applied sciences

MDPI

Article

Comparing the Effectiveness of Speech and Physiological Features in Explaining Emotional Responses during Voice User Interface Interactions

Danya Swoboda [1], Jared Boasen [1,2], Pierre-Majorique Léger [1,*], Romain Pourchon [3] and Sylvain Sénécal [1]

[1] Tech3Lab, HEC Montréal, Montréal, QC H3T 1T7, Canada; danya.swoboda@hec.ca (D.S.); jared.boasen@hec.ca (J.B.); sylvain.senecal@hec.ca (S.S.)
[2] Faculty of Health Sciences, Hokkaido University, Sapporo 060-0812, Japan
[3] Deloitte, Montréal, QC H3T 1T7, Canada; rpourchon@deloitte.ca
* Correspondence: pierre-majorique.leger@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) methods. 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. Thus, we sought to contribute to the emerging literature regarding voice user interface evaluation and, consequently, assist user experience professionals in their quest to create optimal vocal experiences. To do so, we compared the effectiveness of physiological features (e.g., phasic electrodermal activity amplitude) and speech features (e.g., spectral slope amplitude) to predict the intensity of users' emotional responses during voice user interface interactions. We performed a within-subjects experiment in which the 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 insights favoring efficient voice user interface evaluation.

Keywords: voice user interface; implicit measures; emotional valence; emotional arousal; user experience

check for updates

Citation: Swoboda, D.; Boasen, J.; Léger, P.-M.; Pourchon, R.; Sénécal, S. Comparing the Effectiveness of Speech and Physiological Features in Explaining Emotional Responses during Voice User Interface Interactions. *Appl. Sci.* **2022**, *12*, 1269. https://doi.org/10.3390/app12031269

Academic Editor: Ilsun Rhiu

Received: 21 December 2021
Accepted: 19 January 2022
Published: 25 January 2022

1. Introduction

The history of interface design has primarily revolved around Graphical User Interfaces (GUI), resulting in longstanding and familiar frameworks [1]. 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 [2,3]. 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 [4]. By 2024, this number is projected to reach 8.4 billion, a number greater than the world's population [4]. With this said, a set of validated voice user interface heuristics and guiding principles has yet to breakthrough.

Research within the field has recently tried to address this matter. For example, Nowacki et al. [5] developed an adapted version of Bastien Scapin's ergonomic criteria to vocal interfaces. On the other hand, Seaborn and Urakami [6] 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 and Urakami [6] highlighted the fact that numerous studies in relation to voice UX rely heavily on self-reported measures, defined by the users' own report of their states of being. According to the authors, self-reported measures, based on psychometric scales, were widespread and consequently called for the development of measures, such as behavioral measures, to support findings. Self-reported measures fall within the realm of explicit measures, characterized by the conscious behaviour of subjects under scrutiny. Explicit measures, such as self-reported measures, are often adopted due to their inexpensive nature [7]. 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 [8]. 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 [9]. As opposed to explicit measures, implicit measures tap into the subconscious behavior of users and can be reflected in speech or physiology. In the study of emotions during voice user interface interactions, the measure of speech is an obvious implicit measurement target due to the vocal nature of the interaction. The data obtained from this measure can be analyzed under various lenses through extracted speech features, including, for instance, pitch and fundamental frequency. However, physiological measures such as electrodermal activity (EDA) and facial micro expressions and their respective features (e.g., phasic EDA and the valence of facial micro expressions, respectively) have the potential to be equally revealing of emotional events [10,11].

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 [12–16], and do not additionally utilize implicit measures. Thus, combining self-reported measures in addition to utilizing implicit methods to evaluate voice user interface interactions is rare and constitutes an important gap in the literature.

Addressing this gap could potentially improve the insight UX professionals can obtain when studying voice interfaces within a business context. Indeed, 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 measurement methods may fail to fully reveal the underlying emotions experienced by users. Comparing and contrasting the strength or effectiveness of speech and physiological measures through their respective features when observing emotional dimensions could 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 human–computer interaction (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 of EDA and automatically analyzed facial micro expression (AFE), during emotionally charged voice user interface interactions. The effectiveness of eight extracted speech and three extracted physiological features 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 provide support for the inclusion of physiological measurements in UX evaluations of voice interfaces.

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.

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, the authors suggest that those regarding voice user interface evaluation lack consensus amongst UX [6], 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 the function of GUIs [1]. 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 [1]. 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 [9,15]. Implicit measures allow for real-time and precise data free of retrospective and cognitive biases to be collected [9]. 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 [9,17–19]. Thus, by including implicit measures, a more thorough understanding of the users' emotions and, consequently, their experiences, may be noted.

2.1. Speech Features

When considering implicit methods, one obvious choice for assessing changes in the 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 [20–25]. In the field of emotion detection and speech research, common prosodic features such as fundamental frequency (F0) (e.g., minimum, maximum, mean, jitter) and energy (e.g., loudness, shimmer) as well as duration are often observed and considered among the most common [21,26,27]. 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 [28,29].

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 [30,31]. Deriving from F0, pitch period entropy (PPE) is a measure that denotes the impaired control of F0 during sustained phonation [32,33]. 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 [34]. As for spectral entropy, it can assess silence and voice region of speech [35]. 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.

Speech Features	Definition
Fundamental frequency (F0)	The rate of vocal fold vibration.
Pitch period	The fundamental period of the signal.
Pitch period entropy (PPE)	The impaired control of F0 during sustained phonation.
Spectral slope	The observed tendency to have low energy during high frequencies.
Spectral spread	The total bandwidth of a speech signal using spectral centroid.
Spectral centroid	A measure used to evaluate the brightness of a speech.
Spectral entropy	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’s disease [32,33]. When assessing affective states, various speech features have been used simultaneously by researchers. As seen within a study by Papakostas et al. [36], spectral entropy, alongside spectral centroid, spectral spread, and energy, was observed in the aim of analyzing speakers’ emotions. In research by Lausen and Hammerschmidt [26], 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 [37]. In line with SER systems, emotion voice conversion is meant to generate expressive speech from neutral synthesized speech or natural human voice [38]. For example, research by Xue et al. [39] analyzed F0, power envelope, and 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 of valence and arousal serving as a control to the degrees of emotion. Valence refers to the degree of pleasure or displeasure, whereas arousal denotes the levels of alertness [40]. 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, and spectral skewness, in addition to prosodic features of energy and pitch [41].

In the context of voice user interface evaluation, a study by Kohh and Kwahk [42] 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 and Kwahk's study [42], 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:

Hypothesis 1 (H1). *There is a relationship between the amplitude of targeted speech features and the emotional intensity of users during voice user interface interactions.*

2.2. Physiological Features

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 [39]. Two common physiological indices used to measure the valence and arousal dimensions defining affective state are facial micro expressions and EDA. Often captured via a webcam, facial micro expressions are generally quantified using some form of automated facial micro expression (AFE) analysis software and assessed through the lens emotional valence. Facial expression analysis remains one of the most reliable ways to physiologically measure emotional valence, as facial muscles' micromovements will involuntarily occur as a direct result of changes in affective state [10]. Indeed, in one study, it was found that data captured via facial micro-expressions were more effective in measuring instant emotions and pleasure of use in comparison to a user questionnaire [43].

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 [44]. Simply put, it has been described as how users feel [45]. The dimension of valence can be studied alone or as a complementary construct to arousal, as 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 [11,46,47]. Indeed, it relates to the sympathetic nervous system, an automatic response to different situations [48]. The easy to use and reliable physiological measure has been widely used in NeuroIS research [48–52]. 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 in turn sensitive to changes in emotional arousal [53,54]. As suggested in the literature, it commonly infers levels of arousal through the measure of skin conductance [46].

The arousal levels measured via EDA range from very calm to neutral to highly stimulated [55]. It has been suggested to be an ecologically valid portrait of the user's arousal, while being non-invasive and free of overt recoded behaviour [18]. 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 [56].

2.3. Combination of Speech and Physiological Measures

Emotion is often expressed through several modalities [57]. 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 [58,59]. Indeed, research has suggested that EDA dynamics are strongly influenced by respiration and speech activity [54]. 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. [59], 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 Prasetio et al. [60] 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 [61]. To assess users' emotional states in naturalistic video sequences, a study by Caridakis et al. [61] combined information from both facial expression recognition and speech prosody feature extraction. A study by Castellano et al. [57] 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 [59], 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. [62], 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	Contribution
Greco et al. (2019)	Improved the recognition of human arousal level during the pronunciation of single affective words.	EDA Speech Features (F0 and MFCC)
Prasetio et al. (2020)	Developed a speech activity detection 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)

EDA: electrodermal activity; MFCC: Mel Frequency Cepstral Coefficients.

