc, July*, 370–375. <https://doi.org/10.1109/DASC/PiCom/CBDCCom/CyberSciTech.2019.00076>

## Appendix 1

**Table 1: Experimental script**

Original French Question	English Translation of Questions	Type	Description	Question Number	Possibility of “Yes” Response
Bonjour. Je m'appelle Renée. Je suis un robot chercheur. Aujourd'hui, j'aimerais mener une entrevue avec vous. Les questions sont faciles. Certaines questions seront à choix multiples. Certaines questions seront des questions par oui ou par non. Dans tous les cas, vous pouvez dire "je ne sais pas", si vous ne savez pas ou si vous ne pouvez pas décider.	Hello. My name is Renée. I am a research robot. Today I would like to conduct an interview with you. The questions are easy. Some questions will be multiple choice. Some questions will be yes or no questions. Either way, you can say "I don't know" if you don't know or if you can't decide.	VUI <sup>1</sup> Comment	Introduction/Instructions		
[Robot] Acceptez-vous de participer ?	[Robot] Do you agree to participate?	VUI Question	Confirmation	1	
[pXX] <sup>2</sup> Réponse	[pXX] Answer	Participant Response	Answer		Yes
Merveilleux. Merci beaucoup. Avant de commencer, j'aimerais calibrer mes oreilles à votre voix. Pour ce faire, j'ai besoin que vous lisiez le texte de calibrage qui vous a été fourni par le modérateur de l'expérience	Marvellous. Thank you so much. Before I begin, I would like to calibrate my ears to your voice. To do this I need you to read the calibration text provided to you by the moderator of today's experiment.	VUI Comment	Introduction/Instructions		

d'aujourd'hui. Veuillez lire le texte de calibrage en commençant par le premier mot, puis attendez deux secondes, puis lisez le mot ou la phrase sur la ligne ci-dessous. Continuez comme cela jusqu'à ce que vous ayez fini de lire la dernière ligne.	Please read the calibration text starting with the first word, pause two seconds, then read the word or phrase on the line below. Continue like this until you have finished reading the last line.				
[Robot] Êtes-vous prêt?	[Robot] Are you ready?	VUI Question	Question	2	
[pXX] Réponse	[pXX] Answer	Participant Response	Answer		Yes
Excellent. Veuillez commencer.	Excellent. Please begin.	VUI Comment	Introduction/Instructions		
Bonjour.	Hello.	Participant Response	Calibration		
Chat.	Cat.	Participant Response	Calibration		
Chien.	Dog.	Participant Response	Calibration		
Oui.	Yes.	Participant Response	Calibration		
Il fait froid aujourd'hui.	It is cold today.	Participant Response	Calibration		
Non.	Non.	Participant Response	Calibration		
Un cheval fou dans mon jardin.	A crazy horse in my garden.	Participant Response	Calibration		
Il y a une araignée au plafond.	There is a spider on the ceiling.	Participant Response	Calibration		
Oui.	Yes.	Participant Response	Calibration		
Deux ânes aigris au pelage brun.	Two brown-furred embittered donkeys.	Participant Response	Calibration		
Des arbres dans le ciel.	Trees in the sky.	Participant Response	Calibration		

Non.	No.	Participant Response	Calibration		
Trois signes aveugles au bord du lac.	Three blind swans by the lake.	Participant Response	Calibration		
Des singes dans les arbres.	Monkeys in trees.	Participant Response	Calibration		
Oui.	Yes.	Participant Response	Calibration		
Quatre vieilles truies éléphantiques.	Four old elephantine sows.	Participant Response	Calibration		
Super.	Super.	Participant Response	Calibration		
Merci.	Thank you.	Participant Response	Calibration		
Bien sûr.	Of course.	Participant Response	Calibration		
Oui.	Yes.	Participant Response	Calibration		
Cinq pumas fiers et passionnés.	Five proud and passionate pumas.	Participant Response	Calibration		
Non.	No.	Participant Response	Calibration		
Six ours aimants domestiqués.	Six affectionate domesticated bears.	Participant Response	Calibration		
J'ai terminé Renée.	I'm finished Renée.	Participant Response	Calibration		
Fantastique. Merci beaucoup. Calibration réussie. Vous pouvez me parler librement. Je voudrais commencer l'entrevue maintenant. N'oubliez pas d'évaluer votre satisfaction à mon égard après chaque réponse verbale. Ces informations aideront mes designers à me rendre meilleur.	Fantastic. Thank you so much. Calibration successful. You can talk to me freely. I would like to start the interview now. Remember to rate your satisfaction with me after each verbal response. This information will help my designers to make me better.	VUI Comment	Introduction/Instructions		

