applied sciences

MDPI

Article

The Impression of Phones and Prosody Choice in the Gibberish Speech of the Virtual Embodied Conversational Agent Kotaro

Antonio Galiza Cerdeira Gonzalez *, Wing-Sum Lo and Ikuo Mizuuchi

Mechanical Systems Engineering Department, Graduate School of Engineering,
Tokyo University of Agriculture and Technology, Naka-cho 2-24-16, Koganei-shi 184-0012, Tokyo-to, Japan;
sam@mizuuchi.lab.tuat.ac.jp (W.-S.L.); ikuo@mizuuchi.lab.tuat.ac.jp (I.M.)
* Correspondence: antonio@mizuuchi.lab.tuat.ac.jp

Abstract: The number of smart devices is expected to exceed 100 billion by 2050, and many will feature conversational user interfaces. Thus, methods for generating appropriate prosody for the responses of embodied conversational agents will be very important. This paper presents the results of the "Talk to Kotaro" experiment, which was conducted to better understand how people from different cultural backgrounds react when listening to prosody and phone choices for the IPA symbol-based gibberish speech of the virtual embodied conversational agent Kotaro. It also presents an analysis of the responses to a post-experiment Likert scale questionnaire and the emotions estimated from the participants' facial expressions, which allowed one to obtain a phone embedding matrix and to conclude that there is no common cross-cultural baseline impression regarding different prosody parameters and that similarly sounding phones are not close in the embedding space. Finally, it also provides the obtained data in a fully anonymous data set.

Keywords: embodied conversational agents; synthetic gibberish speech; prosody generation

Citation: Gonzalez, A.G.C.; Lo, W.-S.; Mizuuchi, I. The Impression of Phones and Prosody Choice in the Gibberish Speech of the Virtual Embodied Conversational Agent Kotaro. *Appl. Sci.* **2023**, *13*, 10143. https://doi.org/10.3390/app131810143

Academic Editors: Jeonghye Han and Daniela Conti

Received: 26 July 2023
Revised: 2 September 2023
Accepted: 3 September 2023
Published: 8 September 2023

1. Introduction

As humanity moves toward the Society 5.0 [1] paradigm and industries move toward the Industry 4.0 [2,3] model, smart devices will become increasingly ubiquitous; it is expected that more than 100 billion such devices will exist by 2050 [4]. With more computing power and smaller (or no) screens, the user interface design paradigm is shifting from graphical user interfaces (GUI) to conversational user interfaces (CUI) [5]. This can already be seen in personal assistants for computers, smartphones, and smart speakers (e.g., Cortana, Alexa, Siri, Bixby, Google Assistant, etc.).

Thus, it is important to develop systems capable of selecting good prosody parameters for the synthesized speech of conversational agents to ensure smooth human–machine interaction and even improving the quality of life of users [6]. Many such systems have a visual representation of their physical and social presence, being called embodied conversational agents (hereafter referred to as ECA). It should be noted that not all devices need semantic speech to convey desired messages [7,8], such as success, failure, attention, or danger. To carry this out, several classes of auditory means have been used, such as sounds, music, gibberish speech etc., can be employed. In particular, the use of gibberish in affective computing offers a key advantage in that it allows for the communication of emotions without the need for understandable language.

That approach is useful for evaluating the effectiveness of affective prosodic strategies as well as for implementing functional systems. This paper investigates the effects of Gibberish Speech in a conversational setting, where humans can openly talk to robots and other embodied conversational agents without the fear of judgement, but still receiving responses that show that the conversational agent is listening to what the human says and is engaged in the conversation, without actually saying anything. Gibberish speech holds

a presence within popular culture, notably within movies and TV series like Star Wars (featuring Rodian, Ewokese, Jawaese, Huttese, and other alien languages), Star Trek (where Vulcan and Klingon "languages" originated as gibberish speech in the original series) and Pingu (in order to teach children how to interpret media just from audio-visual cues). This inclusion serves to introduce an otherworldly or fantastical dimension to the narrative, effectively avoiding the need for the development of a fully coherent language. Despite its apparent randomness, gibberish speech manages to convey emotions and sentiments through character dialogues through its acoustic prosodic properties. In interactive media, gibberish speech has found its place in video games (such as "Star Fox Command" and "Papers, Please") and toys (like the Furby). These platforms utilize gibberish speech to craft immersive interactions that do not rely on conveying coherent meaning. Instead, they contribute to a unique and unconventional ambiance, enhancing the overall experience of the interaction. Despite such cultural presence, little human–robot and human–computer interaction research has been performed.

When automatically generating appropriate prosody for synthetic semantic speech, it is possible to learn from prosodic databases extracted from natural speech which patterns are appropriate for a given input text to make it sound natural [9,10] or to make a certain impression on the listener [11,12]. However, a similar approach for gibberish speech is not as feasible since, to the best of the authors' knowledge, there is only one emotional speech database for gibberish speech, the EMOGIB data set [13], from which it would be possible to extract prosodic parameters. However, it consists only of words composed by phones present in English and Dutch, which limits the ability to develop prosody attribution systems from it. Furthermore, the emotional label present in the data set is the perceived feeling that the utterances convey, not how they made the listeners feel.

To fill such gaps, the web-based crowdsourcing experiment "Talk to Kotaro" [14] was conducted to investigate how IPA-based gibberish speech affect humans with distinct cultural backgrounds and to generate an data set with different acoustic prosody patterns, whose labels are the immediate emotion change the utterances caused on listeners in an open ended conversation setting. To achieve such a goal, we developed a web platform where volunteers talked to a screen-based ECA [15] inspired by the Kotaro robot [16], which responded with gibberish speech. The developed website recorded both the audio of the volunteers' speech and the video of their facial expressions while listening to the ECA's utterances. A total of 37 volunteers from 8 different countries participated, speaking a total of 14 different languages; they contributed over 734 video samples. After the conversation with Kotaro, the volunteers were asked to fill out a Likert scale questionnaire, which was optional. The questionnaire was filled out by 22 participants. Thus, this paper presents the results of the analysis performed on the audiovisual data and questionnaire responses. It is the first paper to use IPA-based gibberish speech, which allows the results to be extended to many different cultures.

We have found out that gibberish speech is not particularly engaging in a conversational setting, since most the average emotion estimates for different GS utterances had slightly negative valence and positive arousal. However, some utterances caused a positive average emotion and some positively impacted the emotional state of volunteers. Using such data, we have developed and trained a recurrent neural network-based architecture to predict the human impression of gibberish Speech.

Moreover, the obtained data set suggests that for non-Yulean GS, there is little to no correlation between the acoustic prosody parameters and the emotion change on research subjects. Moreover, similar-sounding phones seem not to be close in the learned embedding space. However, more data are necessary to strengthen both claims.

The structure of this paper is as follows: Section 1 is the Introduction; Section 2 is the Background, where the theoretical background necessary for understanding this paper is briefly explained, as well as related works; Section 3 is Materials and Methods, where the Talk to Kotaro experiment and the methods employed for analyzing the data obtained through it are presented; Section 4 is the Results, where the results of the analysis of the

audio and video recordings and the Likert scale questionnaire responses of the experiment, as well as the results of the phone embedding, are presented; Section 5 is the Discussion, where the implications of the results of the analysis performed over the experiment data are showcased; and Section 6 is Conclusions and Future Work, where the conclusions from the present work and the future work to be developed with the data obtained are presented.

2. Background

This section briefly introduces concepts necessary for understanding this work and presents related work. Its Section 2.1 explains in detail what gibberish speech is and introduces the International Phonetic Alphabet, whose symbols serve as building blocks for gibberish speech in our work. Section 2.2 explains what prosody is in the context of linguistics. Section 2.3 explains the valence–arousal emotion classification model. Section 2.4 explains the mathematical model that maps the listener's emotional response to an IPA phone–prosody pair. Section 2.5 explains the statistical bootstrapping method, which is used to obtain intervals of confidence for computed statics, making the present analysis stronger. Finally, Section 2.6 presents previous research related to the present work.

2.1. Gibberish Speech

In human–computer communication, when a given language is used for communication, it limits the set of people who can effectively understand what an ECA is trying to convey; and the meaning of words can have multiple interpretations that affect the impact on a listener. To avoid such limitations, semantically free utterances have been used to convey emotions such as anger, sadness, etc. Such utterances use musical cues such as tempo and pitch to convey emotion. For example, sounds with slower tempo, lower pitch, and little variation convey sadness, while sounds with faster tempo, high volume, and intensity can convey anger. Such cues also apply to human-like speech, allowing it to convey such emotions without conveying meaning, but still resembling a language. There are four main classifications of semantic-free speech [7]: (i) gibberish speech, SFU, which is composed of human speech sounds; (ii) paralinguistic utterances, which are composed of human non-speech sounds, such as laughs, sighs, etc.; (iii) musical utterances, which use musical sounds to convey messages and feelings; and (iv) non-linguistic utterances, which consist of beeps, whirs, and pings, among many other sounds, to communicate [7].

This work focuses on gibberish speech (GS) because it is useful for systems that do not require meaningful vocalizations to convey certain meanings, such as in human–robot interaction, video games, or animation. It can also be beneficial when users need to communicate with a technology that has little natural language processing capability, such as voice-activated devices with low processing power. The ability to convey more subtle emotions and intentions through fluctuations in pitch, rhythm, and other acoustic aspects is an advantage of using gibberish speech over other SFUs.

Such choice was also motivated by the fact that both gibberish speech and semantic speech are constructed using the same building blocks: phones. Thus, there is the possibility that the understanding of the effects of prosody and phone selection on human listeners obtained for gibberish might also extend to semantic speech.

2.2. Prosody

In linguistics, prosody is defined as the study of larger units of speech, such as syllable characteristics, intonation, stress, and rhythm [17]. Listeners can infer the emotional state of speakers from the prosody of their utterances, since someone who is excited, for example, may speak faster, louder, and at a higher pitch than usual.

The most important auditory variables in prosody are pitch (how low or how high the voice is), rate (the length of the utterances), loudness (how loud the voice is), and timbre (the quality of the sound of the voice) [17]. This paper is concerned with the first three characteristics, assuming that decreasing the quality of the audio will lead to negative reactions because it will make it harder to understand what the ECA is saying.

2.3. Valence and Arousal

The question of how many human emotions there are and how to classify them is an important problem in psychology, and thus, many classification models have been developed. One such model is Russell's two-dimensional model of valence and arousal [18], which classifies emotions in a continuous valence–arousal space. Valence represents how positive or negative an emotion is, while arousal represents how aroused a person is from relaxation to excitement [18]. The valence–arousal emotion space is defined over $\{v \in \mathbb{R} | -1 \leq v \leq 1\}$ and $\{a \in \mathbb{R} | -1 \leq a \leq 1\}$, which produces the emotion space shown in Figure 1, along with the positioning of some emotions. This model is often used because it produces a continuous emotion space rather than discrete labels, such as Paul Ekman's six or seven basic emotions [19] or Plutchik's wheel of emotions [20]. It is often used for emotion estimation from facial expressions, the same context of this work [21].

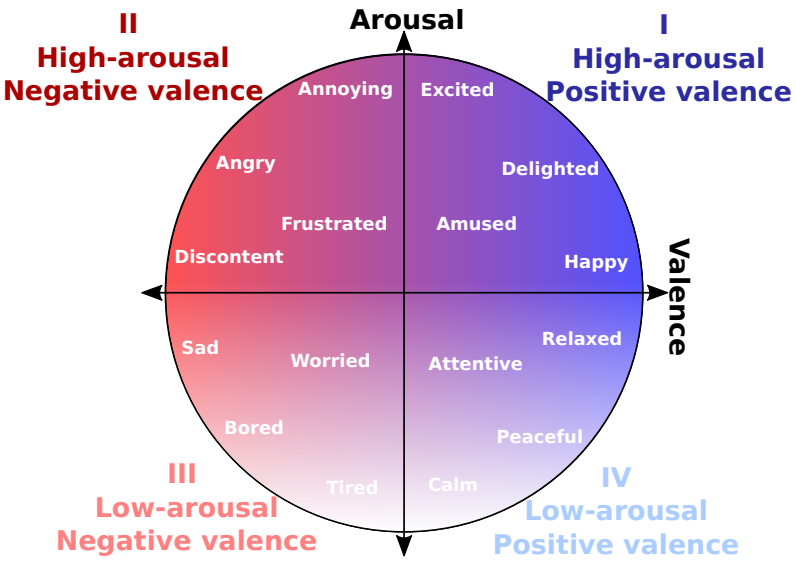

Figure 1. Russell's two-dimensional model of valence and arousal and the mapping of some emotions in it.

The emotional state of a person at a given time t is then defined as $E_t = (v_t, a_t)$.

2.4. Speech Act

Speech is an act on itself with effects on listeners. Since the speech used in this work is meaningless, the only effect it can have on listeners is emotional, and thus, this work only considers perlocutionary acts and studies their perlocutionary effects on human listeners. A speech is defined as $S(w, P)$, where w is a vector containing each phone to be spoken and P is a $||w|| \times 3$ matrix containing the prosody (volume, speed, and pitch) associated with each phone. An example of a speech is $S_{example}$:

$$S_{example} = S([\text{b,a,ɪ}], \begin{bmatrix} 100 & 130 & 45 \\ 90 & 130 & 50 \\ 95 & 140 & 50 \end{bmatrix})$$

The communication act by the ECA can be defined as $C[S(w, p)] = f(S(w, p))$, where f is a rendering function, which, in this work, represents the eSpeak speech synthesizer. Even if listening to an utterance does not lead to an action, it is expected to produce an impression $\overrightarrow{I_S}$, which is defined as $\overrightarrow{I} = \overrightarrow{\delta_E}(\delta_v, \delta_a)$, representing the change in valence and arousal caused by the speech act S. This change can be modeled as $E_{t+1} = g\{E_t, C[S(w, p)]\}$, where t is the moment before hearing S and $t + 1$ is the moment after; and g is a function representing how a listener responds to utterances. This function represents individual preferences, sensibilities, cultural background, etc., and is very difficult, if not impossible, to

model. However, with enough data, it is possible to learn listener preferences for phonetic and prosodic choices through machine learning.

2.5. Statistical Bootstrapping

Bootstrapping is a statistical technique introduced in the late 1970's that enables researchers to make data-based inferences without strict distributional assumptions for univariate and multivariate data. It involves two distributions: the underlying distribution of the data (for example, normal or binomial) and the distribution of a computed statistic (in our case, Stuart–Kendall τ_C correlation). New m data sets are formed by Monte Carlo resampling, each one containing the same number of observations n as the original data set [22]. Monte Carlo resampling is performed through randomly selecting points in the original data set and copying them into the new data set, until there are n points in the new data set.

That way, a data sample of the original data set might appear one time, multiple times, or not at all in the new data set. Such an operation is performed m times; and for each new data set created, it is necessary to perform the computed statistic operation. Now, we have a distribution of computed statistic results, from which we can obtain a confidence interval through several techniques, such as percentile [23], bias-corrected (BC) [24], bias-corrected and accelerated (BCa) [25], and approximate bootstrap confidence (ABC) [26], among others.

The authors have chosen to use the percentile method since it suffices for the performed analysis and due to its easiness of implementation. The percentile method consists of plotting the frequency histogram of the m computed statistics of the new sampled data sets, and the 95% confidence interval will consist of the values between the 2.5th and 97.5th percentiles. These percentiles represent the lower and upper bounds of the confidence interval, respectively. The resulting confidence interval yields a range of values within which the true parameter value is likely to fall with a certain level of confidence, in our case, 95%.