With this said, EDA, facial micro expressions, and speech are capable of explaining emotional states, both in isolation and in conjunction with each other. As stressed, facial expressions and electrodermal activity are indicative of a user's valence and arousal levels, making them pertinent measures to the study of emotions. This leads us to our following replication hypotheses:

Hypothesis 2a (H2a). *There is a relationship between the amplitude of the extracted EDA features and the emotional intensity of users during voice user interface interactions.*

Hypothesis 2b (H2b). *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 [63]. As suggested by Scherer [63], F0, energy, and rate are considered the most indicative of arousal. More precisely, high arousal is associated with high mean F0, F0 variability, fast speech rate, short pauses, increased voice intensity, and increased high frequency energy [64–72]. Indeed, emotions associated with high levels of physiological arousal, such as anger, fear, joy, and anxiety, have depicted increases in mean F0 and F0 variability, in addition to vocal intensity [30]. 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 [30]. With this said, vocal aspects can covary with emotional attributes, which reflect and communicate arousal levels associated to emotional reactions [63]. Across studies, results regarding arousal and speech remain consistent [73].

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 [68,69,72,74,75]. In others, valence is not associated to specific patterns of vocal cues [65,67,70]. Moreover, research has suggested that valence values are better assessed using facial features in comparison to acoustic features [76,77]. 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.

Hypothesis 3 (H3). *Physiological features are more explicative of emotional voice interaction events in comparison to speech features.*

3. Methods

3.1. Experimental Design

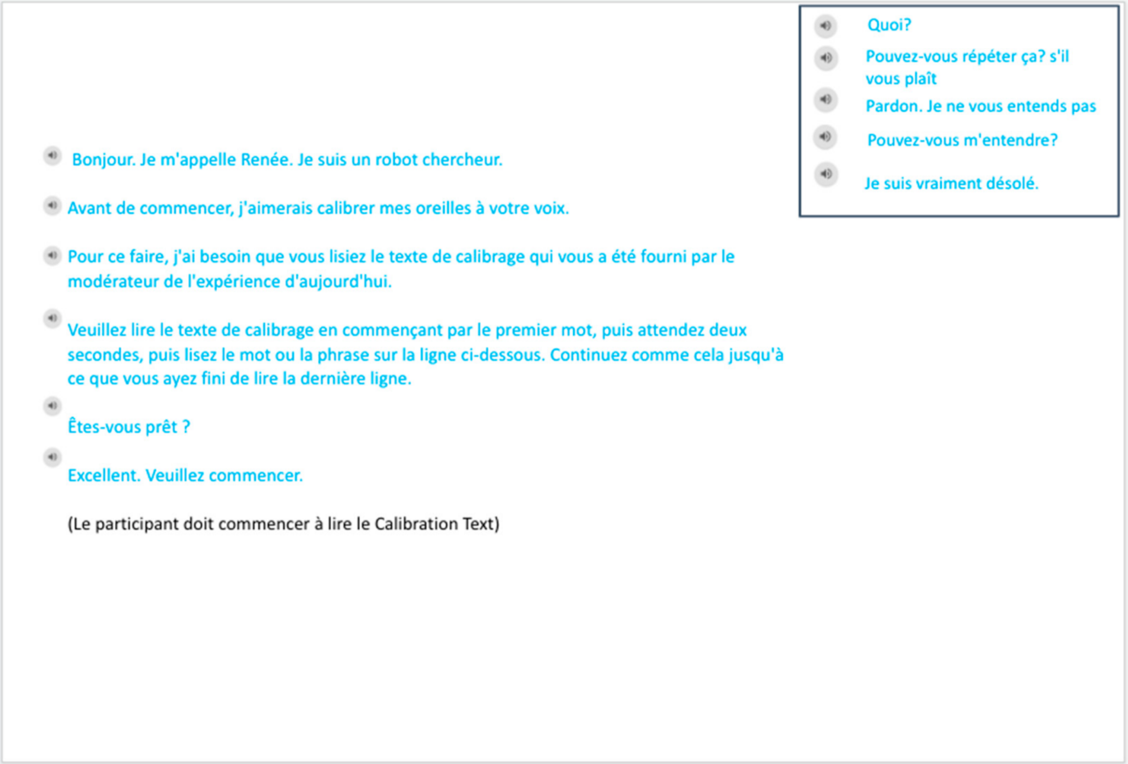
To test our hypotheses, we conducted a one-factor within-subject remote laboratory experiment in which speech and physiological responses, including EDA and facial expressions, were recorded during voice user interface interactions that were purposely designed to elicit intense emotional responses. Considering the nature of the COVID-19 pandemic, a remote experimental laboratory was made mandatory. The experiment followed guidelines established for remote data collection [78,79].

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). All participants were adept at using computers and had no trouble with using the software and tools required for the study. 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. 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/, accessed on 20 December 2021) 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.



Quoi?

Pouvez-vous répéter ça? s'il vous plaît

Pardon. Je ne vous entends pas

Pouvez-vous m'entendre?

Je suis vraiment désolé.

Bonjour. Je m'appelle Renée. Je suis un robot chercheur.

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 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.

Êtes-vous prêt ?

Excellent. Veuillez commencer.

(Le participant doit commencer à lire le Calibration Text)

Figure 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 elicit 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 participant's 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 and a chart featuring the number of questions posed can be seen in Tables A1 and A2, respectively, in Appendix A, in which both the original French dialogue used for the experiment and the translated English version are featured.

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, Santa Clara, CA, USA). 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 were 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.

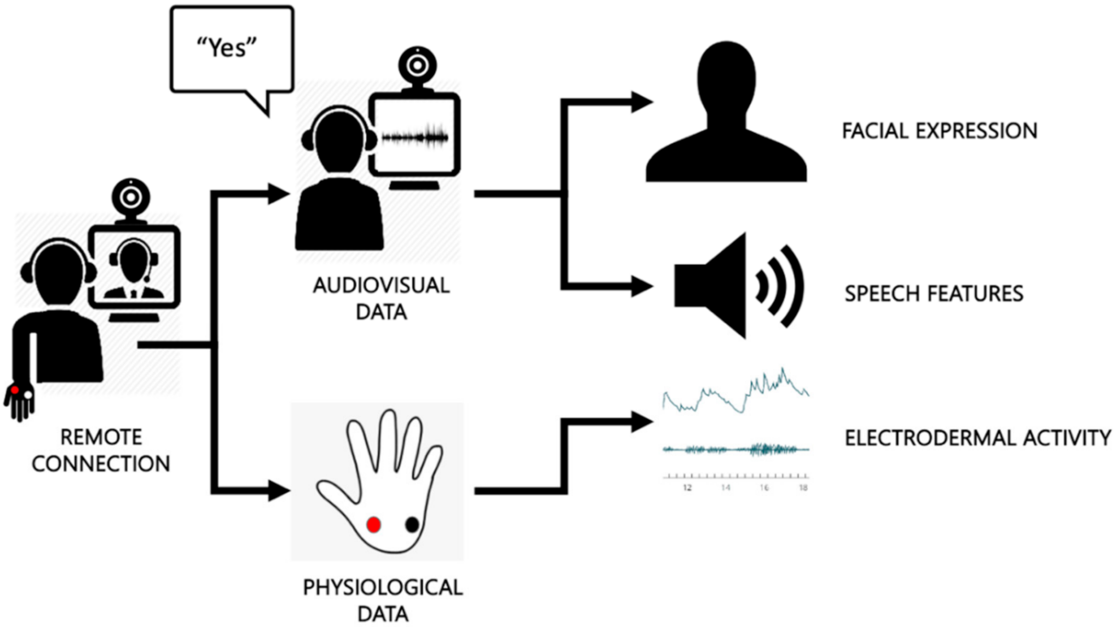


Figure 2. An overview illustration of the experimental setup.

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., Lisboa, Portugal) technology to record biosignals. EDA was captured via electrodes placed on the lower part of participant's palm, as depicted within the illustration featured in Figure 2, above. A photographic image of the placement is also found in Figure 3, below.

