Between Phones and Video-Chats

A Computer-Mediated Communication Experiment

Bachelor Thesis
Faculty of Science, University of Bern

submitted by

Leonard Felix Hertzberg

from Bern, Switzerland

Supervision:

PD Dr. Kaspar Riesen
Institute of Computer Science (INF)
University of Bern, Switzerland

Abstract

This thesis examines computer-mediated communication within the domain of human-computer interaction. Prior research has addressed the question of how humans modify their speaking behavior in different computer communication settings over a relatively short period of time. Consequently, numerous open questions remain unanswered, largely due to the rapid advancements in software and hardware technologies. Variations in conversation contexts arise from the fact that non-verbal communication is often severely limited in modern communication. Thus, the focus of this thesis is the development of a communication interface that reintroduces facial expression and head-nod channels into the telephone setting, resulting in a hybrid software between voice-calling and video-chatting. Using the "turn-taking" model to conceptualize conversations, differences in key data values are measured, such as the length of speaking turns, silences, and overlaps in speech. Furthermore, the perceived differences by the participants are recorded, and therefore, answering the question: If there is a need for this hybrid communication software. With the help of the deep learning library of "DeepFace," the emotions of the participants are displayed to the conversation partner, and a series of tasks is completed. While generally the results were not statistically significant, some key differences were spotted. Silences between people speaking and overlapping speech were, on average, longer in the normal telephone control setting. The expected change in the length of people's speaking turns was not recorded. The feedback on the software leaves room for improvements, but was also noted as helpful and possibly suited for real-life applications down the line.

Contents

Chapter 1

Introduction

Human-computer interaction is a broad field that combines many different research fields such as computer science, psychology, linguistics, philosophy, engineering, and design. As computers are more and more integrated in every part of human life, this list of research fields contributing to Human-Computer-Interaction is not finished, and we can analyze interactions of humans and computers from the perspective of many more fields. It encapsulates everything that has to do with humans and computers interacting. Its primary focus is on every interface between the two instances. Naturally, since the invention of computer devices such as the telephone or fax, communication is also part of this research field.

A typical human-computer interaction in communication consists of a person who chooses a modality to interact with some sort of computer. These modalities can be visual, haptic, auditory, or even gestures. Every interaction between them also needs an interface. For example, the telephone has buttons, a screen, audio sensors, and small speakers to function. At the end of the interaction, there is usually an interaction partner, such as a neural network or robotics, as well as other humans connected with an interface. These subcategories are broad; in this paper, the focus lies on interpersonal communication between two humans.

Another key distinction in computer-mediated communication is synchronicity, as computers give the freedom of pausing conversations when desired. Email and texting are such asynchronous forms of communication; on the other hand, telephone calls or video-chats are modeled after real-life conversations and appear to be very similar to them. To conceptualize communication, the model of turn-taking as originally proposed by Sacks et. al. [1] is the most common and best documented. In short, people in conversation take turns speaking and listening. There are also forms of non-verbal communication accounted for, which will be the focus of this thesis. The idea is to take the telephone setting, and add these non-verbal communication channels in the form of facial expressions and head-nods.

To achieve this, we need the help of artificial intelligence. It is generally defined as computers that can perform tasks that usually require human intelligence to do. Similarly to humans, computers also need training and learning to do these tasks. As emotion-detection is very much an "intelligent" task, this approach seems promising. Just like humans, computers can detect certain patterns in a face and can learn their meaning. This is generally described as pattern recognition and is a broad branch within AI, where computers learn certain patterns through training, to spot them and their meaning, in real-world applications. And therefore, emotion-detection, which in this thesis is not developed ourselves, but an open-source library is used, which uses deep convolutional neural networks.

The goals of this thesis are then to identify communication differences between the two different forms of telephone conversations with different backchannel feedback introduced. Furthermore, the need and usefulness of this piece of software are being analyzed, and whether it could have real-life applications in the workplace or the school setting. For example, when needing better communication options than just telephone over long distances, but for privacy or personal reasons, video chatting is not preferred. From a scientific standpoint, the differences in turn lengths, turn frequency, silence lengths, and overlap lengths are interesting pieces of information that can help us better our understanding of interpersonal communication with computers.

The thesis is organized in the following manner. In the next chapter, the theory of the model of turn-taking is explained in detail, as well as a quick look at the workings of the emotion-detection. After that, the solution is presented, and how the software and experiments are set up. This led us to the statistical and qualitative results of the experiments, which are interpreted and contextualized in the conclusion chapter.

Chapter 2

Theory

Computer-mediated communication is not technically a new field. As the invention of the telephone goes back to 1889, communication mediated by some sort of computer has existed for over a hundred years now. Despite this, not much research has been designed to analyze the difference in computer-mediated communication with varying modes of communication channels. Although it has been gaining focus since the start of the domination of social media. As software as well as hardware tech are constantly evolving, it's hard for research to keep up with these changing communication settings. Yesterday, the internet was a huge new thing, and when gaining some clarification, the next huge new thing in artificial intelligence communication was presented. However, AI communication is not examined in this paper, as we stick to humans conversing with one another.

