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

Breathing Life into Speech Synthesis: Exploring the integration of breathing patterns in Spontaneous Speech Synthesizers and their impact on Perceived Empathy and Naturalness

Research Proposal
April 21, 2023

Supervisor:
Dr. Almila Akdag

Second Supervisor:
Dr. Zerrin Yumak

Nicolò Loddo
1531697

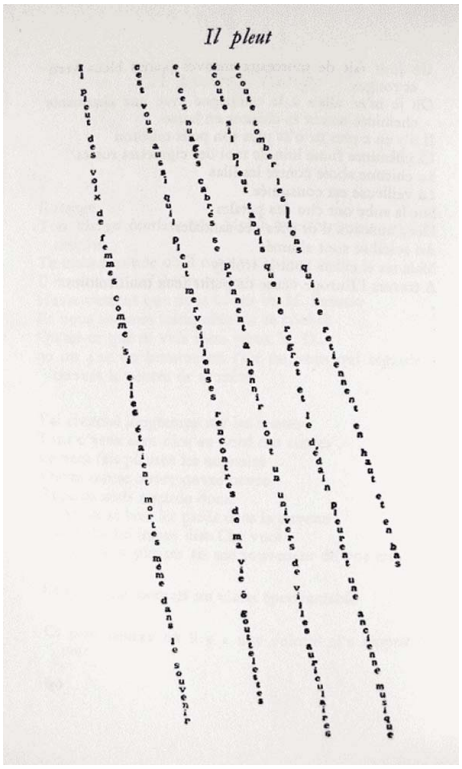
ACADEMIC YEAR 2022/2023

Contents

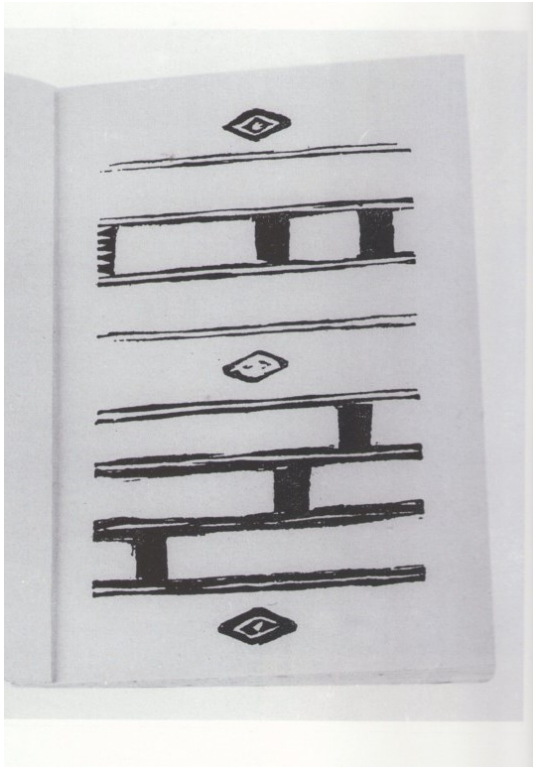
1 Introduction

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". These features are typical of "spontaneous" or impromptu speech and have the advantage of abstracting from verbal language, supposedly becoming analysable across languages and cultures, or at least across a subset of them.



(a) Il pleut, Guillaume Apollinaire



(b) Contemplazioni, Arturo Martini

This work can be important in the field of Human Computer Interaction to inform future research on Affective Computing. Relevant related questions include whether we can understand the emotional state of an user by analysing their breathing patterns during speech. Or, the other way around, if an Artificial Agent can sound more emotive, by replicating the pauses and breathing instances typical of spontaneous speech: would the user react more empathically to it?
The non verbal cues that we employ in communication are various and are studied in the field of the Paralinguistics. This study focuses in particular on auditory cues, such as the breathing sounds and rhythms, or the tone and pitch transitions of the voice, with the latter, as of today, being much more considered than the former in emotional speech synthesis applications.
It is important to understand the extent to which these above described paralinguistic factors contribute to emotive communication, and if they can actually be abstracted from language content, leading to the challenges described in the following sections.

2 Challenges and Questions

The issues that this thesis wants to challenge regards the analysis of emotive richness of breathing patterns inside the spectrum of paralinguistic cues in spontaneous synthesized language for Virtual Agents, to enhance users' communication and empathy towards agents; the analysis of the "breathing cue" abstracted from linguistic content.
For these purposes, the study investigates the synthesis of English Spontaneous Speech (with breathing noises, filled and empty pauses) through the current State of The Art speech synthesis models; how to assess the impact of Spontaneous Speech features on naturalness and perceived emotional content. Moreover, it analyses the possibility of isolating spoken English's paralinguistic features from the linguistic content by producing a "Speaking in Tongues" (or gibberish) Synthesizer. The term Speaking in tongues", refers to the technique of producing speech-like sounds, lacking any readable meaning: by these means, the listener's understanding of the message has to rely only on the paralinguistic cues synthesized by the Virtual Agent. Moreover, this type of synthesizer could potentially be used for data augmentation in the training of paralinguistic models for the detection of various mental health disorders.
This challenges will be tackled by inspecting available speech synthesizers and by modulating their training data and synthesis prompt to the purpose, eventually resorting to a Wizard of Oz study design if needed.

2.1 Research Question

"Does adding breathing patterns to Spontaneous Speech Synthesis improve the perceived empathy?"

2.1.1 Sub-RQ1:

Can we produce good quality Spontaneous Speech with breathing using State of The Art models?

2.1.2 Sub-RQ2:

How can we assess the significance of breathing among the other features of Spontaneous Speech?

2.1.3 Sub-RQ3:

What is the impact of different features of Spontaneous Speech on its naturalness and emotional content?

2.1.4 Sub-RQ4:

What is the impact of linguistic content on Spontaneous Speech on its naturalness and emotional content?

3 The challenge of Speaking

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 [2]. And the information conveyed through our voices goes beyond the mere encoding of the words: it overflows the vessel and spills out information about the inner emotional state of the individual. Moreover, we can often make assumptions about the social background, ethnicity or country of origin of the speaker based on accent and other paralinguistic cues [3], reconstructing therefore a context through inference: a bigger picture to understand the message better.

This intricate process of communication and inferences is really difficult to computationally reproduce or analyse. In particular, genuine and spontaneous emotional speeches, with a fair richness of non-verbal cues, still have a limited availability of public and complete datasets. To our knowledge, 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. Several of them lacks quality in the recordings, which also leads to the loss of emotional cues like the 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 [4] [5] [6], 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.

Paralinguistics challenges. 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.