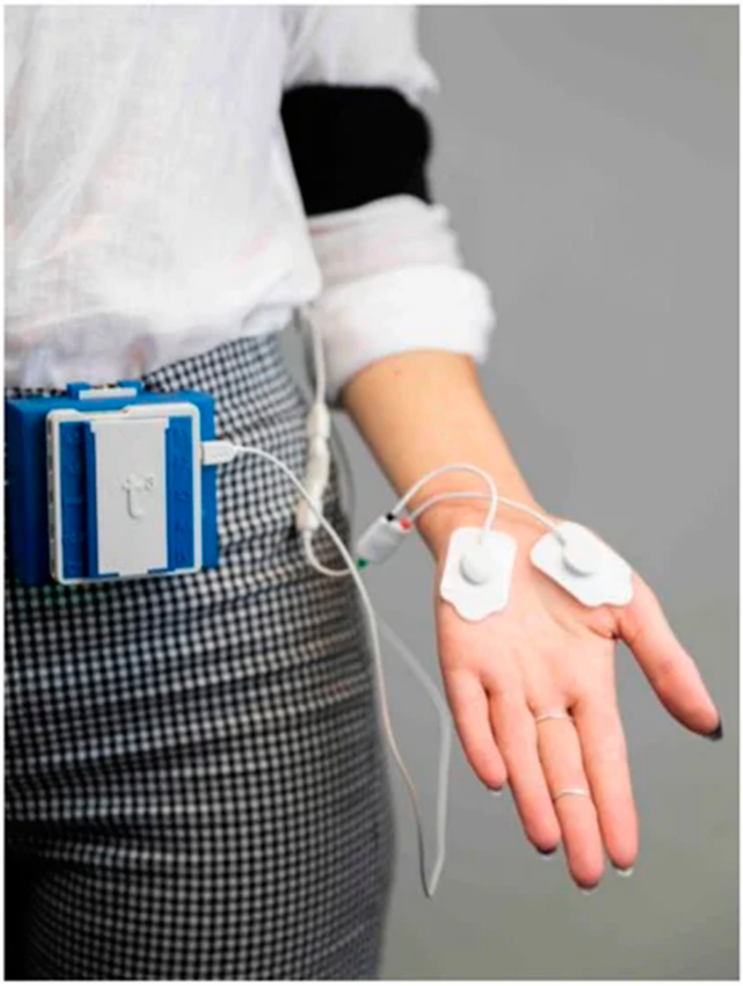


Figure 3. The electrodes placed on the participant's non-dominant hand are connected to sensor cables wired to the COBALT Bluebox device. Image source: Brissette-Gendron, R., Léger, P.M., Courtemanche, F., Chen, S.L., Ouhnana, M., and Sénécal, S. (2021). The response to impactful interactivity on spectators' engagement in a digital game. Multimodal Technologies and Interaction, 4(89), 89–89. https://doi.org/10.3390/mti4040089, accessed on 20 December 2021.

3.6. Experimental Procedures

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 consented to participation in the experiment, as well as to the recording of the session, screen, and physiological data. The moderator also validated that the informed consent form, sent 24 h 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 had conformed 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 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 recorded [80]. 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 in 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. [81], the synchronization technique used ensured the Bluetooth low energy (BLE) (Montréal, QC, Canada) signals were sent to the lightbox and BITalino device in range [82].

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 were provided to the participants. More precisely, the participant was informed that 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 to with either a "yes" or "no" response. If these answers did not apply to the question posed, the participant was instructed to answer with 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, UT, USA), an online survey tool. (Results from the sliding scales were purposely omitted from this study due to inconsistencies regarding evaluation time gaps between interactions). 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.

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 min. Once the final audio recording was played, the moderator turned on his or her camera and readjusted the sound preferences back to microphone setting. A summary of the procedures is found in the graphical representation in Figure 4.

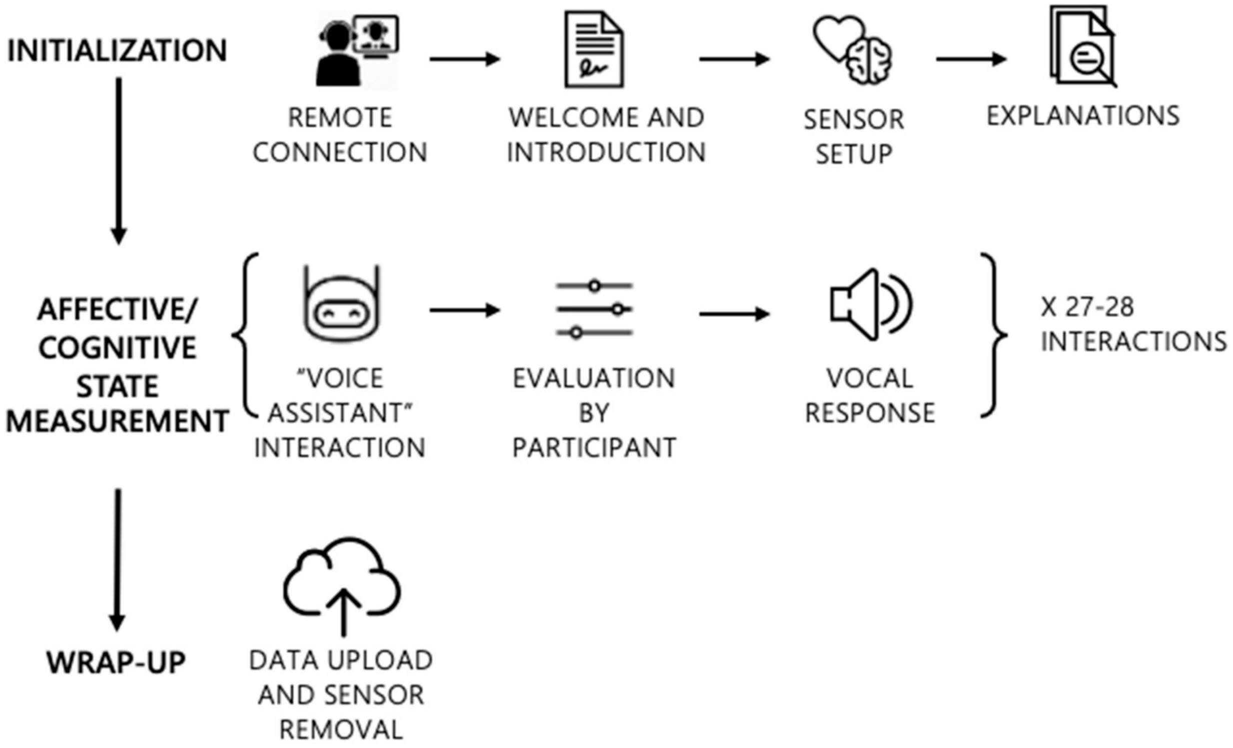

INITIALIZATION
REMOTE CONNECTION
WELCOME AND INTRODUCTION
SENSOR SETUP
EXPLANATIONS
AFFECTIVE/ COGNITIVE STATE MEASUREMENT
"VOICE ASSISTANT" INTERACTION
EVALUATION BY PARTICIPANT
VOCAL RESPONSE
X 27-28 INTERACTIONS
WRAP-UP
DATA UPLOAD AND SENSOR REMOVAL

Figure 4. Graphical summary of the experiment procedures.

3.7. Third-Party Emotion Evaluation

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 ms 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 the online survey tool Qualtrics. The survey used to record the evaluations was built into the platform and embedded on the page using custom HTML code. 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 [83]. 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 nine-point scale, although five and seven-point variants may also exist [84]. For this experiment, we opted for the nine-point scale in order to offer further precision and remain consistent with previous observer-based studies utilizing this measure [85,86].

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 bias, as evaluator fatigue upon the same final evaluation was avoided.

3.8. Data Processing and Feature Extraction

As a result of the recorded experiments, two raw data streams, 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, the open-source audio software Audacity (Muse Group, New York, NY, USA) was employed. In parallel, the video was processed using FaceReader 8 ™ (Noldus, Wageningen, The Netherlands) software, resulting in a time series data stream for AFE-based valence. The output, or time points, 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 response.

3.8.1. Speech Features

The onset of the participants' speech response was manually identified for every interaction where the response was "yes" and defined as the time point where the participants' speech envelope exceeded .10 decibels. This was performed as such 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 time point was identical to that of the participant's response. The time window consisted of the moment from the onset of participant responses until 500 ms 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) (http://sp-tk.sourceforge.net/, accessed on 20 December 2021). Congruent with existent research on emotion and speech [87], 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 among the most commonly analyzed with the study of emotion in speech [87].

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 [88,89]. The AFE analysis was subsequently conducted 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 [55,90], allowing researchers to distinguish between a set of discrete emotions, such as angry, happy, disgusted, sad, scared, and surprised.

The time-series data, from the onset of the voice user interface speech until the onset of participant response, were 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 were further processed for each interaction tested within the statistical analyses.

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, were 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 [11]. In other words, phasic EDA stems from faster changing elements of the signal, known as the Skin Conductance Response (SCR) [11]. As for the z-score, it requires the mean and standard deviation to be used in place of a hypothetical maximum [11].

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 were further processed to derive phasic and z-score features, serving as targets for an arousal assessment, for each interaction tested within the statistical analyses.

3.9. Statistical Analyses

Using SPSS® (IBM, New York, NY, USA), 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 [91,92]. ICC scores allow for both the degree of correlation and agreement between measurements to be reflected within a reliability index [93]. 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, and 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 α = 0.05, resulting in a significance threshold of $p \leq 0.0045$ [94].

4. Results

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.

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 [95]. Based on analysis standards, inter-rater agreement for STEE was considered adequate, as it fell within the 0.60 and 0.74 range [95]. 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 Figure 5a–d below, in which four distinct line graphs depicting the mean scores per evaluator, participant, and dimension are presented.

Table 3. Results of the ICC scores.

Dimension	ICC Scores	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
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

STEE: short-term emotional episodes.