Most past research in this field has focused on communication in a setting of learning and teaching, for communication at the workplace, or even in interviews. As these types of communication have a clearer structure and purpose, it is easier to have comparability among different people. In this thesis, only dyadic conversations are analyzed. Although the whole process could also be applied to group conversations, this adds a whole new layer of complexity and is out of the scope of this project. Although it could be an interesting research approach for the future.

Since computer-mediated communication that transports the spoken word is analyzed, the need for a model to conceptualize this arises. In the following sections, turn-taking is explained and used as the base model for communication.

2.1 Turn-Taking

The turn-taking model was originally proposed by Sacks et al. in 1974 [1]. Since then, it has been one of the most applied and preferred models for the analysis of communication and communication systems. The model has undergone decades

of research, proving its broad application and validity. While many other models fail to have universal application among different cultures and different languages, turn-taking has also proven helpful in this regard [2] [3].

2.1.1 Turns and Rules

The model of turn-taking lives on its adaptability and broad application. Its strength is that it does not follow a predetermined structure for a conversation. The model states that people cannot listen and speak at the same time. Since that is the case, participants in a conversation need to figure out who takes on the role of speaker at any given time and who functions as the listener. The model suggests that humans take turns in speaking and listening, and every time these roles change, a turn has been completed. A turn can itself consist of many different components, called "turn-constructional units". These consist of resources that include: "lexis (or words), phonetic and prosodic resources (rhythm and intonation), syntactic, morphological and other grammatical forms, timing (e.g. very slightly delaying a response), laughter and aspiration, gesture and other bodily movements and positions (including eye gaze)" [4]. Some of them are discussed in greater detail later. Sacks et al. specify three rules for when such a turn is over, and the roles can change:[1] [2]

1. Rule: Person A (currently speaking) selects person B to be next in line to speak their turn. This may be verbal, by directly naming the person that should speak next, or indirectly, by addressing a person with gestures, mimics, or by gazing into the person's eyes. If the conversation is dyadic, then the explicit addressing of a person defaults to the conversation partner.

2. Rule: This rule is only used if the first rule is not applied. That means the current speaker (Person A) does not specify a new speaker. Then the other participants of the conversation can self-select to have a turn and be the new speaker. The first, or sometimes the loudest person to start, is the new speaker. Overlaps can occur here when two or more people start speaking at the same time, which is discussed in detail later.

3. Rule: This rule is used only when the first and second rules are not applied. This means that neither Person A appointed a new speaker, nor did a conversation participant self-select themselves. In that case, person A can continue to speak after a gap, or the conversation can be over.

Criticism of this conversation model has been expressed, especially when trying to model computer-mediated communication. Simpson [5] argues that to model

computer-mediated communication, a different approach is needed, since turn-takings cannot be as effectively seen until they are sent, because non-verbal clues are missing. Simpson proposes a different model, that of a conversational floor, although his study is focused on text-based communication. As this paper analyzes telephone conversations, the turn-taking model is best suited, since the differences from regular face-to-face conversation are interesting as well, and it can analyze both settings.

2.1.2 Silences

Originally, Sacks et al. [1] proposed three different categories for acoustic silence: pauses, gaps, and lapses. Pauses refer to the silence between the turns of the same speaker. Gaps refer to short silences between turns and speaker changes. And lapses refer to longer silences between turns and speaker changes. This classification has its problems, since a silence between the same speaker speaking was always classified as a pause, even though it could have been a rule three application and another person could have spoken, but did not [6]. Furthermore, the same is true for gaps, as silences between speakers should count as gaps even though the turn did not change. Many different names for acoustic silences have been proposed, most of which describe the same three differences originally explained by Sacks et al. As proposed in [6], the original names are kept, pauses, and gaps, but actual turn-takings are not considered, and only whether the person speaking has changed. Furthermore, it's not distinguished between lapses and gaps. Providing us with the definition of a gap being the time passing between different people's utterances. And pauses are the quantified silence between the same person speaking. For the measurements later only, the gaps are interesting for comparison, as the duration of the pauses should not vary much.

2.1.3 Overlaps

Overlaps are occurrences when two or more people speak at the same time. These can, of course, arise at transitions between turns and speakers, for example, if in a group setting, rule two is applied. But also, during someone's turn, if another person tries to interrupt or give feedback (backchannels). This begs the question of whether overlaps and silences are two separate things. When the time passing between turns is quantified, a positive value is a silence and therefore a gap. But if there is an overlap of speakers, the amount of time two people were uttering is conceptualized as the negative value of a silence. This means that the gaps between turns are overlaps, when there was no silence detected. This results in a continuum

of values that are counted as overlaps if they are negative and as gaps when they are positive.

2.1.4 Prediction or Anticipation

If these gaps and overlaps are quantified, they show some interesting characteristics: According to [6], these gaps are typically quite short, with most of them consisting of about 100 - 300 ms of silence. Interestingly, language interpretation has been estimated to be at least 600 ms long [7]. This begs the question of why these gaps appear to be shorter than the time needed to understand the last word that has been said. Researchers have two different opinions on that matter: First, there is the projection theory, proposing that the next speaker is anticipating the end of the turn, with cues given by the current speaker (originally proposed by Sacks et al.). Secondly, there is the signal theory, which tries to explain this apparent paradox by stating that the next speaker reacts to a specific signal, given by the current speaker, that they are about to finish their turn.