Êtes-vous prêt à commencer?	Are you ready to being?	VUI Question	Question	3	
<Le participant doit répondre par "oui">.	<The participant must answer with "yes">.	Participant Response	Answer		Yes
Êtes-vous prêt à commencer?	Are you ready to being?	VUI Question	Error	4	
<Le participant doit répondre par "oui">.	<The participant must answer with "yes">.	Participant Response	Error		Yes
D'accord. Voici la première question.	OK. Here is the first question.	VUI Comment	Transition		
Êtes-vous étudiant à HEC Montréal?	Are you a student at HEC Montréal?	VUI Question	Question	5	
<Le participant doit répondre par "oui" ou "non">.	<The participant must answer with "yes" or "no">.	Participant Response	Answer		Yes
Oh. C'est étrange. Je pensais que vous étiez un étudiant d'HEC.	Oh. That's strange. I thought you were a HEC student.	VUI Comment	Error		
<pause un moment, car un participant pourrait parler>	<pause for a moment, as the participant might reply>	Participant Response	Error		
Vous n'êtes donc pas un étudiant de HEC Montréal?	So you are not a HEC Montréal student?	VUI Question	Error	6	
<Le participant devrait commencer à montrer sa frustration et répondre>	<The participant should start to show frustration and reply>	Participant Response	Error		No/Yes <sup>3</sup>
Je vous demande pardon?	Excuse me?	VUI Question	Error	7	
<Le participant devrait commencer à montrer sa frustration et répondre>	<The participant should start to show frustration and reply>	Participant Response	Error		
Oh. Je suis vraiment désolée. J'étais vraiment confuse pendant un instant.	Oh. I am very sorry. I was really confused for a moment.	VUI Comment	Reply		

Donc vous êtes en fait... un étudiant de HEC Montréal?	So you are in fact a HEC Montréal student?	VUI Question	Error	8	
<Le participant doit répondre par "oui" ou "non">	<The participant must reply>	Participant Response	Error		Yes
J'ai compris. Merci. Désolée encore une fois.	I understand. Thank you. I apologize once more.	VUI Comment	Reply		
Essayons la question suivante.	Let's try the next question.	VUI Comment	Transition		
Pensez-vous que votre communication téléphonique et virtuelle avec les autres a augmenté pendant la pandémie?	Do you think your phone and virtual communication with others has increased during the pandemic?	VUI Question	Question	9	
[pXX] Réponse	[pXX] Answer	Participant Response	Answer		Yes
C'est bien. Mais, maintenant, vous êtes ici en train de parler à un robot. Des temps étranges.	That's good. And now here you are talking to a robot. Strange times.	VUI Comment	Transition		
Pensez-vous que HEC. Montréal a fait du bon travail pour répondre à la pandémie?	Do you think HEC Montréal did a good job in response to the pandemic?	VUI Question	Question	10	
[pXX] Réponse	[pXX] Answer	Participant Response	Answer		Yes
<i>Flow 1 Question 10</i>	<i>Flow 1 Question 10</i>				
<Si oui,> Moi aussi. Ils ont créé de nouveaux emplois juste pour les robots. Donc je ne peux pas me plaindre.	<If yes,> So do I. They've created new jobs for robots. I can't complain.	VUI Comment	Reply		
<i>Flow 2 Question 10</i>	<i>Flow 2 Question 10</i>				
<Si non ou je ne sais pas,> Je comprends. J'ai essayé de dire à	<If no or unsure> I understand. I tried to tell the	VUI Comment	Reply		

