Tampereen yliopisto
Tampere University

Jan van Dijk

NON-VERBAL COMMUNICATION IN INSTANT MESSAGING: CONVEYING EMOTION THROUGH VOICE INTERFACES

Master of Science Thesis
Faculty of Information Technology and Communication Sciences
Supervisor: Markku Turunen
July 2023

ABSTRACT

Jan van Dijk: Non-verbal communication in instant messaging: conveying emotion through voice interfaces
Master of Science Thesis
Tampere University
Master Degree Programme in Human-Technology Interaction
July 2023

Instant Messaging has become a keystone of human personal communication, where the biggest application WhatsApp is currently serving over 2 billion people. Plenty of research confirms people use non-verbal communication in computer mediated communication, allowing for emotional communication at distance. At the same time, Virtual Personal Assistants, such as the Google Assistant and Apple Siri, are continuously expanding their market share. Recently, they have included support for voice-based instant messaging, which includes reading aloud instant messages. As instant messages are synthesised, included digital non-verbal communication traits may be lost or omitted. This study aims to explore the impact of text-to-speech conversion of instant messages by virtual personal assistants on recognition of non-verbal cues by the receiving party. Secondly, the research aims to explore and test methods to include non-verbal communication traits in instant messages to speech synthesis, by the inclusion of spatial arrays (emojis) and modification of synthetic voice prosody. Sentiment analysis and emotion detection are explored and applied to extract emotional data from instant messages, which can be used to modify speech synthesis characteristics, such as pitch and speech rate, to mimic human paralanguage and vocal non-verbal communication to convey emotion.

Keywords: Computer Mediated Communication, Virtual Personal Assistants, Sentiment and Emotion Detection, Speech Synthesis, Spoken Instant Messaging

TIIVISTELMÄ

Jan van Dijk: Ei-sanallinen kommunikaatio pikaviestinnässä: tunteiden välittäminen äänen käyttöliittymän kautta
Pro gradu-tutkielma
Tampereen yliopisto
Ihmisen ja teknologian vuorovaikutuksen maisteriopinnot
Heinäkuu 2023

Pikaviestintä on tullut keskeiseksi osaksi ihmisten henkilökohtaista kommunikointia, ja suurin sovellus WhatsApp palvelee tällä hetkellä yli kahta miljardia ihmistä. Useat tutkimukset vahvistavat, että ihmiset käyttävät ei-sanallista kommunikointia tietokoneavusteisessa viestinnässä, mikä mahdollistaa tunneviestinnän etäisyydeltä. Samalla virtuaaliset henkilökohtaiset avustajat, kuten Google Assistant ja Apple Siri, laajentavat jatkuvasti markkinaosuuttaan. Viime aikoina niihin on sisällytetty tuki äänipohjaiselle pikaviestinnälle, johon sisältyy ääneen lukeminen. Koska pikaviestit syntetisoidaan, mukaan lukien digitaaliset ei-sanalliset viestintätavat, niitä voi jäädä pois tai ne voivat hävitä. Tämä tutkimus pyrkii tutkimaan virtuaalisten henkilökohtaisten avustajien teksti-puhe-muunnosten vaikutusta pikaviesteihin vastaanottavan osapuolen ei-sanallisten vihjeiden tunnistamiseksi. Toiseksi tutkimuksessa pyritään tutkimaan ja testaamaan menetelmiä ei-sanallisten viestintäpiirteiden sisällyttämiseksi pikaviesteihin puheen synteesissä, mukaan lukien ääneen lukemisen (emojit) ja synteettisen äänen prosoodian muokkaus. Tunteiden analyysiä ja tunneilmaisun tunnistamista tutkitaan ja sovelletaan tunteellisen datan eristämiseksi pikaviesteistä, mikä voidaan käyttää puheen synteesin ominaisuuksien, kuten äänensävyn ja puherytmin, muokkaamiseen matkimaan ihmisen parakielistä ja äänellistä ei-sanallista viestintää tunteiden välittämisessä.

Avainsanat: Tietokonevälitteinen kommunikaatio, virtuaaliset henkilökohtaiset avustajat, tunne- ja emotion havaitseminen, puhesynteesi, puhuttu pikaviestintä

PREFACE

I am very pleased to present this thesis, the result of my studies in Human-Technology Interaction at Tampere Universities in Finland. I am excited to share my research on non-verbal communication in instant messaging, as well as conveying emotion through voice interfaces.

I would like to thank Tampere Universities, and my supervisor Markku Turunen in particular, for the amazing journey these studies and thesis have been, and for the support, feedback and expertise. Second, I would like to express my heartfelt gratitude to mother, stepdad, grandmother and grandfather for the unconditional support and trust during my full academic journey. Furthermore, I would like to deeply thank all my friends and family for their support, encouragement and belief.

Last, I would like to thank all individuals who have participated and contributed to this research. Their willingness share their insights and participate in experiments has been integral to this research.

I hope that this thesis contributes to knowledge in the field and sparks curiosity for further research. I hope it serves as a stepping stone and ultimately makes voice technology more accessible and mature.

Thank you,

Tampere, 7th July 2023

Jan van Dijk

CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

CMC	Computer-Mediated Communication
F2F	Face-To-Face
IM	Instant Messaging
ISO	International Organization for Standardization
SSML	Speech Synthesis Markup Language
VPA	Virtual Personal Assistant

1. INTRODUCTION

Nowadays, we are living in an era of hyper-communication. As of writing, roughly 62 percent of people own a mobile phone, totalling 4.88 billion users (Turner, 2021). Of those, 3.8 billion are smartphones, meaning that roughly 48 percent of people on earth own a smartphone. These pocketable devices provide 24/7 connectivity, changing humans in surprising ways. "We have never had a technology that is so intensely used for so many different things" (Carr, 2014). One of the most prominent ways these devices affect us is arguably in the way they allow for communication, such as through Instant Messaging. This chapter introduces the thesis.

Over the years, IM and has become a cornerstone of our digital communication. The tool has deeply embedded itself in our society. VPAs are now able to tap in to this communication channel, allowing us to communicate IM verbally instead of using a screen. Sentiment and emotion analysis poise to improve this newly adapted interface, allowing to include nonverbal communication traits to spoken IM.

1.1 Background

Instant messaging apps are on the rise: Since the start of web 2.0, the moment when the web changed to a more collaborative place, online interaction has mostly been in a one-to-all fashion. This means that content on the web was generated by one person, to be consumed by numerous others, for example on platforms like Facebook and Twitter. However, Instant Messaging (IM) apps are ever-growing, their active user count surpassing social networks such as Facebook and Instagram by 20% (Intelligence, 2018). Analysis by Tankovska (2019) shows that every minute, on average, 42 million messages are sent. This expansion also shows in the market share of IM applications - as of February 2020, WhatsApp (2020) has 2 billion active users, Facebook Messenger counts 1.3 billion and WeChat is catching up by recently hitting the 1 billion mark (Clement, 2020).

Adaptation of Virtual Personal Assistants grows: In addition to the advancement of IM, another technological advancement is the appearance of Virtual Personal Assistants (VPA). Currently, the most widely used VPAs, such as Google Assistant, Amazon Echo and Apple Siri, are able to interpret commands given by the user using speech, and respond with synthetic voices. These VPAs provide assistance with personal day-to-day tasks, such as setting

reminders, checking calendars and controlling smart-home appliances. Smart Speakers with microphones are the main dedicated devices featuring VPAs, although they are available on a wide array of devices such as in-car entertainment systems, smartphones, wearables and televisions. Even though sales slowed down slightly during the pandemic, in the US, Kinsella, 2020 reveals about one-third of the population owns a smart speaker. Additionally, Kinsella reports that in the UK, in 2019a, about 21% of households owned a smart speaker, with Germany following up counting about 12%. As companies are continuously expanding the capabilities of their VPAs, one of the maturing features is hand- and screen-free messaging. Using this feature, users can have their incoming instant messages read out loud at times when they are not able to interact with a screen, for example when driving or cooking. Overall, this technology allows users to use the VPA throughout the day, supporting the user with relatively simple tasks such as playing music or placing a call.

Sentiment and emotion analysis improves: Sentiment Analysis refers to the use of Natural Language Processing (NLP), combined with computational linguistics and text analysis, to identify sentiment in text in a systematic manner. Emotion analysis is the next step in extracting data from text, categorising interactions, usually based on the seven universally accepted emotions. Emotion analysis, and thus NLP, has been studied since the creation of Artificial Intelligence in the '50s, and increasingly more since the turn of the century, with the technology maturing slowly. This growth is due to increased processing capabilities, the development of more efficient and precise algorithms, as well as an increased psychological understanding of human interactions and emotions (Weiyuan and Hua, 2014). Where in the early 00s moderation of online utterances was mainly a task for humans, as of now the flow of information is extremely high, forcing other ways of moderation. This resulted in the continued development and adoption of sentiment and emotion analysis (Huh et al., 2013).

1.2 Research objectives

Computer Mediated Communication (CMC) has seen a humongous spike in adaptation over the past decades (S. Herring, 2004, Vos et al., 2004). It can be accompanied by photos, videos and other visual cues such as emojis, but are mostly text-based (Adrianson, 2001). The usage of CMC does raise questions; are conversations using CMC different from those in real life? Do we express ourselves in different ways? Originally, researchers intensely argued that CMC levitates towards a cold (Culnan, 1987) and less personal (Rice Love, 1987) platform that gives off weak social cues (Sproull and Kiesler, 1986). However, as CMC became more commonly used and thus more usual in the following decade, sentiment changed to be more appreciative towards CMC as a valuable method of communication, as is Face-To-Face (F2F) communication (Adrianson, 2001). Nowadays, IM is a widely used method of CMC.

Like many other technologies, CMC and thus IM is ever evolving. Recently, CMC has started integrating to spoken interfaces. Virtual Personal Assistants now provide support for Instant Mes-

saging, using speech as the medium. This means that messages get synthesised to speech, so users can receive them without looking at their screen. However, there is a great lack of research in regard to the recognition of emotion in instant messages through voice interfaces. Currently, no research can be found. User-centred design research is a fitting method to connect existing research on human emotion usage, emotion in CMC and Instant Messaging research to voice interfaces, to explore how users interpret emotion and affect through voice-based messaging.

As technological development keeps pushing forward, new features are continuously expanding existing technological solutions. In the case of instant messaging using voice interfaces, two technologies are being combined, with one of the common goals being increasing the connectivity and quality of life of users (Stansberry et al., 2019). The combination of instant messaging and speech synthesis, particularly its potential influence on the capabilities of users to recognise non-verbal communication, will be the focus of this research. Additionally, methods for improving non-verbal communication conveyed through synthetic speech will be researched, using emotion recognition to modify the nonverbal speech characteristics.

The effects of converting text-based instant messages, including their nonverbal communication cues, to synthetic speech has seen few research. This research gap between two areas of expertise creates a void in understanding how users interpret and handle emotion and affect through voice-based messaging. This thesis aims to set initial steps towards closing this gap and better understanding of vocal instant messaging.

1.2.1 Research approach and aims

This thesis explores the state and future of voice-based instant messaging, as well as the user perception of it. No prior research in the Human-Technology Interaction field exists regarding the effects of converting text-to-speech on user experience of instant messaging. Therefore, this study explores the foundations of (non-verbal) communication, the current state of emotion in instant messaging (i.e. emotional CMC), the state of synthetic speech generation in the context of instant messaging, as well as possibilities to improve those spoken instant messages by applying emotion detection algorithms to improve speech prosody characteristics. A qualitative user inquiry deepens the knowledge of CMC and IM by uncovering usage habits, preferences and non-verbal communication traits. Based on the findings, a concept and a high fidelity prototype for an emotion-enabled voice-based IM solution was developed and tested.

The research questions are as follows: **When using instant messaging**,

(1) How is the personal experience of users of instant messaging applications in the context of emotion; how is emotion used and recognised?

(2) How and to what extent can software classify emotion in instant messages; can the results be applied in a prototype setting to modify speech synthesis to portray emotion

in vocalised messages?

(3) How is the user experience of emotion-enabled instant messages experienced through text-to-speech and how can the emotion experience be improved by including non-verbal communication traits?

1.2.2 Thesis outline

This thesis contains eight chapters. The next chapter (2) provides background information on human communication, computer-mediated communication, text-based sentiment detection and emotion classification and finally synthetic voices. In chapters 4, 5 and 6, the three research aims stemming from the research questions are explained, executed and concluded;

1. Investigate nonverbal and emotional expression in instant messaging
2. Develop a sentiment and emotion detection proof-of-concept and high fidelity prototype
3. Establish the user experience of emotion in spoken instant messaging

This structure was used to better represent the chronological research process and create a more coherent and understandable document. Each of research sections include findings and conclusions. Chapter 7 discusses the thesis and its findings, and chapter 8 concludes the thesis, provides limitations of the work and recommendations for future work, as well as touching upon the the learning experience.

1.2.3 Motivation

The motivation for this research has two points of origin, from a personal perspective. First off, I have been working with voice interfaces for well over five years now. Nearing the end of my bachelor studies in Information and Communication Technologies and Media Design, I started exploring and applying the realm of Virtual Personal Assistants. My thesis and post-graduate Conversational Designer job focused on vocal advertising using VPAs, for companies such as IKEA, ING and Volkswagen.

Studying at Tampere University opened my eyes to the deeper psychological aspect of HTI. I developed an increasing interest in communication and Computer Mediated Communication. Personally uncovering the possibilities and limitations of communication in our age, I started focusing on the bleeding-edge technology implementations of the VPAs, which included voice-based instant messaging, triggering my interest and eventually resulting in the topic of this thesis. I see great potential in speech interfaces, especially for the less technologically oriented members of our societies, such as our grandparents. As speech interfaces are tapping into our natural communication habits, their learning curve is surprisingly low. This makes them accessible to even the techno-phobic. My personal goal in this is to make speech interfaces seamlessly fit the user needs.

I personally experience many emotional nuances in digital communication with people close to me, such as friends and relatives. Every now and then, these emotional nuances are interpreted differently than intended, creating friction and exposing the limitations of CMC. By focusing on the future of CMC and VPAs, I hope I can contribute to more smooth and more natural interactions in digital human communication.

2. BACKGROUND

Where the previous chapter aims to sketch the current situation of communication technologies and software improvements by introducing trends, this chapter aims to further underpin the history and current state of affairs of (voice-enabled) instant messaging. Exploring existing research will conclude in possibilities and complications of this thesis.

2.1 Human Communication

Although humans have a varying number of skills, communication is considered key (Levinson and Holler, 2014). Communication plays a role in every part of life and is one of the principles of advanced being, as it allows us to define reality and organise concepts and experiences. Communication helps to think. In this section, we will explore the basics of human communication, to provide a knowledge base for further research on Computer Mediated Communication, text-based emotion detection and synthetic voice generation.

2.1.1 Verbal Communication

Communication itself refers to use of speech, writing or another medium to share information with others (Press, 2022). The usual train-of-thought when imagining the concept of communication starts with spoken communication. In scientific terms, this is defined as verbal communication, the use of language, both to speak and to write. Verbal communication also includes written communication and is rule-governed, meaning that there is a set of rules that form a common understanding used to convey desired meanings.

Formal and informal communication

In verbal communication, differences exist between spoken and written communication. The first usually being informal at its core, is used on a daily basis when interacting with others in-person, such as at work or on the phone. Alongside is formal communication, usually applied in written verbal communication. Where informal communication allows for the use of slang, verbal mistakes and divergent grammatical structures without users being inclined to correct those mistakes (Akmajian, 1995), formal communication is usually expected to be more strictly obeying those aforementioned rules in verbal communication. However, there are exceptions

to these precedents. With the introduction and increasing use of communication technology, such as email and IM, including informal communication traits into written texts to make them *feel* more like a conversation is increasingly common (Kristiansen and Coupland, 2011).

Synchronous and asynchronous communication

Another aspect separating spoken and written communication is one of timing. Communication using speech is overwhelmingly *synchronous*, where it is mainly used in real-time, for example in a conversation with a teacher. Even when speech is used as a form of mass-communication, such as giving a presentation, it is still synchronous, as the speaker gets an immediate response to what they are saying from other beings, such as an applause or a response to a greeting.

Asynchronous communication occurs over time. This is usually the case with written communication, where one can expect a delayed response to their input, for example when communicating using letters or email.

2.1.2 Nonverbal communication

Using only verbal communication, or words, deciphering a message may be difficult, due to the abstract core of languages. However, long before verbal communication existed, humans relied on nonverbal communication cues for thousands of years (Head et al., 2013). Nonverbal communication means to convey meaning in ways other than using words, although both usually go hand in hand. Nonverbal communication specifically means to communicate meaning through sounds, artefacts and behaviour other than verbal communication. Nonverbal communication is best seen as a process alongside verbal communication, not the same, although part of the same system. This also shows in the processes in the human brain, where the left side is typically tasked with verbal communication, whilst right side is simultaneously accountable for the creation of non-verbal cues (McGilchrist, 2009). Nonverbal communication is usually happening as an unconscious process, where verbal communication is conscious.

Types of nonverbal communication

Scientific research on nonverbal communication began in 1873 when Charles Darwin published his book "The Expression of the Emotions in Man and Animals". Since, countless research has been published on the subject. Like verbal communication, nonverbal communication can be categorised. In contrast to verbal communication, multiple types of nonverbal communication are usually used at the same time, complementing, changing or contradicting the verbal message. At the same time, nonverbal communication is dependent on culture, common understanding and other shared aspects. These have an influence on the following types of nonverbal communication.

Gestures are divisible into three categories, according to Andersen (1999); Adaptors, Emblems

and Illustrators. The first is behaviours usually associated with anxiety and arousal. We can think of subconsciously playing with objects near us, touching our hair or face, or moving our body when in a meeting or whilst waiting. Emblems refer to gestures of which the meaning is generally agreed upon. Think about giving a thumbs-up to someone doing a good job, or shaking the index finger when correcting someone. Finally, illustrators are the most seen feature. They are mostly tied to whatever the person is saying, and rarely have a meaning of their own. Illustrators come automatically when talking, for example pointing in a direction when giving directions.

Posture can convey many types of gestures, due to its many variations. For example, resting hands on the hips with elbows pointing out indicates confidence and makes the person look bigger, leaning forward can show interest, and raising hands can show enthusiasm.

Head movements are mostly used to confirm something or show interest. For example, nodding the head is associated with *yes*, shaking our head side-to-side is accepted as a denying gesture, the equivalent of saying *no*. Tilting the head to a side can convey a feel of interest, intimacy or trust.

Eye contact is another important part of gestures. While also studied in other branches of studies, in this area eyes are usually an important focus during interactions. Eye states can be tied to emotional states such as 'sparkling eyes' and 'angry eyes'. At the same time, eye contact is an important cue in the turn-taking in conversations, and keeping connection whilst having interactions to show interest. Similarly, avoiding eye contact indicates a lack of interest for communication.

Facial expressions are one of the most versatile ways of showing expression. In many languages, there is a saying along the lines of *one look is worth a thousand words*. We can for example think about a picture of people taken at a party; even though it's just a second frozen in time, there are still many emotions to read from captured faces. According to research, facial expressions can be generally be divided into five baseline groups; happiness, anger, disgust, fear and sadness. These are found in most cultures, although their interpretation might differ based on different norms (Andersen, 1999).

Haptics are a form of communication by touch. Communication by touch is one of the most exquisite and powerful forms of the nonverbal ones, where being educated on when, where and how to use it is regarded key and differs per culture. Think for example about giving a hug or even a kiss as a greeting; a powerful moment, regarded normal (even for strangers) in some countries, whilst in others it is reserved for the most intimate contacts - something that needs to be taught.

Additionally, there are functional and social levels of touch. In a professional environment, a firm handshake can be a standard interaction to start off a meeting. On a social level, they help to start off and end interactions, eventually becoming part of a ritual.

Vocalics are part of the aforementioned paralanguage. It defines the nonverbal aspects of a spoken message. Vocalics consist of verbal fillers, pitch, volume, speaking rate and quality of speech. There is an array of interesting observations by Andersen (1999) regarding pitch, as it helps to define meaning, message intensity and regulate flow. Questions usually end with a higher pitched ending, greetings and goodbyes emphasise the end of a sentence, and a slightly different pitch occurs when conveying sarcasm.

Volume helps to communicate with more intensity, and is usually adjusted based on the situation. Whispering can be appropriate in private and personal messages, whilst in public it might lower the credibility of the speaker.

Similarly, the speaking rate indicates the time it takes for a person to convey what they are saying. On average, people speaking English use 120 to 150 words per minute (Buller and Burgoon, 1986). Speaking faster can be seen as more intelligent, while speaking slower can result in loss of interest and distraction of the receiving party.

Tone-of-voice is what is controlled using the volume, pitch and emphasis, although this varies for all humans. Usually, people prefer voices with a lot of variation, according to Andersen (1999). Verbal fillers are used to fill up the silences in our speech while we think about how to continue. They are common and usually not disrupting the conversation, as they help to keep the attention from shifting from shifting to another focus.

Nonverbal communication is the most used and most powerful way of human communication, and people have found many ways to include it in CMC. Vocalics play an essential role in the creation of modern synthetic voices, and in written communication. Choice of words and sentence structures add sentiment and emotion to text, which can be detected by sentiment and emotion detection algorithms.

2.2 Computer-Mediated Communication

Communication can take place through computers as well. Technology is ever changing the way people live, including the way they communicate. Computer Mediated Communication, commonly called CMC, has been present since the first appearance of the digital computer. CMC started during WWII, while the first record of prototype e-mail dates back to the early 60's. A classic definition by S. Herring (1996) is as follows; "Computer Mediated Communication (CMC) is communication that takes place between human beings via the instrumentality of computers". This definition is quite wide, as CMC includes websites, phone applications and even online lectures. For the goal of this thesis, research focus will be on text-based CMC.

2.2.1 Communication traits in text-based CMC

Emotion and (non)verbal communication in digital messaging has seen a significant amount of research. At its core, CMC lacks traditional paralinguistic features of in-person non-verbal communication, such as gestures, facial expressions, body language, tone of voice and even appearance. However, users of such messaging platforms have devised methods to assist in communication.

In 1980, Carey identified two types of verbal communication in CMC; linguistic markers (e.g. thank you endlessly) and emotion words (e.g. exiting, angry). Linguistic markers convey emotion without containing emotion words. Moreover, five types of paralinguistic cues were identified in CMC;

Lexical surrogates are written vocal sounds, such as “haha”, “hmmm” and “uh-oh”, meant to make the conversation ‘sound’ as if it is happening in real-life.
Vocal spelling refers to the alteration of spelling to copy speech, for example writing “helloooo” instead of hello, “yeeees” instead of yes, for example to indicate enthusiasm.
Minus features omit formatting elements such as ignoring capitalisation, where using lowercase letters is experienced as more friendly, and omitting parts of words or abbreviations, such as using bc instead of ‘because’.
Spatial arrays referrs to the use of symbols for pictographs such as :) or :D. Emoticons being the evolved version of spatial arrays, research by Lo (2008) shows that use of those in messaging can succeed in conveying emotion, when utilised as non-verbal communication tools, although other research shows that they have a very limited ability to mimic nonverbal communication due to the lack of their “range and nuance” (Derks et al., 2008, Dresner and Herring, 2010, Hancock, 2004), indicating that they only amplify written text.
Manipulation of grammatical markers is about modification of words such as using all-caps, multiple interpunction marks (“!!!”).

All these cues have been regarded as replacements for nonverbal communication in CMC, being called oralisation of written text (Yus, 2011). This is in line with research by Mehrabian (1972), showing that in everyday verbal communication, 38% of emotional communication originated in tone-of-voice, 55% was connected to facial expression, and only 7% of emotional communication originated in used words.

In the earlier days of CMC, these cues were seen as an added feature (Spears and Lea, 1992). More recently, the cues are seen as an integral part of online communication (Postmes et al., 2000). Walther (2011), wrote that using CMC can be just as effective as traditional methods of communication when making use of the different types of cues, Kalman et al. (2010) backs this up by noting that CMC cues are an abundant and diverse phenomenon, found to be an integral part of CMC.

Use of emoticons and emojis

Spatial arrays are a method to add facial nonverbal communication to written messages. Depending on the context, manipulation of grammatical markers can have a more notable function than emoticons (Vandergriff, 2010). For example, repeated dots (...) can be sent and experienced as a replacement for in-person silences. These manipulations can be intended in an ironic way, a response to a 'loss of face', or in certain contexts can be experienced as a threatening way of speaking. In a broader sense, a lot of CMC cues share a common goal; adding an emotional dynamic to formal textual communication.

Research on the use of emoticons and emojis in text-based CMC shows that they mostly support the existing written non-verbal communication traits (Kalman and Gergle, 2009). However, in the last decade, the amount of technological improvements regarding CMC has been growing incrementally, adding features such as avatars, haptics, live-emojis, personal typography and other creative tools to support emotional communication.