2.2 Backchannels

Backchannels, in general, are a way to communicate and send feedback to the other person who is currently speaking, without stopping their flow of speech and without interrupting their turn. In the past, many different definitions of essentially the same thing have been proposed with many different names [8]. But the principle is always very similar: these backchannels are most used to signal the speaker that we are still listening and that the speaker should continue to speak [9]. Other uses include signaling confusion, disinterest, or communicating that we want to have a turn to speak. These are extremely important because otherwise, every time we want to convey information, a turn would need to take place. For example, when sending feedback that we are still paying attention. And with that, making conversation much longer and less dense with information. Many functions of backchannels have been proposed, according to [10], the following seven categories from [11] are best for recognizing emotional function and are more descriptive: discourse markers, dispreference markers, hesitation markers, assessment markers, assessment tokens, continues, and newsmakers. Backchannel feedback is placed in very short pauses made by the speaker, [12] calls them "inter-personal units"(IPU), as earlier defined, these are also a form of pause.

When comparing backchannel feedback from different countries, languages, and cultures, there appear to be important differences regarding length, volume, and

frequency [13]. More research is needed on cultural differences regarding communication styles. The experiments in this paper are all done in the Swiss-German dialect.

2.2.1 Verbal Backchannel

Backchannels have many forms, and researchers have discussed what counts as a feedback channel. The first and most obvious one is the verbal backchannel. For example, words uttered by the listener while not on their turn and not breaking the turn of speaker. This verbal feedback most often consists of short words, for example, for signaling social awareness, these can include "yeah", "yes", and "aha-aha". These can also be non-lexical words and just agreeing sounding syllables [14]. Of course, almost any word can be uttered through a backchannel to present quick feedback to the speaker.

2.2.2 Non-Verbal Backchannel

Other than verbal backchannels, non-verbal backchannels as a concept enclose every feedback that is given other than the spoken word. Head movements, facial expressions and mimics, eye contact, and in general body language and gestures are all part of this concept. Head nods, or vertical head movements, are the most common form of this non-verbal backchannel[10], with one study finding that they make up around 42% of all the feedback given through backchannels[15]. Another common form is facial expressions such as smiling [16] or, for example, the "I don't get it look", to signal the speaker that they need to clarify what they are currently saying [15]. Similarly, the gaze can be a way to convey information to the speaker but also a way for the speaker to get feedback on what they are currently saying, more on that in the section "Feedback-Inviting-Cues". Combined, all the non-verbal feedback given exceeds the verbal ones, with one study in the Slovak language measuring them to be 61.1% and 38.9% respectively [9].

The frequencies of different backchannels change with contrasting topics of conversation. When having to do a task with lots of explaining dialog, more feedback needs to be given that the listener is following the conversation, than in communication about a random topic [9]. Also [17] notes that there are changing frequencies in backchannels, especially in non-verbal cues, in telephone settings depending on: gender, role, topic of conversation, mode of interaction, personality, and conflict handling styles of the interactants. When designing the experiment, these factors were kept in mind, but most of them are hard to control for.

Focusing on all backchannels at once is out of scope for this project. Most

research focuses on head nods and gazing, with not many laying their main focus on facial expressions. Furthermore, head nods are already an often chosen focus in backchannel research, and the same is true for the gaze. Also, gazes might be hard to artificially reintroduce and keep their subtle meaning. Thus, the focus of facial expressions is chosen over the other non-verbal backchannels.

2.2.3 Feedback-Inviting-Cues

As early as 1972, [18], it was discovered that feedback through backchannels is often invited by the speaker using signals with the self-describing name "Feedback-Inviting-Cues". They are used to indicate a specific point where the listener is encouraged to give feedback, without handing over the current turn. These can themselves be verbal and non-verbal. [12] finds five events in the English language that occur in higher frequencies before feedback is given through backchannels: a final rising intonation, a higher intensity level, a higher pitch level, a lower value of noise-to-harmonics ratio, and a longer pause duration. Of course, these signals can occur simultaneously. As these events tend to focus on verbal and pause cues, others have proposed that the speaker has many more ways to initiate a feedback-reaction from the listener: [9] found that 67% of backchannels usually take place within two seconds of mutual eye contact, with many of them being in the range of 500 ms.

Since Feedback-Inviting-Cues are generally not described as being linked with facial expression and the focus of them being very much on acoustic channels and gazes, it is to be expected that the frequency of turns changing will not be affected compared to the normal telephone setting.

2.3 Differences in Communication Settings

In a setting of dyadic telephone communication, verbal backchannels are, for the most part, identical to those in face-to-face interactions, as they are still detected and transmitted. Although backchannel feedback is more noticeable in phone conversations since visual cues are entirely lacking [19]. One thing worth pointing out, however, is the delay that occurs when transmitting speech via telephone, so one could argue that minor differences occur. For example, a very short pause extension as a Feedback-Inviting-Cue may not be picked up by the listener. Furthermore, overlaps tend to happen more often as the delay can be mistaken as the conversation partner not taking the next turn, when we expected them to. For non-verbal backchannels, the telephone setting eliminates every single one of them, and does

not provide a way to convey mimics, gestures, or gazes.