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

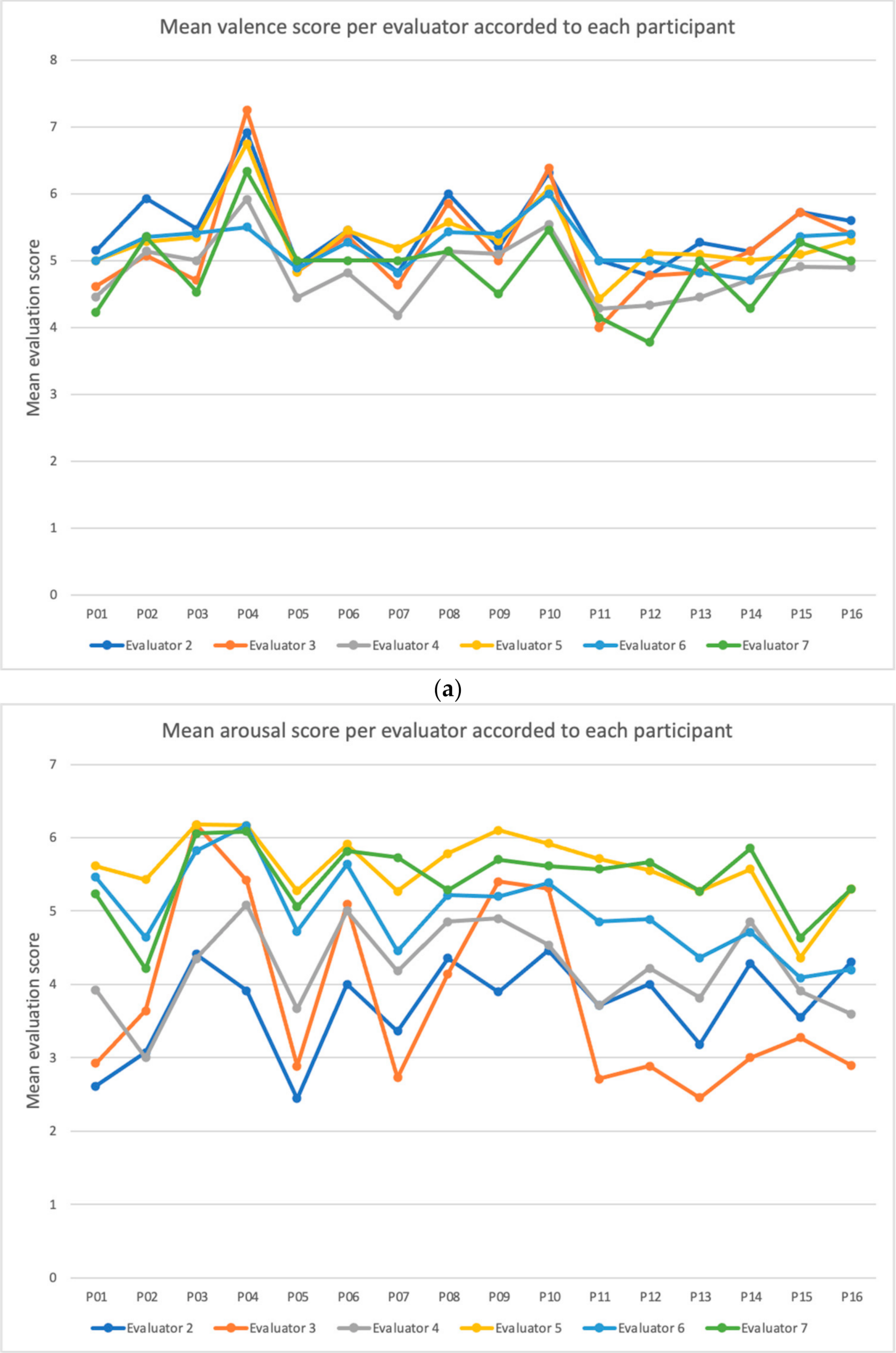
	Mean	Minimum	Maximum	Range	Maximum/ Minimum	Variance
Valence	5.187	4.862	5.516	0.654	1.135	0.061
Arousal	4.640	3.676	5.601	1.926	1.524	0.665
Control	5.537	4.723	6.404	1.681	1.356	0.542
STEE	−0.057	−0.0101	0.027	0.128	−0.0263	0

STEE: short-term emotional episodes.

4.2. Multiple Linear Regression

Table 5 presents 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 with 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

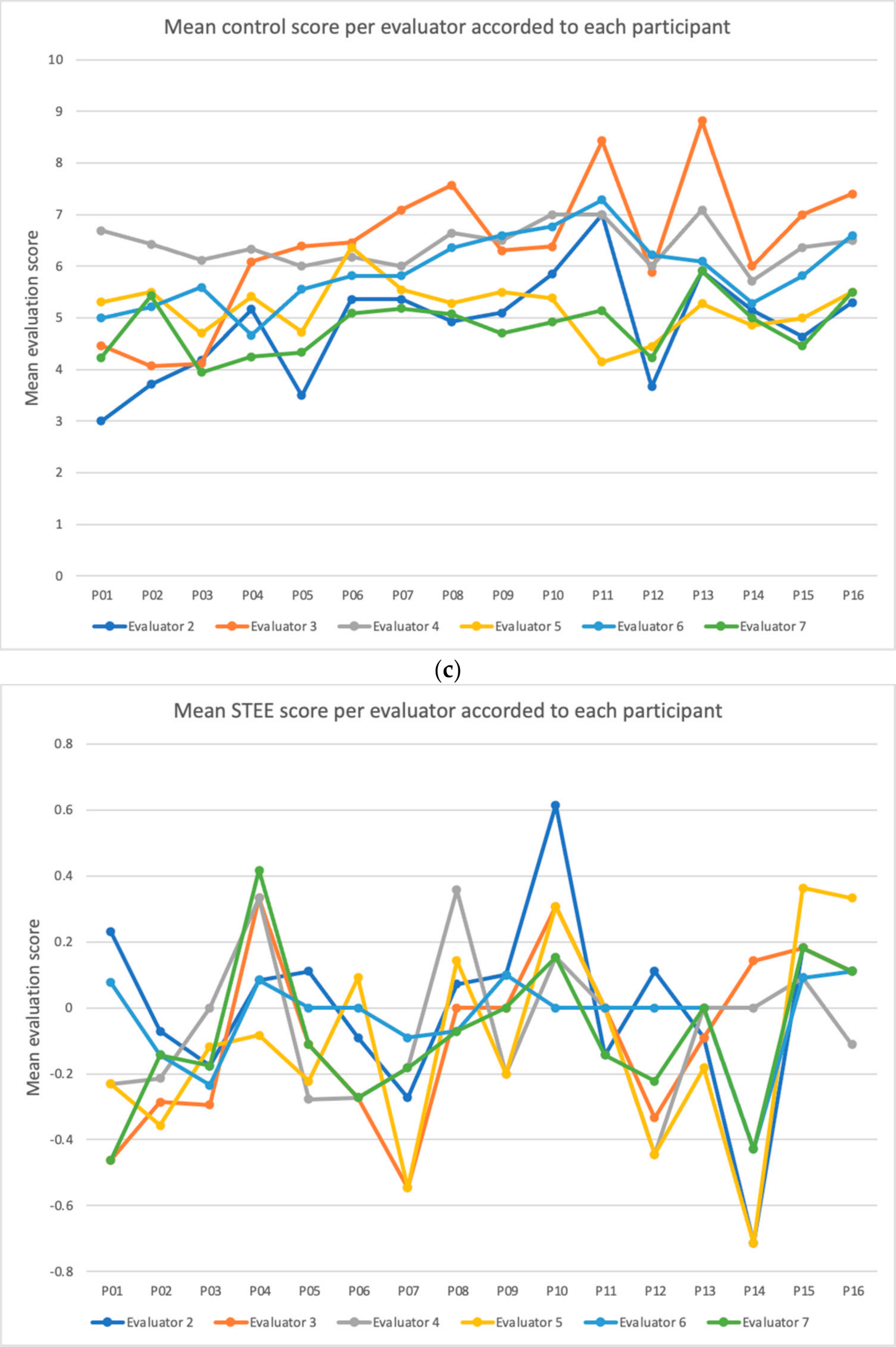
0.008 and 0.040 (see Table 7). The R-squared values associated with spectral slope were 0.048 and 0.028 for spectral spread. However, neither factor was considered statistically significant following the Bonferroni correction. Lastly, multiple linear regression revealed significant relationships between 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.



(**a**)

(**b**)

Figure 5. *Cont.*



(**c**)

(**d**)

Figure 5. (**a**–**d**) The evaluator scores of the dimensions of valence, arousal, and control are a function of the 9-point SAM scale, whereas the evaluator score of the dimension of STEE ranges from −2 to 2. (**a**) Mean valence score per evaluator accorded to each participant. (**b**) Mean arousal score per evaluator accorded to each participant. (**c**) Mean control score per evaluator accorded to each participant. (**d**) Mean STEE score per evaluator accorded to each participant.

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.

Table 5. Regression results of Valence dimension.

Factor	Estimate	SE [1]	DF [2]	T Value	*p* Value	R^2 Value [3]
AFE-based valence	3.076	0.351	129	8.770	<0.001 *[4]	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 [5]	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

[1] SE: Standard Error. [2] DF: Degree of Freedom. [3] R^2: R-Squared. [4] Significant factors following the Bonferroni correction, with threshold of 0.004, identified with *. [5] PPE: Pitch Period Entropy.

Table 6. Regression results of Arousal dimension.