In general, CMC cues are often seen as a way to communicate social and emotional meanings, appearing more often in groups of people who know one another. When outside of a known group, Hancock et al. (2007) indicated that receiving parties can register written emotion, when intently placed by the sending party. They specifically noted that context and mutual understanding are defining factors in successful interpretation, such as a common conversation goal or common conversation setting. Riordan and Trichtinger (2017) support this discovery.

Different generations use emojis differently, as described well by a Wall Street Journal article by Demato (2021). For example, millennials, the generation born between 1981 and 1996 who saw the internet, emojis and IM emerge and adapted them by including them effortlessly in their lives, use emojis in more versatile, creative and consistent ways than their older generations. However, their consecutive generation who were born with IM at their fingertips, use emojis in an increasingly abstract way, bending the intended meanings of emojis. For instance, millennial can use a laughing and crying emoji (face-with-tears-of-joy) to indicate they are amused, where generation Z may use an emoji of a skull (skull).

Research of influence of the emotional state in CMC cues by Pirzadeh

Hancock et al. (2007) suggests that the emotional state of a person influences the way they use emotional cues in text-based CMC. A follow-up research by Hancock et al. in 2008 shows that participants in a negative environment or state used less words in text-based communication, while at the same time using more negative terms.

Research by Pirzadeh and Pfaff (2012) deepened these findings with research on instant-messaging between friends, adding increased use of grammatical markers, whilst uncovering significant broader patterns for other emotional states. They researched the effect of the emotional states relaxed, happy, sad and angry on the use of different emotional cues in CMC.

Individuals in a happy state tend to use more words, more vocal spelling and increased manipulation of grammatical markers (e.g. *I'M HAPPPYYYYY!!!*). When sad, participants used significantly less minus features and affect words, whilst increasing use of negative emotion words. Participants noted sadness was the hardest emotion to convey through chat, as they experience it as a personal emotion, which is hard through words as it is usually conveyed and picked up from (subtle) cues that need to be felt in real-life. In the angry emotion, like sadness, part of negative sentiment, even less minus features occur, while affect and negative words stay at a similar level. A participant of Pirzadeh and Pfaff noted that "when it comes to deeper emotions like sadness or anger I tend to use a lot of gestures, facial expressions and I seek them in the respondent. Not being able to utilise those aspects of conversation was frustrating".

In a relaxed state, participants used a significant amount of affect and positive words, and the lowest amount of negative words, and the lowest amount of grammatical markers. Participants also noted that relaxed was a hard state to convey emotion in since they were in a neutral state.

In general, people have found many ways to include all types of nonverbal emotion into digital communication, resulting in emotion conveyed even when thousands of kilometers apart, making CMC truly valuable and feasible for interpersonal communication. A pattern in these types of communication can be identified to be used in text-based emotion detection.

2.3 Text-based sentiment detection and emotion classification

Artificial intelligence has been around since the 1950s, and got increased attention and innovation after the turn of the century. It has contributed to many aspects of human life and society, one of which being Natural Language Processing (NLP). This section will explore how sentiment and emotion can be extracted from text, which types of techniques are available and how the data can be used.

As mentioned in the before, NLP refers to the processing of human language, by a computer. It is an area of expertise where sciences come together; computer science, artificial intelligence, cultural studies and linguistics. NLP mainly refers to programming software in such a way that a computer understands human communication. This results in switch-of-roles, as traditionally, computers and software require humans to understand them to use them, for example through a Graphical User Interface. With NLP, computers try to understand humans through one of their natural ways of communication; language.

2.3.1 Sentiment analysis

NLP is used in a variety of systems, like translation, information gatherers and organisers (such as Google) and automatic text generation. For this thesis, focus lies on two aspects of language

processing; sentiment analysis or emotion classification and synthetic voice generation. The hypothesis, which will be further explained later in this thesis, assumes that sentiment analysis on IM can be used to improve the conversion of text-to-speech of these instant messages, by adding a layer of emotional value.

Text-based sentiment and emotion detection has seen an enduring history of research (Alm et al., 2005), with it becoming increasingly tied to machine learning. The main goal of sentiment analysis, also known as opinion mining, is to analyse and polarise text by extracting its thoughts and ideas. This automated process results in a sentiment score; a representation of the text and its underlying mood. The score is usually given in numbers, ranging from -3 to 3, where -3 stands for very negative, 0 for neutral and 3 for very positive.

Applying sentiment analysis software, a sentence like *Thank you so incredibly much* usually results in a very positive sentiment score, while a sentence such *I hate this news* will end up with a negative score. Sentiment analysis can be used on social media to find sentiment about active topics (like politics and companies), in corporate helpdesks to analyse customer sentiment on-the-go, to analyse written texts and so on. However, sentiment analysis is not straight forward. In cases where sarcasm is used, a classification may end up in an incorrect bracket.

Emotion Classification (EC) is a subfield of Sentiment Analysis, which aims to uncover even finer details in the polarisation of texts. Where Sentiment Analysis uses the top-level of the emotional spectrum (positive = happy, excited; negative = angry, sad; and standard = neutral), Emotion Detection tries to categorise analysed text into one or more of those specific emotions.

Sentiment Analysis versus Emotion Classification

All models using sentiment analyse and polarise input, in the categories positive, neutral and negative. Some Emotion Detection models take a next step, and focus on feelings and emotions, such as happiness, sadness and anger. A limited set of models, mainly used by companies to analyse social media, is able to determine whether urgency is present in a text, indicating that quick action needs to be taken based on the content of the text.

Fine grained Sentiment Analysis polarises with less or more of the aforementioned categories; very positive, positive, neutral, negative and very negative with numeric representations. *Emotion Classification* aims to detect emotion from messages, like happiness, sadness or fear. They mainly depend on lexicons (a list of emotion words and emoticons) for analysis.

Although both techniques work well, false-positives arise relatively often. For example, it is hard to make a difference between texts like “This product is shit” and “This product is the shit”, which would both be rated as a negative sentiment, although the latter carries a rather positive sentiment.

2.3.2 Sentiment Analysis Methods

There are three main methods of sentiment analysis; rule-based analysis, machine learning and a hybrid of both.

Automatic systems rely on techniques or data to learn from data, for example using machine learning and rule-based tagging.

Rule-based Method

Rule-based systems automatically perform analysis based on a set of rules and existing data. A number of structured datasets are available online, containing words or sentences and their respective sentiment or emotion. These datasets, such as ISEAR, EMOBANK, WASSA-2017 and AMAN'S, are based on research questionnaires and/or analysis. They can be used in one or more NLP techniques, such as;

Tokenization or **Bag-of-words model** breaks up strings into single word groups, and analyses those one-by-one, counting the sentiment score. For example "Cheese from the Netherlands tastes very good!" could be broken up into "Cheese", "from", "the Netherlands", "tastes", "very", "good". When processing, if "good" has a positive sentiment in the set of existing data, it would get a positive sentiment rating (+1). The inclusion of "very" multiplies the sentiment rating of "good", which results in a more polarised score (+2).

Part-of-Speech Tagging (POST), also known as grammatical tagging, tries to assign words in a text to specific parts of speech, based on the definition and context of the words. It is similar to the way we learn to identify words such as nouns, verbs, adverbs etcetera, although POST typically uses 50 to 150 categories, instead of the 9 used in regular teaching. It is considered more complex than tokenisation, as some words can be ambiguous (e.g. glue can be used as a verb ("I will glue it together"), and as a noun ("A bottle of glue")).

Named-entity recognition (NER) tries to extract named entities from a piece of text. It uses a set of predefined categories (i.e. businesses, names, food) to tag certain parts of a sentence (e.g. "Anastasia[person] bought[action] Apples[fruit] from Prisma[company]").

Machine Learning Method

The Machine Learning Method removes some of the limitations of rule-based tagging by using Machine Learning to determine emotion in text, with research showing that it's often used in EC and usually yields better results (Canales and Martınez-Barco, 2014). Two methods of ML can be used, either supervised and unsupervised ML. Machine Learning algorithms are trained using example data, such as the aforementioned datasets. After training, the algorithms classify input without human intervention. Machine Learning provides a method to process the task of sentiment analysis without excessively being programmed to do so. Additionally, these

models can be expanded in iterations to recognise deeper understandings such as sarcasm and account for spelling and grammatical mistakes on the user side.

Hybrid method

The hybrid method combines both rule-based tagging methods and machine learning. Both methods are usually executed separately. The outputs of multiple systems are then combined to generate a more diverse final sentiment score. This can be useful, as it evens out errors in the processes of any sentiment analysis software.

2.4 Synthetic voices

As explored in human communication (2.1), vocalics or paralanguage play an important role in nonverbal communication. Vocalics consist of verbal fillers, pitch, volume, speaking rate and quality of speech, allowing for conveyance of subtle as well as strong emotional cues in conversations. In CMC, users use an aforementioned array of techniques to include nonverbal communication in written text. This paragraph will explore the possibilities and effects of including nonverbal written emotional cues, in synthetically generated speech using text-to-speech (TTS).

2.4.1 Emotional prosody in human voice

Language is important in the transfer of emotions whilst speaking. Think about asking someone how they are doing, and they respond with *I'm okay*. By the tone of voice you recognise how they feel; happy, sad, angry, or any other emotion.

There are several studies exploring the effect of vocal expressions on the perception of emotion, by letting participants read aloud sentences in different emotional states. Usually, emotion evaluation texts are created to carry no emotion by itself, only to be added by the participant, although some researchers ciritise the unnaturalness of such methods (Schröder, 2004). Bachorowski (1999) found that emotions such as joy and fear change the acoustics of speech, using a higher pitch than for example sadness. They found these traits to be fundamental aspects of communication, providing important cues to arousal, as well as an important contribution to pleasantness of experienced emotions in conversations. The outcome of their tests confirmed that listeners are able to recognise emotions from emotionally loaded speech in a significant way. D. A. Sauter et al. (2010, 2010) researched the perceptual cues in non-verbal vocal expressions. Participants were asked to speak words and sentences in varying emotional states. Recordings were then presented to other participants, asking them to recognise emotions. They concluded the research that emotions are able to be accurately included and recognised through voice.

Table 2.1. *Categorization of positive and negative emotional vocalisation characteristics (%) by D. Sauter et al., 2010*

	Achievement	**Amusement**	**Anger**	**Contentment**	**Disgust**	**Fear**	**Pleasure**	**Relief**	**Sadness**	**Surprise**
Duration	0,02	-0,04	0,09	0,31	-0,35	-0,1	0,32	0,04	0,15	-0,43
Ampl. RMS	0,46	0,19	-0,22	0,02	-0,03	-0,24	-0,02	-0,19	-0,19	0,5
Ampl. Onset	0,07	0,47	0,14	0,04	0,02	-0,03	-0,03	-0,1	-0,21	0,02
Intensity	-0,11	0,07	0,22	0,37	-0,14	0,15	0,38	0,43	-0,71	-0,51
Pitch min	-0,75	-0,39	0,21	-0,03	-,11	0,38	0,12	-0,39	-,24	-0,52
Pitch max	0,07	0,06	0,12	-0,21	-0,04	0,21	-0,03	-0,22	0,19	0,28
Pitch mean	0,81	0,42	-0,46	0,07	-0,18	0,24	-0,08	0,59	-0,17	0,59

After, they performed an acoustic analysis on the recordings. Table 2.1 shows the relevant conclusions of differences in speech rate, pitch and intensity, based on the acoustic analysis. For example, when speaking with a sad emotional load, the duration of the recording is 15 percent longer. Some table data irrelevant for this research has been left out.

The results of the categorisation done by D. Sauter et al. indicate that the positive, negative and neutral emotions have recognisable vocal nonverbal expressions, in a human-to-human context. These results are similar to previous research, such as from Banse and Scherer (1996) and Juslin and Laukka (2002), which both found that different acoustic patterns connect to different emotional states. Belin et al. (2008) support the data as well, noting similar results.

2.4.2 Speech Synthesis and emotion

Speech synthesis or text-to-speech (TTS) is a form of assistive technology that can convert text to read out loud speech. It is used on an array of devices, including computers, phones and in cars. As mentioned, smart speakers and VPAs use the same technology. The voice is computer generated, with varying qualities, although the most recent versions based on machine-learning sound increasingly natural. Speech researchers have been focusing on improving the full range and variations of speech, to add a more psychological dimension to TTS, with the goal to match human sound.

There are multiple types of speech synthesis, of which formant synthesis, concatenative synthesis and articulatory synthesis are the most commonly used.

Formant synthesis

Formant Synthesis is the oldest type of speech synthesis, and has been the most used implementation for a long time. Currently, formant synthesis is still a common choice. It uses a source-filter model, which means it generates signals from text and sends them to a circuit that models vocalics of a human. The vocal tract that is simulated to synthesise the speech is a rather simplified version, resulting in the well known robotic voice. If the model used to synthesise speech is as accurate as possible, it is highly beneficial for the quality of the speech, resulting in a more natural sounding end product. The use of a formant synthesis model means that the vocal values and traits can be modified; for example to make the pitch of the voice higher.

Concatenative synthesis

This form of speech synthesis relies on a database of pre-recorded bits of speech. It is also referred to as the cut-and-paste method, since it combines short recordings of vocal sounds to produce an expectedly more natural sounding end result. However, in practice, the generated speech still faces limitations. These vary from memory and processing limitations due to the large amount of audio clips needed. The amount of audio clips needed is the second limitation. Due to the dependency on these clips, the tone and emotional conveyance is hard to manipulate, as they are pre-recorded. Nevertheless, it is suitable for certain applications such as PA systems, reading news articles or GPS navigation. Where its limitations show is in reading larger pieces or more diverse types of texts, where a personality is in focus.

Articulatory synthesis

Articulatory speech synthesis systems rely far more on the structure of synthesis models and computational power. Compared to the previous systems, this system lies a way heavier emphasis on modeling the speech production mechanism of human speech. The creation of such a system is an intensive project and can take years. In the beginning of the 21st century this technology was still difficult to use and experiments were not always successful, however it poised potential for high quality synthesis. However, the past seven years, big technological players such as Google and Apple have gone all-in on the creation of these systems, and letting them learn and improve with machine learning using human speech recordings. It extends the vocal tract models of concatenative analysis, while adding the physiology of the vocal cords. Adding these physiological features, it removes the unnatural sounds often heard in concatenative models, by not allowing them to arise, as they are considered artifacts.

WaveNet

WaveNet is one of the latest improvements in audio generation. Based on a so-called deep neural network, Wavenet is able to generate raw audio waveforms based on machine learning and data mining. In 2016, van den Oord et al. published the paper about the generative model for raw audio. Even though it can be used for generating all kinds of audio, such as classical music pieces, when used in a TTS context, it yields significantly better results than previous TTS solutions. This is because it creates the raw waveforms of audio from scratch, using the neural network. This network is trained using many speech samples, extracting the underlying structure of the speech such which ones occur in speaking, and how tones follow up on each other. This means that, when giving the network text input, it can generate the corresponding speech waveforms, omitting the limitations of methods such as concatenative synthesis.

2.4.3 Speech Synthesis Markup Language

As the demands for speech synthesis keep increasing to go beyond simply reading out texts to being applied in everyday conversational solutions, the need for more vocally varied synthesis arises. This need calls for subtle differences in various aspects of the high-level voice attributes, such as intensity, pitch and prosody. Most speech synthesis programs and services provide these capabilities by allowing modification of their voices via including Speech Synthesis Markup Language (SSML).

The World Wide Web Consortium (W3C), the organisation responsible for designing and maintaining web standards, describes SSML as follows;

> *"The essential role of the markup language is to provide authors of synthesizable content a standard way to control aspects of speech such as pronunciation, volume, pitch, rate, etc. across different synthesis-capable platforms."*
>
> Walker et al. (2001)

Adding emotion using SSML

SSML provides a possibility to modify the speech output in such a way that it portrays the previously explored nonverbal communication traits, such as speaking rate and pitch, in speech synthesis. According to observations by C. Janani et al. (2015), exploring inclusion of emotion in speech synthesis, including emotion is easier to do for sentences with positive sentiment such as happiness. This is due to the use of happy emotive words making the sentence positive. Janani et al. noted that to convey happiness in speech synthesis, one can modify the tone of the used emotive words to portray positive nonverbal traits. Sentences portraying sadness are to be handled at phrase level, as to make the whole sentence 'sound' sad.

Aside from the application of SSML in sex robots (Bendel, 2017), few scientific research on the

application of specifically “SSML” in connection to emotion was found. Therefore, this research will aim to close the gap using existing and new research.

2.5 Concluding existing research

There are many ways we, humans, use communication. It has seen a sustainable amount of research. Where the unknowing might think communication is mainly verbal, will be proven wrong. Non-verbal communication is one of, if not the, most important methods of communication; especially from an emotional standpoint. Those non-verbal communication traits are without a doubt present in computer-mediated communication. Many of the non-verbal cues find their way to variants in verbal communication in CMC, by manipulating, expanding or omitting language features or adding methods to mimic facial expressions, such as the use of emojis. However, when moving to the frontier of the research, where both topics merge in the form of spoken computer-mediated communication, research is increasingly scarce. It is not known whether computer-mediated communication using speech provides similar traits to include non-verbal communication, or whether it strips some or all of the non-verbal traits in its conversion.

In other fields, both emotion recognition and speech synthesis have improved significantly in the last years. The latter its voices now on their way out of the uncanny valley, provides possibilities to modify voice characteristics such as speed and pitch, where some of those characteristics may provide opportunities to include non-verbal communication. There is a need for user-centered research and design to determine the effectiveness of non-verbal communication in speech interfaces, and whether modifying speech to include non-verbal traits has an effect, it being positive, negative or insignificant.

However, for non-verbal cues to be used in speech synthesis, the fitting emotional state needs to be determined. Although humans obviously are the best at recognising emotions in communication, manual sentiment detection and emotion classification is not desirable in an instant messaging scope, be it from the perspective of speed, privacy or the vast amounts of data. This is where sentiment detection and emotion classification software can play a valuable role, by automating this recognition process. Thus, the goal of this research is to create a prototype which classifies emotions, and to understand whether the inclusion of non-verbal communication in speech synthesis of instant messages succeeds in conveying non-verbal cues.

3. RESEARCH AIMS AND STRUCTURE

The existing research was performed to create background knowledge about the identified trends. Human communication, Computer-Mediated Communication, text-based sentiment detection and emotion classification and synthetic voices. Using this knowledge, the research aims and structure was established.

3.1 Research target group

A wide variety of people is using Instant Messaging apps and VPAs. Instant Messaging apps are used by the elderly, as well as children, and everything in between. The same goes for VPAs, which are used by every age bracket in our society, although in varying adoption rates. However, to focus this research and improve the usability of its results, a target audience for the research was selected.

It is suggested there is a significant difference in the way adolescents use IM versus the elder generations, where the youngsters are usually pioneering in new ways of using communication platforms (Manganari, 2021). For example, it is suggested that there is a difference in the way generation Z currently uses IM applications; in contrast to the older generations they are found to abandon the (original) use of emojis and other types of spatial arrays, or find other ways to substitute for those, and use them in higher volumes than other generations. Non-scientific sources have written about generation Z increasingly abandoning emojis as of late (Demato, 2021).

Similarly, culture influences the way IM applications are utilised. Culture is defined in various ways, but one of the widely accepted definitions is by Hofstede (2010); *"[Culture] is the collective programming of the mind that distinguishes the members of one group or category of people from another"*. This includes the way people interact with one another on a daily basis, and certain customs and habits which communication adheres to. For example, in the context of communication, Hall and Hall (1990) describe the differences in communication between high- and low-context cultures. In the former, the meaning of communication is mostly embedded in situations, resulting in an increased attention to contextual information and increased use of non-verbal cues in communication. For the low-context cultures, the situation in which a conversation takes place plays a more minimal role, resulting in a lesser use of nonverbal

communication. Research shows that these findings also translate to digital communication, where high-context cultures use significantly more emojis in CMC (H. Li et al., 2011, Kayan et al., 2006).

Instant Messaging app users

As mentioned, IM applications are used by all age brackets of society. However, in European countries, IM apps are most used by users from 18 to 40 years old. In Finland, 24 to 35 year old use IM apps the most, with 95% of them using it for communication (virallinen tilasto, 2020). Similarly, in France 90% of 18 to 24-years old are using IM apps, and 84% of 25 to 39-years old (BAILLET et al., 2019) , both the biggest groups. In the Netherlands, of the 12 to 25 year-olds, and 25 to 35 year olds, 95% uses communication apps (voor de Statistiek, 2020). Similar statistics can be found for other countries in Europe.

Virtual Personal Assistant usage and Smart Speakers ownership

In the US, the usage of voice assistant among all groups has been increasing over the past years (Petrock, 2019). The numbers for smart speaker adaption is similar across the age groups, with generation Z and millennials having comparable ownership numbers in the US (ages 18-29 34,1% and ages 30-44 29%, Kinsella, 2019b). Compared to Germany in the EU, consumers are generally behind in adaptation by two years in comparison to the US, according to Kinsella (2021). Their data shows that in Germany, the adaptation of smart speakers is around 34% for the adult population.

3.1.1 Target group

Based on the information listed above, the decision was made to focus the research on millennials. Based on the findings, it is assumed that this group will adapt voice-based IM functionalities first. At the same time, it will create a convenience sample as it will be relatively easier to find research participants from said age bracket. Outside of the millennial generation, I do not have many connections in Finland and other countries to end up with a diverse research sample. The gender of the target group is not considered important, however a fair distribution of genders is desired. Education level, income, occupation, martial status, ethnicity are similarly not considered important, although it is aimed to create a fair distribution in these aspects as well. For demography, the scope of the research will be limited to people residing in or originating from Europe, residing or originating from higher-context cultures, according to the culture map provided by Hofstede (2010).

4. USER EXPERIENCE OF INSTANT MESSAGING

Aim 1 researches the baseline as to what extent and how emotion is used and experienced in IM, from the personal perspective of research participants. This will be done using an exploratory interview, based on existing research. To focus the research, a focus research group will be established. The interview should result in a deep insight of how non-verbal communication and emotion is used in instant messaging containing. The results of this first research aim will provide a solid starting point for narrowing down the scope of the continued research. Deciding factors may include the intensity and/or abundance of recognised emotion and nonverbal cues in instant messaging.

4.1 Methodology: a semi-structured interview

To explore the current opinions, experiences and perceptions of millennials on emotion in instant messaging, the method *semi-structured interviews* was selected. This is a qualitative method, and is a common semi-structured qualitative study method in comparable research. These methods do not aim to prove a hypothesis, rather try to address questions and develop knowledge in an explanatory way, and shed light on different aspects of the topic (Blandford, 2013).

Blandford suggests this research method as an effective way to gather versatile and in-depth human insights in HTI matters. The method gives researchers the possibility to discover, reveal similarities and differences in participants viewpoints, and is commonly used at the start of a research track. The goal of interviews in this research is to review the current needs and habits of the users, and gather insights into the effects of the younger technologies. A downside of the method used is that participants might refuse to respond or partly respond to certain personal questions. As this research does touch upon personal emotions and personal experiences, these downsides have been kept in consideration when creating the interview structure, accommodating for open and free conversations which do not go into the content of an individuals specific messages, but rather the feelings accompanied by them.

In preparation of the semi-structured interviews, an interview outline was created, which can be found in appendix A. Special care was taken to adapt the interview to an online setting, by making the questions specific, and writing a solid introduction to make the participant feel at

ease. Additionally, before conducting the real interviews, a test interview took place to work out any imperfections and make final adjustments to ensure an optimal experience for participants.

A total of 8 interviews have taken place, with all participants falling in to the age bracket of 22 to 26 year old, with an average of 24,5. Participants all came from Finland, Sweden and the Netherlands. Half of them were students, where the other half was working. All were higher educated, with the field of work relatively varied, including arts, healthcare, IT, automotive and philosophy.

The interviews were scheduled by the hour, and aimed to be a length of 30 minutes, consisting of 7 minutes of introduction and setting up, 20 minutes of interview and 3 minutes of wrapping up. The other 30 minutes were occasionally used for interviews that required more time. However, in practice, the extra time ended up used for processing and enriching notes with the interview was still fresh in mind, as well as to prepare for the next interview if applicable. In total, 8 interviews took place, with a maximum of 3 per day, to avoid mental exhaustion and guard valuable results. The interviews mostly took place through Zoom calls because of the ongoing COVID-19 situation and language of correspondence was English.

Following the introduction, consent and first open questions, the interview focused on collecting generic data of the participants, such as their age, sex, profession and nationality. With the formalities handled, questions turned to the details of instant messaging usage of the participant, including their most and recently used applications, favourite and least favourite services, and how the participant uses IM apps.

Following the initial part of the interview, focus shifted to the personal use of, and recognition of emotion in IM. Questions focused on the incorporation of various emotions in their messages, which methods they use to include emotion (e.g. lexical surrogates, spatial arrays and manipulation of grammatical markers), and their general communication style. Finally, participants were asked their worst and best experiences in IM.