Gaps between turns are reported to be up to four times as large in face-to-face settings as they are in telephone settings [20]. Two explanations are given, although not enough tested: First, as non-verbal backchannels are missing, and people cannot use simple head nods or mimics to convey that they are still listening, verbal feedback and turns need to be more frequent and faster. Second, while communicating via telephone, it is usually the main task that the participants are doing. With face-to-face dialog, we often engage in other activities at the same time, thus adding to the gaps between turns.

When reintroducing facial expression backchannels and head nods to the telephone setting, it will be interesting to see what happens to these longer reported gaps, if they have no influence and it will be the same as normal telephone communication, or if a change in gap duration can be observed.

2.4 Emotions Essential for Communication

When picking up and displaying emotions, the question of which facial expressions to use arises. Many different sets of emotions were used in the past, for example, in [21] they used: anger, boredom, sadness, calm, surprise, happiness, and fear. Ultimately, the limit is the technology that is going to be used, and adaptations to the set of emotions present in the software are needed. This is discussed further in the following section.

2.5 Emotion Detection with DeepFace

DeepFace is a real-time, deep learning based face recognition software that is also able to analyze faces on their emotional state. As a dataset, they use the “Fec2013”, which consists of about 295 MB of space. There are 28 thousand training and three thousand testing images in this dataset. Each image is stored in 48×48 pixels, resulting in 2304 pixels per instance. Additionally, each image is annotated with an emotion type, and a classic train and test split is done. [22]

The network works with applied convolutional neural networks, or CNN. To train such a network, the first step is always one or more convolutional layers, where specified parts of the image get compressed and reduced to a set size. These can be repeated, and after a certain depth, the network is called ”deep”. This is done to separate certain features in the image. After that, these matrices are used to classify the original image into categories. For example, the network gets a happy face image and marks the key features, such as the mouth, that smiles. Through

	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Angry	214	9	53	30	67	8	86
Disgust	10	24	9	2	6	0	5
Fear	45	2	208	29	89	45	78
Happy	24	0	40	696	37	18	80
Sad	65	3	83	56	285	10	151
Surprise	7	1	42	27	9	303	26
Neutral	45	2	68	65	88	8	331

Table 2.1: Confusion Matrix

training, a smiling mouth is associated with the emotion "happy". Many more features can lead to the end classification. These emotion categories are picked and trained as stated above, with the DeepFace model, the following seven classes are predicted: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.

To test the accuracy of such a network, cross-entropy loss functions are used, where a value of zero would be a perfect prediction and a value of one a perfect mismatch. For visualization, a confusion matrix can be calculated, which shows its performance as it was trained on the train set and tested on the test set (Table 2.1). On the diagonal, there are the correctly identified emotions, and elsewhere, the incorrect guesses. So, cell (1, 2) was "Angry", but falsely identified as "Disgust". This accuracy would be good enough for our purpose. Although the set of emotions is suboptimal, while Anger and Disgust are not ever useful, confusion, for example, is a critical emotion missing from the set. To have meaningful results, we may need to tweak those sets of emotions or train some of them ourselves.

Also, noting that the non-verbal feedback-inviting-cues are often so subtle that it is almost impossible to reintroduce them automatically.

2.6 Measurements and Hypotheses

The goals of my experiments are to figure out the following questions and differences in measurements.

1. Automate the emotion detection in a web application in real time, such that the conversation partner can react to the feedback. The following questions should be answered:

   (a) Does the real-time element work as intended, and is it helpful regarding the flow of the conversation?

   (b) What's the scope of this application, and can it be used in everyday life?

   (c) Can the application minimize misunderstandings?

2. The Evaluation of the following data with comparison to the experiment without the application:
   (a) change in length of the turns.
   (b) change in length of the pauses and overlaps.
   (c) change in the frequency of the different backchannels.
   (d) change in the frequency of Feedback-Inviting-Cues.
3. Problems to solve in development:
   (a) Try different modes of displaying the detected emotions, such as avatars or emojis.
   (b) Feedback-Inviting-Cues may be hard to pick up. Experiment with buttons to address this issue.
   (c) In general, if the emotion detection does not work as intended or is too inaccurate, also experiment with "Feedback-Buttons", to give non-verbal feedback.

Hypotheses:

1. The extra feedback channels should be helpful, but very dependent on the execution of the software.
2. The length of the turns should vary slightly, such that the turns are longer when having more non-verbal feedback channels.
3. Less misunderstandings, which should show themselves in fewer overlaps with the non-verbal feedback channels present.
4. Less silence between turns, so also overall fewer gaps between turns (time between turns: silences and the negative overlaps values).
5. Turns per minute are expected to be lower with the emotions present because some turns changing could be eliminated through the reintroduced backchannels.

Chapter 3

Solution

3.1 Software

To analyze the non-verbal Backchannel feedback, the need for some sort of interface for displaying this to the conversation partner arises. The simplest way of conveying emotions and head-nods is with smileys. There are more sophisticated ways of showing these emotions, for example, with a full 3-D Model of the head, like an avatar that mimics all the head and face features. This would most certainly lead to more accurate results; however, the distinction between just video calling would then be very slim. So, for displaying emotions, smileys are used and for the head-nods, two little black stripes above and below the smiley-head to indicate vertical movement (Figure 3.1).

