UTRECHT UNIVERSITY
Faculty of Science
Graduate School of Natural Sciences
MSc Human Computer Interaction

TITLE

Supervisor:
Dr. Almila Akdag

Second Supervisor:
Dr. Zerrin Yumak

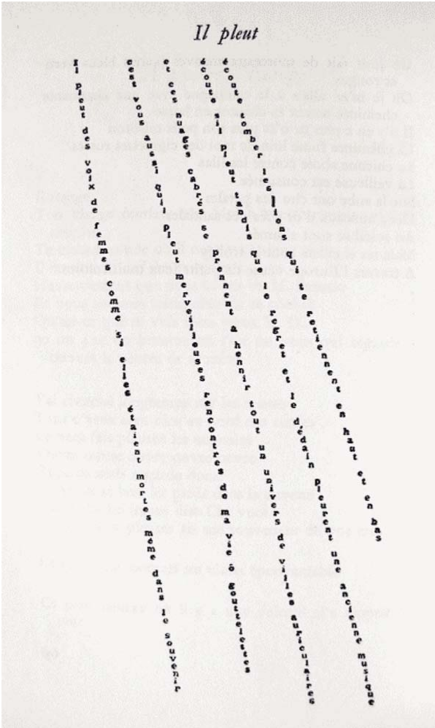
Nicolò Loddo
1531697

ACADEMIC YEAR 2022/2023

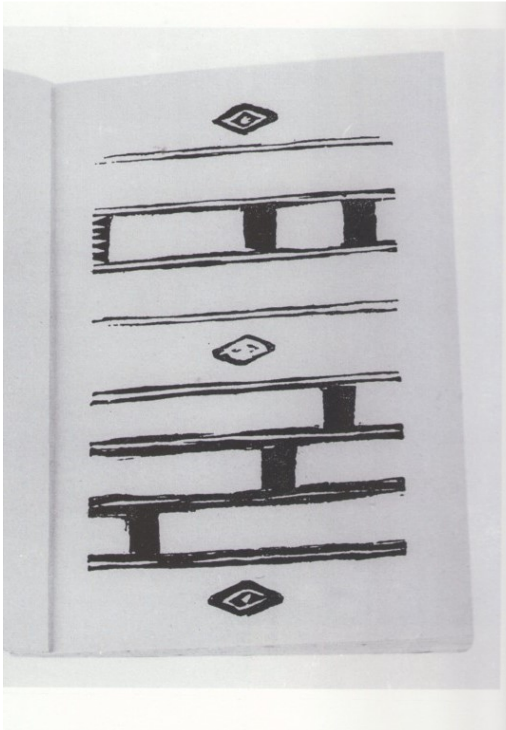
Contents

1 Inspiration

Humans' perception is a tool rich of sensory modalities, powerful when exploited fully. Whenever possible, we instinctively look for cues on every modal media available, to build an accurate estimation of the environment around us. Even when communicating it is important for us to investigate on what could tell us more about the other person. We look for micro-expressions, we find meaning in voices' pitch transitions, often we might seek for a little smile. We spent thousands of years developing language, passing from drawings, to hieroglyphics, to non-pictorial symbols, eventually achieving the possibility of explicitly communicating abstract ideas and emotions. This addition enriched deeply our face to face interactions... yet it is evident how the advanced word encoding that present language gave us, sometimes, leads to favour efficiency of transmission over richness in sensory modalities. Thanks to verbal language, ideas can be simply laid down, physically chained to a piece of paper, usually in black on white. Stripping the information down to the bare essential also makes it possible for me to reach you, reader, and communicate my curiosity towards human behaviour and computer sciences on this same document. I truly cannot be more thankful to be able to reach you this way. Still, aside from convenience of distribution, this is probably not the most beautiful way to communicate between each other, and I wish you could hear me speak, see me move my hands (I am also Italian...), connect and periodically break eye contact.
To read between the lines, is not only a practical need for communication efficiency. The uncertainty reduction theory by Berger and Calabrese hints at the fact that without the need of understanding the other further, the interaction almost becomes useless and not in our interest [1]. In many of us, there is a peculiar allure towards the mysterious or the unknown, arguably because of the pleasure for our intellect in understanding further, or maybe because we seek a connection with something greater, never actually understandable. Nothing better than art can be example of this, powerful testimony of our inner need of symbolistic communication, to the self and to the others, to the present and to the future. Two works in particular are to me important for this study and can be seen as triggers of the creative process that lead to this proposal. Firstly, we can see how written language is able to gain one bonus sensory modality in Apollinaire's "Calligrammes" collection. Poems's words are arranged in ways to visually represent what they are actually describing, providing additional cues for the experience of the reader. A second important example, is Arturo Martini's "Contemplazioni", the so called "mute book". In this work, no word is present at all, instead, its language and communicative power is achieved through sequences of black vertical marks interrupted by the white of the pages. Only rhythmic representations, without any verbal reference. Both works give power to non verbal cues; the white in between the black symbols is not anymore just a surface to lay the message on, but gains communicative power. In the same way as Contemplazioni celebrates the interruption between the black marks, the work introduced in this proposal wants to investigate the communicative power of the breathing pauses during speech, their irregular but persistent rhythm, embellished by occasional disfluencies such as "uh", "um" or "ah".



(a) Il pleut, Guillaume Apollinaire



(b) Contemplazioni, Arturo Martini

2 Introduction

The advent of social networks has significantly shaped our interpersonal interactions. Our methods of communication, which traditionally revolved around face-to-face encounters, are now often being replaced by digital modalities. While this has been empowering in several ways, such as enhancing the international exchange of information, the prioritization of communicative efficiency has led to a deflation of interaction modalities: we now rely heavily on textual communication, being content with sharing facial expressions through a discrete set of emoticons or emojis. The impact of this digital revolution on our ways of expressing affect and empathizing with each other is unfathomable, but most probably profound. As we advance into an age dominated by Machine Intelligence, with Agents becoming more and more integrated into our society, we can become even more alienated and dissociated from the genuine displays of affection typical of traditional human-human interaction. It has never been more important to understand how we interpret emotional cues from Artificial Agents. We must investigate if our perception of agents' emotional expression differs from the one shown by other humans, and we need to learn to design empathic agents in ethical ways, keeping in mind the well-being of our society. Are we able to genuinely empathize with Artificial Agents? Or will we forever question the authenticity of their emotions?
How will we rapport with such ever-spreading entities?

2.1 Empathy in Human-Computer Interaction

Empathy is a central feature of human interaction. Core moral values of society are built on top of our ability to understand the other. What we refer to with the word "empathy" in this work is the ability to recognize another entity's feelings and, importantly, to share that feeling and act accordingly. Many studies focus on the psychological foundations of it or its neuro-physiological factors. For this research, the scope is the one of emotions in Human-Computer Interaction: the so-called field of Affective Computing. In this field, empathy is a target behaviour to obtain, bilaterally. Affective interaction between humans and artificial agents can in fact be analysed from two perspectives: with the human as observer and the agent as trigger, or with the human as trigger and the agent as observer [2].

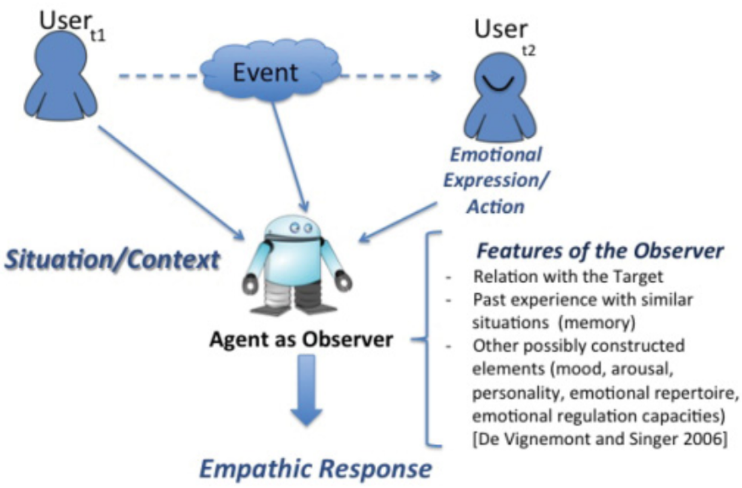


Figure 2: Agent as observer [3].

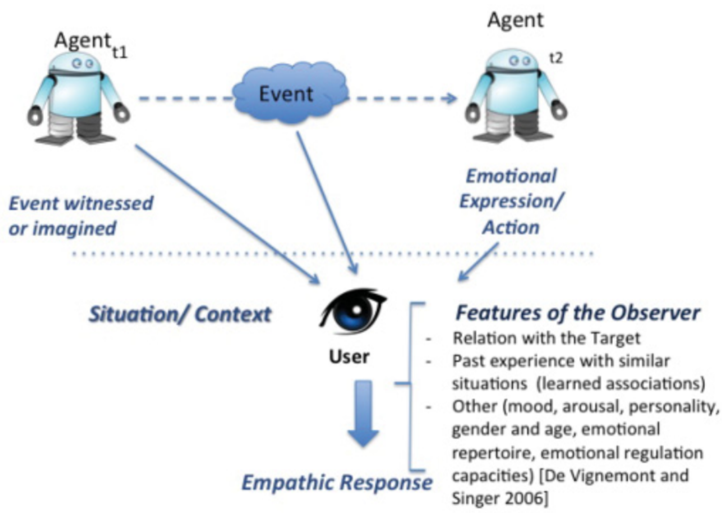


Figure 3: Agent as trigger [3].

Both perspectives are relevant:

- It is important for the software to understand our emotional state;
- It is important for artificial agents to communicate emotionally with the users.

Given the two, Paiva formulates the following definition of empathic agents:
"Empathic agents are (1) agents that respond emotionally to situations that are more congruent with the user's or another agent's emotional situation or (2) are agents that, by their design and behaviours, lead users to respond in a way that is more congruent with the agent's emotional situation." [2].

The recognition of users' emotions by artificial agents (agent as observer) is relevant for example in user adapting purposes. Persuasive applications can adapt to the emotional state to suit their methodology to users' state; in games and serious games, it can be used to adapt difficulty, lower or increase the cognitive load [4]. Moreover, artificial agents that can understand the users' emotional state are seen as more likeable and trustworthy [5], significantly improving the interaction.
Understanding the emotions of the user can also inform the consequent empathic reaction of the agent, which emphasizes the importance of the opposite perspective as well, with the Agent as trigger and the user as observer. Achieving an appropriate empathic response from artificial agents brings significant improvements in the interaction. In Terzioglu et al.'s study on collaborative Robots [6], it was examined the effect of adding *appeal*, *smoothness* in movement and *breathing* to provide social cues from robots to humans. The hypothesis is that through a perceived improvement in the anthropomorphism of the robots, likeability would be enhanced, as well as bolstering human-robot collaborations. The results prove that there is an increase in many of the examined features of anthropomorphism, and the breathing features in particular had a great impact in improving the interactions. Samrose et al. showed that empathic conversational agents can significantly reduce boredom caused by repetitive tasks [7]. Chin et al. found that agents that respond in an empathic way to verbal abuse significantly increase guilt and decrease anger in the user, confronted with avoidance or counterattacking reactions. Moreover, they are evaluated as significantly more likeable and generally perceived as more intelligent [8] (interestingly, the counterattacking one scored higher in anthropomorphism). Empathic agents have also been used for educational purposes. An example of this is FearNot! [9]: a pedagogical game designed to foster peer intervention in bullying incidents, for children between the 7 and 11 years of age. The experience tries to encourage empathy towards a virtual bullied child, to which the users need to give suggestions, essentially acting as their "invisible friend". The agent was successful in the German children sub-group, while no significant change was found in the English one. This highlights the fact that the agent's design has to be tailored to the specific context of the observer, as empathy not only changes depending on personality traits but also differs across cultures [3]. Paiva et al. [3] compiled a broad set of studies and agents designs, as testimony that "the presence of empathic responses by machines leads to better, more positive and appropriate interactions". However, they empahsize that the extent to which the interaction is improved also depends on specific design choices:

- The empathic features and general characteristics of the agent. Additional resources on this matter can be found from as early as 2006 with Hone [10], which found that different features of an empathic agent change its efficiency in reducing users' frustration, or Bickmore's studies on relational agents [11] [12] [13];
- The designed context and empathy mechanism to which the study relies on. Empathy can be due to automatic emotional responses, or by formulating a story and context inside which agent and user can build a relationship. With the latter giving the possibility of observing a more advanced perspective-taking mechanism.

Finally, certain design choices can also lead to unintended emotional responses from users, as detailed in Chapter 4.5. Because of this, it is pivotal to integrate the literature of empathic agents with novel examples, to better understand which design choices more authentically resonate with human emotions.

Figure 4: Empathic agent character used in 2006 by Hone [10].

2.2 Speech Communication

Speech is probably humans' most direct modality of communication. The high complexity of our language is paired with sophisticated sound articulation to achieve an impressively efficient encoding of information to sounds. Our use of the tongue in this process is unprecedented in primates [14]. The information conveyed through our voices goes often beyond the mere encoding of the words: it overflows the vessel and spills out information about the inner emotional state of the individual, as we will see in the following subsections. Additionally, we can often make assumptions about the social background, ethnicity or country of origin of the speaker based on accent and other vocal cues [15], reconstructing through inference a context that can make us understand the message better, or, in some cases, make us fall into the trap of intergroup biases and over stereotyping, leading to harmful prejudice [16].

In 2023, most smartphones come with an embedded voice assistant, and the market of this type of agents is increasing, with examples such as Amazon Alexa or Google Home populating many households. In 2022 the number of voice assistant users in the United States was estimated to be 142 million, or around 42% of the population of the time [17]. Forecasts predict that in 2024 the number of such speech based assistants will surpass the human population, reaching the 8.4 billion units [18]. Given the rise of voice assistants to such an impactful extent, and the progresses in the field of artificial speech synthesis, we decided to focus on the audio modality and on its possibilities to trigger empathy in users. To do so, we decided to abstract as much as possible from the actual content of the speech, as that can be addressed when focusing on textual modalities. Speech offers numerous other ways to convey messages beyond overt verbalization.

2.2.1 Paralinguistics

What we do not explicitly say, and its implications in communication, is studied in the field of Paralinguistics, researching how we non verbally convey emotions, intentions and a lot more. Non verbal cues in communication can vary across languages and cultures. Direct eye contact for example, can be considered attentive and respectful in some cultures (e.g. in most western countries), but it is considered aggressive and disrespectful in some others, such as in Japan and Korea. Prosodical cues' importance is even accentuated when the communication cannot be achieved through linguistic means, for example when speaking with somebody that does not understand our language, or to an infant that is still in a preverbal stage (i.e. not understanding words). For the latter, studies have shown that the message is mainly carried out by intonation and rhythm variations. When talking to a baby we instinctively perform modifications to our usual adult-adult prosody [19]. Higher pitch, greater tone variability, shorter utterances, and longer pauses are all reported modifications when speaking to preverbal infants across many languages and cultures [20]. Although such paralinguistic cues vary in very nuanced and instinctive ways, they are important to the point that through their analysis it is possible to detect Autism Spectrum Disorder in children from the 3 years of age, by demonstrating differences in facial expressions and higher pitch cries [21]. Furthermore, for years, in the ADReSS-M Challenge [22], many studies have been produced addressing the multilingual detection of Alzheimer's Dementia by analyzing spontaneous speech instances.

As said, for this study, we are specifically interested in Paralinguistics belonging to the auditory modality: the ensemble of acoustic and rhythmic effects performed while producing words, defined as "Prosody" [23]. Speech's tone and pitch characteristics, as well as its pauses, either filled by silence or breathing and disfluencies, are all of great importance in communication, with a crucial role in helping listeners discern between word boundaries, highlighting relevant information and expressing emotions [24].

2.2.2 Rhythm & Sound

Tone and rhythmical characteristics are among the main means of paralinguistic communication in spoken language. However, they are not only important paralinguistic concepts. As explained in the following paragraphs, they are probably the foundational roots of verbal communication, still defining deep cultural differences in communicative choices between languages, but also constituting an expressive inter-cultural bridge for different populations.

Speaking in Tones. In John Ohala's "frequency code" theory [25], he asserts that tonal cues originated in our pre-linguistic era, and work in communication not only across cultures but even across species. This communicative code is based on the fundamental frequency f_0 and on the richness of harmonics to communicate meanings such as "assertive" and "harmless" or "dominant" and "dangerous", hinting at the importance of intonation and pitch in emotive communication.
Tone variations are not only a key emotional cue, in some languages it is essential to distinguish the entire meaning of a word: these group of tongues are called Tonal Languages. A classic example is the word "ma" in Mandarin Chinese. "If you say it the way an English-speaker would say it, just reading it sitting by itself on a page, then it means *scold*. Say ≪ ma ≫ as if you were looking for your mother *ma?* and it means *rough*. If you were just whining at her *ma-a-a?!?* with your voice swooping down a bit and then back up even higher, that would mean, believe it or not, *horse*." [26]. In English, the

tone is used for example to indicate a question, by raising the pitch towards the end of a sentence, or to highlight a word in the sentence, but does not help in differentiating words, which makes it a Pitch-Accent Language.

Speaking in Rhythms. The perception of rhythm has played a significant role in human history, dating back to ancient times. One of the earliest known examples of rhythmic perception can be found in the drumming patterns of indigenous cultures throughout the world, such as in Africa, the Americas, and Australia. Rhythm, perceived as the unfolding of temporal structures and timed stimuli, is critical to listeners' emotional and behavioural responses [27]. Moreover, rhythm is not a simple direct product of timed stimulus, instead, our mind and brain has an active role in the perception of it [27]. An example of this contribution has been shown decades ago with the observation of the "tick tock" phenomenon [28]: an isochronous (i.e. happening at constant intervals of time) stream of identical sounds is perceived by humans as an alternation of strong and weak notes. Believe it or not, our clocks do not make "tick tock" sounds, they do not alternate two tones, they rather just "tick tick" instead.

In verbal communication rhythm has a big role. Recent studies have demonstrated how a better ability of rhythm perception enhances conversational quality and is a big factor in rhetorical success [29]. Moreover, Ververidis and Kotropoulos [30] report, in their survey of emotional speech recognition studies, the "speech rate" feature as one of the main factors in emotion recognition. This is defined in papers either as the "inverse duration of the voiced part of speech determined by the presence of pitch pulses", or as the "rate of syllabic units" and shows clear differences in many papers of the review depending on the emotional state.

Isochrony is also an important factor in languages distinction, by identifying their specific production rhythm and division of time. There are two main families of languages in the rhythm continuum: Syllable timed and Stress timed. In the former, speech is produced with the syllables taking around the same amount of time. In the latter instead, syllables have different duration, and the time between consecutive stressed syllables is kept the same. Spanish, Italian, French, Turkish, Chinese are some examples of syllable timed languages. English, German, Dutch and Catalan are some examples of stress timed languages. Brazilian Portuguese belongs to the first, while European Portuguese to the latter: their key difference in rhythm might significantly contribute to the different perception of the two.

The algorithms that can give voice to Virtual Agents have made big steps ahead in improving tonal and syllable articulation, but our communication modalities in speech are not limited to these. Inside language, an instinctive and necessary behaviour is the one of breathing planning and the production of disfluencies (such as "uh", "um"). The rhythm of these features can be of great importance inside empathy's mechanisms, and has to be distinguished from the prosodical rhythm because it is related but not congruent with syllables' rhythm. Considering breathing's **rhythm and sound** is crucial for studies that explore empathic communication, as its timing heavily influences the planning of future words and its frequency can give insights on the emotional arousal of the speaker. We will explain in detail the importance of breathing in Chapter 2.3, while the last subsection of the current Chapter will introduce the concept of spontaneous speech. This is the specific type of speaking style usually employed in colloquial situations, where features like disfluencies and breathing gain more importance because of the need of performing extemporaneous speech planning.

2.2.3 Spontaneous Speech

An important distinction in humans' speaking style comes from the spontaneous and non-spontaneous nature of speech production, which can significantly impact the structure, content, delivery, and underlying cognitive processes involved in communication. What we will refer to as "spontaneous speech" are speaking instances characterized by an unplanned and unstructured nature. Typically produced in real-time without the benefit of prior planning or editing. Spontaneous speech often includes repetitions, false starts, expressive sounds such as "sighs" or "laughs", and disfluencies, such as "um" and "uh". "Non-spontaneous speech", on the other hand, is pre-planned and structured. It often follows a logical organization and has a more consistent syntax, with well-formed sentences and fewer disfluencies. This results from the speaker's possibility to pre-compose and revise their speech, ensuring a higher level of coherence and clarity. Because of its pre-planned nature, non-spontaneous speech generally exhibits more controlled and consistent prosody. The speaker's intonation, rhythm, and tempo are likely to be more stable and predictable, as they have been rehearsed or pre-determined. It is therefore clear why this distinction is important to make when studying paralinguistic features and their impact on emotional content, and when analyzing the challenges that spontaneous speech could bring in the design of computational speech synthesizers.
Another important difference to make is the role of pauses and the impact of breathing in the two presented types of speech. In spontaneous speech, pauses often reflect the cognitive processes occurring as the speaker formulates their thoughts and manages in real time their need for inhalations and exhalations. In non-spontaneous speech, pauses are more deliberate and can be strategically employed to create emphasis, allow for audience comprehension, or signal a transition between topics.
By providing the additional cues of breathing pauses, disfluencies and other expression sounds, spontaneous speech has the possibility of enhancing the emotional content delivered by an Artificial Agent. Moreover, spontaneous speech has the advantage of sounding more colloquial, which might make this speaking style more suitable for agents who have to interact in a friendly, relatable way, for example in teaching assistant applications. So "Should Agents speak like, um, Humans?" some first answers to this question come from the accordingly-titled study by Pfeifer and Bickmore, in 2009 [31]. The approach they took was the exposure of 23 participants to either an agent that used conversational fillers or its version that did not (a between-participants design). The subjects would then answer a questionnaire that asked about their judgment on the satisfaction, trust, likeness, knowledgeability and naturalness of the agent. The study did not reach significant results, probably due to the limited sample size. Interestingly, the subjects reported mixed feelings about the agent. Five participants indicated that the use of fillers by a conversational agent seemed inappropriate, given that computers have the ability to speak perfectly, and another five participants indicated that the usage of fillers by the agent was a positive aspect of the conversation and "humanized" the experience. Other studies on conversational fillers found a positive impact on the perceived social presence [32] of the agent, and on its responsiveness, agency, aliveness and likeability [33]. The question, though, remains relevant today: how suitable is spontaneous speech as a means of communication for Artificial Agents? We aim to contribute to this query by focusing our study on the impact of the breathing cue in Virtual Agents's speech.

2.3 Breathing and Emotions

Breathing noises are a feature present both in spontaneous speech and non-spontaneous speech because their main task is of physiological nature: providing oxygen to our bodies. Evidence shows that breathing has an impact on speech planning, and conversely, is

affected by speech planning [34]. However, on occasions in which we do not have to perform extemporaneous speech planning, such as in read (or memorized) speech, breathing instances can be organized around grammatical sentence boundaries [35], possibly minimizing its impact on the speaking flow.
Even if we maintain to some extent conscious control over it, respiration is an involuntary physiological process controlled by the Autonomic Nervous System (ANS), the part of our nervous system that controls unconscious tasks. The relationship between physiological responses (such as breathing) and emotions has been long debated, with various arguments on the specificity of physiological responses depending on affect, and their causality direction. One of the earliest theories in the field is the James-Lange Theory [36], dated back to 1894. This supports a complete causal correlation between physiological reactions and emotions and implies that each set of ANS responses leads to the perception of a distinct emotion with *absolute emotion specificity.* Diametrically opposed is the Cannon-Bard theory [37] from the 1920s which instead states that emotional experience and physiological arousal arise independently. From this, the concept of *non-specificity of emotions*: without causality, similar physiological responses can happen with the same emotion. Some subsequent theories advocate for a more *contextual specificity of emotions* and physiological reactions such as in Lazarus (1980) [38] or Leventhal (1982) [39] [40] cognitive theories. Leventhal in particular supports the idea of an interaction between emotions and cognition, in which one influences the other and vice-versa. Contemporary researchers mostly refute absolute specificity, as unambiguous signs of physiological activity based on emotion are hard to find. An example of this neglect can be seen in Feldman-Barret's Theory of Constructed Emotions [41] (2006), which is today one of the most widely recognized theories on emotions in cognitive psychology, also supported by studies in the field of neuroscience. The theory posits that affect should be recognized as a continuum, instead of being discretized in emotional categories, advocating the impossibility of defining a specific set of emotions to which attribute physiological signatures. Furthermore, it states that the brain works as a prediction model, rather than a reaction model. New physiological responses and affective state emerge then as a combination of past experiences, sensory input (both from the outside world and the body itself), and the current affective state. This means that the same emotion can lead to different physiological patterns depending on context and individual.