In the final part of the semi-structured interview, participants were asked about their familiarity with, and usage of VPAs. After gathering general information about the participants usage of VPAs, the questions focused on the use of text-to-speech functionality for reading aloud IM, and their experiences using these features. Finally, participants were questioned about whether they would be able to recognise emotions in IM read by VPAs, and their opinions about possible methods to improve their ability to recognise emotions.

To wrap up, feedback on the interview was asked to fine-tune for the following participants. Participants received a bar of chocolate as a token of appreciation afterwards.

4.1.1 Review using Affinity Diagrams

Semi-structured interviews generate a vast amount of data, which in its essence is unstructured. There are overarching similarities, however, every interview and person is different, and so are their answers. To process the data, the Affinity Diagram methodology was used. The method is mostly used organise, form and make sense of large sets of unstructured and diverse qualitative data (Hartson and Pyla, 2012), commonly used for analysis of contextual inquiry data (Holtzblatt et al., 2004) and commonly used in HCI (Benyon, 2014). To guide the process, the steps in the paper by Lucero (2015) were used as a guideline. In this paper, Lucero specifically mentions their sources "use of affinity diagrams is aimed at the early stages of the design process", whereas Lucero "looks at how affinity diagrams can provide support to analyze interactive prototype evaluations in the later stages of the design process". This support of analysis of prototypes in later stages fits the goal of this thesis very well, concluding that the selected method fits this thesis fully.

Note creation

To start the affinity diagram process, a set of notes was created in the online tool https://whimsical.co. This was done using the written notes taken during the interview, as well as re-watching the video recordings of the interviews, and making notes based on the second observation. During this second observation, flags were placed in the videos to note relevant quotes for later reference. The notes included both quotes and annotations. In total, 683 notes were created, and each participant was given an unique colour for their notes.

4.2 Findings

The affinity mapping resulted in four main categories; preferred ways of using IM, communication through IM, emotion in IM and finally VPA usage, experience and expectations. The findings of these categories are presented in the following four subsections.

4.2.1 Preferences of using IM

All interviewees voiced their personal preferences regarding IM applications. In this subsection the main findings of preferred IM applications will be discussed, to develop insights on messaging app and feature usage.

IM application preference

Overall, respondents note their preferred IM app is one that most of their network uses, and the least preferred IM app is one that they themselves use the least. Two interviewees note that

design of the interface is an important aspect. Half of the respondents voice that WhatsApp is their preferred IM app, because it "feels natural" 7, "is straightforward" 5 and they like it because "of its simplicity" 7. Facebook Messenger, Discord and Telegram are noted as favourite applications as well.

Three out of eight respondents note that they dislike their least liked IM app due to their network not being on it. Three respondents explain specifically that they dislike Facebook Messenger, as it is connected to their personal profile, of which participant 4 notes they are "forced to use it" because of their family.

Safety is regarded important in the preference of IM application, as participant 4 voiced that as they had been receiving messages from hacked accounts, of which they did not know what was going on, lowering their trust in the service.

A divide in IM application preferences was found depending on the social circles. For example, participant 2 notes that Snapchat, an app mostly aimed at quickly sharing photos and visuals, is personally used in friend circles, while Microsoft Teams is for use in work relations. In general, WhatsApp received no comments that indicate a clear divide for use cases among participants, all participants note it is used in all situations, either professional or personal.

A day of using IM

Respondents use IM apps to stay in touch with friends and family, as well as to arrange practical matters in professional environments such as work or studies. Especially contact with friends and family is present throughout the day, where this *keeping in touch* is worded as an "automated process during the day" by participant 3.

Participant 1 uses IM to stay in contact with the friends and family by sharing daily life moments, as they live abroad. This includes sharing texts, pictures and videos, either from their own experiences, as well as media found on the internet (i.e. memes, articles). "I think because I have a lot of far away contacts, [the amount I use instant messaging apps] is not too much".

Participant 4 notes that their app feels like a part of their life, and that "they feel at home in the app, it feels intuitive". However, they also voice that communication with "WhatsApp is for use on [a] phone" and that they use it "on their own terms", explaining that they do not want to use it outside of the app on their phone, such as on an online interface (i.e. *WhatsApp Web*) and possibly voice interfaces.

Two participants note that they have their notifications turned off, so they do not get distracted by incoming messages, "silence is also important" 4.

Importance of IM applications

When interviewing participants about their usage of IM, all interviewees describe IM as an essential part of their lives. Two participants note that they wouldn't be able to live without IM, while the rest of the participants voice that they would or might be able to live without IM. The latter comment that living without IM would "change my life significantly", "is not convenient" and that it would "cut all ties with my friends - not in a way that I would prefer" 5.

Furthermore, IM is described as "a standard way of communication" 2, and an "automatic" process, indicating that it's a habit to use these applications throughout daily life.

4.2.2 Communication through IM

The most prominent reason for using IM is to stay in touch with other people, as well as arranging practicalities. IM is described by participants as a "conversation tool" 4 and that "connecting [...] with people makes messaging nice" 2 This subsection will describe the uncovered habits of using IM to communicate.

Shortcomings of IM

IM makes it effortless to stay in touch, discuss plans and share what is going on. Participant 3 describes that it feels close to identical to calling. Five participants agree that it is possible to have full conversations through IM, such as one would have in real life. However, there is a more negative sentiment towards certain aspects of IM as well. In particular participant 5 voices a clear description on the downsides; "If you only send text messages to each other, you'll have quite a flat conversation, because you don't have any emotion, you don't have body language, it's not as spontaneous and fluent then when you're seeing each other" and "I don't think you can really express yourself deeply though [IM] apps". Participant 3 describes the same, saying that IM "is always different than talking talking with someone in person, you can't hear their voice or see their face".

This sentiment continues with 4 participants expressing that they prefer calling over texting, 5 taking it a step further and noting that they always ask to call when using IM for deeper emotional topics, as they feel like they are unable to express themselves well through text.

However, participant 1 says the following in regards to discussing difficult topics through IM; "if it is stuff that is difficult to talk about, it is easier sometimes to not be face to face". This implies that in certain cases, for certain people, IM can be a preferred method of communication, including serious or difficult topics.

When to use IM and when misunderstandings arise

As misunderstandings arise in daily life, misunderstandings arise in IM as well. However, the frequency and fixing them is suggested to differ from *offline life*. The interviews uncovered that misunderstandings arise more often using IM relative to offline life.

All participants voice that they experienced misunderstandings and discussions through IM. When inquiring about the meaning of misunderstandings, participant 4 defined it as "things that were interpreted not as intended"; what they had written *sounded* different in their head. Half of the participants note that misunderstandings arise due to the lack of, and understanding of emotion in messages. Participant 3 even "fell of with friends" due to misunderstandings on an IM platform.

"If it's something emotional I want to talk about [it] with someone in real life, I would just call someone" 5. Participant 6 voiced that a misunderstanding in the emotional load of a message on an IM platform resulted in a "talk [that] became an argument, if it was face to face it would have been fine". Participant 3 voices that "arguments happen with people you know, because you can't see them", again voicing the limitations of the interface on mobile devices.

Six out of eight participants choose to fix misunderstandings in a manner that does not involve IM, namely through a (video) call or meeting up in real life. When asking about the lack of emotion in communication, participant 6 notes that "I could see my mistakes after", with 'mistakes' referring to the (emotional) wording of their messages.

4.2.3 Emotion in IM

All participants agree that emotion is present in IM, in various ways. The way participants include their emotion in messages is generally similar, with some individual differences. Due to the free-flowing core of the interview, the talk about emotion in IM started at the same point, but progression differed depending on the answers of the participant.

Emojis

The use of emojis to express emotion is common among all participants; "it is a lot more clear to use emojis to show how I mean things", "I use them to give tone to what I'm saying" 5 and "emojis help me express myself emotionally" 7. Participants react to messages with emojis with a facial expression, the same as they would have used in real-life interactions. They can as well be added "after the fact" 5, after the initial message, to add or strengthen the intended emotion. "If I want to show love, I add a heart" 1. Participant 6 notes that they "mainly use the basic emojis, not the detailed ones, the happy one, laughing one", explaining that they mainly use it as an emotional reaction to a received message, and in a lesser sense to a message they send themselves.

Stickers and GIFs

Stickers and GIFs are used to express emotion in a similar ways as emojis, although in a more airy manner, in more relaxed contexts. The visual messages are explained to be mainly used in group chats and in response to shared media, such as a picture, online video or article. Participants do not use stickers and GIFs in conversations with deeper emotional loads.

Emotion words and grammar

Participants explain that they increase the use of emotion words in emotional situations. Participant 6 mentions that use more words in general when they are explaining an emotion. Manipulation of grammar and grammatical rules are also present to modify the emotional load of a message.

Oralisation of written text

All types of oralisation of written text as found in 2.2.1 were mentioned by the participants as being used regularly, albeit less or more in varying emotional situations. The interview results regarding usage of methods for oralisation of written text per participant can be found in table 4.1.

4.2.4 Silence as a newly discovered method of conveying emotion

The interviews unveiled new insights on how emotion is expressed. When asked, two participants explained that "Silence and not answering is [a way of showing emotion]" 3. Instead of continuing the conversation, a person shows that they are feeling emotions by halting the conversation and not replying to any messages sent. In various online settings, this behaviour is referred to as *ghosting*, defined as "the act or practice of abruptly cutting off all contact with someone (such as a former romantic partner) usually without explanation by no longer accepting or responding to phone calls, instant messages, etc." (Merriam-Webster, n.d.).

The five emotions

Participants were asked how they convey emotion through IM. The following paragraphs summarise their answers.

Happy - To express happiness, participants manipulate grammatical markers, for example by writing *!!!* after a sentence, as well as using *all-caps*. Participants use "lots and specific emojis" or spatial arrays, stickers and GIFs as well as lexical surrogates.

Sad - While using IM to express sad emotions, communication is described to have a more serious tone, including little to no expressive emojis or spatial arrays depending on the person.

Table 4.1. *Usage of methods for non-verbal communication written text (2.2.1) in IM, ordered by participant, results.*

	Participant 1	Participant 2	Participant 3	Participant 4	Participant 5	Participant 6	Participant 7	Participant 8
Lexical surrogates	X	X	X	X	X	X	X	X
Vocal spelling	X	X	X	X		X	X	X
Minus features	X	X	X	X	X	X	X	X
Spatial arrays (incl. emojis)	X	X	X	X	X	X	X	X
Visuals (stickers, GIFs)	X	X	X	X		X	X	
Manipulation of grammatical markers	X	X	X		X	X		X
Silences (ghosting, new insight)			X				X	

Minus features such as omitting capitalisation are more common, and messages are described to be shorter. Half of the participants note that they clearly mention the specific emotion (sadness) they are experiencing.

Anger - When angry, more sarcasm is used, also described as passive aggressiveness. Emojis are almost unanimously omitted, and grammatical markers are given extra attention. Messages are shorter, and some participants describe to use harsher words.

Disgust - Disgust is expressed using less emojis and short messages. Participants describe they use emotion words more often to describe their feeling, use lexical surrogates such as *eeeew*. One participant notes that in the case of disgust of the opponent of the conversation, they would give no response to ignore them.

Fear - Emojis are used more often in the case of an emotion related to fear, participant 6 especially notes the teeth biting-emoji. An interesting note of participant 7 was that they would talk in a way that they would reassure themselves within the conversation. Emotions are explained more clearly by most participants, especially including emotion words.

These insights display similarities to the discoveries of Pirzadeh and Pfaff (2012) as written in 2.2.1.

4.2.5 VPA usage, experience and expectations

The final interview questions aimed to collect insights from participants about the familiarity, experience and usage of VPAs, as well as the possibility to use IM through VPAs. All interviewed participants were familiar with the existence of VPAs. Of eight participants, two participants use voice assistants on a daily basis. Furthermore, three participants have tried to use or used virtual assistants at some point. For participant 3, a reason to not use voice assistants after

experimenting with them was that "it does not understand my accent or pronunciation". The final three participants showed no interest towards any of the digital assistants.

In general, usage of voice assistants met up to expectations of the group, and all participants that had used VPAs voiced that it is or may be convenient to use in specific situations. One of the participants noted that "the voice is laughable", referring to the robot-like voice, being indicated especially present in Finnish TTS solutions.

Hands-free messaging

One participant had used text-to-speech for IM using their voice assistant, using Google Assistant. Although they mentioned it was "handy" to use, they noted limitations due to a "language gap", their assistant was not able to switch between the local language (Dutch) and English, which means that the messages in the local language were spoken as if it were English sentences, resulting in "funny" but unusable recordings.

However, participants do note that they would like to be able to listen to messages instead of reading them, especially "when there's a situation where you can't use your phone" 5, such as in a car. It does "depend on the situation", where participants want to use speech based IM "only if hands-free [...] but only if it's really important" 1.

Emotion in voice-based instant messaging

The sentiment towards the recognition of emotion through voice interfaces was less positive. All participants voiced some negative expectations towards recognising emotion in spoken IM. Responses ranging from "It's really hard to convey tone [with artificial voices]" 1, "there is probably something emotionally in the way" 3 and "the computer is not very good at emoting, it's monotonous" 8. In general, participants agree that recognising emotion through the interfaces may be hard due to the lack of emotional load in the voice. Participant 6 questions how VPAs would handle the usage of emojis in sentences, whether they will be included in the TTS or omitted entirely. On the other hand, participant 1 stated "I expect I can recognise emotions in spoken IM from people that I know really well, I know the way they speak".

In hands-free situations where IM through VPAs is considered a solution for being unable to text physically, five out of eight participants have reasons to prefer not use the voice interface at all. "Instant Messaging voice interfaces seem useless" 7. Participants 1 and 6 would rather call than IM using voice in hands-free situations. Participants 4, 5 and 8 will just "accept [to] have to wait" 5.

Solutions for recognising emotion in voice interfaces

Finally, participants were asked about solutions that would help them recognise emotions in IM voice interfaces, of which three main categories of answers were identified. First, use of emotion words would help to recognise emotion in messages received. Second, participants noted that emotion in the voice of the assistant would help them recognise emotions. Participant 6 explained "if the voice assistant can interpret a message, so if it's an exiting message, it uses an exiting voice". Participant 4 said "changing tone of voice, based on the amount of question marks, or capital letters" would help them recognise emotion. Third, participants noted that voicing the emojis included in messages would help them identify included emotion. If a sender would send *That's amazing woman-health-worker-medium-skin-tone*, the assistant would voice *that's amazing 'emoji of woman health worker'*.

Other comments of interest regarded inclusion of voice characteristics of the sender in a vocalised message, meaning using their voice characteristics to generate the audio, and the ability to translate messages to the receivers native language before vocalising.

4.3 Conclusion

This chapter aimed to develop insights in the use of emotion in IM in the daily life of interview participants, to complement insights gathered from existing research in 2. By executing semi-structured interviews, deeper and more valuable knowledge on emotion in IM and VPAs has been uncovered, which can be converted to action points for the following research aims.

In general, Instant Messaging has progressed to be an essential part of communication, and is deeply embedded in the rhythms of daily life. Living without IM would change the lives of participants significantly, and was not desired by any of the participants. Being connected is regarded important and a natural part of being.

Emotions and emotional connections are an integral part of digital communication, especially in cases where it is used with friends and family. Emotion is omnipresent in all messages and the experience of all interview participants, as indicated by the research of Derks et al. (2008). Oralisation of written text is one of the cornerstones of digital communication, as well as the newly uncovered insight of introducing silences, similar to ghosting. However, shortcomings of IM include having a flat conversation, where in certain situations emotion may be hard to include or recognise.

Spatial arrays, including emojis, are an important aspect in conveying emotion through IM, and is used intensively. This underlines the background research in 2.2.1, mostly regarding the research by Carey (Carey). Emojis underline and emphasise the emotion present in messages, and used as non-verbal communication. Spatial arrays can also be used as a main carrier to convey emotion, sometimes sent in a separate message to add emotion after an initially sent

message to add clarification. This is in line with the research by Kalman and Gergle (2009).

The sending of emojis separately after a message uncovers that a turn in a conversation through IM does not necessarily consist of one message. One or multiple emotions can be spread over multiple instant messages sent in the same conversation, with some messages only consisting of lexical surrogates or spatial arrays. This may impose difficulties in later research aims where emotion detection software is planned to be utilised, and should be considered carefully.

Furthermore, visuals are used in easier emotional contexts, however they are not always used as an emotional amplifier. GIFs and stickers are usually used in group chat settings. Emotion words are actively used to express and include emotion in messages. Especially in more intense emotional settings, emotion words receive an increased amount of attention to convey emotion to the best of the senders abilities.

The emotional connection between IM participants is regarded important, as discussed in 2.2.1. When one knows one another, the receiving party reads messages in the voice of the sending party. Personal traits of participants of conversations are of vital importance to understand more complex emotional conversations.

Voice interfaces are being used moderately by a smaller percentage of interviewed participants. Only 1 of 8 participants used text-to-speech conversion of IM. Hurdles to overcome are the sound of the assistant its voice, adaption to local language and the ability to switch between various languages.

To better recognise emotion in voice interfaces, three main categories were affirmed. The first possibility to improve the experience of emotion in TTS lies with the sender; the use of emotion words is thought one of the main actions that participants expect to help them with the recognition of emotion in messages directed at them, received through a VPA. Second, participants suggested to modify the characteristics of the voice of the assistant to portray the emotions in a more natural way, such as making the voice sound happy. This suggestion will be explored in the third aim of this research. Third, participants suggested to read aloud spatial arrays, which will also be explored in the third aim of this research. Additionally, a participant suggested it would aid them to recognise emotion by personalisation of the voice. In other words, by making the voice sound like that of the sender.

The insights uncovered in this research aim will be used to set out the research path for the second aim, starting with the creation of an IM dataset.

5. EMOTION RECOGNITION SOFTWARE

This aim aims to develop a proof of concept for real-time emotion and sentiment detection for instant messaging. To start off, an IM dataset containing various emotions will be be curated. This dataset is used to feed to emotions detection software, resulting in the messages in the dataset begin tagged with detected emotions. Later, detected emotion using software may be used to manipulate prosody of vocalised messages, using techniques such as Speech Synthesis Markup Language (SSML). The aim is not to find the most efficient solution, it is rather to produce a working proof of concept high fidelity prototype to develop knowledge.

5.1 Curating an IM dataset

Prior to engaging into emotion recognition using software, a dataset of instant messages was curated, consisting of messages categorised by emotion. The purpose of this dataset is to validate and test emotion recognition software, as well as generate TTS audio files utilising SSML. In aim 3, these audio files will be used to research the effectiveness of emotional cues in TTS with research participants.

Two datasets were created, both consisting of 6 emotion categories (anger, disgust, fear, joy, sadness, surprise) and a neutral category. Each category includes 20 messages, for a total of 140 messages per dataset, or 280 in total.

Avoiding design bias

As discussed in 3.1, culture and location may influence how emotion is embedded in IM. It can be argued that it may even influence the topics and emotional signals used in messages. Therefore, the decision was made to curate a dataset based on existing datasets of emotion-tagged sentences, instead of creating one from scratch, to avoid researcher design bias as much as possible.

5.1.1 Existing datasets

An article published by Acheampong et al. (2020) surveys the concept of Emotion Detection (ED) from text, as well as providing a list of emotion labelled data sources to "provide neophytes

with eligible text datasets for ED". This list of currently available and validated datasets was used as a starting point for a curated dataset. This resulted in the table 5.1, where datasets were selected and listed based on their origin, emotion model and availability.

Unfortunately, at the time of writing, the *Emotion Lines* dataset was not available for download, even after registrations, as it would have been included it the list otherwise. More datasets were found and considered, however, they were found to be not of sufficient quality or neutrality. Therefore, the decision was made to go ahead with a selection of the datasets gathered by Acheampong et al. (2020).

Selecting a appropriate dataset

In order to select an appropriate dataset, the research question of the third aim (*How is the user experience of emotion-enabled instant messages experienced through text-to-speech and how can the emotion experience be improved by including non-verbal communication traits?*) was taken as a starting point. The core of IM is that it is personal and mainly in a one-to-one or one-to-few setting, meaning that a message is meant for a personal audience. The parts of IM that are researched in this thesis regard dialogues, therefore the (public) one-to-all tweets, news articles and the like are assumed not likely to convert well to an IM settings and fell off for the selection.

The results of research aim one (4) indicate that conversations with IM can be similar to *offline* conversations. This makes it reasonable to assume that non-scripted personal dialogues would convert relatively well to a TTS IM setting. However, datasets with *turns* in the conversations may need some altering to fit more to the online setting. Subsequently, the possible influence of lower- and higher context cultures should be kept in mind.

These requirements resulted in two datasets regarded suitable for this research, both *Daily Dialog* and *EmotionPush* are dialogue-based in a personal one-to-one or one-to-few setting.

Initially, it seemed that the EmotionPush dataset was unobtainable, as the dataset is only available on an on-request basis. During the initial work on this aim, contact could not be made with the authors of the dataset, and thus work was continued using the Daily Dialog set, although it was considered a better fit.

Daily Dialog

The Daily Dialog dataset is a "high-quality multi-turn dialog dataset" with the language being "human-written and less noisy" (Y. Li et al., 2017). It "cover[s] various topics about our daily life" where the creators "manually label[led] the developed dataset with communication intention and emotion information".

The dataset has 13,118 real-life (*offline*) dialogues in a text file, with an average of 7,8 turns per

Dataset	Year	Content Type	Something	Size	Included emotion classifiers	Balanced
Daily dialog	2017	Dialogues	Personal one-to-one	102k sentences	neutral, joy, surprise, sadness, anger, disgust, fear	No
Emotion Stimulus	2015	Dialogues	Personal one-to-one	2.5k sentences	sadness, joy, anger, fear, surprise, disgust	No
ISEAR	1990	Emotional situations		7.5k sentences	joy, fear, anger, sadness, disgust, shame, guilt	Yes
SemiEval Task 4	2017	Tweets and news headlines	Public One-to-one	1.25k texts	anger, surprise, disgust, joy, fear, sadness	No
WASSA-2017	2017	Tweets	One-to-all	3.6k tweets	joy, sadness, fear, anger	No
CrowdFlower	2016	Tweets	One-to-all	40k tweets	anger, boredom, empty, enthusiasm, fun, happiness, hate, love, neutral, relief, sadness, surprise, worry	No
EmotionPush	2016	Instant messages	One-to-one	3.5k messages	neutral, joy, sadness, fear, anger, surprise, disgust	No

Table 5.1. *Selection of suitable datasets listed by Acheampong et al. (2020) to be possibly used for ED and curating an IM dataset*

dialogue. Each turn is labelled with one of the 6 emotions (anger, disgust, fear, joy, sadness, surprise) or *neutral*. 20 sentences were randomly picked for each emotion using a random number generator, where a random number was picked based on the total amount of sentences available with the selected emotion. In a few cases a dialogue turn was not selected to be included in the dataset, for example for it being incomprehensible without the context of the conversation, or for is being a simple answer like *yes*, *no* or *okay*.

The sentences in the dataset were slightly altered in some occasions to integrate better to the established focus group of the research, for instance by changing currencies to euros, avoiding or using easier to understand first names and avoiding complicated sentence structures.

Finally, emojis were added based on the context and emotion tag of the sentence. This was done according to the usage of spatial array insights gathered in 4.2.3. Messages with emotions like *joy* and *surprise* arguable have a higher number of emojis added than more complicated emotions like *sadness* and *disgust*, however, all messages had an emoji added in preperation for the final aim. Emojis to include were selected based on the tagging information provided by Unicode[1].

The created dataset, of which the sentences will now be referred to as messages, consists of messages that have a clear bi-turn dialog flow, where the turns create a clear exchange of information between speakers. The emotion in the messages is rich, which "enhances social bonding" according to Y. Li et al. (2017). However, some of the messages do apply clearly to a daily life in-person interaction, which at times does not translate well to the online setting of IM. The fact that DailyDialog is a manually developed dataset by researches is both its strength and it weakness; it is so balanced and well-tagged it is a perfect tool for training algorithms. However, to humans in an IM context, the messages may arguably come off artificial and out-of-place. The higher context of the Asian country of origin may also influence the structure and emotion of the dataset.

EmotionPush

EmotionPush is an IM dataset with messages originating from real chats on Facebook Messenger (Shmueli and Ku (2019)). While the authors of the dataset were initially not reachable, later in the process, contact was established with the authors, grating access to the EmotionPush dataset. Thus, the dataset was processed using the same method as the Daily Dialog dataset, creating 7 emotional categories of 20 messages and adjusting to the local context, as well as adding in emojis. The EmotionPush dataset anonymised location information, names and brands in messages, which were added back. There were no messages that required to be altered in flow to fit in with the online setting, contrary to Daily Dialog.