Factor	Estimate	SE [1]	DF [2]	T Value	*p* Value	R^2 Value [3]
AFE-based valence	1.755	0.365	129	4.810	<0.001 *[4]	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 [5]	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

[1] SE: Standard Error. [2] DF: Degree of Freedom. [3] R^2: R-Squared. [4] Significant factors following the Bonferroni correction, with threshold of 0.004, identified with *. [5] PPE: Pitch Period Entropy.

Table 7. Regression results of Control dimension.

Factor	Estimate	SE [1]	DF [2]	T Value	*p* Value	R^2 Value [3]
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 [4]	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

[1] SE: Standard Error. [2] DF: Degree of Freedom. [3] R^2: R-Squared. [4] PPE: Pitch Period Entropy. Note: No feature was considered statistically significant.

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 indicative of a user's arousal during voice user interface interactions. Thus, H2a is not supported.

Table 8. Regression results of STEE dimension.

Factor	Estimate	SE [1]	DF [2]	T Value	*p* Value	R^2 Value [3]
AFE-based valence	0.936	0.171	129	5.480	<0001 *[4]	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 [5]	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

[1] SE: Standard Error. [2] DF: Degree of Freedom.[3] R^2: R-Squared. [4] Significant factors following the Bonferroni correction, with threshold of 0.004, identified with *. [5] PPE: Pitch Period Entropy.

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 ($p < 0.0001$). This fact remained valid following the Bonferroni correction. The R-squared value associated with 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, H2b is supported.

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 ($p < 0.0001$), 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 ($p < 0.0001$) (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 of 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 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 with 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 the physiological feature's EDA z-score in explaining control ratings. Hence, the relationship between speech factor spectral slope is stronger than the 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 Figures 6 and 7.

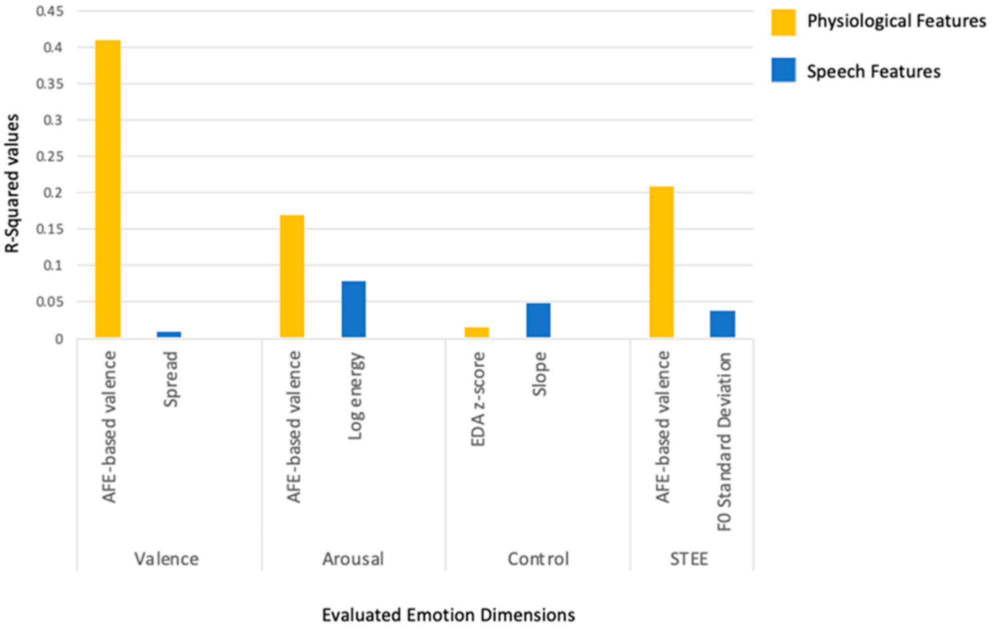


Figure 6. Bar chart of relationship strengths between physiological and speech features per dimension.

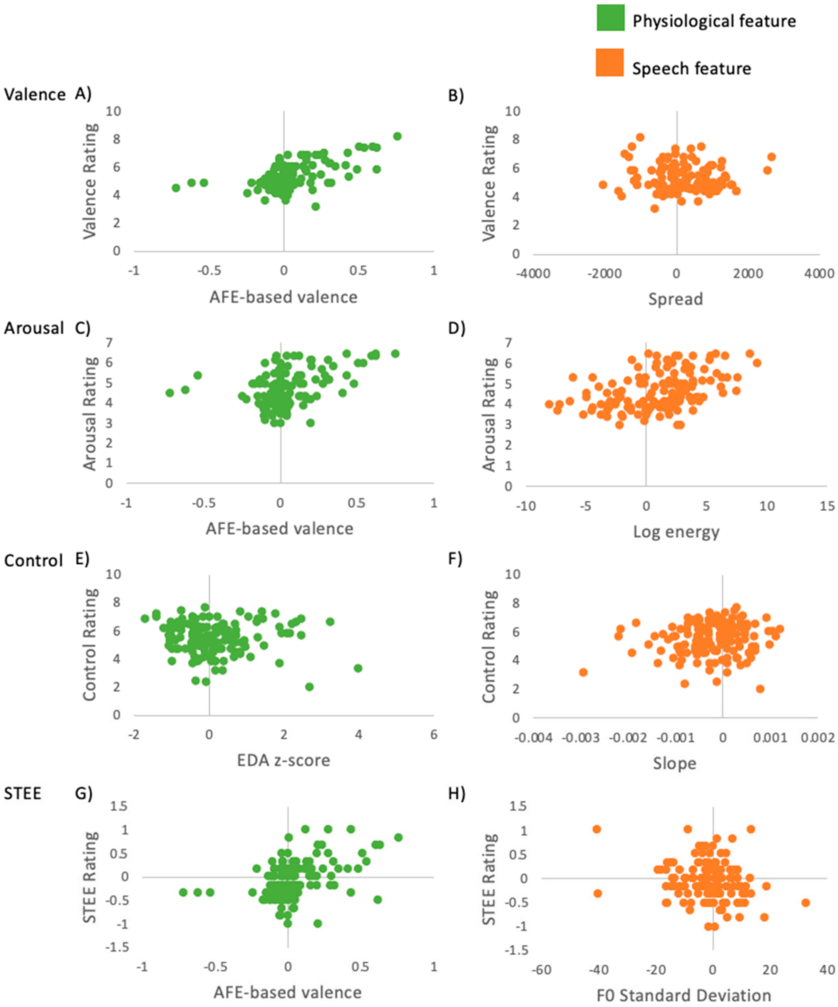


Figure 7. Boxplots of evaluator rating and select physiological and speech features.

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' statuses 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
H2a	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
H2b	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

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 the speech features of 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 [74]. 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 with high frequency energy [63,68,69,72]. 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 (H2a). 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 [96]. 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 through various modalities, including facial expressions and speech [58,59]. Enhanced arousal levels influence the intensity of facial reactions [97]. 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, H2a 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 (H2b). This result supports previous findings in emotion literature suggesting facial expression to be more indicative of valence than speech features [76,77]. Indeed, results correspond 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 [10]. Thus, H2b is supported.

On the contrary, research has suggested that there are no specific vocal cues associated with valence [65,67,70]. Moreover, the effects of valence are often vocally unapparent, as they are masked by other emotional dimensions such as arousal and dominance [98]. 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 [90]. Considering the time points chosen for speech analysis, STEEs were most likely excluded because 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.

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 [13,99,100]. Data obtained from explicit measures relying on self-reported measures can be flawed, as users are at risk of cognitive and retrospective biases [9]. By including implicit measures, our study avoids such biases while taking into account real-time, subconscious reactions linked to important emotional states [9]. 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 complementary to the current literature.

Secondly, few studies have observed the users' speech features during voice user interface interactions, and less have done so in combination with physiological measures, as research within the study of emotion through speech tends to focus on single sensor data [101]. Thus, utilizing multiple physiological measures within this field of research is 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 such as 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 is in line with previous research [26,34,36,43].

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 dominant. 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 its counterparts of valence and arousal [102]. 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 the speech feature of 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 [73]. Indeed, when observing the dimension of control in relation to spectral slope, research by Schröder et al. [69] suggest that low dominance is accompanied by a flatter spectral slope, contrary to results obtained by Banse and Scherer [103]. 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.

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% from 2020 [104]. 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 [105]. 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.

6. 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 herein 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 were 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 time points 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 [96], in conjunction to the time points chosen, indicative 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.