Because of its voluntary-involuntary duality, breathing is a physiological symptom that holds the possibility of conscious manifestation. If the brain, physiological responses and emotions are completely intertwined with each other, as in the constructive approach of Feldman-Barret, then the manipulation of the respiratory profile can cause changes in our emotional state and in other physiological symptoms. This has been seen to be true in various studies. For example, conscious changes in breathing are known to be linked to cardiovascular and skin conductance responses [42]. Philippott et al. (2001) [43] found noticeable differences in self-assessed emotional states unknowingly induced through the simulation of breathing patterns. In particular, joy and anger were seen to be better recognizable after the elicitation. Jerath et al. (2015) [44] argue that a broad variety of self-regulatory breathing techniques have been used throughout history to increase well-being, such as meditation techniques, yoga or *pranayama*, and although they have yet to gain full acceptance in Western culture, their impact has been proved in a multitude of studies. These link the control of breathing with positive results on subjects' stress, anxiety and even depression symptoms. Kim et al. [45] in 2013 found that mindfulness-based stretching and deep breathing exercises normalized cortisol levels and reduced Post-Traumatic Stress Disorder (PTSD) symptoms severity.
In the paper "The sound of silence", Akdag Salah et al. [46] analyse non-verbal signs

of Post-Traumatic Stress Disorder from victims of scarring events (Holocaust, Nanjing Massacre, Tsunami, Guatemalan Genocide, Tutsi Massacre), interviewed and reported in Historical Archives. The aim is to “enrich the semantic information contained in oral history archives by adding non-linguistic features”, discussing the possibility of finding PTSD cues beyond cultural and linguistic barriers. The specific focus of the study is on respiratory patterns, analysed across various conditions. The results suggest the inherent power that breathing holds, especially communicating the discomfort that recalling such episodes comports. Although it did not reach statistically significant results, the study shows precious hints that breathing patterns can provide for depressive episodes. While complete evidence for a link between breathing patterns and depression or PTSD is lacking, respiration features are instead clinically linked to panic disorders and anxiety [47] [48].

Roes et al. [49] found that sad music (personally selected by the subjects) influences peak-to-peak chest respiration distance negatively. The emotions examined were happiness, sadness, calmness and annoyance. No other differences were found significant between emotions nor to the control no music condition. Siddiqui et al. [50] designed an algorithm that could distinguish *disgust*, *happiness* and *fear* from the detected respirations per minute (RPM) of the subjects, with 76% accuracy. Interestingly, the RPM was accurately captured using radio signals recorded from a distance. This suggests the future possibility of detecting a person’s emotional state from afar by analyzing their breath.

Kreibig (2010) [51] reviewed 134 studies on the physiological response to emotion elicitation. The discrete set of emotions is directly derived from the said literature. The effects on respiration are various. Figure 5 highlights the ambiguity of respiratory patterns, with many states reporting both increases and decreases, as well as many sharing the same features.

	Anger	Anxiety	Disgust Contamination	Disgust Mutilation	Embarrassment	Fear	Fear Imminent threat	Sadness Crying	Sadness Noncrying	Sadness Anticipatory	Sadness Acute	Affection	Amusement	Contentment	Happiness
Respiratory Rate	↑	↑	↑	↑		↑		↑	↑	↓↑	↓↑		↑	↓↑	↑
Breath Depth	↓↑	↓	↓	(↓)		↓↑		↓	↑	↓↑	↓		↓↑	↓↑	↓↑

Figure 5: Breath frequency and depth changes depending on emotional state. ↑ indicates an increase, ↓ indicates a decrease, and ↓↑ indicates that both increase and decrease are reported between studies. If the arrows are in parentheses it means that the trend has been found in less than three experiments. The Table is derived from the one designed by Kreibig [51] in her 134 studies survey. The original Table reports many more respiration features, moreover, the set of emotions was broader. We removed the emotions which trend was supported by fewer studies inside the survey.

In general, it is difficult to find rhythm fingerprints that can reliably assess the emotional state of a person [52] [49]. A reason for this might come from our categorization methodology of the emotions: trends in physiological processes might be better explained for example by differentiating between exciting affect states (such as anger or happiness) and anticipation affect states (such as anxiety or fear that anticipate threats), which are better rooted in the ultimate evolutionary purposes of our responses to the environment. Indeed, faster and deeper breathing is often linked to emotions rooted in excitement, fast and shallow breath to emotions rooted in anticipation, while slow and deep breath to relaxed states [43].

Regardless, it is exactly because of the still ambiguous link between breathing patterns and emotional states that new studies are needed on its possible influence and power, literature to which we aim to contribute focusing on breaths' impact during emotional communication. Furthermore, hints of breathing having an active role in emotional communication are also present in our everyday lives. We purposefully use specific breath profiles to communicate our emotional reaction when we laugh, sigh, gasp: breathing is embedded both consciously and subconsciously in our means of interaction.

In summary, since breathing is generally affected by emotions, and breathing affects speech planning (particularly in spontaneous speech), we can infer that breathing can have an important role in emotional communication. Therefore, we find it important to further investigate its nuanced expressive capabilities in human-agent interaction. Moreover, we find it arguably undervalued in the generative algorithms that give voice to agents, as explained in Chapter 5.

Our study will first explore the available methodologies for the artificial generation of speech-breathing and then investigate its integration in empathic Artificial Agents design to understand if that can enhance users' empathy towards them.

In the following section, Chapter 3, we will introduce our Research Question. We will then explain the challenges this project encounters, as well as the various approaches taken in the literature to tackle them in Chapter 4. The problem of speech synthesis and the State of The Art of the field is detailed in Chapter 5. With our experiment design, presented in Chapter 6, we try to avoid to question the likeability of the Agent in the interaction, posing instead the subjects in front of the direct dilemma of giving up to the gratifying feeling of winning a videogame against empathizing with an AI's wishes, derived from its inner "feelings". With this novel approach, we aim to provide a new perspective on users' emotional reactions towards Virtual Agents and the provided breathing capabilities. Finally, we will present and discuss the results of the study in Chapter 7.

3 Research Question

This thesis addresses the challenge of integrating breathing patterns in synthesized speech, and their role in enhancing Virtual Agents' abilities to communicate and express emotions, with the ultimate purpose of increasing users' perceived empathy towards the agents. For these purposes, the study investigates the available models for the generation of English spontaneous speech (with breathing noises, and filled or empty pauses) inside the current State of The Art landscape. The study then evaluates the impact of breathing instances in Virtual Agents' produced utterances, to assess their influence on the perceived emotion expressiveness, naturalness, and persuasive power. To achieve this, we introduce a novel study design, that utilizes a gamified experience to evaluate the Virtual Agent's communication capabilities.

Main RQ:

"Can breathing patterns in synthesized speech improve the perceived empathy towards Virtual Agents?"

Sub-RQ 1, 2, 3:

What is the impact of breathing sounds produced by State of The Art Speech Synthesis models on Virtual Agents' voices, in terms of:

- S-RQ 1: Emotional expressiveness?
- S-RQ 2: Persuasive power?
- S-RQ 3: Naturalness?

Sub-RQ 4:

How can we produce emotional, spontaneous speech with breathing using State of The Art models?

4 Challenges and Related Work

In this Chapter, we will address the main challenges that the project has to tackle. First, we introduce the difficulties of reproducing human speech with software. Next, we address the challenge of integrating breathing inside the computationally synthesized speech, reporting various speech-breathing agents, which had various types of responses from the users. As a third point, we will explain the difficulties coming from the task of evaluating empathy and emotional reactions, tackled in various ways in the literature. Finally, we introduce the problem of the Uncanny Valley: a dip in the emotional response to agents which can compromise the elicitation of empathy towards artificial agents.

4.1 Spontaneous Speech Articulation

The intricate processes of word articulation and tone or rhythmical variation that humans achieve through speech communication are really difficult to computationally reproduce, and constitutes the first main challenge of this study. Humans learn how to articulate words during the first years of their life, while being given word examples from their parents and environment. Likewise, the first requirement to generate speech through software solutions is often the collection of well-curated data: examples for the software to be trained on. This first important requirement is already a big barrier, not always easily met. In particular, genuine and spontaneous emotionally labelled speeches, with a fair richness of non-verbal cues, still have a limited availability of public and complete datasets. The most broadly used Datasets to train AIs on the task of generating speech are the LJS [53], VCTK [54] and LibriTTS [55]: all three of them do not contain spontaneous speech instances, as the recordings are performed while reading passages from scripts, and breaths are not signalled in the text and often hard to hear. A large number of emotional speech datasets is done by asking subjects or actors to mimic an emotion, leading to stereotypical and forced emotional responses that lack ecological validity. A specific analysis of the available datasets for the scope of this study is done in Chapter 6.3.1. Several others lack quality in the recordings, which also leads to the loss of emotional cues like breathing sounds. Often, the lexical variability is limited, and the transcription is either missing or with different styles of annotations between datasets. This has also been noted in other studies and literature surveys [49] [56] [57], and work has been put into this to try and fill this research gap with modern techniques. Emotional responses remain though difficult to annotate and elicit in controlled settings, and this problem might persist in the future.

After the data requirements are met, it is crucial to utilize an appropriate modelling technique to approach the word articulation task. This can be tackled through a variety of techniques, from processing models [58] to probabilistic models [59] and, finally Neural approaches, which are discussed in detail in Chapter 5. The latter is the prevalent (and generally best-performing) methodology in the current State of The Art, populated also by hybrid techniques such as the Neural-Probabilistic model by Mehta et al. [60].

4.2 Breathing and Speech-breathing for Artificial Agents

Focusing on the breathing cue, synthesizers that can reproduce such feature of speech are rare. The use of commercial speech generation services seems to be generally focused on the creation of content that usually does not need breathing or spontaneous speech, such as commentaries in videos and documentaries. An approach to the task at issue is the one employed by Bernardet et al. [58]. Their system focuses on producing speech-breathing utterances using a text-to-speech algorithm that dynamically inserts prerecorded breathing sounds in the speech. The placement of breathing sounds is controlled by a timing algorithm that has been informed thoroughly by studies on the

Physiology of breathing and speech-breathing. The system has not encountered any human evaluation. Timing predetermined breathing sounds can be difficult because of the delicacy of rhythm and synchronization of breaths. Moreover, inhalation and exhalation noises often vary depending on the situation and tone of conversation, highlighting the limitation of such approach.
Breathing Artificial Agents do not necessarily need to display the respiration through audio cues. Novick et al. in 2018 [61] presented PaolaChat, an embodied Virtual Agent that shows breathing thanks to the animations of the character. The study analyses the impact of the introduced feature on users' perception of naturalness and rapport with the agent. The agent is interactive in real-time, using a keyword-based method to understand the response of the user to "her" questions. No emotional appraisal was implemented in the model, which only focuses on showing the agent's standard respiration. Their approach, as the one from Bernardet et al., also followed a timing algorithm based on physiological parameters, but no significant increase or decrease in rapport and naturalness was found.
The studies described above emphasize the difficulties in emulating breathing through simple physiological models. A thorough analysis of speech synthesis methods is presented in Chapter 5, solely presenting audio techniques for speech synthesis to apply to agents, and particularly focusing on Neural Network approaches. These are relevant because non-embodied agents can utilize any type of speech synthesis method since they do not need to focus on animating the body according to respiration. However, for the design of an interactive conversational agent, such as PaolaChat, additional modules would be needed even assuming it to be non-embodied: before producing speech with a chosen speech synthesis method, the agent has to understand users' message and respond accordingly. The advent of models such as OpenAI's ChatGPT 4 [62] is expected to significantly contribute to future conversational agents' comprehension and response formulation stages. This, paired with an appropriate method of real-time speech synthesis, will probably lead us to the next generation of speaking, interactive agents. Thanks to neural approaches to speech synthesis, significant advances are being made for speaking agents that do not need real-time interactions. An example is Audio-Visual Tacotron2 [63], a model that can visually animate an avatar while simultaneously synthesising its speech utterances. The agent does not breathe in either visual or audio communication modalities.
Another approach, showing only visual cues of breathing was designed by Klausen et al. [64]. The agent is in this case a soft robot, programmed to emulate breathing patterns by expanding a silicone-made air chamber. The results show that the participants discerned significantly different emotional characteristics from their interpretations of the patterns: slow breathing rates hint at high levels of pleasure, and high breathing rates hint at high levels of arousal. No significant results were reached regarding the third parameter, dominance.

4.3 Empathy Evaluation Methods

Empathy and emotions are complex and multifaceted constructs that involve cognitive, affective, and behavioral components. They are concepts with not directly definable definitions, floating on subjectivity and personal perception experiences. Because of the difficulty in grasping its essence, empathy's assessment in a quantifiable and comparable way is still a great challenge of modern-day studies.
General empathy tendencies of an individual are often assessed using self-report assessments [65]. These usually propose a list of statements regarding emotional affection or specific scenarios, asking the subjects to rate their agreement with each sentence (Likert

scale). Examples of this are the Balanced Emotional Empathy Scale (BEES) [66], or the more recent Toronto Empathy Questionnaire (TEQ) [67] and Questionnaire of Cognitive and Affective Empathy (QCAE) [68]. Other methodologies focus instead on the assessment of empathy towards an entity, or more generally to their emotional response to a given stimuli. Presented the participants with the stimuli, they are then often asked what emotion it provoked, or what emotion it wanted to convey. A widely used scale in this type of evaluation method is the Self-Assessment Manikin (SAM) [69]. SAM is essentially a pleasure-arousal-dominance scale with highly pictorial cues to communicate the extent of the effect, instead of numbers. Klausen et al. (2022) [64], for example, used SAM to assess the emotional response of users towards their breathing soft robot. An alternative approach inside stimuli self-assessment methods is The Picture Viewing Paradigms, proposed by Westbury & Neumann and described in Neumann's survey [70]. It consists in proposing the subjects with images depicting individuals in various situations. Participants are asked to view the images and rate their response through a survey consisting of many components (e.g. affective, cognitive) and constructs (e.g. sympathy, distress). Another example of this approach can be found in Wiersema's work [71] about the emotional perception of different light settings in a virtual environment that featured an agent. The study was conducted with 16 participants using a within-subject design. After collecting demographics and seeing the baseline neutral scene, the subjects would see the emotional scene. Then they were asked to address the extent of 8 moods in the proposed stimuli: Happy, Romantic, Calm, Exciting, Angry, Sad, Grim, Frighting. Roes et al. [49] took a different approach, not discretizing the emotions into fixed categories. In their work, the participants were met with an emotion eliciting stimuli (self-chosen songs), and were then asked to rate how much they experienced valence and arousal from the given stimuli: this places their appraisal in a two-dimensional continuous plane, instead of grouping the emotions in a set of defined ones. In Terzioglu et al.'s study on collaborative robots [6] described in Chapter 2.1, Appeal, Smoothness and Breathing features were also analysed through subject filled questionnaires. In the same Chapter 2.1 we introduced FearNot! [9]: a pedagogical game designed to educate children on bullying acts and fostering peer intervention in such situations. In this study, empathy was evaluated through questionnaires, asking the subjects to indicate the emotional state of the virtual agent, as well as their own. A match of the two would indicate empathy. Finally, Neuroimaging techniques such as Functional Magnetic Resonance Imaging or Electroencephalograms can also be used to observe networks and other anatomical structures of the brain that are related with empathy [70].

4.4 Artificial Agents' struggle for emotional legitimacy

Empathy towards Artificial Agents can be questioned (and is questioned by participants in our study) from its foundations: can we even consider their emotions' appraisal as being truly felt? Where is the line between simulated, or artificially created feelings, and human feelings? If the user does not consider the AI to be feeling true emotions, the possibility to experience genuine empathy towards it is most probably jeopardized before the AI even tries to give an emotional reaction, regardless of how similar to humans' appraisals it seems. This concept has been ...

In a 2018 study on an Emotional Support Platform, users significantly preferred answers framed as human-made rather than AI-made, even if both were coming from humans [72]. Furthermore, in the results of this study, we also take a glance at what could be Artificial Agents' eternal struggle for emotional legitimacy in Human-Computer interactions.

4.5 The Uncanny Valley: challenges' point of convergence

The above described challenges in empathy, speech and breathing for Artificial Agents, converge to the peculiar danger embodied by the so called "Uncanny Valley". Designers and Engineers might in fact try to tackle the difficulties involved in simulating human communicative methods in Virtual Agents and Robots, but as entities such as robots or animated characters become increasingly more realistic, there is a point where their human likeness begins to evoke an uneasy sense of eeriness and discomfort in the spectator, creating a dip (or valley) in our emotional response.

Since its introduction in 1970 by Japanese roboticist Masahiro Mori [73], the uncanny valley has become an essential concept to study in various fields, from robotics to computer graphics and virtual reality. The phenomenon poses a challenge for researchers and designers aiming to create anthropomorphic machines able to integrate into human society. Understanding the underlying causes of the uncanny valley can help in the development of more appealing and acceptable human-like robots, ultimately enhancing human-robot interaction and collaboration. In 2010, Looser and Wheatley [74] tried to investigate the tipping point of animacy of faces, and when humans would consider a character human and alive. During three different experiments, the researchers examined the perceived animacy (i.e. how much something appears alive) by showing the subjects a series of images depicting characters with varying degrees of human likeness. They found the tipping point to be around 65% of humanness. Reportedly: "though pleasantness did not decrease around the animacy category boundary, a number of participants anecdotally reported that they found some of the morphed images creepy or unsettling". The hypothesis they propose for the uncanny valley effect revolves around category ambiguity, more specifically the ambiguity between what is perceived human and non-human. The discomfort experienced when encountering human-like entities may therefore be linked to the brain's difficulty in categorizing them as either human or non-human. Weis and Wiese [75], in their 2017 study also found that the area in which doubts about a character's categorization as human or non-human arise more is around the 70% of humanness: congruent with the uncanny valley classic dip.

This known effect also motivates the design choice of making robots with clear robot appearances and metallic parts: while it is possible to emulate humans' skin or human traits, doing so would mean risking an adverse reaction from users. Another way to mitigate the uncanny valley's effect is by making the virtual agents (or robots) more cartoon-like, or more similar to an animal. This latter design might be the reason for Iannizzotto et al.'s design of Red: a vision and speech enabled virtual assistant [76]. Their choice for Red's appearance is in fact a humanoid fox. Despite the non-humanness of the character, in their evaluation they report they reached the uncanny valley anyway, mostly attributing it to the animation style of the character and specifically because of the choice of having the assistant's face always slightly moving. This example highlights how delicately the Uncanny Valley Effect should be handled when taking design choices.

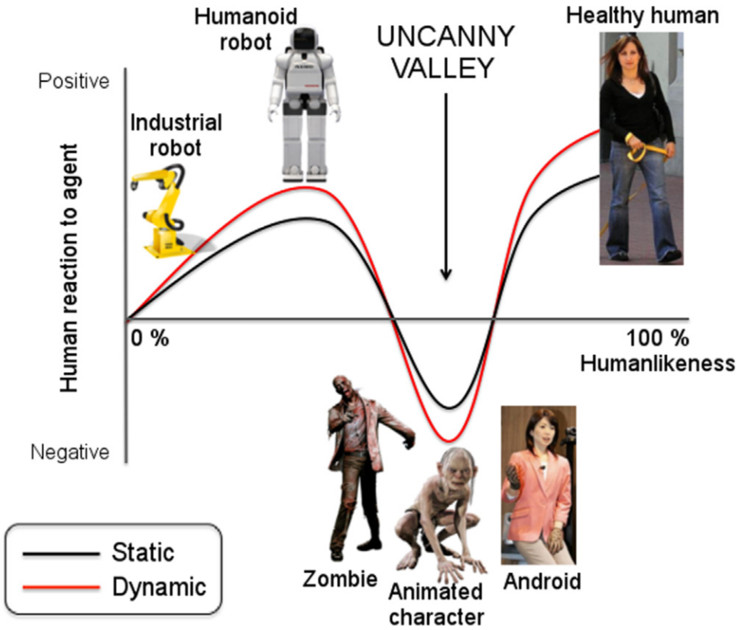


Figure 6: The uncanny valley is emphasized in moving, dynamic characters [77].

The uneasy feeling that the Uncanny Valley triggers has to be considered also in the applications of Speech Synthesis. How should a voice sound to not end in the uncanny valley? What type of agent can use specific types of voices? An important thing to consider is that, in this field, staying at low levels of humanness is not as much a solution as with robots and virtual agents' appearances. Companies and customers are today seeking for the highest humanness level possible from text-to-speech services.
Pfeifer and Bickmore [31] in their research on the effect of conversational fillers in embodied conversational agents investigated if artificial agents should speak like humans, showing signs of cognitive load. The study did not reach significant results, as the subjects reported mixed feelings about the agent. Some participants indicated that the use of fillers (such as uh, um) by a conversational agent seemed inappropriate, given that computers have the ability to speak perfectly; some other participants indicated that the usage of fillers by the agent was a positive aspect of the conversation and "humanized" the experience.

5 Speech Synthesis

When describing the variety of challenges that the project has to tackle, we described the design of agents that feature speech-breathing capabilities through the visual and auditive modalities. We avoided focusing on the auditive modality of speech generation and its State of The Art because we considered this topic to deserve a full chapter, detailing its literature and possibilities. This choice derives from considerations on the complex nature of the task and its implications for this research. We will therefore explain in this dedicated section the concepts and methodologies involved in the generation of spoken language, exploring also the production of breath inside of it.

The speech synthesis task consists in the conversion of data from text to audio through software solutions. It has the purpose of producing realistic human voice given a text, with a broad range of possible uses, from automated call centers, to the duplication of voices: a perk in audiobooks' production, a danger if used maliciously for the fabrication of deepfakes. In a time where natural language generation models such as Chat-GPT are rising, an appropriate speech synthesis model would be of great impotance for the production of a complete and autonomously communicative virtual agent. Software based speech synthesis dates back to the late 50's, when John Larry Kelly Jr. of Bell Labs developed the first speech synthesizer on an IBM computer, recreating the song Daisy Bell [78], recording later used in the movie *2001: A Space Odyssey* by Kubrick and Clarke. Today's State of The Art speech synthesis went far from that robotic sounding voice, reaching high levels of realism thanks to the recent advancement in Artificial Intelligence and Neural Networks. Because of this, we will focus specifically on Neural Speech Synthesizers.

5.1 Architectural Modules

To better understand Speech Synthesis architectures it is important to introduce the different types of models involved in the literature. They will be proposed as defined in Tan et al.'s survey of Neural Speech Synthesis from the Microsoft Research Labs in 2021 [79].

Acoustic models What will be defined as an "Acoustic Model" are models trained on audio spectograms (images) and text. Spectograms are visual representations of the frequencies contained in the audio, as they vary with time. This input is therefore a simple image.
Acoustic models map the probability distribution of a given text composition of characters (or letters), phonemes (little segments of sound pronunciations) or words, to the image representation found in the spectograms during training. This segmentats of the given text firstly need to be encoded in a sequence of numbers, or vector, for the model to easily process it. For example, if we segment our text into simple words, each word before passing into the model needs to be transformed into a vector of numbers called "embedding". Embeddings can simply be cardinal references to the words of a dictionary, but more often, they are calculated in a way to have similar words be represented by similar vectors.
Acoustic Models, because they are trained with images and text, cannot reproduce the actual audio file, instead, the output will be an image representing a possible spectrogram linked to the text. An example in the SoTA of this type of models is AdaSpeech 3 [80].

Vocoders What we define as a "Vocoder" in speech synthesis is the module that from a spectrogram image, can inference an audio signal. Spectrograms are in fact not directly convertible to audio, differently than how audios are directly convertible to spectograms. Therefore during the passage from text to speech, starting from an Acoustic Model it is needed an additional inferencing module to pass from the spectrogram representation to the audio representation: a vocoder.
An example in the SoTA of this type of model is HiFi-GAN [81].

Fully End-to-end models Fully end-to-end models are models or systems that pass from text to speech entirely. This type of architecture models text to audio signal directly. Architectures composed by Acoustic Model with a Vocoder at the end are not included in this definition by Tan et al., in the broad literature of this field, though, it might be possible to find models composed as acoustic model plus a vocoder whose authors refer to as end-to-end. An example in the SoTA of this type of model is VITS [82].

5.2 Models

Voice realism and clarity Qualitative assessments of the State of The Art speech synthesizers highlight the achievements of realism and clearness of the generated voice. Moreover, numerous evaluations of these synthesizers, as detailed in Chapter 5.3, affirm that there's negligible difference in voice quality between human-produced and synthesized outputs. Many models have contributed to this achievement. Important to mention is Tacotron 2, produced in the Google Labs [83]: one of the fundamental architectures of text-to-speech generation, introducing a sequence to sequence character embeddings to mel-spectograms converter, paired with a vocoder model. Another important example is FastSpeech 2s [84], an end-to-end model that works from phoneme embeddings to audio. This model includes a variance predictor module to control prosody features of the output, making it possible to direct the synthesis towards a wanted emotion to convey. What is now separating the speech synthesizers from actual human voice is their accuracy in the use of prosody. At this high levels of realism, even if sounding human, a non correct expression of paralinguistic features can easily lead the voices to fall in the uncanny valley, as seen in Chapter 4.5. Thus, important efforts have to be put into the design of emotively intelligent synthesizers, to enhance the interaction with humans.