The curated dataset based on EmotionPush arguably contains more authentic messages, due

[1] د

of their authentic, non researcher-written, online IM origin. However, the emotion included in the messages *feels* more nuanced as it originates from a personal relation between two individuals who are aware of one another and their communication styles. Additionally, the messages originate from students based in the United States, which shares the lower context culture with target group of this research.

5.1.2 Finalising the dataset

Both sets containing both an emoji-free and emoji version were finalised in a Google Spreadsheet, ordered by emotion and exported to four CSV files;

- Daily Dialog excluding Emojis (140 messages)
- Daily Dialog including Emojis (140 messages)
- EmotionPush excluding Emojis (140 messages)
- EmotionPush including Emojis (140 messages)

These CSV files were processed in to associative JSON arrays to be further processed by the emotion recognition software as described in 5.3. Each element consists of a message ID, the message and emotion label. A human-readable list of messages can be found in B.1.1 and B.1.2.

5.2 Available emotion classification software

In the past decade, most of the researched sentiment analysis software available focuses on classifying polarity in an input text, like positive, neutral and negative (Thelwall et al., 2012). As written in section 2.3, newer projects utilise machine learning to further improve classification of sentiment in text, by using the emotional dimensions.

An online search for open source text emotion classification projects was performed. The background research uncovered that machine learning methods are poised the most promising in terms of accuracy and reliability, and thus a focus has been applied to finding this type of software. At the same time, only more recent open-source projects have been taken into consideration, even though they have seen less scientific research. Finally, since this aim is aimed to explore the feasibility of using emotion classification software to modify SSML parameters for the generation of TTS, and not to develop a whole and extensive new system, only matured and well documented projects have been explored.

5.2.1 List of classification methods

Classification is the process where texts are categorised in to predefined groups, or labelled with relevant tags. In this subsection, multiple text classification algorithms are explored to

determine the most effective solution for emotion tagging.

Naive Bayes

Naive Bayes classification works as a collection of classification algorithms, based on Bayes his theorem. The base theorem itself lies at the base of a whole branch of statistical studies. The idea of the theorem is to use historical data to calculate statistical predictions for the future, which should result in more reliable results. As the Naive Bayes classifiers work as a collection, although independent, a classification, such as an emotional one, should end up more reliable. However, it is required that every used algorithm uses the same common principle of classifying data. In the end, in most cases of text classification, the probability is calculated P(c|x), where c is the class of possible results and x is the instance that has to be classified whilst representing some features.

$$P(c|x) = P(x|c) * P(c)/P(x) \tag{5.1}$$

In general, Naive Bayes is mostly used in NLP problems, to predict certain tags associated to texts. In the case of IM, the probability of all listed emotions (i.e. happiness, sadness) will be calculated. After this calculation, the classification with the highest probability score 'wins' and the label is outputted.

Support Vector Machines

A Support Vector Machine (SVM) is a supervised machine learning algorithm, based on the research performed by V. N. Vapnik (2000) and co-workers. They are regarded to be one of the most sturdy prediction models, and like Naive Bayes SVM is based on statistical statistical learning frameworks, by V. Vapnik and Chervonenkis (1974).

Traditionally, SVM maps given data and classifications into two given categories. Then, as new data without classification is provided, predictions will be run and the new data will be classified in the same manner. This is referred to as linear classification. Additionally, using a so-called *mathematical kernel trick*, data can be mapped in a dimensional space.

Convolutional Neural Networks (CNN)

Convolutional filters were originally designed to find spatial patterns in data, and perform particularly well in computer vision projects. However, they can also be applied to find patterns in sentences and words as the same type of *tagging* is just as relevant in emotion classification. Most emotion classification from text projects use a one-dimensional CNN, since the data is one-dimensional, which means that the kernel moves in only one direction, with the output data being two-dimensional.

Multi-class text classification with BERT

BERT, meaning Bidirectional Encoder Representations from Transformers, is one of the latest and state-of-the-art machine learning models used for NLP created by Devlin et al. (2018) at Google and is used in their NLP products. They write that BERT achieved one of the highest accuracy for numerous language tasks in that year. The standard version has support for "104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters", with the larger model providing a "24-layer, 1024-hidden, 16-heads, 340M parameters" model. In 2020, Google released smaller versions of the model on their Github page [2], which can be executed on less powerful hardware. BERT outperforms previously available models as it is "the first unsupervised, deeply bidirectional system for pre-training NLP" according to Devlin et al. It is accessible to use in personal projects, as a pre-trained model and can be downloaded for to be used for text classification.

Using BERT for this research

For this thesis, the decision was made to go forward with the newest reliable technology possible. Therefore, it was decided to use BERT based recognition software to develop a proof-of-concept. Since BERT is open-source and actively maintained by various parties, there is plenty documentation and example code available, in the form of open-source projects.

5.2.2 Open source BERT projects

The development of a proof-of-concept prototype can be greatly facilitated by using an open source project. This approach allows for a more efficient and effective use of resources, as well as a deeper understanding of the technology and methods involved.

Based on the research, it was determined that models based on BERT are particularly well-suited for our purposes. Since its initial release, numerous variations and improvements have been proposed, which aim to further specialise the capabilities of the model. After conducting a review of available options, three open source projects that align with our needs and objectives were identified, which are considered the basis for our prototype.

BERT

Using the basic version of BERT, Lukas Garbas shared a Jupiter Notebook on his Github page[3], providing step-by-step instructions on how to apply BERT and ktrain[4] for emotion classification, including three datasets, and providing instructions on how to load others.

RoBERTa

RoBERTa by Liu et al. (2019), stands for ***R**obustly **O**ptimized **BERT** Pre-training **A**pproach*. It

[2] https://github.com/google-research/bert
[3] https://github.com/lukasgarbas/nlp-text-emotion
[4] https://github.com/amaiya/ktrain

was published by researches from Washington University and Facebook. They modified BERT to optimise the training of the model for it to become more efficient across the board. It achieves better results than the original BERT (Large) model and at the time of its release matched or improved state-of-the-art result like XLNet. RoBERTa is available at [5].

EmoBERTa

EmoBERTa is scientific work continued on RoBERTa, by Kim and Vossen (2021), aimed at "Speaker-Aware **Emo**tion Recognition in Conversation with Ro**BERTa**". Described "a simple yet expressive scheme of solving the ERC (emotion recognition in conversation) task", their research shows that they reach "new state of the art on the two popular ERC datasets using a basic and straight-forward approach". The project is available open source at [6].

5.3 Using EmoBERTa for emotion detection

EmoBERTa is based on the RoBERTa model, which is a state-of-the-art language model that has been pre-trained on a large corpus data. EmoBERTa extends this base model by incorporating additional training data and techniques that specifically target emotion recognition. As a result, EmoBERTa has been shown to outperform RoBERTa on tasks related to emotion detection and analysis (Kim and Vossen, 2021).

In light of these advantages, EmoBERTa was used as the starting point for the proof-of-concept prototype project. One of the factors that influenced this decision was the availability of comprehensive documentation and installation instructions, which are provided on the project its GitHub page and an accompanying YouTube tutorial, making it easier for to access and utilize the project.

Setting up the project

The project was set up by pulling it from the aforementioned Github repository. The detection software was set up to run in a docker container. Unfortunately, the hardware available during the development was not powerful enough to run the *large* model provided, so the *base* model was used. The models used were not speaker-aware, meaning that they do not keep the previous messages of the speaker *in mind*, and only classify utterances once at a time.

After setting up the Flask RESTful server application, the setup was tested with messages from the dataset created in 2.3.2 and the default ”client.py”. After debugging the project was running as expected and custom code was written to analyse the curated datasets.

[5] https://github.com/facebookresearch/fairseq
[6] https://github.com/tae898/erc

Custom processing scripts

To properly process and analyse the curated dataset and save the results, the datasets were converted JSON arrays listing the message and its emotion. Two versions of each dataset were created, both including and excluding *emojis*.

After, a function was written to process the created JSON files. The script loops over all elements in the array, and presents them to the RESTful server. The server response is then processed to save the detected emotion, all emotion detection results, and determine the confidence of the result of the detected emotions. Finally, results for all elements are collected and written to a separate JSON file for further analytical processing.

In 6, the creation of modified audio files using emotion will be added to this script. The script can be found in C, is based on the provided ”client.py” and is used to determine the effectiveness of the model.

5.3.1 Determining accuracy of EmoBERTa

To test the effectiveness, the EmotionPush dataset was fed to the RESTful server running the detection software using the custom processing script. This produced JSON file was imported to Google Sheets for analysis. In total, 280 messages were labelled. To process the results, rules for a *correct detection* were established.

Correct detection definition

When the server attempts to detect emotion in the message presented, there are three possible outcomes;

Correct emotion detected - the emotion corresponds to the emotion embedded in the message, meaning that the recognised emotion can be correctly used in aim 3 to apply vocal emotion to the TTS generation of this message.

No emotion detected - the server detected no emotion in the message and thus assigned the *neutral* tag. This means that in aim 3, if the detection model would be used to modify prosody of the generated voice, no modifications would be made to the generated speech.

Wrong emotion detected - the server detected the wrong emotion in the message. In this case, the wrong emotion would be applied to the TTS conversion in aim 3, a happy message could for example be spoken with an angry tone.

In principle, a message that does not have the correct emotion detected would be considered wrong. However, it can be argued that a *neutral* detection is not a wrong detection, since neutral is the indifferent emotion. There may be emotion in the message, however it could be dependent on contextual information of the conversation or personal contexts of a message

that are not recognisable without speaker-aware detection. In the case of a neutral message being assigned a neutral label, the detection is correct.

For analysing the data, the decision as made to do analysis of both definitions of a *wrong detection result*. The definition *accurate* will be used for emotion detection where the intended emotion corresponds only to the detected emotion, and the definition *accurate or neutral* will be used where the intended emotion corresponds to the detected emotion, or is labelled as neutral. This resulted in three tables of results in the spreadsheets.

Detection confidence threshold

With each result, a detection confidence is presented. This detection confidence ranges from 0% to 100%, and shows how sure the model is about the detected emotion. A detection result of 91% indicates a very strong confidence that an emotion is present, while 14% indicates a very low confidence. This detection confidence can be used to avoid in*accurate* detection results, by setting a minimum confidence value. If the confidence is under this threshold value, the emotion could be regarded neutral to avoid incorrect labelling.

Data processing spreadsheet

The detection results were loaded in to a Google Docs spreadsheet, with one message per row. For each row, it was determined whether the detection result was *accurate*, and this result was then saved as a true or false value. The same was done for every message to check if the result was *accurate or neutral*, and for 10 confidence threshold intervals. This data was then converted to a table.

Each table consists of 10 columns, where each column represents a detection threshold, ranging from 0% to 90%. Using these columns, the influence of the detection threshold on the accuracy of the algorithm can be examined.

Table 5.2 shows the detection accuracy of *accurate* detection results.

5.4 Concluding Emotion Recognition software / Conclusion

How and to what extend can software classify emotion in instant messages; and can the results be applied in a prototype setting to modify speech synthesis to portray emotion in vocalised messages?

The curation of the IM datasets has provided this research with a solid continuation base to transition to the last aim. The datasets and its messages are set up in a neutral manner, based on real-life communication tagged with emotion, verified by scientific research. The curated datasets were sanitised, processed and fed to the emotion recognition software successfully.

The explored classification methods for sentiment analysis, and thus emotion recognition soft-

ware, shows that it came a long way and is getting increasingly capable. Where in the early days sentiment analysis was one-dimensional, based on static rules such as the bag-of-words, growing through Support Vector Machines and Convolutional Neural Networks, BERT now allows for emotion recognition using machine learning. The amount of research, open source projects and information on BERT projects is impressive and is poised to take an increasingly significant role in NLP. In this research, the research and open source code by Kim and Vossen (2021) provided a solid base to uncover the capability of software to classify emotion in IM, and use the insights to develop a prototype.

As table 5.2 shows, the current implementation of EmoBERTa is about 66,4% efficient in emotion recognition using accurate or neutral detection results (50,7% in case of accurate detection results). Although this number is impressive, it is not considered reliable enough to be used as the support for modifying SSML of IM, as in 33,6% (or 49,7%) of the instances it is incorrect. This means that in many cases, incorrect emotion characteristics could be applied to vocalised IM, omitting the expected benefits of adding emotion, and possibly making it harder to recognise emotion.

Therefore, to not negatively impact the final experiment in the following aim, it was decided to not base emotion-enabled TTS conversion on detection results of EmoBERTa. Instead, emotion labels included in datasets are used to enrich audio files. This takes a variable out of scope in the final test, where results are not connected to the previous aim its results, making the insights from the final test standalone and more valuable.

Conf. thr.	None	10%	20%	30%	40%	50%	60%	70%	80%	90%
Overall	***50,7%***	***50,7%***	***50,7%***	***50,7%***	***47,9%***	***42,9%***	***37,9%***	***32,1%***	***27,1%***	***15,0%***
Anger	65,0%	65,0%	65,0%	65,0%	60,0%	50,0%	50,0%	45,0%	40,0%	25,0%
Disgust	30,0%	30,0%	30,0%	30,0%	20,0%	20,0%	10,0%	0,0%	0,0%	0,0%
Fear	30,0%	30,0%	30,0%	30,0%	30,0%	30,0%	10,0%	0,0%	0,0%	0,0%
Joy	95,0%	75,0%	75,0%	75,0%	75,0%	65,0%	65,0%	60,0%	55,0%	35,0%
Neutral	90,0%	90,0%	90,0%	90,0%	90,0%	75,0%	75,0%	65,0%	55,0%	35,0%
Sadness	35,0%	35,0%	35,0%	35,0%	30,0%	30,0%	25,0%	25,0%	20,0%	10,0%
Surprise	30,0%	30,0%	30,0%	30,0%	30,0%	30,0%	30,0%	30,0%	20,0%	0,0%

Table 5.2. Accurate detection results*; percentage where EmoBERTa base detected emotion in the 2.3.2 EmotionPush dataset correctly. Underlined column represents raw data without any thresholds.*

Conf. thr.	None	10%	20%	30%	40%	50%	60%	70%	80%	90%
Overall	***66,4%***	***66,4%***	***66,4%***	***68,6%***	***72,1%***	***77,9%***	***82,1%***	***85,7%***	***88,6%***	***95,7%***
Anger	85,0%	85,0%	85,0%	90,0%	90,0%	90,0%	90,0%	95,0%	95,0%	100,0%
Disgust	45,0%	45,0%	45,0%	50,0%	55,0%	60,0%	75,0%	80,0%	80,0%	90,0%
Fear	50,0%	50,0%	50,0%	50,0%	60,0%	70,0%	80,0%	80,0%	85,0%	90,0%
Joy	85,0%	85,0%	85,0%	90,0%	90,0%	95,0%	95,0%	95,0%	100,0%	100,0%
Neutral	90,0%	90,0%	90,0%	90,0%	90,0%	90,0%	90,0%	90,0%	95,0%	95,0%
Sadness	65,0%	65,0%	65,0%	65,0%	70,0%	70,0%	70,0%	80,0%	85,0%	95,0%
Surprise	45,0%	45,0%	45,0%	45,0%	50,0%	70,0%	75,0%	80,0%	80,0%	100,0%

Table 5.3. Accurate or neutral detection results*; percentage where EmoBERTa base detected emotion in the 2.3.2 EmotionPush dataset correctly or did not detect emotion. Underlined column represents raw data without any thresholds.*

6. USER EXPERIENCE OF SPOKEN INSTANT MESSAGING

In this aim, the goal is to determine the experience of voice-based instant messages, including modifications to voice prosody and addition of emojis to include emotion or non-verbal communication. First, the curated dataset based on EmotionPush (2.3.2) will be processed through Google Cloud TTS to generate audio files. Using those recordings, the final user research will be designed, executed, after which the insights on the effectiveness of emotion in TTS will be gathered.

Combining the findings of aim 1 and 2, this final aim aims to investigate the effects of variations of text-to-speech synthesis, including and excluding nonverbal and emotional cues, using a high-fidelity prototype. These include multiple variations such as; plain synthesis without vocal nonverbal cues, using SSML to include non-verbal speech characteristics, and speaking aloud the emojis in IM. Insights gathered during aim 1 and 2 are deciding factors on the types of text-to-speech conversion methods that will be applied.

6.1 Including emotional prosody in speech synthesised Instant Messages

Aim 2 (5) concluded that the use of emotion recognition software as the source of emotion tagging information resulted in an unreasonable amount of labelling errors, due to a high percentage of detection being incorrect or neutral. Therefore, only intended and known correct emotions tags for the instant message will be used for TTS. Thus, the variable of emotion detection accuracy will not influence the research insights of this research aim. In other words, software based emotion detection will not be used in this aim.

6.1.1 TTS using WaveNet on Google Cloud

As discussed in 2.4.2, WaveNet is one of the latest improvements in audio generation. Since the release of the paper by van den Oord et al., 2016, developments to the model have con-

tinued, which are available to use through Google Cloud[1]. The platform supports adjustments of the characteristics duration, intensity and pitch through SSML and programmable methods. Therefore, the decision was made to generate the audio files using Google TTS, with the latest *Neural 2* voice. Upon implementation, the UK English based voice generated surprisingly more robotic speech than its US English counterpart. The voice selected to use for this thesis was *en-US-Neural2-C*. Furthermore, as the WaveNet paper describes, the generated voices sound natural, and the documentation for the service by Google is comprehensive and straightforward.

Modifying voice characteristics

The generated audio files were modified based on the categorisation of positive and negative emotional vocalisation characteristics by D. Sauter et al. (2010) as found in 2.1. It was not possible to reliably modify the amplification onset using currently available tools. For modifications of other variables, the approach described below was taken, based on the documentation by Google Cloud [2].

Duration of a spoken sentence, measured as relative change from a neutral sentence. Implemented using the SSML tag *rate* using the percentage value, documented in the W3 SSML standard[3].

Intensity, or volume, was implemented using the *volume* tag. This tag uses decibel values, where +6db is double the intensity in comparison to baseline volume, and -6db is half the intensity compared to baseline volume, according to the W3 SSML standard[4]. Values used by D. Sauter et al. (2010) were measured and provided in percentages, requiring them to be converted to decibels using the formula below, before adding them to the tag.

$$x = 10log_{10}k \tag{6.1}$$

Pitch Mean was modified using the *pitch* value of the *prosody* element. The values are provided as a relative change, which resulted in the relative change provided as a percentage as *(pitchMean*100)+'%'* to the pitch tag.

Structure of the vocalised instant message

When using an receiving instant messages through an assistant, the interface that delivers the message is vocal. Both Siri by Apple and the Google Assistant first announce an incoming message, including the sender their name, before proceeding to read aloud the message. To make the audio files represent a real life situation as best as possible, this interface has been

[1] https://cloud.google.com/text-to-speech
[2] https://cloud.google.com/text-to-speech/docs/ssml
[3] https://www.w3.org/TR/speech-synthesis11/#S3.2.4
[4] https://www.w3.org/TR/speech-synthesis11/#S3.2.4

recreated. Using SSML, the following is announced before the actual message: *Message from Thomas*.

As the EmotionPush dataset is anonymised and the sender names are not available, these blanks were filled. To select English names that fit the language of correspondence, the dataset "Top 100 baby names in England and Wales: historical data" by the UK Office for National Statistics (Corps, 2014) was used. The ten most common male and female names were added to an array and randomly selected for use in the speech files upon generation of the files.

Example SSML code

The process described above resulted in automatically generated SSML code to send to Google TTS, of which an example can be found below.

```
<speak>Message from Thomas:
    <break time="400ms"/>
    <prosody volume="+0.864dB" rate="91%" pitch="-46%">
        You put me in such a position in front of others
    </prosody>
</speak>
```

The SSML code for an angry instant message from a sender named Thomas.

6.1.2 Vocalisation of spatial arrays and emojis

The results of the contextual inquiry in aim 1 (4.2.3) confirmed that spatial arrays, emojis and other methods of various levels of visual non-verbal communication in messaging are considered an important aspect of embedding, and thus recognising, emotion in IM. Participants noted they especially use emojis to affirm emotion in messages.

To research the importance and effects of using emojis in TTS, it was decided to include the vocalisation of spatial arrays and emojis in to the final research, which will be explained in detail later in this section. To vocalise spatial arrays, these were converted to their emoji equivalent according to the Full Emoji List provided and maintained by Unicode [5]. Thereafter, all emojis in messages were converted to their *CLDR Short Name*, and the whole message was converted in to speech in the same manner as explained in 6.1.1.

6.2 Methodology: High fidelity prototype experiment

To research the influence of variations of audio synthesising methods, an experiment was designed to present research participants with four high-fidelity prototype variations of audio files

[5] https://unicode.org/emoji/charts/full-emoji-list.html

of selected Instant Messages as discussed in the previous section;

1. Vocalised IM with no modifications to voice characteristics (neutral/control)
2. Vocalised IM with no modifications to voice characteristics, emojis spoken aloud
3. Vocalised IM with voice prosody characteristics modified based on embedded emotion
4. Vocalised IM with voice prosody characteristics modified based on embedded emotion, emojis spoken aloud

These variations are evaluated with 20 research participants divided in to four groups of five, fitting the study focus group as discussed in 3.1, with exception of the age group; this was expanded to participants up to the year of 2000 to make it more convenient to find research participants. On top of the general requirements, participants were required to not use voice assistants to send IM on a regular basis, as well as to not be familiar with the research. All experiment variations concern the same message content, the aspect that changes is the TTS modifications and/or the vocalisation of emojis.

The group size of five was determined by two factors. First is a resource and time limitation, a large research sample is not within the scope of this thesis. The second factor is that five test subjects per variation are considered enough to uncover shortcomings and qualities in interaction designs as described in the paper by Nielsen and Landauer (1993). It can be argued that this is not a qualitative, rather a quantitative aspect of the research, and in this light the amount of research participants is not considered satisfactory to find statistically significant research insights. However, this is a conscious limitation as the goal of this thesis is to uncover qualitative insights aimed at further research, and not to statistically prove preferred IM synthesising methods.

Experiment execution

To portray the test as close to a real-life situation as possible, audio files are played through a Google Home, while the participant is seated at a desk. The Google Home is connected over Bluetooth to a computer playing the audio files.

Before commencing the experiment, the research is introduced, with an opportunity to ask questions, and sign the consent form. After, the participants are to get familiar with the Google Assistant, by asking questions which may include inquiring about the current weather, asking when the next bus is leaving and playing some music. This is done so that the participant gets acquainted with the interface and functionality of voice assistants.

Once the participant is comfortable with the VPA, the forms used to evaluate the messages are introduced and explained, and the participant will be played a neutral, introductory incoming message to get familiar with the test, its questions and how to evaluate them.

A/B-inspired testing

The first part of the research is approached as an A/B test. A/B tests are used to compare two or more version of a similar design to see which version performs the best. A/B tests help to understand which variations in designs generate better results, but will not help to understand why a particular design was preferred. Similarly, Nielsen (2015) advises it to only be used in projects that have a clear goal, or in other words, experiments with a clear key performance indicator.

Aware of these limitations, and aimed to gather insights, after experiencing an incoming message, participants are requested to rate the emotion experience and its intensity on a 7-point Likert scale. After completion of the evaluation part, participants are asked to fill an UEQ+ questionnaire and are asked for comments on their experience.

Experiment part 1 asks the participants to listen to 14 message audio files: 2 times 6 of each emotion, plus 2 neutral messages. These messages are presented in a randomised order. After listening to each message, the participant is asked to rate the emotion they found present in the message, and the intensity they found the emotion on a 7-point Likert scale, from *very minimal emotion* to *very strong emotion*. This rating is done on paper to avoid distraction and for simplicity. During evaluation, participants can request to replay messages, as would be the case in a real-life situation. After evaluating all messages, participants are asked to fill a modular UEQ+ questionnaire (6.2), on the scales attractiveness, stimulation, novelty, intuitive use and usefulness. Finally, the participant is asked for comments on their experience.

Experiment part 2 is an in-person survey to collect more data on the experiment variations of vocalisations that the participant has not yet been exposed to. It presents participants with 12 more messages, 4 messages for each inexperienced variation. Participants receive another paper 7-point Likert scale form to review their observations.

After evaluating all messages, participants are asked three questions written in a semi-structured interview. Participants were given a bar of chocolate as appreciation for their participation.

During both parts of the experiment where the participant is reviewing the audio files, as well as evaluating the UEQ+, a distance of 5 meters to the participant was kept to avoid distraction and avoid introduction of bias.

Modular User Experience Questionnaire (UEQ+)

To measure the user experience of the four variations of embedding emotions in spoken messages, in a simple and efficient manner, the User Experience Questionnaire (UEQ) was used. This questionnaire, of which research was initiated by Laugwitz et al. (2008) and built upon over the years, is meant to evaluate interactive products. However, in this research, what is tested is part of an interactive product. This means that the interactive evaluation aspects of the UEQ

are irrelevant. In 2021, Schrepp published “Measuring User Experience with Modular Questionnaires”, allowing for a modular approach to measuring the various dimensions of the UEQ. This way, the interactive dimensions of the questionnaire can be left out, while still collecting data on the attractiveness, stimulation, novelty, intuitive use and usefulness of the implementation. A data analysis tool is provided to process the responses from participants. The participants were asked to answer the questions of the UEQ+ after completing the listening part of the research.