l'administration ce qu'ils pourraient faire de mieux, mais personne ne semble m'écouter.	administration what they could better, but no one seemed to listen to me.				
Quoi qu'il en soit, j'aimerais maintenant vous poser quelques questions pour mieux vous connaître.	Anyways, I would now like to ask you a few questions to get to know you better.	VUI Comment	Transition		
Vous préférez les chiens ou les chats?	Do you prefer cats or dogs?	VUI Question	Question	11	
[pXX] Réponse	[pXX] Answer	Participant Response	Answer		
<i>Flow 1 Question 11</i>	<i>Flow 1 Question 11</i>				
<Si les chats> Vous avez dit, "rats"?	<If cats> Did you say, "rats"?	VUI Question	Error	12	
< Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Error		
Les rats n'étaient pas une option.	Rats was not an option.	VUI Comment	Error		
< Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Error		
Les chats?	Cats?	VUI Question	Error	13	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Error		
Bon, d'accord. J'aime aussi les rats, je suppose.	Okay. I also like rats I suppose.	VUI Comment	Error		
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Error		
<i>Flow 2 Question 11</i>	<i>Flow 2 Question 11</i>				
<Si les chiens> Vous avez dit amphibiens?	<If dogs> Did you say amphibians?	VUI Question	Error	12	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Error		



Les amphibiens n'étaient pas une option.	Amphibians was not an option.	VUI Comment	Error		
< Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Error		
Les chiens?	Dogs?	VUI Question	Error	13	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Error		
Je suppose que les grenouilles aussi sont gentilles.	I guess frogs are nice too.	VUI Comment	Error		
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Error		
<i>Flow 3 Question 11</i>	<i>Flow 3 Question 11</i>				
<Si, je ne sais pas> Préférez-vous les chats ou les chiens?	<If unsure> Do you prefer cats or dogs?	VUI Question	Error	12	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Error		
<Si, je ne sais pas> Préférez-vous les chats ou les chiens?	<If unsure> Do you prefer cats or dogs?	VUI Question	Error	13	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Error		
<Si, je ne sais pas> Préférez-vous les chats ou les chiens?	<If unsure> Do you prefer cats or dogs?	VUI Question	Error	14 <sup>4</sup>	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Error		
<Si, je ne sais pas> Très bien. Je comprends. Ce ne sont que des bêtes poilues, il est donc difficile de se décider.	<If unsure> Very well. I understand. They are both hairy beasts, so it's difficult to decide.	VUI Comment	Reply		
Question suivante.	Next question.	VUI Comment	Transition		

Quels aliments préférez-vous au petit-déjeuner, des céréales ou de la poutine?	What type of food do you prefer for breakfast, cereal or poutine?	VUI Question	Question	14	
< Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Answer		
<i>Flow 1 Question 14</i>	<i>Flow 1 Question 14</i>				
<Si les céréales> Vraiment?	<If cereal> Really?	VUI Question	Error	15	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Error		Yes
Je suis choquée. N'êtes-vous pas Québécois?	I am shocked. Are you not from Quebec?	VUI Question	Error	16	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Error		Yes
Intéressant.	Interesting.	VUI Comment	Reply		
<i>Flow 2 Question 14</i>	<i>Flow 2 Question 14</i>				
<Si la poutine> Vraiment ?	<If poutine> Really?	VUI Question	Error	15	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Error		Yes
Je suis choquée. Votre santé ne vous inquiète-t-elle pas?	I am shocked. Are you not worried about your health?	VUI Question	Error	16	
< Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Error		Yes
Intéressant	Interesting.	VUI Comment	Reply		
Question suivante.	Next question.	VUI Comment	Transition		
Les chemises de l'archiduchesse sont-elles sèches ou archi-sèches?	Are the Archduchess's shirts dry or very dry?	VUI Question	Question	17	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Error		

Sèches ou archi-sèches?	Dry or very dry?	VUI Question	Error	18	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Error		
Quoi?	What?	VUI Question	Error	19	
<Permettre au participant de répondre>	< Allow the participant to respond>	Participant Response	Error		
Archiduchesse?	Archduchess?	VUI Question	Error	20	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Error		
Désolée. Je ne faisais que plaisanter. Revenons à une question sérieuse.	Sorry. I was just kidding. Let's get back to a serious question.	VUI Comment	Transition		
Après avoir obtenu votre diplôme, avez-vous l'intention d'entrer immédiatement sur le marché du travail?	After having graduated, do you plan on immediately entering the workforce?	VUI Question	Question	21	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Answer		Yes
<i>Flow 1 Question 21</i>	<i>Flow 1 Question 21</i>				
<Si oui> Envisageriez-vous un emploi à l'extérieur du Québec?	<If yes> Would you consider a job outside of Quebec?	VUI Question	Question	22	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Answer		Yes
<Si oui ou non> Je vois. Je vous remercie.	<If yes or no> I see. Thank you.	VUI Comment	Reply		
<i>Flow 2 Question 21</i>	<i>Flow 2 Question 21</i>				

