

Master Thesis

April 29, 2019

# Boosting Collaboration

Automatically Measuring Meeting Quality Based  
on Speech Features

**Claudia Vogel**

of Grabs SG, Switzerland (13-556-055)

**supervised by**

Prof. Dr. Thomas Fritz

André Meyer



University of  
Zurich<sup>UZH</sup>





Master Thesis

---

# Boosting Collaboration

Automatically Measuring Meeting Quality Based  
on Speech Features

**Claudia Vogel**



University of  
Zurich<sup>UZH</sup>



**Master Thesis**

**Author:** Claudia Vogel, [claudia.vogel@uzh.ch](mailto:claudia.vogel@uzh.ch)

**Project period:** 01.11.2018 - 30.04.2019

Software Evolution & Architecture Lab  
Department of Informatics, University of Zurich

---

# Acknowledgements

First of all, I would like to thank Prof. Dr. Thomas Fritz for giving me the opportunity to write this master thesis at the software evolution and architecture lab (s.e.a.l.) and for his great ideas and his excellent support. Many thanks to André Meyer for his valuable comments, suggestions and ideas. Further appreciation goes to the company and all its employees, which enabled us to hold the study. Many thanks to all the participants for your time and effort. In addition, I would like to thank Prof. Dr. Volker Dellwo and Thayabaran Kathiresan of the Institute of Computational Linguistics of the University of Zurich for supporting me with the analysis of human speech and giving me interesting insights. The last thanks go to my family and friends for the motivation and strength they gave me.



---

# Abstract

Meetings are a big part of work life. While they are important, meeting participants often perceive them as a waste of time. Attending bad meetings can not only negatively affect the general mood of employees but also lower their job satisfaction. Researchers have looked into an automatic analysis of meeting quality based on features that can be extracted from the speech. However, all of these studies are limited as they often depend on the human coding of events from recorded audio data or on usage of scenario meetings in which each participant is given a specific role.

This thesis explores whether it is possible to measure participants' perceived quality of meetings using automatically extractable features of audio recordings from non-scenario meetings. To investigate this, we conducted a multi-day field study in a company based in Germany. The gathered data consists of 25 raw audio recordings from the meetings as well as 78 answers of the post-meeting survey that assesses the perception of meeting quality. We developed an approach to extract speech-related features of a raw audio file which we used to analyze the collected audio recordings. The results of our survey indicate that an open and positive meeting atmosphere and a lively exchange in meetings are both important factors contributing to participants' meeting quality. Results further suggest that the number of speaking turns is the only factor that we captured automatically and that is related to the meeting quality. Nevertheless, we see a potential to increase the meeting quality in the future, either by providing awareness about participant's meeting behavior or by reflecting on the meeting quality over time.





---

# Zusammenfassung

Mitarbeiter sind oft frustriert über Meetings und nehmen sie als Zeitverschwendung wahr. Die Teilnahme an schlechten Meetings kann die allgemeine Stimmung der Mitarbeiter negativ beeinflussen. Forscher haben sich mit einer automatischen Analyse der Meetingqualität beschäftigt, die auf Merkmalen basiert, die aus der Sprache extrahiert werden können. Allerdings sind diese Studien begrenzt, da sie oft von der menschlichen Kodierung von Ereignissen aus aufgezeichneten Audiodaten oder von der Verwendung von Szenario-Meetings abhängen, in denen jedem Teilnehmer eine bestimmte Rolle zugewiesen wird.

Diese Arbeit untersucht, ob es möglich ist, die von den Teilnehmern wahrgenommene Qualität von Meetings durch automatisch extrahierbare Merkmale von Audioaufnahmen aus nicht szenariobasierten Meetings zu messen. Um dies zu untersuchen, haben wir eine mehrtägige Feldstudie in einem deutschen Unternehmen durchgeführt. Die gesammelten Daten bestehen aus 25 Rohaudioaufnahmen der Meetings sowie 78 Antworten der Umfrage, die nach dem Meeting ausgefüllt wurde und die Wahrnehmung der Meetingqualität ermittelt. Wir haben einen Ansatz entwickelt, um sprachbasierte Merkmale einer Rohaudiodatei zu extrahieren, mit dem wir die gesammelten Audioaufnahmen analysiert haben. Die Ergebnisse unserer Umfrage zeigen, dass eine offene und positive Meeting-Atmosphäre und ein reger Austausch in Meetings wichtige Faktoren sind, die zur wahrgenommenen Meetingqualität der Teilnehmer beitragen. Die Ergebnisse deuten weiterhin darauf hin, dass die Anzahl der Redeanteile mit einer höheren Zufriedenheit der Teilnehmer verbunden ist. Basierend auf unserer Analyse war dieses sprachbasierte Merkmal jedoch das einzige, das mit der Zufriedenheit der Teilnehmer zusammenhing. Dennoch sehen wir eine Möglichkeit, zukünftig die Meetingqualität zu erhöhen, entweder durch die Bewusstseinsbildung über das Meeting-Verhalten der Teilnehmer oder durch die Reflexion der Meetingqualität im Verlauf der Zeit.