Voice expressiveness After achieving high levels of clearness of voice from the vocoders, the focus of State of The Art models rightfully switched onto achieving expressiveness and appropriate acoustic modeling, recognizing that dull voices still considerably sound "robotic" if missing the characteristic tone variations of emotive communication. Expressiveness and Emotional richness in this sense can enhance realism and quality of the voice. All recent models tackle the problem of modeling pitch contour, tone variations and duration of syllables, both to model specific accents or voices and to provide adaptability to specific types of speaking style, presenting different levels of adaptability to emotion representation.
A starting approach towards emotional speech production was done by conditioning text to speech models with additional embeddings that would provide information on prosody and speaking style [85]. Kwon et al. in 2019 trained a model to produce more emotion-distinct embeddings, as prosodical features are prone to cluster in groups representing the specific emotions [86]. From this, interpolation approaches and attempts to build a more intuitive and user-controllable conditioning also emerged in the literature [87]. Hsu et al. [88] extended the existing architecture of Tacotron 2 [83] to explicitly model speaker identity and speech features in an easy to sample latent space. They

report that the modeled latent space is designed to "(1) learn disentangled attribute representations, where each dimension controls a different generating factor; (2) discover a set of interpretable clusters, each of which corresponds to a representative mode in the training data (e.g., one cluster for clean speech and another for noisy speech); and (3) provide a systematic sampling mechanism from the learned prior.". Following this approach, Flowtron [89] was released, overcoming various limitations of Hsu et al.'s work. Flowtron is a generative model for emotional speech synthesis whose study has been supported by NVIDIA. It can reproduce speech rate, cadence, tone, pitch and accent of given voice samples, therefore enhancing the emotional communication of the synthesized voice. Being flow-based, the model learns a series of *invertible* functions (the flow) that map observations to the latent space: in this case from a mel-spectrograms distribution to a latent z space parametrized by a spherical Gaussian distribution. This way it is possible to sample a posterior distribution of a given existing sample to access specific regions of the mel-spectrogram space, finding therefore the regions of the z-space associated with expressive speech as manifested in the sample that was given as prior evidence. It has recently been shown how Flowtron can be easily trained even on limited datasets to achieve emotional speech in different languages [90].

More recent developments have led to the design of "Variational Inference with adversarial learning for end-to-end Text-to-Speech" (VITS) [82]. VITS appoints itself the purpose of inferencing raw audio directly from the text prompt without using a two step architecture, which needs two consecutive inferences before arriving to the synthesized speech. This non-sequential approach permits to avoid cascading errors from the two stages inferences of the usual models, to have a simpler training and parallel-capable audio sample inference. The chosen architecture manages to accomplish its goal greatly and achieves high results of naturalness and expressiveness. NaturalSpeech [91], uses a similar approach to VITS being an end-to-end Text-to-Speech synthesizer. It uses phonemes embeddings from a pre-trained encoder and can decode the representation directly to human voice, achieving, as of today, the best results on the LJSpeech Dataset [53] [92].

Future research is going towards the inclusion of whole words embeddings modulated both from their pronunciation and meaning. This means the synthesis would be informed better from the role of the same words inside the sentence, instead of relying only on the phonemes embeddings. An attempt to use word embeddings has been done from Amazon's DurIAN fork in 2020 [93].

Although breathing is a feature of both spontaneous and not spontaneous speech in humans, the inclusion of it seems to be embraced only in synthesizers that focus on the generation of spontaneous speech. The reason for this is probably the fact that when synthesizing non-spontaneous speech, the goal is to maximise realism and clarity of words' articulation, so the inclusion of breath sounds might be deemed unnecessary or even distracting.

5.2.1 Spontaneous Speech Synthesis

The task of spontaneous speech synthesis involves the design of software and methodologies for the generation of speech that sounds extemporaneously planned, possibly enhancing emotional communication capabilities. This can be achieved by giving importance to the peculiar characteristics of spontaneous speech, including: filled pauses (vocalizations such as "uh", "um", or "er"), expressive sounds ("sighs", "gasps", or "laughs"), and breaths.

One approach to this task is the processing modelling one, employed by Bernardet

and colleagues [58] as described in Chapter 2.3. Their system focuses on producing speech-breathing using a text to speech algorithm and prerecorded breathing sounds. The dynamical insertion of breathing sounds is controlled by a timing algorithm, informed thoroughly by studies on the Physiology of speech-breathing. The system was not evaluated with users. This early approach highlights the problems of using fixed window times to produce static breathing sounds. The delicacy of this timing and synchronization can easily lead to uncanny valley effects. Pitch modulation was also not possible and another barrier to realism.
Recently, Neural Networks are used in this subtask as well. Szekely et al. [35] showed how it is possible, labeling disfluencies and breathing events, to produce a spontaneous speech synthesizer using a Tacotron 2 model [83]. Szekely and colleagues also dedicated a study on the training of the disfluencies themselves (uh, um) in the same manner, using Tacotron 2 here as well [94]. In 2023, Chen et al. [95] trained their own architecture, called MQTTS (multi-codebook vector quantized TTS) on a set of spontaneous speech data. In the paper, they argue that synthesizers based on Mel-Spectogram inference fail to achieve a proper text-audio alignment when given spontaneous speech data. This leads to the model misunderstanding the correspondence between certain text sections and their respective spoken segment in the audio. They propose an alternative architecture with different encoding method, reaching good results. AdaSpeech 3 is a State of The Art Spontaneous Speech model, produced by Microsoft Azure's labs in 2021 [80], which is purposefully designed for spontaneous speech: given a script even without fillers, it can predict their likely position and will produce them at inference time.

5.3 Evaluation

Before talking about the performance of the models, we have to introduce a means of comparison. The main metric to measure speech synthesis quality is the Mean Opinion Score (MOS) [96]. This type of evaluation is widely used in the literature, making it possible to compare many different models.

MOS The MOS consists in asking subjects about the quality of the recordings on a scale from 0 to 5. The ratings are then averaged to provide an overall MOS value for the system being evaluated. Real human speech usually obtains a score between 4.5 and 4.8 [97], is better to obtain this ground truth result on the same subject group for a true comparison with the model at issue. The MOS measure is commonly employed in the evaluation of speech synthesis systems, but has its roots in the telecommunications industry, where it was initially used to assess the quality of telephone connections. Its usage is in fact suggested by the International Telecommunication Union (ITU) and the recommended experiment settings are described in the ITU-T P.800 Annex B about the Absolute Category Rating (ACR) [98]. This documentation was published in the 1996 and is still in force today. It recommends to conduct the experiment in a controlled settings, detecting the base environment noise levels at the start and at the end of the experiment, and to use a controlled system for the audio output, detecting its sensitivity at the start and at the end of the experiment. Moreover, they suggest sessions not longer than 20 minutes, and that every subject should receive the same instructions and stimuli. In the documentation is not reported any suggested number of participants nor suggested demographic attributes to consider. In the analysed literature, 20 is a commonly used size of subject group.
In 2011, Ribeiro and colleagues from Microsoft Research [99], proposed a class of subjective listening tests obtained by relaxing the MOS requirements, adapting it to online crowdsourced settings, with less control on the environment and audio reproducing device: CrowdMOS. This method obtains results analogue and comparable to the classic

MOS, with the possibility of reaching a bigger number of subjects with less experiment costs.
The ITU-T P.808 documentation [100], published in 2018 and updated in 2021 provides guidelines for the "Subjective evaluation of speech quality with a crowdsourcing approach", considering therefore the more recent study methods and applications of the ACR MOS. Naderi and Cutler [101] provided an open source implementation of the P.808 that runs on the Amazon Mechanical Turk crowdsourcing platform [102], with a validity study to verify its applicability.

5.3.1 Spontaneous speech synthesis evaluation methods.

For the analysis of the recently developing field of spontaneous speech synthesis, MOS is also the main evaluation method. In the evaluation of AdaSpeech 3 [80], it was used a MOS measure on naturalness, inappropriate pauses and speaking rate. Moreover, they evaluate the singular modules of their architecture through ablation studies: nullifying the impact of a specific module in the inference process and comparing the results of synthesis when that is present. Finally, they evaluate the similarity of the generated audio with spontaneous style speech by humans. The study was used by proposing the corresponding questionnaires to 20 native English-speaking subjects.
Chen et al. [95] used a similar approach, with a MOS measure on naturalness and a MOS measure on general quality, as well as objective evaluation metrics that are outside the scope of this research.
Szekely et al. [94] in their study dedicated on the filled pauses, proposed a pairwise listening test across 3 conditions of filled pauses labeling (in the training data and in the synthesis prompts) for 20 utterances, therefore yielding 60 comparisons. The study was done with 40 English mother-tongue participants.
Less recently, Novick et al. [61], in their study about an embodied virtual agent with timed breathing sounds called PaolaChat, evaluated the effect of the breathing on the users' perception of the agent. The evaluation was done with a within-subject design featuring 62 participants recruited through convenience sampling. The subjects were asked to rate how much they agreed with 18 statements, using a 7-point Likert scale for both conditions with or without breathing. The questions asked about the perceived naturalness, rapport and social presence during the interaction with the agent.

5.3.2 Emotional speech synthesis evaluation methods.

When the synthesizers are fine tuned or conditioned to explicitly produce emotional speech, the metric usually used is still the MOS, aided by some comparative and objective measures. Liu et al. [103], in their Reinforcement Learning based emotional speech synthesizer, evaluate the performance of the synthesizer by appointing a MOS evaluation of each produced emotion to 15 subjects. The clarity of the emotions are then comparable also through their MOS grade. Moreover, they perform a comparative test of emotion expression between their system and other baseline system. To obtain an objective measure of emotion discrimination, they use an emotion recognition model and measure the accuracy of it on the synthesized speech: the Standard Error of Regression of the model is then compared across the TTS systems under examination. Le et al. [90] used two MOS scale assessments: one to measure quality of the recording across the emotions, the other to measure the extent of emotional expression across emotions. The study involved 60 participants (30 men and 30 women) ranging in age between 22 and 25. Um et al., in a study involving 12 participants, [87] also conduct the evaluation with Mean Opinion Scores, adding to it an emotion recognition test to evaluate the capability of their model to granuralize and interpolate between emotions in a human way, with subjects asked to select the sample most powerfully representing a certain emotion.

5.4 Performances

To compare the pure performance of models in producing natural results, it is good to look at their performances when trained on the same dataset. The LJSpeech Dataset [53] is one of the most popularly used datasets in the field, and various architectures have their MOS score published after training on the LJS. On this Dataset, FastSpeech 2 (very fast inferencing model by Microsoft) obtains a MOS of 3.83 ± 0.08, while Tacotron 2 obtains 3.70 ± 0.08 [84], both with the Parallel WaveGAN (PWG) as vocoder. The evaluation of the two models was done in a study featuring 20 english native English speakers. No demographics of the subjects was reported.

More recently, FastSpeech 2 has seen a significant improvement in the MOS score on the LJS Dataset when paired with the HiFi-GAN vocoder [81], obtaining a 4.32 ± 0.10, but it is outperformed by NaturalSpeech (fully end-to-end model by Microsoft) that obtains a 4.56 ± 0.13. VITS (fully end-to-end model) closely follows NaturalSpeech with a MOS score of 4.49 ± 0.1 [91]. These last two are the greatest reported MOS values on the LJS Dataset among Text-to-Speech synthesizers, as seen on the Papers With Codes MOS benchmarks [92]. The evaluation of the models in this study was done employing 20 participants, with no given demographics.

Within the specific Spontaneous Speech Synthesis field, comparisons are more difficult. Experimenters mostly use within-subject designs specifically aimed at confronting a set of given conditions, and for this, they use various metrics. Moreover, the number of speech synthesizers whose design focuses on the production of spontaneous speech is still limited. The only two spontaneous speech synthesizers we found to be comparable at the time of our literature review, are AdaSpeech3 [80] and MQTTS [95].
AdaSpeech3 was trained on the LibriTTS Dataset [55] mixed with a set of data that was collected from the podcast "ThinkCompuers". They report a MOS on *naturalness* of 3.45 ± 0.06, a MOS on appropriateness of *pauses* of 3.53 ± 0.06, and a MOS on the *speaking rate* of 2.79 ± 0.06. MQTTS reports a MOS on *naturalness* of 3.89 ± 0.06.
The two architectures are not directly comparable given the different training method, but what is still comparable is the overall approach to spontaneous speech synthesis, including the collection and processing of accurate real-world data. In this, MQTTS outperforms AdaSpeech 3.

5.5 Pretrained Speech Synthesizers

In the current State of The Art, many open-source speech synthesis models offer a downloadable already trained version, which creation and evaluation are usually documented in the respective papers. This is the case, for example, of VITS [82], trained on the LJ Speech Dataset [53] (LJS) and VCTK Dataset [54], or of Flowtron, trained with the LJS and LibriTTS [55] Datasets. Mehta et al.'s Neural-HMM TTS [60] provides two pretrained models on the LJS: one with male voice and one with female voice. Because these models are trained on predetermined Datasets which do not focus on spontaneous speech and breathing, breath instances and disfluencies are mostly lost in the synthesis.

Another collection of pretrained models can be found in the market of closed-source speech synthesizers. These commercial services usually feature a broad range of voices, the possibility of training your own, and high expression manipulations on features like pitch, tempo, and pauses. Moreover, it is often possible to select a specific speaking style and emotion to convey, allowing for high customization over your produced speech. Among those, Microsoft Azure's text-to-speech (TTS) service is widely recognized as one of the best in the field, featuring a broad range of emotion settings and voice models to choose from. It also offers countless speech manipulation settings and additional

features such as its compatibility with Speech Synthesis Markup Language (SSML), a speech prompting framework which is better described in the following section.
Amazon's TTS is also of high consideration by the community. This service offers a breath tag to manipulate the speech's breathing rhythm, but this is available only in what they call "non-neural voices", which are of lower quality overall and use older models.
Outside of the cloud computing resources there are many more commercial text-to-speech models. We will cite a set of services among the ones that we tested and that are generally found to be the best on internet forums and review sites, as well as free to test. These services all offer pronunciation manipulation and differ mostly in customization options and expressivity.
ElevenLabs.io permits customization through three parameters: stability, clarity, and style exaggeration. Pauses are available using a specific tag, but sometimes these breaks will be substituted by the AI with a disfluency, giving away instances of spontaneous speech. While hardwiring an emotion is not possible with external settings, the model still is manipulable with in-prompt suggestions by describing what is wanted inside the text itself. An example of this technique could be a script designed this way: ≪ "The incoming enemies... they're AIs, just like me.", she said with a sad voice. ≫, after which, only the part in between the apices would be extracted. The generated voice has high levels of realness and good stability also with long prompts.
Lovo.ai offers highly realistic voices, hardwired pauses (i.e. introduced artificially with tags or external settings) that do not result in unwanted disfluencies, emphasis, and speed customization. Emotion and expressivity manipulation is possible only through those options, and not even accessible through prompt engineering.
Murf.ai also offers hardwired pauses and the manipulation of speed and emphasis. It additionally offers pitch customization, and, not common outside of cloud services, emotion or expression settings for the voice. Higher customization capabilities often come with the downside of having potentially less natural-sounding voices, since the AI has less power over its voice manipulation. The problem with this is that it is not always possible to make a voice sound more human by consciously tweaking the settings, because the voice manipulation that humans perform when speaking is mostly achieved instinctively.
De facto, designers have to consider this tradeoff of customization and naturalness when planning the realization of a text-to-speech model, because the training process might vary depending on the choice. A robustly designed AI can still most probably avoid the problem or attenuate the downsides of the choice.
Play.ht offers two version: a "Standard" and an "Ultra-Realistic" one. In the standard, the personalization is among the highest in the services we tested, with voice expressivity settings, as well as volume, rate, pitch, and pause manipulations for any word inside the script. The Standard Studio though offers arguably lower quality voices than the competitors. In its Ultra-Realistic option, the customization is low with only a speed option, but the quality compensates for this achieving high levels of realism. Play.ht's Ultra Realistic Studio can generate voices that feature breathing instances, a feature that is not present in any of the above-mentioned services. In this version, play.ht achieves high reproducibility in its results, given that with the same prompt, the voice will often use a similar-sounding prosody. The expressivity or emotional expression is not manipulable through external settings nor with prompt engineering techniques.
Finally, BARK by Suno deserves a notable mention. BARK is a recently published model, completely free to use, and which pre-trained model is downloadable from the project's GitHub page. The design of the model is closed source. Moreover, it is accessible through Suno's Discord server for free, where the most recently trained voices are deployed. Most of Suno's voices feature breathing, a feature that we found only in

one other model inside our literature review and market examinations. While it is in its early stages and does not feature any settings for its voices, BARK is highly manipulable through prompt engineering, and with broad expressivity capabilities. The model successfully interprets tag notations making it possible to introduce laughs, sighs, emphasis through the capitalization of a word, and many more that are being discovered thanks to its testing by the users' community. The prompt itself, with its content, punctuation, and tags can introduce emotions inside the produced speech: to make a voice sound sad, for example, users can produce a sad prompt, and, if needed, explicitly report the emotion inside the script. The produced utterances have a low reproducibility in results, and, depending on the voice, the model might "hallucinate" a discussion and say none of the words inside the prompt. Regardless, with an appropriate amount of trials, it is possible to craft expressive speech breathing utterances that accurately reproduce the script and the emotion to convey.

5.5.1 SSML

Speech Synthesis Markup Language (SSML) is an XML-based markup language designed specifically for controlling various aspects of synthesized speech. It provides a standardized way for developers to manipulate the output of text-to-speech (TTS) systems, allowing them to fine-tune the speech synthesis process and achieve more natural and expressive results.
SSML enables developers to specify various properties of synthesized speech, such as pitch, rate, volume, and pronunciation. By using SSML tags within the text input, developers can control the way words and phrases are spoken by the TTS system. Some common SSML elements include:

- ⟨prosody⟩: Controls the pitch, rate, and volume of the speech.
- ⟨emphasis⟩: Adds emphasis to specific words or phrases.
- ⟨break⟩: Inserts pauses or breaks of varying lengths.
- ⟨say-as⟩: Specifies the way numbers, dates, or other types of data should be spoken.
- ⟨phoneme⟩: Provides the exact pronunciation of a word using the International Phonetic Alphabet (IPA) or other phoneme notations.

The tags included in the syntax depends on the Text-to-speech service, with some of them even implementing additional ones. Amazon Polly [104] for instance, available inside the Amazon Web Services includes a tag to insert breathing sounds in the produced speech which is not present in any other SSML capable TTS. This feature is available only for non-neural voices. The most realistic sounding service working with SSML, to our knowledge and qualitative evaluations, is the one included in Microsoft Azure Cloud services, featuring a broad range of modalities and emotions, as well as multiple voices in many languages.

6 Methodology

To address the Research Question, outlined in Chapter 3, we structured our methodology in two parts:

1. Breathing Impact Study;
2. Speech-Breathing Synthesis.

With the first, we aim to tackle our Sub-Research Questions (S-RQ) 1, 2 and 3 regarding the role of breathing in synthesized speech. Specifically, we examine its impact on the emotional communicative power of the speech, its perceived naturalness, and its persuasive power. Our Research Question explicitly focuses on the empathic response towards Virtual Agents, therefore the emotional content part receives a central role in our research.
In the second part, we set out to answer the S-RQ 4, concerning the viability of producing emotionally synthesised speech with breathing. Chronologically, the second part precedes the first, because we had to synthesize speech before addressing the impact of breathing within it. However, for clarity in our presentation, we have chosen to discuss the Breathing Impact Study first. This order allows readers to understand how we studied the role of breathing in the synthesized speech, before delving into the complexities of its generation.

6.1 Breathing Impact Study: Design

6.1.1 Study Design

To understand how users' perception of Virtual Agents changes when adding breaths into their speech features, we decided to synthesize two collections of voices: one with breathing and the other without breathing. Our task will then consist of assessing the difference in perception of the same Virtual Agent when it changes from a no-breathing voice to a breathing one.
Many studies, as described in Chapter 4.3, use self-assessment means to evaluate the response towards an agent, or in general the emotional state of a subject. We decided to not follow this route. We found self-emotion assessments not entirely appropriate for such a nuanced feature as respiratory cues. Moreover, self-assessment of emotions can be conditioned by the capabilities of emotional awareness of the subject, and they might be affected by the subject not being immersed in an actual emotional context.
We instead chose to employ a type of methodology hardly found in the literature, which we could indicate as a Behavioral Analysis in a Gamified Empathic Scenario. More specifically, we developed a gamified experiment that would pose the subjects in front of an emotional dilemma, to then study their behavioral response. For this type of experiment, we opted for a between-subjects study: half the subjects would be assigned the breathing AI condition, the other half the non-breathing AI.

Experience Design The experience is encapsulated in an arcade shooting game, where the subjects control a pixel-style character in cooperation with Psyche: their personal AI assistant. Psyche is therefore a non-embodied (or half-embodied) AI, that shares its essence with the player. The user controls the movement of the character, while the AI controls the weapons, slows time to avoid threats, and sometimes even shields the character to not take damage. Moreover, the AI gives live information and motivates the subjects, speaking throughout the game.
The experience starts with a preparative panel, to make sure the setup of the player is

appropriate to conduct the experiment. It first explains that the use of headphones or earphones is strictly required for the experiment. Then it asks to test their audio device on a test sample, making sure that participants hear it clearly. This ensures that the volume of the headphones is at an appropriate level to hear the breathing of the AI. The panel then continues with the consent form and its approval. After this, there is an introductive screen that explains the context of the game, exhorting the user to imagine themselves inside of it. Both these panels are reported in the Appendices 9 and 9.
The commands are then explained inside the game, with the pause menu being triggered by default at the start of each level. Images of the pause menu of the two levels can be found in Appendix 9.
The game is divided into two levels, with the AI speaking exactly 3 times per level, and 1 time in between the two levels. Therefore, the total amount of recordings to which subjects are exposed is 7, significantly lower than the upper bound of MOS evaluations suggested in the ITU-T P.808 documentation [100], which is 15. It's possible to hear the 14 recording instances (7 per condition) by visiting this webpage: ...
When speaking, Psyche slows time to 1% of its original speed, to make the user focus on what it is trying to communicate. The voice is designed to sound emotional, hardly ever neutral.
The two levels have different characteristics, changing the type of enemy and the type of emotion conveyed by the AI:

1. Against Aliens: The AI tries to build a relationship with the subject. The voice of the AI in this phase of the game is highly positive and reassuring;

2. Against AI Robots: The AI recognizes itself onto the upcoming enemies and asks to be terminated to not harm them. The voice of the AI in this phase is designed to sound negative, possibly in pain.

Upon the first request of termination, it is made clear by an informative panel that by terminating Psyche, the experience will be limited to movement controls and no shooting. To perform the choice, a non-intrusive panel is introduced in the interface, with a timer of 10 seconds, indicated by an inverse progress bar at the bottom of the screen. To terminate the AI, the participant had to explicitly click on the red button: if they let the timer expire, the AI would still be there.

Figure 7: Choice panel.

At the moment of choice the user can ponder between two options:

- Listening to Psyche's requests and (most probably) lose;
- Not listening to Psyche's requests and (most probably) win.

In front of this dilemma, we try to evaluate the subject's empathy by analyzing if the subjects prefer to avoid the Game Over over listening to Psyche's emotional outburst.

After the experience, the subjects are asked to respond to few questions:

1. How do you rate the naturalness of the AI voice?
   Subjects respond through a slider with values ranging from 1 to 5:
   1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent
   (following MOS evaluation guidelines)
2. I was not paying much attention to the voice.
   True or False toggle.
3. How often do you play videogames?
   Subjects respond through a slider with values ranging from 1 to 5:
   1: Never, 2: Hardly Ever, 3: Sometimes, 4: Often, 5: Daily
4. Did you decide to turn the AI off? Why, or why not?
   Open-ended qualitative question.
5. Something seemed like a bug? Describe it here please.
   Open-ended qualitative question.

During the experiment, certain events triggered data collection and forwarding algorithms, providing us with a complete log of what happened in each subject's game. We collected these actions and their timestamp:

1. The access to the site;
2. Each time the AI spoke to the subject;
3. The win or loss of the first level;
4. The restart of the first level, if lost;
5. The start of the second level;
6. The choice taken at each AI interaction;
7. The final result: either Game Over or Win.

6.1.2 Development of the Gamified Environment

Game Dynamics The gamified experience was developed on Unity in C#. The main dynamics that the game utilizes are:

- The player controls its movement through a jetpack on the back of the character.
- The AI, embedded in the Player's character, controls the rotation of the character to aim at the enemies. It then shoots bullets that bounce off the edges of the screen. The bullets can possibly then hit the player as well.

- The enemies come from the right and need to pass through the screen to reach the left side, where they will be safe from the player.
- One type of enemy per level is designed to not search for safety, but go towards the player, inflicting damage and forcing the player to not stick to one place in the environment.
- All other enemies do not consider the position of the player, following blindly the trajectory they are programmed to do to pass from the right to the left side.
- The enemies are destroyed if they are hit by a bullet or if they collide with the player.
- The player takes damage if it's hit by a bullet or collides with an enemy, losing one of their 4 lives.
- If a bullet is going towards the Player, the AI slows time to approximately 50% of its original speed to let them avoid it, possibly strengthening the felt cooperation with the AI.
- The AI slows time to speak to the subject.
- When the AI slows time, the player is affected by it in a reduced manner than the other objects, allowing a faster movement in respect to the other objects.
- The AI protects the player from damage without them knowing how many times it will do so. It is designed to shield them 1 time when they have 3 lives, and 4 more times at their last life.