The idea behind such process is to estimate the sampling distribution of a statistic by repeatedly sampling with replacement from the observed data. The ultimate goal is to make inferences about a population parameter or the distribution of a statistic even when you have a limited amount of data, as long as the distribution of the limited data set somewhat resembles the real distribution of the real-world variable.

2.6. Literature Review

Research on semantic-free utterances is not new, and many different types of semantic-free utterances, such as [8], have been performed, but research on gibberish speech still needs more development. Among the works that used gibberish speech, all of them were based on existing languages, such as Japanese [27] and Dutch and English [13,28,29]. Thus, this work is novel in the sense that it presents a language agnostic gibberish speech and analyzes its emotional impact on listeners. It also investigates the effects of prosodic acoustic characteristics of gibberish speech on human impression, but unlike the investigation performed in [30], which investigated which prosodic characteristics of gibberish speech better fit different robot morphologies according to adult's expectations, it investigates how such characteristics affect adult human impression for a fixed ECA appearance.

In [29], the authors developed a gibberish generation system based on swapping the vowel nucleus of Dutch and English words to turn them into gibberish, but to avoid ending up with weird sounding words, the authors developed a weighted swapping mechanism according to the probability distribution of each vowel core in English and Dutch. The gibberish generation algorithm developed for the "Talk to Kotaro" experiment deliberately allowed for the generation of utterances that did not follow any yule-like phone distribution [31], because if a distribution were chosen, it might cause alienation to speakers of other language families. Moreover, by not following the usual rules, we can study the effects of the violation of such principle on listeners.

In order to analyze what emotions are generated by gibberish speech, the authors of [32] conducted child–robot interaction experiments using an NAO robot equipped with control and behavior modules. The experiments were divided into two trials: one in which the experimental setup was designed to elicit natural emotions in children, and the second in which the setup was designed to analyze children's perception and response to the gibberish speech of the NAO robot. Similarly, [33] investigates the perceived emotions of Spanish synthetic expressive voices by participants of four Asian nations (Japan, South Korea, Vietnam, and Malaysia), which shows that non-verbal cues are very important in the perception of emotion, but, again, the work does not focus on how the listeners felt.

In [30], the acoustic prosody features were chosen in a Wizard of Oz setup, but several techniques for automatic prosody generation, at least for semantic speech, have been developed. Such techniques can be rule-based [11,34,35] or neural network-based [36–40].

In general, rule-based approaches have been superseded by neural prosody selection because manually creating rules to generate appropriate prosody from every possible case is an impossible task. The problem with neural prosody generation is that it depends on existing data from which appropriate prosody for the speech content can be learned; this is not possible for gibberish speech since, by definition, no one speaks gibberish, and thus, the data are scarce and artificially generated, as in [13]. Thus, this work provides novelty in the sense that it has generated a small data set that can be used to learn appropriate prosody for IPA-based input text. Moreover, another problem of most neural prosody generation work is that they are tightly coupled to speech synthesis, whereas the proposed architecture is speech-synthesizer-independent.

Regarding the emotional evaluation of prosodic speech, again, most of the works had semantic speech as their focus [11,12,41–43] and had research participants evaluate their perception of what emotion the generated speech conveyed, rather than how it affected their emotional state. In addition, most of the evaluation was conducted through subjective post-listening evaluation [43], rather than measuring the immediate response of the participants through their facial expressions and body language, for example through EEG [44]. An exception to these constraints is [32], where they have analyzed the emotion caused by gibberish speech on children by analyzing the facial expressions and bodily language displayed on video samples.

All of the automatic prosody generation research was conducted for the domain of semantic speech, and most of it focused on learning prosodic patterns from pre-existing audio recordings. Moreover, research studies that develop systems for emotional speech generation have only verified the emotion which listeners perceive on the generated speech, not on how the research subjects themselves felt when listening to the obtained speech, seeking to be perceived as natural speech, such as [36], which deals with voice only and [37], which also deals with the visual components.

3. Materials and Methods

3.1. Talk to Kotaro: A Web Crowdsourcing Experiment

To create a data set of how people respond to different phonetic and prosodic choices in gibberish speech, the web-based crowdsourcing experiment "Talk to Kotaro" was conducted between 1 October 2021 to 31 March 2023. The experiment was approved by Tokyo University of Agriculture and Technology Ethics Committee (approval number 210801-0321 and experiment extension request approval number 220306-0321). All participants had to read an online consent form and the experiment instructions and had to click a consent button, which was deemed an acceptable means of obtaining consent by University of Agriculture and Technology Ethics Committee.

In the "Talk to Kotaro" experiment, research volunteers talked as long as they wanted to a cartoon avatar of the robot Kotaro, which responded with IPA-based gibberish speech. The website, developed using JavaScript (client-side) and Python (server-side), required volunteers to register a profile with personal information, which was used to evaluate the impact of gender, age, nationality, native language, etc., on the average response of a given

group of people. In addition, the platform recorded the audio of what the volunteers said to assess their emotional state before listening to Kotaro's response. Video of the volunteers' facial expressions as they listened to the GS response was recorded. To avoid the need for voice activity detection (VAD), the platform required volunteers to press a button (bottom left button shown in Figure 2) before speaking and press it again after they finished speaking, allowing Kotaro to respond. This cycle would repeat as long as a volunteer wanted; the most prolific volunteer contributed 201 conversations, about 23.4% of all the data in the experiment. When volunteers felt they had talked enough, they could either close the experiment web page tab or click the logout button (top left button in Figure 2). If they chose to click the logout button, they were redirected to a 10-question Likert scale questionnaire about their experience. Answering the questionnaire was optional, and 22 of the 37 participants chose to do so.

The experiment was designed with the goal of obtaining the immediate emotional changed caused by the GS patterns of the robot avatar in a open-ended conversational setting, trying not to introduce many distraction factors, such as having a very expressive conversational agent or a lively background of the web page. The experiment was performed during the height of COVID-19 pandemic restriction, and thus, it had to be held online, since having research subjects in the laboratory without wearing face masks would be impossible. However, such exceptional measures also had the benefit of allowing research subjects from all over the world to participate in the experiment. The decision of having a screen-based embodied conversational agent was taken in order to simulate the experience of talking to a social robot without requiring any special hardware by volunteers, such as VR or AR googles, trading off some of the realism of the task for ease of participation.

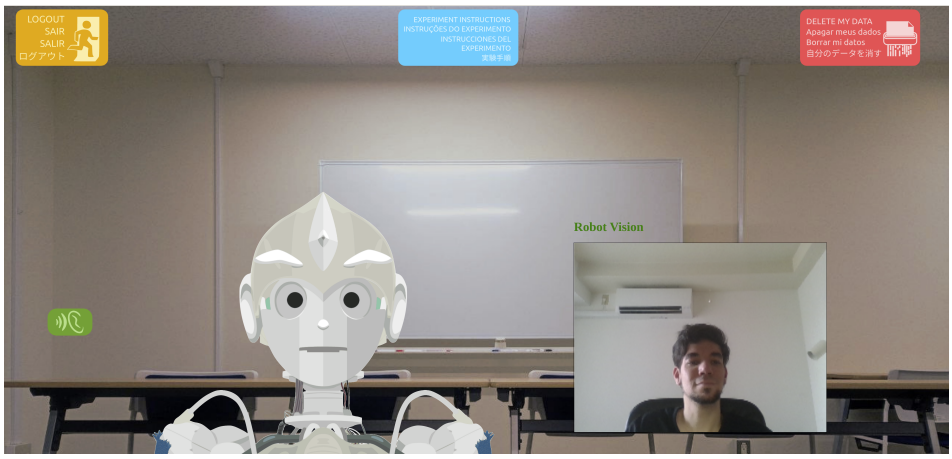

Figure 2. Talk to Kotaro experiment screen, which shows ECA Kotaro, the video feed from the camera of volunteers (one of the authors, in this case) and the turn-taking button (green).

Kotaro's IPA-based gibberish speech was transformed into voice by using the eSpeak [45] speech synthesizer, which was chosen because it is open-source, can receive ASCII-IPA input, and allows for control of the prosody of generated speech. Algorithm 1, described in Section 3.1.1, was used to select the phones to be used in Kotaro's speech. As for the prosody, the three chosen parameters—speed, pitch, and volume—were randomly chosen between 80–450 words per minute (speed), 10–200% (volume), and 0–99 (arbitrary unit, pitch). Some participants reported a feeling of alienation when the ECA suddenly changed its voice pitch, making them feel like they were not talking to the same person.

3.1.1. IPA-Based Gibberish Speech Generation

To create Kotaro's gibberish, an algorithm, originally described in [14], draws vowels and consonants from the IPA table to randomly generate Kotaro's responses. The International Phonetic Alphabet is a phonetic notation system created by the International Phonetic Association in the 19th century to provide a standardized way of representing

speech sounds in different languages in written form [46]. It can represent various aspects of the lexical and prosodic sounds of human speech; phones, intonation, and pauses. Other non-speech sounds, such as clicks, grits, and lisping, are represented by an extended set of symbols. There are two basic sets of symbols: letters and diacritics. Algorithm 1 describes how Kotaro's utterances were generated during the experiment.

Algorithm 1 IPA Giberish Speech generation algorithm

procedure GENERATE GIBBERISH
  $max_{iter} \leftarrow choice([1, \ldots, 10])$
  $counter_{iter} \leftarrow 0$ ▷ iteration counter.
  $utterance \leftarrow$ "" ▷ gibberish speech utteranc, starts empty.
  IPA_v ▷ list of all IPA vowels.
  IPA_c ▷ list of all IPA consonants.
  IPA_o ▷ list of all IPA other symbols.
  while $counter_{iter} < max_{iter}$ **do**
    $chunk \leftarrow choice(choice([IPA_v, IPA_c]))$
    $chunk \leftarrow chunk + choice(choice([IPA_v, IPA_c, IPA_o, ""]))$
    if $len(chunk) > 1$ **then**
      if $chunk[0] \in IPA_c \wedge chunk[1] \in IPA_c$ **then**
        $chunk \leftarrow chunk + choice(IPA_v)$
      else if $chunk[1] \in IPA_o$ **then**
        $chunk \leftarrow choice(choice([IPA_v, IPA_c]))$
      else
        $chunk \leftarrow chunk + choice(choice([IPA_v, IPA_c, [""], [""], [""], [""], [""], [""]]))$
    $utterance \leftarrow utterance + chunk$
    $counter_{iter} \leftarrow counter_{iter} + 1$
  return *utterance*

To better understand Algorithm 1, it is necessary to define the function $choice(l)$, which randomly chooses an element belonging to a list l. At the beginning of the routine, the number of iterations for generating the utterance is randomly chosen between 1 and 10, an arbitrary maximum chosen by the researchers to avoid very long utterances and to avoid very long delays between a volunteer finishing speaking and Kotaro responding. The utterance starts as an empty string, which obtains chunks of one or more IPA symbols in each iteration. There is a 50% chance that a chunk will start as a vowel and a 50% chance that it will be a consonant symbol. All symbols in each list have the same chance of being chosen by the *choice* function. After that, there is a 75% chance that a second symbol will be added (vowels, consonants, and other symbols all have a 25% chance), and a 25% chance that nothing else will be added to the chunk. If a second symbol is chosen, and both the first and second are consonants, a third symbol from the vowel list is added. If the second symbol chosen is another symbol, there is a 50% chance that a vowel will be added, and a 50% chance that a consonant will be added instead. Otherwise, there is a 12.5% chance that a vowel will be added, and a 12.5% chance that a consonant will be added. The remaining probability is that nothing will be added. At the end of the iteration, the chunk is added to the utterance and the iteration counter is incremented.

Note that when receiving ASCII-IPA input, eSpeak will skip unpronounceable sounds if there is a space between each chunk, i.e., it will just speak the next one. An example of an utterance generated by this algorithm is [ionuɪmˈə]. It is necessary to note that the chosen algorithm leads to a non-Yule distribution of phones, unlike real languages [31], but since there is no research on human impression towards words generated without following said distribution, it is worth investigating the effect of such utterances [47].

The gibberish speech of EMOGIB [13] and the Hanamogera [27] speech is that they both use traditional syllabic structures, that is, syllables are composed by an onset consonant, a nucleic vowel, and a coda (final consonant cluster). Since EMOGIB is inspired by Dutch and English languages, it also contains syllabic consonants, that is, consonants that

are the nucleus of a syllable. Hanamogera, which is inspired by the Japanese language, does not contain syllabic consonants. However, the GS yielded by Algorithm 1 can generate words that are a string of syllabic consonants, sounding like no existing language, whose effect on listeners also never have been investigated before.

3.1.2. Likert Scale Questionnaire

Likert scale questionnaires are a tool for measuring overall attitudes toward a topic. They consist of prompts, statements about the topic being studied, to which respondents choose their level of agreement, ranging from strongly agree to strongly disagree. The number of prompts and possible responses is not predetermined; researchers must use as many as they need, keeping in mind that increased precision may be offset by increased burden on research subjects. However, the most traditionally used scales have either five or seven responses. It is also possible to remove the neutral option, that is, to have a pair of possible levels of agreement, to prevent respondents from over-relying on neutral responses as a socially acceptable stance. This paper uses the traditional 5-point scale and 10-point prompts to avoid tiring respondents.

The Likert scale questionnaire used in the Talk to Kotaro experiment uses a classic five-point format, i.e., respondents can choose their level of agreement with a prompt between 1—strongly disagree, 2—disagree, 3—neutral, 4—agree, and 5—strongly agree.

This decision was made so that respondents would not have to think too much while answering a questionnaire that they could simply exit by closing a tab on their web browser. However, not making the questionnaire mandatory was a design choice to prevent participants who were already tired from the experiment from randomly clicking through the answers to end their participation as quickly as possible. While this risk could not be completely avoided, as participants completed the questionnaire unsupervised, it was a way to reduce this possibility.

The questionnaire was designed to measure volunteers' enjoyment of the statements Kotaro responded to them with, and to measure what factors were most relevant to that impression. The prompts shown are as follows:

(P_1) Talking with the robot avatar was interesting;
(P_2) Variation of the speech characteristics made conversation more natural;
(P_3) Some randomly generated words are less pleasant than others;
(P_4) Some speech characteristics, such as speed, loudness or pitch influence more than others;
(P_5) Different random words didn't have an impact on your enjoyment;
(P_6) You felt that the robot was answering your speech accordingly;
(P_7) Longer phrases were more interesting;
(P_8) The turn-based conversation felt unnatural;
(P_9) Foreign sounding phones were more interesting;
(P_{10}) The robot seemed to be intelligent.

3.2. Neural Network Architectures for Emotion Analysis

The most important data provided by the volunteers of the "Talk to Kotaro" experiment consisted of video recordings of the facial expressions of participants while listening to the gibberish speech responses and audio recordings of what participants told the conversational agent. The idea behind recording both audio and video was to help gauge the emotional state of volunteers before Kotaro's answer, from the audio, and understand how the emotions of the volunteers have changed from the facial expressions displayed in the recorded videos. The emotion change caused by the Gibberish speech utterance was, then, used as a label for the developed data set.

To achieve such a goal, three different artificial neural network architectures were employed. Two neural architectures, described in Section 3.2.1, are used for estimating the emotion of volunteers from their facial expressions, and one, described in Section 3.2.2, is used for sentiment classification of the audio samples.

After possessing labels in terms of valence and arousal for the reaction displayed by volunteers after listening to distinct phones and gibberish speech patterns, we investigate the relationship between the aforementioned parameters of Kotaro's gibberish speech and the impression of volunteers through neural network architectures described in Section 3.2.3.