Author Contributions: Conceptualization, D.S., J.B., S.S. and P.-M.L.; methodology, D.S., J.B., S.S. and P.-M.L.; validation, J.B., S.S. and P.-M.L.; formal analysis, D.S. and J.B.; investigation, D.S. and J.B.; resources, P.-M.L., S.S. and J.B.; data curation, D.S., J.B., P.-M.L. and S.S.; writing—original draft preparation, D.S.; writing—review and editing, P.-M.L., J.B. and S.S.; supervision, J.B., P.-M.L. and S.S.; project administration, J.B.; funding acquisition, P.-M.L., S.S. and R.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the NSERC Industrial Research Chair 505259-16 Léger grant and by Deloitte Digital.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of HEC Montreal (Certificate #2021-4289, 18 December 2020).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data is available upon request.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

Table A1. 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 [1] Comment	Introduction/ Instructions		
[Robot] Acceptez-vous de participer?	[Robot] Do you agree to participate?	VUI Question	Confirmation	1	

Table A1. *Cont.*

Original French Question	English Translation of Questions	Type	Description	Question Number	Possibility of "Yes" Response
[pXX] [2] 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 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.	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. 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.	VUI Comment	Introduction/ Instructions		
[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		

Table A1. *Cont.*

Original French Question	English Translation of Questions	Type	Description	Question Number	Possibility of "Yes" Response
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 éléphantesques.	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	

Table A1. *Cont.*

Original French Question	English Translation of Questions	Type	Description	Question Number	Possibility of "Yes" Response
[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 [3]
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. Apologies 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

Table A1. *Cont.*

Original French Question	English Translation of Questions	Type	Description	Question Number	Possibility of "Yes" Response
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
Flow 1 Question 10	*Flow 1 Question 10*				
[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		
Flow 2 Question 10	*Flow 2 Question 10*				
[Si non ou je ne sais pas,] Je comprends. J'ai essayé de dire à l'administration ce qu'ils pourraient faire de mieux, mais personne ne semble m'écouter.	[If no or unsure] I understand. I tried to tell the administration what they could do better, but no one seemed to listen to me.	VUI Comment	Reply		
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		
Flow 1 Question 11	*Flow 1 Question 11*				
[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		

Table A1. *Cont.*

Original French Question	English Translation of Questions	Type	Description	Question Number	Possibility of "Yes" Response
[Permettre au participant de répondre]	[Allow the participant to respond]	Participant Response	Error		
Flow 2 Question 11	*Flow 2 Question 11*				
[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		
Flow 3 Question 11	*Flow 3 Question 11*				
[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 [4]	
[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		

Table A1. *Cont.*

Original French Question	English Translation of Questions	Type	Description	Question Number	Possibility of "Yes" Response
Flow 1 Question 14	*Flow 1 Question 14*				
[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		
Flow 2 Question 14	*Flow 2 Question 14*				
[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	

Table A1. *Cont.*

Original French Question	English Translation of Questions	Type	Description	Question Number	Possibility of "Yes" Response
[Permettre au participant de répondre]	[Allow the participant to respond]	Participant Response	Answer		Yes
Flow 1 Question 21	*Flow 1 Question 21*				
[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		
Flow 2 Question 21	*Flow 2 Question 21*				
[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		

Table A1. *Cont.*

Original French Question	English Translation of Questions	Type	Description	Question Number	Possibility of "Yes" Response
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		

[1] VUI: Voice User Interface. [2] [pXX]: Participant Number. [3] Certain participants answered with a "yes" response despite it being a typical "no" response. [4] 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 A2 below.

Table A2. 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

References

1. Murad, C.; Munteanu, C. Designing Voice Interfaces: Back to the (Curriculum) Basics. In Proceedings of the CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–12. [CrossRef]
2. Bastien, J.M.C.; Scapin, D.L. A validation of ergonomic criteria for the evaluation of human-computer interfaces. *Int. J. Hum.-Comput. Interact.* **1992**, *4*, 183–196. [CrossRef]
3. Nielsen, J. Usability inspection methods. In Proceedings of the CHI94: ACM Conference on Human Factors in Computer Systems, Boston, MA, USA, 24–28 April 1994; pp. 413–414.
4. Statista. The Most Important Voice Platforms in 2020. Available online: https://www.statista.com/chart/22314/voice-platform-ranking/ (accessed on 10 July 2021).
5. Nowacki, C.; Gordeeva, A.; Lizé, A.H. Improving the Usability of Voice User Interfaces: A New Set of Ergonomic Criteria. In *Design, User Experience, and Usability. Design for Contemporary Interactive Environments, Proceedings of HCII 2020: International Conference on Human-Computer Interaction, Copenhagen, Denmark, 19–24 July 2020*; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Cham, Switzerland, 2020; Volume 12201, pp. 117–133. [CrossRef]
6. Seaborn, K.; Urakami, J. Measuring Voice UX Quantitatively: A Rapid Review. In Proceedings of the Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohoma, Japan, 8 May 2021; pp. 1–8.
7. Alves, R.; Valente, P.; Nunes, N.J. The state of user experience evaluation practice. In Proceedings of the NordiCHI 2014: The 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational, Helsinki, Finland, 26–30 October 2014; pp. 93–102. [CrossRef]
8. Hura, S.L. Usability testing of spoken conversational systems. *J. Usability Stud.* **2017**, *12*, 155–163.
9. Ortiz de Guinea, A.; Titah, R.; Léger, P.-M. Explicit and implicit antecedents of users' behavioral beliefs in information systems: A neuropsychological investigation. *J. Manag. Inf. Syst.* **2014**, *30*, 179–210. [CrossRef]
10. Den Uyl, M.J.; Van Kuilenburg, H. The Facereader: Online Facial Expression Recognition. In Proceedings of the Measuring Behavior 2005, Wageningen, The Netherlands, 30 August–2 September 2005.
11. Braithwaite, J.J.; Watson, D.G.; Jones, R.; Rowe, M. A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology* **2013**, *49*, 1017–1034.
12. Clark, L.; Doyle, P.; Garaialde, D.; Gilmartin, E.; Schlögl, S.; Edlund, J.; Aylett, M.; Cabral, J.; Munteanu, C.; Edwards, J.; et al. The state of speech in HCI: Trends, themes and challenges. *Interact. Comput.* **2019**, *31*, 349–371. [CrossRef]
13. Lopatovska, I.; Williams, H. Personification of the Amazon Alexa: BFF or a mindless companion. In Proceedings of the 2018 Conference on Human Information Interaction & Retrieval, New Brunswick, NJ, USA, 11–15 March 2018; pp. 265–268.