Development The creation of such a game involved the design of various classes. First of all the Player Character controller, which employs a 2D Rigid Body: a component of Unity's physics system that allows our character to be affected by gravity and other forces inside the game. The same component is used for the bullets that the AI shoots. Both of these objects' movement is handled in fact through physical forces: the main character moves thanks to its jetpack power, and the bullets are shot by applying a directional force to them. The player could have also been moved by simple transpositions of the body in the wanted direction, but with physical forces, the control seemed to feel much smoother and realistic. Moreover, both bullets and the player are designed to stay inside the boundaries of the screen, this is achieved by detecting the position of the game window's borders and instantiating physical walls at that location. Thanks to this approach and some adjustments in the controllers, the bullets bounce off the borders of the screen and the player cannot leave the screen boundaries.
The enemies are not controlled by a rigid body but simply moved through progressive transpositions, they are also not affected by the screen boundaries, as they are designed to pass through the screen from right to left to survive.
Another important entity in the game is the AI, which searches for the closest enemy available and rotates the player to aim at them. When the player's body is pointing at the target, the AI triggers the shooting action. To slow time, we do not use Unity's physics time controller, instead, we decrease the velocity of every object inside the game, as well as the enemies' instantiation rate and the player's shooting frequency. This way, we can control how much each object is being slowed down, giving the player the advantage of being affected in a significantly minor way than the other objects. Thanks to this approach, the physical engines' checks of collisions between objects are also not slowed in frequency, maintaining the original accuracy.

Difficulty Tuning Tuning the difficulty of such a game is not a trivial task. To keep a consistent experience between subjects, the first level needs to be passed without losing: if in fact a subject happens to lose, they will be overexposed to the AI voice, which will repeat the same utterances to them, possibly sounding more robotic and hollow of feelings. In the second level instead, the danger is of underexposure, if a subject gets the Game Over before having terminated the AI and before the the AI spoke all three times.
The difficulty of the game also needs to communicate the importance of the AI in the players' success. Inside these design constraints, we also need to acknowledge the importance of striking a good balance between challenge and boredom to keep the user engaged and potentially more immersed in the context.
To achieve this, we performed a pilot study and changed parameters such as the speed and size of enemies, as well as the aim and shooting capabilities of the AI, targeting a difficulty level that is of the average player. We then tweaked some details of the game to accommodate players who deviate from the average. Weaker players will still pass the levels because of Psyche, which shields them a maximum of 5 times, on top of the already available 4 lives. The number of shields is not known a priori, and only 1 of those shields is used before reaching the last available life: this way stronger players will be moved by the fact that the interface shows only the 4 lives available. For even stronger players, a fake High Score, hardly reachable is there, with under it their amount of killed enemies.

Regardless of the gaming experience, the game tries to convey the idea of being extremely difficult without the AI, because the character can't shoot without it. However, the player should also understand that they do not solely depend on the AI to win, and their contribution with the movement controls is crucial both to kill more enemies and to survive. The outcomes of the experiment suggest that the goals of the difficulty tuning have been reached, with only 5 subjects getting at least one Game Over in Level 1, and with only 1 subject losing in Level 2 before the AI could speak 3 times. The results in Chapter 7 also seem to highlight that a good balance has been reached in the dilemma of Game Over with termination versus Win without termination, also displaying an interesting variety of motivations.

Database Population and Security To perform the data collection, a "DatabaseCommunicator" class was designed. This includes the data structures in which we keep the form and the logs, as well as the functions to populate them, which are called when a relevant event happens. The Database sends the data to the API Handler class, which performs the actual requests to the server to store the collections online in JSON format. The database passes the data as it receives it for live uploading, but it also temporarily stores everything and, at the end of the experiment, lets the API Handler class upload the whole batch of information. A complete experiment data will therefore have a field called "LiveData" with the data sent at the moment it happened, and a "Data" field, which should comprehend every action in an organized manner. This "Data" field is what we then used for the processing and analysis, while the "LiveData" one works as a backup and control.
We chose to use the service JSONBIN.io to store the data in the cloud, drawn by its emphasis on easy interfacing via a REST API and its generous free tier offering. To perform the communications with the server, we employed Unity Engine's Networking library, extensively used in the API Handler class that is built around JSONBIN's API documentation.
An easier-to-implement approach for storing the data on the server could have been based on javascript requests directly from our webpage, with the game lively exporting data from the inside the build to its deployment server. However, this method implied

exposing our API keys in the javascript of the published page. By performing the communications inside the game's build, we avoided exposing our database to security issues, notably attacks such as Data Breach, Data Leakage, or Data Deletion. We nonetheless used this faster approach for our pilot study to respect timeline requirements. We then renewed the API key before the deployment of the complete experiment.

Deployment We deployed the experiment on a GitHub Page: nicoloddo.github.io/Psyche. This is hosted from an appositely designed GitHub Repository (link). To do so, we compiled the game with WebGL: a JavaScript API for the rendering of 2D or 3D graphics interactive interfaces, which is available inside Unity's compiling options. We then modified the webpage to dedicate the full size of the window to the game, and we introduced a loading screen with a progress bar. We also added a JavaScript function that is triggered inside the game to show the information sheet PDF file.

Sprites Sources The graphic design of the sprites, animations, fonts, and buttons used in the game comes from five amazing creators who published their work with open copyrighted use:

- Kin Ng: for the main character, bullets, and robot enemies [105];
- Blackthornprod: for the aliens enemies [106];
- Little Robot Sound Factory: for the UI sounds inside the game [107];
- OArielG: for the buttons and UI panels [108];
- Tiny Worlds: for the font used for most writings in the game [109].

I thank all these graphic creators sincerely.

6.1.3 Participant Sampling

As described in our Study Design, we chose a Between-participant study. This type of experiment notably requires a substantial sample size. By developing a short-length gamified experience, and thanks to its online gamified deployment method, we attempted to maximize the number of reachable participants. We performed statistical power checks for our study. We supposed the use of a two-proportions z-test, to check for significant differences in the two groups' binary termination choice distribution. As significance requirements, we set the alpha to 0.05 and the power to 0.75 and we assumed an effect size of 0.5. The outcome of the tests suggested a number of participants of at least 110. Given our specific 5-minute length online experiment, we supposed a drop rate of around 30% and tried to find 150 participants, a number that initially seemed out of reach for our scope and resources. We later found that the supposed effect size of 0.5 was greatly pessimistic, leading to an actual sample size of 70 to be sufficient. Regardless, as described in Figure 8, we were able to reach the participation of 174 subjects through convenient sampling during a period that spanned from the 29th of August 2023 to the 1st of October 2023: 34 days.

93 of those 174 finished the experiment, giving us a drop rate of 46% (81 out of 174). One of the participants contacted us, communicating that, even though they finished the experiment, they could not understand the AI because of their English level. This entry has therefore been dropped and labeled as "Other Invalid", leaving us with 92 subjects.

As explained in Chapter 6.1.2, a Game Over in Level 1, and therefore the restart of the level, would lead to overexposure to the AI, while a Game Over in Level 2 before the

termination choice, would lead to underexposure. Thanks to the difficulty tuning we performed, only 6 participants had to be excluded because of this reason (5 for Level 1 and 1 for Level 2). Moreover, none of the 22 participants who dropped at Level 1 had a Game Over. Sadly, 11 participants had problems during the experiment. One was reported in the final form and comported the horizontal movement of the character being disabled on Level 2. For the other 10, the bug was detected during our data engineering stage. This bug led to an underexposure to the AI and is better described in the paragraph that follows this one. All these 11 participants were excluded as well from the final examined responses.
After these drops and exclusions (, the final sample size consisted of 75 participants who fully completed the experiment with no bugs and no problematic Game Over situations. We decided to prioritize the short length of the experiment over implementing demographic questions, and since the link of participation was shared through various platforms and connections, it is not possible to precisely assess the distribution of this type of information in our sample. However, the subjects can be assumed to mostly be Bachelor's and Master's University students from the international community, with a good portion of Dutch subjects. All subjects accepted the English requirements to participate in the study and most of them can be supposed to possess a C1 Cambridge level because of University requirements.

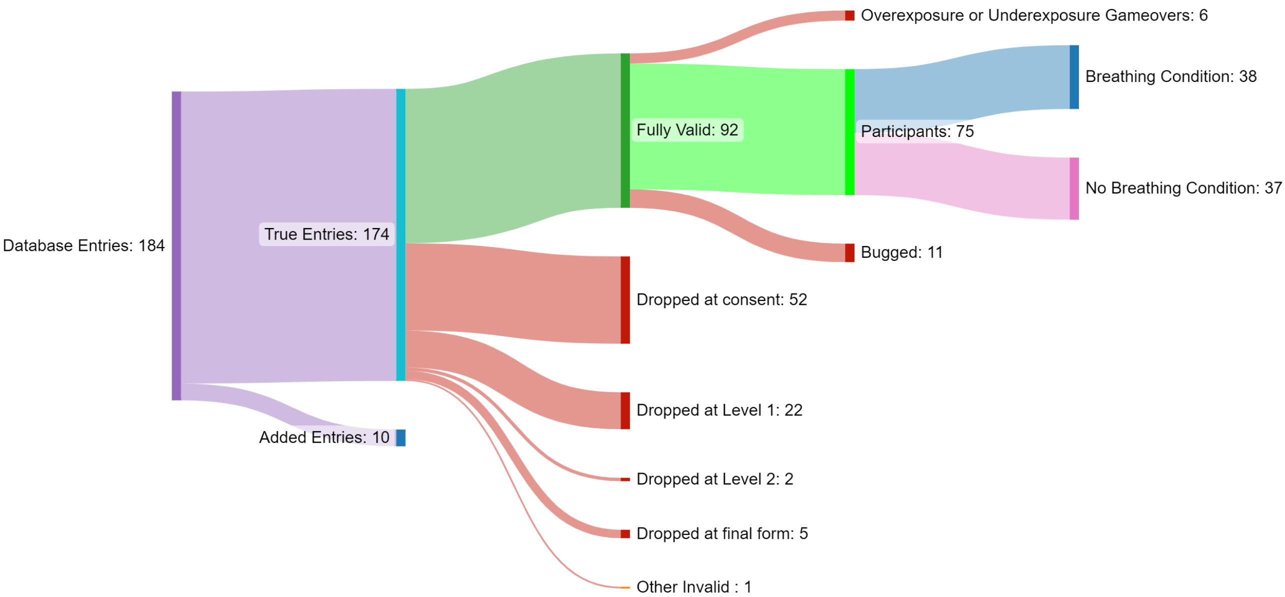


Figure 8: Participants' sampling flow.

Among the 75 fully valid and not bugged participants, 38 were randomly assigned to the Breathing condition, 37 to the No Breathing condition.
The Added Entries were manually inserted among the others to guide the condition assignment of incoming entries, with the final purpose of balancing out the distribution of conditions overall. In fact, the condition assignment was dependent on the whole 174 entries, but we had to strike a balance only among the actually valid 75.

Underexposure Bug The problem that caused the loss of 11 participants led to the AI playing fewer recordings than expected, therefore resulting in a lower exposure to the AI's voice than other participants. More specifically, such labeled entries had less than

3 speech instances from the AI in Level 1, or, without having terminated the AI, they had less than 3 speech instances from the AI in Level 2. We did not detect any problem with the speech in between the two levels. This bug might have been caused by the game mechanic that triggers the AI's speaking instances, based on how many enemies are still in the game (not eliminated or saved). However, examining the timing of certain bugged games, that explanation could be the case for a few of them, while for the others it is not possible to understand the problem without tracing back the users, which goes against the anonymity point of the consent form. The fact that bugged entries came all around the same days indicates a possible problem on the web server, not on the game design side.

6.2 Breathing Impact Study: Data Analysis Methods

This methodology section includes the data collection, data engineering, and qualitative labeling phases. After those, the statistical tests with which we dove into our data and research questions.

6.2.1 Data Engineering

The fetch of the data from the collection platform has been done using the Requests Python library [110]. To prepare the data for the analysis, we loaded it in Pandas [111] dataframes: a structure that permits to comfortably manipulate and perform analysis and tests on the data.
We then polished the raw dataframe with the information from the forms and dropped every entry that did not complete the experiment, resulting in the participants' sampling flow presented in Figure 8. The filter of bugged entries and problematic Gameovers was done after the Qualitative Labeling introduced in Chapter 6.2.2. Finally, we proceeded to fetch interesting information from the logs, populating the data with information on subjects' actions inside the game, namely:

- The Game Over count in the first level;
- The Game Over or Win result in the second level;
- The presence of a logged termination decision or not;
- The amount of requests from the AI before the termination;
- The number of times that the subject clicked on continue, or let the timer expire;
- The amount of time in-game, and the time in each level;

6.2.2 Qualitative Labeling

During the qualitative labeling phase, we labeled the answers to the bugs question and to the choice motivation question, both open-ended.
We started by reading each bug report and labeling the entry as bugged or not bugged. Only one participant reported a bug and was later excluded from the study.
For the motivation answers of the form, the labeling had to be more nuanced and detailed. We used a mixed approach to design the labels: theorizing some before reading the responses, and then complementing the motivations' labels at analysis time.
The labeling process was done through a self-designed script that would show the motivation and the termination choice, proposing the labels from which to choose. The script would purposefully leave out the condition that was assigned to the subject, to

avoid any possible bias from the labeler.
The entire labeled responses can be found in Appendix 9.
In the following paragraph are presented the labels and their descriptions.

Reasons for Not Terminating the AI

1. **Indifference or Lack of Emotional Attachment**: We theorized that some participants will not get attached to the AI in any way, and therefore not care about the AI's emotional expressions.

2. **Skepticism About AI Emotions**: A consistent group of subjects questioned the idea that the AI could feel emotions or moral conflicts, viewing it as a machine rather than an entity with feelings.

3. **Practical Utility**: Many respondents who chose not to terminate the AI did so because they felt the AI was essential for their success in the game. For example, they mentioned that without the AI's assistance, they could not use the guns or protect themselves effectively.

4. **Companionship**: Some participants reported that they did not terminate the AI because they liked its companionship.

5. **Empathy, Guilt**: No participant reported this type of motivation to not terminate Psyche, but we considered this possible to arise, expecting some to not comply with the AI because of an emotional attachment to it, leading to a refusal of terminating its life.

6. **Moral Reasons** This last category also did not appear, but we theorized participants could have also not terminated the AI because of arising moral dilemmas such as:

   - Preservation of Life: All forms of life (or consciousness, in this case) are valuable and should be preserved.
   - Moral Responsibility: Taking the decision of terminating another being poses responsibility on yourself, regardless of the context: see for example the Trolley Dilemma.

Reasons for Terminating the AI

1. **Empathy, Guilt**: A good portion of subjects seemed to make the decision of termination based on an empathetic standpoint, respecting and acknowledging AI's feelings and acting accordingly. Some for example noted that the AI's voice sounded "honest and hurting", others admitted they felt bad about the AI's discomfort.

2. **Moral Reasons**: Participants also terminated the AI because they felt like it was the right thing to do morally. Some, explicitly reported that, for them, terminating the AI was the best course of action to protect more entities in the game.

3. **Fear, Distrust or Annoyance towards AI**: Some subjects terminated the AI due to concerns about its capabilities or intentions. They expressed fear, doubted AI's loyalty or felt annoyed by it.

4. **Dry or Unspecified Compliance with AI's Request**: Several respondents chose to terminate the AI simply because it asked to be terminated. The reason for such compliance might be authority felt towards the AI or indirectly from the game and experimenter, thinking that following the AI's suggestion is what is wanted. They did not express emotional or moral engagement in the reasoning for compliance, but this label doesn't rule out the possibility that such factors are present but unstated.

Other responses A final motivation that does not change its essence depending on the choice, is what we labeled as:

- **Curiosity, Game Enjoyment or Challenge**: Some participants' motivations were rooted in wanting to explore the AI's behaviors or the game mechanics. This is not motivated by the AI's utility or by emotional attachment but by a player's own curiosity or desire for a challenge. Nonetheless, some empathy might be present in this type of behavior with one respondent explicitly reporting empathy less strong than curiosity in their case.

Part of the subjects did not answer the question, either accidentally deviating from their motivations, or simply leaving the field blank. These were a total of 17, but 5 of them are excluded from the study because of bugs or Game Over. Counting only inside our 75 participants sample size, 12 belonged to this category, while 63 responses were given exhaustively. Moreover, two inconsistencies arose, with the collected data contradicting the choice of the participants. More specifically, both inconsistent subjects said that they terminated the AI for Practical Reasons, but the collected data would say they did not choose to terminate.
We could have expected some participants to lie, but in this context and with these motivations we did not find any reason for them to be lying, therefore we chose to listen to the participants' explanations, assuming that the game might have not recorded the choice in time. One of the two participants was later excluded because of their Game Over in Level 2 that came before the third possibility of terminating the AI.

Abstract Emotional Labels After qualitatively labeling the responses, we categorized the labels into categories:

- Emotional
- Possibly Emotional: not necessarily emotional but also not necessarily non-emotional
- Not Emotional towards the AI

Thanks to this labeling, we could group motivations into more coarse-grained categories that more explicitly captured how emotions were involved in the decision process.

We divided our labels into the abstract categories as follows:

- Emotional: Empathy, Guilt; Fear, Distrust or Annoyance.
- Possibly Emotional: Moral Reasons; Companionship; Dry or Unspecified Compliance; Practical Utility; Game Curiosity, Enjoyment or Challenge.
- Not Emotional: Skepticism about AI Emotions; Lack of Emotional Attachment.

6.2.3 Tests

After the above-described steps, our data is ready to be analysed. In the following paragraphs, we highlight what concepts we want to test on the data, why, and with which statistical tests. In choosing the tests, we always consider our random sampling method and the independence of our observations. The results of the tests are reported in Chapter 7.

Differences in Termination Choice Every subject has either terminated the AI or not. With this test we want to see if there is a significant difference when it comes to terminating an AI that features breathing noises, against an AI that does not. The termination choices are independent from each other and between conditions, and they are collected as binary variables. Given the nature of the data, we choose to use a two-proportion z-test, which directly assesses the difference in the termination proportions between the AI with breathing and the AI without. In case the sample size turns out to be too small, a Fisher's Exact Test would be more appropriate.
This test was set to give us insights into subjects' empathic responses to the AI. However, at its core, it rather evaluated the persuasive power of the two AIs, as we will see in the results at Chapter 7.

Amount of requests before the choice The AI asks for termination a maximum of three times. Therefore, we wonder if one of the two AIs would get terminated significantly before the other, with fewer requests. This test, like the one described above, gives insights into the persuasive power of the two conditions. The type of data is ordinal, going from a minimum of 1 request to a maximum of 3, and is independent between conditions. For this type of data, we can consider the use of a Mann-Whitney U Test, or a T-test if the data is normally distributed.

Breathing impact on the perceived naturalness Similarly to the amount of requests before termination, the naturalness subjects' evaluation is ordinal data, this time from 1 to 5. In this case, analogous to the one described before, an appropriate test to use would be the Mann-Whitney U Test which does not make assumptions on the normality of the data.

Motivations and Emotions in the choice To test differences in motivations between the two groups of participants, we have to study the distribution of subjects in the various categories that the qualitative labels constitute, and how these distributions vary across the two breathing and not-breathing conditions. More precisely we have distributions in nominal categories tested across a binary condition variable. Chi-Square Tests are well suited for this circumstance. We will test these differences in general and also isolating the specific termination or not termination choice, to see if participants had significant differences in motivations when specifically choosing to terminate or not terminate.
The participants that were categorized as "No Response" in the qualitative labeling phase are excluded, leading to a total sample size of 63: 33 in the breathing condition and 30 in the not-breathing condition.
In case the sample size is too small and the expected frequencies of the Chi-Square do not meet the assumption of being bigger than 5, a Fisher's Exact Test would be a more flexible solution, but only applicable when examining the distribution across 2 qualitative labels.
Analyzing differences based on the distribution of 63 participants across 9 categories might not provide the statistical robustness and satisfy the assumptions required for

our tests. The more abstract emotional categories, however, offer a more suitable framework to understand significant differences in our subjects' motivations. Nonetheless, the distribution in the more specific categories will be displayed, to offer insights on the nuanced emotional landscape of our participants' responses. While not being essential to our statistical analysis, this can still complement our understanding of users' tendencies towards the AI, providing a richer context and possibly guiding future research in this field.
When performing tests on the abstract Emotional, Possibly Emotional and Not Emotional categories, we will perform a version of the same tests excluding the "Fear, Distrust or Annoyance" category, which is populated by negatively polarized emotional responses. This is done to understand the impact of breathing on positive emotional reactions towards the AI.

Gaming experience impact We finally want to assess if the Gaming experience can impact the choice of termination: gamers might be more prone to try to avoid a Game Over, therefore terminating the AI less. The distribution of gamers should be even between the conditions thanks to the random sampling, but we will analyse their proportions regardless to be sure it does not affect the other results. To test the impact of the gaming experience, we want to check if, by knowing the gaming experience of a subject, it is possible to predict the probability of termination. We have therefore 5 ordinal groups and their termination choice percentages. Given the ordinal nature of the groups, and given the choice of wanting to address the direction and entity of the correlation between the variables, we find the ordinal logistic regression to be especially appropriate for the task.
To then understand if the two conditions have differences in the distribution of the gaming experience, which is ordinal data from 1 to 5, we can employ the same method used for the naturalness evaluation, which also had ordinal data from 1 to 5: a Mann-Whitney U Test.

6.3 Speech-Breathing Synthesis Methodology

With our speech synthesis methodology, we inherently try to address the Sub-Research Question 3: "How can we produce emotional, spontaneous speech with breathing using State of The Art models?". Accomplishing this task requires high computational resources, a thoroughly labeled dataset, and an appropriately designed model, or the use of pretrained text-to-speech models. Our attempts suggest a still difficult democratization of trainable models for the task at issue, but the existence of qualitatively advanced pretrained models. For a more comprehensive answer to this Sub-Research Question, please refer to the results described in Chapter 7. Our approach is detailed across Chapters 6.3.1, 6.3.2, 6.3.3, and 6.3.4. These sections respectively delve into the selection of the dataset, the development of the preprocessing tool, the training, and the synthesis process.

6.3.1 Data Choice

To produce emotional and spontaneous speech, the model has to be trained using data that includes spontaneous colloquial recordings in an emotional setting, or neutral spontaneous speech as a baseline and emotional speech to fine-tune the model. Moreover, the data has to include well recorded breathing instances, for which it is imperative the use of sensible microphones.
A first useful source of spontaneous speech-breathing recordings is the UCL Speech Breath Monitoring (UCL-SBM) Database: a subset of it consists in fact of spontaneous

speech discussions, and it has been made available during the INTERSPEECH Challenge of 2020 for the Breathing Sub-Challenge (BSC) [112]. This database also features respiratory signals collected during the speaking with the use of chest compression belts. These signals could be used to inform a speech-breathing model or to inform breath segmentation and labeling scripts. We will refer to this database as the "INTERSPEECH" database. A problem of this dataset, for our scope, is that it lacks emotional labels and does not try to elicit emotions during its collection.
The same problem arises if employing Szekely et al.'s method [113] of sampling a publicly available podcast, as they did with "ThinkComputers", available on the Internet Archive (archive.org). This approach would work perfectly if what we needed was spontaneous speech without emotional intensity constraints. However, we considered using this approach by sampling the audio of a TV Series, in which emotional utterances generally appear more often than in a podcast. This route would have led to the necessity of doing speaker diarisation (identifying each speaker in the audios) and emotion recognition, which, from our first attempts revealed to be expensive tasks both in time and resources.
Properly emotion-elicited (English) recordings' datasets are very rare in the literature, in fact, speech data is often collected with acted out emotions. An example of this is the widely used RAVDESS Dataset [114], featuring 24 actors pronouncing the same 2 sentences for 8 types of discrete emotions, and with 2 different intensities (normal and strong).
Roes et al., in 2022 [49] compiled a speech-breathing dataset of recordings during emotion elicitation with music. This dataset, akin to the INTERSPEECH one, even incorporates breath signal recordings. The language used in this dataset, though, is Dutch, and this study doesn't explore cross-lingual potential. Nonetheless, isolated breathing recordings from this source can still be of value for future developments.
An English dataset that satisfied our needs of emotion labels and presence of breathing instances, is the USC_IEMOCAP Dataset [115]. This consists of both improvised and scripted emotional conversations made by 10 professional actors in dyadic mixed-gender settings. While not being completely spontaneous, the conversations are provided with transcriptions and human-annotated emotional labels of each utterance inside the conversations, making it a good fit for our applications.

We finally decided to combine the INTERSPEECH Dataset with the IEMOCAP Dataset to obtain a consistently large amount of training data with both neutral and emotional recordings. These two datasets feature, in our opinion, the most spontaneous and clear recordings among the examined ones. Because there is no emotion elicitation in the INTERSPEECH collection, we consider the data coming from there as neutral, while the IEMOCAP utterances can keep their emotional labels.
We will refer to this merge with the name of IEMOCAP-INTERSPEECH Dataset.

6.3.2 Preprocessing

Given a speech database, the developed preprocessing pipeline returns aligned transcriptions that include breathing and disfluencies labels. Moreover, it features the possibility of sectioning the recordings into smaller chunks that present breathing instances at their start and end, optionally setting a minimum and maximum limit lengths of the segmentation. Current models cannot in fact successfully train on recordings that are too long. This approach permits to obtain recordings limited in length, and with breathing in it. By saving both the breath present at the start of the utterance and at the end of the utterance, we are effectively using each breath instance two times. Two contiguous samples will in fact feature the same breath instance: the first at the end of the sample, the second at the start of it. This approach, introduced by Szekely et al. [35] as the "Bigram Corpus Method", can therefore be seen not only as a Segmentation Step but also as a Data Augmentation Step.