At the core of this software is the detection of the emotions of both participants. For this task, the open-source library DeepFace is used. Pictures get taken of a person every quarter of a second. Then every screenshot is analyzed by DeepFace on its dominant emotion, and on the percentages of the seven basic emotions. These emotions in DeepFace are: happy, sad, surprise, disgust, angry, fear, and neutral. As disgust, anger, and fear were never really used when testing speaking with another person, they were removed, and the values of them were combined into the one key emotion missing in this set, being confusion. This worked great, and when testing this setup, it could detect the expressed emotion most of the time. The last modification that was made was mapping the sad emotion to the neutral state, as some neutral faces were always incorrectly classified as sad.

For better results, key variables were implemented that helped to improve the results, and all of them were optimized through testing. The "DominantEmotion-Threshold" is defined to accept a dominant emotion only when the value exceeds it. It landed on 40 out of 100 for good accuracy. Next, the "EmotionFrameCounter" indicates how many consecutive frames need to reach the same emotion-decision to

Figure 3.1: Emotion and Head-Nod Display

display it. As only four pictures per second are taken, this must be rather low and came out to two consecutive frames. This means an emotion of the participant gets displayed to the conversation partner when two consecutive frames are identified as an emotion that differs from the current one displayed. For better results, the frames per second analyzed and the threshold could be amplified. After that, the change in emotion is locked for one second, to minimize stuttering of the emotions, but still detect every change. This variable is called "EmotionDelay", and one second seemed to be most natural.

For detecting vertical head movements, the same analyzing function from DeepFace can be used, which gives the x and y coordinates of the center of the face. Then it is just a matter of comparing different y-levels of consecutive frames to check if the participant is nodding. The threshold value of this is quite finicky, as a small value tends to produce false positives, but the head movements in conversations are mostly tiny. This variable seemed to deliver the best result with a value of 7 pixels.

Now, both participants can run the software, and it identifies their emotions and vertical head movements. The HTTP client "Axios" was used for fetching the information of the emotions via the LAN. For displaying them, "React JSX" was used. The product contains, on the top left, a field for displaying the tasks (explained later), on the top right, the detected emotion and head-nod of the conversation partner, and on the bottom left, your own detected emotion as feedback (Figure 3.2).

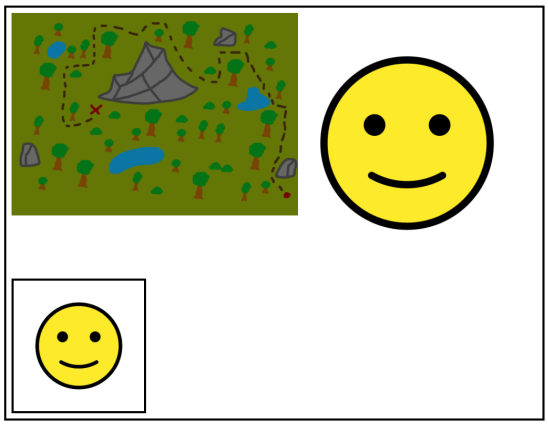

Figure 3.2: Experimental Setup

3.2 Experimental Design

Ten people in groups of two participated in the experiment. Two tasks were designed, where there is always an explainer and a follower role. Both roles are alternated, resulting in four different conversations and experiment runs. Both tasks needed to be challenging enough to encourage complicated conversations, but not so challenging that the smileys are ignored.

- Task 1 - Treasure Hunt (Figure 3.2): Person A is given a map with the marked start and finish location and a path between them. Person B has the same map, but only the starting position is marked. The task is now for person A to give instructions on where to walk along the path to find the treasure. The task ends when person B can point to the location of the buried treasure. Again, it is not about the task itself, but it gets the participants communicating.

- Task 2 - "Taboo": Person A receives a list of words and must describe them to person B without using the word itself or parts of it. The task is to get as many words correct in one minute.

One of each task, two in total, was run as a control experiment without the emotion feedback channels and only with the map or words displayed. The other two were run with the emotion detection present (Figure 3.2). The total four tasks were audio recorded and then evaluated. Additionally, the participants were given interview questions about the quality of the software and their experience with the experiments.

Considerations of the experiment:

- Gender: three female and seven male identifying people participated in the experiments.
- Roles of conversation: The roles of the participants were changed such that both participants were the explainer and follower for both tasks.
- Personality and conflict handling styles were not considered.
- Emotion feedback was alternated between the tasks, to eliminate effects of the variants of the tasks themselves. For example, if one path was easier to follow than the other one.
- Language: All of the experiments were done in the Swiss German dialect.

Chapter 4

Results

4.1 Empirical Results

The experiment featured ten people, who, in pairs, solved four tasks, giving us a total of 20 tasks done. In total, 774 turns were completed across all experiments, while 408 of them took place in the emotion-setup and 366 of them in the emotionless-setup. As both empirical and qualitative interview data were collected, let us first have a look at the measurements of the experiment (Figure 4.1). To interpret the collected values, the Mann-Whitney U test was performed. For all graphs, the p-values were far from significant, although mean differences are present (Figure 4.2).