14. Garg, R.; Moreno, C. Exploring Everyday Sharing Practices of Smart Speakers. In Proceedings of the IUI Workshops, Los Angeles, CA, USA, 20 March 2019.
15. Sciuto, A.; Saini, A.; Forlizzi, J.; Hong, J.I. "Hey Alexa, What's Up?". In Proceedings of the Designing Interactive Systems Conference 2018—DIS '18, Hong Kong, China, 9–13 June 2018. [CrossRef]
16. Lopatovska, I.; Oropeza, H. User interactions with "Alexa" in public academic space. *Proc. Assoc. Inf. Sci. Technol.* **2018**, *55*, 309–318. [CrossRef]
17. Ortiz de Guinea, A.; Webster, J. An investigation of information systems use patterns: Technological events as triggers, the effect of time, and consequences for performance. *MIS Q.* **2013**, *37*, 1165–1188. [CrossRef]
18. Dirican, A.C.; Göktürk, M. Psychophysiological Measures of Human Cognitive States Applied in Human Computer Interaction. *Procedia Comput. Sci.* **2011**, *3*, 1361–1367. [CrossRef]
19. Ivonin, L.; Chang, H.-M.; Díaz, M.; Català, A.; Chen, W.; Rauterberg, M. Beyond Cognition and Affect: Sensing the Unconscious. *Behav. Inf. Technol.* **2014**, *34*, 220–238. [CrossRef]
20. Cordaro, D.T.; Keltner, D.; Tshering, S.; Wangchuk, D.; Flynn, L.M. The voice conveys emotion in ten globalized cultures and one remote village in Bhutan. *Emotion* **2016**, *16*, 117–128. [CrossRef]
21. Juslin, P.N.; Laukka, P. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychol. Bull.* **2003**, *129*, 770–814. [CrossRef] [PubMed]
22. Kraus, M.W. Voice-only communication enhances empathic accuracy. *Am. Psychol.* **2017**, *72*, 644–654. [CrossRef]
23. Laukka, P.; Elfenbein, H.A.; Thingujam, N.S.; Rockstuhl, T.; Iraki, F.K.; Chui, W.; Althoff, J. The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features. *J. Personal. Soc. Psychol.* **2016**, *111*, 686–705. [CrossRef]
24. Provine, R.R.; Fischer, K.R. Laughing, smiling, and talking: Relation to sleeping and social context in humans. *Ethology* **1989**, *83*, 295–305. [CrossRef]
25. Vidrascu, L.; Devillers, L. Real-life emotion representation and detection in call centers data. In *Affective Computing and Intelligent Interaction*; Tao, J., Tan, T., Picard, R.W., Eds.; Springer: Berlin, Germany, 2005; pp. 739–746.
26. Lausen, A.; Hammerschmidt, K. Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. *Humanit. Soc. Sci. Commun.* **2020**, *7*, 2. [CrossRef]
27. Johnstone, T.; Scherer, K.R. Vocal communication of emotion. *Handb. Emot.* **2000**, *2*, 220–235.
28. Tahon, M.; Degottex, G.; Devillers, L. Usual voice quality features and glottal features for emotional valence detection. In Proceedings of the 6th International Conference on Speech Prosody, Shanghai, China, 25 May 2012; Volume 2, pp. 693–697.
29. Shilker, T.S. Analysis of Affective Expression in Speech. Ph.D. Thesis, Cambridge University, Cambridge, UK, 2009.
30. Bachorowski, J.A. Vocal Expression and Perception of Emotion. *Curr. Dir. Psychol. Sci.* **1999**, *8*, 53–57. [CrossRef]
31. Li, S.Z.; Jain, A. Fundamental Frequency, Pitch, F0. In *Encyclopedia of Biometrics*; Springer: Boston, MA, USA, 2009. [CrossRef]
32. Little, M.A.; McSharry, P.E.; Hunter, E.J.; Spielman, J.; Ramig, L.O. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Trans. Bio-Med. Eng.* **2009**, *56*, 1015. [CrossRef]
33. Arora, S.; Baghai-Ravary, L.; Tsanas, A. Developing a large scale population screening tool for the assessment of Parkinson's disease using telephone-quality voice. *J. Acoust. Soc. Am.* **2019**, *145*, 2871–2884. [CrossRef]
34. Mannepalli, K.; Sastry, P.N.; Suman, M. Emotion recognition in speech signals using optimization based multi-SVNN classifier. *J. King Saud Univ. Comput. Inf. Sci.* 2018, *in press*. [CrossRef]
35. Toh, A.M.; Togneri, R.; Nordholm, S. Spectral entropy as speech features for speech recognition. *Proc. PEECS* **2005**, *1*, 92.
36. Papakostas, M.; Siantikos, G.; Giannakopoulos, T.; Spyrou, E.; Sgouropoulos, D. Recognizing emotional states using speech information. In *GeNeDis 2016*; Springer: Cham, Switzerland, 2017; pp. 155–164.
37. Wani, T.M.; Gunawan, T.S.; Qadri, S.A.A.; Kartiwi, M.; Ambikairajah, E. A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access* **2021**, *9*, 47795–47814. [CrossRef]
38. Robinson, C.; Obin, N.; Roebel, A. Sequence-to-sequence modelling of f0 for speech emotion conversion. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12 May 2019; pp. 6830–6834.
39. Xue, Y.; Hamada, Y.; Akagi, M. Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space. *Speech Commun.* **2018**, *102*, 54–67. [CrossRef]
40. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161–1178. [CrossRef]
41. Zhu, C.; Ahmad, W. Emotion recognition from speech to improve human-robot interaction. In Proceedings of the IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Fukuoka, Japan, 5–8 August 2019; pp. 370–375. [CrossRef]
42. Koh, Y.; Kwahk, J. B3-1 Analysis of User's Speech Behavior Pattern after Correction: Focusing on Smartphone Voice User Interface. *Jpn. J. Ergon.* **2017**, *53*, 408–411. [CrossRef]
43. Zaman, B.; Shrimpton-Smith, T. The FaceReader: Measuring instant fun of use. In Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles, Oslo, Norway, 14 October 2006; pp. 457–460. [CrossRef]
44. Lang, P.J.; Bradley, M.M.; Cuthbert, B.N. Emotion, motivation, and anxiety: Brain mechanisms and psychophysiology. *Biol. Psychiatry* **1998**, *44*, 1248–1263. [CrossRef]

45. Burton-Jones, A.; Gallivan, M.J. Towards a deeper understanding of system usage in organizations. *MIS Q.* **2007**, *31*, 657–679. [CrossRef]
46. Dawson, M.E.; Schell, A.M.; Filion, D.L. The electrodermal system. In *Handbook of Psychophysiology*; Cacioppo, J.T., Tassinary, L.G., Berntson, G.G., Eds.; Cambridge University Press: Cambridge, UK, 2007; pp. 159–181. [CrossRef]
47. Bethel, C.L.; Salomon, K.; Murphy, R.R.; Burke, J.L. Survey of psychophysiology measurements applied to human-robot interaction. In Proceedings of the RO-MAN 2007—The 16th IEEE International Symposium on Robot and Human Interactive Communication, Jeju, Korea, 26–29 August 2007; pp. 732–737.
48. Riedl, R.; Léger, P.M. *Fundamentals of NeuroIS: Information Systems and the Brain*; Studies in Neuroscience, Psychology and Behavioral Economics; Springer: Berlin/Heidelberg, Germany, 2016. [CrossRef]
49. Léger, P.-M.; Davis, F.D.; Cronan, T.P.; Perret, J. Neurophysiological Correlates of Cognitive Absorption in an Enactive Training Context. *Comput. Hum. Behav.* **2014**, *34*, 273–283. [CrossRef]
50. Vom Brocke, J.; Riedl, R.; Léger, P.-M. Application strategies for neuroscience in information systems design science research. *J. Comput. Inf. Syst.* **2013**, *53*, 1–13. [CrossRef]
51. Giroux-Huppé, C.; Sénécal, S.; Fredette, M.; Chen, S.L.; Demolin, B.; Léger, P.-M. *Identifying Psychophysiological Pain Points in the Online User Journey: The Case of Online Grocery*; Springer: Cham, Switzerland, 2019; pp. 459–473.
52. Lamontagne, C.; Sénécal, S.; Fredette, M.; Chen, S.L.; Pourchon, R.; Gaumont, Y.; De Grandpré, D.; Léger, P.M. User Test: How Many Users Are Needed to Find the Psychophysiological Pain Points in a Journey Map? In Proceedings of the International Conference on Human Interaction and Emerging Technologies, Nice, France, 26 July 2019; Springer: Cham, Switzerland, 2020; pp, 136–142
53. Hassenzahl, M.; Tractinsky, N. User Experience—A Research Agenda. *Behav. Inf. Technol.* **2006**, *25*, 91–97. [CrossRef]
54. Boucsein, W. *Electrodermal Activity*; Springer: Boston, MA, USA, 2012.
55. Ekman, P.; Friesen, W.V. *The Facial Action Coding System*; Consulting Psychologists Press: San Fransisco, CA, USA, 1978.
56. Leite, I.; Henriques, R.; Martinho, C.; Paiva, A. Sensors in the wild: Exploring electrodermal activity in child-robot interaction. In Proceedings of the 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Tokyo, Japan, 3–6 March 2013; pp. 41–48.
57. Castellano, G.; Kessous, L.; Caridakis, G. Emotion recognition through multiple modalities: Face, body gesture, speech. In *Affect and Emotion in Human-Computer Interaction*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 92–103.
58. Gross, J.J.; Muñoz Ricardo, F. Emotion regulation and mental health. *Clin. Psychol. Sci. Pract.* **1995**, *2*, 151–164. [CrossRef]
59. Greco, A.; Marzi, C.; Lanata, A.; Scilingo, E.P.; Vanello, N. Combining Electrodermal Activity and Speech Analysis towards a more Accurate Emotion Recognition System. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Berlin, German, 23–27 July 2019; pp. 229–232. [CrossRef]
60. Prasetio, B.H.; Tamura, H.; Tanno, K. Embedded Discriminant Analysis based Speech Activity Detection for Unsupervised Stress Speech Clustering. In Proceedings of the 2020 Joint 9th International Conference on Informatics, Electronics and Vision and 2020 4th International Conference on Imaging, Vision and Pattern Recognition, ICIEV and IcIVPR, Kitakyushu, Japan, 26–29 August 2020. [CrossRef]
61. Caridakis, G.; Malatesta, L.; Kessous, L.; Amir, N.; Raouzaiou, A.; Karpouzis, K. Modeling naturalistic affective states via facial and vocal expressions recognition. In Proceedings of the ICMI'06: 8th International Conference on Multimodal Interfaces, Banff, AB, Canada, 2–4 November 2006; pp. 146–154. [CrossRef]
62. Alshamsi, H.; Kepuska, V.; Alshamsi, H.; Meng, H. Automated Facial Expression and Speech Emotion Recognition App Development on Smart Phones using Cloud Computing. In Proceedings of the 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2018, Vancouver, BC, Canada, 1–3 November 2019; pp. 730–738. [CrossRef]
63. Scherer, K.R. Vocal affect expression: A review and a model for future research. *Psychol. Bull.* **1986**, *99*, 143–165. [CrossRef]
64. Breitenstein, C.; Van Lancker, D.; Daum, I. The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. *Cogn. Emot.* **2001**, *15*, 57–79.
65. Davitz, J.R. (Ed.) *The Communication of Emotional Meaning*; Mcgraw Hill: New York, NY, USA, 1964.
66. Levin, H.; Lord, W. Speech pitch frequency as an emotional state indicator. *IEEE Trans. Syst. Man Cybern.* **1975**, *5*, 259–273. [CrossRef]
67. Pereira, C. Dimensions of emotional meaning in speech. In Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, Newcastle, UK, 5–7 September 2000.
68. Scherer, K.R.; Oshinsky, J.S. Cue utilization in emotion attribution from auditory stimuli. *Motiv. Emot.* **1977**, *1*, 331–346. [CrossRef]
69. Schröder, M.; Cowie, R.; Douglas-Cowie, E.; Westerdijk, M.; Gielen, S. Acoustic correlates of emotion dimensions in view of speech synthesis. In Proceedings of the Seventh European Conference on Speech Communication and Technology, Aalborg, Denmark, 3–7 September 2001.
70. Apple, W.; Streeter, L.A.; Krauss, R.M. Effects of pitch and speech rate on personal attributions. *J. Personal. Soc. Psychol.* **1979**, *37*, 715–727. [CrossRef]
71. Kehrein, R. The prosody of authentic emotions. In Proceedings of the Speech Prosody 2002 International Conference, Aix-en-Provence, France, 11–13 April 2002. [CrossRef]
72. Pittam, J.; Gallois, C.; Callan, V. The long-term spectrum and perceived emotion. *Speech Commun.* **1990**, *9*, 177–187. [CrossRef]