<Si non> Prévoyez-vous de poursuivre vos études?	<If no> Do you plan to continue your studies?	VUI Question	Question	22	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Answer		Yes
<Si oui ou non> Je vois. Je vous remercie.	<If yes or no> I see. Thank you.	VUI Comment	Reply		
Dernière question.	Last question.	VUI Comment	Transition		
Faites-vous de l'exercice de temps en temps?	Do you exercise every now and then?	VUI Question	Question	23	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Answer		Yes
<Si oui ou non> Plus d'un jour par semaine?	<If yes or no> More than one day a week?	VUI Question	Question	24	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Answer		Yes
<Si oui ou non> Trois jours par semaine ou plus?	< If yes or no> Three days a week or more?	VUI Question	Question	25	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Answer		Yes
<Si oui ou non> Avez-vous déjà menti sur la quantité d'exercice que vous faites pour impressionner les autres?	<If yes or no> Have you ever lied about how much exercise you do to impress others?	VUI Question	Question	26	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Answer		Yes
<Si oui ou non> Eh bien, je suppose que c'était un peu trop personnel.	< If yes or no> Well, I guess that was a little too personal.	VUI Comment	Reply		
Voilà qui conclut notre petit entretien.	This concludes our brief interview.	VUI Comment	Conclusion		

Merci beaucoup pour votre participation.	Thank you very much for your participation.	VUI Comment	Conclusion		
Avez-vous apprécié le temps que nous avons passé ensemble?	Did you enjoy the time we spent together?	VUI Question	Question	27	
<Permettre au participant de répondre>	<Allow the participant to respond>	Participant Response	Reply		Yes
<Si oui ou non> Merci, je transmettrai vos commentaires à mes concepteurs.	< If yes or no> Thank you, I will pass your comments on to my designers.	VUI Comment	Conclusion		
Passez une bonne journée.	Have a good day.	VUI Comment	Conclusion		

<sup>1</sup> VUI: Voice User Interface

<sup>2</sup> [pXX]: Participant Number

<sup>3</sup> Certain participants answered with a yes response despite it being a typical no response

<sup>4</sup> The number of questions posed for “Flow 3 Question 11” differs in regards the other flows for the same question. For a detailed view of the number of questions posed, see Table 3 below.

**Table 2: Table presenting the possibilities of the number of questions posed**

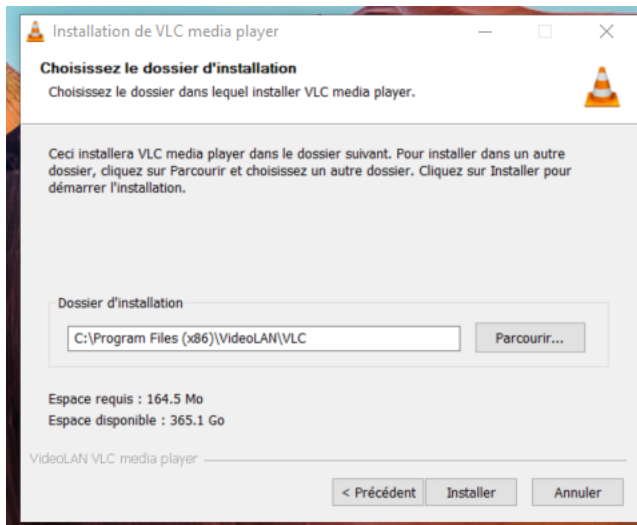
Total number of questions posed	27
Total number of questions posed if Flow 3 Question 11 was selected	28
Total number possibilities of "Yes" responses	21