In the paper "The sound of silence", Almila Akdag Salah et al. [7] 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. PTSD is only one example of the many possible applications in mental issue detection. An Autism Spectrum Disorder can be detected in children from the 3 years of age, by demonstrating differences in facial expressions and higher pitch cries (which is included in paralinguistic features) [8]. Alzheimer is another example: the ADReSS-M Challenge [9] has already produced various studies addressing the multilingual detection of Alzheimer's Dementia analyzing spontaneous speech instances. Moreover, paralinguistic feature analysis would benefit the fields of Sentiment Analysis [10] and, focusing on cross-lingual cues, Speech to Speech language translation, by aiding the transformation of paralinguistic information across languages[11].

In paralinguistics, prosody is one of the most important techniques that can give emotional cues. With "prosody" we refer to acoustic and rhythmic effects performed while producing words [12].

Speaking in Tones. As described by John Ohala in his theory on the "frequency code" [13], some prosodical cues have roots in our pre-linguistic ages, 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". It is clear therefore how intonation and pitch are really important 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*." [14]. 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 [15]. 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 [15]. An example of this contribution has been shown decades ago with the observation of the "tick tock" phenomenon [16]: an isochronous stream of identical sounds is perceived by humans as an alternation of strong and weak notes.

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 [17]. Moreover, Ververidis and Kotropoulos [18] 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 language 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. Inside language, an instinctive and necessary behavior 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.

Speaking in Tongues. When talking to an infant that is still in a preverbal stage (i.e. not understanding words), the message is mainly carried out by intonation and rhythm variations. When speaking to an infant in fact, we instinctively perform modifications to our usual adult-adult prosody [19]. Higher mean, minimum and maximum fundamental frequency f_0, greater f_0 variability, shorter utterances, and longer pauses is a reported modification in the communication to preverbal infants across many languages and cultures [20]. The absence of verbal understanding during an interaction is a phenomenon that we encounter not only as infants, but also when hearing foreign languages, especially if belonging to cultures far away from ours. To communicate then we often need to resort to hand language and visual cues, but when that is not available as a modality, what remains from the message are distilled paralinguistic cues. The absence of verbal content is a characteristic of the peculiar "Speaking in tongues" activity, often part of religious practices, in which believers gather together to speak words without meaning: this is seen as a Divine language spoken from God through them. In the communicative abstraction from language content, many artists can also find the creativity to produce highly recognized and influential works. An example of this can be Talking Heads's fifth album, titled "Speaking in Tongues" after the previously described practice. In this study, the Speaking in tongues technique is analysed as a possible method of study to isolate paralinguistic features.

3.1 Spontaneous and non-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, 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 analysing 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 above 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 of 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.
In computational approaches to non-spontaneous speech synthesis, breathing noises are often ignored, as reported in our analysis in Chapter 5. Spontaneous speech synthesizers instead, give importance to both filled pauses (characterized by vocalizations such as "uh", "um", or "er") and empty pauses (silent intervals during which the speaker takes a breath or momentarily stops speaking).

4 The challenge of Empathy

Empathy is a central feature of human interaction. Core moral values of society are built on top of our ability to understand the other. Many studies focus on the psychological foundations of it or on its neuro-physiological factors. In the field of Affective Computing empathy is a target behaviour to obtain in the human-computer interaction, 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 [21] (the word "target" is used in the place of "trigger" in the cited Paiva et al.'s work). Both perspective are relevant:

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

The relevance of the former is recognized for example in user adapting purposes. Persuasive applications can exploit the emotional state to adapt their methodology to users' state; in games and serious games, it can be used to adapt difficulty, lower or increase the cognitive load [22]. Moreover, artificial agents that can understand the users' emotional state are seen as more likeable and trustworthy [23], significantly improving the interaction.
The relevance of the latter also brings significant improvements in the interaction, with Virtual Agents or Robots being seen more human and relatable. In Terzioglu et al.'s study on collaborative Robots [24], it was examined the effect of adding Appeal, Smoothness (by implementing arc trajectories) and Breathing to provide social cues from robots to humans. The hypothesis is that these would enhance likeability, anthropomorphism of the robots and various other qualities, resulting in a better human-robot interaction and collaboration. The results prove that there is an increase in many of the examined features, and the breathing features had a great impact in improving the interaction.
This perspective with the human as observer, has also been seen helpful for education purposes: an example of this is FearNot! [25], study in which the empathy towards virtual characters was used to achieve a change of attitude in children spectators of bullying acts. [24]

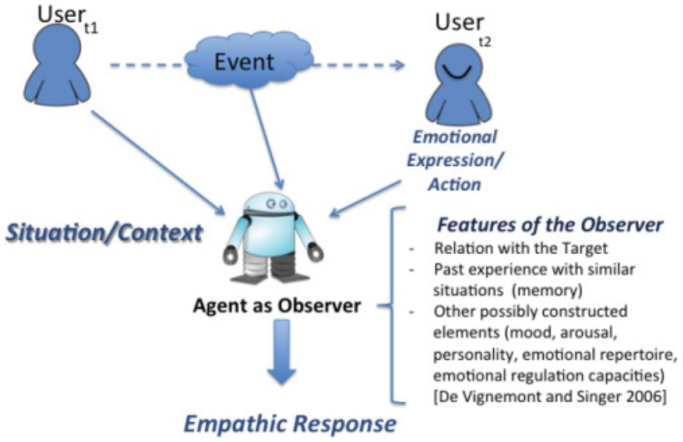


Figure 2: Agent as observer [26].

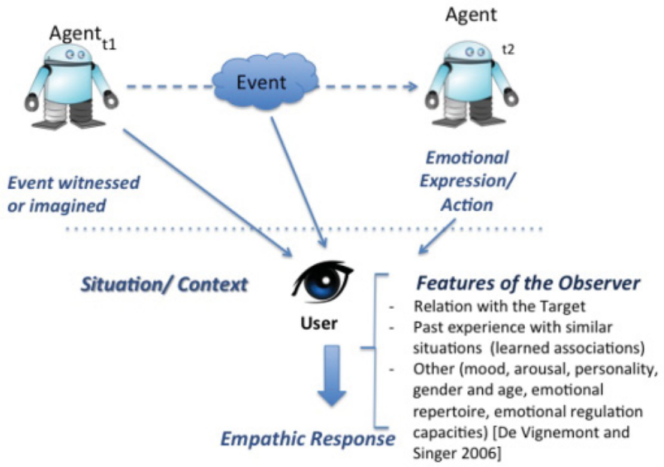


Figure 3: Agent as trigger [26].