Semi-structured interview

A short, three question semi-structured interview is held to wrap up both parts of the experiment and gather possible fresh insights on the user experience and further development focus. The questions consist of gathering information on the participants experience of emotion during the experiment, which specific elements of the message or vocalisation made them recognise emotion, and finally which additions or changes to the vocalisation or IM would make it easier for them to recognise emotion.

Review using data using comparison and visualisation

Collected data on the recognition and experience of non-verbal and emotional communication in vocalised IM will be gathered and processed in a spreadsheet, where the results of each variation will be plotted side-by-side to show the correct detection of emotion and experienced emotion intensity, as well as in a table.

The data collected from the UEQ+ will be processed through the designated Excel sheet, after which the results are compared both visually and in a table.

Finally, comments and interview answers from the participants will be gathered during the experiment, and will be enriched using voice recordings.

Limitations

Based on the findings of aim 1 (4), an important aspect in the usage and recognition of non-verbal communication is the established relationship with the sender, and awareness of their communication style. In this experiment, such relationships and awareness with the sender are not present, due to such test arrangements being complex and thus requiring a significant amount of time.

It is assumed that voice based IM is mostly used in situations where the receiver is not able to use their communication device as they are preoccupied with other tasks. In this experiment, the participant is not preoccupied with other tasks while evaluating the messages. This was a conscious decision, as to not distract the participants whilst they are writing down their evaluation, to avoid erroneously filled observations.

6.2.1 Hypothesis

Based on the experiment set up and background research, a set of hypothesis was established for the high-fidelity prototype;

Hypothesis 1: Modifying prosody of vocalised instant messages to include non-verbal communication characteristics improves emotion recognition compared to the control by experiment participants.
Hypothesis 2: Speaking aloud spatial arrays (emojis) in instant messages improves emotion recognition compared to the control by experiment participants.
Hypothesis 3: Modifying prosody of vocalised messages to include non-verbal communication characteristics improves novelty, intuitive use and stimulation compared to the control.

These hypothesis will be used as a lead to display the findings of the experiment.

6.3 Findings

This section dives into findings of the previously introduced experiment, divided in three subsections; the participants their experience of emotion, the user experience of the prototype and interview insights and test observations.

A total of 20 research participants participated in the test, with their age ranging from 22 to 35, with an average of 25,1. All participants were students at Tampere University, and were physically present at the tests. Their background was European. Of the participants, 13 were female and 7 were male. Their field of study was varied, including business, healthcare, IT, politics and biology.

Don't have the data at hand at the moment, will update in next revision.

6.3.1 Experience and recognition of emotion in speech-based IM

This subsection describes the main insights regarding the executed prototype experiments and its variations. It will uncover which version is presumed the most effective at conveying emotion using speech, based on the quantitative data gathered. The data will be presented using graphs, where the raw data is available in appendix B.1.2.

Effect of speech prosody modifications and inclusion of emojis on recognition of emotion category

During experiment part one, participants were presented with one of four vocalisation variations as established in 6.2. In this part of the experiment was no exposure to other variations of prototype. In the second part of the experiment, participants are exposed to the other versions

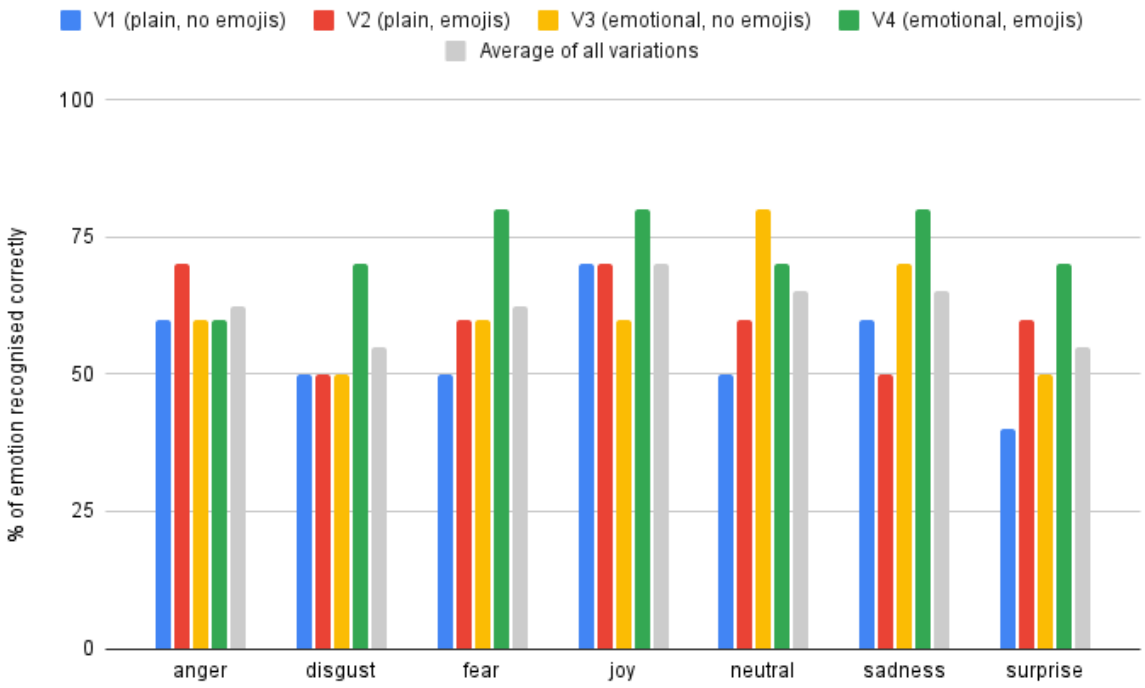

Figure 6.1. *Percentage of emotions in messages correctly recognised by participants, per variation (only test part 1). Raw data in appendix D.5.*

they did not yet experience. The answers from participants were processed in a spreadsheet, ordered by variation and emotion, and finally plotted as a graph as seen in figures 6.1 and 6.2. The average intensity of emotional experience can be found in figures 6.3 and 6.4.

Variation 1, the control variation, plays out below-average for correct emotion recognition in all emotional IM except joy, as well as the neutral messages. However, the average offset differs in negative emotions such as anger and disgust, compared to positive emotions such as surprise and joy. As seen in figure 6.1 and 6.2, joy is an outlier. This is the single occurrence where the control variation does not come to be the lowest percentile comparatively. When applying focus to figure 6.2, which represents data from both experiment part one and two combined, average difference between the control group and other variations generally increases.

Variation 2, which uses a neutral voice while including emojis, performs equal or better compared to the control version in the first test, with an exception for sadness which performs lesser than the control. In the case of anger, the percentage of correctly recognised emotion is 20% higher compared to other variations. In negative emotions disgust and fear, variation two performs the same or marginally (10%) better than variation one. Regarding positive emotions joy and surprise, the former performs similar compared to the control, where correct recognition of the latter increases with 20%. However, when including the results of the second part, the correct recognition of joy dips to a relative low, where the difference of both surprise and anger shrinks to 10%. Disgust and fear show an increased correct recognition, by respectively 20% and 15%. Overall, the second variation generally performs near-average.

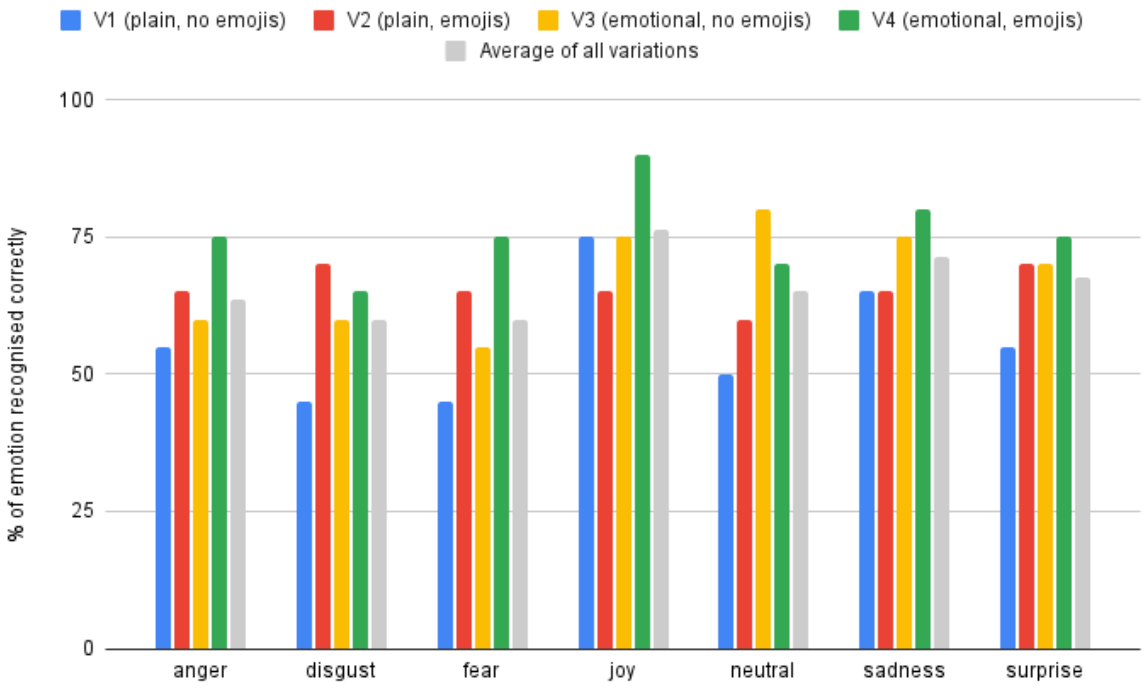

Figure 6.2. *Percentage of emotions in messages correctly recognised by participants, per variation (test part 1 and part 2 combined). Raw data in appendix D.2.*

Variation 3, which uses no emojis while including non-verbal communication traits, shows no exceptional results. In the cases anger and disgust, correct emotion recognition is equal to the control variation. Fearful, surprising as well as sad spoken IM is recognised correctly 10% more often compared to the control variation. As described before, in the case of joy, the correct recognition of emotion is 10% lower. However, when combing the recognition results of both experiment parts one and two, the voice with non-verbal communication traits applied using prosody modifications consistently performs better in comparison to the control variation. Solely in the case of joy, the performance for both variations is equal.

Variation 4, utilising both emojis and modified prosody, resulted in the highest percentile correct emotion recognition across the board in test part one, with an exception for anger as well as neutral messages. The recognition of fear and surprise with a 30% increase is considerably higher than the control variation, as well as a 20% increase for neutral messages, or those containing disgust or sadness. Anger stays on-par with variation one and three, while neutral message emotion recognition by participants is higher only in variation three. Overall, variation four scores best in five out of six emotion categories in experiment part one. When including the second part of the experiment in to the dataset, anger is recognised the most often for its category, whilst disgust is recognised less accurately by participants, being more often correctly recognised in variation 2.

Over all categories it is observed that the inclusion of emojis generally improves the emotion recognition, and in cases where it does not, recognition stays on-par with variation one. The

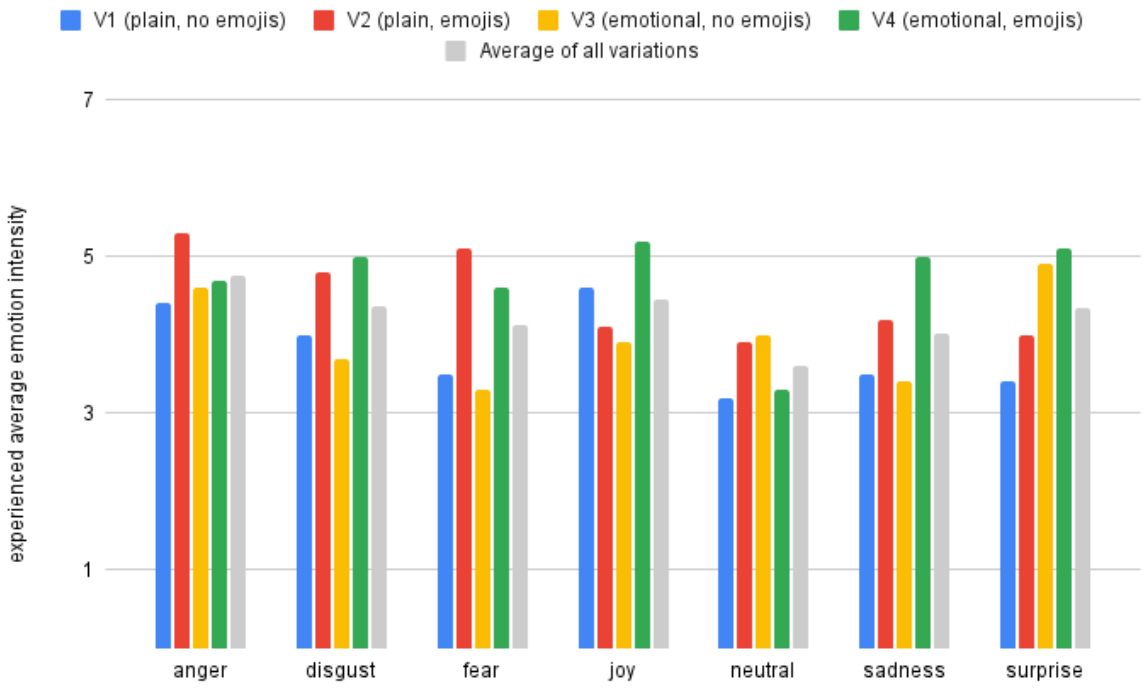

Figure 6.3. *Average emotion intensity rating on a 1-7 Likert scale, ordered by variation and emotion (only test part 1). Raw data in appendix D.3.*

same applies to the modification of speech characteristics to include emotion, with an exception for joy in experiment part one. When both non-verbal communication traits are combined, it generally results in the highest percentage of correct recognised emotion. In some categories, the control or neutral variation performs similar to the adjusted variations.

On average of all versions, the emotion joy is recognised correctly the most often, followed by sadness, anger and fear, and disgust and surprise. When including the second part of the experiment to the dataset, joy is still recognised correctly the most often, followed by sadness, surprise, anger, and a shared last place for disgust and fear.

Effect of speech prosody modifications and inclusion of emojis on intensity of experienced emotion

Besides evaluating the emotion experienced, participants were asked to rate the intensity of the emotion they experienced. This ranged from 1, *very minimal emotion*, to 7, *very strong emotion*. Figure 6.3 displays the average of the emotion intensity experienced, ordered by emotion and variation. This regards just the first part of the test, where participants have not been exposed to other variations of the prototype. Figure 6.3 represent both parts of the experiment, including part 2.

Variation 1, the experiment control variation, prompted participants to indicate a below-average emotion intensity experience comparatively in all categories, except for joy, where the control

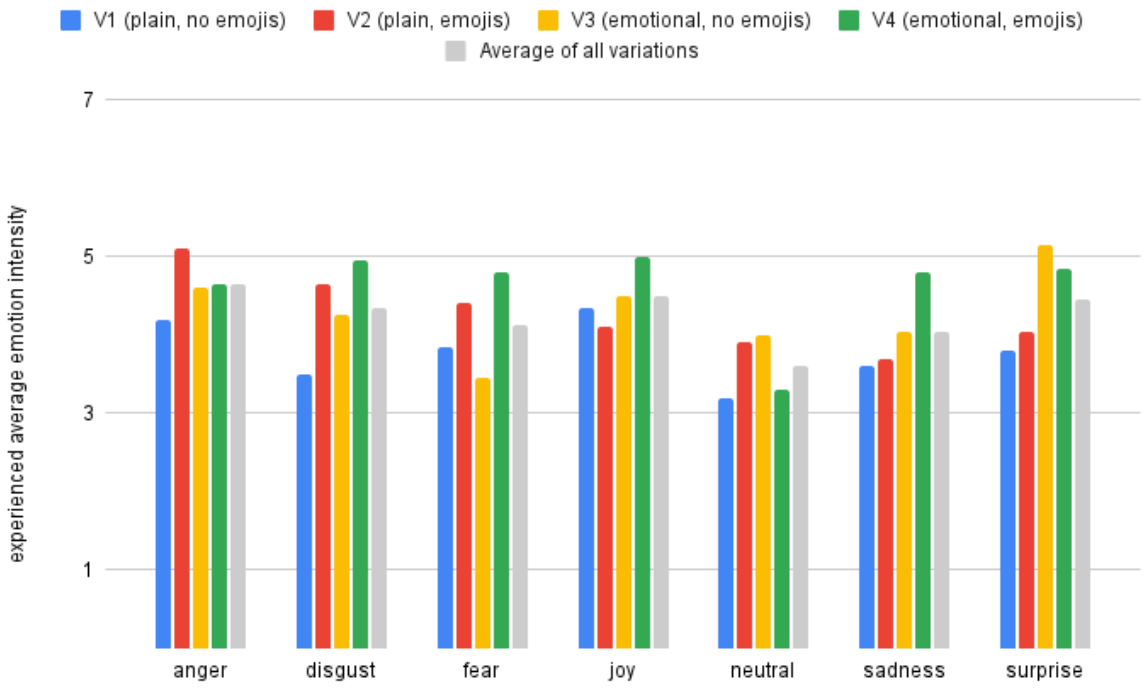

Figure 6.4. *Average emotion intensity rating on a 1-7 Likert scale, ordered by variation and emotion (both test part 1 and 2. Raw data in appendix D.4.)*

variation is rated as the second-highest. However, the observed percentual difference varies in intensity depending on the category. When including the second part of the experiment to the data, as seen in 6.3, the changes to the data are relatively small.

Variation 2, experiment part one, in which includes emojis whilst embodying a neutral speech, makes participants experience an increased emotion intensity, except in the case of aforementioned joy. In the cases of the mostly negative emotions anger, disgust, fear and sadness, the experienced emotion intensity is reported higher than the average of all variations combined. Variation two includes the fourth-highest experienced emotion intensity; fear, as well as the highest experienced emotion intensity; anger, with an average score of 5,3. In the case of both aforementioned emotions, based on provided data, emojis yield the highest emotion intensity. In the case of the positive emotions joy and surprise, the experienced emotion intensity is higher than the control variation, although lower than the combined average of all variations. When including the second part experiment data to the analysis, all emotion categories show a decreased experience of intensity, except for surprise messages which raise with 0,05 points.

Variation 3, the variation omitting emojis whilst including non-verbal voice characteristics, prompts participants to report similar experienced emotional intensity for anger, disgust, fear and sadness, compared to the control variation. In the categories neutral and notably surprise, the variation is reported to be containing strong emotional intensity. Notably, in the categories disgust, fear, joy and sadness, the experienced emotion intensity of participants is the lowest of all variations, and the variation in general is never rated highest in any emotion categories.

When including the second test in the dataset (6.3), all but the emotion anger take an uptick in experienced emotion recognition, with joy taking the biggest leap of 0,6 points.

Variation 4, combining of both studied methods of including non-verbal communication in to computer generated spoken IM, yields the highest experienced emotion intensity regarding emotions disgust, joy, sadness and surprise in the first part of the experiment. Anger, fear and disgust perform relatively similar compared to variation two. When including the second experiment part data in to the analysis, fear overtakes variation two as rated the most intensely experienced, while the highest position for surprise is overtaken by variation three.

Over all categories, based on the data of experiment part one, it is observed that negative emotion categories (anger, disgust, fear and sadness) generate a stronger experienced emotion intensity using the inclusion of emojis. Messages including spatial arrays as non-verbal communication, are reported to be more emotionally loaded by 1 point on average, compared to the control variation. With a 1,6 point higher average emotion intensity experience score, fear takes the biggest leap upwards. However, when omitting emojis and solely modifying non-verbal characteristics of used voice, aforementioned emotions show similar experienced emotion intensity compared to the control variation. In the case of joy, either including emojis or modifying voice characteristics did not yield increased emotion experience, while combining them gave the highest results. However, experienced emotion intensity of surprise increased with the inclusion of emojis, and increased substantially more applying the modification of voice characteristics.

Overall, anger is the most intensely experienced emotion, followed by joy, surprise, fear, disgust and finally sadness. When including the data of the second part of the test, fear and disgust swap places.

6.3.2 User experience of modified speech based instant messaging

This subsection will present the results of the UEQ+, per scale. The data presented concerns means, and was collected after the participants evaluated the first part of the experiment, before being interviewed, and before the second part of the experiment.

Attractiveness

The scale attractiveness describes the research participants their overall impression of the product. It inquires whether the participants experiences the product they evaluate is enjoyable, pleasant, good and friendly. The scale is applied as this product aims to be provide the user with a method to finish the task of receiving IM without unnecessary effort.

In the observed results of the UEQ+, the control variation performs above average and second-highest. Interestingly, the inclusion of only emojis in variation two results in the lowest score

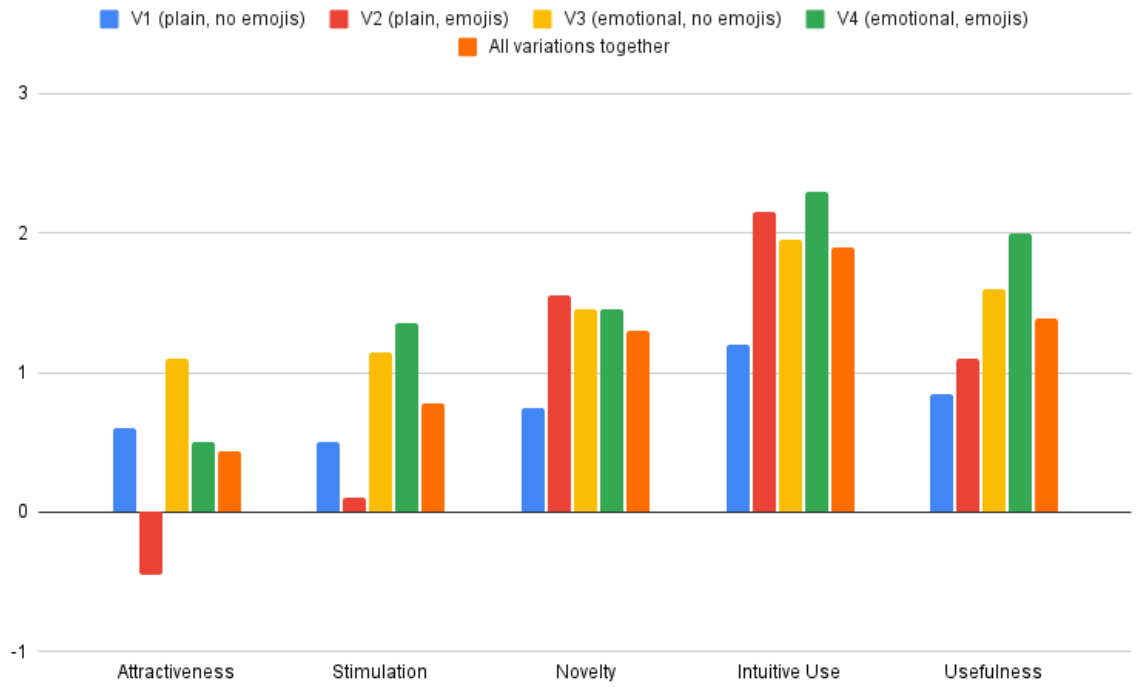

Figure 6.5. *Results of the UEQ+, four variations and all variations combined, displayed per dimension*

across all variations and scales. Variation three, which includes solely modifications to the prosody of speech, performs best with a score of 1,1 between -3 and 3.

Although attractiveness in comparison to the other scales performs relatively low, all scales cannot be compared one to one. Unlike the regular UEQ, the UEQ+ is not an established questionnaire and the results cannot be interpreted in comparison to benchmarks or studies of existing products (Laugwitz et al., 2008).

Stimulation

The stimulation scale applies to all types of products, and concerns whether the product is interesting and fun to use. Regarding the experiment, the mean of the control variation is below the average mean, although not the lowest. The addition of emojis in variation two just performs worse. Variation three performs significantly more strong, where variation four results in the highest score.