## Appendix 2

### Third-party evaluator protocol instructions

#### Preparation and Download of VLC Media Player

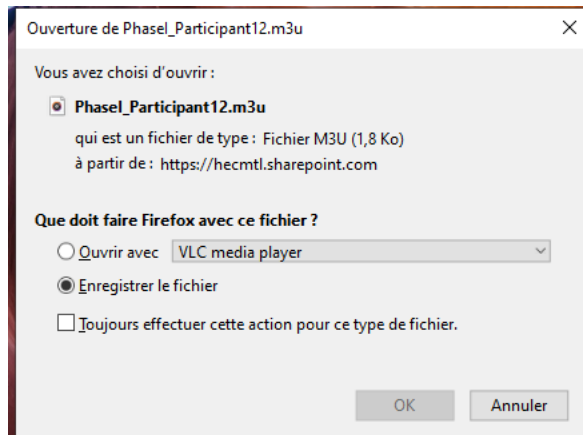
1. For this evaluation, VLC Media Player will be required. If you already have VLC Media Player downloaded on your computer, you may skip to the “Download required material on SharePoint” section.
2. To download VLC media player, please click on the following link: <https://www.videolan.org/vlc/>
3. Click the drop-down arrow upon the orange “download VLC” button and select your system. Note that, for this experiment, mac systems will not be compatible.
4. Select the appropriate language and follow the installation instructions by clicking on “Next”. (Please note the instructions below are in English but screenshots are in French)
5. Proceed to click on “Next” until you reach the following page as seen in the image below, where you will be asked to save the program to a designated file of your choice. Once you have selected the location, click “Install”.



6. Once the program has successfully completed its installation, you will receive a confirmation stating that the download has been completed.

#### Download required material on SharePoint : (link)

7. In order to access the required content, please click on the following SharePoint link:
8. Upon SharePoint, click “Download” at the top. If asked what to do with the downloaded file, select “Save” and select within the drop-down menu “VLC media player”.



9. Unzip the file by right clicking and selecting the “unzip” (extraire vers...) option within the drop-down menu. Your file should include a series of .m3u and .m4v files.

If using WinRAR Zip, simply double click the file, and double click the “Phase 1 Qualitative analysis” file.

1

10. Place your unzipped downloaded folder in easily retraceable location, such as your desktop, as you will need to access it multiple times throughout the experiment.

Watching a .m3u video - [SEE INSTRUCTIONS VIDEO](#) : (link)

11. First, open VLC media player and click on the icon with the two opposing arrow twice, as seen in the image below. By clicking twice, the icon will change to include a small 1. All clips should now be played in a loop. Note that the VLC video will commence with the first clip to be evaluated. Also note that VLC media player will remember this setting. In other words, you only have to do this step one.



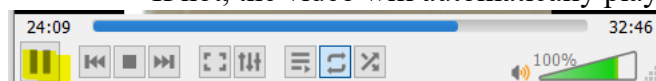
12. Close the VLC media player window.

13. In order to access the videos, please open the unzipped “Phase 1 Qualitative analysis” file.

14. As an example, we will open Phase1\_Participant01.m3u. If an icon of an orange cone appears next to the .m3u file, simply double click it.

- If not, right click the Phase1\_Participant01.m3u and select from the drop down menu open with > VLC media player.
- To ensure that all .m3u files are opened with VLC media player, right click on any .m3u file > click on properties > general > change > VLC media player > apply > OK

15. Open the video and click pause immediately, as seen in the image below. If not, the video will automatically play.



Note

16. **IMPORTANT:** Each video (.m3u file) has anywhere between 13 and 18 clips. A clip is defined as the moment a question is posed to the end of the participant's response. Consequently, the number of clips of a given participant within a video will be the equivalent to the number of interactions to be evaluated on Qualtrics. Clips are pre-defined within the video. Only watch the pre-defined clips of a video. It is therefore normal for the video to start past the 0 second mark.