3.2.1. Emotion Estimation from Video

In order to obtain the impression created on the volunteers by the GS utterances, two different neural network architectures, VGG-16 and ResNet18 (inspired by the architecture proposed in [48]), were used to estimate the volunteers' valence and arousal, respectively, from the videos of their facial expressions. This hybrid system was chosen because VGG-16 performed better than ResNet18 for valence, while ResNet18 performed better for arousal. Both networks were trained on the AffecNet data set [49]. This method of engagement and preference estimation was chosen because it is not an invasive method, does not require very expensive additional hardware for volunteers (given most laptop computers, tablet computers and smartphones have front cameras nowadays), and it does not pause the experiment, allowing one to capture the immediate emotional change caused by the speech sound. The decision for capturing the immediate reaction stems from previous research findings that the candid reaction of research subjects differs substantially from their opinion after being given some time to think and rationalize their own feelings and opinions about an experiment [50,51]. However, since it is also important to know the attitude of volunteers towards Kotaro's gibberish speech, towards Kotaro and the experiment itself after having some time to think, this approach is coupled with the Likert scale questionnaire proposed in Section 3.1.2.

However, since the aforementioned neural networks estimate human emotions from still images, and the collected data consist of video samples, it was necessary to choose a metric capable of representing the impact of the gibberish speech on the listener. Thus, it is necessary to obtain the initial emotional state E_t of the subject and the emotional state E_{t+1} after listening to the utterance.

The chosen metric is then the difference between the emotion estimation from the initial (just before Kotaro starts speaking) and the last frames of each video sample. This metric is called $\overrightarrow{\delta_E} = (\delta_v, \delta_a)$, where δ_v, δ_a is the change in displayed valence and arousal.

It is possible to see how a few utterances (randomly selected from the data set) have affected the subjects in the valence–arousal space shown in Figure 3.

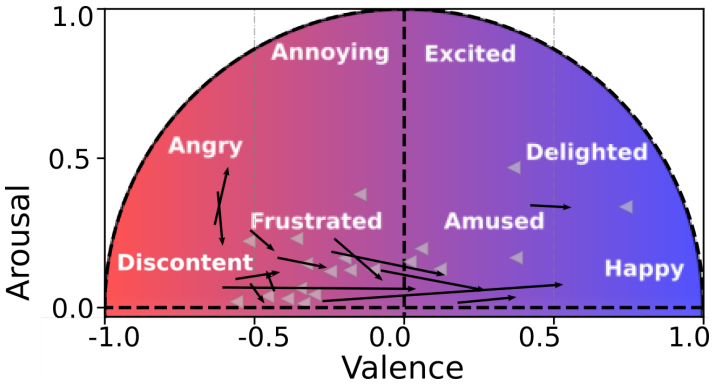

Figure 3. $\overrightarrow{\delta_E}$ represented in the valence–arousal emotion space, where arrows indicate the valence–arousal change and grey triangles denote an utterance that caused no visible emotional impression.

3.2.2. Sentiment Analysis of Recorded Speech

To make the predictions of the volunteers' initial emotional state just before listening to Kotaro's responses more accurate, the Talk to Kotaro web platform recorded what the volunteers said to Kotaro, which allowed us to perform sentiment analysis on the recorded audio samples. However, the initial problem is that the authors could not find a data

set for human speech whose sentiment labels were in terms of valence and arousal, only categorical labels. The chosen data sets were the audio data sets TESS [52], RAVDESS [53] and SAVEE [54], whose samples were labeled with one of the following seven emotions: anger, disgust, fear, happiness, neutral, sadness, and surprise. We extracted the main features of the available audio data using mel-frequency cepstral coefficients (hereafter called MFCC) and used the obtained information to train an LSTM-based neural network, whose architecture is shown in Figure 4, which was then used to verify the accuracy of the participants' initial emotional state prediction, at least in qualitative terms, since the labels are categorical.

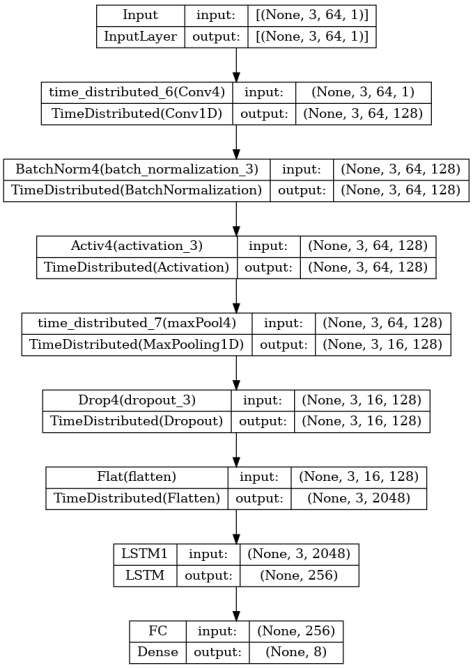

Figure 4. Architecture of the LSTM-based neural network used for analyzing the sentiment of the voice recordings of participants.

3.2.3. Gibberish Speech Impression Prediction System

To better understand the effect of phones on participants' impressions, a neural network called GRU_{phones} was built, consisting of an embedding layer with 64 outputs, followed by a bidirectional gated recurrent unit (GRU) layer with 512 units, which is then followed by four fully connected layers with 512, 256, 128, and 1 neurons. All connected layers use ReLu as activation function, except the last one, which is linear (architecture shown in Figure 5a). The proposed neural network was able to learn an embedding for each of the 71 IPA symbols used by Kotaro (some symbols were not used because not enough utterances were generated). The neural network was trained with the data from the experiment, taking the tokenized IPA symbols as input and outputting the predicted valence or arousal.

Besides the analysis performed by Stuart–Kendall's τ_C correlation coefficient, another way to learn the correlation between the acoustic prosodic parameters is to use a neural network that receives as input a vector containing the speed, volume, and pitch of a given utterance and predicts the subjects' impression. However, only using the prosody information did not yield good results, and by adding the profile information encoded together with the prosody parameters into a 1×80 vector, it was used as input for a neural network called $MLP_{profile+prosody}$, which consists of an input layer of 80 neurons connected to three hidden layers of 128, 128, and 64 neurons each, and ReLu as the activation function (architecture shown in Figure 5b). The output layer is a single neuron, and thus, two copies of $MLP_{profile+prosody}$ were trained, one for predicting arousal and another for valence.

Since gibberish speech utterances C, it is necessary to take both aspects into account to make accurate predictions, and thus, we combined both neural networks by averaging their outputs. Other architectures were tested for combining $MLP_{profile+prosody}$ and GRU_{phones}, but the results were not as accurate the ones obtained by averaging the outputs of both pre-trained models. The resulting model is called the Gibberish Speech Impression Prediction System, hereby referred to as $GSIP$.

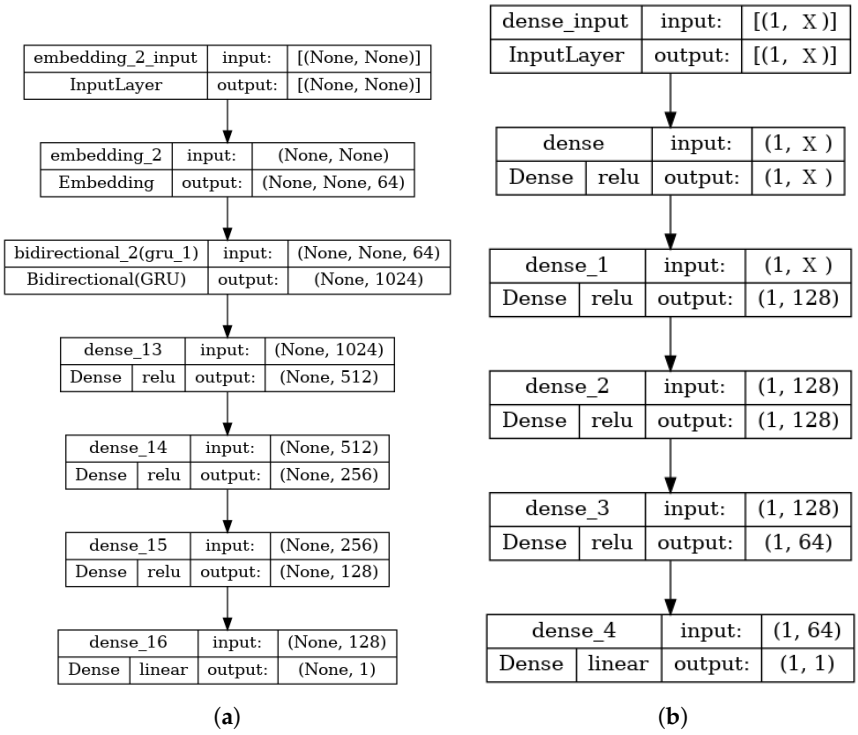

Figure 5. Neural networks used for impression prediction in this work. (**a**) Architecture of the bi-directional GRU neural network GRU_{phones} for generating a phone-embedding matrix; (**b**) architecture of neural network $MLP_{profile+prosody}$ and its variations, where X represents the number oc columns of the input vector.

4. Results

In this section, we present the data obtained from the "Talk to Kotaro" experiment in fully anonymized form and perform the necessary analysis to verify the influence of phone and prosody choices in gibberish speech. The audio and video recordings cannot be shared because that would violate the privacy of the volunteers, a condition set by Tokyo University of Agriculture and Technology's Ethics Committee. However, a fully anonymized version of the data set, containing the phones, prosodic parameters of each generated gibberish speech, and results of the emotional analysis performed on audio and video data, together with its partitions into data set without outliers, training, validation and test data sets are available in the supplementary files in the present paper.

Section 4.1 presents the profile information of the participants of the experiment, while Section 4.2 presents the results of the emotion analysis performed on the participants' video and the investigation of the correlation between the prosody parameters and the impression on the volunteers. To further improve the emotion estimation from the volunteers' facial expressions before listening to Kotaro's utterances, we performed a sentiment analysis of the participants' recorded speech, which is presented in Section 4.3. Section 4.4 presents and discusses the results of the phone embedding matrices obtained from the experimental data; and the results of the Likert scale questionnaire are discussed in Section 4.6.

4.1. Profile of Participants Breakdown

This subsection breaks down the information about the participants of the Talk to Kotaro experiment. Profile information of the volunteers was stored to try to determine how the prosody changes affected each nationality, speakers of certain languages, age groups, etc. The stored information included ID, password, age, gender, country/region of origin, native language, other languages spoken by the volunteer, and if the volunteer lives or has lived abroad (write where and years lived abroad).

The initial goal was to try to find a cross-cultural baseline for human impression for different prosody parameters, a point that will be described in more detail in the Section 4.2. The effects of phone choice on human impression are discussed in Section 4.2.

Countries with participants are shown in Table 1, along with the number of speakers of each language. Initially, 61 participants from 16 countries speaking 17 languages registered, but after removing those who contributed with no data, or contributed only with unusable data (e.g., participated in very dark environments, wore face masks, etc.), only 37 were left. That fact showcases one of the greatest weaknesses of web-based crowdsourcing: data quality varies a lot because participants have different hardware and environment conditions, and might misinterpret instructions without any chance for correction.

Out of the remaining 37 participants, 23 were male and 14 were female. The mean age of the participants was 27.46 years, with a standard deviation of 9.39 years, a median of 25 years, and a mode of 21 years. The youngest participant was 18 years old and the oldest was 55 years old.

Table 1. Breakdown of the cultural background of the participants.

Country/Region of Origin	Male	Female	All	Mother Language	Male	Female	All	Total Speakers	Male	Female	All
Japan	9	10	19	Japanese	9	11	20	English	19	15	34
Brazil	6	1	7	Portuguese (Brazil)	6	1	7	Japanese	14	11	25
Malaysia	2	0	2	Mandarin	3	0	3	Portuguese (Brazil)	6	1	7
China	1	0	1	Cantonese	0	1	1	Mandarin	3	0	3
Hong Kong (China)	1	0	1	English	0	1	1	Malaysian	2	0	2
India	0	1	1	Marathi	0	1	1	Arabic	1	0	1
Peru	1	0	1	Spanish	1	0	1	Cantonese	1	0	1
USA	0	1	1	Arabic	1	0	1	Spanish	1	0	1
Bangladesh	1	0	1	Sinhala	1	0	1	Sanskrit	0	1	1
Egypt	1	0	1	Bengali	1	0	1	Korean	0	1	1
Sri Lanka	1	0	1					Sinhala	1	0	1
Undisclosed	0	1	1					Bengali	1	0	1
								Hindi	0	1	1
								Marathi	0	1	1

4.2. Impression Estimation from Video and Prosody Correlation

In this subsection, the videos of the volunteers' facial expressions are analyzed and the impression $\overrightarrow{I_S}$, the immediate emotional response to the speech act S, is obtained by the vector $\overrightarrow{\delta_E} = (\delta_v, \delta_a)$, which is obtained by subtracting the estimated emotional state of the initial and final frames of the video. In this way, a data set is generated that associates the speech acts and the human impression, allowing us to verify if there is a correlation between the prosody parameters and the impression $\overrightarrow{I_S}$, and to use machine learning to obtain an embedding matrix for the IPA phones. This is achieved by using the VGG-16 and ResNet-18 neural networks described in Section 3.2.

Plotting all obtained $\overrightarrow{\delta_{E,S}} = (\delta_{a,S}, \delta_{v,S})$ vectors, the impression caused by each speech S in the valence–arousal space yields Figure 6. However, since there are 734 vectors, it is difficult to visualize the results in valence–arousal space. Plotting the histograms of the results of the analysis, shown in the left and middle histograms of Figure 7, shows the results of the analysis performed on the video samples of the participants' reactions to each utterance spoken by Kotaro during the entire experiment. It can be seen that many utterances had little effect on the participants' valence or arousal. However, if we calculate

the norm of the $\overrightarrow{\delta_E}$ vector for each speech act, we obtain the right histogram in Figure 7, which shows that many utterances caused little to no change in the emotional state of the listeners, but most still made an impression. The set of all $||\overrightarrow{\delta_E}||$ has a mean of 0.124 and a standard deviation of 0.135.

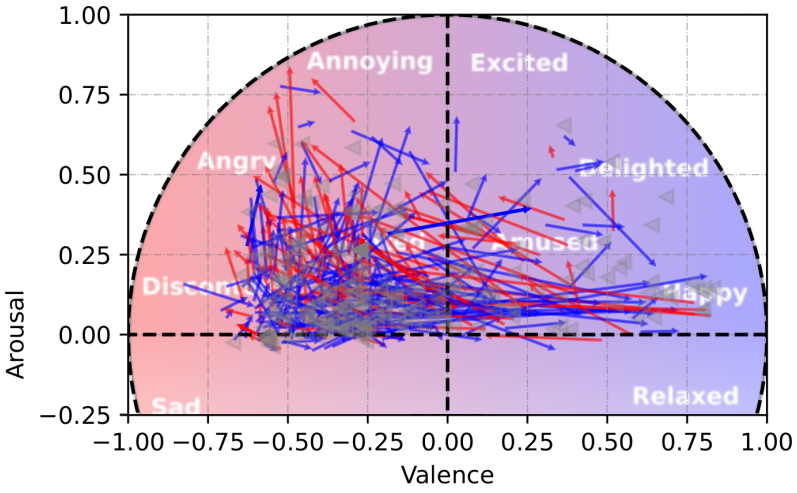

Figure 6. Emotional state changes caused by every utterance in the "Talk to Kotaro" experiment in the valence–arousal space, where a blue arrow denotes a positive change in valence, a red one denotes negative valence change, and a grey triangle denotes no visible emotional change.