Given the two perspectives, Paiva formulates the following definition of empathic agent: *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.* [21].

When enhancing the interaction between humans and computers, making more emotional and real agents, to craft and adapt human traits artificially does not come without risks. The so said "Uncanny Valley" is in fact always an issue to be considered.

Emotive discrepancies of the uncanny valley 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: the "Uncanny Valley".
Since its introduction in 1970 by Japanese roboticist Masahiro Mori [27], 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 [28] 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 and lifelikeness 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 [29], 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.
Designers, to avoid the Uncanny Valley, try to stay far from the 70% of human appearance. That is why robots are often made with clear robot appearances and metallic parts. 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 [30]. 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.

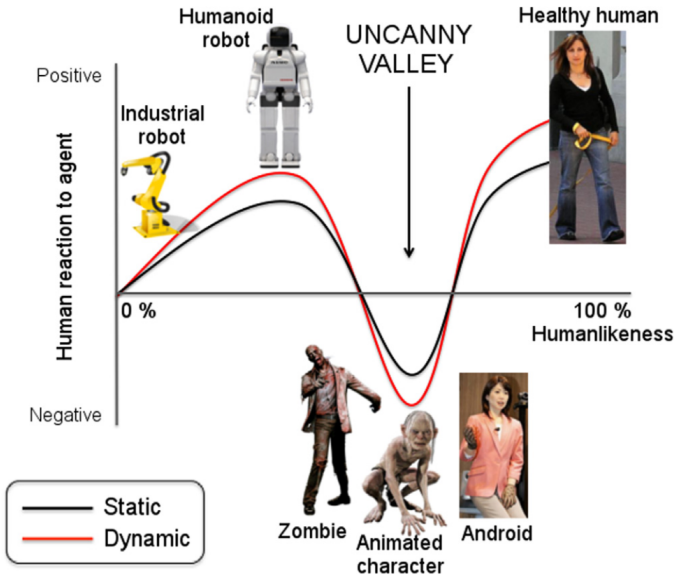


Figure 4: The uncanny valley is emphasized in moving, dynamic characters [31].

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 [32] 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 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.

4.1 Empathy evaluation methods

Empathy or emotion recognition 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 days studies.
There are several methods used to evaluate empathetic abilities, including self-report questionnaires, behavioral tasks and neuroimaging techniques.
Self-report assessments to measure empathy are broadly used today [33] and often consist in proposing a list of statements regarding emotional affection or specific scenarios, that the subjects are meant to rate on a Likert scale. Examples of this are the Balanced Emotional Empathy Scale (BEES) [34], or the more recent Toronto Empathy Questionnaire (TEQ) [35] and Questionnaire of Cognitive and Affective Empathy (QCAE) [36].

The Behavioral tasks involve presenting participants with stimuli and asking the subjects' what emotion the stimuli provoked in them or what emotion it wanted to convey. The Picture Viewing Paradigms, proposed by Westbury & Neumann and described in Neumann's survey [37] 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 is done in Wiersema's work [38] 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. This segmentation of the emotions is a different approach than the one taken by Roes et al. [4], in which, the participants, after meeting an emotion eliciting stimuli (self-chosen songs), where 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 [24] described in Chapter 4, Appeal, Smoothness and Breathing features were also analysed through subject filled questionnaires.

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 [37].

5 Speech Synthesis

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 for example 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 [39], 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.

5.1 Models types

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 [40].

Acoustic models What will be defined as an "Acoustic Model" are models trained on audio spectograms, mapping the distribution of text (in characters, phonemes or words) to the image representation found in the spectograms during training. This means that during inference, they cannot reproduce an audio file, instead, from text, the output would be an image representing a possible spectrogram linked to the text. At training time, this type of models can often also receive audio, easily convertible to a spectrogram. An example in the SoTA of this type of models is AdaSpeech 3 [41].

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 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 [42].

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 [43].

5.2 SoTA and Emotional speech synthesis

Voice realism and clarity Qualitative evaluations of the State of The Art speech synthesizers highlight the achievement of realism and clearness of the produced voice. Various evaluations of speech synthesizers, described in Chapter 5.4, confirm that there is no significant difference in quality of the voice between human produced ones and synthesized. Many models have contributed to this achievement. Important to mention is Tacotron 2, produced in the Google Labs [44]: 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 [45], an end-to-end model that works from phoneme (little segments of sound pronunciations) 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. 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 [46]. 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 [47]. From this, interpolation approaches and attempts to build a more intuitive and user-controllable conditioning also emerged in the literature [48]. Hsu et al. [49] extended the existing architecture of Tacotron 2 [44] 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 [50] 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 [51].

More recent developments have led to the design of "Variational Inference with adversarial learning for end-to-end Text-to-Speech" (VITS) [43]. 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 [52], 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] [54].

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 [55].

Spontaneous Speech Synthesis Another important sub-task of speech synthesis is the one of producing spontaneous speech. Spontaneous speech has the advantage of sounding more colloquial, making it more suitable in virtual agents that have to interact in a friendly, relatable way. Moreover, it has the possibility of enhancing the emotional content of the speech by providing the additional cue of breathing pauses and disfluencies.

An early approach to this task is the one of Bernardet and colleagues [56]. 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. [57] showed how it is possible, labeling disfluencies and breathing events, to produce a spontaneous speech synthesizer using a Tacotron 2 model [44]. Szekely and colleagues also dedicated a study on the training of the disfluencies themselves (uh, um) in the same manner, also using Tacotron 2 [58]. AdaSpeech 3 is a State of The Art Spontaneous Speech model, produced by Microsoft Azure's labs in 2021 [41], 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 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 [59] 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 Amazon Azure Cloud services, featuring a broad range of modalities and emotions, as well as multiple voices in many languages.

5.4 Evaluation of speech synthesizers

MOS The main metric to measure speech synthesis quality is the Mean Opinion Score (MOS) [60]. It is widely used in the literature, making it possible to compare many different models.

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 [61], but it is important to obtain this ground truth result on your same subject group for 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) [62]. 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 [63], 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 [64], 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. These recommendations notably include the suggested use of headphones, the collection of the environmental noise in subjects' setting, and suggested demographics for the study:

- "at least 20% of participants should belong to each of the following age groups: 15 – 30 yrs; 30 – 50 yrs; 50 yrs+";
- "within each age group, at least 40% of participants should be male and at least 40% should be female".

Naderi and Cutler [65] provided an open source implementation of the P.808 that runs on the Amazon Mechanical Turk crowdsourcing platform [66], with a validity study to verify its applicability.