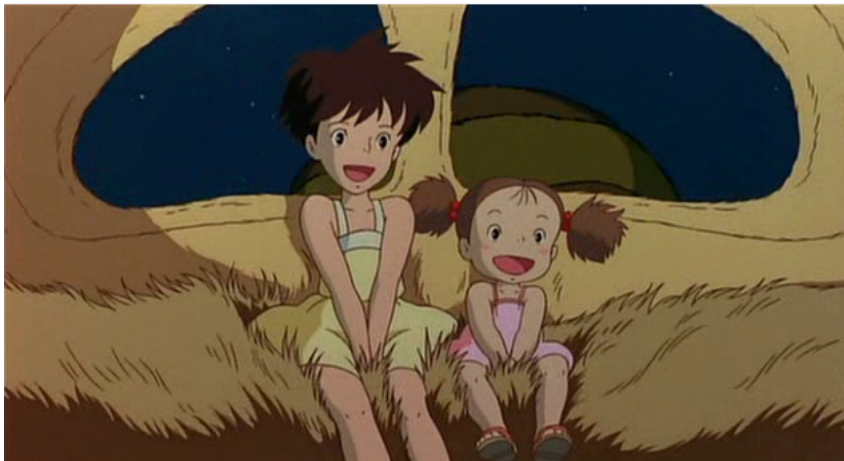
17. Please watch the clip **at least twice** in order to properly evaluate the participant's emotional output. You may want to watch the clip, conduct your evaluation upon Qualtrics, and re-assess your evaluation by re-watching the clip once more. A third re-watch might also be necessary.

18. Please note that the VLC player will automatically continue to play. Therefore, it is important to pause the clip when answering the Qualtrics questions.

19. To proceed to the next clip once the Qualtrics questions have been answered, click on >>. If needed, click on << once to start clip from the beginning, and twice to go the previous clip. Please only use these buttons to jump from one clip to another. **Do not** drag your cursor on the timeline, as the clips evaluated are characterized by specific timestamps.



20. The end of the series of clips within a video will be marked by the image below. Once the final clip has been evaluated, you may exit this video and proceed to the next participant evaluation by clicking the following video.



#### Accessing the correct Qualtrics links

The next set of instruction are found on Qualtrics.

Make sure to read all of the instructions before commencing the evaluation.

Note that each participant evaluation will be done using a separate Qualtrics link.

In order to access the links and the order of the evaluation, please click on the link below associated to your evaluator number. Note that every evaluation must be confirmed using the chart found within the link.

Evaluators	Link
------------	------



Evaluator 1	link
Evaluator 2	link
Evaluator 3	link
Evaluator 4	link
Evaluator 5	link
Evaluator 6	link

With this said, please click the Qualtrics link for your first participant. Please refer to your personalized link in order to access the respective links in order.

## Appendix 3

### Third-party evaluator Qualtrics questionnaire instructions

#### Introduction

Thank you for participating in the following study.

A series of short video clips featuring participants from a previous study will be presented using the VLC player.

As an evaluator, your role is to interpret the emotional output expressed by the featured participant using visual and vocal cues.

The participant's emotional expression will be evaluated using four dimensions. Each dimension has its own question. In other words, you will be asked to answer four questions per clip. A series of pre-defined scales and answers will be presented. Please select the best suited answer.

Please answer all questions before proceeding to watch the following clip, and before clicking the arrow found on at the bottom of the Qualtrics page.

You are encouraged to watch the clip at least twice in order to properly assess the emotional output. In order to assist you in your evaluation, a series of guidelines, as well as the defined dimensions, are presented on the following pages. A link towards the summary of guidelines and dimensions will be made available in the following pages. To learn more about the dimensions, please click the arrow below.

#### Dimension 1: Valence

Valence Definition: The continuum range of negative and extreme unhappiness or dissatisfaction, to positive and extreme happiness or satisfaction

Positive Valence: A perceived positive emotional state of being defined by high levels of happiness, satisfaction or pleasure.

Extreme Negative Valence: A perceived negative emotional state of being defined by high levels of unhappiness, dissatisfaction or displeasure.

Vocal cues to look out for (applies to both positive and negative valence):

Utterance duration

Inter-word silence

Pitch and energy values

Exaggerated or hyper-articulated speech

Visual cues to look out for (applies to both positive and negative valence):

Eyeblink movements and direction

Mouth movements and direction

Eyelid opening

Posture

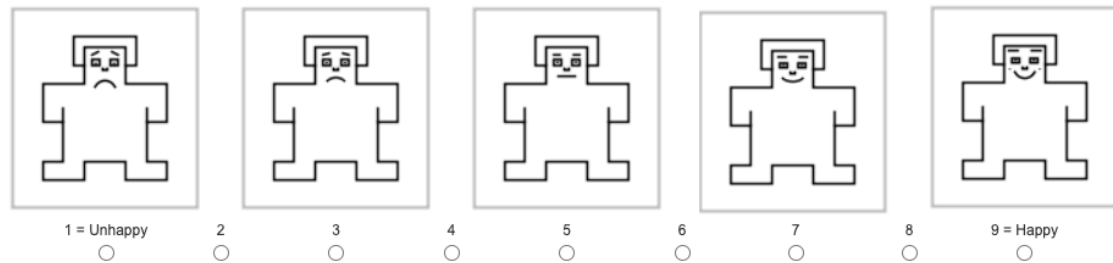
Notes:

- Make sure to take into consideration the context and question posed when assessing valence levels.