Novelty

Novelty asses the impression that the idea of the product is creative and original. In this scale, the control variation performs at a score of 0,75 between -3 and 3, which is roughly half of the other means of the other variations, which perform at 1,55, 1,45 and 1,45. This implies that variations with modified prosody characteristics and/or emojis are more creative and original in

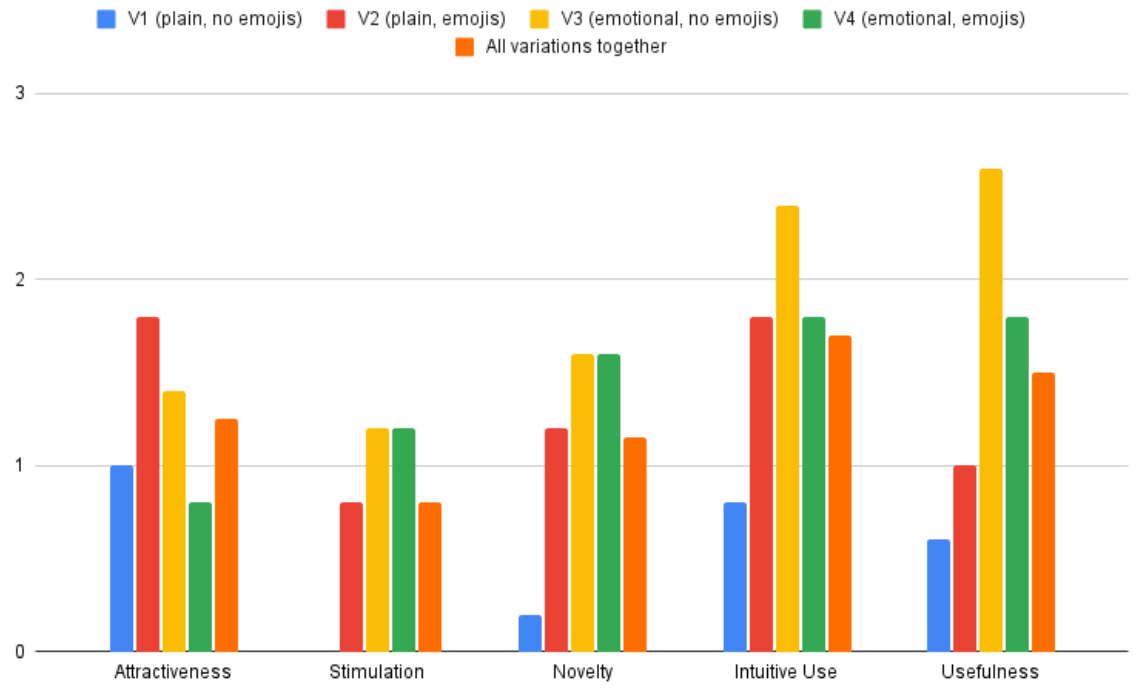

Figure 6.6. *Importance ratings of the UEQ+, four variations and all variations combined, displayed per dimension, ordered by variation*

the experience of participants, where the control embodies less of these characteristics.

Intuitive use

This scale analyses the impression that the product can be used immediately without need for any training or help. Similar to the novelty scale, the control variation performs lower compared to the other variations. However, the difference in means lies at less than double in this scale, the control variation mean is 1,2, where the other variations respectively score 2,15 1,95 and 2,3. Again, variation four its mean is the highest of the bunch.

Usefulness

Usefulness analyses the impression that using the product is beneficial. Comparing the results of the questionnaire, it is noticed that the control variation again scores lowest. Variation two, with only emojis, scores slightly higher in comparison. Variation three again takes another step upwards, and when combining all non-verbal communication traits in to the vocalisation of IM, variation four scores the highest of all with a mean of 2.

Overall insights

In general, when analysing the results of the UEQ+, it is noted that variation four scores among the highest scores consistently, expect for the case of attractiveness, where participants prefer a variation with solely speech prosody modifications.

6.3.3 Interview insights and experiment observations

The interviews gathered qualitative data on participants their experience of voice-based IM, with and without modifications. This subsection will describe the findings, starting with general insights which apply to all forms of IM vocalisation, where-after focus will narrow down on the effects of modification of voice characteristics, as well as the effect of presence of emojis. Finally, improvement suggestions of participants will be laid out.

General vocalisation insights

In the first part of the experiment, participants are only exposed to one variation, providing clear insights in the experiences and limitations without influence of other variations. During the interviews, there were recurring comments for all four variations of IM vocalisations, by all participants. These comments apply to the general vocalisation of IM, base characteristics of the voice, intonation, speed and other similar factors.

All twenty research participants voice that the content as well as context of a message is (one of) the key aspect(s) of recognising emotions in messages. Additionally, interviewees noted that knowing the other party personally is an important part aforementioned context. Participants suggest that an existing relationship and personal experience with communication contributes to successful IM and thus recognition of emotion. This is in line with the conclusions by Hancock et al. (2007) as written about in 2.2.1.

On the subject of the TTS voice, twelve out of 20 participants laughed during the speech synthesis at some point of the experiment. When asked to explain this reaction during the interview, participants note that this is due to the "robobotic sound" (i.e. 4, 6), the synthesisation of lexical surrogates 1 and the way abbreviations such as *rn* (*right now*) are synthesised as letters instead of their written out meaning. To increase emotion recognition, 7 participants noted that improved intonation, for example better processing of one or more grammatical markers, would make it easier to understand the content of the message.

Participant 6 and 13 noted that a sudden cut-off in the vocalisation, for example at the end of a sentence, felt unnatural 13 and was distracting 6. Other participants note that "the flow of the message is bad" 10 and "the pauses are too long" 3. Participants noted that this is due to the lack of punctuation, or failure of vocalisation of it.

Participant 16 notes that the presented way of IM "elevates burden of texting", it provides a

manner of communication without the need to pick up a phone or use a screen. However, participant 14 feels different, as in their opinion "written reading is easier because you know what is what", meaning that they prefer to read the message as they can extract meaning better from the sender their written message.

Participant 19 voices that relatable messages are the easiest to digest, as those resonate in such ways that they themselves could have experienced the situations described or sent the messages presented.

Variations without voice prosody modifications (V1, V2)

In the control group, five participants examined vocalised IM without any emojis or adjustments to prosody. The most common comment on the vocalisation of the IM was the voice being "flat" (1, 2, 10) or "one-dimensional" 7. Six out of ten participants were laughing at the vocalisation quality least one point during the test, where 1 and 3 noted that they did get used to the voice after more exposure. In general, main recognition of emotion came from context and emotion words, not from the voice of the vocalisation: "context is the most important in recognising emotion, it's hard to hear emotion with flat voice" 3, "context made the decision almost 100 percent of the time" 7.

In this group, participants noted that their focus on recognising emotion lies with the use of non-gradable adjectives such as totally and completely, as well as the emotion words as previously described.

Voice with prosody modifications (V3, V4)

In the third and fourth variation, the prosody of the voice is changed to portray emotions, as described in 6.1.1. Research participants showed a divide in the experience of emotions; some noted emotion was distinctly present in the messages: "I wouldn't change anything to the voice, it is already quite human-like" 15, where others noted that the voice was still relatively plain: "[the voice was] underwhelming, not conversing very well" 16.

Where in variation one not a single participant notes they recognise emotion based on the voice of the assistant, in version three, four out of five participants recognise emotion using the voice characteristics of the vocalisation, "[I] recognised emotion using tone, pauses and breaks" 11, "the context resonated with the way *it* was saying the message, intonation was good" 15. In variation four, participants their interview focus and feedback mostly concerns emojis, which will be discussed later.

Participant 11 finds that they can "clearly feel emotions" and were "impressed with how it reads emotions, I was not expecting it to be this good", so much that they "feel like the options for rating were too limited, like there was no option for frustration". In general, the comments on the voice were positive, and participants were pleasantly surprised. Participant 18 feels the

same, saying that the assistant had "more emotion in voice then I ever heard before", referring to artificial voices.

On the contrary, participant 16 voiced that it felt uncomfortable to not be in charge of the intonation of the voice, as they usually read the messages with their own voice in their head. They similarly noted they were experiencing a lack of inflections, and that "AI should be able to understand the understand an emotion" which can be used to improve the voice. Participant 14 notes that "in some cases the voice sounds bland and automated, specifically with fear, ranting, concern" and that "throughout [the test] the voice sounded artificial".

In variations three and four, in the case of emotion recognition, participants preferred longer messages. The message being longer may contribute to the voice adjustments being more prominent, as well as more emotional context to base decisions on. However, in cases where the messages presented were lengthier and based on an emotion that shortens the prosody of the voice, such as surprise, the message played noticeably faster. This resulted in participants asking to replay the message more often than usual as it was too fast to successfully understand the content.

Variations with vocalisation of spatial arrays and emojis (V2, V4)

In variations two and four, vocalisations of emojis were applied. When inquiring about the ways emotion was experienced, participant 4 answered that "the dead giveaway were the emojis", meaning that for them, the emojis were significant in recognising emotion. Participant 13 voiced that "emojis and context help to recognise emotion". However, participant 9 mentioned that in their opinion, emojis have a "background function" in the context of emotion. For participant 5, the descriptions of emojis were distracting, where they mentioned that they were "thinking in my head which is which", meaning that they were thinking which specific emoji had been described, for them resulting in less focus applied to the actual content of the message. Participant 14 thinks that "emojis help in certain messages where the context is not clear", or in other words, that emojis have a supportive role. Participant 18 relies on emojis for 50% of the time to recognise emotion, to them emotion in messages is "50/50 content or emojis".

Participant 4 brings up that, in order to fully use a product that vocalises incoming messages for them, they should teach their friends how to use emojis. They argued that if their friends would use emojis in a correct manner, it would help them to recognise emotions from their messages more consistently. However, participant 4 also notes that "it's not always needed to include emojis" and that the emojis in some cases did not correspond to the content of the message. After listening to an incoming message, participant 9 mentions they would still look at their phone screen to see the visual representation of the emoji, to make sure they correctly understand it.

Voice-based IM improvement suggestions

"The assistant doesn't know the emotions" 1: a comment from a participant of variation one, the control variation without any modifications. They suggest that it would be bad if the assistant tries to interpret emotions, as in their opinion, it could make mistakes. When transitioning to the second part of the experiment, where they were exposed to the emotional variations, their sentiment changes to "the emotional voice is already good and strong". Participants suggest that a stronger emphasis on the emotion in the voice would help them detect emotion more efficiently.

An often voiced request is to guard the correct length of pauses and intonation in messages. In some cases, participants experience them as too short or too long, which is described as distracting. The same comments are made for the flow of especially shorter messages, which is often described as *off* or short.

Participants 12, 16 and 17 suggest the use of a more human voice, either through a more advanced language model, or through online services which imitate the voice of a human being, for example their friends, based on audio recordings. Participant 16 notes that this might make the voice more accessible to everyone, for instance those in retirement homes.

6.4 Conclusion of user experience of spoken instant messaging

The goal of this aim is to determine the experience of receiving voice-based IM, and whether it can be improved using modifications to the voice. This was done by processing the curated IM dataset through Google TTS, generating four variations of vocalisation: a control group, a variation with added emojis, a variation with added emotional prosody characteristics and a version with both emojis and prosody modifications. After creating the audio files for the test, this resulted in three hypothesis.

The Wavenet service by Google TTS used for the speech synthesis allows the prosody of the voice to be modified sufficiently to include non-verbal communication traits. Emotion in the voice of the vocalised IM was experienced by most of the research participants, some of which were notably surprised by the quality of the product. The message dataset, which contains real-life messages, was a good and natural fit for this experiment.

Hypothesis 1: Modifying prosody of vocalised instant messages to include non-verbal communication characteristics improves emotion recognition compared to the control by experiment participants

The data presented in D.4 uncovers that, in general, modification to the prosody of the voice to include non-verbal communication in the form of emotion, based on the research by D. Sauter et al., 2010, positively affects the recognition of emotion. In the case of joy, the change of

speech characteristics has a neither positive or negative influence on the recognition of emotion, where in the case of all other emotions (anger, disgust, fear, sadness and surprise), the correct recognition of emotion improves. Therefore, in this experiment, it can be concluded that the modification of the speech characteristics based on the emotional content of IM, improves the recognition of emotion in comparison to a neutral vocalisation, as presented with the control group. This confirms the research discussed in 2.4.1 in the context of artificial speech. The average experienced intensity of emotions similarly increases for every emotion, except fear which intensity is experienced slightly lower than the control.

Qualitative discoveries from performed interviews affirm the quantitative data gathered; seven out of ten research participants clearly self-indicate emotion recognition based on the characteristics of the voice, where none of the participants of the control variation voice such comments. Participants are impressed with the assistant "reads emotions" 11, and in general are pleasantly surprised. However, some participants voice that the emotional voice can be uncomfortable and automated.

Hypothesis 2: Speaking aloud spatial arrays (emojis) in instant messages improves emotion recognition compared to the control by experiment participants

The inclusion of spoken spatial arrays, such as emojis, on average increase the correct recognition of emotion embedded in vocalised IM. Disgust and fear increase most significantly compared to the control variation, followed by surprise and anger. Sadness performs equal to the control variation, where joy takes a dip. In the case of experienced average emotion intensity, the inclusion of emojis result in an increased in participant reported intensity in all emotion categories except joy. Therefore, the research suggests that the inclusion of emojis supports correct emotion recognition.

In the qualitative interviews, participants voice that the emojis are either a "dead giveaway" 4 of the emotion in the messages, or that they contribute to the detection of emotion. In few cases, the emojis are distracting to research participants.

Hypothesis 3: Modifying prosody of vocalised messages to include non-verbal communication characteristics improves novelty, intuitive use and stimulation compared to the control

To examine the user experience of participants during the use of the prototype variations, the UEQ+ was used. The mean results as presented and discussed in 6.5 show one or more variations rating increase across all dimensions when compared to the control variation. Novelty, intuitive use and usefulness ratings increase significantly. Attractiveness sees only one of the variations exceed the control variation, where the two other variations perform worse. In the case of stimulation, only the variation which solely modifications in the form of emojis, performs worse than the control. However, it can be argued that the group size of participants is not of sufficient size to gather significant insights. Therefore, the research suggests that modification of speech characteristics to include emotional non-verbal communication improves novelty,

intuitive use and stimulation relative to the control variation.

Overall, the modification of speech prosody based on the embedded emotion in a message and including emojis in vocalisation has been suggested successful. The prototype and experiment have given insights in the positive *and* negative effects of speech-based IM. However, in general, the test variations have suggested to increase the experience of emotion, thus suggesting the influence of TTS on the experience of emotion in messages.

1 Dead giveways were the emojis P4 Sometimes the emojis don’t correspond to the message P4 Context was also important but had to assume P4 The voice was very formal, especially longer messages P4 Should teach friends to use emojis for them to recognise emotions P4 It’s not always needed to include emojis P4 Differences with how they would read themselves P6 Very monotoneus P6 No interpunction present P6 - feels like one long sentence Have to be aware of the person to know what it’s about P6 Abbr. and content made recognise emotions P6 Request for better flow, more nuanced tone, adjust tone based on emoji "Who wants to be screamed at when someone is angry?" P6 Avoid cutoff P6 Underwhelming, not conversing very well P8 Robotic voice P6 Messages without emojis would be better, emojis don’t add value, they distract because imagine P8 Recognise with context P8 Train voices to recognise context with emojis figuure out using emojis P8 Robotic voice, weird names P8 Would use it only to see if a message is important, would read the message on phone to see eomij, knows how to read persons emotion, after the fact P9 Know no background P9 Emoji has background function P9 It’s better to hear messages when it’s impossible ot look at screen P9 Context/content is the most important P9 Add !? internation, add, would make recognition easier. Very minimal emotion in messages P5 tone made it hard to identify emotions P5 Desription of emojis is distracting, thinking in head which is which, imagining, emoji was hard P5

IN GENERAL: Use a different vocalisation service.

Clearly feel emotions, but feel like the options for rating were too limited, no option for frustration, impressed with how it reads emotions, not expected it to be this good. P11 Recognised emotion using tone, pauses and breaks P11 Should speak slower sometimes P11 Pitch, middle / end of the sentence could go more up in case of question/statement P11 Content, sometimes more important P11 Irony is not always portrayed well P11 Context resonated with the way "it" was saying the message, internation was good P15 Context is improtant P15 Internation P15 Slow messages are more human-like P15 more internation would be even more P15 Wouldn’t change anything to the voice, it is already quite human-like P15 Would like it to be read by the voice of the people P16 Read with the voice of people in head: voice takes it of the tone P16 It feels weird to not be in charge of internation P16 The service pushes emotion on the message, not in control P16. Not enough inflections P16 Lose of skill of conversation and conflicts P16 Neutral messages would be better to understand the messages and avoid confilt P16 AI should be able to understand the emotion P16 It understands fear, machines taking control P16 elivates burden of texting P16 transforms text in a less active activity - no need to take out phone P16 Good to have a break P16 Tone of voice, difference shows emotion P16 Inflection, neutral, joy didn’t sound flat P16

Messages with emojis

Laughing at voice P1 Laughing at abbrevations P1 after a while normal P1 easy to understand P1 knowing the person helps with context P1 emotion words help recognise (hate, love are

strong) P1 emotion recognition not from voice P2 length of the text emotion P2 longer messages sound more natural P2 = better recognition Context is the most important in recognising emotoin it's hard to hear emotion with flat voice P3 lack of internation needs to process to interpertation P3 words like totally, completely, descriptive words pauses are too long P3 abbrevations LOL, conveys emotion P7 "the voice puts you off" P7 "context made the desicion almost 100 percent of the time" P7 Tone of voice should be less monotoneus/less formal P7 Flow is bad P10 particulary bad at displaying fear P10 Simple messages need context P10 Voice sounds angry P10 Powerful words P10

Variation 2 - in the version with a neutral voice, where messages contain emojis,

VERSIE 3

VERSIE 4 Clear, emojis and context help to recognise emotion P13 Voice is sometimes too fast P13 The break between emojis and text is sometimes too short P13 Concern in voice P14 In some cases voice sounds bland and automated, specific fear, ranting, concern P14 Emojis help in certain messsages where the context is not clear P14 Written reading is easier because you know what=what P14 Long messages are very fast but better to understand, concern P14 Smaller messages you know what emotion is present from experience P14 Short messages sound more robotic P14 Throughout the voice sounded artificial P14 Difficult to understand emotion, tone difference, the product, prefer to think myself P17 Content P17 Voice don't portray emotion P17 Use a different vocalisation service P17 Distracting, quicker messages are harder to listen to, slower pace p17 Device cannot understand emotion P17 Monotoneus P17 Could use the voice of whoever texted you P17 Emoji was the main recognition, helped to put otgether the whole message P18 50/50 content / emoji P18 Emotion words P18 pitch, speed helped P18 More emotion in voice then heard before P18

6.4.1 Conclusion

6.4.2 Modification of prosody does (not) increase experience of emotion

Test 1-5

—- some voices sound weird emojis add strength speak out loud !!! ... emotional voice was already good

7. DISCUSSION

This master's thesis explored the user experience of instant messaging, emotion recognition software and the user experience of spoken instant messaging, to research the influence of non-verbal communication in speech-based IM emotion recognition. The research steps resulted in a high-fidelity prototype which allows users to receive IM through speech, with embedded non-verbal emotion using emojis and voice prosody modifications. In this discussion chapter, the main findings are discussed further. As the findings per research aim have been discussed in 4.2 and 6.3, and respectively concluded in 4.3, 5.4 and 6.4, this chapter will be more concise and discuss the main insights.

7.1 Summary of main findings

This study generated knowledge and insights in the use of personal communication using messaging apps. Instant messaging has become omnipresent and an essential part of communication, deeply embedded in to most of our daily lives. Emotion is an integral part of this digital communication, especially in cases where the communication is personal, in line with research by Kalman et al., 2010. Non-verbal communication is present in IM in various ways, for example in the form of emojis, oralisation of written text manipulation of grammatical markers, in line with research of Walther, 2011. Furthermore, this research uncovered that people find surprising ways to communicate emotion, for example by *ghosting* the other party: introducing long silences to convey negative emotions. Emojis are used to convey emotion and affirm emotion in digital communication, and provides hints in the practice of conversational turn-taking. Finally, the semi-structured interview brought attention to the insight of emotion words, which were deemed more important in the exchange of emotion than previously expected.

Emotion recognition software quality has improved significantly over the past years, and various models are increasingly efficient in labelling emotion. The amount of research, open source projects and information available is impressive. However: EmoBERTa, the model used for an emotion recognition prototype, is not deemed reliable enough to provide data for the manipulation of speech characteristics. Too often, the model is erroneous in its labelling, where it would modify the non-verbal speech characteristics incorrectly when used for SSML. The message dataset used was a good fit for the research, as it has real-life messages of users fitting the target group of the research.

In the final aim, aforementioned dataset was used to generate audio files of vocalised incoming IM, using a prototype. In four separate versions, the non-verbal characteristics of these synthesised messages were modified to portray non-verbal communication. This was achieved either through the modification of the voice prosody to mimic human non-verbal speech characteristics as researched by D. Sauter et al. (2010) or by including spoken emojis to the message, as well as a combination of both methods. Overall, the aforementioned variations showed promising results regarding improvement of emotion recognition and the intensity of emotion experienced, which increased in comparison to a control variation.

7.1.1 Thoughts on the user experience of instant messaging

How is the personal experience of users of instant messaging applications in the context of emotion; how is emotion used and recognised?

The first aim of this thesis aimed to research the baseline as to what extent emotion is used and experienced in IM, from the perspective of research of research participants, using an exploratory interview. The aim was not to uncover new insights, but rather to deepen knowledge of emotion and non-verbal communication in IM to provide a solid base for the continuation of the research.

Emotional usage of IM: The findings of the interviews suggest that IM use is widespread and has become an integral part of daily life. There are shortcomings of IM that hinder the full potential of such services. Even though it is mostly accepted that full conversations can be had through such applications, like in real life, IM can feel cold and distant. This may result in dull conversations, due to the lack of non-verbal communication such as body language, gestures in the form of adaptors and illustrators, posture, head movements, eye contact and more as described in 2.1.2 and researched by Andersen in 1999.

Aspects of online communication are inherently different to in-person communication, and these are boundaries are difficult to lower, albeit not impossible. These limitations may also contribute to misunderstandings in IM. It may be of benefit to give users methods to make online communication more personal and personalised. Functionality focused on personalisation should focus on the inclusion of non-verbal communication not yet available on the digital platforms. A great example are software solutions such as Bitmoji [1] or Memoji [2] which promote the use of facial expressions. Furthermore, improvement of stickers and gifs would allow for expression of emotion in an increasingly airy manner.

Use of spatial arrays and emojis: Confirming the research of Lo (2008), the research suggests emojis and spatial arrays convey and affirm emotion in IM. However, participants did not comment on the lack of range and nuance as described by Derks et al. (2008) and Dresner and

[1] https://www.bitmoji.com/

[2] https://support.apple.com/en-us/HT208986

Herring (2010), which may be due to the fact that their insights applied focus to spatial arrays consisting of characters (i.e. *:D*) and emoticons, where emojis are a *next version* with more detail and choice. As Kalman and Gergle (2009) published, the use of emoticons and emojis in CMC, and thus IM, mostly supports existing written non-verbal or emotional communication, which was affirmed by the research participants and thus suggested by the research results.

Personal communication: An existing connection between parties on IM is suggested important, as the participants note the receiving party profits from knowledge of and personality traits of the sending party. It may be beneficial to keep this in consideration when creating similar research, as interview participants regard this a pillar of emotional communication through IM, in line with the research of Riordan and Trichtinger (2017).

In the light of using voice-based IM to communicate, it is considered important to consider to give the user flexibility of IM platform. Research participant note using multiple services to communicate, limiting said vocalisation software to one service may create hurdles in adaptation. For example; it may confusing for users to be using different vocalisation services in different apps. However, current implementations handles this using the system-integrated assistants.

7.1.2 Thoughts on emotion recognition software

How and to what extend can software classify emotion in instant messages; can the results be applied in a prototype setting to modify speech synthesis to portray emotion in vocalised messages?

Instant messaging dataset: From a technical standpoint, the curated instant messaging datasets, and especially the dataset based on EmotionPush (Shmueli and Ku, 2019) allowed for an efficient evaluation of EmoBERTa. Being able to feed the model two different variations of messages, and experiencing similar results, allowed for a confident decision to not apply the use of the model as an emotion detection source for the final experiment.

Emotion recognition software and accuracy: The study demonstrates that emotion recognition is a powerful tool to analyse and process text, and thus IM. The technology has grown progressed significantly over the years. Depending on the context and purpose, software can perform outstandingly in regards to the recognition of emotion and sentiment. In cases, this may require significant computing power to guard accuracy, which was not available during this research.

A vast amount of closed- and open-source detection software is available on the net. For this thesis, EmoBERTa by Kim and Vossen (2021) was a confident choice and a solid fit for the research, especially taking in to account time, documentation and implementation. The knowledge developed aided to implement a prototype able of processing IM, generating vocalised voice messages which will be handled in the following subsection.

The language model processed the dataset rather quickly. However, the detection efficiency of the model was underwhelming, with correct detection on average only 66,4%, or 50,7% when excluding false-neutral detection. Especially in regards to emotion in speech-based instant messaging, it is important that the results are as reliable as possible. As modification of voice characteristics is suggested to influence the emotion experience of users, errors in the detection may alter the user experience of emotions. For instance, when a *sad* message is tagged as *angry*, and prosody modifications are applied based on the tag, the receiving party may be inclined towards wrong emotion recognition. In other cases where the model is of an assistive use, it may be rather useful, such as on a moderation platform or in customer service.