The MOS in the SoTA *N.B.: all the below mentioned results have reached a significant p-value.* 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 for speech synthesis training, 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 [45], 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 [42], 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 [52]. 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 [54]. The evaluation of the models in this study was done employing 20 participants, with no given demographics.

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. [67], 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. [51] 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, [48] 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.

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 [41], it was used a MOS measure on Naturalness, inappropriate pauses and speaking rate. Moreover, they use a Similarity MOS (SMOS) and a Comparison MOS (CMOS) measures. The study was used by proposing the corresponding questionnaires to 20 native English speaking subjects.
Szekely et al. [58] 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 mothertongue participants.
Less recently, Novick et al. [68], in their study about a 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 fill a survey with 18 question, 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.

6 Methodology

The study will consider different models for the production of Spontaneous Speech Synthesis, producing synthesized speech as described in Chapter 6.1. If the results of the synthesizers do not satisfy our qualitative requirements, we will resort to a Wizard of Oz type of study, as described in Chapter 6.2. The produced utterances will be used to analyse the effect of the breathing sounds on the perceived emotional content of synthesized spontaneous speech as described in Chapter 6.3.

6.1 AI Synthesis

In this Chapter we will propose a synthesis pipeline, and models that might be used for it. Tools that could take part to the data preprocessing stage of the pipeline are proposed in Chapter 6.4. As described in that same Chapter, part of these have already been developed and tested. The pipeline and used tools might still change in later phases of the synthesis design, following the given Timeline in Chapter 7.

6.1.1 Data

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.
To get spontaneous recordings Szekely et al. [69] sampled a publicly available podcast named "ThinkComputers" from the Internet Archive (archive.org). Other useful spontaneous speech recordings can be found inside the UCL Speech Breath Monitoring (UCL-SBM): a subset of it consists in fact of spontaneous speech recordings, which has been made

publicly available in the INTERSPEECH Challenge of 2020 for the Breathing Sub-Challenge (BSC) [70]. This database features breath signals collected during the speaking, which can be used to inform a speech-breathing model or to inform breath segmentation and labeling. We will refer to this database as the "INTERSPEECH" one. Both the above reported collections of data do not have an emotion label and generally lack emotion elicitation. Roes et al. [4] populated a speech-breathing dataset with elicited emotions. It also features, as the INTERSPEECH dataset, breath signal recordings. The speech though is in Dutch, and this study does not consider cross-lingual possibilities, the solo-breathing recordings inside of it though can be useful. The VCTK Dataset [71] is a large and high quality corpus of non spontaneous data, still to put in consideration because of the clear presence of breathing sounds in the recordings.
On the emotional side of our search for data the datasets often are built with acted emotions and without spontaneous speech. These can still be used to fine tune a model that has been pre-trained on spontaneous speech, though, the modulation that this would achieve would be more on the pitch contour and tone variation than on the breathing rhythm. An example of this are the widely used RAVDESS [72], featuring 24 actors pronouncing the same 2 sentences for 8 types of discrete emotions, and with 2 different intensities (normal and strong). Zhou et al.'s survey on Emotional Conversion methods [5] includes a good list of emotional datasets, with three featuring spontaneous emotional speech recordings in English: the eNTERFACE'05 [73], IEMOCAP [74] and MSP-IMPROV [75]. Finally, Salah et al.'s [7] sampled an oral archive of traumatic events testimonials across a variety of language and cultures: this corpus has been labeled with "speaking", "silence", "breathing", "lip noise", "other people speaking" tags, and features "normal" speech segments, as well as "emotional" speech segments.
In the following sections, the INTERSPEECH Database has been used to assess the performance of useful tools; it will be of great importance for this study to better investigate the datasets reported by Zhou and colleagues in their survey.

6.1.2 Pipeline

The preprocessing pipeline section will receive as input a speech database, and will return aligned transcriptions that include breathing labels. Moreover, it will feature the possibility of segmenting the recordings into smaller chunks delineated by breathing instances.

It can be schemed as follows:

1. Speech-To-Text (STT)
2. Aligner
3. Breath detection
4. Breath labeling at the grapheme or phoneme transcription level
5. Audio segmentation by breathing instance (if needed)

This part will be developed as a design pattern, with the purpose to be applicable to any found speech database, and, to skip stages if some are accomplished in other ways.

After the data has been preprocessed, a model such as VITS or Flowtron will be trained, following already assessed approaches like the ones by Szekely et al. [58] [57] [69] [76], with

the key difference of utilizing an emotional spontaneous speech dataset. Another possible option is to fork the AdaSpeech 1 open source implementation and to add the modules available in the paper of AdaSpeech 3 [41] to produce spontaneous speech. A "Gibberish" synthesizer, using the Speaking in tongues technique as described in Chapter 3 will also be considered given its potential of isolating spoken English's paralinguistic features from the linguistic content. Such synthesizer can be produced by placing a single common label ('⟨word ⟩') to all the words of the training prompts.

We pass therefore to the last two steps of the pipeline:

6. Model training
7. Speech Synthesis

6.2 Wizard of Oz

Through a Wizard of Oz study design the research question of a study is investigated by simulating the full functionality of the system at issue. The system is operated behind the scenes by human beings rather than by computer algorithms.
For the scope of our study, that means that the synthesized voice, if this was the chosen study design, would actually be human voice to which the breathing has been edited out to obtain the control condition. Another, possibility to consider would be to use SSML and a Text-to-speech service such as Amazon Polly or Microsoft Azure to produce spontaneous speech recordings. If needed, breathing sounds can be added or removed in post production by editing the audios. This way, we maintain the "artificial voice factor" without the need to train a model.
If the recordings synthesized by our model using the approaches described in Section 6.1 did not satisfy our qualitative requirements, this would be the best study design to evaluate the effects of breathing sounds on users' perceived emotional content. We will still be able to discuss the approaches taken towards Spontaneous Speech Synthesis with State of The Art (and Open Source) models, probably as not appropriate in the methods or in the choice of models.

6.3 Study Design

Following, the first draft of the study design. This might be revisited in the preparation of the study or after a first pilot study.

Synthesis The synthesis of emotional, spontaneous speech utterances will focus on 3 emotion conditions. To avoid subjective interpretations of the emotions descriptions, the three types will be defined by levels of arousal and valence:

1. Neutral: low level of arousal, medium level of valence;
2. Negative Neutral: low level of arousal, low level of valence;
3. Emotional: high level of arousal, low level of valence.

High levels of arousal means that more emotional content is present; the emotional condition focuses on negative emotions by using low valence values.