Firstly, we can look at the average length of the turns. Overall, the participants were speaking about 2.87 seconds on every turn. This, however, is heavily dependent on the role in the conversation, as on average the explainers talked for 4.63 seconds at a time and the explainers for only 1.11 seconds at a time. This is a huge difference of about 3.52 seconds on average. This was to be expected and makes perfect sense, since the explainer had more information on the tasks and the followers only needed to react to them and signal, they are still following the instructions. Naturally, the explainers talked for longer at a time. When focusing on the difference between emotion- and emotionless-setup, there is almost no difference when it comes to the lengths of the turns.

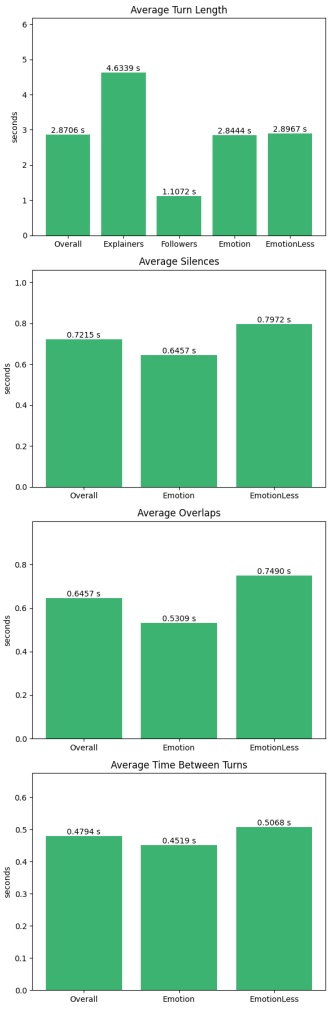

Figure 4.1: Results

Mann-Whitney U Test

	p-Value	U Statistic	Maximum U
Turn Lengths	0.4186	72‘130.5	149‘285
Silence Lengths	0.1682	44‘351.5	94‘792
Overlap Lengths	0.5931	1‘874.5	3‘969
Silence/Overlap Lengths Combined	0.7212	63‘779.5	129‘551

Figure 4.2: Mann-Whitney U Test

This is also supported by a very high p-value of 0.3631 and a U statistic of 72’130.5 compared to a maximum U of 149’285, suggesting little to no difference.

Secondly, we can analyze the results of the average silences between participants’ utterances. Here is the first actual difference between the emotion- and emotionless-setup, as the average length of the silences between turns was 645.7 ms for the emotion-setup and 797.2 ms for the emotionless-setup. This results in an average difference of 152.5 ms of silence. Although this looks promising, the p-value is again very high with 0.1682, but with a U statistic of 44351.5 and maximum U 94792, there are differences present, although not significant.

Thirdly, the average overlap time behaved similarly to the average silences. Here, the difference between emotion and emotionless is even bigger, with the average length of an overlap being 218.1 ms shorter with the emotions present. The emotion- and emotionless setups were 530.9 ms and 749 ms average overlap, respectively. If both numbers are combined, we end up with the fourth graph, the average time between turns. As both silences and overlaps are longer with the emotionless setup, the picture on this graph is less clear, although there is still a noticeable difference. The emotion-setup was, on average, 451.9 ms time between turns, and the emotionless-setup was 506.8 ms long. This results in a difference of 54.9 ms on average. It is very interesting to see is, that the average time between turns does not show a big difference, but the silences and overlaps were both longer in the emotionless setup, such that they had longer overlap-moments and longer silences where nobody was speaking. A similar picture is again drawn when looking at the p-values, as for overlaps it is very high with 0.5931, and silences and overlaps combined an astronomically high p-value of 0.7212, suggesting no significance at all.

Overall, the average length of the time passing between turns is already shown in the graphs. As a nice visualization, we can plot the frequency of the length of the gaps and overlaps for overall, emotion, and emotionless experiments (Figure 4.3). We can see that most gaps were between zero and 800 ms of silence. Also, if the participants happened to overlap, it was rather short, with most of them being less

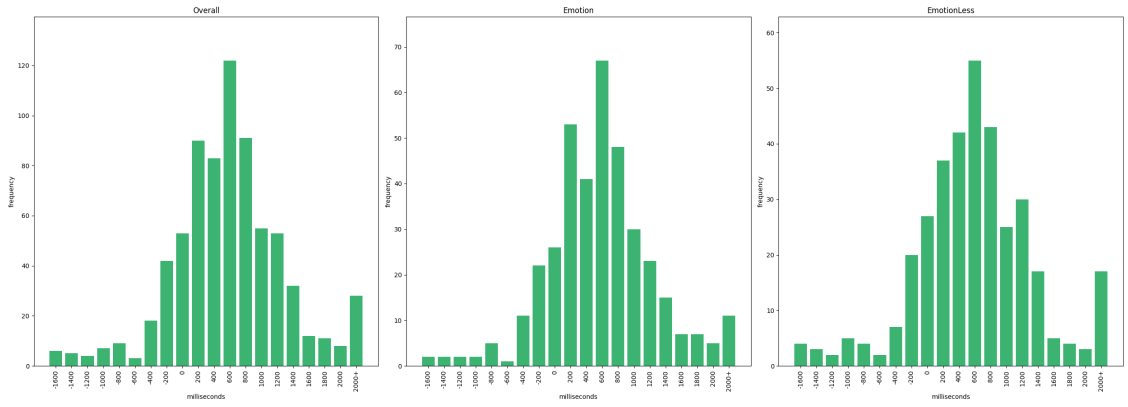

Figure 4.3: Frequency of Overlap and Gap Lengths

than 600 ms.