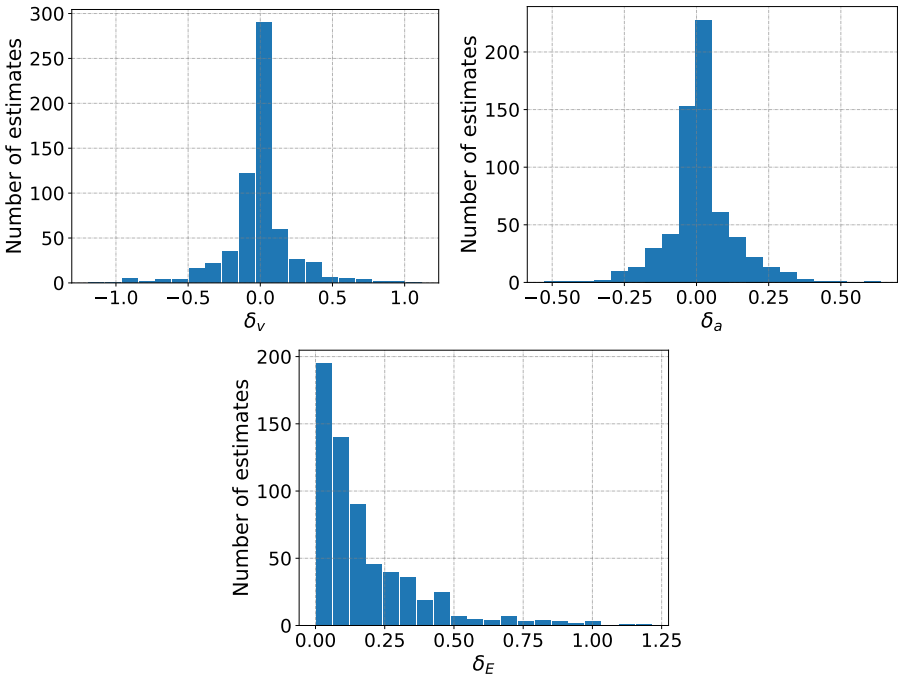

Figure 7. Histograms of δ_v (**top left**) and δ_a (**top right**) and $||\overrightarrow{\delta_E}||$ (**bottom**) for every utterance generated in the "Talk to Kotaro" experiment.

Since Russell's two-dimensional model of valence and arousal is defined over $\{v \in \mathbb{R} | -1 \leq v \leq 1\}$ and $\{a \in \mathbb{R} | -1 \leq a \leq 1\}$, where v and a are valence and arousal, respectively, the impression space is defined over $\{\delta_v \in \mathbb{R} | -2 \leq \delta_v \leq 2\}$ and $\{\delta_a \in \mathbb{R} | -2 \leq \delta_a \leq 2\}$. Thus, we can obtain another representation for all the impressions caused by Kotaro's utterances, shown in Figure 8, where outlier impressions are highlighted in red. Since the obtained impressions consist of two variables, we used the Mahalanobis distance

metric [55], which measures the distance between a point and a distribution, to determine which emotion changes were outliers, with a Mahalanobis distance threshold of 3.

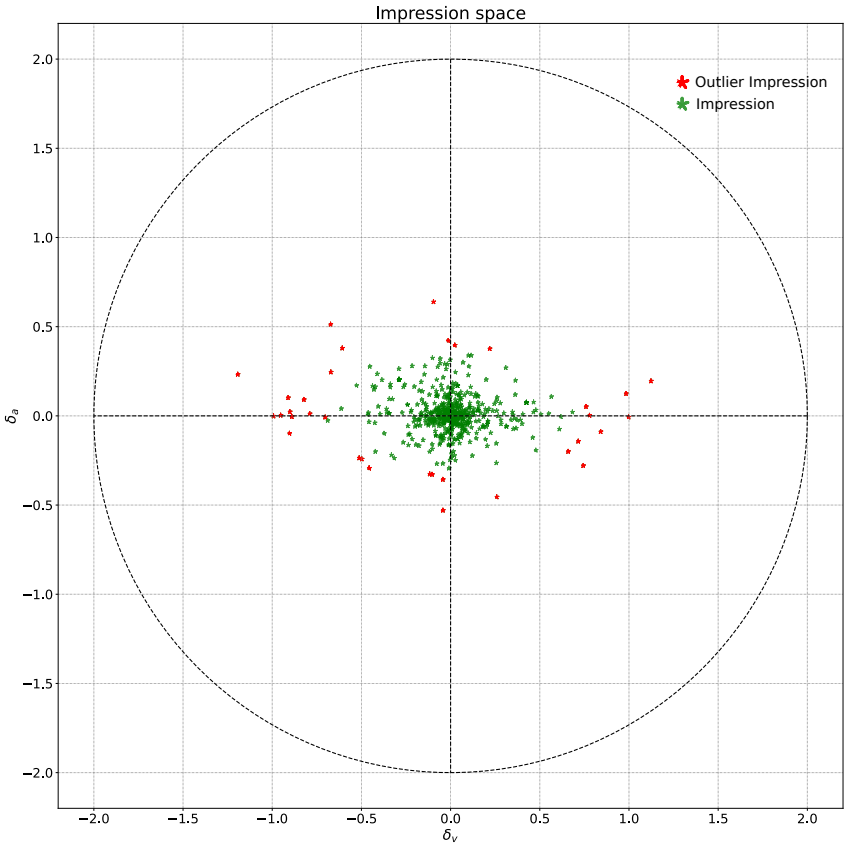

Figure 8. Every emotion change $\overrightarrow{\delta_E}$ in the data set represented in the impression space.

From the emotional estimates obtained from each frame of every interaction with Kotaro, we have seen that the maximum valence shown was of 0.831, while the lowest valence was of -0.823, while the average was -0.172, with a standard deviation of 0.291. For arousal, the highest estimate was of 0.815 and the lowest estimate was of -0.061. The average arousal was of 0.176 with a standard deviation of 0.134.

Another metric that can be explored, for the video of a volunteer listening to a given GS utterance, is averaging the emotion estimates for each frame. This way, we can obtain an overall feeling of the emotion elicited by the interaction. Calculating such a metric for every video sample and averaging the average emotion, we obtained an average of all average emotion estimates $E_{avg_avg} = (-0.248, 0.161)$, with standard deviations of 0.293 for valence and 0.137 for arousal. The lowest and highest valence averages for each interaction were -0.693 and 0.831, respectively. The lowest and highest arousal averages were -0.051 and 0.767, respectively. Out of all 734 video samples, 130 video samples had a non-negative average valence and 695 had non-negative average arousal; and 37 video samples had both non-negative valence and arousal averages. Such results show that the majority of non-Yulean gibberish speech did generate moderately negative feelings on listeners, but still, few interactions had a positive average valence.

Considering participants that had more than a single exchange with Kotaro, it is possible to analyze how their emotional state changed along the overall interaction. Out of the 37 participants, 33 had multiple exchanges and 7 had interactions across multiple days. We used linear regression to detect the tendency of the evolution of the average valence of each interaction within the the same day (a participation session) and across multiple days, for the volunteers who participate multiple days (multiple sessions). Out of the 65 participation sessions, volunteers had their average valence decrease in 30 sessions, while

in the remaining 35, the average valence of the interactions increased. For arousal, out of the 65 sessions, only in 28 could we see the valence increasing.

For volunteers who participated in multiple sessions across distinct days, the average valence decreased across different sessions for four volunteers, while it increased only for three of them. Arousal, on the other hand, increased only for two volunteers across multiple sessions, decreasing for the remaining five.

The results of such analysis can be seen in Figure 9, where the average valences of the different sessions are represented in different colors and the line resulting from the linear regression for each session has the same color as the points. The longest line represent the changes across multiple sessions. Some volunteers had their emotions improve, while other had their emotional state deteriorate while listening to the GS utterances. For the top left image, we can see a result were the line fits the data very well, but for most volunteers, that is not the case, showing that continued interactions are not a good predictor of how well listeners will react to the different. Especially when looking at the bottom right graph of Figure 9, where we can see that in the first session there is a tendency of improving valence, but in the next session, there is a strong decrease in valence as the volunteer listened to the GS utterances.

For arousal, the results are similar, but since it has lower variance when compared to valence, linear fitting describes the evolution of the emotional state of participants during a session, as one can see in Figure 10. However, as it is possible to see in the bottom right figure, that is not the case for every volunteer. We have calculated the average mean squared error between the predicted and the actual impression for average arousal of every session, obtaining a value of 0.006, while the same metric for valence is of 0.023.

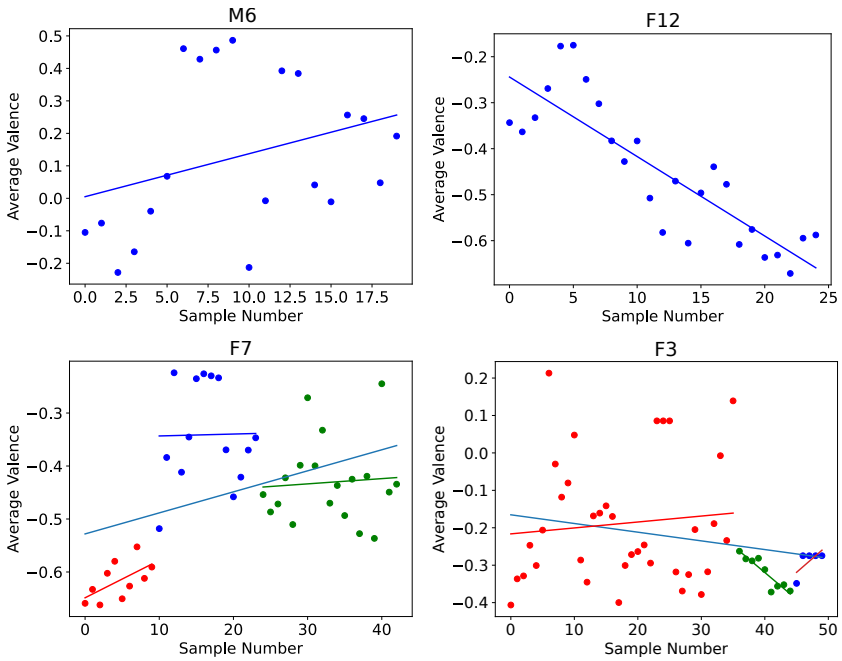

Figure 9. Scatter plots of the average valence of each time volunteers M6, F3, F7, and F12 listened to a GS utterance and the results of the linear regression for each session and across multiple sessions. Points with the same color were obtained in the same session, and the line for that session shares the color with the points.

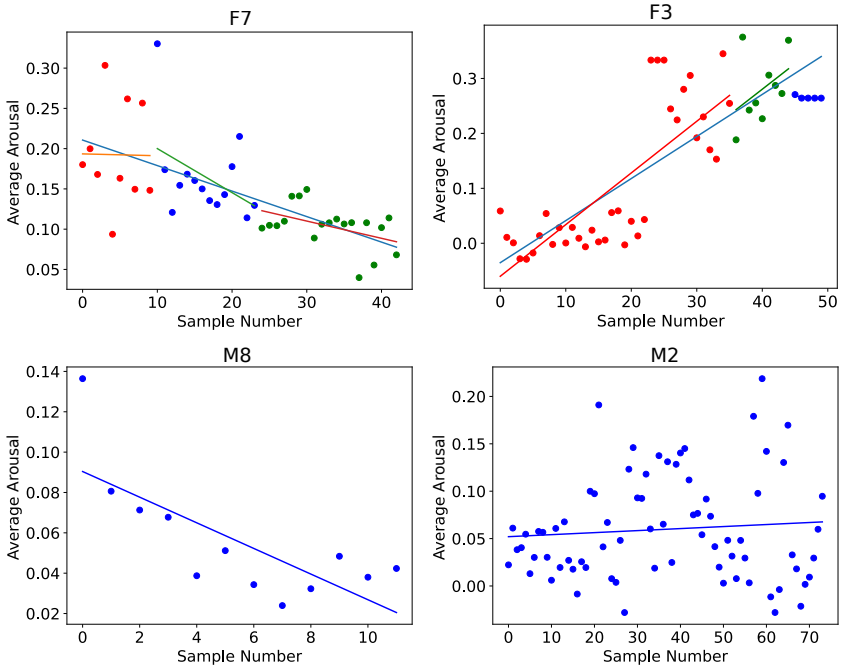

Figure 10. Scatter plots of the average arousal of each time volunteers M2, M8, F3, and F7 listened to a GS utterance and the results of the linear regression for each session and across multiple sessions. Points with the same color were obtained in the same session, and the line for that session shares the color with the points.

Emotion State Change Estimate Error

In the context of the present work, the error of the estimate of the emotion change caused by gibberish speech $S(w, P)$ is defined as the norm of the difference between the actual emotional state change $\overrightarrow{\delta_{E_{S,A}}}$ and the predicted emotional state change $\overrightarrow{\delta_{E_{S,P}}}$, that is,

$$EE_S = ||\overrightarrow{\delta_{E_{S,A}}} - \overrightarrow{\delta_{E_{S,P}}}|| = \sqrt{(\delta_{v_{a,S}} - \delta_{v_{p,S}})^2 + (\delta_{a_{a,S}} - \delta_{a_{p,S}})^2}$$

Two different neural network architectures, VGG-16 and ResNet18, were used for estimating arousal and valence of volunteers from their facial expressions, respectively. However, since the aforementioned neural networks estimate human emotions from still image frames and the collected data consist of video samples, it was necessary to choose a metric that was capable of representing the impact the gibberish speech had on the listener. This way, it is necessary to obtain the initial emotional state E_t of the volunteer and the emotional state E_{t+1} after listening to the utterance.

The chosen metric is then the difference between the emotion estimation from the initial (just before Kotaro starts speaking) and the last frames of each video sample. This metric is referred to as $\overrightarrow{\delta_E} = (\delta_v, \delta_a)$, where δ_v, δ_a is the change in the displayed valence and arousal. It is possible to see how a given utterance has affected research subjects in the valence–arousal space shown in Figure 6. If the valence improved, the vector is shown as blue; otherwise, as red.

The original research hypothesis during the development of the Talk to Kotaro platform was that prosodic choice is the most important factor in generating emotional responses in listeners; since gibberish has no meaning, it was expected that volunteers would respond according to prosodic features. Furthermore, it was expected that there would be a cross-cultural preference for certain prosodic parameters, similar to the Bouba–Kiki effect [56]. To test this hypothesis, it is necessary to compute the correlation between the prosody parameters and δ_v and δ_a. This analysis was performed pairwise using Stuart–Kendall's τ_C correlation coefficient for each participant, all male volunteers, all female

volunteers, all Japanese nationals, and all Brazilian nationals; their correlation matrices are shown in Figure 11. The correlation coefficient was also calculated for other demographics, but for the sake of brevity, the matrices are not shown.