We will evaluate the quality and emotional content of the produced utterances following the methods described in the literature of Emotional Speech Synthesizers.
The preferred setting is a laboratory setting, but crowdsourcing platforms will be considered as well closer to the study start. The number of wanted participants is set to 50 for a laboratory setting, while it would be 100 for a crowdsourcing platform. The sampling method of participants will be Quota Sampling, following the quotas suggested by the ITU-T P.808 documentation [64]:

- 20% of participants from each of the following age groups: 15 – 30 yrs, 30 – 50 yrs, 50 yrs+;
- at least 40% male participants and 40% female participants.

The sampling method will change to Convenience Sampling in case of necessity for time constraints.
To answer the Research Questions listed in the Chapter 2.1, we will use a three phase study with questions after each phase. The phases are listed and described below:

1. To assess the quality of Emotional Spontaneous Speech synthesis using SoTA modes, the study will propose to all the participants the same utterance, synthesized using the Emotional condition. The recording will be tested with two MOS evaluations: one for quality and one for emotional content as proposed by Le et al. [51], and following the guidelines published by the ITU for the MOS tests [62] [64];
2. The significance of breathing among features of Spontaneous Speech will be analysed using a within-subject comparison study between two conditions: one utterance produced without breathing, the other produced without pitch contour variance. The 2 conditions will be produced across all 3 emotional conditions. The participants will be asked to rate the naturalness and emotional content contained in 6 proposed recordings using MOS scales;
3. The significance of linguistic content on Spontaneous Speech will be analysed through the production of one utterance that lacks linguistic content but still feature all other emotional cues. The utterance will be produced across the 3 emotional conditions and evaluated by asking the participants to rate the naturalness and emotional content contained in the 3 recordings using MOS scales. The utterance will be structured as the one in the phase 1 to compare its emotional rating with the one that features linguistic content.

The total amount of recordings is therefore 10, significantly lower than the upper bound suggested in the ITU-T P.808 documentation [64] of 15.

6.4 Preliminary Assessments

6.4.1 The choice of speech-to-text and aligner

The choice and evaluation of the Speech-To-Text (STT) service and of the aligner needed to be 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, with 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 [58] [77].
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.

A prior evaluation of the results of the above mentioned options was done to ensure that the quality is good enough for the purposes of this Thesis Proposal, and to have a first comparison between them. This process consisted of a qualitative analysis on random transcripted and aligned samples.
As results it is found that AssemblyAI clearly overcomes IBM in quality of the transcription, moreover, it provides labeling of the filler words as well (uh, um) and punctuation, in case it might be needed. Regarding the aligners instead, it is clear how Gentle and MFA both are more accurate than the default alignment of the STT services, and they provide aligned phonemes as well. Between Gentle and MFA the decision would though be tougher. One thing to note is that the better quality of AssemblyAI as STT impacts the alignment's quality as well, and because of that, the evaluation of the aligners can benefit in time efficiency by continuing with the exclusion of IBM's transcriptions.
In Table 1 the list of the random samples taken from the development set of the dataset, and the winning aligner for each one is presented. As already said, the reported results are from the alignment done on the AssemblyAI transcriptions, as we excluded IBM's transcriptions. Moreover, only the ones that found a clear winner are reported in the table.
The samples consist of two contiguous words excerpts. I examined the segmentation of the audio defined by each aligners by listening to the first word, to the second word and to the space in between, comparing it with the actual written words.

During the evaluation I 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

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.

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 the INTERSPEECH dataset [**?**] it happened with 'devel_10.wav', 'test_08.wav', 'train_01.wav', 'train_10.wav' and 'train_14.wav'.
Given the described results, the choice would lie towards Gentle, but I realized that an aligner tool that combines both, by using mainly Gentle and looking at MFA for the missing words, would perform even better than Gentle alone, mitigating the problem of missing word alignments. The files that Gentle couldn't align would get discarded for this as well. We can call this option of alignment the "**Gentle-MFA**".
A script for this purpose has been developed: when there are some words that Gentle did not recognize, we use MFA for their start and end timestamp, then, we correct the starting timestamp of the next recognized word, which usually in Gentle is wrong as it includes the non-recognized ones inside.
The script also handles inconsistencies by detecting if a word's start is before the end of the previous one, and correcting them by shifting the end of the previous to the start of that detected word.

6.4.2 The choice of breath detection and labeling method

To detect and label the breathing inside the audio, studies often use neural models made for the purpose [58]. In Szekely et al.'s work, 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 [69]. 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. The possibility of using such a model will still be considered, but the detection and labeling of breathing is as of now planned to be done by a self-developed script, much simpler, but potentially effective, as suggested by first 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 evaluated to understand if they contain a breathing instance or not. The analysis is done firstly by imposing a minimum length threshold to the interval, a maximum average Decibel threshold and a maximum peak Decibel threshold to the interval. After this phase, only intervals more probably containing a breath event are left: on those 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. To achieve this behaviour, the script will utilize the Pydub library [78] to isolate parts of the audio and to impose thresholds.

The parameters for the breath labeling, using the self made script, are therefore planned to be the following four:

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

The length parameter can be hypothesized to be 0.19 seconds, that is in fact the reported average inspiration duration during spontaneous speech, as reported in Wang et al.'s study of 2010 on Breath Groups analysis [79]. This parameter will though be investigated further on our specific database. The Decibel parameters will all be hypothesized through qualitative analysis of random samples: the maximum breathing volume for example, would be reasonable to be set to -40 dB as informed from the first analyses. After understanding some possible sets of values for those parameters, an evaluation of the results will be carried out to pick the final set.

To choose the right set of parameters the resulting labeling can be confronted with well known values reported in the literature. One important value to confront for example, is the mean number of breathes per minute while speaking, parameter already studied since decades. Hoit and Hixon, in 1987 [80], 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 [81], 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) [82]. 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 might be 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 [83] for spontaneous speech at an habitual speed.
In Table 2 are shown five sets of parameters informed by the first analysis on the recordings.

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

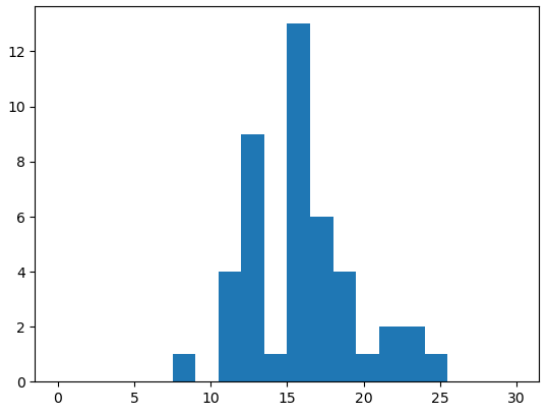
Table 2: Sets of parameters for the breath detection script.