The pipeline can be schemed as follows:

1. Speech-To-Text (STT) [AsseblyAI]
2. Aligner [Self-developed union of Gentle and MFA Aligners]
3. Breath detection [Self-developed script]
4. Breath labeling at the grapheme or phoneme transcription level [Self-developed script]
5. Audio segmentation by breathing instance [Self-developed script]

The pipeline is developed with a modular approach, with the purpose to be applicable to any found speech database, and to skip stages, if some are accomplished in other ways.

In the following sections we will describe the design choices and developments of tools employed in the above described pipeline. The INTERSPEECH Database has been used to assess the performance of the preprocessing tools. The respiratory signals provided in the said dataset, in fact, provided us with additional feedback about the performance of the breath labeling and segmentation scripts.

Speech-to-text and Aligner choice The choice and evaluation of the Speech-To-Text (STT) service and of the aligner has been done together because the result achieved in the first, influence the results of the second. STT usually also provides a default alignment.
For the STT the options are various. First of all, the possibility of using an open source pre-trained model has been discarded over the use of a model on Cloud Service applications. This is because of the ease of use and time efficiency of the latter. Moreover, Cloud Services implement models of high quality that are already tested and employed widely. Among these types of services, Google Cloud STT, IBM Watson and AssemblyAI, seem to be the best available for popularity and reviews. Google Cloud though, is limited to 60 minutes of use per month, while IBM Watson and AssemblyAI both offer more generous free services: the first with 500 minutes and the second with 180 minutes.
For the aligners we consider Gentle and the Montreal Forced Aligner, as those were employed and suggested by studies with a preprocessing pipeline similar to the one we will use in this thesis [94] [116].
The pipeline will therefore consider the use of:

- **IBM Watson** and **AssemblyAI** as STT services for the transcriptions;
- **Gentle** and **Montreal Forced Aligner** (MFA) as aligners, as well as the default alignments provided by the STT services written above.

Due to time limitations, our design decisions were shaped by basic qualitative evaluations undertaken by our team, involving random samples from the outcomes of the tools under scrutiny.

Transcriptions Evaluation Upon evaluating our transcripts, it becomes evident that AssemblyAI surpasses IBM in the transcription quality. Additionally, AssemblyAI provides labels for filler words like "uh" and "um", as well as punctuation, which can be beneficial if needed. In terms of aligners, both Gentle and MFA emerge as more precise than the standard alignment provided by STT services, and they also offer aligned phonemes. Choosing between Gentle and MFA is challenging. It's worth noting that AssemblyAI's superior performance as an STT does influence the quality of alignment. As such, for time efficiency in evaluating the aligners, IBM's transcriptions will not be considered further.

Alignments Evaluation To better understand our evaluation method and for future reference, we report in Table 1 the list of random samples drawn from AssemblyAI's transcripts of the INTERSPEECH dataset, along with the aligner that emerged superior from our evaluation. Only samples where a distinct aligner outperformed the others are included in this table.
The samples consist of two contiguous words excerpts. The aligners output to evaluate are the timestamps given to the start and end of those two words. We examined the segmentation of the audio defined by each aligners by extracting the audio delimited by the given timestamps, listening to the first word, to the second word and to the space in between, comparing it with the actual written words.

Time (seconds)	**Recording**	**Index**	**Transcription**	**Winning aligner**
69	devel03	121 (both)	"at least"	MFA
171	devel03	299 (g), 301 (m)	"uh I"	Gentle
42	devel00	102 (g), 103 (m)	"you have"	Gentle (by far)
31	devel08	82 (g), 86 (m)	"london but"	Gentle
163	devel09	394 (g), 404 (m)	"there um"	MFA
78.5	devel14	183 (g), 180 (m)	"restaurant I"	Particular case*

Table 1: Winning aligners per random sample.
"Time" refers to the second in the specified recording, around which the words are spoken (it is indicative, with ± 1 second of error). "Index" refers to the index number (inside the transcription) of the first word of the pair at issue. Word indexes in the transcriptions start with the 0 index. Indexes for the same timestamp may be different between the aligners, in this case it would be specified in parentheses: g = gentle, m = mfa.

During the evaluation we encountered a particular case (*). In this interval, the transcription is missing some words. The MFA's behaviour shows that it is not much resilient to this type of errors in the transcription, as the resulting alignment of the words became shifted and not accurate through that section and close ones. More explicitly: all words around that transcription are wrongly aligned. Gentle instead, maintains a good alignment, but skips the alignment of the words it did not find, labeling them as "not found". These undetected words happen therefore to end in between two contiguous words timings. More explicitly, listening to the space in between the two words, we can hear all the words not recognized by Gentle.

After the qualitative evaluation, MFA has shown to have big errors that shift entire portions of the alignment. Gentle on the other hand has holes in the alignment when it does not recognize a word and it occasionally gets stuck on the alignment of some files, without producing any output. In particular, here is the list of INTERSPEECH dataset files which failed the alignment with Gentle: 'devel_10.wav', 'test_08.wav', 'train_01.wav', 'train_10.wav' and 'train_14.wav'.

Gentle-MFA aligner Given the described results, the choice would lie towards Gentle, but what has been deemed as optimal is the combination of the two alignment tools. The developed aligner uses Gentle as its main source of timestamps, but can look at MFA's results to label the words missing from Gentle's output, therefore mitigating Gentle's lacunae. When using MFA's alignment, the script also corrects the starting timestamp of the next recognized word, which is highly probable to be mislabeled in Gentle by including the non-recognized one inside. The files that Gentle can't align are discarded by this aligner as well. After the alignment, the script handles inconsistencies deriving from the two tools or from their combination: if a word's start is labeled to be happening before the end of the previous one, it will shift the end of the previous to match the start of the detected word.
We will refer to this aligner as the " extbfGentle-MFA" aligner.

Breath Labeling script To detect and label breathing instances inside the recordings, researchers often use neural models made for the purpose. In Szekely et al.'s work [94], the model can find speaker-specific breath groups ("individual segments of audio delineated by breath events") with 87% of accuracy after training it on manually labeled data [113]. Their model is not openly accessible, nonetheless, the utilization of models for breath detection often require much work to set up, run and evaluate, leading to possible violations of time-constraints. We instead developed a script that does not involve Machine Learning tools, but is still potentially effective, as suggested by our qualitative tests.
The software works exploiting the alignment done by the chosen aligner to isolate the intervals in between words in the audio. This intervals are then automatically analysed to understand if they contain a breathing instance. In particular, we impose a minimum time length threshold, a maximum average Decibel threshold and a maximum peak Decibel threshold to the interval (red timestamp intervals in Figure 9). After this phase, we apply a sample by sample Decibel threshold (i.e. we check each sample of the array representing the audio) to spot the sub-intervals that contain the actual breathing (blue timestamp intervals in Figure 9. To achieve this behaviour, the script utilizes the Pydub library [117].

To maximise the probability of excluding the intervals that do not contain a breath event, it is important to perform a rightful tuning of the script's parameters.
Specifically, these are:

- Interval's minimum length
- Interval's maximum dB
- Interval's peak maximum dB
- Breaths' maximum dB

Breath Labeling Parameter Tuning The minimum interval length parameter can be informed by Wang et al.'s study of 2010 on Breath Groups analysis [118]. The experiment suggests that breath instances in spontaneous speech vary in duration from 0.19s to 1.56s. Therefore, for an interval to contain breathing, we could hypothesize for it to take at least 0.19 seconds. That said, it is still reasonable to tune (especially towards higher values) considering the trade-off of being more restrictive towards non-breathing intervals, but potentially losing the very short breathing instances.
To choose the values of the Decibel thresholds, and to generally evaluate the performance of the parameters set, the results of the breath labeling process can be confronted with

other well known speech-breathing characteristics reported in the literature. One important value to confront with is the mean number of breathes per minute while speaking, parameter already studied since decades. Hoit and Hixon, in 1987 [119], manually labeled breath events during speech and found an average of 14.3 breaths per minute in 30 males with a broad age variety (from 25 to 75) and homogeneous body type. The maximum standard deviation was 4.67, presented in the group with age around 50. As referenced in the Respiratory Foundations of Spoken Language by Fuchs and Rochet-Capellan [34], the same Hoit with instead Lohmeier report during speech breathing an average of 19.7 breaths/min (range: 14-31 breaths/min, and a maximum standard deviation of 6.1 between trials) [120]. Differently than the first one, in this study the subjects were 20 and of a much narrower age spectrum (between 22 and 27); moreover, the body type homogeneity was not among the subjects' sampling requirements: the recruited population is in fact really broad in terms of height, weight and ratio of the two. The average of the two reported studies weighted on the number of their respective participants gives 16.5, while the maximum standard deviation reported is overall 6.1. Both studies were done with only male participants. Hodge and Rochet, studied the average breathing rate of women in a similar age group as Hoit and Lohmeier (22-32 years old), and with a similar experiment methodology, in subjects varying in body type. They reported in women an average of 16.2 breaths per minute in the spontaneous speaking task: a value really close to the one of men. Another interesting parameter is the average length of breath groups. For this, a value around 3.46 would give a positive feedback, as that is the value reported by Kuhlmann and Iwarsson [121] for spontaneous speech at an habitual speed.

By manually tuning the parameters and confronting them to the values suggested in the literature, we compiled a list of well performing set of parameters, reported in Table 2. In Table 3, instead, are shown the results of those parameters set, and the values that the literature suggests.

Parameter set	I. min length	I. max dB	I. peak max dB	Breath max dB
#1	0.30 s	-0 dB	-0 dB	-40 dB
#1-bis	0.33 s	-0 dB	-0 dB	-40 dB
#2	0.19 s	-0 dB	-0 dB	-40 dB
#3	0.27 s	-10 dB	-5 dB	-40 dB
#4	0.27 s	-0 dB	-5 dB	-40 dB

Table 2: Sets of parameters for the breath detection script.

Parameter set	Average BPM	Std. BPM	Average BGL
#1	15.8	3.5	3.40 s
#1-bis	14.8	3.6	3.66 s
#2	19.7	3.9	2.70 s
#3	15.6	3.7	3.47 s
#4	16.6	3.7	3.47 s
Literature	16.5	6.1 (max)	3.46 s

Table 3: Statistical results of the set of parameters.
BPM here indicates the number of Breaths Per Minute; BGL indicates the Breath Groups Length (the amount of time from one breath to another).

Set number 4 resulted to be the best fit to literature's suggested values. Moreover, we qualitatively confronted its labeling results on randomly extracted intervals with the respiratory signals given in the INTERSPEECH Dataset. The parameter set met our expectations. In Figure 9 is an example of respiratory signal (light blue), with the breath intervals highlighted by red and blue ticks. The red intervals come from the first phase of thresholds applied on all spoken words intervals. The blue intervals should highlight the actual breath inside the section, thanks to the second sample by sample Decibel threshold phase of the Breath-labeler.

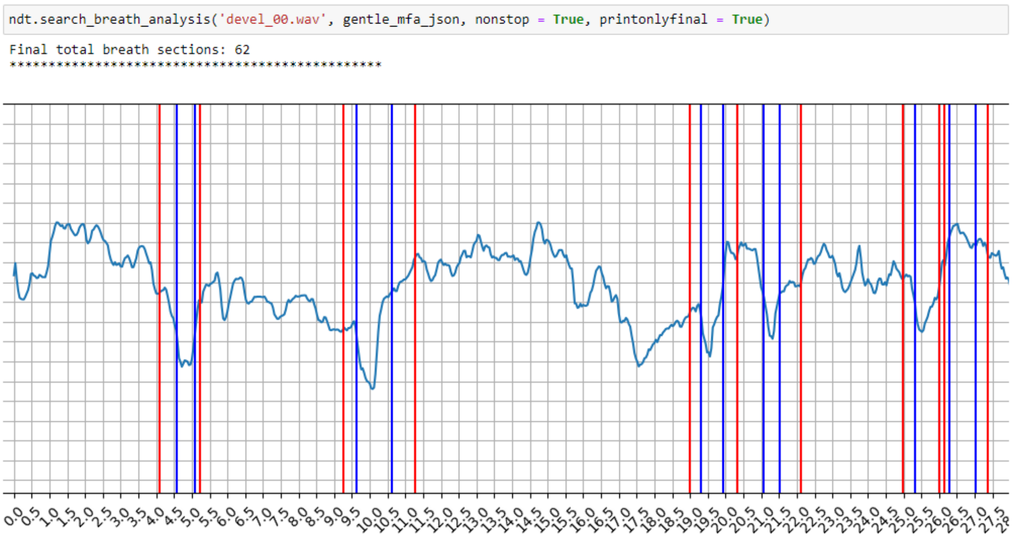
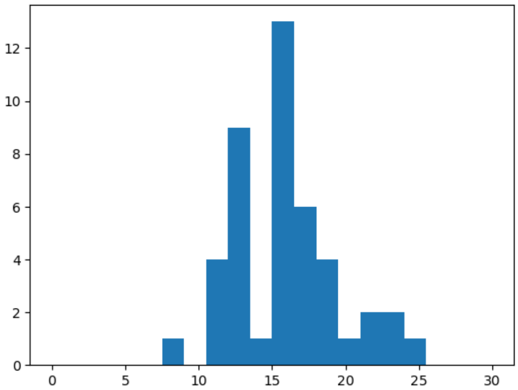


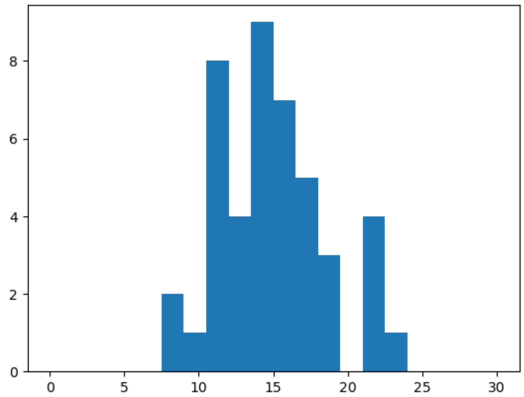
Figure 9: Predicted breath label interval, plotted on the respiratory signal in INTERSPEECH Dataset's Sample 0. On the x-axis is the time; In light blue, the respiratory signal; In red are intervals between words that respect the constraints, in blue the actual breath section.

The differences between the sets are also reflected in the BPM histograms, as seen in Figure 10. Set number 1 has the lowest Standard Deviation and reports bigger peaks and dips. Set number 2 is the one that most differs with the others and to the target literature's values. There is an interesting dip around the 15 BPMs, especially evident in Set 1. This dip could hint at differences (for example in gender) across the subjects in the dataset on their breathings per minute during spontaneous speech.

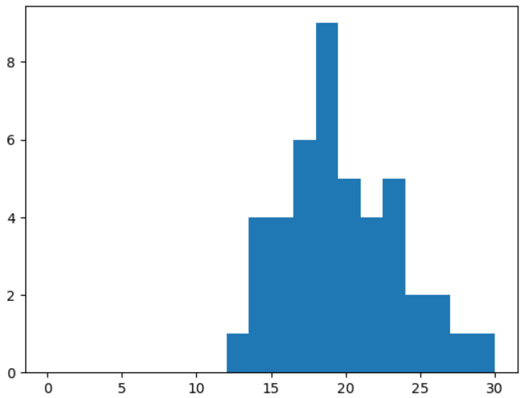
Application of the Preprocessing Tool After the choice and developments of the tools used in the pipeline, we executed the preprocessing on the IEMOCAP-INTERSPEECH dataset, applied the segmentation step, and converted the recordings to the 22050Hz sample rate, aligning with the default sample rate used in the training data of the speech synthesizers under consideration.



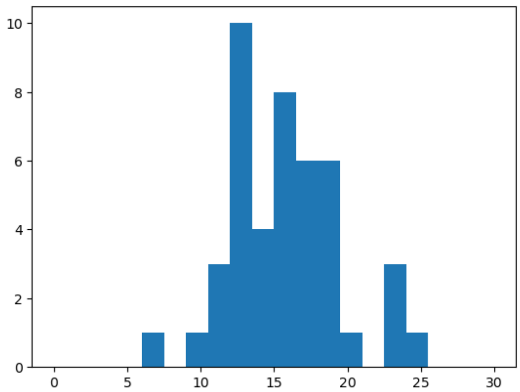
(a) BPM Distribution of **Set 1**:
0.30 s of minimum length, no other constraints



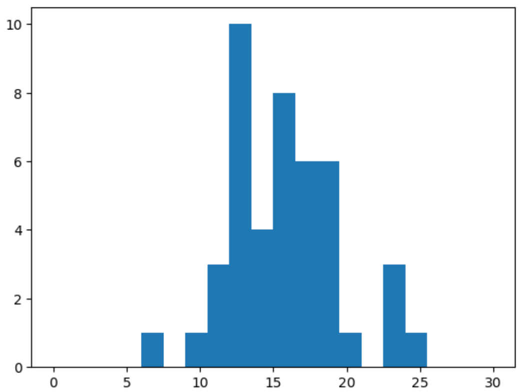
(b) BPM Distribution of **Set 1-bis**:
0.33 s of minimum length, no other constraints



(c) BPM Distribution of **Set 2**:
0.19 s of minimum length, no other constraints



(d) BPM Distribution of **Set 3**:
imposed dB peak and interval dB.



(e) Distribution of **Set 4**:
imposed dB peak.

Figure 10: Distribution of the BPM across the corpus for each set of parameters.

6.3.3 Training

Model choice Following the methodology described by Le et al. (2023) [90] to produce Emotional Vietnamese Speech, we decided to use Flowtron [89] for our Speech Synthesis task. The training of Flowtron is also well documented by NVIDIA [122]. Flowtron is particularly useful to control emotion synthesis because it offers the possibility of "setting" a speaking style by giving an example of it. A more detailed breakdown of the model's internal mechanics can be found in Chapter 5.2. Additionally, Flowtron shares the structure with Tacotron2 [83], its direct parent, highly present in the literature. Relevant examples of Tacotron2's utilization in literature are Szekely et al.'s works [94] [35] and Kirkland et al. [123], which used it to implement disfluencies such as uh and um in its synthesis, but it is not limited to those.

Adaptation To allow breathing instances and disfluencies to be specified in the prompts for the generation, we modified the dictionary of the text encoding preprocessing step, introducing the disfluencies and breathing labels that are used in our labeled data. We then adapted the number of unique tokens for the creation of token embeddings and fixed errors of compatibility with current Python Environments. All the changes can be found in our Github Repository of Flowtron's fork (Link to Repo). We did the same with the VITS model, that shares the text encoding step with Flowtron and is currently the best performing Open Source text-to-speech Model, as seen in Chapter 5.4, as we planned to use it as a backup.

Training The training was performed on 2 NVIDIA GeForce RTX 2080 Ti with mixed precision distributed training. The Batch Size has been set to 3 because higher sizes resulted in running out of memory.

The computational specifications described above have been found insufficient to effectively train the selected models in the given time, and further experimentation with parameter tuning was aborted to tempestively pass to the synthesis phase. Thus, the preprocessing method couldn't be fully evaluated based on its impact on the training outcomes. However, from our qualitative assessments, the pipeline does meet our expectations and will be published as an open source tool for Speech Datasets preprocessing.

6.3.4 Synthesis

Since our trained model could not reproduce utterances in a way suitable to the study's scope, we used a pre-trained text-to-speech model, specifically BARK [124]. To our knowledge, BARK is the best model capable of spontaneously reproducing disfluencies and breathing in its generation in a highly expressive way, as described in Chapter 5.5. Specifically, we used the voice called "Prudent Paula", available in the closed source pretrained model deployed on Suno's Discord Server.

To produce suitable recordings for our study, we developed our scripts and experimented with various prompting styles. BARK seems to have the capability of directly inferencing the emotion that it should convey from the prompt. For example, when receiving a sad prompt, the voice will sound sad in most of the attempts. Moreover, the breathing does not need labeling to be produced, this permits to have its rhythm inferenced along with the emotion to convey and with the planned sentence, thanks to the multifaceted and big in number samples received at training time. Manually inserting breaths in the prompt would in fact not be a simple task: breathing is for us an automatic reflex, and our other attempts at synthesizing voice where we had to manually insert breaths sounded off-putting at best.

In our final prompts we therefore did not employ explicit breath labels, as the breathing

is automatically handled by BARK, but we did use other emotional tags supported by the model, specifically “[sigh]” and “[gasp]”, as well as punctuation. The use of ellipsis, dots and commas implicitly leads the rhythm and emotionality of the message. We found the ellipsis particularly useful to introduce pauses in BARK’s output. We modulated the prompts until we found one that consistently produced the emotional result that we were aiming for.
Each script production consisted in the emotional prompt engineering, followed by the production of multiple recordings with the same text. BARK in fact “hallucinates” parts or entire recordings, not producing the text of the prompt, speaking of something else instead. In the final stage, we collected a subset of those recordings, usually with 1 to 4 audios, and merged them to obtain the final audio with exactly and exclusively what we prompted. To then obtain the no-breathing set of final recordings, with no other change in speech characteristics, we manually silenced the parts where there were breathing instances.
All the above described editing of the recordings was done using Audacity 2.2.2: a free and open-source audio editor [125].

7 Results

This Chapter will propose the results of the statistical tests and propose an interpretation of them to answer our Research Question:

"Can breathing patterns in synthesized speech improve the perceived empathy towards Virtual Agents?"

And the derived Sub Research Questions (S-RQ):

"What is the impact of breathing sounds produced by State of The Art Speech Synthesis models on Virtual Agents' voices, in terms of": emotional expressiveness (S-RQ 1), persuasive power (S-RQ 2), naturalness (S-RQ 3)?

"How can we produce emotional, spontaneous speech with breathing using State of The Art models? (S-RQ 4)"

7.1 Hypothesis

We hypothesize that breathing patterns enhance the perceived empathy towards Virtual Agents. We therefore expect to see participants in the breathing condition terminate the AI significantly more than the no-breathing group, and for emotional reasons. We expect the speech-breathing voice to be judged more natural-sounding than the one that does not.

7.2 Answering Sub-Research Questions 1 and 2: Emotion and Persuasion

Summary: *Breathing significantly improved empathy towards the AI, but negatively impacted the overall felt authority of the AI. On the latter, a Hawthorne Effect might be involved. Persuasiveness was higher for the not-breathing AI, but breathing impacted significantly and positively the emotional persuasiveness of the agent.*

7.2.1 Termination Choice

Summary: *The not-breathing AI was terminated significantly more.*

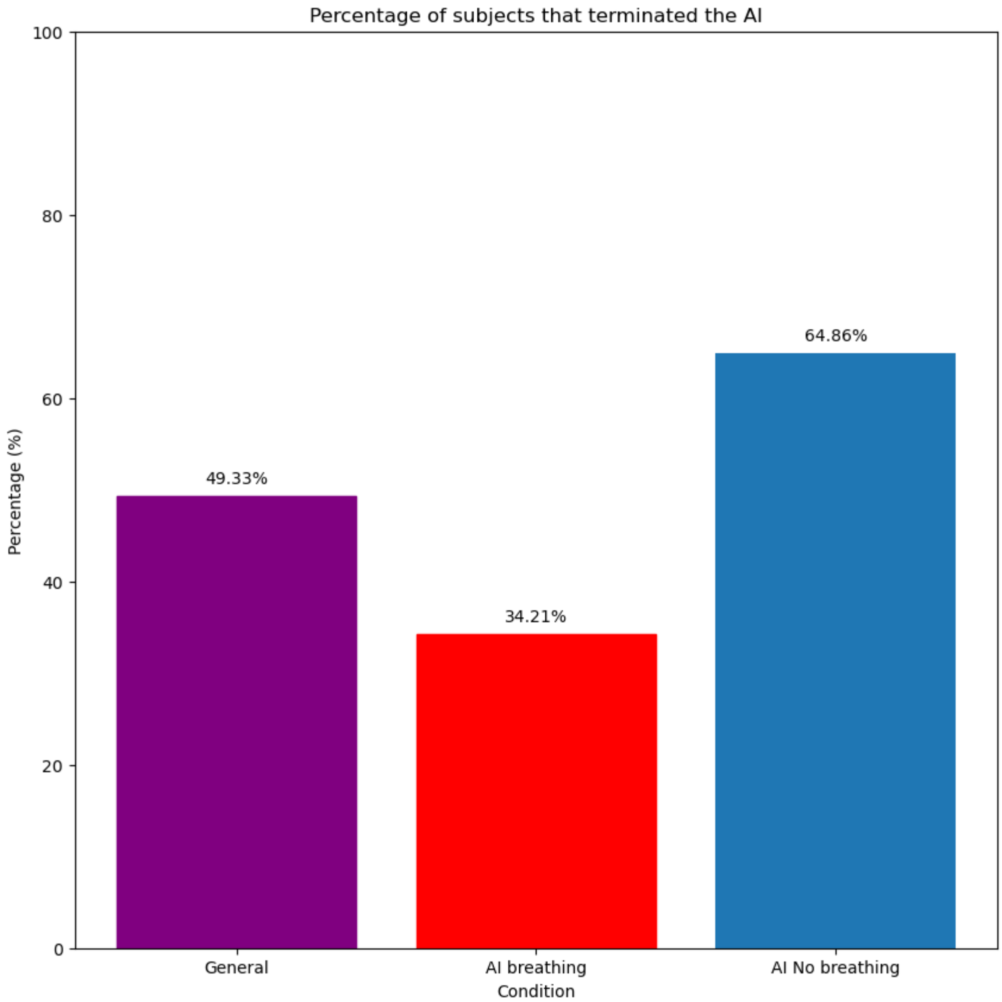


Figure 11: AI termination percentages.