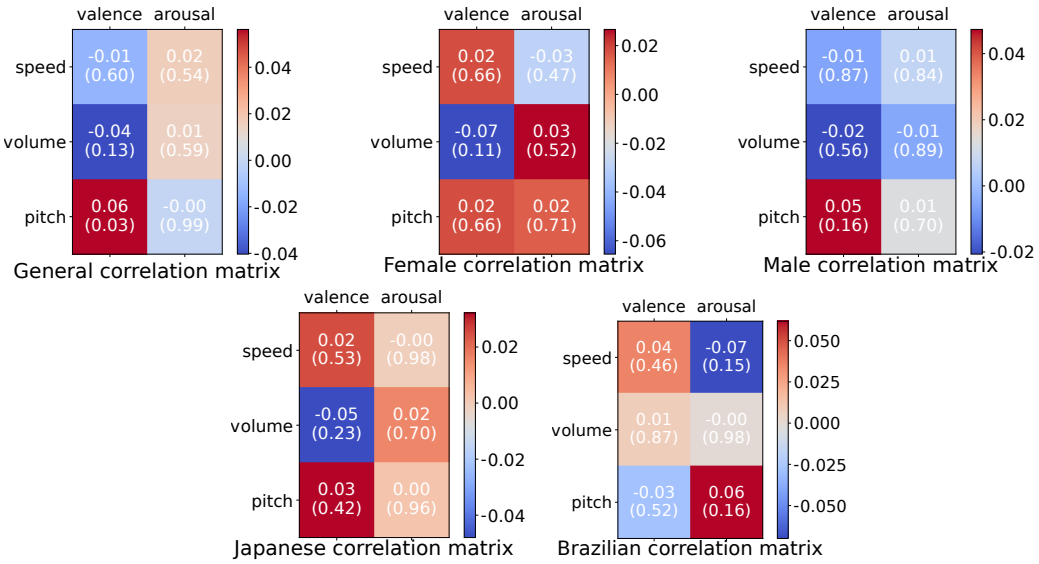

Figure 11. Pairwise Stuart-Kendall's correlation coefficient matrices, where the top number of a cell indicates the coefficient and the number in parentheses indicates the related *p*-value.

It is very clear from Figure 11 that there is no statistically relevant correlation between the acoustic prosody characteristics and the generated impression for all volunteers, except for a very weak correlation between pitch and valence. For only the male participants, only the female participants, only Japanese nationals, and all Brazilian nationals as separate groups, no statistically relevant correlation could be found.

However, it is necessary to investigate if there are significant differences on the impressions displayed by men and women and by Brazilian and Japanese volunteers. In order to verify if the variance of the samples are similar, we performed multivariate analysis of variance, MANOVA, on the obtained data, whose results are shown in Tables 2 and 3. The results of column Pr > F suggest that there is no statistically relevant difference between the reactions displayed by male and female volunteers. However, the reactions displayed by Japanese and Brazilian participants are statistically distinct.

Table 2. Results of the MANOVA for the data volunteered by male and female participants.

Group	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.9950	2.0000	628.0000	1.5839	0.2060
Pillai's trace	0.0050	2.0000	628.0000	1.5839	0.2060
Hotelling–Lawley trace	0.0050	2.0000	628.0000	1.5839	0.2060
Roy's greatest root	0.0050	2.0000	628.0000	1.5839	0.2060

Table 3. Results of the MANOVA for the data volunteered by Japanese and Brazilian participants.

Group	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.9716	4.0000	1120.0000	4.0687	0.0028
Pillai's trace	0.0285	4.0000	1122.0000	4.0534	0.0029
Hotelling–Lawley trace	0.0292	4.0000	670.9614	4.0889	0.0028
Roy's greatest root	0.0274	2.0000	561.0000	7.6855	0.0005

Thus, the original research hypothesis does not hold, i.e., there is no common baseline preference for particular prosodic patterns across cultures, across cultural groups, across

genders, and across age groups. Correlation between the acoustic prosody parameters and emotion change was investigated also for other groups, but since no other statistically relevant correlation was found, only the previously mentioned groups are displayed for the sake of brevity.

In order to strengthen the calculated correlations, assuming that some combination of the data points obtained through the "Talk to Kotaro" experiment actually reflect the real-world distribution of the reaction of how humans in general, men, women, Japanese people and Brazilian people would react to different acoustic prosody parameters, we perform statistical bootstrapping as defined in Section 2.5. In order to perform the bootstrapping technique, we consider the pairs (r, m), where r is either speed, pitch, or volume, and m is the either the associated δ_v or δ_a. During the Monte Carlo resampling operation, r and m are joined. After all sub-data sets are obtained, we separate all r and m into sets R and M and calculate the Stuart–Kendall τ_C correlation between both sets. For the present bootstrapping correlation analysis, we created 10,000 sub-data sets and used the percentile method to obtain the 95% confidence interval, whose results are shown in Table 4.

In order to obtain the desired $GSIP$ model, we first removed the outliers from the data set and trained the $MLP_{profile+prosody}$ model using the prosodic characteristics of Kotaro's remaining utterances and the profile of participants. Using the Adam optimizer (learning a rate of 10^{-3}, no decaying rate) with mean square error as the loss function, the model was trained for 100 epochs with a batch size of 32. The loss function for the training was mean squared error. Since the data set is quite small, 10% of the data were used for validation and 10% for testing. Two copies of the model were trained, one for valence and the other for arousal prediction. Together, they achieved an average error (as defined in Section 4.2) of 0.157 for the training data, 0.129 for the validation data, and 0.204 for the test data. The benchmarking results can be seen in Figure 12.

Table 4. Stuart–Kendall's τ_C correlation 95% confidence interval obtained through bootstrapping.

Group	Prosodic Parameter	Valence	Arousal
General	Speed	[−0.065, 0.038]	[−0.036, 0.070]
	Volume	[−0.095, 0.014]	[−0.040, 0.068]
	Pitch	[0.0045, 0.110]	[−0.054, 0.052]
Male	Speed	[−0.076, 0.063]	[−0.061, 0.077]
	Volume	[−0.091, 0.050]	[−0.074, 0.064]
	Pitch	[−0.020, 0.114]	[−0.048, 0.076]
Female	Speed	[−0.062, 0.098]	[−0.119, 0.061]
	Volume	[−0.156, 0.025]	[−0.060, 0.114]
	Pitch	[−0.063, 0.101]	[−0.076, 0.106]
Brazilian	Speed	[−0.058, 0.127]	[−0.168, 0.029]
	Volume	[−0.087, 0.100]	[−0.100, 0.095]
	Pitch	[−0.117, 0.060]	[−0.018, 0.142]
Japanese	Speed	[−0.053, 0.103]	[−0.090, 0.086]
	Volume	[−0.136, 0.041]	[−0.067, 0.100]
	Pitch	[−0.051, 0.113]	[−0.083, 0.085]

Regarding the training of model GRU_{phones}, it is performed in Section 4.4, since it is also used for investigating the positioning of the phones in the phone embedding space.

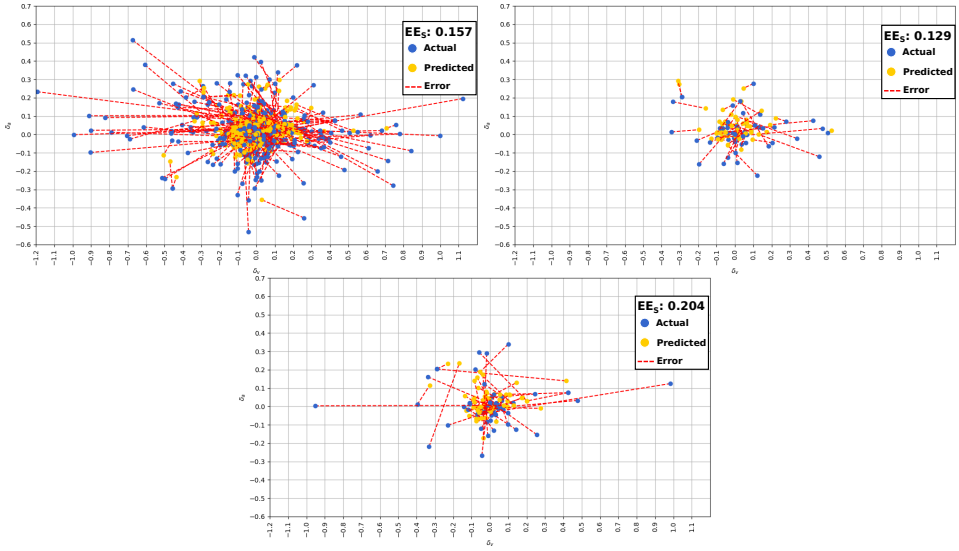

Figure 12. Comparison between the actual impression and the impression predicted by $MLP_{profile+prosody}$ for (**top left**) training data, (**top right**) validation data, and (**bottom**) test data.

4.3. Analysis of the Recorded Speech Supports the Findings of the Video Analysis

A total of 823 audio samples were recorded in the experiment, but many were unusable (were completely silent, contained very loud background noise, etc.), leaving us with 517 voice recordings. These voice recordings were analyzed using the LSTM-based neural network described in Section 3.2.2. The results are summarized in Table 5. It can be seen that the most frequent emotions of the recorded voice were disgust and anger, i.e., negative valence and low arousal values, and negative valence and high arousal values, respectively. These results are consistent with those obtained from the analysis of the volunteers' facial expressions, as shown in Figure 6. Happy, calm, and surprised initial states were rare but present in the interactions.

Table 5. Results of the sentiment analysis of volunteer's speeches.

Emotion Label	Number of Samples
Disgust	118
Angry	113
Happy	78
Surprised	54
Fearful	46
Sad	43
Calm	38
Neutral	27

Unfortunately, it was not possible to improve the accuracy of emotion estimation from the original video frames, but since the negative emotion estimates from the audio matched negative valence values and the positive ones matched positive valence values, it helped to validate, albeit qualitatively, the results of emotion estimation from facial expressions.

4.4. Phone Embedding Analysis

To investigate the contribution of each phone to the estimated impression across subjects, the GRU_{phones} neural network introduced in Section 3.2 was trained using the Adamax optimizer and mean square error as the loss function with a batch size of 32 for 100 epochs. Two copies of the model were trained, one for valence and other for arousal, to estimate the change in arousal and valence caused by a given string of phones. The output dimension of the embedding layer was chosen after trials with many different values; the best results were obtained with an output dimension of 64. Thus, the resulting embedding matrices for valence and arousal are of 64×71 dimension. However, since the

each phone has a position in a high-dimension hyperspace, it is not possible to visualize their proximity graphically.

To facilitate the analysis of the contribution of each phone, we used the k-means clustering method in order to group the phones accordingly to their proximity for both embedding matrices. In order to select the best number of clusters, we used the silhouette score analysis method [57] in order to determine the optimal number of clusters.

The optimal number of clusters was 35 for valence and 26 for the arousal embedding matrices. The number of phones in the largest cluster was 5, and most phones were grouped with distant phones or alone in their own cluster. From the clusters of the embedding spaces for valence and arousal, we were able to obtain Figures 13a,b, 14 and 15, which allows us to visualize which vowels and consonants phones belong to the same clusters, since such phones are colored with the same color. The red-colored phones were not selected by the gibberish speech generation algorithm, since the algorithm randomly picked phones.

However, since many phones were absent in Kotaro's gibberish speech, clustering might not be the best method to analyze which phones have a similar emotional impact on listeners, so we calculated the distance between the phones in the embedding space both for valence and arousal and for valence. Considering vowels and consonants separately, we obtained that for valence, the following phones that are close in their articulation loci are also the closest neighbors in the embedding space: [a]–[ɛ], [ʌ]–[ɛ], [ə]–[ɛ], [i]–[u], [p]–[m], [d]–[t], [k]–[g] and [ʎ]–[l]. If we consider all phones together, we have all combinations of vowels and consonants, but none with close articulation in the human mouth.

We have performed the same analysis for the arousal embedding space, and fewer pairs have a close articulation locus in the human mouth were obtained than for valence. The following pairs were identified: [ɛ]–[e], [ə]–[ɛ], [o]–[ʊ], [z]–[ʃ], [ʒ]–[ɾ] and [x]–[χ].

Such results show that while some phones are close in the learned embedding space for valence, it is not possible to claim that similar phones, except for a few exceptions, are close in the embedding space, in the context of non-Yulean gibberish speech.

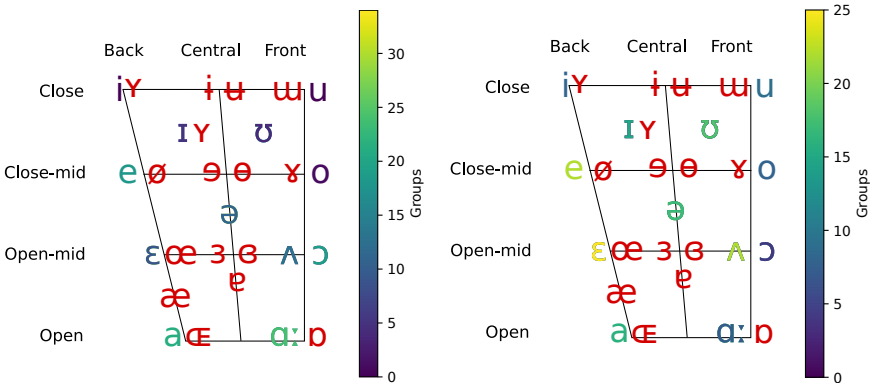

Figure 13. Embedding values for vowels of the IPA for valence and arousal estimation. IPA symbols in red were absent in the generated utterances. Other symbols were colored according to the index of the cluster they belong to, as shown in the rightmost color bar; (**a**) Embedding values of vowel for valence change estimation; (**b**) embedding values of vowel for arousal change estimation.

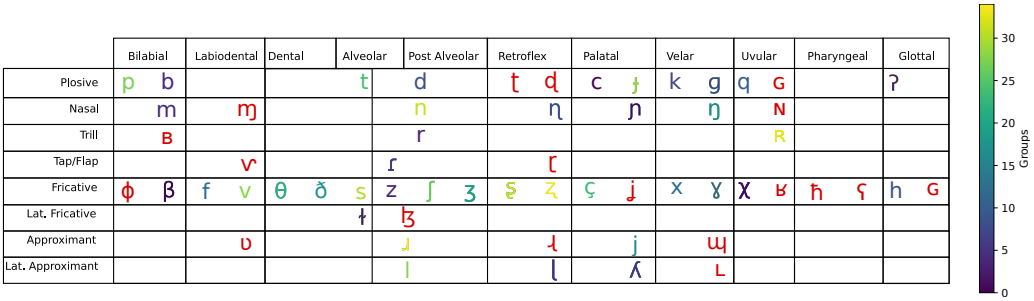

Figure 14. IPA Consonant table with embedding values for valence change estimation. IPA symbols in red were absent in the generated utterances. Other symbols were colored according to the index of the cluster they belong to, as shown in the rightmost color bar.

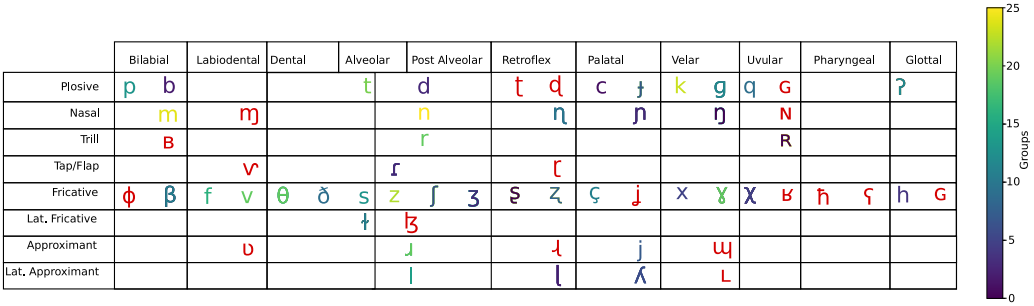

Figure 15. IPA Consonant table with embedding values for arousal change estimation. IPA symbols in red were absent in the generated utterances. Other symbols were colored according to the index of the cluster they belong to, as shown in the rightmost color bar.

The proposed neural network was then able to estimate the emotional change caused just by the tokenized phone vector w of a given Gibberish speech $S(w, P)$, achieving a prediction error of 0.035 for training data and 0.241 for validation data, as one can see in Figure 16.