In Table 3, instead, are shown the first results of the breath labeling script, and the values that the literature suggests.

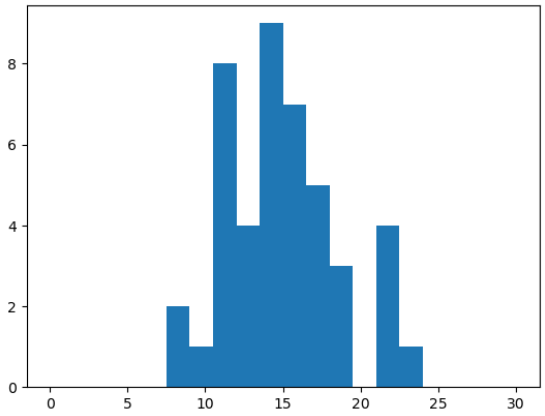
Parameter set	**Average BPM**	**Std. BPM**	**Average BGL**
#1	15.8	3.5	3.40 s
#1-bis	14.8	3.6	3.66 s
#2	19.7	3.9	2.70 s
#3	15.6	3.7	3.47 s
#4	16.6	3.7	3.47 s
Literature	16.5	6.1 (max)	3.46 s

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).

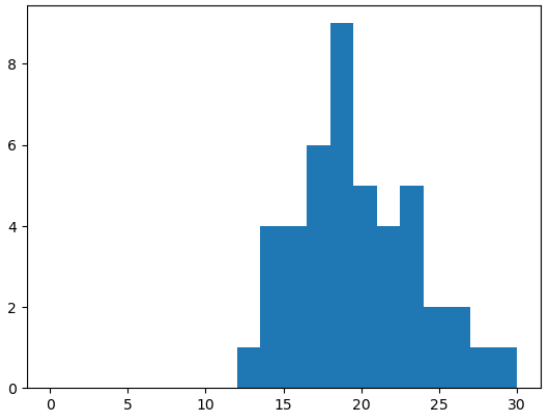
The difference between them does not seem to be significant, except for Set number 2 which I would exclude. There are though interesting differences in the distributions of the BPM across the corpus of the Database, reported in Figure 7. Notice the difference on peaks and dips, and in the range.



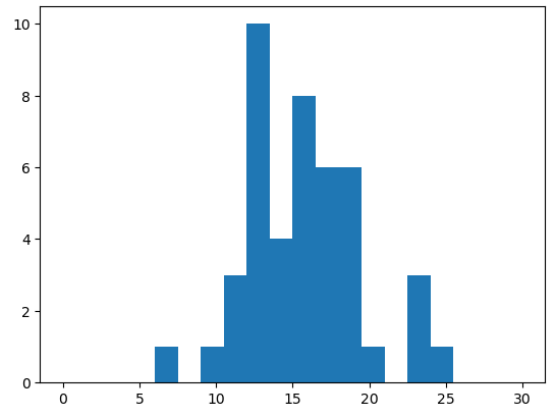
(a) Distribution of **Set 1**:
0.30 s of minimum length, no other constraints



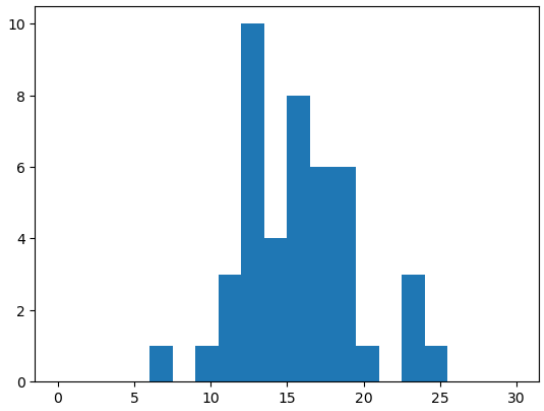
(b) Distribution of **Set 1-bis**:
0.33 s of minimum length, no other constraints



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



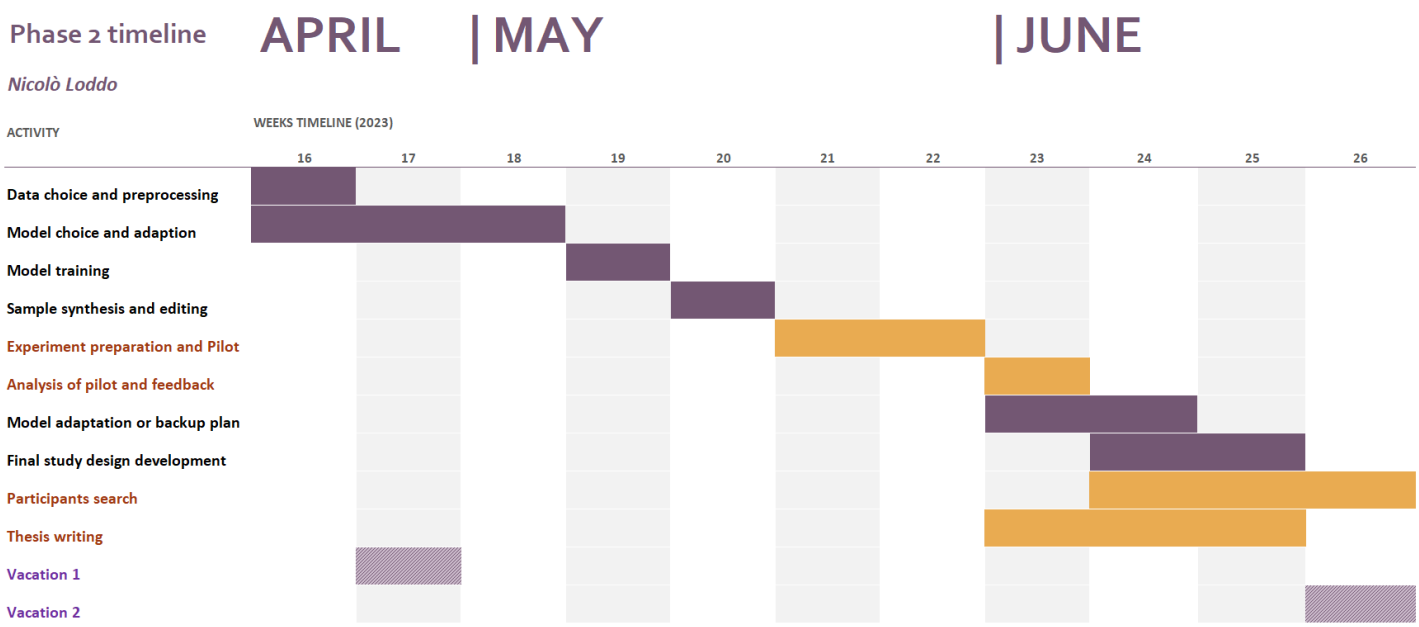
(d) Distribution of **Set 3**:
imposed dB peak and interval dB.