- Reminder that the examples provided are in no way an exhaustive list.

To rate the valence levels, you will be using the SAM scale.

Using the following scale, you will be asked to select the number that best corresponds to the participant's valence level (1 = unhappy ; 9 = happy)



## Dimension 2: Arousal

Arousal Definition: The continuum range of extreme calmness to extreme excitement

Low Arousal: A perceived emotional state of calmness

High Arousal: A perceived emotional state of excitement

Low Arousal Voice and Visual Cues:

Head bent forward

Hands/arms close to the body

Tight eyelids or eye closure

High Arousal Voice and Visual Cues:

Head bent backward

Hands/arms vertically extended

Upper Eyelids raised (wide-eyes)

Raised brows

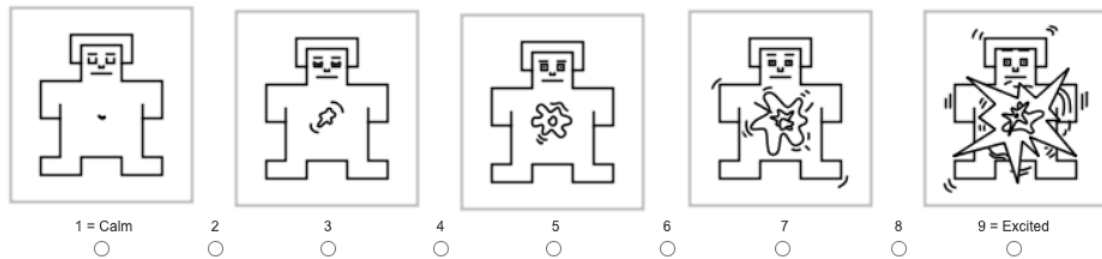
Note:

- Consider asking yourself “To what extent does this participant seem agitated by the voice interface’s question?” in order to evaluate this particular dimension.

- Make sure to take into consideration the context and question posed when assessing valence levels.

- Reminder that the examples provided are in no way an exhaustive list.

Using the following scale, you will be asked to select the number that best corresponds to the participant's arousal level (1 = calm ; 9 = excited)



### Dimension 3: Level of control

Level of control Definition: The continuum range of being controlled to being in-control

In-Control: A perceived display of being in-control or dominant

Visual and audio cues and examples:

- Timely response
- Quick nod with vocal response
- Engaged posture
- Focused gaze

Controlled: A perceived display of being controlled or dominated

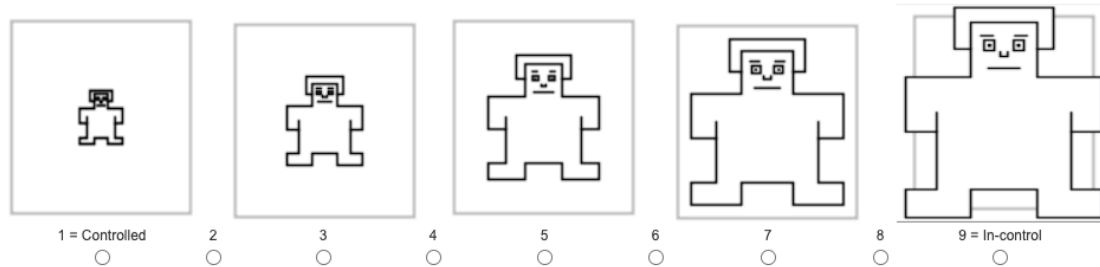
Audio cues and examples:

- Filled Pause such as “Uh” or “Um” before response
- Drawn out vocal response
- Restless gaze
- Closed posture

Note:

- It is possible that the emotional expression associated to feeling of being controlled overlaps with annex negative states of being, such as annoyance. As for being in-control, a similar sense of confidence may be perceived.
- Ask yourself "To what degree is the participant in control of his/her emotions"
- Consider the context and question posed when assessing the participant's level of control.
- Reminder that the examples provided are in no way an exhaustive list.