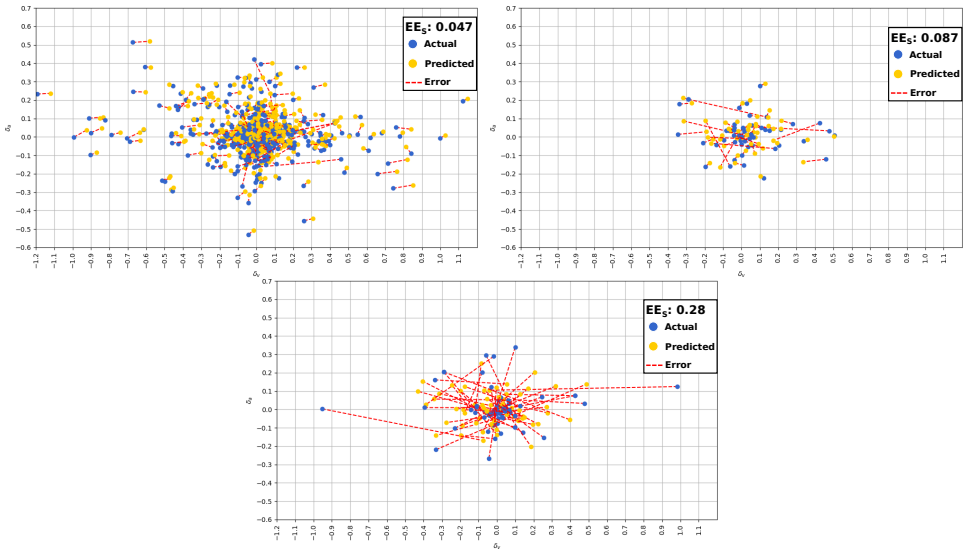

Figure 16. Comparison between the actual impression and the impression predicted by GRU_{phones} for (**top left**) training data, (**top right**) validation data, and (**bottom**) test data.

4.5. GSIP Evaluation

With both $MLP_{profile+prosody}$ and GRU_{phones} pre-trained, we further trained the combined models, using the standard gradient descent method (learning rate of 0.01 and no momentum) for 100 epochs with a batch size of size 32. It achieved an an average error of

0.141 for training data, 0.139 for validation data, and 0.19 for test data, as one can see in Figure 17.

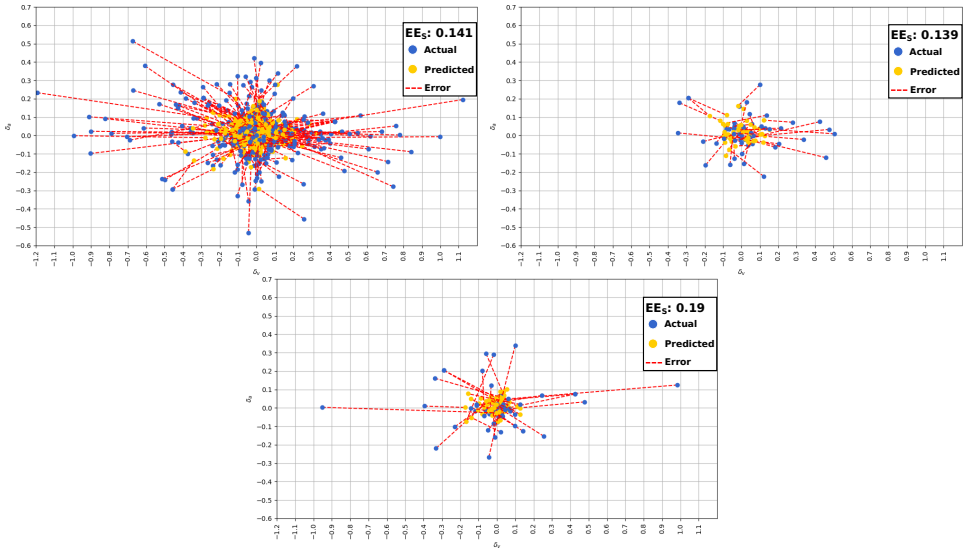

Figure 17. Comparison between the actual impression and the impression predicted by *GSIP* for (**top left**) training data, (**top right**) validation data, and (**bottom**) test data.

4.6. Likert Scale Questionnaire Analysis

Out of a total of 37 research volunteers, only 22 (13 male and 9 female) answered the optional Likert scale questionnaire after participating in the experiment. With the results, it is possible to perform a post hoc analysis of the internal consistency of the questionnaire. Cronbach's alpha was chosen to measure the consistency of the questionnaire prompts; we obtained a Cronbach's alpha of 0.752, with a 95% confidence interval of $[0.562, 0.881]$. The internal consistency of the questionnaire is, thus, considered to be sufficient, and we can proceed with the analysis of the responses of the volunteers.

Given that prompts P_5 and P_8 were worded negatively, the responses must be inverted before any analysis is performed. Prompt P3, although seemingly negatively worded, does not change its meaning when inverted, i.e., if it had been worded as "Some randomly generated words are more pleasant than others", it would not have changed participants' responses, since some words being less pleasant than others already implies that some are more pleasant. The same is not true for P_5 and P_8, which become "Different random words had an impact on your enjoyment" and "The turn-based conversation felt natural". To obtain the inverted responses IR from the actual responses AR, the following calculation must be made: $IR = MS - AR + 1$, where MS is the maximum score of the highest level of agreement; in this paper, it is 5.

To obtain the overall attitude toward a prompt, it is necessary to calculate the weighted average, where the value of a given item is multiplied by the number of respondents who chose that level of agreement, summed for each item, and divided by the total number of respondents in the questionnaire.

Results of the analysis performed on all prompts can be seen in the box and whisker plots shown in Figures 18 and 19, and the bar plots of each response by male and female volunteers can be seen in Figures 20 and 21. The overall attitude towards the Talk to Kotaro experiment was mostly neutral or slightly negative. Such results were expected after the emotion estimation analysis performed in Section 4.2, since most of the average emotion shown by participants during the experiment had negative valence and low but positive arousal.

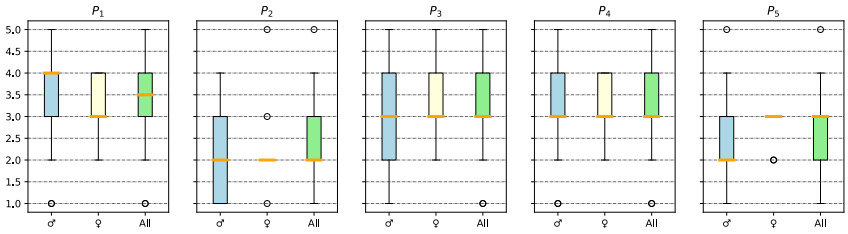

Figure 18. Male (blue), female (yellow), and everyone's (green) responses to the optional Likert scale questionnaire's prompts 1 to 5. The median value of the responses is highlighted in orange, outliers are represented by small circles.

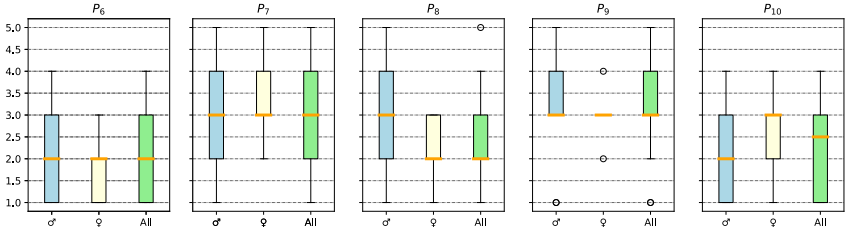

Figure 19. Male (blue), female (yellow), and everyone's (green) responses to the optional Likert scale questionnaire's prompts 6 to 10. The median value of the responses is highlighted in orange, outliers are represented by small circles.

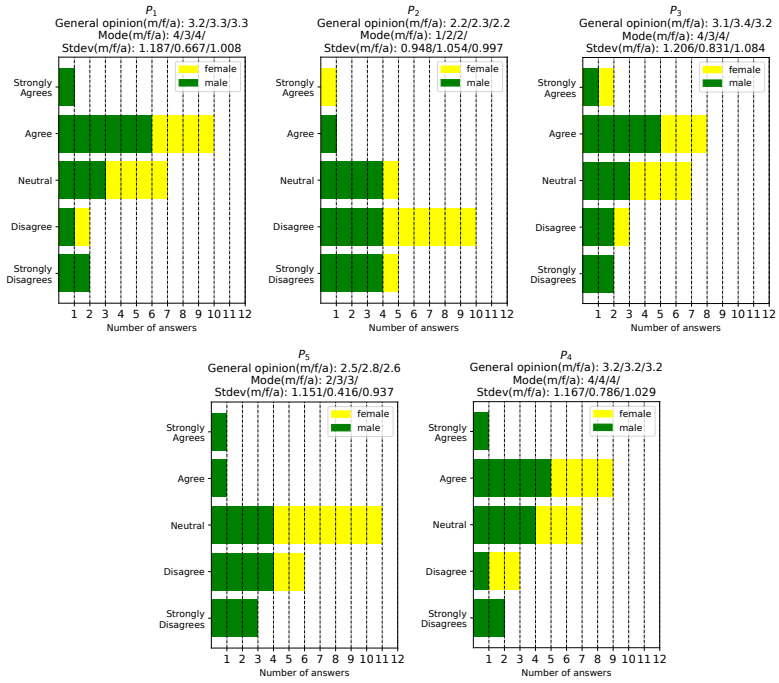

Figure 20. Bar plots of the male and female responses to prompts 1 to 5 of the optional Likert cale questionnaire.

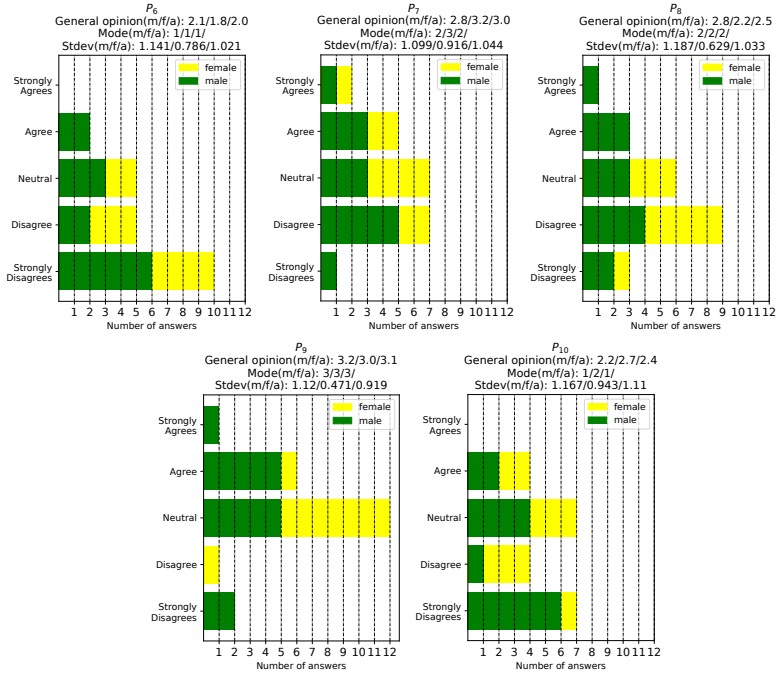

Figure 21. Bar plots of the male and female responses to prompts 6 to 10 of the optional Likert scale questionnaire.

P_1—Talking with the robot avatar was interesting

The general opinion is that talking to the robot avatar was just slightly above neutral, but it must be noted that the most common answer for male respondents was actually that they agreed that talking to Kotaro was an interesting experience, while most women tied between finding it neutral or slightly interesting experience. There are many factors that could have contributed to such results, but in line with the opinions of participants in [27], talking to a gibberish-speaking robot does not lead to a very enjoyable conversation, even if it is more interesting than the nodding robot. However, it must be noted that the attitude towards the experience of talking to Kotaro was worse than the attitude towards talking with the Hanamogera-speaking NAO robot in [27]. Such result suggests that GS that has phone distribution tends to perform better than GS that does not. We are, however, cognizant of the fact that more embodied conversational agents tend to elicit higher engagement and better impression on research subjects, and thus, this result warrants further investigation.

P_2—Variation of the Speech Characteristics Made Conversation More Natural

Volunteers, both male and female, felt that the random prosody variations used by Kotaro for his gibberish speeches did not make the conversation feel natural. This result was somewhat expected, since the avatar could suddenly change its voice from a very high pitch to a very low pitch, sounding like a completely different entity. Such a result is supported by previous research works, such as [58], where pitch inflection is identified as a very important factor in voice recognition. Another point is that low volume and high speed may have affected the overall experience, since people usually do not suddenly change the speed or volume of their speech unless there is a context for doing so.

P_3—Some randomly generated words are less pleasant than others

The most frequent answer for participants was "3—neutral", suggesting that research subjects could not see much difference on how distinct words generated by Algorithm 1 made them feel. This results suggests that non-Yulean gibberish speech words could not pick the interest of research subjects, again, in a similar fashion to the Hanamogera GS words in [27].

Moreover, since the algorithm also created some unusual combinations that were described by some participants to be "alien-like", generated words might have caused estrangement on participants.

This result is also consistent with the data shown in Section 4.2, since most of the utterances left a neutral-to-negative impression on participants.

P_4—Some speech characteristics, such as speed, loudness, or pitch influence more than others

The analysis of this prompt was of particular interest since there was little to no correlation between speed, pitch, volume, and valence, and arousal. The question was somewhat divisive among the participants, since the most common response was "4—agree" (10 responses), although the seven neutral responses, "3—disagree", and "2—strongly disagree", skewed the overall attitude towards neutrality. The overall attitude agrees with the result of the emotion analysis from the video samples and with the lack of correlation between the acoustic prosody parameters and the impression of volunteers.

P_5—Different random words didn't have an impact on your enjoyment

While the previous prompt analyzed the effect of prosody choice, this prompt analyzes the effect of phone choice. It is very similar to prompt P_3, but phrased differently to validate the results obtained. Since the prompt is negatively worded, in order to be comparable to the others, it is necessary to rephrase it as "Different random words had an impact on your enjoyment" and invert the responses.

The results were consistent for the overall attitude of all participants together and for female volunteers. However, for male volunteers, the overall impression worsened, since fewer male participants agreed with the random words. Such a result, even if unexpected, is more aligned with the emotion analysis from the video samples, but it shows that some volunteers might not be so sure of their opinion about the impact of phone choice in their impression.

P_6—You felt that the robot was answering your speech accordingly

This question, along with P_{10}, tests the perceived intelligence of Kotaro. The results indicate a highly negative perception, with the majority of male and female respondents strongly disagreeing with the prompt. The results suggest that the use of non-Yule-like distributions of phones and randomly changing prosody patterns leads to a poor opinion of the agent's intelligence. Participants were likely aware of the random selection of phones and prosody patterns, which contributed to their negative perceptions.

P_7—Longer phrases were more interesting

The overall opinion that longer sentences are more interesting than shorter ones was rather neutral, but one can see that male participants had a worse attitude towards longer utterances, suggesting that men would prefer shorter gibberish utterances as a response.

P_8—The turn-based conversation felt unnatural

Another negatively worded prompt, P_8, needs to be inverted to allow for a closer comparison with other prompts in the questionnaire. It then becomes "The turn-based conversation felt natural", which tries to capture the effect that pressing a button to talk and having Kotaro answer might have had on the volunteers' impressions. The general attitude is that the chosen turn-based conversation system felt unnatural. This was to be expected, since humans are very good at taking turns in conversation; the average silence between turns is within a range of 250 ms from the cross-language mean of 208 ms [59]. However, overall male impression was rather neutral, suggesting that such effect might be not as strong for male participants.