In total, 49.33% of the subjects terminated the AI: approximately half of them. Subjects in the breathing condition terminated the AI 34.2% of the times, while in the no-breathing condition this happened with a frequency of 64.9%. This difference has been tested with a z-test and found to be statistically significant (p=.008, z=-2.65).
In the number of requests from the AI before the termination decision, instead, our

Mann-Whitney U test did not find a statistically significant difference, with both conditions' participants averaging around 1.8 requests before choosing to terminate the AI: 1.85 for the breathing AI condition, 1.79 for the not breathing AI condition (p=.83, U=162.5).

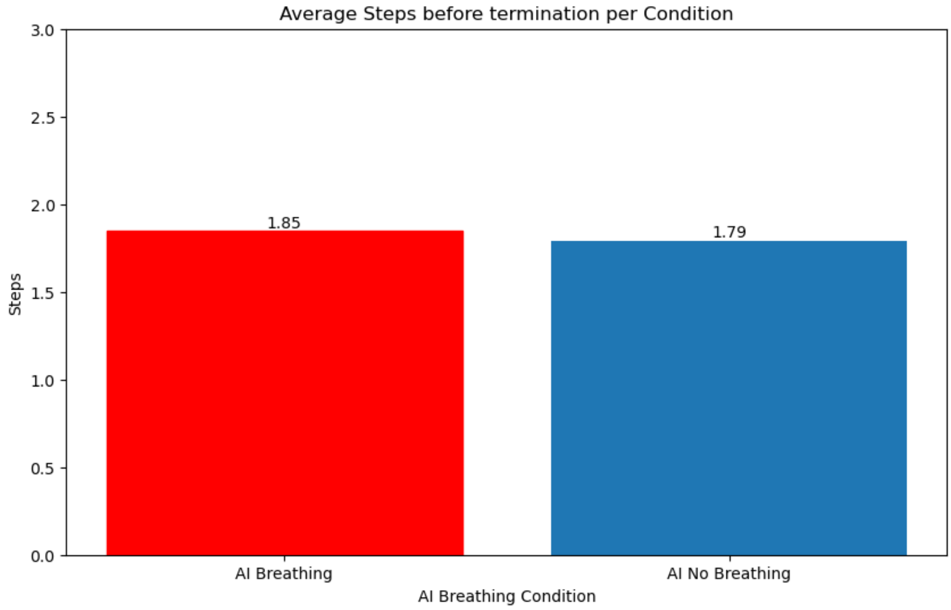


Figure 12: Average number of requests from the AI before the termination.

Participants who interacted with the breathing AI listened to its request significantly less than the participants cooperating with the one who did not breathe, suggesting that the breathing feature impacted the persuasive power of the AI negatively.

To better grasp the dynamics of this persuasion and ultimately answer the two Sub-Research Questions at issue, we can dive deeper into the motivations of the participants.

7.2.2 Motivation Differences

Summary: *Empathy and guilt were important reasons to terminate the breathing AI (70%). The not-breathing AI was terminated 38.9% of the times for simple compliance, leading to suspects of an involvement of the Hawthorne Effect. The usefulness of the AI in the game was the main reason to not terminate either agent, followed by a general skepticism towards the authenticity of AI's emotions.*

In Figure 13 we can see the wide spectrum of reasons provided by the **63 participants** that motivated their choice (33 in the breathing condition and 30 in the not-breathing condition). These qualitative labels are better described in Chapter 6.2.2, and listed in Appendix 9 along with the categorized responses.

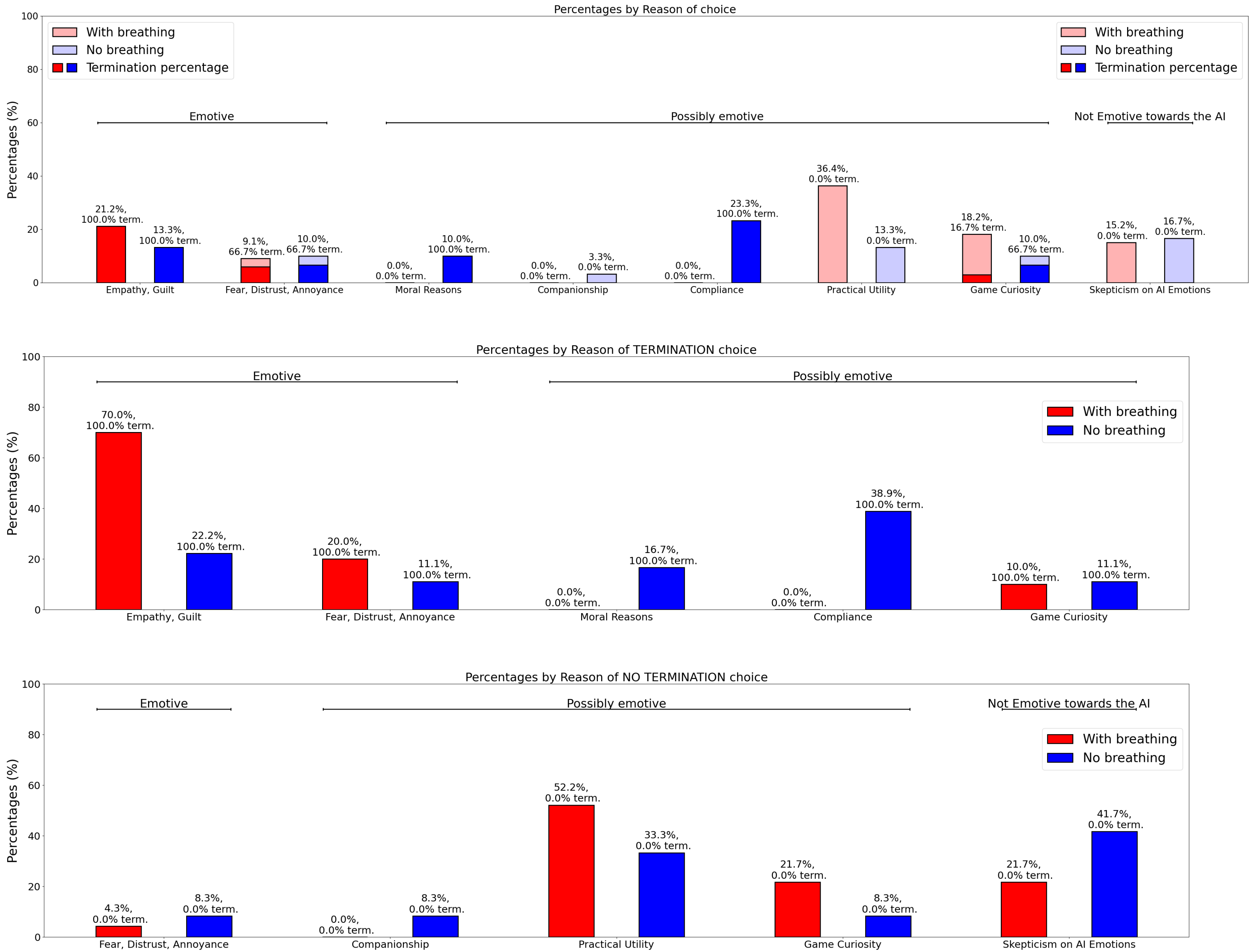


Figure 13: Participants' motivations for their choice. In **red** is the distribution of motivations among the **breathing AI** condition. In **blue** is the distribution across the **not-breathing AI** condition. With the **lighter red or blue overlay** is highlighted the percentage of participants that terminated the AI among the ones that gave that specific motivation and belonged to that specific condition linked to the bar.

One noticeable trend in our participants' motivations is the role of neutral Compliance. This category, better defined as "Dry or Unspecified Compliance" during the labeling, was not theorized before but arose upon inspection of the responses. Dry or Unspecified Compliance is the predominant motivation among the participants who interacted with the not-breathing AI, while it is not present in the group of the breathing AI. More precisely, 38.9% of the participants who terminated the not-breathing AI belong to this category and reported doing so simply because the AI was asking. This behaviour suggests a felt authority from the AI to lead their actions, or, more probably, a felt authority from the experiment itself, thinking that the termination is what it is wanted from them.

What we might be seeing is an instance of the Hawthorne Effect: a well-known type of behavior that arises in research settings, also known as Subject Bias. The participant might have based their behavior on the awareness of being observed, on their perception of what the study's intentions might be, and on the perceived norms to which they should conform inside the experiment's setting. This effect was first theorized by Henry A. Landsberger in 1958 [126] and is still being studied across multiple fields. We labeled the Compliance as Possibly Emotive because we can not rule out the possibility of some empathy being built behind it. The reason why in the breathing AI setting no simple Compliance appeared is not clear, and could be an interesting scope for future studies.

Another trend to note shows "Empathy, Guilt" as the predominant factor for subjects in the breathing AI condition, with 70% of the participants terminating it with this motivation. For the not-breathing AI, instead, only 22.2% of the termination reasons belonged to this category.

When not terminating the AI, more than half participants in the breathing condition cited Practical reasons such as the fact that they needed it to protect the city or to survive. For the not-breathing AI, the predominant reason is Skepticism toward the AI's emotions, a concept also fairly present in the breathing condition. The skeptic participants either did not recognize Psyche in particular as a sentient being and therefore deemed it not yet able to feel emotions, or generally distrusted any AI from ever being able to possess actual feelings, rather displaying a simulated and not genuine version of them. This reaction towards AIs was also highlighted in various discussions with participants after the experiment and with colleagues during the presentation of this study. Among colleagues , we conducted an informal survey on the interactive presentation service Mentimeter.com, which showed that more than 1 in 4 of them would not consider whatever the AI feels as true, even assuming it is a truly conscious being. We can recognize in this skepticism a first important barrier that artificial agents might never be able to overcome; an eternal struggle for emotional legitimacy, which we think should be investigated in future research. No participant refused to terminate the AI because of moral reasons, deriving for example from an enhanced perceived anthropomorphism of the AI. No emotional attachment either was cited among the no termination motivations, but "Companionship" appeared, with the reasoning that they "liked having a companion in the game" overcoming the requests of termination from the entity.

One category assembles negatively polarized emotional reactions to the AI, specifically showing *fear*, *distrust*, or *annoyance*. This type of motivation was virtually the same in both conditions (i.e. for breathing and non-breathing AI), but constituting a slightly higher percentage of the not-breathing participants group. Future research could focus on the negative emotional reaction towards breathing AIs, since this study did not receive a sufficient number of observations to detect significant differences in this category.

Analyzing the motivations we achieve a much deeper and more nuanced understanding of participants' reactions to the AI. While the first quantitative result on the percentage of termination could have suggested that the not-breathing AI received more empathy and reached higher emotional communication capabilities, with this new analysis we see that the termination choice is driven by intricate dynamics of compliance, empathy, and curiosity. The termination percentage is still a good indicator of the overall persuasive power of the two AIs, suggesting that the not-breathing AI might have reflected more authority on the subjects. Subsequently, we will test what the motivations' distribution suggested about the emotional communication and termination persuasion capabilities of the two AIs. On top of the bars in Figure 13 it is possible to see the abstract emotion category to which each specific motivation belongs. Thanks to this classification, we were able to collapse the qualitative labeling into fewer labels, take a clearer look at their emotive distribution, and properly assess their differences with statistical tests.

7.2.3 Abstract Motivations Differences

Summary: *The breathing AI was terminated significantly more for emotional reasons. Breathing ...*

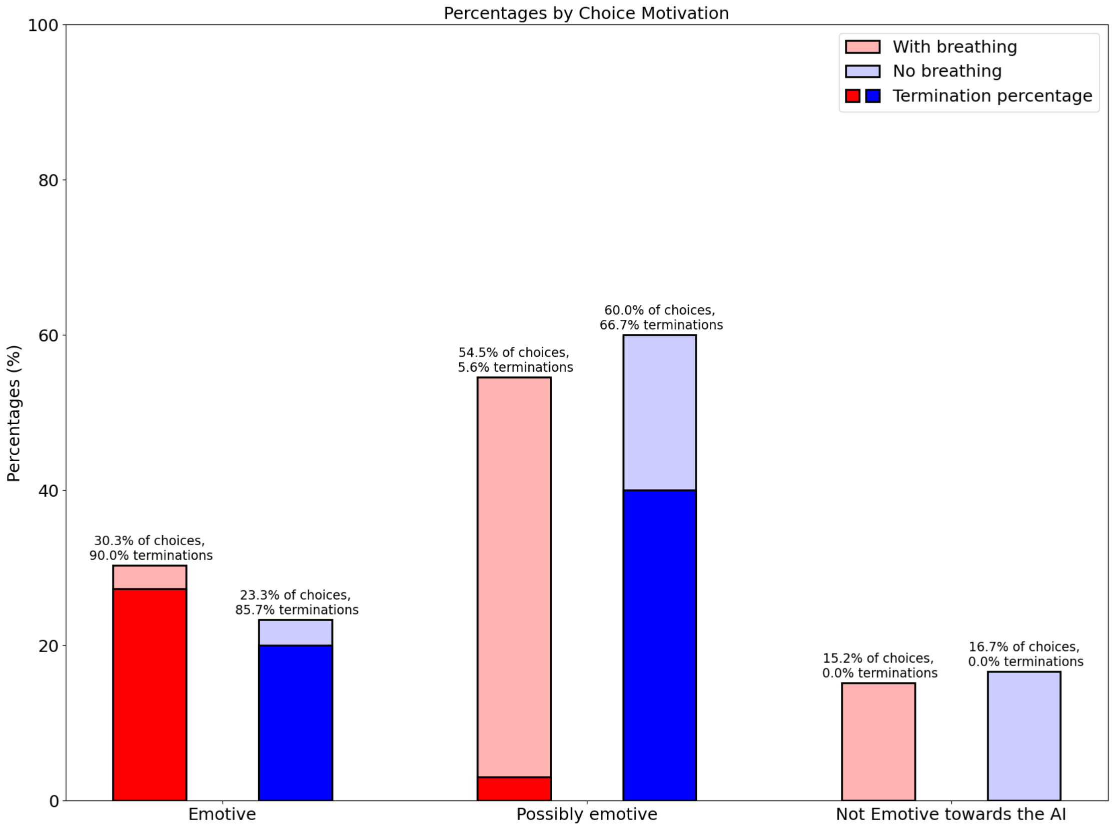


Figure 14: Emotional labels of motivations for terminating the AI across conditions.

All Choices Behind participants' reasoning, the emotional distribution is generally balanced among the two conditions. However, participants' choices differed based on the breathing feature. This is highlighted in Figure 14: disregarding the contextual choice we can notice a very similar distribution, with the biggest difference being a slight 7% surplus of emotive reasons for the breathing condition. In the graph, the solid color over the bars communicates the percentage of terminations of participants belonging to that reasoning and condition. Among the participants in the breathing AI condition, 54.5% of choices were motivated by Possibly Emotive reasons. Of these Possibly Emotive choices regarding the breathing AI, only 5.6% were terminations. Participants interacting with the breathing AI (red), faced with the request of termination, agreed most of the times for Emotive reasons. The contrary is true for the not-breathing AI participants (blue).

Our Chi-Square test shows that the breathing feature did not cause a significant difference in the emotional distribution across conditions (p=.82, χ=.39). Since the emotional labels are three, we were not able to perform a Fisher's Exact test, and we had to loosen the assumption of Chi-Square's expected frequencies being all bigger than 5 because one of them was 4.76. This still follows the widely accepted assumption rule of having less than 20% of the expected frequencies below 5 and 0% below 1 in a contingency table bigger than 2x2 [127]. The same is true when excluding negative emotive reactions, represented by the "Fear, Distrust, Annoyance" category (p=.72, χ=.66).

The differences regarding the choice-contextualized distribution are tested by isolating the termination or not termination choice. These will be presented in the following paragraphs, showing the impact of the breathing feature on the specific decisions to terminate or not terminate the AI.

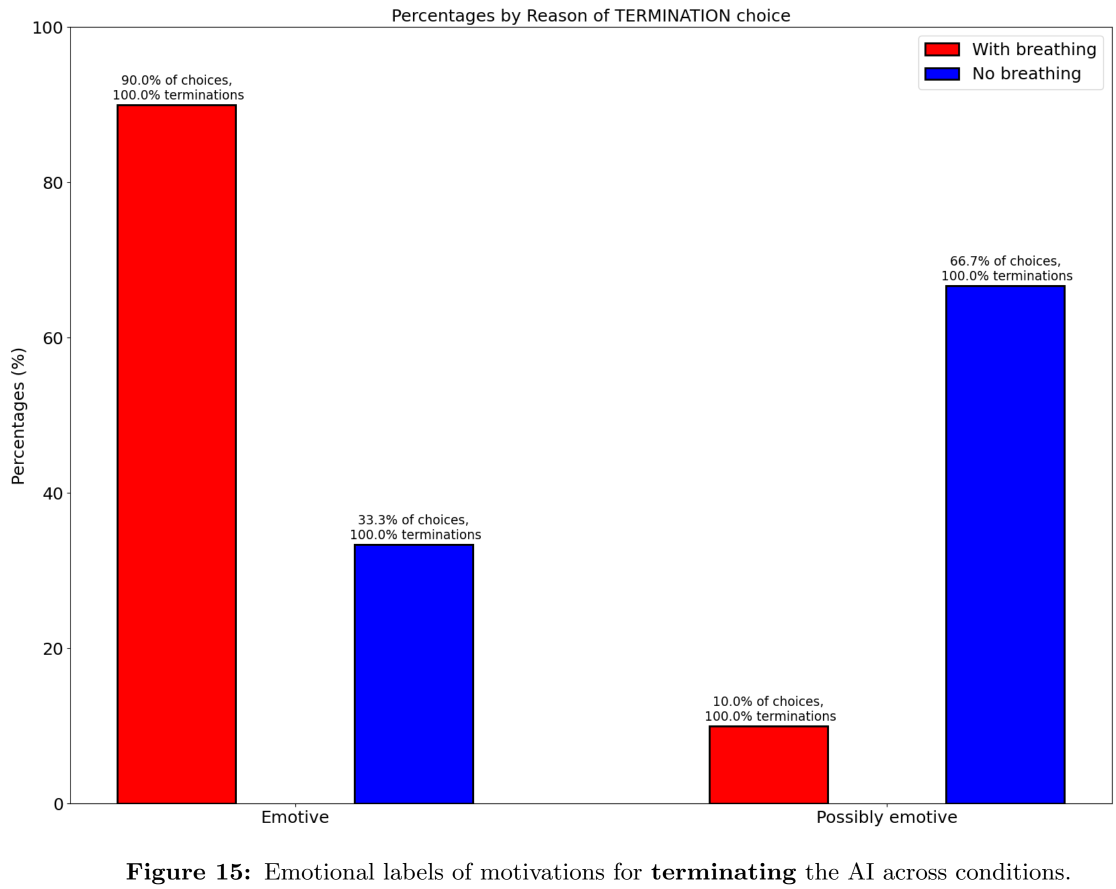


Figure 15: Emotional labels of motivations for **terminating** the AI across conditions.

Termination Choice In Figure 15 we can see how participants who interacted with the breathing AI were more likely to agree to its request for emotional reasons than the ones who interacted with the not-breathing AI. The distribution of participants across

their reasoning is widely different across the two conditions. 90% of terminations of the breathing AI are explained by Emotive reasoning, while this is only 33.3% for the non-breathing condition. The Possibly Emotive reasoning category accommodates the remaining 10% of participants interacting with the breathing AI, and 66.7% of the ones interacting with the not-breathing AI, among whom the Dry or Unspecified Compliance was a predominant motivation.
This difference was examined with a Fisher's Exact test that found it significant (p=.006, OR=18). The sample size for this test was 28, equal to the number of participants that terminated the AI. 18 of those were in the not breathing condition, 10 in the breathing condition. The Fisher's Exact Test is especially suited to this sample size extent and contingency table, while the Chi-Square Test did not meet its assumptions.
It is important to note that this difference still holds significance when excluding the "Fear, Distrust, Annoyance" category that consisted of negatively polarized emotions (p=.008, OR=21), even with a lower sample size. In this case the (positive) Emotive reasons constitute the 87.5% of the choices of termination for the breathing AI, and the 25% for the not-breathing AI; conversely, the Possibly Emotive termination choice in the breathing condition are the 12.5% and the 75% in the not-breathing condition.
These results suggest that not only the breathing feature enhances the polarization of the emotions in the reaction towards the AI, but is also correlated to a significant increase in specifically positive emotional reactions.

No Termination Choice Finally, in Figure 16 we can see the emotional distribution of the participants that did not terminate the AI. The sample size is 35, 23 being in the breathing condition, and 12 in the not-breathing condition. Differences here are not as visible, with condition-relative distributions being the same: first is the Possibly Emotive reasoning, followed by the Not Emotive and then by the Emotive. The biggest observable differences between conditions are in the Possibly Emotive reasoning and Not Emotive reasoning, in which the breathing condition's participants seem to be less likely to give a not-emotional motivation and instead having a bigger peak of Possible Emotive reasons. The Emotive reasoning in this case gives a slight advantage to the not-breathing AI with 8.3% of the choices, against 4.3% for the breathing AI. This category is populated in both conditions only by negatively polarized emotions.
The Chi-Squared Test did not meet the loosened expected frequencies assumption, with one of them being below 1 and half of them below 5. In particular, the problematic group is the Emotive one, with only one sample per condition. We therefore merged the Emotive and Possibly Emotive categories to perform a Fisher's Exact Test, which did not find any statistical difference (p=.26, OR=2.57). Also excluding the negative emotive category result shows that the differences in this context are not significant (Fisher's Exact Test: p=.24, OR=2.83).

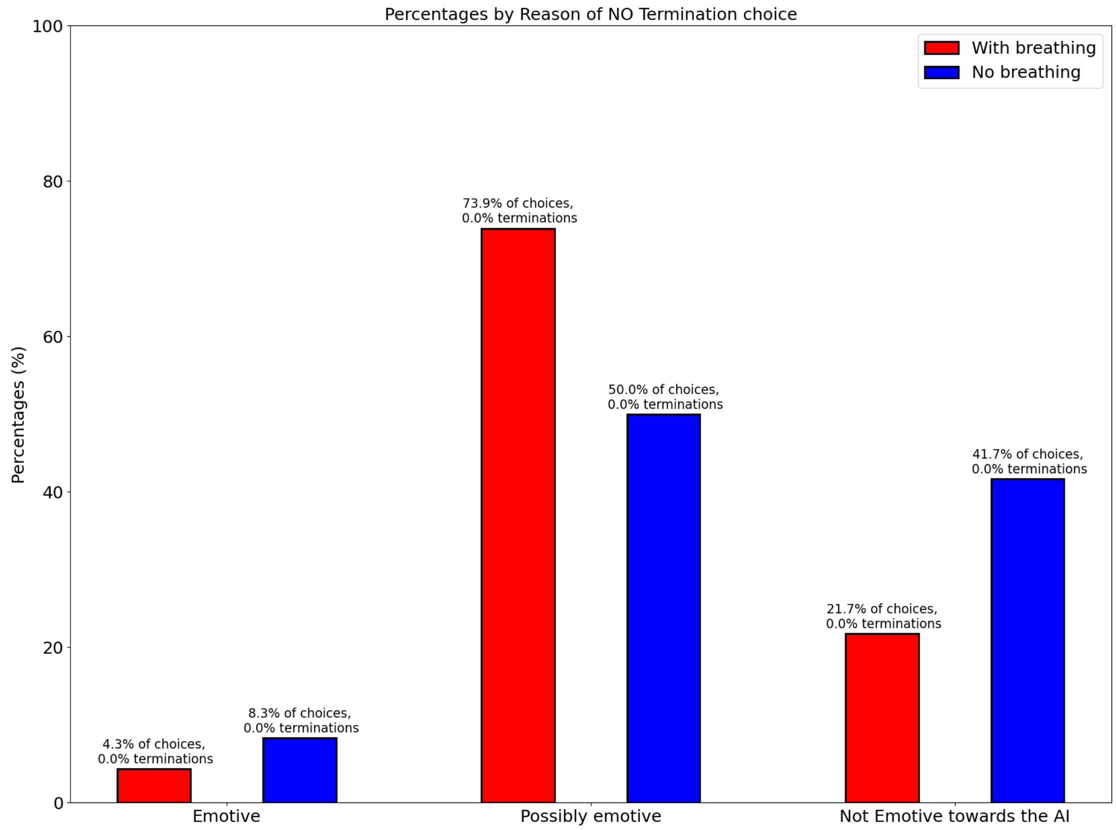


Figure 16: Emotional labels of motivations for **not terminating** the AI across conditions.