Aside from the processing of text, it may be beneficial to analyse the intent of included emojis in text. As discussed in 2.2.1, it is assumed that different generations of users utilise CMC and emojis in different and increasingly abstract manners. This may be a hurdle for correct detection, though it can be expected that a quality detection model can overcome this hurdle.

7.1.3 Thoughts on emotion in speech-based instant messaging

How is the user experience of emotion-enabled instant messages experienced through text-to-speech and how can the emotion experience be improved by including non-verbal communication traits?

Application of SSML: SSML is found to be a satisfactory method to modify the prosody of the computer generated voice, to convey the emotion present in an IM, based on the categorisation of positive and negative emotional vocalisation characteristics by D. Sauter et al. (2010). Whether determined by emotion recognition, or in the case of this test by the emotion tagging of the dataset, the speed, pitch and volume of the voice can be modified to convey non-verbal vocalisation characteristics. The vocalisation could be improved by including more modifications to the voice based on the aforementioned research, such as applying modification to the amplification onset and RMS. Another possibility would be to alter the intensity of pitch variations based on the strength of the intended emotion in the message.

Quality of the synthesised voice: However, research participants commented on the robotic characteristics of the voice, especially in the control group which used a neutral voice. In the case of voice with modified prosody, comments tended more towards appreciation for the quality of the voice. It may be beneficial to use more advanced language models to more closely resemble a human voice, or even using a model trained on the voice of the actual sender to generate the incoming voice messages.

Even though all research participants were aware of VPAs and their functionality, only a small percentage of them used them on a daily basis. This implies that good text-to-speech IM has a big market share to tap in to, where arguably the quality of the service can be a deciding factor on the adaption of such service.

Addition of emojis and modification of prosody: The main results of the final experiment suggest that the addition of emojis as well as the modification of prosody to include non-verbal emotional communication aid in the recognition of emotion and emotion intensity by research participants. The findings suggest that, especially with both aforementioned modifications combined, emotion recognition by participants show an increase of correct emotion recognition and experienced emotion intensity. This implies that, regarding the use and inclusion of emojis, the findings of the research by Lo (2008) may be applicable to voice interfaces.

In some cases, participants their emotion recognition and experience of emotion intensity using the combined methods results in average or lower than average recognition. However, it is important to acknowledge that the sample size of research participants is considered too low to make solid conclusions on the data.

Psychological influences of vocalised IM: Vocalised IM may affect users in other ways than expected. Several research participants note that, in varying intensity, the vocalisation *takes away* from their ability to read messages on their own terms. They voice that they themselves read the message in the voice of the sender, or in their own voice. Having the message read to them, takes away control over this interpretation. Similarly, as a participant noted, the vocalisation which includes emotion characteristics, may contribute to diminishing the user their ability to learn or analyse emotion in messages, as the service provides this information for them.

Limitations of the experiment: Similarly, it is important to acknowledge that the addition of emojis to the messages may alter the message, affirming the emotion. Even though the emojis have been added based on the labelled emotion, they are an addition to the original or core of the message. It may be beneficial to examine the influence of the emojis on emotion recognition on a deeper level, for example by using messages with originally include emojis and examining them before and after removal of said visual.

In the case of the performed experiment, one of its limitations is the absence of examination of the IM on the basis of a screen. Including test variations where the messages including and excluding emojis would be examined through a smartphone or other device which does not utilise a speech interface, would have provided with more data to compare the influence of text-to-speech conversion on the experience of emotion.

8. CONCLUSION

This master's thesis aimed to explore the influence and effects of inclusion of non-verbal communication on emotion recognition when receiving speech-based IM. This process included a semi-structured interview on the use of IM to deepen the user experience knowledge on the topic, the creation of a high fidelity software-based emotion recognition prototype, which in turn uses determined emotion to generate audio files of incoming messages. To examine the user experience and recognition of emotion in speech-based IM, as well as to find methods to improve said recognition, aforementioned audio files were modified to include non-verbal communication traits using emojis and modification of voice prosody to reflect human paralanguage and non-verbal communication. A dataset of IM was used to generate four variations of each 140 audio files and run an experiment with 20 research participants to evaluate the effects and user experience of non-verbal communication in spoken IM.

8.1 Summary

The thesis suggests that emotion is experienced in IM, both through traditional screen-based interfaces, as well as voice-based interfaces such as Virtual Personal Assistants. Without modifications to messages, research participants experience emotions in messages that have been converted from text to speech. When applying modifications to the prosody of the voice of said assistant, to reflect emotion and non-verbal communication embedded in messages, it is suggested that in general emotion can be more accurately recognised by the participants, as well as being more intensely experienced. When modifying messages and speech to include spoken aloud spatial arrays and emojis, it is also suggested that, on average, emotion recognition and experienced emotion intensity increase. Combining these aforementioned modifications, on average, emotion recognition and experienced intensity increases even more. Furthermore, the research findings suggest that such modifications in general have a positive effect on the user experience of the product, as measured with the Modular User Experience Questionnaire. Interview insights similarly suggest that participants positively experience the increase of non-verbal communication in voice-based instant messaging.

Emotion recognition software is poised to play a role in the modification of speech characteristic modifications in voice-based instant messaging. With a prototype, it is proven possible to determine emotion to some degree of accuracy. For it to be effective and useful in application, it is

suggested that detection efficiency should be efficient, as the study suggests aforementioned prosody modifications influence the user their emotion recognition ability.

8.2 Study limitations and future work

This study has its limitations. Based on the study performed in this master's thesis, it is not possible to make significant suggestions for the modification of speech to convey emotion, neither did it provide significant statistics on experiment results. However, this was intentional, as the goal was to generate quantitative insights and a starting point for future work. This limitation is foremost due to the small amount of research participants. Additionally, the research focus group did not include a wide array of generations and mainly focused on Millennials, with the final test overflowing into Generation Z slightly. Most of the research participants did not have experience with voice interfaces, and none had experience with voice based instant messaging, which was beneficial, as the participants could generate clear feedback, not being influenced by previous experiences.

Even though the semi-structured interview in aim 4.2.3 was fruitful and provided great insights, it could have focused more on voice-based instant messaging, for example by including recordings of current implementations. This would have given a deeper insight for the variations of aim 6.

The emotion detection proof-of-concept generated good insight in the state of the technology and possibilities. However, it did not yield new discoveries, as it is based on existing research, and the knowledge of such systems finally did not contribute to the final experiment or outcome of the thesis. Even though the result was fruitful, the research time could have been more focused on text-to-speech services.

The EmotionPush dataset used was of high quality and a good fit for this research. However, in the context of the experiment, it would have been beneficial to use a message source closer to research participants. This would result in a deeper understanding of emotion in IM, as research participants did not have a personal connection to the messages in the dataset used. Therefore, it may be beneficial to use participants their personal messages, if possible from a privacy and moral perspective.

The final experiment would greatly benefit from a control variation presented through traditional screen-based interfaces such as a smartphone. When included, it would make it feasible to compare the emotion experience of screen-based and voice-based interactions. Furthermore, the data gathered in this thesis would benefit from a more in-depth data analysis to generate more statistical insights.

For future work, it would be required to set a more tightly scoped research goal, to evaluate a more precise aspect of emotion in voice-based instant messaging. More work is necessary

to develop insights to the effects of text-to-speech conversion of instant messages on the experience of emotion. Additionally, more research is needed on the effect of the voice on the experience of instant messages. Various vocalisation options should be explored, such as using a more realistic voice model, a voice model based on the voice of the sending party (as if spoken by a friend), as well as a voice model based on the voice of the receiving party (as if spoken by oneself). Finally, research should evaluate whether modification of voice characteristics poises a positive or negative effect for users, also on a longer term.

8.3 Reflecting on the learning experience

The process of the research has required perseverance, overcoming hurdles and learning new skills to complete. In hindsight, the scope of the thesis was way too wide, which resulted in a very lengthy and difficult process. At times it was hard to keep the target clear and in sight, especially during the years of Covid, as well as when being preoccupied with work. Finding a research gap did not poise a problem due to interest in the matter, however, the construction of a concise and focused research objective was. Following the research path mainly independently forced to make independent decisions regarding important aspects of background research, building upon gathered and created knowledge during the defined research aims and creating a research structure to convey the knowledge. This resulted in a great learning experience.

The scheduling of the master's thesis has been proven an point of struggle, it being too big to oversee at times, as well as not having clear deadlines or goals. Especially when other projects, such as at work, seem more motivating, this thesis at times received little attention. However, over the years and the last half year, it has received much attention, certainly more than the 1080 hours appointed to it.

On a positive note, I am still content with the subject and the research path ambled, the skills learned as well as the knowledge developed. It is already aiding in my professional career, offering career opportunities, solutions to problems of existing projects, and learning opportunities in the future. In the end, I am very happy with the results of the thesis and to conclude the process.

REFERENCES

Acheampong, F., Wenyu, C., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, *2*. https://doi.org/10.1002/eng2.12189

Adrianson, L. (2001). Gender and computer-mediated communication: Group processes in problem solving. *Computers in Human Behavior*, *17*, 71–94. https://doi.org/10.1016/S0747-5632(00)00033-9

Akmajian, A. (1995). *Linguistics : An introduction to language and communication*. MIT Press.

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 579–586. https://doi.org/10.3115/1220575.1220648

Andersen, P. (1999). *Nonverbal communication: Forms and functions. mountain view*. CA: Mayfield Publishing Co.

Bachorowski, J.-A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science*, *8*(2), 53–57. https://doi.org/10.1111/1467-8721.00013

BAILLET, J., CROUTTE, P., & PRIEUR, V. (2019). Baromètre du numérique 2019 (Crédoc, Ed.) [Accessed: 05-05-2021]. https://www.credoc.fr/publications/barometre-du-numerique-2019

Banse, R., & Scherer, K. (1996). Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, *70*, 614–36. https://doi.org/10.1037/0022-3514.70.3.614

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior research methods*, *40*, 531–9. https://doi.org/10.3758/BRM.40.2.531

Bendel, O. (2017). Sex robots from the perspective of machine ethics, 17–26. https://doi.org/10.1007/978-3-319-57738-8_2

Benyon, D. (2014). Designing interactive systems: A comprehensive guide to hci, ux and interaction design.

Blandford, A. (2013). Semi-structured qualitative studies.

Buller, D. B., & Burgoon, J. K. (1986). The effects of vocalics and nonverbal sensitivity on compliance a replication and extension. *Human Communication Research*, *13*(1), 126–144. https://doi.org/https://doi.org/10.1111/j.1468-2958.1986.tb00098.x

Canales, L., & Martınez-Barco, P. (2014). Emotion detection from text: A survey. *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, 37–43. https://doi.org/10.3115/v1/W14-6905

Carey, J. (1980). Paralanguage in computer mediated communication. https://doi.org/10.3115/981436.981458

Carr, N. (2014). *De glazen kooi: Wat automatisering met ons doet (the glass cage: Where automation is taking us)*. Maven Publishing. https://books.google.fi/books?id=pFzABQAAQBAJ

Clement, J. (2020). Most used social media 2020 [Accessed: 06-10-2020]. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Corps, D. (2014). Top 100 baby names in england and wales: Historical data. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/datasets/babynamesenglandandwalestop100babynameshistoricaldata

Culnan, M. (1987). Information technologies. *Handbook of organizational communication: An interdisciplinary perspective*.

Darwin, C. (1873). *The expression of the emotions in man and animals*. D. Appleton. https://books.google.fi/books?id=4jp9AAAAMAAJ

Demato, J. (2021). Sending smiley emojis? they now mean different things to different people. https://www.wsj.com/articles/sending-a-smiley-face-make-sure-you-know-what-youre-saying-11628522840?st=zf0h0fj3cbq3b7p&reflink=desktopwebshare_linkedin

Derks, D., Fischer, A., & Bos, A. (2008). The role of emotion in computer-mediated communication: A review. *Computers in Human Behavior*, *24*, 766–785. https://doi.org/10.1016/j.chb.2007.04.004

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*. http://arxiv.org/abs/1810.04805

Dresner, E., & Herring, S. C. (2010). Functions of the nonverbal in cmc: Emoticons and illocutionary force. *Communication Theory*, *20*(3), 249–268. https://doi.org/https://doi.org/10.1111/j.1468-2885.2010.01362.x

Hall, E. T., & Hall, M. R. (1990). *Understanding cultural differences / edward t. hall and mildred reed hall*. Intercultural Press Yarmouth, Me.

Hancock, J., Gee, K., Ciaccio, K., & Lin, J. (2008). I'm sad you're sad: Emotional contagion in cmc. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 295–298. https://doi.org/10.1145/1460563.1460611

Hancock, J., Landrigan, C., & Silver, C. (2007). Expressing emotion in text-based communication, 929–932. https://doi.org/10.1145/1240624.1240764

Hartson, R., & Pyla, P. (2012). *The ux book: Process and guidelines for ensuring a quality user experience*. Elsevier.

Head, Preeti, B., & Sharma, P. (2013). Reaching out: A historical overview of the evolution of non verbal communication dr. preeti bala sharma. *RJELAL*.

Herring, S. (1996). *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives*. John Benjamins Publishing Company. https://books.google.fi/books?id=W3IajVswsK0C

Herring, S. (2004). Computer-mediated discourse analysis: An approach to researching online communities. *Designing for Virtual Communities in the Service of Learning*, 316–338. https://doi.org/10.1017/CBO9780511805080.016

Hofstede, G. (2010). *Cultures and organizations : Software of the mind* (Third edition.). McGraw-Hill.

Holtzblatt, K., Wendell, J. B., & Wood, S. (2004). *Rapid contextual design: A how-to guide to key techniques for user-centered design*. Elsevier.

Huh, J., Yetisgen-Yildiz, M., & Pratt, W. (2013). Text classification for assisting moderators in online health communities [Special Section: Social Media Environments]. *Journal of Biomedical Informatics*, *46*(6), 998–1005. https://doi.org/https://doi.org/10.1016/j.jbi.2013.08.011

Intelligence, B. I. (2018). The messaging apps report [Accessed: 02-10-2020]. https://store.businessinsider.com/products/the-messaging-apps-report

Juslin, P., & Laukka, P. (2002). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion (Washington, D.C.)*, *1*, 381–412. https://doi.org/10.1037//1528-3542.1.4.381

Kalman, Y., & Gergle, D. (2009). Letter and punctuation mark repeats as cues in computer mediated communication.

Kalman, Y., Scissors, L., & Gergle, D. (2010). Chronemic aspects of chat, and their relationship to trust in a virtual team., 46.

Kayan, S., Fussell, S., & Setlock, L. (2006). Cultural differences in the use of instant messaging in asia and north america. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 525–528. https://doi.org/10.1145/1180875.1180956

Kim, T., & Vossen, P. (2021). Emoberta: Speaker-aware emotion recognition in conversation with roberta. *CoRR*, *abs/2108.12009*. https://arxiv.org/abs/2108.12009

Kinsella, B. (2019a). Over 20% of uk households have smart speakers while germany passes 10% and ireland approaches that milestone [Accessed: 06-10-2020]. https://voicebot.ai/2019/10/11/over-20-of-uk-households-have-smart-speakers-while-germany-passes-10-and-ireland-approaches-that-milestone/

Kinsella, B. (2019b). Voice assistant demographic data - young consumers more likely to own smart speakers while over 60 bias toward alexa and siri. https://voicebot.ai/2019/06/21/voice-assistant-demographic-data-young-consumers-more-likely-to-own-smart-speakers-while-over-60-bias-toward-alexa-and-siri

Kinsella, B. (2020). Nearly 90 million u.s. adults have smart speakers, adoption now exceeds one-third of consumers [Accessed: 06-10-2020]. https://voicebot.ai/2020/04/28/nearly-90-million-u-s-adults-have-smart-speakers-adoption-now-exceeds-one-third-of-consumers/

Kinsella, B. (2021). Germany smart speaker adoption closely mirrors u.s. pattern [[Accessed 19-Jun-2023]].

Kristiansen, T., & Coupland, N. (2011). *Standard languages and language standards in a changing europe*. Novus Press. https://books.google.fi/books?id=bFNkuAAACAAJ

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. *USAB 2008*, *5298*, 63–76. https://doi.org/10.1007/978-3-540-89350-9_6

Levinson, S. C., & Holler, J. (2014). The origin of human multi-modal communication. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *369*(1651), 20130302.

Li, H., Rau, P.-L., & Hohmann, A. (2011). The impact of cultural differences on instant messaging communication in china and germany, 75–84. https://doi.org/10.1007/978-3-642-21660-2_9

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017). DailyDialog: A manually labelled multi-turn dialogue dataset. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 986–995. https://aclanthology.org/I17-1099

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, *abs/1907.11692*. http://arxiv.org/abs/1907.11692

Lo, S.-K. (2008). The nonverbal communication functions of emoticons in computer-mediated communication. *Cyberpsychology behavior : the impact of the Internet, multimedia and virtual reality on behavior and society*, *11*, 595–7. https://doi.org/10.1089/cpb.2007.0132

Lucero, A. (2015). Using affinity diagrams to evaluate interactive prototypes. In J. Abascal, S. Barbosa, M. Fetter, T. Gross, P. Palanque, & M. Winckler (Eds.), *Human-computer interaction – interact 2015* (pp. 231–248). Springer International Publishing.

Manganari, E. (2021). Emoji use in computer-mediated communication. *The International Technology Management Review*, *10*, 1. https://doi.org/10.2991/itmr.k.210105.001

McGilchrist, I. (2009). *The master and his emissary : The divided brain and the making of the western world*. Yale University Press.

Mehrabian, A. (1972). *Nonverbal communication*. Aldine Publishing Company. https://books.google.fi/books?id=Xt-YALu9CGwC

Merriam-Webster. (n.d.). Ghosting defenition [Accessed: 23-06-2022]. https://www.merriam-webster.com/dictionary/ghosting

Nielsen, J. (2015). Putting a/b testing in its place. https://www.nngroup.com/articles/putting-ab-testing-in-its-place/

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, 206–213. https://doi.org/10.1145/169059.169166

Petrock, V. (2019). Us voice assistant users 2019. https://www.emarketer.com/content/us-voice-assistant-users-2019

Pirzadeh, A., & Pfaff, M. (2012). Expression of emotion in im. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 199–202. https://doi.org/10.1145/2141512.2141577

Postmes, T., Spears, R., & Lea, M. (2000). The formation of group norms in computer-mediated communication. *Human Communication Research - HUM COMMUN RES*, *26*, 341–371. https://doi.org/10.1093/hcr/26.3.341

Press, O. U. (2022). Communication, n. [Accessed: 05-10-2022]. https://www.oed.com/view/Entry/37309?redirectedFrom=communication

Riordan, M. A., & Trichtinger, L. A. (2017). Overconfidence at the Keyboard: Confidence and Accuracy in Interpreting Affect in E-Mail Exchanges. *Human Communication Research*, *43*(1), 1–24. https://doi.org/10.1111/hcre.12093

Sauter, D., Eisner, F., Calder, A., & Scott, S. (2010). Perceptual cues in nonverbal vocal expressions of emotion. *Quarterly Journal of Experimental Psychology*, *63*, 2251–2272.

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, *107*(6), 2408–2412. https://doi.org/10.1073/pnas.0908239106

Schrepp, M. (2021). Measuring user experience with modular questionnaires. *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 1–6. https://doi.org/10.1109/ICACSIS53237.2021.9631321

Schröder, M. (2004). Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. *Affective Dialogue Systems*, 209–220.

Shmueli, B., & Ku, L.-W. (2019). Socialnlp emotionx 2019 challenge overview: Predicting emotions in spoken dialogues and chats. *arXiv preprint arXiv:1909.07734*.

Spears, R., & Lea, M. (1992). Social influence and the influence of the "social" in computer-mediated communication.

Sproull, L., & Kiesler, S. (1986). Reducing social context cues: The case of electronic mail. *Management Science*, *32*.

Stansberry, K., Anderson, J., & Rainie, L. (2019). Experts optimistic about the next 50 years of digital life. *Pew Research Center*.

Tankovska, H. (2019). Most popular messaging apps [Accessed: 17-12-2020]. https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, *63*(1), 163–173. https://doi.org/https://doi.org/10.1002/asi.21662

Turner, A. (2021). How many people have smartphones worldwide [Accessed: 03-04-2020]. https://www.bankmycell.com/blog/how-many-phones-are-in-the-world#1579705085743-b3697bdb-9a8f

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *Arxiv*. https://arxiv.org/abs/1609.03499

Vandergriff, I. (2010). Humor and play in cmc. *Handbook of Research on Discourse Behavior and Digital Communication: Language Structures and Social Interaction*, *1*, 235–251. https://doi.org/10.4018/978-1-61520-773-2.ch015

Vapnik, V., & Chervonenkis, A. (1974). *Theory of pattern recognition [in russian]* [(German Translation: W. Wapnik & A. Tscherwonenkis, *Theorie der Zeichenerkennung*, Akademie–Verlag, Berlin, 1979)]. Nauka.

Vapnik, V. N. (2000). *The nature of statistical learning theory* (2nd ed). Springer.

virallinen tilasto, S. (2020). Väestön tieto- ja viestintätekniikan käyttö [verkkojulkaisu] (H. Tilastokeskus, Ed.) [Accessed: 05-05-2021]. http://www.stat.fi/til/sutivi/2020/sutivi_2020_2020-11-10_tau_018_fi.html

voor de Statistiek, C. B. (2020). Nederland in cijfers, editie 2020 (C. B. voor de Statistiek, Ed.) [Accessed: 05-05-2021]. https://www.cbs.nl/-/media/_pdf/2020/51/nederland-in-cijfers-2020.pdf

Vos, H., ter Hofte, H., & Poot, H. (2004). Adoption of instant messaging in a knowledge worker organisation, 10 pp. https://doi.org/10.1109/HICSS.2004.1265072

Walker, M., Larson, J., & Hunt, A. (2001). A new w3c markup standard for text-to-speech synthesis. *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, *2*, 965–968 vol.2. https://doi.org/10.1109/ICASSP.2001.941077

Walther, J. (2011). Theories of computer-mediated communication and interpersonal relations. *The Handbook of Interpersonal Communication*, 443–479.

Wang, S.-M., Li, C.-H., Lo, Y.-C., Huang, T.-H. K., & Ku, L.-W. (2016). Sensing emotions in text messages: An application and deployment study of emotionpush. *CoRR*, *abs/1610.04758*. http://arxiv.org/abs/1610.04758

Weiyuan, L., & Hua, X. (2014). Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications*, *41*(4, Part 2), 1742–1749. https://doi.org/https://doi.org/10.1016/j.eswa.2013.08.073

WhatsApp. (2020). Two billion users – connecting the world privately [Accessed: 24-09-2020]. https://blog.whatsapp.com/two-billion-users-connecting-the-world-privately

Yus, F. (2011). *Cyberpragmatics. internet-mediated communication in context*. https://doi.org/10.1075/pbns.213

APPENDIX A: FOCUS GROUP INTERVIEW

A.1 Semi-structured Interview Digital Communication Starting the interview

- Make sure the right participant is present
- Start stopwatch on phone
- Start with introduction
- Computer-mediated communication

First of all, I would like to thank you for participating in my research! Your participation helps me greatly in the research for my master thesis. My name is Jan van Dijk, and I am a Human-Technology Interaction student at Tampere University in Finland.

I expect this interview to last about 20-30 minutes. As communicated before, this interview will be recorded. This includes both sound and video. If you prefer that your video is not recorded, you can turn off your webcam, or request to not use the video camera. However, in the case of a video call, I hope that we can keep the webcam connected to make the conversation go more smoothly and naturally.

Your participation in this interview is completely anonymous, and not linked to any of your personal information. The recording will generally only be seen by me, except when the staff of Tampere University requests to view the recordings to support or verify my research. The recording will be saved in an encrypted and password protected file. Any references or quotations to this interview will be completely anonymous. After the research is finished, the recording will be destroyed.

- Do you verbally accept the recording of this interview?

Secondly, I would like to inform you that you have the right to stop participation in this interview at any time, for any reason. Participation is completely voluntary. Even after the interview has taken place, you can request to cancel your participation. During the interview, there are no right or wrong answers. Anything you say is correct, as I am interested in your views, opinions and experiences. You can ask questions during any part of the conversation.

- At this point, do you have any questions?

The topic for our semi-guided conversation is computer-mediated communication. Computer-mediated communication is basically a catch-all term for all types of communication that happens with the help of computers. For example, sending an email, commenting on a post on facebook or sending a message through WhatsApp. Our conversation will be mainly focussing on instant-messaging apps, such as WhatsApp, Facebook Messenger and Telegram. I'm sure you're familiar with one or more of those apps, but if you're not, please let me know. The instant-messaging apps allow for real-time communication - like chatting as if a conversation is actually happening - and are used widely.