The chosen empirical approach was done using real-world data, which is a positive. The Mann-Whitney U test is robust and simple, but generally yields lower power findings, as was seen. Interpreting the results is clear when comparing means, but it does not explain the underlying reasons why. To better this understanding, the following key takeaways from the interviews can be consulted.

4.2 Qualitative Results

Key Takeaways from the interviews:

- The tasks were rated as easy and not complex.
- When asked what would make the task easier, every answer focused on making the task easier and not communication.
- When directly asked about the difference in telephone and face-to-face communication, and which would be easier to do the tasks, the answers were almost fifty-fifty split. Face-to-face would be easier, but the answers were mixed. An explanation could be that the tasks were designed to work on the computer, so it was seen as easier.
- The answer to whether the smileys were at all helpful was also mixed; some said yes, while others were fully focused on the tasks and were not paying much attention to them, or found them rather distracting.
- Also, the fluidity of the conversation was reported to be almost the same regardless of the emotions present.

- What emotion was most useful, and were all named about the same amount. This appears to be very personal.

- What was distracting or bad: The emotions were not correctly displayed sometimes, also the "abrupt" changing of emotions confused some. Almost half reported having no negative impacts on the conversation.

- When asked about real-life applications for this software, the answers were again mixed; some were open to it, while others failed to see the need for a hybrid approach between telephone and video chatting.

4.3 Revision of Hypotheses

First Hypothesis: "The extra feedback channels should be helpful, but very dependent on execution of the software." - The answers were rather mixed, while some found it helpful, many critiqued the technical limitations of the software or the inaccuracy. The idea was to find a middle ground between feedback-less telephone calls and video chatting, while some criticized this approach and preferred either telephone or video chatting, others liked the anonymity with the added emotions.

Second Hypothesis: "The length of the turns should vary slightly, such that the turns are longer when having more non-verbal feedback channels." - This cannot be confirmed by the data. The original idea was that since more non-verbal feedback channels are present, fewer turns need to take place, since the turn does not need to change when only nodding, for example, contrary to saying "yes". Emotionless was even a bit longer than with the emotions were present.

Third Hypothesis: "Less misunderstandings, which should show themselves in fewer overlaps with the non-verbal feedback channels present" - Statistically, this was at least a little indicated by the data. The overlaps were shorter with the emotions present. On average, 218.1 ms longer were the overlaps with the emotionless setup, although without statistical significance.

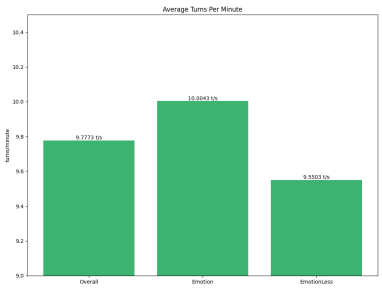

Figure 4.4: Turns Per Minute

Fourth Hypothesis: "Less silence between turns, so also overall less gaps between turns (time between turns: silences and the negative overlaps values)" - Similarly to the third hypothesis, this was also at least hinted at by the data, as both silences and overall time between turns were shorter when

the emotions were present. With 152.5 ms shorter silences on average, and 54.9 ms shorter time between speaking turns. But also, there is no statistical significance.

Fifth Hypothesis: "Turns per minute are expected to be lower with the emotions present because some turns changing could be eliminated through the reintroduced backchannels." - Finally, this was not confirmed by the data; in fact, the opposite was true. More turns per minute were recorded across all tasks when the emotions were present. Although this difference is rather small, with approximately half a turn per minute more when the emotion feedback was present (Figure 4.4).

4.4 Comparing to Research

My gaps between turns, also when even the overlaps are counted negatively, are much longer than typically observed in the literature. From 100 - 300 ms to my average of 479 ms overall. My idea is that the internet telephone used adds to this delay in the gaps between turns. Also, the topic of conversation was rather demanding, such that the participants had to focus on the tasks and were more focused on thinking about what was being said, rather than just replying quickly like in a casual conversation. Other results are hard to compare as my experiment setup is rather unique in setup and design.

Chapter 5

Conclusion

Despite some drawbacks to my approach, the results gathered from the thesis and experiments are pleasing. Although some hypotheses were not supported by the measurements, others were backed by the data, despite non-significance. What was most challenging about the developed solution was the dependence and inaccuracy of the chosen emotion detection from DeepFace. Many factors and additions made this better, but certainly not perfect. The emotion detection works fine most of the time, but it has its fair share of false positives and negatives, which confused most participants. In two cases, so much so that they chose to ignore the emotional feedback given and only focus on the verbal communication. Despite this, the data did show a change between the two experimental setups.

Another drawback the smileys brought was the "sharp" changes in emotions. Whereas humans tend to make smooth transitions with their emotions, the smileys always "jump" to the next one. This negative effect is amplified by the fact that, despite the "emotion-delay", the emotion detection still tended to be quite "jumpy" and restless. This, however, is in direct clash with the need to pick up, even small smiles or frowns. This could be optimized further with more time, adding parameters, and increasing the frames per second for emotion detection.