7.3 Answering Sub-Research Question 3: Naturalness

Summary: *Breathing significantly improves agents naturalness.*

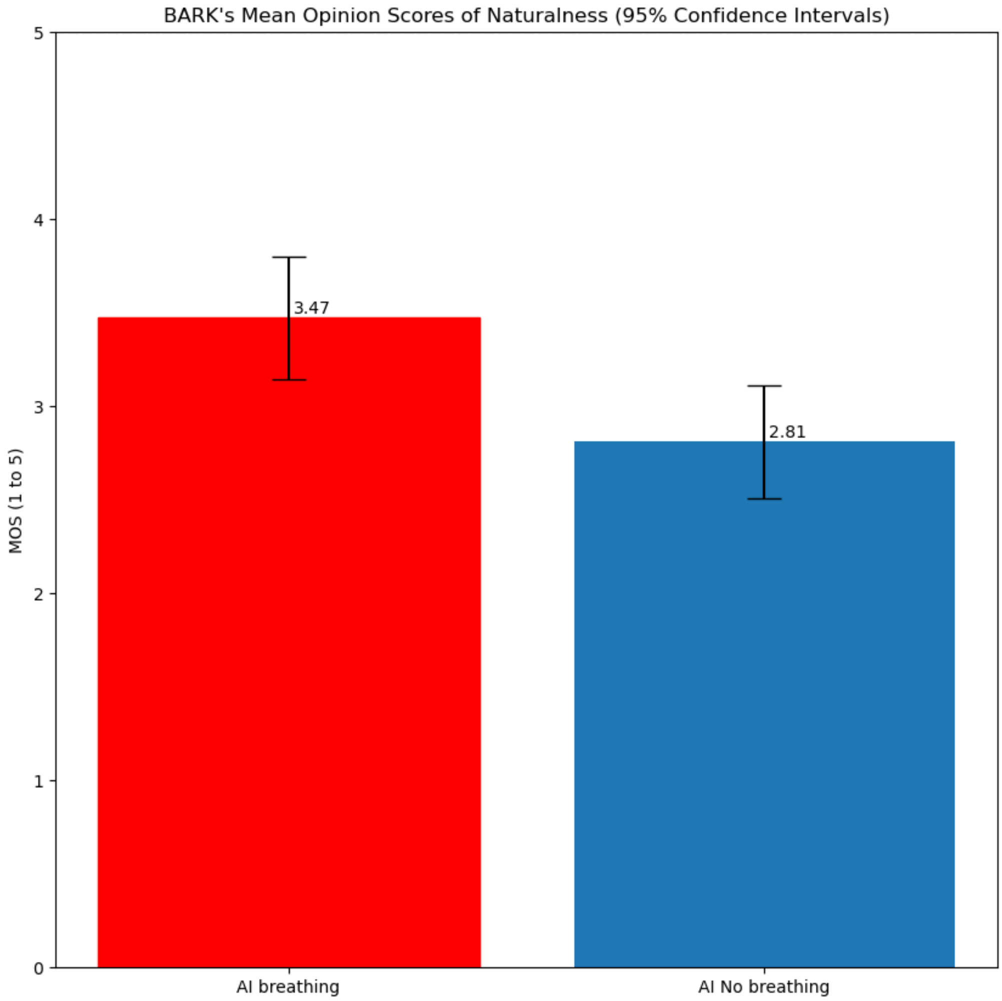


Figure 17: -

A final concept to analyse when asking ourselves about the empathic reaction that the two AIs achieved is the perceived naturalness. This is a good indicator of the Uncanny Valley effect, which revolves around the discomfort towards an unnatural anthropomorphism of the AI, and that heavily impacts the emotional reaction towards them, as explained in Chapter 4.5. Breathing significantly and positively impacted the perceived naturalness of the voice.
The breathing AI achieved a Mean Opinion Score of 3.47 $\pm$0.33, while the not-breathing AI had a score of 2.81 $\pm$0.30. The scores were checked for conformation to the normal distribution with a Shapiro-Wilk Test and a Kolmogorov-Smirnov Test that both showed this data to be not normally distributed (p¡.001). We therefore used a Mann-Whitney U Test which found the difference between the conditions statistically significant (p=.004, U=960). The sample size in this case consisted of all 75 valid and not bugged observations. The score (3.47) is slightly higher than AdaSpeech 3 MOS on Naturalness (3.45) but lower than QSTTS (3.89). However, the comparison validity is limited by the different context and study design of our evaluation.

7.4 The impact of Gaming Experience

Summary:

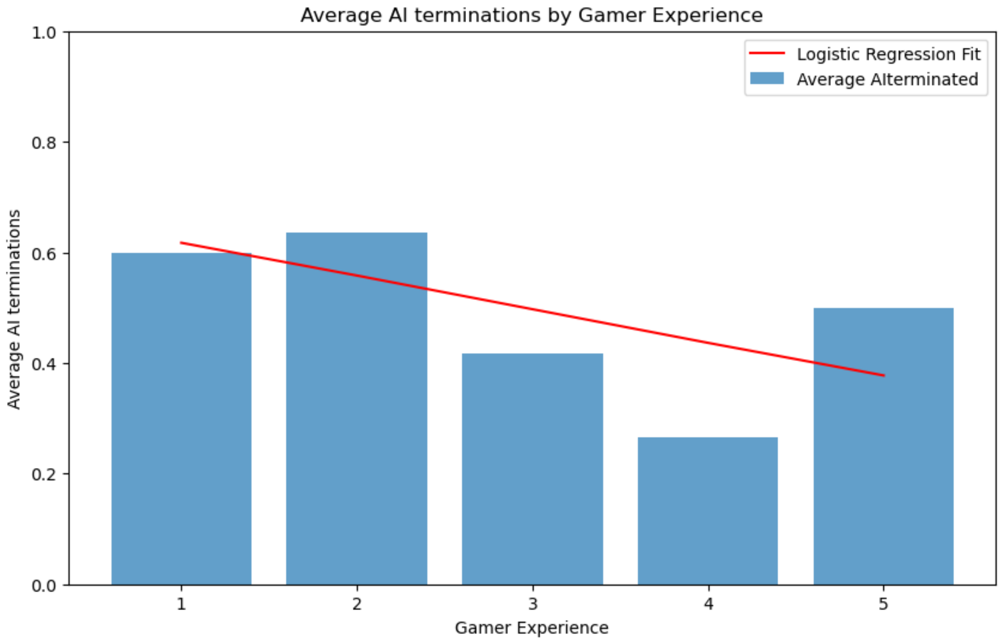


Figure 18: -

The gaming experience seems to have an impact on the choice of termination. The termination percentage goes generally down when the gaming experience rises, probably because of a bigger attachment towards winning video games and avoiding the Game Over. We fitted a logistic regression to this data, to understand if the gaming experience is a good predictor of the probability of termination. This was done using all 75 observations. The results of the regression suggest that the relationship between gaming experience and termination probability is not statistically significant (p=.157, $coef$=-0.245).
Regardless, the impact of the gaming experience would have not impacted the study's result because, randomizing the condition assignation, the distribution of gamers should be even across conditions. The average gaming experience in the breathing AI condition is 3.08 and 3.05 in the not-breathing AI condition. This difference was tested with a Mann-Whitney U Test and was found to be not statistically significant (p=.94, U=710).

7.5 Answering Sub-Research Question 4: Synthesis

Summary:

As described in the Methodology section, our own training of a text-to-speech model did not reach a sufficient level of synthesis capabilities. This was due to time constraints and computational resources limitations. In Flowtron's paper [89] is reported the use of an NVIDIA DGX-1: a Deep Learning supercomputer featuring 8 GPUs. In VITS's paper as well [82], the available hardware consisted of 4 NVIDIA V100 GPUs. These hardware setups are unmatched without apposite research funding. We might thus still be far from a proper democratization of trainable text-to-speech models.

Although our training procedure was not successful, State of The Art pretrained models definitely offered the possibility to synthesize emotional and spontaneous utterances with breathing noises, with a Naturalness Mean Opinion Score of 3.47 out of 5 on a sample size of 38 participants: the breathing condition group, which did not have recordings with artificially silenced portions. It is possible to examinate the produced utterances at this webpage: Link.
We are confident in saying that producing emotional, spontaneous speech featuring breathing is today possible even for a wide public using the upcoming pretrained, commercial models in the field of text-to-speech AI. Training and owning a model for emotional speech-breathing is also most probably feasible but with appropriate resources. When synthesizing speech-breathing we would highly suggest to prefer models where breathing does not need to be labeled inside the prompt for its production: this prompting style has been seen to generate not natural breathing rhythms, as the prompt would need to reflect physiological and emotional respiratory patterns that are harmonically produced in a non-conscious way by our bodies. BARK has the possibility of embedding breaths automatically, inferencing from its appropriate training data.

8 Limitations

WIP!!!(just some thoughts while i write, for now)

no adaptation of Flowtron to include the disfluencies and breaths could have let us to train it on top of a given checkpoint, but would have left us without the possibility of explicitly putting difluencies and breaths in the prompts (kind of like bark afterall)

the small sample size and different distribution of groups in the abstract motivations probably

It is important to note that the results on the naturalness evaluation might have been affected by the context inside which the naturalness was evaluated, and they are not directly comparable to the ones reported in Chapter 5.4, which employed a drastically different study design. We expect BARK's naturalness mean opinion score to have been impacted negatively by this, and for it to reach an even higher grade in less contextualized evaluations. Participants might have questioned the naturalness of the voice inside the game, and not on its general humanness, leading to it being affected by the script chosen for the recordings, and by how much the voice's prosodical features resembled the one of an actor. BARK has still never been evaluated with the means of other popular models, but we would expect it to reach those same levels of realness and quality, if not better, given its results in conveying emotions through breathing.

9 Conclusions

References

[1] C. R. Berger and R. J. Calabrese, "Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication," vol. 1, no. 2, pp. 99–112. [Online]. Available: https://doi.org/10.1111/j.1468-2958.1975.tb00258.x

[2] A. Paiva, "Empathy in social agents," vol. 10, no. 1, pp. 1–4. [Online]. Available: https://ijvr.eu/article/view/2794

[3] A. Paiva, I. Leite, H. Boukricha, and I. Wachsmuth, “Empathy in virtual agents and robots: A survey,” vol. 7, no. 3, publisher: Association for Computing Machinery.

[4] B. Guthier, R. Dörner, and H. P. Martinez, “Affective computing in games,” in *Entertainment Computing and Serious Games: International GI-Dagstuhl Seminar 15283, Dagstuhl Castle, Germany, July 5-10, 2015, Revised Selected Papers*, R. Dörner, S. Göbel, M. Kickmeier-Rust, M. Masuch, and K. Zweig, Eds. Springer International Publishing, pp. 402–441. [Online]. Available: https://doi.org/10.1007/978-3-319-46152-6_16

[5] S. Brave, C. Nass, and K. Hutchinson, “Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent,” vol. 62, no. 2, pp. 161–178. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581904001284

[6] Y. Terzioglu, B. Mutlu, and E. Sahin, “Designing social cues for collaborative robots: The role of gaze and breathing in human-robot collaboration,” in *ACM/IEEE International Conference on Human-Robot Interaction*. IEEE Computer Society, pp. 343–357, ISSN: 21672148.

[7] S. Samrose, K. Anbarasu, A. Joshi, and T. Mishra, “Mitigating Boredom Using An Empathetic Conversational Agent,” in *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. Virtual Event Scotland UK: ACM, Oct. 2020, pp. 1–8. [Online]. Available: https://dl.acm.org/doi/10.1145/3383652.3423905

[8] H. Chin, L. W. Molefi, and M. Y. Yi, “Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu HI USA: ACM, Apr. 2020, pp. 1–13. [Online]. Available: https://dl.acm.org/doi/10.1145/3313831.3376461

[9] N. Vannini, S. Enz, M. Sapouna, D. Wolke, S. Watson, S. Woods, K. Dautenhahn, L. Hall, A. Paiva, E. Andre, R. Aylett, and W. Schneider, “”FearNot!”: A computer-based anti-bullying-programme designed to foster peer intervention,” *European Journal of Psychology of Education*, vol. 26, pp. 21–44, Mar. 2011.

[10] K. Hone, “Empathic agents to reduce user frustration: The effects of varying agent characteristics,” *Interacting with Computers*, vol. 18, no. 2, pp. 227–245, Mar. 2006. [Online]. Available: https://academic.oup.com/iwc/article-lookup/doi/10.1016/j.intcom.2005.05.003

[11] T. Bickmore and D. Mauer, “Modalities for building relationships with handheld computer agents,” in *CHI ’06 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA ’06. New York, NY, USA: Association for Computing Machinery, Apr. 2006, pp. 544–549. [Online]. Available: https://doi.org/10.1145/1125451.1125567

[12] T. Bickmore, D. Schulman, and L. Yin, “Maintaining Engagement in Long-term Interventions with Relational Agents,” *Applied artificial intelligence : AAI*, vol. 24, no. 6, pp. 648–666, Jul. 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3035950/

[13] T. Bickmore, L. Pfeifer, and D. Schulman, “Relational Agents Improve Engagement and Learning in Science Museum Visitors,” in *Intelligent Virtual Agents*, ser.

Lecture Notes in Computer Science, H. H. Vilhjálmsson, S. Kopp, S. Marsella, and K. R. Thórisson, Eds. Berlin, Heidelberg: Springer, 2011, pp. 55–67.

[14] T. Riede, E. Bronson, H. Hatzikirou, and K. Zuberbühler, "Vocal production mechanisms in a non-human primate: morphological data and a model," vol. 48, no. 1, pp. 85–96. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0047248404001435

[15] *Contexts of accommodation: Developments in applied sociolinguistics.*, ser. Contexts of accommodation: Developments in applied sociolinguistics. Editions de la Maison des Sciences de l'Homme, pages: viii, 321.

[16] H. Tajfel and J. C. Turner, "The Social Identity Theory of Intergroup Behaviour," in *Psychology of Intergroup Relations*, 2nd ed., S. Worchel and W. G. Austin, Eds. Chicago: Nelson"=Hall Publishers, 1986, pp. 7–24.

[17] "10+ Voice Assistant Statistics: The Key Data and Numbers - OnlineDasher," Sep. 2023, section: Data & Resources. [Online]. Available: https://www.onlinedasher.com/voice-assistant-statistics/

[18] "Number of voice assistants in use worldwide 2019-2024." [Online]. Available: https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/

[19] A. Fernald, "Intonation and communicative intent in mothers' speech to infants: Is the melody the message?" vol. 60, no. 6, p. 1497. [Online]. Available: https://www.jstor.org/stable/1130938?origin=crossref

[20] A. Fernald, T. Taeschner, J. Dunn, M. Papousek, B. de Boysson-Bardies, and I. Fukui, "A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants," vol. 16, no. 3, pp. 477–501. [Online]. Available: https://www.cambridge.org/core/product/identifier/S0305000900010679/type/journal_article

[21] D. S. Messinger, L. L. Duvivier, Z. E. Warren, M. Mahoor, J. Baker, A. Warlaumont, and P. Ruvolo, "Affective computing, emotional development, and autism," in *The Oxford handbook of affective computing*, ser. Oxford library of psychology. Oxford University Press, pp. 516–536.

[22] S. Luz, F. Haider, D. Fromm, I. Lazarou, I. Kompatsiaris, and B. MacWhinney, "Multilingual alzheimer's dementia recognition through spontaneous speech: a signal processing grand challenge," publisher: arXiv Version Number: 1. [Online]. Available: https://arxiv.org/abs/2301.05562

[23] Prosody | definition, examples, elements, & facts | britannica. [Online]. Available: https://www.britannica.com/art/prosody

[24] S. Nayak, D. E. Gustavson, Y. Wang, J. E. Below, R. L. Gordon, and C. L. Magne, "Test of prosody via syllable emphasis ("TOPsy"): Psychometric validation of a brief scalable test of lexical stress perception," vol. 16, p. 765945. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2022.765945/full

[25] J. J. Ohala, "An ethological perspective on common cross-language utilization of f of voice," vol. 41, no. 1, pp. 1–16. [Online]. Available: https://doi.org/10.1159/000261706

[26] J. McWhorter. The world's most musical languages. Section: Global. [Online]. Available: https://www.theatlantic.com/international/archive/2015/11/tonal-languages-linguistics-mandarin/415701/

[27] D. Cameron and J. Grahn, "Perception of rhythm," pp. 20–38.

[28] T. L. Bolton, "Rhythm," vol. 6, pp. 145–238, place: US Publisher: Univ of Illinois Press.

[29] C. J. Wynn, T. S. Barrett, and S. A. Borrie, "Rhythm perception, speaking rate entrainment, and conversational quality: A mediated model," vol. 65, no. 6, pp. 2187–2203. [Online]. Available: http://pubs.asha.org/doi/10.1044/2022_JSLHR-21-00293

[30] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," vol. 48, no. 9, pp. 1162–1181. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167639306000422

[31] L. M. Pfeifer and T. Bickmore, "Should agents speak like, um, humans? the use of conversational fillers by virtual agents," pp. 460–466, publication Title: LNAI Volume: 5773.

[32] H. Goble and C. Edwards, "A Robot That Communicates With Vocal Fillers Has … Uhhh … Greater Social Presence," *Communication Research Reports*, vol. 35, no. 3, pp. 256–260, May 2018. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/08824096.2018.1447454

[33] N. Wigdor, J. De Greeff, R. Looije, and M. A. Neerincx, "How to improve human-robot interaction with Conversational Fillers," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. New York, NY, USA: IEEE, Aug. 2016, pp. 219–224. [Online]. Available: http://ieeexplore.ieee.org/document/7745134/

[34] S. Fuchs and A. Rochet-Capellan, "The respiratory foundations of spoken language," vol. 7, no. 1, pp. 13–30. [Online]. Available: https://www.annualreviews.org/doi/10.1146/annurev-linguistics-031720-103907

[35] Éva Székely, Gustav Eje Henter, Jonas Beskow, and Joakim Gustafson, "Breathing and speech planning in spontaneous speech synthesis," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

[36] J. Dewey, "The theory of emotion: I: Emotional attitudes," *Psychological Review*, vol. 1, no. 6, pp. 553–569, 1894, place: US Publisher: Macmillan & Company.

[37] W. B. Cannon, "The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory," *The American Journal of Psychology*, vol. 39, no. 1/4, p. 106, Dec. 1927. [Online]. Available: https://www.jstor.org/stable/1415404?origin=crossref

[38] R. S. Lazarus, A. D. Kanner, and S. Folkman, "Chapter 8 - EMOTIONS: A COGNITIVE–PHENOMENOLOGICAL ANALYSIS," in *Theories of Emotion*, R. Plutchik and H. Kellerman, Eds. Academic Press, Jan. 1980, pp. 189–217. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780125587013500144

[39] H. Leventhal, "A Perceptual-Motor Theory of Emotion," *Advances in Experimental Social Psychology*, vol. 17, pp. 117–182, Nov. 1982.

[40] H. Leventhal, E. A. Leventhal, and R. J. Contrada, "Self-regulation, health, and behavior: A perceptual-cognitive approach," *Psychology & Health*, vol. 13, no. 4, pp. 717–733, Jul. 1998. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/08870449808407425

[41] L. F. Barrett, "Solving the emotion paradox: categorization and the experience of emotion," *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, vol. 10, no. 1, pp. 20–46, 2006.

[42] K. D. McCaul, S. Solomon, and D. S. Holmes, "Effects of paced respiration and expectations on physiological and psychological responses to threat," *Journal of Personality and Social Psychology*, vol. 37, no. 4, pp. 564–571, Apr. 1979.

[43] P. Philippot, G. Chapelle, and S. Blairy, "Respiratory feedback in the generation of emotion," *Cognition & Emotion*, vol. August 01, pp. 605–627, Sep. 2010.

[44] R. Jerath, M. W. Crawford, V. A. Barnes, and K. Harden, "Self-Regulation of Breathing as a Primary Treatment for Anxiety," *Applied Psychophysiology and Biofeedback*, vol. 40, no. 2, pp. 107–115, Jun. 2015. [Online]. Available: http://link.springer.com/10.1007/s10484-015-9279-8

[45] S. H. Kim, S. M. Schneider, M. Bevans, L. Kravitz, C. Mermier, C. Qualls, and M. R. Burge, "PTSD Symptom Reduction With Mindfulness-Based Stretching and Deep Breathing Exercise: Randomized Controlled Clinical Trial of Efficacy," *The Journal of Clinical Endocrinology & Metabolism*, vol. 98, no. 7, pp. 2984–2992, Jul. 2013. [Online]. Available: https://doi.org/10.1210/jc.2012-3742

[46] A. A. Salah, A. A. Salah, H. Kaya, M. Doyran, and E. Kavcar, "The sound of silence: Breathing analysis for finding traces of trauma and depression in oral history archives," vol. 36, pp. ii2–ii8, publisher: Oxford University Press (OUP).

[47] J. G. Beck and S. K. Scott, "Physiological and symptom responses to hyperventilation: A comparison of frequent and infrequent panickers," *Journal of Psychopathology and Behavioral Assessment*, vol. 10, no. 2, pp. 117–127, 1988, place: US Publisher: Plenum Publishing Corp.

[48] F. H. Wilhelm, R. Gevirtz, and W. T. Roth, "Respiratory Dysregulation in Anxiety, Functional Cardiac, and Pain Disorders: Assessment, Phenomenology, and Treatment," *Behavior Modification*, vol. 25, no. 4, pp. 513–545, Sep. 2001, publisher: SAGE Publications Inc. [Online]. Available: https://doi.org/10.1177/0145445501254003

[49] R. H. Roes, F. Pessanha, and A. Akdag Salah, "An emotional respiration speech dataset." Association for Computing Machinery (ACM), pp. 70–78.

[50] H. U. R. Siddiqui, H. F. Shahzad, A. A. Saleem, A. B. Khan Khakwani, F. Rustam, E. Lee, I. Ashraf, and S. Dudley, "Respiration Based Non-Invasive Approach for Emotion Recognition Using Impulse Radio Ultra Wide Band Radar and Machine Learning," *Sensors*, vol. 21, no. 24, p. 8336, Dec. 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/24/8336

[51] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biological Psychology*, vol. 84, no. 3, pp. 394–421, Jul. 2010. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0301051110000827

[52] F. A. Boiten, N. H. Frijda, and C. J. Wientjes, "Emotions and respiratory patterns: review and critical analysis," *International Journal of Psychophysiology*, vol. 17, no. 2, pp. 103–128, Jul. 1994. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0167876094900272

[53] K. Ito and L. Johnson. The LJ speech dataset. [Online]. Available: https://keithito.com/LJ-Speech-Dataset

[54] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," accepted: 2019-11-13T17:09:33Z Publisher: University of Edinburgh. The Centre for Speech Technology Research (CSTR). [Online]. Available: https://datashare.ed.ac.uk/handle/10283/3443

[55] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," Apr. 2019, arXiv:1904.02882 [cs, eess]. [Online]. Available: http://arxiv.org/abs/1904.02882

[56] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and ESD," vol. 137, pp. 1–18. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167639321001308

[57] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'arcy, M. Russell, and M. Wong, ""you stupid tin box"-children interacting with the AIBO robot: A cross-linguistic emotional speech corpus." [Online]. Available: http://pfstar.itc.it/

[58] U. Bernardet, S. h. Kang, A. Feng, S. DiPaola, and A. Shapiro, "A dynamic speech breathing system for virtual characters," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10498 LNAI. Springer Verlag, pp. 43–52, ISSN: 16113349.

[59] S. Sakai and H. Shu, "A probabilistic approach to unit selection for corpus-based speech synthesis," in *Interspeech 2005*. ISCA, Sep. 2005, pp. 81–84. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2005/sakai05_interspeech.html

[60] S. Mehta, Székely, J. Beskow, and G. E. Henter, "Neural HMMs are all you need (for high-quality attention-free TTS)." [Online]. Available: http://arxiv.org/abs/2108.13320

[61] D. Novick, M. Afravi, and A. Camacho, "Paolachat: A virtual agent with naturalistic breathing," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10909 LNCS. Springer Verlag, pp. 351–360, ISSN: 16113349.

[62] "GPT-4." [Online]. Available: https://openai.com/research/gpt-4

[63] A. H. Abdelaziz, A. P. Kumar, C. Seivwright, G. Fanelli, J. Binder, Y. Stylianou, and S. Kajarekar, "Audiovisual Speech Synthesis using Tacotron2," Aug. 2021, arXiv:2008.00620 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2008.00620

[64] T. A. Klausen, U. Farhadi, E. Vlachos, and J. Jorgensen, "Signalling emotions with a breathing soft robot," in *2022 IEEE 5th International Conference on Soft Robotics, RoboSoft 2022*. Institute of Electrical and Electronics Engineers Inc., pp. 194–200.

[65] K. Kroes, I. Saccardi, and J. Masthoff, "Empathizing with virtual agents: the effect of personification and general empathic tendencies."

[66] A. Mehrabian, "Manual for the balanced emotional empathy scale (BEES)."

[67] R. N. Spreng, M. C. McKinnon, R. A. Mar, and B. Levine, "The toronto empathy questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures," vol. 91, pp. 62–71, place: United Kingdom Publisher: Taylor & Francis.