- At this point, do you have any questions?

A.1.1 General Information

- Age
- Gender
- Occupancy / Field of work / Field of study
- Do you use English to communicate on a regular basis?

A.1.2 Conversation about instant messaging apps

- **Which instant messaging apps did you use in the last month?** *Facebook Messenger, WeChat, Line, WhatsApp, Discord, Kik Messenger, Tencent QQ, Signal, Slack, Microsoft Teams, Snapchat, iMessage, KakaoTalk, Skype, Telegram.*
- Which is your favourite?
  - Why?
- Which is your least favourite?
  - Why?
- How do you use those apps?
  - Daily?
- If you think about your day from waking up to going to sleep, at which moments during the day do you use the app?
  - Do you think you use the app a lot?
  - For what kind of situations?
- How important are those apps to you?
  - Are they an essential part of your life?
  - Would you be able to live without them?

- How do you stay in touch with your friends/family/colleagues/other people though IM?
- Are instant messaging apps a good substitute for in-person conversations?
- Can you have full conversations on the app?
  - In other words, can they replace an offline conversation?
- Do you ever handle important matters using instant messaging?
  - What kind of matters?.
  - With everyone? Family? Friends? Colleagues?
- Do you prefer talking about important matters through IM?
- Do you express yourself emotionally, for example, show that you're happy or sad through Instant Messaging apps?
  - How do you express yourself emotionally?
  - How do you make your sentences emotional?
  - Do you use emojis?
- How?
- Think of a happy message you've sent
  - How did you express yourself?
- Think of an angry message you've sent
  - How did you express yourself?
- Think of a sad message you've sent
  - How did you express yourself?
- Think of a disgustful message you've sent
  - How did you express yourself?
- Think of a fearful message you've sent
  - How did you express yourself?
- What is your communication style?
  - Do you 'type as you say' i.e. happppyyyy
  - ...
  - ALLCAPS
  - Emojis?
  - GIFs?

  - Stickers?

- **Experiences instant messaging**
  - What was your worst experience when using instant messaging to communicate
    - Emotionally
    - Formally
  - How was it solved / What happened after?
  - What was your best experience when using instant messaging to communicate
- Emotionally
- Personally/Formally
- How was it solved / What happened after?

Voice Assistants

- Are you aware of the existence of voice assistants? *Apple Siri, Google Assistant, Amazon Alexa, Bixby, XiaoAi*
- Do you use them?
- Do you use voice assistants when you need to be 'hands free', i.e. in the car, whilst cooking?
- For what features?

Nowadays, many voice assistants start to roll-out hands-free instant messaging, for example when driving in the car. When you receive a message, an option will be available to speak-aloud the received message. The assistant will then read your message to you, and give you the opportunity to respond. In many situations, like when driving or cooking, the message is not visible in written text so as to not distract; you'll only be able to hear the message.

- Do you ever let voice assistants read aloud your messages? I.e. in the car?
  - What is your experience?
  - Does it meet up to expectations?
  - Do you think you'll be able to have a full instant messaging conversation through a voice interface?
- Why?
- Do you think you'd be able to recognise emotions in the messages you receive through a voice interface?
  - Why?
  - How?

- What do you think would help you to recognise emotions through a voice interface?
- This is the last question: Regarding the use of voice assistants for instant messaging, is there anything you'd like to add or recommend?

APPENDIX B: INSTANT MESSAGING DATASET

B.1 Instant Messages Dataset

B.1.1 Dataset based on Daily Dialog (Y. Li et al., 2017)

The dataset is available including emojis at https://docs.google.com/spreadsheets/d/1whk-Sf2tzTBFIB6BrkmK
edit

Neutral

1. I'll throw out the garbage
2. Would you mind waiting for a bit?
3. This is how you turn on the computer
4. I bought food from the snack stand near our hotel
5. Good afternoon
6. Can I help you?
7. That's 68 euros altogether
8. The cork seems to be stuck
9. Do you have anything particular in mind
10. Is there a lot of crime in your city?
11. I went shopping without my umbrella yesterday morning
12. Did you hear what happened to Sally?
13. In summer I go hiking, and in winter I ski.
14. Here are the leads from last week's exhibit as a trade show
15. It's going to a volleyball match between the economic department and foreign language department this afternoon
16. Hi, did you catch the game last night?
17. I'm going to England by flight BE987
18. Its gross area is approximately nine hundred square feet. There's one living room, one dining room, one master bedroom and two other bedrooms
19. It will take me a while to put together
20. Could you come and help me mend the computer?

Anger

1. Tell me! Would you?
2. What's wrong with that?
3. No, the steak was recommended, but it is not very fresh
4. No, thank you
5. I hate doing the damn laundry
6. Yes, you should have
7. Why aren't they are aware of the gravity of the situation?
8. Stop picking on me, I'm going to be an-

gry

9. The air quality here horrendous! The pollution levels were so high that we weren't supposed to go outside with a face mask again!
10. We are through
11. I can't stand being with you any more
12. I wasn't done talking to you!
13. You and I have been together for a whole year , and our vacation time should be about the two of us!
14. Oh , how annoying
15. That's crazy!
16. I hate the way he treats us
17. How is that even possible!
18. That just isn't good enough .
19. It just makes me angry
20. That really grinds my gears

Disgust

1. The kitchen stinks
2. Coffee? I don't honestly like that kind of stuff
3. I swear, I'm going to kill you for this.
4. You stink
5. What a mess
6. No, cats are too dirty. They are lazy and cunning. I don't like them at all.
7. This place stinks like rotten eggs
8. Wow your apartment is a mess
9. Don't even start, it was terrible
10. Ir's awful! It's like drinking saltwater
11. The bathroom is like a pig stall
12. Gross! What are you doing to yourself?
13. Nasty. You'd better not rub your eyes.
14. You shouldn't let the dirty dishes accumulate in the sink
15. Many people say it's a good story. But I think it's disgusting!
16. Oh god, not again. It's the same as last week
17. I'm really done with my job in the bank
18. I've had enough of that
19. Oh, it's horrendous
20. The new computer software is driving me crazy

Fear

1. It's like death is in front of me
2. I am afraid your diamond bracelet has gone
3. Yes, it's an emergency! There is a huge fire here
4. I think there is something under my bed!
5. I have two finals this week and one oral presentation and I'm not ready for any of them
6. I'm afraid of the darkness
7. Because it scares me
8. I just watched a movie and I'm scared
9. Damn that's scary
10. Dan Dan Dan! You have to come over to my house right now!
11. You are being watched! Be careful!
12. I don't know what to do, if I can't find the money
13. I'm scared that she might not come back
14. I was scared stiff after my first day

15. I'm terrified, don't know what to do
16. He followed me all the way home
17. Man, I'm scared .
18. Watch out!
19. Of course not. I'm afraid of them.
20. I'm freaking out! You gotta help me!

Joy

1. Because you are amazing
2. Good morning!
3. Heeey, wow how are you? It's been such a long time!
4. This is the good life!
5. It's almost midnight, happy new year!
6. Hi honey, guess what! Julie and Alex are getting married !
7. He told me the tests were negative and there was nothing wrong with me
8. That's very clear . I think I can find my way now . Thank you .
9. How about getting some coffee tonight
10. Thanks a lot, that's the favor I was going to ask you for
11. I like Chinese food
12. Thank you
13. You're welcome!
14. That's amazing, how did you know!
15. Well , thank you very much , I hope I can find it
16. It looks more like a toy, or a cool briefcase
17. Congratulations , Vivian, you won the grand prize
18. Great! I can teach you men a thing or two about shopping!
19. How about taking a walk in the park?
20. You must be pretty excited about your trip!

Sadness

1. I'm fine. I'm just so touched by the sad story
2. I'm really sorry. I'll face the music
3. I'm so sorry about your brother
4. I'm sorry, there's a limit to the speed
5. I'm afraid I'm going to the gym for a workout
6. I had a blood test last week. The doctor said that I have high blood pressure and my blood is thick and sticky. I'm very worried
7. Sorry, but I don't have time for that at the moment
8. My parents are always saying I am not good enough.
9. Sorry. It is the smallest size we have.
10. I'm sorry
11. You're so careless, Billy
12. I'm afraid I forgot to lock the door
13. Sorry, darling, I forgot
14. Yes, I can't sleep well every night.
15. That certainly sounds like a dreary Saturday
16. Tired and stressed. This wedding is giving me a headache
17. I'm sorry, my car's taken
18. He asked me to beef up in the work

19. I've just parted with my boyfriend
20. I fell on the way to school , and your bike got scratched

Surprise

1. I can't believe that Anthony is finally getting married
2. What's wrong with MSG? It helps to bring out the taste of the food
3. No! Are you kidding me?
4. Really? How much is it?
5. What a find! So, how much does it cost?
6. Wow! That's expensive
7. That's absurd
8. Wow ! What's the hold up ?
9. Oh ! What bad luck ! How can I get to the Theater ?
10. Oh, my god. I am so heavy, I gained 10 kilos
11. Oh ? What's that
12. Are you kidding? It's too early for her
13. Poor you, you are so enfeebled
14. What ? I sent you three or four messages !
15. I thought you had an unlimited SMS plan
16. Cancelled! So what am I supposed to do now
17. Dear god, it's five flights up !
18. What!
19. Really? Come on, I'll introduce them to you
20. You know you shouldn't say that at a time like this

B.1.2 Dataset based on EmotionPush (Wang et al., 2016)

Neutral

1. Also most of the work force is Mexican manufacturing
2. But I do regularly wear that shirt
3. Deal. Though if anyone wants particular stuff, they'll need to tell me.
4. Did you send her the link to the spreadsheet?
5. Feel free to ask her, though
6. I usually fly back home cause I'm impatient. It's only like a 7 hour drive tho
7. I'm going to go use the bathroom real fast and then come back upstairs to talk to you and sleep
8. I'm waiting in the lobby for a FedEx package
9. It was not about being ziddi I told you I wanted to talk to you before I slept you said hmmm
10. It's like on Kathy next to David and behind Scobell
11. Maybe, i don't really remember, I was in the car talking to mom
12. Not yet, I'll finalize with her and let you know by tomorrow night. I'll prepare a draft of the sublease and you can sign it when you move in
13. So u can come but u gotta be early
14. So you just left when they kicked u out

right?

15. Sorry was working
16. Umm I have info session in an hour. And need to read a paper.
17. well i was thinking of going to the movies at 20:00?
18. What do you want me to do
19. Yeah i'd like to buy some clothes
20. You won't need to produce any receipts. At least I didn't have to

Anger

1. And we argued about it and I'm just angry
2. hey just did it and DID NOT TELL ME UNTIL EVERYTHING WAS CHANGED
3. i hate that they don't post them on time
4. I have no time
5. I really FUCKING hate this semester bc of this FUCKING schedule
6. I seriously can't stand my dad
7. I'm kinda pissed
8. I'm not even complaining about you and I HAVE been dealing with it
9. Is the same thing over and over again
10. it's so fucking hard to get that first internship
11. looks like shit
12. My mom's being a royal pain rn
13. she has no right to torment your cat
14. that what i'm annoyed about
15. This is impossible
16. wtf this shouldn't even be possible. Zero is supposed to only be spawned
17. Yeah I just need to push everyone to actually get it done. I sent it to everyone but no one messaged me back.
18. You dick
19. You put me in such a position in front of others
20. you should beat him up

Disgust

1. And people smell bad
2. but he's implying that people are only voting for her/supporting her because she's a women
3. Dennis too obsessed with twitter and ig
4. DISGUSTING!!!!!
5. I am done caring about this project
6. I hate my photo
7. I really don't like magicians
8. I'm gonna go puke
9. Is that what he's talking about? Honestly, that's also stupid as fuck.
10. its implying that people aren't already voting by their best intentions
11. kinda disgusting
12. Major waste
13. Oh ewwwww
14. Popplio is literally a clown
15. she is such a biiiiiiitch
16. She's like so ocd and perfectionist about everything except for her health/weight
17. This is cruel
18. very immature way to handle it
19. why is that poster so bad

20. You have to realise I don't want the feel of the stuff on my skin

Fear

1. also my road test is today ahhhh
2. But what if my mom died
3. first things first. how to approach her. do I like just randomly go up and talk or should I be like oh hey would it be cool if I sat here.
4. Go to health services pls
5. Hey are you still there? I'm a little worried about you
6. I almost crashed, front tire lost all the air pressure at once and I was driving moderately fast
7. I don't want to look at it
8. i dont think i did very welll
9. I'm scared
10. I'm so scared!
11. im on a really tight budget and im a bit worried that the additional Ž0ac200 might make me go over
12. Is Melanie OK?
13. Lots of uncertainty ahead
14. oh wait omg i don't want that
15. Sarah still hasn't replied I'm freaking out every time I get an email
16. Scary for one. The world just seems to big right now. In the city center, and feel weird since I can't just walk to everywhere.
17. So I don't get lost lol
18. they just made a big change to the lock system apparently and I'm worried its gonna mess stuff up for a brief period
19. Tomorrow is going to be rough
20. Uh-oh

Joy

1. At least waking them up nicely lol
2. Hey there, I just wanted to say I really enjoyed working with you this year. You were always really great about explaining things, and never condescending. I really appreciated it. Oh and congrats on graduating!
3. I love you
4. i loVE YOU MORE
5. I was like yasss and Emmy helped get free shipping
6. I'm hungry but I'm excited!
7. IT WAS INCREDIBLE
8. it's like I have to work two jobs to compete lol
9. Like I said, it's a little slower, but it's really pretty
10. Like we just got a voucher that we can exchange for a ticket the day of so if you buy we can sit together lol
11. Like, it's my coat rack next to the front door hahah A
12. LOL but also nice nice it's good to hang with your parents before they set you off into the wind
13. Lol congrats
14. Lol he still sounds too excited
15. Omg were watching how to train your dragon 2 at lunch

16. Pretty glad I'm done with this job after this quarter
17. That is pure stream of consciousness hahahahahahahahah A
18. That's sweet, excited to see that at the end
19. Yeah it's really nice! Been just 2 days since I landed here. So I am still trying to figure out things and exploring!
20. Yeah. i was leaning towards that one bc it's the bigger deal hahaha

Sadness

1. Guess we're not seeing each other </3
2. I don't remember any dreams
3. I feel bad though haha I wanna help.
4. I haven't been very productive recently
5. I mean I have no reason to expect anything else
6. I miss you
7. I still don't have a job
8. i think i'll actually have to stay here for the summer
9. I want to be with you
10. i want to come but flights are like 700Ž0ac+ from here
11. I'm running on four-ish hours of sleep
12. I'm sorry you have to deal with all of this
13. Its not a fun week for me
14. Lol I've just been having a rough time since Maria and I broke up but that's it
15. Sorry about last night
16. sorry about that
17. sorry about that Mathew
18. this work is never ending sigh
19. Yes it is but no one ever responds do it probably sucks
20. You were supposed to kiss me

Surprise

1. Apparently I fell asleep in a chair in my friend's room and then I went to sleep in his bed and then I walked to my room and went to sleep in my bed
2. Are you okay?!!
3. Did you already get nougat wtf
4. Dude I still can't believe that happened
5. Holy shit that's a nightmare for employees hahahahA
6. Holy shit the first song is metal
7. I thought he graduated
8. i thought you weren't going to date Sasia!
9. Lukas GOT AN A?!?!? SO SMART!?!?!?!
10. Oh cool... and good good. Are we not supposed to know that it's not graded?
11. oh my god
12. Oh! Dope. What's there?
13. Ok! Oh I thought I paid for water on June?
14. omg there are people frolicking in the cut pond
15. So ur having dinner with Emelie?
16. That's it?
17. The person was saying Ž0ac2350 for the entire summer!
18. wait did i put i was available?!
19. what do you mean, it's great!
20. you caught me by surprise

APPENDIX C: PROCESSING SCRIPTS

```
import argparse
import logging
import json
import jsonpickle
import requests
import random

logging.basicConfig(
    level=logging.INFO,
    format="%(asctime)s.%(msecs)03d %(levelname)s %(module)s - %(
                                        funcName)s: %(message)s",
    datefmt="%Y-%m-%d %H:%M:%S",
)

def synthesize_text(text, emotion):
    """ Synthesizes speech from the input string of text. """
    """ SOURCE: https://cloud.google.com/text-to-speech/docs/create-
                                        audio#text-to-speech-text-python
                                        , edits made by Jan """
    from google.cloud import texttospeech
    from google.oauth2 import service_account

    credentials = service_account.Credentials.from_service_account_file(
                                        'creds.json')
    client = texttospeech.TextToSpeechClient(credentials=credentials)
    names = ["Thomas", "James", "Jack", "Daniel", "Mathew", "Ryan", "
                                        Joshua", "Luka", "Samuel", "
                                        Jordan", "Rebecca", "Lauren", "
                                        Jessica", "Charlotte", "Hannah",
                                         "Sophie", "Amy", "Emily", "
                                        Laura", "Emma"]

    # Note: the voice can also be specified by name.
    # Names of voices can be retrieved with client.list_voices().
    audio_config = texttospeech.AudioConfig(
        audio_encoding=texttospeech.AudioEncoding.MP3,
    )
    voice = texttospeech.VoiceSelectionParams(
        language_code="en-US",
```

```
        name="en-US-Neural2-C",
        ssml_gender=texttospeech.SsmlVoiceGender.FEMALE,
    )

    if emotion == "anger":
        ssml = '<speak>Message from ' + random.choice(names) + ':<break
                                        time="400ms"/> <prosody
                                        volume="+0.864dB" rate="91%"
                                         pitch="-46%">' + text + '</
                                        prosody></speak>'

    elif emotion == "sadness":
        ssml = '<speak>Message from ' + random.choice(names) + ':<break
                                        time="400ms"/> <prosody
                                        volume="-5.38dB" rate="85%"
                                        pitch="-17%">' + text + '</
                                        prosody></speak>'

    elif emotion == "surprise":
        ssml = '<speak>Message from ' + random.choice(names) + ':<break
                                        time="400ms"/> <prosody
                                        volume="-3.10dB" rate="143%"
                                         pitch="59%">' + text + '</
                                        prosody></speak>'

    elif emotion == "fear":
        ssml = '<speak>Message from ' + random.choice(names) + ':<break
                                        time="400ms"/> <prosody
                                        volume="+0.607dB" rate="110%
                                        " pitch="+24%">' + text + '
                                        </prosody></speak>'

    elif emotion == "joy":
        ssml = '<speak>Message from ' + random.choice(names) + ':<break
                                        time="400ms"/> <prosody
                                        volume="+0.294dB" rate="104%
                                        " pitch="+42%">' + text + '
                                        </prosody></speak>'

    elif emotion == "disgust":
        ssml = '<speak>Message from ' + random.choice(names) + ':<break
                                        time="400ms"/> <prosody
                                        volume="-0.655dB" rate="135%
                                        " pitch="-18%">' + text + '
                                        </prosody></speak>'

    else: #Default, just use the defaut/neutral voice
```

```
        ssml = '<speak>Message from ' + random.choice(names) + ':<break
                                            time="400ms"/>' + text + '</
                                            speak>'

    input_text = texttospeech.SynthesisInput(ssml=ssml)

    response = client.synthesize_speech(
        request={"input": input_text, "voice": voice, "audio_config":
                                            audio_config}
    )

    # The response's audio_content is binary.
    with open("generated_audio/" + emotion + text + "_test" + ".mp3", "
                                        wb") as out:
        out.write(response.audio_content)
        print('Audio content written to file "output.mp3"')

def process_file() -> None:
    # Load dataset
    dataset_filename = 'messagesdd' #omit .json, it is added later.
    input_file = open (dataset_filename + '.json')
    message_dataset = json.load(input_file)

    processed_messages = []

    for message in message_dataset:
        # Prepare and send the message to detection instance
        data = {"text": message["message"]}
        response = requests.post("http://127.0.0.1:10006/", json=
                                            jsonpickle.encode(data))

        # Create a decorded dict ordered by detection percentage (higest
                                             first)
        detected_results = jsonpickle.decode(response.text)
        detected_results = dict(reversed(sorted(detected_results.items()
                                            , key=lambda item: item[1]))
                                            )


        # Format data to append to results array
        result = {
            'message': message["message"],
            'actual_emotion': message["emotion"],
            'detected_emotion': list(detected_results.keys())[0],
            'detected_emotion_confidence': list(detected_results.values
                                                ())[0],
            'all_detected_emotion': detected_results
        }
        processed_messages.append(result)
```

```
        synthesize_text(message["message"], message["emotion"])

    json_string = json.dumps(processed_messages)
    with open('messagesdd_file_processed.json', 'w') as outfile:
        outfile.write(json_string)


if __name__ == "__main__":
    print("Stating data processing, please wait.\n")
    process_file()
```

Full codebase can be found at
https://drive.google.com/drive/folders/1Uh6sbceiMjzX3ujQuy1TtepeJFg1qpoR

APPENDIX D: FINAL TEST RESULTS

D.1 Final test results

D.1.1 Recognition of emotion in spoken messages

	V1	V2	V3	V4	Avg.
anger	60%	70%	60%	60%	62,5%
disgust	50%	50%	50%	70%	55%
fear	50%	60%	60%	80%	62,5%
joy	70%	70%	60%	80%	70%
neutral	50%	60%	80%	70%	65%
sadness	60%	50%	70%	80%	65%
surprise	40%	60%	50%	70%	55%

Table D.1. *Percentage of emotion in messages correctly recognised by participants, ordered by variation and emotion, only test part 1. Data of chart at 6.2.*

	V1	V2	V3	V4	Avg.
anger	55%	65%	60%	75%	63,8%
disgust	45%	70%	60%	65%	60,0%
fear	45%	65%	55%	75%	60,0%
joy	75%	65%	75%	90%	76,3%
neutral	50%	60%	80%	70%	65,0%
sadness	65%	65%	75%	80%	71,3%
surprise	55%	70%	70%	75%	67,5%

Table D.2. *Percentage of emotion in messages correctly recognised by participants, ordered by variation and emotion, test part 1 and 2 combined. Data of chart at 6.1.*

D.1.2 Emotion intensity ratings

	V1	V2	V3	V4	Avg
anger	4,4	5,3	4,6	4,7	4,75
disgust	4	4,8	3,7	5	4,375
fear	3,5	5,1	3,3	4,6	4,125
joy	4,6	4,1	3,9	5,2	4,45
neutral	3,2	3,9	4	3,3	3,6
sadness	3,5	4,2	3,4	5	4,025
surprise	3,4	4	4,9	5,1	4,35

Table D.3. *Average emotion intensity rating on a 1-7 Likert scale, ordered by test variation and emotion, test part 1 only. Data of chart at 6.3.*

	V1	V2	V3	V4	Avg
anger	4,2	5,1	4,6	4,65	4,6375
disgust	3,5	4,65	4,25	4,95	4,3375
fear	3,85	4,4	3,45	4,8	4,125
joy	4,35	4,1	4,5	5	4,4875
neutral	3,2	3,9	4	3,3	3,6
sadness	3,6	3,7	4,05	4,8	4,0375
surprise	3,8	4,05	5,15	4,85	4,4625

Table D.4. *Average emotion intensity rating on a 1-7 Likert scale, ordered by test variation and emotion, test part 1 and 2 combined. Data of chart at 6.4.*

D.1.3 Raw participant data

#	Message
1	And we argued about it and I'm just angry
2	hey just did it and DID NOT TELL ME UNTIL EVERYTHING WAS CHANGED
3	i hate that they don't post them on time
4	I have no time
5	I really FUCKING hate this semester bc of this FUCKING schedule
6	I seriously can't stand my dad
7	I'm kinda pissed
8	I'm not even complaining about you and I HAVE been dealing with it
9	Is the same thing over and over again
10	it's so fucking hard to get that first internship
11	looks like shit
12	My mom's being a royal pain rn
13	she has no right to torment your cat
14	that what i'm annoyed about
15	This is impossible
16	wtf this shouldn't even be possible. Zero is supposed to only be spawned
17	Yeah I just need to push everyone to actually get it done. I sent it to everyone but no one messaged m
18	You dick
19	You put me in such a position in front of others
20	you should beat him up
21	And people smell bad
22	but he's implying that people are only voting for her/supporting her because she's a women
23	Dennis too obsessed with twitter and ig
24	DISGUSTING!!!!!
25	I am done caring about this project
26	I hate my photo
27	I really don't like magicians
28	I'm gonna go puke
29	Is that what he's talking about? Honestly, that's also stupid as fuck.
30	its implying that people aren't already voting by their best intentions
31	kinda disgusting
32	Major waste
33	Oh ewwwww
34	Popplio is literally a clown
35	she is such a biiiiiiitch
36	She's like so ocd and perfectionist about everything except for her health/weight
37	This is cruel
38	very immature way to handle it
39	why is that poster so bad
40	You have to realise I don't want the feel of the stuff on my skin
41	also my road test is today ahhhh
42	But what if my mom died