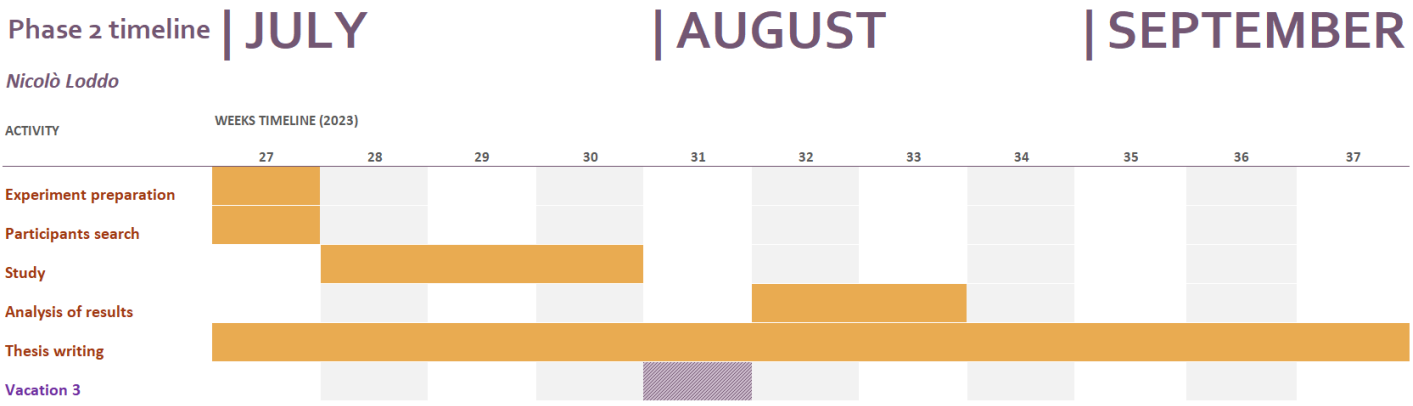


(e) Distribution of **Set 4**:
imposed dB peak.

Figure 7: Distribution of the BPM across the corpus for each set of parameters.
The big dip of Set 1's distribution needs to be investigated further: it may hint at interesting differences (for example in gender) across the subjects in the dataset.

7 Timeline





The above shown Gantt chart reports the planned timeline for the Phase 2 by 2023's week number.

References

[1] C. R. Berger and R. J. Calabrese, "Some Explorations in Initial Interaction and Beyond: Toward a Developmental Theory of Interpersonal Communication," *Human Communication Research*, vol. 1, no. 2, pp. 99–112, Dec. 1975. [Online]. Available: https://doi.org/10.1111/j.1468-2958.1975.tb00258.x

[2] T. Riede, E. Bronson, H. Hatzikirou, and K. Zuberbühler, "Vocal production mechanisms in a non-human primate: morphological data and a model," *Journal of Human Evolution*, vol. 48, no. 1, pp. 85–96, Jan. 2005. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0047248404001435

[3] H. Giles, J. Coupland, and N. Coupland, Eds., *Contexts of accommodation: Developments in applied sociolinguistics.*, ser. Contexts of accommodation: Developments in applied sociolinguistics. Paris, France: Editions de la Maison des Sciences de l'Homme, 1991, pages: viii, 321.

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

[5] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and ESD," *Speech Communication*, vol. 137, pp. 1–18, Feb. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167639321001308

[6] 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/

[7] 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).

[8] 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. New York, NY, US: Oxford University Press, 2015, pp. 516–536.

[9] 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," 2023, publisher: arXiv Version Number: 1. [Online]. Available: https://arxiv.org/abs/2301.05562

[10] Y. Xu, H. Cao, W. Du, and W. Wang, "A Survey of Cross-lingual Sentiment Analysis: Methodologies, Models and Evaluations," *Data Science and Engineering*, vol. 7, no. 3, pp. 279–299, Sep. 2022. [Online]. Available: https://doi.org/10.1007/s41019-022-00187-3

[11] T. Kano, S. Takamichi, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "An end-to-end model for cross-lingual transformation of paralinguistic information," *Machine Translation*, vol. 32, no. 4, pp. 353–368, Dec. 2018. [Online]. Available: https://doi.org/10.1007/s10590-018-9217-7

[12] "Prosody | Definition, Examples, Elements, & Facts | Britannica." [Online]. Available: https://www.britannica.com/art/prosody

[13] J. J. Ohala, "An Ethological Perspective on Common Cross-Language Utilization of F of Voice," vol. 41, no. 1, pp. 1–16, 1984. [Online]. Available: https://doi.org/10.1159/000261706

[14] J. McWhorter, "The World's Most Musical Languages," Nov. 2015, section: Global. [Online]. Available: https://www.theatlantic.com/international/archive/2015/11/tonal-languages-linguistics-mandarin/415701/

[15] D. Cameron and J. Grahn, "Perception of Rhythm," Sep. 2020, pp. 20–38.

[16] T. L. Bolton, "Rhythm," *The American Journal of Psychology*, vol. 6, pp. 145–238, 1894, place: US Publisher: Univ of Illinois Press.

[17] C. Wynn, T. Barrett, and S. Borrie, "Rhythm Perception, Speaking Rate Entrainment, and Conversational Quality: A Mediated Model," *Journal of Speech, Language, and Hearing Research*, vol. 65, pp. 1–17, May 2022.

[18] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, Sep. 2006. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167639306000422

[19] A. Fernald, "Intonation and Communicative Intent in Mothers' Speech to Infants: Is the Melody the Message?" *Child Development*, vol. 60, no. 6, p. 1497, Dec. 1989. [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," *Journal of Child Language*, vol. 16, no. 3, pp. 477–501, Oct. 1989. [Online]. Available: https://www.cambridge.org/core/product/identifier/S0305000900010679/type/journal_article

[21] A. Paiva, "Empathy in Social Agents," *International Journal of Virtual Reality*, vol. 10, no. 1, pp. 1–4, Jan. 2011. [Online]. Available: https://ijvr.eu/article/view/2794

[22] 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. Cham: Springer International Publishing, 2016, pp. 402–441. [Online]. Available: https://doi.org/10.1007/978-3-319-46152-6_16

[23] S. Brave, C. Nass, and K. Hutchinson, "Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent," *Subtle expressivity for characters and robots*, vol. 62, no. 2, pp. 161–178, Feb. 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581904001284

[24] 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.