As a software idea for real-life applications, it still needs some work and refactoring to get the emotion detection better suited for aiding communication. As stated in the results already, most of the participants gave feedback that the application would have a real-life application where more feedback channels are needed, but for privacy reasons, video chatting is not acceptable. However, the form of the emotions displayed with smileys would also need a change, as at least half the participants noted that the emotions were not as clear or "weird".

In the future, many more aspects of interpersonal computer-mediated communication are worth exploring. What should be experimented with are different forms of conveying emotions, not just through smileys. In my opinion, 3-D avatars would

do a much better job at this, and it would also be possible to eliminate the need to detect emotions in the first place if every feature of the face is displayed to the conversation partner. This would also be helpful when adding "Feedback-Inviting-Cues" to the research, as then even gazes could be transmitted. Additionally, the help of buttons to signal conversation cues to the partner could prove helpful and would be interesting to experiment with. The whole setup could theoretically also be applied to group conversations. Where the avatar or emotions of the current speaker are seen by everyone, like office video-calls tend to be. In addition, there are buttons already in use in current software to virtually raise your hand, for example. If the emotion feedback and head nods were to continue to be the sole non-verbal channels, the emotion detection would need some more work. The best thing would be to train a CNN ourselves and directly choose at the beginning which emotions are needed. This would certainly yield more accurate results, although the need for training images would be quite large.

The last point of criticism is the relatively small sample size of my experiment. As every run required two people meeting up at the same time, at the same location, the organization was rather complicated. Also, transcribing the experiment data by hand was a big time investment, as the effort to automate this endeavor seemed almost impossible for the level of detail required. In an ideal world, the number of contestants would be considerably larger. This is also most likely the cause of the non-significance of the results.

Bibliography

[1] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Linguisitc Society of America*, 50(4):696–735, 1974.

[2] Stepgen C. Levinson and Francisco Torreria. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6, 2015.

[3] T. Stivers, N.J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J.P. de Ruiter, K. Yoon, and S.C. Levinson. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences (PNAS)*, 106(26):10587–10592, 2009.

[4] P. Drew. *Turn Design*, chapter 7, pages 131–149. John Wiley Sons, Ltd, 2012.

[5] James Simpson. Conversational floors in synchronous text-based cmc discourse. *Discourse Studies*, 7(3):337–361, 2005.

[6] Mattias Heldner and Jens Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568, 2010.

[7] P Indefrey and W.J.M Levelt. The spatial and temporal signatures of word production components. *Cognition*, 92(1):101–144, 2004.

[8] Keiko Tsuchiya. *Listenership behaviours in intercultural encounters*. John Benjamins Publishing Company, 2013.

[9] Stanislav Ondáš, Eva Kiktová, Matúš Pleva, and Jozef Juhár. Analysis of backchannel inviting cues in dyadic speech communication. *Electronics*, 12(17), 2023.

[10] Saya Ike. The interactional basis of backchannel behaviour in japanese english. , (47):129–138, 2016.

[11] Rod Gardener. *When listeners talk*. John Benjamins Publishing Company, 2001.

[12] Julia Bell Hirschberg and Agustin Gravano. Backchannel-inviting cues in task-oriented dialogue. *Proceedings of Interspeech*, 2009.

[13] S. Benus. The prosody of backchannels in slovak. 2016.

[14] Shoichi Iwasaki. The northridge earthquake conversations: The floor structure and the 'loop' sequence in japanese conversation. *Journal of Pragmatics*, 28(6):661–693, 1997.

[15] Lisda Nurjaleka. Backchannels responses as conversational strategies in the interaction of indonesian speakers in interview setting. *Journal of Research and Innovation in Language*, 5(2):99–114, 2023.

[16] L. J. Brunner. Smiles can be back channels. *Journal of Personality and Social Psychology*, 37(5):728–734, 1979.

[17] Alessandro Vinciarelli, Paraskevi Chatziioannou, and Anna Esposito. When the words are not everything: The use of laughter, fillers, back-channel, silence, and overlapping speech in phone calls. *Frontiers in ICT*, 2, 2015.

[18] Stakey Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292, 1972.

[19] Bettina Heinz. Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of Pragmatics*, 35(7):1113–1142, 2003.

[20] L. ten Bosch, N. Oostdijk, and J.P de Ruiter. *Durational Aspects of Turn-Taking in Spontaneous Face-to-Face and Telephone Dialogues*, pages 563–570. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[21] Min Chen, Ping Zhou, and Giancarlo Fortino. Emotion communication system. *IEEE access*, 5:326–337, 2016.

[22] Sefik Ilkin Serengil. Facial expression recognition with keras. Technical report.

Erklärung

gemäss Art. 30 RSL Phil.-nat.18

Name/Vorname: Hertzberg Leonard Felix

Matrikelnummer: 21-127-303

Studiengang:

Bachelor ☑ Master ☐ Dissertation ☐

Titel der Arbeit: Between Phones and Video-Chats -
A Computer-Mediated Communication Experiment

LeiterIn der Arbeit: PD Dr. Kaspar Riesen

Ich erkläre hiermit, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls der Senat gemäss Artikel 36 Absatz 1 Buchstabe r des Gesetzes vom 5. September 1996 über die Universität zum Entzug des auf Grund dieser Arbeit verliehenen Titels berechtigt ist.
Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.

Ort/Datum 04.08.2025

Unterschrift