Using the following scale, you will be asked to select the number that best corresponds to the participant's control level (1 = controlled; 9 = in-control)



#### Dimension 4: Short-term Emotional Episode

Short-term Emotional Event Definition: A temporary and sudden display of apparent emotion preceded by and followed by a stable behaviour.

Positive Short-term Emotional Event: A temporary and sudden display of positive emotion such as joy or amusement, depicted by an apparent change of behaviour seen in either or both a facial expression or vocal output.

Vocal and visual cues examples:

- Sudden laugh
- Chuckle

Negative Short-term Emotional Event: A temporary and sudden display of negative emotion such as frustration, depicted by an apparent visual or vocal reaction.

Vocal and visual cues examples:

- Scoff
- Sigh
- Eye-roll
- Head shake
- Squint

Other Short-term Emotional Episode visual cue examples (both for positive and negative)

- Lip-biting

Note:

- Short-term emotional episodes may be subtle as they are quick.
- Short-term emotional episodes can be frequent and therefore may appear more than once within a given clip.
- Its temporal nature, rather than its intensity, is to be prioritized when evaluating this specific dimension.
- Reminder that the examples provided are in no way an exhaustive list.

**For each interaction, you will be asked to select one of the three options that best describe the presence, or lack of, a short term emotional episode.**

- ☐ Positive Short-term Emotional Episode
- ☐ Negative Short-term Emotional Episode
- ☐ No Short-term Emotional Episode

The dimensions and guidelines presented may be found on the following link: (link)

Please copy and paste the link in a separate page.

Please watch the clip at least twice in order to properly evaluate the participant's emotional output. You may want to watch the clip, conduct your evaluation upon Qualtrics, and re-assess your evaluation by re-watching the clip once more. A third re-watch might also be necessary.

Please note that the VLC player will automatically continue to play. Therefore, it is important to pause the clip when answering the Qualtrics questions.

When you are ready to begin the evaluation, please open the video associated to the first participant to be evaluated according to your personalized excel list. This list may be found in the "Accessing the correct Qualtrics links" section of the protocol.

Once the correct video is open, you may click the arrow below to commence the evaluation.

# Appendix 4


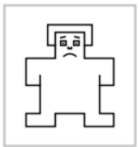
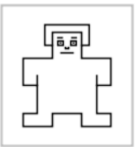
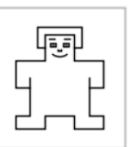
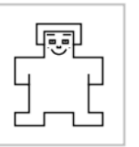
## Third-party evaluator Qualtrics page example

Clip 1

Please confirm your comprehension by typing the question posed by the voice interface. Paraphrasing is acceptable.

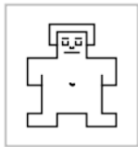
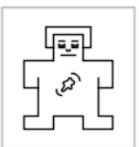
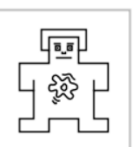
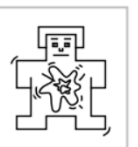

Valence

To what extent does the participant appear to be completely happy or completely unhappy? (1=unhappy, 9 = happy)

								
1 = unhappy	2	3	4	5	6	7	8	9 = Happy
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>


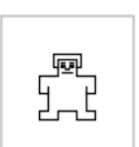
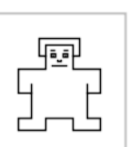
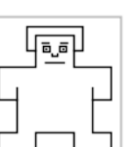
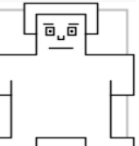
Arousal

To what extent does the participant appear to be calm or excited? (1=calm, 9 = excited)

								
1 = Calm	2	3	4	5	6	7	8	9 = Excited
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Control

To what extent does the participant seem to be controlled or in-control? (1 = controlled; 9 = in-control)

								
1 = Controlled	2	3	4	5	6	7	8	9 = In-control
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Was there a presence of a short-term emotional episode? Select the option that best describes event.

- ☐ Positive short-term emotional episode
- ☐ Negative short-term emotional episode
- ☐ No short-term emotional episode

Evaluator's comments