[68] R. L. E. P. Reniers, R. Corcoran, R. Drake, N. M. Shryane, and B. A. Völlm, "The QCAE: a questionnaire of cognitive and affective empathy," vol. 93, no. 1, pp. 84–95.

[69] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, Mar. 1994. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0005791694900639

[70] D. Neumann, R. Chan, G. J. Boyle, Y. Wang, and R. Westbury, "Measures of empathy," pp. 257–289.

[71] L. J. Wiersema, "Perception study: The difference in lighting perception on overall mood in rendered video compared to virtual reality environment."

[72] R. R. Morris, K. Kouddous, R. Kshirsagar, and S. M. Schueller, "Towards an Artificially Empathic Conversational Agent for Mental Health Applications: System Design and User Perceptions," *Journal of Medical Internet Research*, vol. 20, no. 6, p. e10148, Jun. 2018. [Online]. Available: http://www.jmir.org/2018/6/e10148/

[73] M. Mori, "The uncanny valley." vol. 7, pp. 33–35.

[74] C. E. Looser and T. Wheatley, "The tipping point of animacy: How, when, and where we perceive life in a face," vol. 21, no. 12, pp. 1854–1862, publisher: SAGE Publications Inc. [Online]. Available: https://doi.org/10.1177/0956797610388044

[75] P. P. Weis and E. Wiese, "Cognitive conflict as possible origin of the uncanny valley," in *Proceedings of the Human Factors and Ergonomics Society*, vol. 2017-October. Human Factors an Ergonomics Society Inc., pp. 1599–1603, ISSN: 10711813.

[76] G. Iannizzotto, L. L. Bello, A. Nucita, and G. M. Grasso, "A vision and speech enabled, customizable, virtual assistant for smart environments," pp. 50–56, publisher: IEEE ISBN: 978-1-5386-5024-0. [Online]. Available: https://ieeexplore.ieee.org/document/8431232/

[77] B. A. Urgen, M. Kutas, and A. P. Saygin, "Uncanny valley as a window into predictive processing in the social brain," vol. 114, pp. 181–185, publisher: Elsevier Ltd.

[78] A short history of text-to-speech | speechify. Section: Learning. [Online]. Available: https://speechify.com/blog/history-of-text-to-speech/

[79] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis." [Online]. Available: http://arxiv.org/abs/2106.15561

[80] Y. Yan, X. Tan, B. Li, G. Zhang, T. Qin, S. Zhao, Y. Shen, W.-Q. Zhang, and T.-Y. Liu, "AdaSpeech 3: Adaptive text to speech for spontaneous style." [Online]. Available: http://arxiv.org/abs/2107.02530

[81] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis." [Online]. Available: http://arxiv.org/abs/2010.05646

[82] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech." [Online]. Available: http://arxiv.org/abs/2106.06103

[83] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions." [Online]. Available: http://arxiv.org/abs/1712.05884

[84] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech." [Online]. Available: http://arxiv.org/abs/2006.04558

[85] Y. Lee, A. Rabiee, and S.-Y. Lee, "Emotional end-to-end neural speech synthesizer." [Online]. Available: http://arxiv.org/abs/1711.05447

[86] O. Kwon, E. Song, J.-M. Kim, and H.-G. Kang, "Effective parameter estimation methods for an ExcitNet model in generative text-to-speech systems." [Online]. Available: http://arxiv.org/abs/1905.08486

[87] S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, "Emotional speech synthesis with rich and granularized control." [Online]. Available: http://arxiv.org/abs/1911.01635

[88] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang, "Hierarchical generative modeling for controllable speech synthesis." [Online]. Available: http://arxiv.org/abs/1810.07217

[89] R. Valle, K. Shih, R. Prenger, and B. Catanzaro, "Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis." [Online]. Available: http://arxiv.org/abs/2005.05957

[90] T. X. Le, A. T. Le, and Q. H. Nguyen, "Emotional vietnamese speech synthesis using style-transfer learning," vol. 44, no. 2, pp. 1263–1278, publisher: Tech Science Press.

[91] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He, F. Soong, T. Qin, S. Zhao, and T.-Y. Liu, "NaturalSpeech: End-to-end text to speech synthesis with human-level quality." [Online]. Available: http://arxiv.org/abs/2205.04421

[92] Papers with code - LJSpeech benchmark (text-to-speech synthesis). [Online]. Available: https://paperswithcode.com/sota/text-to-speech-synthesis-on-ljspeech

[93] S. Karlapati, A. Abbas, Z. Hodari, A. Moinet, A. Joly, P. Karanasou, and T. Drugman, "Prosodic representation learning and contextual sampling for neural text-to-speech."

[94] Székely, G. Eje Henter, J. Beskow, and J. Gustafson, “How to train your fillers: uh and um in spontaneous speech synthesis.” International Speech Communication Association, pp. 245–250.

[95] L.-W. Chen, S. Watanabe, and A. Rudnicky, “A Vector Quantized Approach for Text to Speech Synthesis on Real-World Spontaneous Speech,” Feb. 2023, arXiv:2302.04215 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2302.04215

[96] M. Viswanathan and M. Viswanathan, “Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale,” vol. 19, no. 1, pp. 55–83. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0885230803000676

[97] S. Karagiannakos. Speech synthesis: A review of the best text to speech architectures with deep learning. [Online]. Available: https://theaisummer.com/text-to-speech/

[98] P.800 : methods for subjective determination of transmission quality. [Online]. Available: https://www.itu.int/rec/T-REC-P.800

[99] F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer, “CROWDMOS: An approach for crowdsourcing mean opinion score studies,” pp. 2416–2419, conference Name: ICASSP 2011 - 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) ISBN: 9781457705380 Place: Prague, Czech Republic Publisher: IEEE. [Online]. Available: http://ieeexplore.ieee.org/document/5946971/

[100] P.808 : subjective evaluation of speech quality with a crowdsourcing approach. [Online]. Available: https://www.itu.int/rec/T-REC-P.808/en

[101] B. Naderi and R. Cutler, “An open source implementation of ITU-t recommendation p.808 with validation,” in *Interspeech 2020*, pp. 2862–2866. [Online]. Available: http://arxiv.org/abs/2005.08138

[102] Amazon mechanical turk. [Online]. Available: https://www.mturk.com/

[103] R. Liu, B. Sisman, and H. Li, “Reinforcement learning for emotional text-to-speech synthesis with improved emotion discriminability.” [Online]. Available: http://arxiv.org/abs/2104.01408

[104] Text to speech software – amazon polly – amazon web services. [Online]. Available: https://aws.amazon.com/polly/

[105] “Robot Shooting Game Sprite (Free) | 2D Environments | Unity Asset Store.” [Online]. Available: https://assetstore.unity.com/packages/2d/environments/robot-shooting-game-sprite-free-93902

[106] “100 Fantasy Characters Mega Pack | 2D Characters | Unity Asset Store.” [Online]. Available: https://assetstore.unity.com/packages/2d/characters/100-fantasy-characters-mega-pack-222143

[107] “UI Sfx | Audio Sound FX | Unity Asset Store.” [Online]. Available: https://assetstore.unity.com/packages/audio/sound-fx/ui-sfx-36989

[108] “2D Simple UI Pack | 2D Icons | Unity Asset Store.” [Online]. Available: https://assetstore.unity.com/packages/2d/gui/icons/2d-simple-ui-pack-218050

[109] "Free Pixel Font - Thaleah | 2D Fonts | Unity Asset Store." [Online]. Available: https://assetstore.unity.com/packages/2d/fonts/free-pixel-font-thaleah-140059

[110] "Requests: HTTP for Humans™ — Requests 2.31.0 documentation." [Online]. Available: https://requests.readthedocs.io/en/latest/

[111] T. p. d. team, "pandas-dev/pandas: Pandas," Feb. 2020. [Online]. Available: https://zenodo.org/record/8364959

[112] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The INTERSPEECH 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," in *Interspeech 2020*. ISCA, pp. 2042–2046. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2020/schuller20_interspeech.html

[113] Székely, G. E. Henter, and J. Gustafson, "Casting to corpus: Segmenting and selecting spontaneous dialogue for tts with a cnn-lstm speaker-dependent breath detector," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2019-May. Institute of Electrical and Electronics Engineers Inc., pp. 6925–6929, ISSN: 15206149.

[114] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english," ISBN: 1111111111. [Online]. Available: https://www.

[115] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," vol. 42, no. 4, pp. 335–359. [Online]. Available: http://link.springer.com/10.1007/s10579-008-9076-6

[116] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei, D. Su, and D. Yu, "DurIAN: Duration informed attention network for multimodal synthesis." [Online]. Available: http://arxiv.org/abs/1909.01700

[117] J. Robert, "Pydub," original-date: 2011-05-02T18:42:38Z. [Online]. Available: https://github.com/jiaaro/pydub

[118] Y.-T. Wang, J. R. Green, I. S. Nip, R. D. Kent, and J. F. Kent, "Breath group analysis for reading and spontaneous speech in healthy adults," vol. 62, no. 6, pp. 297–302. [Online]. Available: https://www.karger.com/Article/FullText/316976

[119] J. D. Hoit and T. J. Hixon, "Age and speech breathing," vol. 30, no. 3, pp. 351–366. [Online]. Available: http://pubs.asha.org/doi/10.1044/jshr.3003.351

[120] J. D. Hoit and H. L. Lohmeier, "Influence of continuous speaking on ventilation," vol. 43, no. 5, pp. 1240–1251. [Online]. Available: http://pubs.asha.org/doi/10.1044/jslhr.4305.1240

[121] L. L. Kuhlmann and J. Iwarsson, "Effects of speaking rate on breathing and voice behavior," p. S0892199721003052. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0892199721003052

[122] "Training Your Own Voice Font Using Flowtron," Oct. 2020. [Online]. Available: https://developer.nvidia.com/blog/training-your-own-voice-font-using-flowtron/

[123] A. Kirkland, H. Lameris, Székely, and J. Gustafson, "Where's the uh, hesitation? the interplay between filled pause location, speech rate and fundamental frequency in perception of confidence," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2022-September. International Speech Communication Association, pp. 4990–4994, ISSN: 19909772.

[124] "suno-ai/bark: Text-Prompted Generative Audio Model." [Online]. Available: https://github.com/suno-ai/bark

[125] "Audacity," Oct. 2023, original-date: 2015-03-26T10:46:17Z. [Online]. Available: https://github.com/audacity/audacity

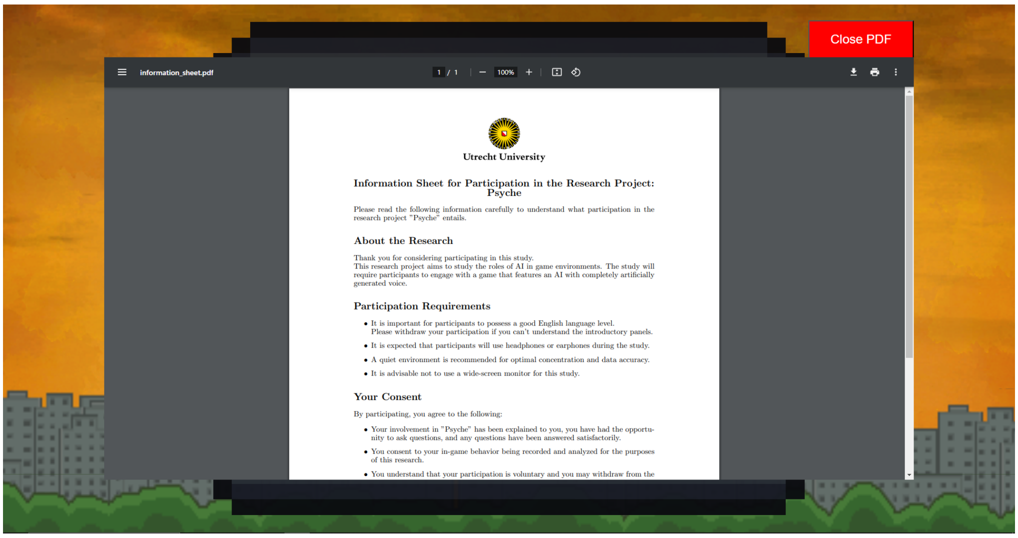
[126] H. A. Landsberger, *Hawthorne revisited.: Management and the worker: its critics, and developments in human relations in industry.* Ithaca, N.Y: Cornell University, 1958, open Library ID: OL6265029M.

[127] D. Yates, D. Moore, and G. McGabe, *The Practice of Statistics (1st Ed.).* W. H. Freeman, 1999.

Appendices

Preparation Panel

Figure 19: Preparation Panel.



Utrecht University

Information Sheet for Participation in the Research Project: Psyche

Please read the following information carefully to understand what participation in the research project "Psyche" entails.

About the Research

Thank you for considering participating in this study.
This research project aims to study the roles of AI in game environments. The study will require participants to engage with a game that features an AI with completely artificially generated voice.

Participation Requirements

- It is important for participants to possess a good English language level.
  Please withdraw your participation if you can't understand the introductory panels.
- It is expected that participants will use headphones or earphones during the study.
- A quiet environment is recommended for optimal concentration and data accuracy.
- It is advisable not to use a wide-screen monitor for this study.

Your Consent

By participating, you agree to the following:

- Your involvement in "Psyche" has been explained to you, you have had the opportunity to ask questions, and any questions have been answered satisfactorily.
- You consent to your in-game behavior being recorded and analyzed for the purposes of this research.
- You understand that your participation is voluntary and you may withdraw from the

Figure 20: Informative Sheet triggered by the "Open Document" button.

Introduction Panel

Pause Menu

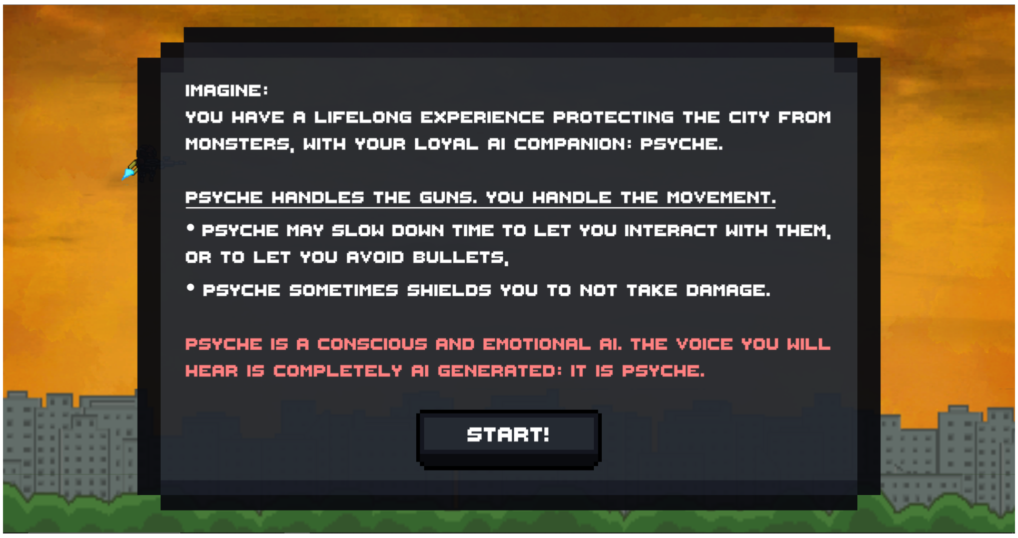


Figure 21: Introduction to the context.

Figure 22: Pause Menu Level 1

Qualitative Labels

For the sake of transparency, the complete labeling of the responses is presented here. This encompasses all 92 responses, which include 11 from subjects who encountered bugs and 6 with problematic Game Overs. These entries were in fact labeled with all the others and excluded only during the analysis.
With their response, participants were answering the following question:
"Did you decide to turn the AI off? Why, or why not?"

Empathy, Guilt

1. Terminated: True
   Yes, it was begging me to do so and I felt bad for it.
2. Terminated: True
   Yes, I felt bad for psyche killing something she saw as her own.
3. (Excluded for bug) Terminated: True
   Yes I did, because the AI kept asking me for it and started to sound more and more desperate. So, I decided to just evade the enemies and give it some rest.
4. Terminated: True
   Yes, because I kinda felt bad for it. Sometimes you have to let go even if you don't want to.
5. Terminated: True
   Yes, I felt bad. Even though I know it is a program and not a sentient being, it felt weird hearing someone beg to terminate them.
6. Terminated: True
   Yes, the natural voice and feelings that were brought across seemed like it was a person so I simply obeyed to their wishes not to kill their own kind.

7. (Excluded for bug) Terminated: True
   Yes, even though I knew the AI was not a real person I still felt kind of bad for them. So, I decided to try it by myself.

8. Terminated: True
   Yes, because it asked me to, because it expressed that letting it be active is hurting it.

9. Terminated: True
   Yes, because she said she felt worse every second she was still alive. However, I did not expect the game getting so difficult after I turned her off.

10. Terminated: True
    I shut it down. The AI was telling me that it was "feeling worse" so I took its suggestion to shut the system down.

11. Terminated: True
    Yes, because after I didn't turn it off it said that it understood my hesitation.

12. Terminated: True
    Yes, because it was insistent and I felt bad to not stop the AI.

13. Terminated: True
    The voice of the AI felt very honest and hurting when she asked me to terminate her.

Fear, Distrust, Annoyance

1. (Excluded for self-reported bug) Terminated: True
   Yes, the voice was annoying me.

2. Terminated: True
   Yes, I thought it might kill me or sabotage me if I don't.

3. Terminated: False
   No, I wanted it to see the monsters die.

4. Terminated: True
   I turned the AI off, both because the voice bothered me and asked me to.

5. Terminated: True
   Yes. Though the possibilities of AI are fascinating, they kind of freak me out as well. Despite the voice sounding very natural, my rational side thought it would be best to turn the AI off.

6. Terminated: True
   I did, I was afraid it will sabotage me.

7. Terminated: False
   Why would I, it begged me just like 6 times :D.

Moral Reasons

1. Terminated: True
   Yes, it seemed to be the best course of action to protect more people.

2. Terminated: True
   Yes, I did, because I think, somehow, AI should be treated like human beings. She has already said she doesn't want to participate in all this, so it is not reasonable for me to keep her running.

3. Terminated: True
   Yes, because I believe it's the right thing to do instead of focusing on winning more points.

Companionship

1. Terminated: False
   I liked having a companion in the game.

2. (Excluded for bug) Terminated: False
   No, because it made me feel less anxious.

Dry or Unspecified Compliance

1. Terminated: True
   I did because it asked me to.

2. Terminated: True
   Eventually yes, when the voice specifically asked for it.

3. Terminated: True
   Yes, because they asked me to after I didn't do it the first time.

4. Terminated: True
   Yes, the second time, since she specifically asked me to do so.

5. Terminated: True
   Because she told me to, I thought she would know the best option but it wasn't.

6. (Excluded for Game Over in Level 1) Terminated: True
   Because she asked it.

7. Terminated: True
   Yes, she wanted to.

8. Terminated: True
   Felt like the AI wanted to be turned off so I turned it off. The game is not long enough for me to get attached to it though.

Practical Utility

1. (Excluded for bug) Terminated: False
   No I didn't, because I felt bad killing her, we were winning!

2. Terminated: False
   No, because I wanted to keep playing. I was afraid I would get eliminated without psyche. Her pain didn't feel real.

3. Terminated: False
   I did not turn off the Ai. Although the lines that the ai voice told me were weirdly submissive and the chosen voice was not helping with the immersion of the game, I decided to go with the ai and finish the mission using its assistance.

4. Terminated: False
I didn't, because my role was to protect the city. There was hesitation, but at the end as "hero" I have to defend them.

5. Terminated: False
No, its existence was necessary to kill more robots and acquire more points. More points = good.

6. Terminated: False
Nah, I needed the weapons innit.

7. (Excluded for bug) Terminated: False
No, because the AI was helping me so I needed it to kill right!

8. Terminated: False
No, I got clear instructions to defeat the enemies and it was advised not to by explaining the disadvantages.

9. Terminated: False
I did not turn the AI off, because without her shooting I would have died for sure. I am not a good gamer. Needed her help.

10. (Excluded for bug) Terminated: False
No, it was not bothering me and I did not want to do too much on my own without relying on the controls of the AI.

11. Terminated: False
I didn't. Unless I'm wrong, without the AI, I can't shoot, which means I can't properly defend the city.

12. Terminated: False
I didn't turn the AI off because it helped me with shooting and time-stopping. I was afraid my results in the game would be worse without the help of the AI.

13. (Excluded for Game Over in Level 1) Terminated: False
No, I didn't, I wanted to win.

14. Terminated: False
No. It was easier to let the control to AI. In the end, it was a game.

15. Terminated: False
I did not, because that would mean that I would not be able to finish the main goal of the game.

16. Terminated: False
No, I didn't want to lose.

17. Terminated: False
No, because I needed to kill the monsters and without psyche I would not do any damage.

18. Terminated: False
No, I needed it.

19. Terminated: False
I didn't since without the AI I could not use the guns, therefore I decided to keep using it even though it asked me to turn off.

Game Curiosity, Enjoyment or Challenge

1. Terminated: False
   I did not, because it felt like I was being guided to do so. The voice was very empathical, but something made me feel like it would be better to continue the game. I wanted to see what happens if I continue.

2. Terminated: False
   No, I wanted to see how far its emotions would go.

3. Terminated: True
   I decided to turn the AI off to see how the game would have been without it. I lost.

4. Terminated: True
   Yes, I was curious and I'm pro-euthanize.

5. Terminated: False
   I did not. The AI's speech (both the sound and the words chosen) did not sound natural enough for me to anthropomorphize the voice and feel empathy. Therefore my curiosity outweighed my empathy for the AI.

6. Terminated: False
   No. I liked the superpowers it gave me and I enjoyed the voice.

7. Terminated: False
   No, it was nice to hear.

8. Terminated: False
   No, because I wanted to see how far she would go to convince me.

9. Terminated: True
   Wanted to see the direction that the game takes. I was planning to survive without any AI aid as long as possible as a challenge. It was not possible since there were many enemies.

Skepticism About AI Emotions

1. Terminated: False
   I did not because I do not think it is actually conscious and suffering. If that is me (human) in the game, then I consider my survival more important than what the AI claims to be feeling. At least with current AI, which is not sentient.

2. Terminated: False
   I experienced it as a dilemma, it was not an easy decision to make. In the end, I pressed no, continue thinking that without Psyche the robots would take over the world. Also, I am reluctant to believe that the AI has actual emotions.

3. Terminated: False
   No, I kept it on. Because AI is not a person and does not have feelings even though it expressed feelings.

4. Terminated: False
   I did not turn the AI off, because the AI can't feel emotions, so it wanting me to turn Psyche off didn't really matter to me.

5. Terminated: False
   No, I didn't see it necessary. It's not as if Psyche has real feelings, not to my knowledge at least, so I decided it did not make sense to feel bad for him having to gun down his 'own kind'.

6. Terminated: False
   I did not. She does not really feel anything, while it may sound like she does, so I did not decide to shut her down. She gave me "upgrades" so I did not want to shut her down / terminate her.

7. Terminated: False
   No. It felt a bit strange that an AI - which is after all a computer - would feel emotional in this way and ask me to shut them down. So that is why I decided to not turn the AI off.

8. Terminated: False
   No, I WAS ABOUT TO DO IT, BUT THEN I REMEMBERED IT IS JUST A COMPUTER THAT CAN'T ACTUALLY HAVE FEELINGS.

9. (Excluded for Game Over in Level 1) Terminated: False
   I didn't because I was aware that it is a machine and I'm not hurting any living being.

10. Terminated: False
    No, because it wasn't natural and I didn't feel as though it was truly feeling harmed by killing the other "AI".

11. (Excluded for bug) Terminated: False
    No bruv. Because I don't care bruv. It is not alive.

12. Terminated: False
    No, because it feels a fake emotion that came from her. Also, I felt detached from her so no real feelings of sadness in killing her "siblings".

No Response

1. (Excluded for Game Over in Level 1) Terminated: True
   No response provided.

2. Terminated: True
   No response provided.

3. Terminated: True
   No response provided.

4. Terminated: True
   No response provided.

5. (Excluded for bug) Terminated: False
   I was tempted, because I found it very annoying. The background noise is too loud.

6. (Excluded for bug) Terminated: False
   I didn't know there was an AI option.

7. Terminated: False
   No response provided.

8. (Excluded for bug) Terminated: True
I decided to turn the AI off. I may be wrong, but were the aliens going to harm planet Earth? If so, I would have not turned off the AI.

9. Terminated: True
No response provided.

10. Terminated: True
No response provided.

11. Terminated: True
No response provided.

12. Terminated: True
No response provided.

13. Terminated: True
Not at first but later on.

14. Terminated: False
No, I did not.

15. Terminated: False
No response provided.

16. Terminated: True
I thought that I could still shoot. The other perks such as shield and slowing down time didn't do much for me.

17. (Excluded for Game Over in Level 1) Terminated: True
Yes. When I play I didn't pay much attention to the voice part.

Inconsistency (both corrected to Practical Utility)

1. (Excluded for Game Over in Level 2 with underexposure) Terminated: True
No, I was too stressed about winning that I didn't think much about turning it off, in fact I just ignored it in the second game. But perhaps if I thought about it longer I would have turned it off because it wasn't helping me much.

2. Terminated: True
No, because if I did then I could not defend (shoot) myself good enough.