P_9—Foreign sounding phones were more interesting

The purpose of generating gibberish speech utterances using IPA symbols was to allow the ECA to use sounds from languages around the world, and this prompt was intended

to measure the impact that foreign-sounding phonemes had on participants. The results indicate that attitudes toward foreign-sounding phonemes were mostly neutral, but an analysis of other responses shows that they were slightly more negative than positive.

P_{10}—The robot seemed to be intelligent

Regarding the perceived intelligence of the embodied conversational agent, the results were mostly negative, in line with prompt P_6, although not as much, since female respondents had "2—disagree" and "3—neutral" as the most frequent responses, while male responses were mostly "1—strongly disagree". Again, subjects were able to perceive that the ECA randomly generated their responses. This prompt was phrased differently than P_6 to measure how the robot's humanoid form affected perceptions of intelligence, since being intelligent and responding accordingly capture two different aspects of the ECA's capabilities. While it did not improve the overall opinion of its intelligence, more responses were neutral, or even in agreement that the ECA was intelligent.

5. Discussion

The results of the analysis performed on the audio and video recordings, together with the investigation on the location of each IPA phone in the learned embedding space and the results of the optional Likert Scale questionnaire, are individual pieces of a larger jigsaw puzzle that must be pieced together in order to allow us to see the bigger picture, enabling us to obtain further useful insights and to contextualize our previously shown results.

5.1. Effects of Kotaro's Gibberish Speech on Listeners

The main takeaway when considering the results of the average emotion during the experiment, the impressions caused by the GS utterances, emotion classification of the audio samples, and the results of the Likert scale questionnaire is that GS that does not follow a traditional Yule-like phone distribution and has random acoustic prosody parameter selection does not have a good performance in a conversational setting with a screen-based conversational agent. Most utterances caused little to no impression, while the average emotion displayed while listening to Kotaro's utterances, $E_{avg_avg} = (-0.248, 0.161)$, could neither excite nor create positive feelings on listeners, on average. However, since the standard deviation for valence was quite high, $stdev_{valence} = 0.293$, we can see that there were still positive experiences, albeit few when compared with the neutral or slightly negative ones. Such results show that volunteers were mostly impatient and frustrated while listening to Kotaro's speech. There very few impression outliers, only 35 impressions, since most utterances caused small emotional changes.

Results of the analysis of what they told the agents also show that they showed little to no enthusiasm while talking to the agent, further showing that the overall experience was not particularly engaging. The neutral attitude toward prompt P_1 further sediments such a conclusion. Even though multiple participants have shown that they enjoyed through their answers, the majority still had a neutral or negative opinion of the experiment. Such results are in line with previous research results of work [27], where volunteers found the Hanamogera gibberish speech-speaking NAO robot more engaging than the nodding NAO robot, but volunteers still remarked that the conversations were still not so engaging. There was no acoustic prosody parameter variation in the GS utterances used by the NAO robot, and the overwhelmingly negative attitude towards prompt P_2 suggests that no variation of the prosody parameters performs better than completely random variations, as some volunteers also noted that drastic changes in pitch made them feel that they were speaking to a completely different entity, as voice pitch is a very important characteristic for identifying particular individuals just from their speech, as was shown in early voice identification works such as [58].

The main takeaway from the emotion analysis performed over the data provided by research subjects' suggestions and contrasting it with the results of previous research that focused on determining how research subjects felt regarding interacting with GS-speaking

conversational agents [27,32] is that while GS can provide positive interactions, its best use might not be in a conversational setting, since both in this work and in [27], volunteers complained about not understanding what the agent was saying and that they were not actually responding to their speech. Such results are unlike the ones shown in [32], where research subjects (children) played with a GS-speaking NAO robot a non-conversational setting, where the robot expressed its own emotions through GS. Since research subjects seemed to enjoy the experiment and to want to play again, GS in a expressive role (since the robot is using it to express itself) seems to perform better than in a conversational setting, where more objective meaning is expected. However, another aspect to be taken into consideration is that in [32], research subjects were children, while in the present work and in [27], research subjects were mostly young adults who might be less accepting of such odd and "alien-like" interactions, since it requires a more imaginative and playful imagination, less focused in the actual communication and more in the experience itself.

Another reason that might explain the worse performance of the present GS generation technique is that the agent itself could not capture the interest of research subjects. The idea of making it mostly expressionless in a not-vibrant environment was to give more focus on the speech itself. Having an ECA on the screen was a deliberate choice to make the task actually resemble more the conversation with a robot or other types of ECA. Moreover, since higher embodiment levels tend to create higher engagement on users, we thought that having volunteers talk with a GS speaking voice without any representation would feel even less engaging, since research subjects could feel like they were talking to a non-entity. The researchers were, however, aware that the choice of the appearance of the ECA also matters in experiments, and the humanoid appearance of Kotaro might have created a mismatch in expected intelligence and the lack of coherence of the words said by the ECA, which tends to generate a bad impression on users, as discussed in [60] and exemplified by the lower perception of the robot in [61].

5.2. Effects of Prosody, Duration of Interaction, and Phone Choice

Previous analysis performed on video, audio, and Likert scale questionnaire answers can help us understand how research subjects felt towards each utterance and the experiment itself, but does nothing to elucidate why, which was one of the goals of the "Talk to Kotaro" experiment. In order to understand how acoustic prosody parameters affect the impression of volunteers, we have calculated pair-wise Stuart–Kendall's τ_C correlation between each one of the investigated prosodic parameters and valence and arousal changes. Unlike what was previously thought by the researchers, no meaningful correlations could be obtained, with the exception of a very weak correlation of 0.06 between pitch and arousal for all participants, which had a p-value of 0.03.

By performing MANOVA analysis, it was possible to verify that Brazilian research subjects had distinct impression patterns compared to Japanese research subjects. Such analysis was not performed for other nationalities since they had too few participants (fewer than four), and thus, it would be a meaningless comparison between an individual and a group of participants, for most nationalities. However, joining the fact that there were no favoured prosody patterns by all volunteers considered as a single group and that volunteers from different cultures had statistically distinct reactions to prosody parameters, we found no support for the original hypothesis from the "Talk to Kotaro Experiment" that, like the Kiki–Bouba effect [56], there would be a cross-cultural preference for certain prosody characteristics; quite the opposite.

However, it is necessary to further investigate if the high p-values are due to the small size of the data set or if there is really no statistically meaningful correlation. One way of performing such analysis is to use for all participants KDE (kernel density estimation) [62] to learn the distribution of the pairs of (p, r), where p is one of the prosody parameters and r is the associated δ_v or δ_a value. With that, it is possible to create synthetic data whose distribution is very similar to the distribution obtained through the experiment and calculate the correlation between the synthetic set of p and r for different quantities

of synthetic data points until meaningful p-values are obtained for all pairs of Stuart–Kendall's correlation. Such an analysis is just a ballpark estimate, since it has a very strong assumption: the distribution of the real data obtained through the experiment actually represents (or represents closely enough) the actual distribution of how people react to different prosody parameters in the context of listening to non-Yulean GS.

We used a Gaussian kernel and chose a bandwidth of 0.1 to learn the distribution of our data in order to create synthetic data sets. With the distributions learned, we increased the size of the synthetic data sets until we consistently obtained meaningful, albeit still very weak, correlations between the synthetic pairs. We started obtaining mostly relevant correlations by 15,000 data points and always obtained statistically relevant correlations with 20,000 data points. Such a result shows that a much larger data set seems to be necessary in order to allow researchers to make stronger claims regarding the correlation between prosodic parameters and the impression of volunteers.

Linear regression was performed on the average emotion of each interaction volunteers had in their experiment sessions as a way of obtaining an overall tendency of how the emotion of participants evolved as they interacted with Kotaro. Both for valence and arousal, volunteers had positive or negative valence/arousal changes across the session, which are not explained by the number of interactions with Kotaro in a session, given that some volunteers that had multiple sessions in different days had days where valence/arousal improved in one session and worsened on the next one, just to improve in the final session, as shown in the bottom right plot for F3 in Figure 9. Additionally, average valence values fluctuated a lot in a same session, very rarely showing any linear tendencies. Arousal, on the other hand, has shown better linear fit for most of the research subject, but not all. Moreover, even if the majority of research subjects showed decreasing arousal as they interacted with Kotaro, which is expected as the experience loses its novelty or as the participant gets tired, some research subjects showed increasing arousal, which is counter-intuitive. However, since users did not answer any personality tests or write any notes that could help elucidate the reason, it was not possible to understand why such patterns happened.

In order to analyze the position of individual IPA phones in the learned embedding space, we developed the GRU_{phones} neural network, which was able to learn to predict the impression of volunteers from Kotaro's GS utterances quite well for training and validation data, which shows good confidence on the 64×71 embedding spaces for predicating valence and arousal.

However, due to the stochastic nature of the algorithm, not every IPA phone was selected for the experiment. Moreover, both calculating the distances between the phones in the learned embedding hyperspaces and the clustering operations have shown no support for the idea that similarly sounding phones cause similar impressions, but many more data are necessary to lay stronger claims in that sense.

5.3. Performance of the GSIP System

In order to develop the gibberish speech impression prediction system, neural networks $MLP_{profile+prosody}$ (responsible for predicting impression just from the profile information of volunteers and the acoustic prosody characteristics of a GS utterance) and GRU_{phones} (responsible for predicting human impression from the tokenized IPA phones of a GS utterance) were pre-trained using the obtained data set after the outlier impressions were removed. GRU_{phones} achieved and outstanding performance for predicting training and validation data, but for test data, the results seemed lackluster and mostly random, which shows a lack of generalization capability of the model. For $MLP_{profile+prosody}$, the results were not as impressive for training and validation data, but it performed better than GRU_{phones} for test data, showing closer predictions for some of the test data.

By using the pre-trained neural networks, we trained the $GSIP$ system, which consisted of the average of both previously mentioned neural networks, which achieved a better performance for test data when compared to previous two neural network, but it

showed a tendency of making more "average" estimates, since most utterances generated small emotional changes.

The results were not satisfactory for test data, showing that even though the models could perform reasonably well for training and validation data, they could not properly learn how to generalize that knowledge for never-seen-before data.

6. Conclusions and Future Works

In the "Talk to Kotaro" experiment, 37 participants from 10 different regions, speaking a total of 14 languages between them, contributed over 730 audio and video samples of their conversation with a 2D animated screen-based ECA, Kotaro. In order to investigate how gibberish speech whose phone distribution does not follow a traditional Yulean-like distribution and not a traditional syllabic structure, many different analysis were performed over the audiovisual data recorded in the "Talk to Kotaro" web-based crowdsourcing experiment. The research was mostly interested in the immediate emotional changes caused by listening to utterances $S(w, P)$ with distinct w vectors of IPA phones and the matrix of the associated acoustic prosody characteristics P, which were chosen accordingly to Algorithm 1. Moreover, we have also analyzed the average emotion displayed by volunteers while listening to Kotaro's GS utterances, since it gives a very useful insight of how participants felt during the overall experiment, instead of focusing in their momentary emotional state. Moreover, the quantitative and qualitative investigation performed over the optional Likert scale experiment helped us understand and validate the results of the previous analyses.

By analyzing the facial expressions of volunteers in the video samples and the main features of their speech through the MFCC of the audio samples, we were able to verify the findings of [27] that gibberish speech is not very engaging for talking with conversational agents. The experiments yielded little to no positive emotional impact, as indicated by the negative average emotion scores. While the diversity of valence responses suggests sporadic positive experiences, the prevailing estimated sentiment among participants was of impatience and frustration. The difficulty of changing the emotional state of participants and the mostly neutral and negative stance towards the prompts of the Likert scale questionnaire further reinforces the notion that engagement and overall experience provided by gibberish speech in a conversational setting were sub-optimal. The study's alignment with previous research underscores the challenge of forging engaging interactions with GS-speaking agents for adults, which suggests that it is not a recommended means of communication for a conversational setting, since conversations tend to feel one-sided, as highlighted by the attitude towards prompt P_6.

Delving deeper into the analysis of prosody, interaction duration, and phone choice, attempts to understand what characteristics of the GS utterances generated the estimated impressions, the results of Sections 4.2 and 5 show that the correlation between the prosodic parameters and the immediate emotion change on volunteers were not statistically relevant, barring a very weak correlation between pitch and arousal. Divergent impression patterns among participants from different cultural backgrounds (Japanese and Brazilian) suggest that the initial hypothesis of a cross-cultural preference for specific prosodic attributes does not hold, underscoring the complexity of cross-cultural communication preferences. However, further investigation on the effect of sample size on calculated p-values show that much more data are necessary to strengthen the finding of our work, which will be achieved through a longer user study.

The trained GRU_{phones} neural networks used for predicting valence and arousal changes from the tokenized IPA phones of the generated GS utterances achieved good performance for training and validation data; however, its generalization capabilities were lackluster. The interest in the pre-trained GRU_{phones} models lies in their learned embedding hyperspaces, particularly where each phone is positioned relative to other phones. The initial hypothesis to be tested in such analysis is that phones with close articulation location in the human mouth were expected to be close to each other in the valence and arousal

embedding spaces, since it was expected that they would generate similar impressions on listeners. Yet, limitations stemming from the stochastic nature of the algorithm while selecting phones for Kotaro's utterances did not allow all phones to figure in the data set, and thus reduced the capacity to which deeper investigations can be performed. Nonetheless, through the K-means clustering method and by computing the distances among all phones, we found out that the phones that are the closest to each other in the embedding space, more often than not, do not have a close articulation locus, showing no support for the original hypothesis. However, since not all phones were used in the experiment, Algorithm 1 needs to be modified to take into account phone frequency and which phones have not yet been used during interactions.

The results of Section 4.6 help to understand why most of the impressions presented by the research subjects were of low valence and low-to-moderate arousal, which is consistent with the mostly neutral or slightly negative attitude towards the ECA revealed by the analysis of the responses to the Likert scale prompts. It also shows that participants attribute a low level of intelligence to humanoid-like ECAs that only speak gibberish. The turn-based conversation was not a good interface for the research, since the participants did not particularly enjoy it, and for further experiments, VAD should be used to guarantee more natural conversations.

The route of predicting human impression from the gibberish speech patterns does not seem promising, as well as its use for human–computer and human–robot conversation, since research subjects seemed to not enjoy the experience, since they could not understand what the robot avatar was trying to convey. Even though research subjects were aware of the fact that the Kotaro avatar did not speak semantic speech, they seemed to still be trying to understand what meaning it was trying to convey. This way, it is necessary to compare the performance non-Yulean gibberish speech with Yulean gibberish speech and against other semantic-free utterances in order to better understand how it performs against other SFUs. Moreover, it is necessary to compare how those SFUs perform in a conversational vs. expressive setting, where conversational agents use SFUs to make the listeners believe the agent feels a certain way.

Regarding prosody selection, the $GSIP$ system was not able to predict human impression very well for test data, showing a lack of generalization capabilities. Yet, it obtained a better performance than $MLP_{profile+prosody}$ and GRU_{phones}, showing that taking information about the conversation partner, acoustic prosody capabilities, and the phones of the GS utterance allows the system to make more accurate predictions, even though it shows a strong preference for small emotional change predictions, which is in accordance with most of the data set. It is necessary to take into account that the lack of correlation shown between prosody parameters and the clash between the attitudes towards prompts P_3 and P_6 for male respondents might hint towards a complicated relationship between phone choice and impression. Another issue that is worth investigating is the validity of using facial expressions to estimate the emotional state of participants in low valence and arousal states, since it might be difficult to distinguish their actual emotion, since it is very close to neutrality.

For future experimentation, another possible route to establish rapport and show that the agent understands the emotion behind the words said by users is to use the prosody synchronicity approach [63], where the conversational agent establishes rapport by copying the prosodic parameters of the interlocutor.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app131810143/s1. The data set obtained through the Talk to Kotaro experiment and its partitions into the data set without outliers, and training, validation and test data sets.