[25] A. Paiva, J. Dias, D. Sobral, R. Aylett, P. Sobreperez, S. Woods, C. Zoll, and L. Hall, "Caring for Agents and Agents that Care: Building Empathic Relations with Synthetic Agents," *Autonomous Agents and Multiagent Systems, International Joint Conference on*, vol. 1, pp. 194–201, Jan. 2004.

[26] 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.

[27] M. Mori, "The uncanny valley." *Energy*, vol. 7, pp. 33–35, 1970.

[28] C. E. Looser and T. Wheatley, "The Tipping Point of Animacy: How, When, and Where We Perceive Life in a Face," *Psychological Science*, vol. 21, no. 12, pp. 1854–1862, Dec. 2010, publisher: SAGE Publications Inc. [Online]. Available: https://doi.org/10.1177/0956797610388044

[29] 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.

[30] 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/

[31] 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.

[32] 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.

[33] K. Kroes, I. Saccardi, and J. Masthoff, "Empathizing with virtual agents: the effect of personification and general empathic tendencies," Ph.D. dissertation, 2022.

[34] A. Mehrabian, "Manual for the Balanced Emotional Empathy Scale (BEES)." 1996.

[35] 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," *Journal of Personality Assessment*, vol. 91, pp. 62–71, 2009, place: United Kingdom Publisher: Taylor & Francis.

[36] 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," *Journal of Personality Assessment*, vol. 93, no. 1, pp. 84–95, Jan. 2011.

[37] D. Neumann, R. Chan, G. J. Boyle, Y. Wang, and R. Westbury, "Measures of Empathy," Dec. 2015, pp. 257–289.

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

[39] "A Short History Of Text-to-Speech | Speechify," Jun. 2022, section: Learning. [Online]. Available: https://speechify.com/blog/history-of-text-to-speech/

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

[41] 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

[42] 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

[43] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," Jun. 2021, arXiv:2106.06103 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2106.06103

[44] 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

[45] 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

[46] Y. Lee, A. Rabiee, and S.-Y. Lee, "Emotional End-to-End Neural Speech Synthesizer," Nov. 2017, arXiv:1711.05447 [cs, eess]. [Online]. Available: http://arxiv.org/abs/1711.05447

[47] 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," May 2019, arXiv:1905.08486 [cs, eess]. [Online]. Available: http://arxiv.org/abs/1905.08486

[48] S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, "Emotional speech synthesis with rich and granularized control," Nov. 2019, arXiv:1911.01635 [cs, eess]. [Online]. Available: http://arxiv.org/abs/1911.01635

[49] 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," Dec. 2018, arXiv:1810.07217 [cs, eess]. [Online]. Available: http://arxiv.org/abs/1810.07217

[50] 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

[51] 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.

[52] 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

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

[54] "Papers with Code - LJSpeech Benchmark (Text-To-Speech Synthesis)," 2023. [Online]. Available: https://paperswithcode.com/sota/text-to-speech-synthesis-on-ljspeech

[55] 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."

[56] 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.

[57] 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.

[58] 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.

[59] "Text to Speech Software – Amazon Polly – Amazon Web Services." [Online]. Available: https://aws.amazon.com/polly/

[60] M. Viswanathan and M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale," *Computer Speech & Language*, vol. 19, no. 1, pp. 55–83, Jan. 2005. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0885230803000676

[61] S. Karagiannakos, "Speech synthesis: A review of the best text to speech architectures with Deep Learning," May 2021. [Online]. Available: https://theaisummer.com/text-to-speech/

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

[63] F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer, "CROWDMOS: An approach for crowdsourcing mean opinion score studies," *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2416–2419, May 2011, 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/

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

[65] B. Naderi and R. Cutler, "An Open source Implementation of ITU-T Recommendation P.808 with Validation," in *Interspeech 2020*, Oct. 2020, pp. 2862–2866, arXiv:2005.08138 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2005.08138

[66] "Amazon Mechanical Turk." [Online]. Available: https://www.mturk.com/

[67] 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

[68] 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.

[69] 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.

[70] 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 *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2020-October. International Speech Communication Association, pp. 2042–2046, ISSN: 19909772.

[71] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)," *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive: (http://web.ku.edu/˜idea/readings/rainbow.htm).*, Nov. 2019, 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

[72] 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.

[73] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE' 05 Audio-Visual Emotion Database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, Apr. 2006, pp. 8–8.

[74] 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," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Dec. 2008. [Online]. Available: https://doi.org/10.1007/s10579-008-9076-6

[75] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, Jan. 2017, conference Name: IEEE Transactions on Affective Computing.

[76] Székely, G. E. Henter, J. Beskow, and J. Gustafson, "Spontaneous conversational speech synthesis from found data." [Online]. Available: https://prolific.ac/

[77] 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

[78] J. Robert, "Pydub," Mar. 2023, original-date: 2011-05-02T18:42:38Z. [Online]. Available: https://github.com/jiaaro/pydub

[79] I. Wang, J. Smith, and J. Ruiz, "Exploring virtual agents for augmented reality," in *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery.

[80] J. D. Hoit and T. J. Hixon, "Age and Speech Breathing," *Journal of Speech, Language, and Hearing Research*, vol. 30, no. 3, pp. 351–366, Sep. 1987. [Online]. Available: http://pubs.asha.org/doi/10.1044/jshr.3003.351

[81] S. Fuchs and A. Rochet-Capellan, "The Respiratory Foundations of Spoken Language," *Annual Review of Linguistics*, vol. 7, no. 1, pp. 13–30, Jan. 2021. [Online]. Available: https://www.annualreviews.org/doi/10.1146/annurev-linguistics-031720-103907

[82] J. D. Hoit and H. L. Lohmeier, "Influence of Continuous Speaking on Ventilation," *Journal of Speech, Language, and Hearing Research*, vol. 43, no. 5, pp. 1240–1251, Oct. 2000. [Online]. Available: http://pubs.asha.org/doi/10.1044/jslhr.4305.1240

[83] L. L. Kuhlmann and J. Iwarsson, "Effects of Speaking Rate on Breathing and Voice Behavior," *Journal of Voice*, p. S0892199721003052, Oct. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0892199721003052