73. Laukka, P.; Juslin, P.N.; Bresin, R. A dimensional approach to vocal expression of emotion. *Cogn. Emot.* **2005**, *19*, 633–653. [CrossRef]
74. Scherer, K.R. Acoustic concomitants of emotional dimensions: Judging affect from synthesized tone sequences. In *Nonverbal Communication*; Weitz, S., Ed.; Oxford University Press: New York, NY, USA, 1974; pp. 105–111.
75. Uldall, E. Attitudinal meanings conveyed by intonation contours. *Lang. Speech* **1960**, *3*, 223–234. [CrossRef]
76. Busso, C.; Deng, Z.; Yildirim, S.; Bulut, M.; Lee, C.M.; Kazemzadeh, A.; Lee, S.; Neumann, U.; Narayanan, S. Analysis of emotion recognition using facial expressions, speech and multimodal information. In Proceedings of the 6th International Conference on Multimodal Interfaces, State College, PA, USA, 13–14 October 2004; pp. 205–211.
77. Busso, C.; Rahman, T. Unveiling the acoustic properties that describe the valence dimension. In Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association, Portland, OR, USA, 9–13 September 2012.
78. Giroux, F.; Léger, P.M.; Brieugne, D.; Courtemanche, F.; Bouvier, F.; Chen, S.L.; Tazi, S.; Rucco, E.; Fredette, M.; Coursaris, C.; et al. Guidelines for Collecting Automatic Facial Expression Detection Data Synchronized with a Dynamic Stimulus in Remote Moderated User Tests. In Proceedings of the International Conference on Human-Computer Interaction, Washngton, DC, USA, 24–29 July 2021; Springer: Cham, Switzerland, 2021; pp. 243–254.
79. Vasseur, A.; Léger, P.M.; Courtemanche, F.; Labonte-Lemoyne, E.; Georges, V.; Valiquette, A.; Brieugne, D.; Rucco, E.; Coursaris, C.; Fredette, M. Distributed remote psychophysiological data collection for UX evaluation: A pilot project. In Proceedings of the International Conference on Human-Computer Interaction, Virtual Event, 24–29 July 2021; Springer: Cham, Switzerland, 2021; pp. 255–267.
80. Figner, B.; Murphy, R.O. Using skin conductance in judgment and decision making research. In *A Handbook of Process Tracing Methods for Decision Research*; Psychology Press: New York, NY, USA, 2011; pp. 163–184.
81. Courtemanche, F.; Fredette, M.; Senecal, S.; Leger, P.M.; Dufresne, A.; Georges, V.; Labonte-Lemoyne, E. Method of and System for Processing Signals Sensed from a User. U.S. Patent No. 10,368,741, 6 August 2019.
82. Courtemanche, F.; Léger, P.M.; Fredette, M.; Sénécal, S. *Cobalt—Bluebox: Système de Synchronisation et d'Acquisition Sans-Fil de Données Utilisateur Multimodales*; Declaration of Invention No. AXE-0045; HEC Montréal: Montreal, QC, Canada, 2022.
83. Bradley, M.M.; Lang, P.J. Measuring emotion: The self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* **1994**, *25*, 49–59. [CrossRef]
84. Betella, A.; Verschure, P.F. The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions. *PLoS ONE* **2016**, *11*, e0148037. [CrossRef] [PubMed]
85. Sutton, T.M.; Herbert, A.M.; Clark, D.Q. Valence, arousal, and dominance ratings for facial stimuli. *Q. J. Exp. Psychol.* **2019**, *72*, 2046–2055. [CrossRef] [PubMed]
86. Jessen, S.; Kotz, S.A. The temporal dynamics of processing emotions from vocal, facial, and bodily expressions. *NeuroImage* **2011**, *58*, 665–674. [CrossRef] [PubMed]
87. Yildirim, S.; Bulut, M.; Lee, C.M.; Kazemzadeh, A.; Busso, C.; Deng, Z.; Lee, S.; Narayanan, S. An acoustic study of emotions expressed in speech. In Proceedings of the Eighth International Conference on Spoken Language Processing, Jeju, Korea, 4–8 October 2004.
88. Skiendziel, T.; Rösch, A.G.; Schultheiss, O.C. Assessing the convergent validity between the automated emotion recognition software Noldus FaceReader 7 and Facial Action Coding System Scoring. *PLoS ONE* **2019**, *14*, e0223905. [CrossRef]
89. Lewinski, P.; den Uyl, T.M.; Butler, C. Automated facial coding: Validation of basic emotions and FACS AUs in FaceReader. *J. Neurosci. Psychol. Econ.* **2014**, *7*, 227. [CrossRef]
90. Cohn, J.F.; Kanade, T. Use of automated facial image analysis for measurement of emotion expression. In *Handbook of Emotion Elicitation and Assessment*; Oxford University Press: Oxford, UK, 2007; pp. 222–238.
91. Hallgren, K.A. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor. Quant. Methods Psychol.* **2012**, *8*, 23–34. [CrossRef]
92. Bartko, J.J. The intraclass correlation coefficient as a measure of reliability. *Psychol. Rep.* **1966**, *19*, 3–11. [CrossRef]
93. Koo, T.K.; Li, M.Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropract. Med.* **2016**, *15*, 155–163. [CrossRef]
94. Bland, J.M.; Altman, D.G. Multiple significance tests: The Bonferroni method. *BMJ* **1995**, *310*, 170. [CrossRef]
95. Cicchetti, D.V. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol. Assess.* **1994**, *6*, 284. [CrossRef]
96. Christopoulos, G.I.; Uy, M.A.; Yap, W.J. The Body and the Brain: Measuring Skin Conductance Responses to Understand the Emotional Experience. *Organ. Res. Methods* **2019**, *22*, 394–420. [CrossRef]
97. Fujimura, T.; Sato, W.; Suzuki, N. Facial expression arousal level modulates facial mimicry. *Int. J. Psychophysiol.* **2010**, *76*, 88–92. [CrossRef]
98. Patel, S.; Scherer, K.R.; Sundberg, J.; Björkner, E. Acoustic markers of emotions based on voice physiology. In Proceedings of the Conference: Speech Prosody, Chicago, IL, USA, 10–14 May 2010; Volume 2010.
99. Easwara Moorthy, A.; Vu, K.-P.L. Privacy Concerns for Use of Voice Activated Personal Assistant in the Public Space. *Int. J. Hum.-Comput. Interact.* **2015**, *31*, 307–335. [CrossRef]

100. Jiang, J.; Hassan Awadallah, A.; Jones, R.; Ozertem, U.; Zitouni, I.; Gurunath Kulkarni, R.; Khan, O.Z. Automatic Online Evaluation of Intelligent Assistants. In Proceedings of the 24th International Conference on World Wide Web—WWW'15, Florence, Italy, 18–22 May 2015. [CrossRef]
101. Ali, M.; Mosa, A.H.; Machot, F.A.; Kyamakya, K. Emotion Recognition Involving Physiological and Speech Signals: A Comprehensive Review. In *Recent Advances in Nonlinear Dynamics and Synchronization*; Kyamakya, K., Mathis, W., Stoop, R., Chedjou, J., Li, Z., Eds.; Studies in Systems, Decision and Control; Springer: Cham, Switzerland, 2018; Volume 109. [CrossRef]
102. Szameitat, D.P.; Darwin, C.J.; Wildgruber, D.; Alter, K.; Szameitat, A.J. Acoustic correlates of emotional dimensions in laughter: Arousal, dominance, and valence. *Cogn. Emot.* **2011**, *25*, 599–611. [CrossRef]
103. Banse, R.; Scherer, K.R. Acoustic profiles in vocal emotion expression. *J. Personal. Soc. Psychol.* **1996**, *70*, 614–636. [CrossRef]
104. Statista. Number of Digital Voice Assistants in Use Worldwide from 2019 to 2024 (in Billions). 2021. Available online: https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/ (accessed on 10 July 2021).
105. Le Pailleur, F.; Huang, B.; Léger, P.M.; Sénéecal, S. A new approach to measure user experience with voice-controlled intelligent assistants: A pilot study. In Proceedings of the HCII 2020: Human-Computer Interaction. Multimodal and Natural Interaction, Copenhagen, Denmark, 19–24 July 2020; Kurosu, M., Ed.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2020; Volume 12182, pp. 197–208. [CrossRef]