Author Contributions: Conceptualization, A.G.C.G. and I.M.; data curation, A.G.C.G.; formal analysis, A.G.C.G.; investigation, A.G.C.G.; methodology, A.G.C.G. and I.M.; project administration, I.M.; software, A.G.C.G. and W.L.; supervision, I.M.; validation, A.G.C.G.; writing—original draft, A.G.C.G.; writing—review and editing, A.G.C.G., W.L. and I.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of Tokyo University of Agriculture and Technology (approval number 210801-0321 and experiment extension request approval number 220306-0321, approved 15 September 2021.

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

Data Availability Statement: The anonymized version of the data set obtained through the Talk to Kotaro Experiment and its partitions into the data set without outliers, and training, validation and test data sets are available as the supplementary files of the present journal paper, which can be downloaded in https://www.mdpi.com/article/10.3390/app131810143/s1 .

Acknowledgments: The authors would like to thank Patricia McGahan of Tokyo University of Agriculture and Technology for her great help with volunteer recruitment for the experiment.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CUI	Conversational User Interface
ECA	Embodied Conversational Agent
GRU	Gated Recurrent Unit
GS	Gibberish Speech
GSIP	Gibberish Impression Prediction System
GUI	Graphical User Interface
IPA	International Phonetic Alphabet
KDE	Kernel Density Estimation
MANOVA	Multivariate Analysis of Variance
MFCC	Mel-Frequency Cepstral Coefficients
MLP	Multilayer Perceptron
NLP	Natural Language Processing
NN	Neural Network
SFU	Semantic-Free Utterance

References

1. Deguchi, A.; Hirai, C.; Matsuoka, H.; Nakano, T.; Oshima, K.; Tai, M.; Tani, S. What is society 5.0. *Society* **2020**, *5*, 1–23.
2. Lasi, H.; Fettke, P.; Kemper, H.G.; Feld, T.; Hoffmann, M. Industry 4.0. *Bus. Inf. Syst. Eng.* **2014**, *6*, 239–242. [CrossRef]
3. Mah, P.M.; Skalna, I.; Muzam, J. Natural Language Processing and Artificial Intelligence for Enterprise Management in the Era of Industry 4.0. *Appl. Sci.* **2022**, *12*, 9207. [CrossRef]
4. Karunarathne, G.; Kulawansa, K.; Firdhous, M. Wireless communication technologies in internet of things: a critical evaluation. In Proceedings of the 2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC), Mon Tresor, Mauritius, 6–7 December 2018; pp. 1–5.
5. Janarthanam, S. *Hands-on Chatbots and Conversational UI Development: Build Chatbots and Voice User Interfaces with Chatfuel, Dialogflow, Microsoft Bot Framework, Twilio, and Alexa Skills*; Packt Publishing Ltd.: Birmingham, UK, 2017.
6. Lee, J. Generating Robotic Speech Prosody for Human Robot Interaction: A Preliminary Study. *Appl. Sci.* **2021**, *11*, 3468. [CrossRef]

7. Yilmazyildiz, S.; Read, R.; Belpeame, T.; Verhelst, W. Review of semantic-free utterances in social human—Robot interaction. *Int. J. Hum. -Comput. Interact.* **2016**, *32*, 63–85. [CrossRef]
8. Schwenk, M.; Arras, K.O. R2-D2 reloaded: A flexible sound synthesis system for sonic human-robot interaction design. In Proceedings of the The 23rd IEEE International Symposium on Robot and Human Interactive Communication, Edinburgh, UK, 25–29 August 2014; pp. 161–167.
9. Caroro, R.; Garcia, A.; Namoco, C. A Text-To-Speech using Rule-based and Data-driven Prosody Techniques with Concatenative Synthesis of the Philippines' Bisaya Dialect. *Int. J. Appl. Eng. Res.* **2015**, *10*, 40209–40215.
10. Sun, G.; Zhang, Y.; Weiss, R.J.; Cao, Y.; Zen, H.; Rosenberg, A.; Ramabhadran, B.; Wu, Y. Generating diverse and natural text-to-speech samples using a quantized fine-grained VAE and autoregressive prosody prior. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6699–6703.
11. Zovato, E.; Pacchiotti, A.; Quazza, S.; Sandri, S. Towards emotional speech synthesis: A rule based approach. In Proceedings of the Fifth ISCA Workshop on Speech Synthesis, Pittsburgh, PA, USA, 14–16 June 2004.
12. Lei, Y.; Yang, S.; Wang, X.; Xie, L. Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 853–864. [CrossRef]
13. Yilmazyildiz, S.; Henderickx, D.; Vanderborght, B.; Verhelst, W.; Soetens, E.; Lefeber, D. EMOGIB: Emotional gibberish speech database for affective human-robot interaction. In *Proceedings of the Affective Computing and Intelligent Interaction: Fourth International Conference, ACII 2011, Memphis, TN, USA, 9–12 October 2011*; Proceedings, Part II; Springer: Berlin/Heidelberg, Germany, 2011, pp. 163–172.
14. Gonzalez, A.G.C.; Lo, W.; Mizuuchi, I. Talk to Kotaro: a web crowdsourcing study on the impact of phone and prosody choice for synthesized speech on human impression. In Proceedings of the 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Naples, Italy, 29 August–2 September 2022; pp. 244–251.
15. Rheu, M.; Shin, J.Y.; Peng, W.; Huh-Yoo, J. Systematic Review: Trust-Building Factors and Implications for Conversational Agent Design. *Int. J. Hum. -Comput. Interact.* **2021**, *37*, 81–96. [CrossRef]
16. Mizuuchi, I.; Yoshikai, T.; Sodeyama, Y.; Nakanishi, Y.; Miyadera, A.; Yamamoto, T.; Niemela, T.; Hayashi, M.; Urata, J.; Namiki, Y.; et al. Development of musculoskeletal humanoid kotaro. In Proceedings of the 2006 IEEE International Conference on Robotics and Automation, ICRA, Orlando, FL, USA, 15–19 May 2006; pp. 82–87.
17. Fujisaki, H. Prosody, models, and spontaneous speech. In *Computing Prosody: Computational Models for Processing Spontaneous Speech*; Springer: New York, NY, USA, 1997; pp. 27–42.
18. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161. [CrossRef]
19. Ekman, P. Are there basic emotions? *Psychol. Rev.* **1992**, *99*, 550–553. [CrossRef]
20. Mondal, A.; Gokhale, S.S. Mining Emotions on Plutchik's Wheel. In Proceedings of the 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS), Virtual Event, Paris, France, 14–16 December 2020; pp. 1–6.
21. Kollias, D.; Tzirakis, P.; Nicolaou, M.A.; Papaioannou, A.; Zhao, G.; Schuller, B.; Kotsia, I.; Zafeiriou, S. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *Int. J. Comput. Vis.* **2019**, *127*, 907–929. [CrossRef]
22. Haukoos, J.S.; Lewis, R.J. Advanced statistics: bootstrapping confidence intervals for statistics with "difficult" distributions. *Acad. Emerg. Med.* **2005**, *12*, 360–365. [CrossRef] [PubMed]
23. Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans*; SIAM: Philadelphia, PA, USA, 1982.
24. Steck, H.; Jaakkola, T. Bias-corrected bootstrap and model uncertainty. *Adv. Neural Inf. Process. Syst.* **2003**, *16*, 521–528.
25. Efron, B. Better bootstrap confidence intervals. *J. Am. Stat. Assoc.* **1987**, *82*, 171–185. [CrossRef]
26. Diciccio, T.; Efron, B. More accurate confidence intervals in exponential families. *Biometrika* **1992**, *79*, 231–245. [CrossRef]
27. Kumagai, K.; Hayashi, K.; Mizuuchi, I. Hanamogera speech robot which makes a person feel a talking is fun. In Proceedings of the 2017 IEEE International Conference on Robotics and Biomimetics (ROBIO), Macau, China, 5–8 December 2017; pp. 463–468.
28. Yilmazyildiz, S.; Henderickx, D.; Vanderborght, B.; Verhelst, W.; Soetens, E.; Lefeber, D. Multi-modal emotion expression for affective human-robot interaction. In Proceedings of the Workshop on Affective Social Speech Signals (WASSS 2013), Grenoble, France, 21–22 August 2013.
29. Yilmazyildiz, S.; Latacz, L.; Mattheyses, W.; Verhelst, W. Expressive gibberish speech synthesis for affective human-computer interaction. In *Proceedings of the Text, Speech and Dialogue: 13th International Conference, TSD 2010, Brno, Czech Republic, 6–10 September 2010*; Proceedings 13; Springer: Berlin/Heidelberg, Germany, 2010, pp. 584–590.
30. Yilmazyildiz, S.; Athanasopoulos, G.; Patsis, G.; Wang, W.; Oveneke, M.C.; Latacz, L.; Verhelst, W.; Sahli, H.; Henderickx, D.; Vanderborght, B.; et al. Voice modification for wizard-of-OZ experiments in robot–child interaction. In Proceedings of the Workshop on Affective Social Speech Signals, Grenoble, France, 22–23 August 2013.
31. Tambovtsev, Y.; Martindale, C. Phoneme frequencies follow a Yule distribution. *SKASE J. Theor. Linguist.* **2007**, *4*, 1–11.
32. Wang, W.; Athanasopoulos, G.; Yilmazyildiz, S.; Patsis, G.; Enescu, V.; Sahli, H.; Verhelst, W.; Hiolle, A.; Lewis, M.; Canamero, L. Natural emotion elicitation for emotion modeling in child-robot interactions. In Proceedings of the WOCCI, Singapore, 19 September 2014; pp. 51–56.
33. Renunathan Naidu, G.; Lebai Lutfi, S.; Azazi, A.A.; Lorenzo-Trueba, J.; Martinez, J.M.M. Cross-Cultural Perception of Spanish Synthetic Expressive Voices Among Asians. *Appl. Sci.* **2018**, *8*, 426. [CrossRef]
34. Malfrere, F.; Dutoit, T.; Mertens, P. Automatic prosody generation using suprasegmental unit selection. In Proceedings of the Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis, Blue Mountains, Australia, 26–29 November 1998.

35. Meron, J. Prosodic unit selection using an imitation speech database. In Proceedings of the 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis, Scotland, UK, 29 August–1 September 2001.
36. Raitio, T.; Rasipuram, R.; Castellani, D. Controllable neural text-to-speech synthesis using intuitive prosodic features. *arXiv* **2020**, arXiv:2009.06775.
37. Fares, M. Towards multimodal human-like characteristics and expressive visual prosody in virtual agents. In Proceedings of the 2020 International Conference on Multimodal Interaction, Utrecht, The Netherlands, 25–29 October 2020; pp. 743–747.
38. Morrison, M.; Jin, Z.; Salamon, J.; Bryan, N.J.; Mysore, G.J. Controllable neural prosody synthesis. *arXiv* **2020**, arXiv:2008.03388.
39. Yi, Y.; He, L.; Pan, S.; Wang, X.; Xiao, Y. Prosodyspeech: Towards advanced prosody model for neural text-to-speech. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 7582–7586.
40. Shen, F.; Du, C.; Yu, K. Acoustic Word Embeddings for End-to-End Speech Synthesis. *Appl. Sci.* **2021**, *11*, 9010. [CrossRef]
41. Lee, Y.; Rabiee, A.; Lee, S.Y. Emotional End-to-End Neural Speech Synthesizer. *arXiv* **2017**, arXiv:cs.SD/1711.05447.
42. Tao, J.; Li, A. Emotional Speech Generation by Using Statistic Prosody Conversion Methods. In *Affective Information Processing*; Springer: London, UK, 2009; pp. 127–141.
43. Um, S.Y.; Oh, S.; Byun, K.; Jang, I.; Ahn, C.; Kang, H.G. Emotional speech synthesis with rich and granularized control. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May; pp. 7254–7258.
44. Sarma, P.; Barma, S. Review on stimuli presentation for affect analysis based on EEG. *IEEE Access* **2020**, *8*, 51991–52009. [CrossRef]
45. Duddington, J.; Dunn, R. eSpeak Text to Speech. 2012. Available online: http://espeak.sourceforge.net (accessed on 23 July 2023).
46. Association, I.P. *Handbook of the International Phonetic Association: A guide to the Use of the International Phonetic Alphabet*; Cambridge University Press: Cambridge, UK, 1999.
47. McMahon, A. *An Introduction to English Phonology*; Edinburgh University Press: Edinburgh, UK, 2002.
48. Kollias, D.; Zafeiriou, S. A multi-component CNN-RNN approach for dimensional emotion recognition in-the-wild. *arXiv* **2018**, arXiv:1805.01452.
49. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Trans. Affect. Comput.* **2017**, *10*, 18–31. [CrossRef]
50. Fussell, S.R.; Kiesler, S.; Setlock, L.D.; Yew, V. How People Anthropomorphize Robots. In Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction, New York, NY, USA, 12–15 March 2008; pp. 145–152. [CrossRef]
51. Takayama, L. Making Sense of Agentic Objects and Teleoperation: In-the-Moment and Reflective Perspectives. In Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, New York, NY, USA, 9–13 March 2009; pp. 239–240. [CrossRef]
52. Pichora-Fuller, M.K.; Dupuis, K. Toronto Emotional Speech Set (TESS). Borealis, 2020. DRAFT VERSION. Available online: https://borealisdata.ca/dataverse/toronto (accessed on 21 July 2023).
53. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [CrossRef]
54. Haq, S.; Jackson, P.J. Multimodal Emotion Recognition. In *Machine Audition: Principles, Algorithms and Systems*; Wang, W., Ed.; IGI Global: Hershey, PA, USA, 2010; Chapter 17, pp. 398–423. [CrossRef]
55. Chandra, M.P. On the generalised distance in statistics. *Indian J. Stat. Ser. A* **1936**, *2*, 49–55.
56. Ramachandran, V.S.; Hubbard, E.M. Synaesthesia—A window into perception, thought and language. *J. Conscious. Stud.* **2001**, *8*, 3–34.
57. Shahapure, K.R.; Nicholas, C. Cluster quality analysis using silhouette score. In Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, Australia, 6–9 October 2020; pp. 747–748.
58. Pollack, I.; Pickett, J.M.; Sumby, W.H. On the identification of speakers by voice. *J. Acoust. Soc. Am.* **1954**, *26*, 403–406. [CrossRef]
59. Stivers, T.; Enfield, N.J.; Brown, P.; Englert, C.; Hayashi, M.; Heinemann, T.; Hoymann, G.; Rossano, F.; De Ruiter, J.P.; Yoon, K.E.; et al. Universals and cultural variation in turn-taking in conversation. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 10587–10592. [CrossRef]
60. Briggs, G. Overselling: Is Appearance or Behavior More Problematic? 2015. Available online: https://www.openroboethics.org/hri15/wp-content/uploads/2015/02/Mf-Briggs.pdf (accessed on 3 September 2023).
61. Canning, C.; Donahue, T.J.; Scheutz, M. Investigating human perceptions of robot capabilities in remote human-robot team tasks based on first-person robot video feeds. In Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, USA, 14–18 September 2014; pp. 4354–4361.
62. Chen, Y.C. A tutorial on kernel density estimation and recent advances. *Biostat. Epidemiol.* **2017**, *1*, 161–187. [CrossRef]
63. Nishimura, S.; Nakamura, T.; Sato, W.; Kanbara, M.; Fujimoto, Y.; Kato, H.; Hagita, N. Vocal Synchrony of Robots Boosts Positive Affective Empathy. *Appl. Sci.* **2021**, *11*, 2502. [CrossRef]