---

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Related Work</b>	<b>3</b>
2.1	Definition of Meeting Quality . . . . .	3
2.2	Measurable Factors Contributing to Meeting Quality . . . . .	4
2.3	Extracting Audio Features - Speech Processing . . . . .	5
2.4	Speech Features Contributing to Meeting Quality . . . . .	7
2.5	Meeting Corpora . . . . .	9
2.6	Meeting Feedback Systems . . . . .	10
<b>3</b>	<b>Approach to Automatically Extract Meeting Quality Features</b>	<b>13</b>
3.1	Open-Source Libraries . . . . .	13
3.1.1	Low-Level Feature Extraction . . . . .	13
3.1.2	Voice Activity Detection . . . . .	14
3.1.3	Gender Detection . . . . .	14
3.1.4	Overlapping Speech Detection . . . . .	15
3.1.5	Laughter Detection . . . . .	15
3.1.6	Emotion Recognition . . . . .	16
3.1.7	Speech Recognition . . . . .	16
3.2	Speaker Diarization . . . . .	17
3.3	Our Approach to Automatically Extract Features . . . . .	22
3.3.1	Implementation . . . . .	22
3.3.2	Extracted Speech Features . . . . .	24
<b>4</b>	<b>Study Method: Field Study</b>	<b>29</b>
4.1	Participants . . . . .	29
4.2	Method . . . . .	29
4.3	Data . . . . .	31
4.4	Analysis . . . . .	31
<b>5</b>	<b>Results</b>	<b>35</b>
5.1	Overview of Captured Meetings . . . . .	35
5.2	Stated Meeting Factors Contributing to Meeting Quality . . . . .	38
5.3	Correlation Between Reported Meeting Factors and Meeting Quality . . . . .	40
5.3.1	Inherent Factors . . . . .	40
5.3.2	External Factors . . . . .	42
5.4	Correlations between Extracted Speech Features and Meeting Factors . . . . .	44

---

5.4.1	Relationships Among Meeting Factors . . . . .	45
5.4.2	Relationships Among Speech Features . . . . .	45
5.5	Correlation Between Automatically Extracted Speech Features and Meeting Quality	46
5.6	Threats to Validity . . . . .	46
5.6.1	External Validity . . . . .	46
5.6.2	Internal Validity . . . . .	47
5.6.3	Construct Validity . . . . .	48
<b>6</b>	<b>Discussion and Future Work</b>	<b>49</b>
<b>7</b>	<b>Conclusion</b>	<b>55</b>
<b>A</b>	<b>Contents on CD-ROM</b>	<b>67</b>
<b>B</b>	<b>Post-Meeting Survey</b>	<b>69</b>
<b>C</b>	<b>Overview Meetings</b>	<b>73</b>
<b>D</b>	<b>Descriptive Statistics Group-Level</b>	<b>75</b>
<b>E</b>	<b>SPSS Output Stepwise Linear Regression Analysis</b>	<b>77</b>

## List of Figures

2.1	Sound Wave . . . . .	6
2.2	Spectrogram . . . . .	6
2.3	Sine Wave . . . . .	7
2.4	Representation of Jitter and Shimmer . . . . .	7
3.1	Segmentation of Audio Recording in its Speakers . . . . .	18
3.2	Screenshot of Diarization Editor . . . . .	24
5.1	Distribution of Participant's Quality of the Time Spent . . . . .	36
5.2	Distribution of Participant's Meeting Satisfaction . . . . .	36
5.3	Distribution of Centrality Features . . . . .	38

## List of Tables

3.1	Example of Speaker Diarization Output . . . . .	19
3.2	Diarization Error Rates of Speaker Diarization Systems . . . . .	22
4.1	Sample Survey Questions . . . . .	30
4.2	Coding of Likert Items . . . . .	32
4.3	Mapping of Subquestions Between Question 2 and 4 . . . . .	33
5.1	Main Findings of Empirical Data Analysis . . . . .	35
5.2	Distribution of Meeting Factor's Effect on Meeting Quality and Occurrence of Factor Expressed by the Means . . . . .	41
5.3	Spearman's Rank Correlations Between Meeting Factors and Participant's Quality of the Time Spent . . . . .	42
5.4	Spearman's Rank Correlations Between Meeting Factors . . . . .	46
5.5	Spearman's Rank Correlations Between Speech Features . . . . .	47
C.1	Overview of Meetings . . . . .	74
D.1	Descriptive Statistics of Group-Level Variables . . . . .	76
E.1	ANOVA . . . . .	77
E.2	Model Summary . . . . .	77
E.3	Coefficients . . . . .	78



# Introduction

In today's work environments, meetings are omnipresent, an integral part of work and an essential tool of an organizational life [NL92, CRAL11]. Employees spend a considerable amount of time in meetings, on average around six hours a week and managers even up to almost 23 hours a week as a study has shown [RSK07]. If they are conducted well, they provide many benefits, such as stimulating creative thinking and generating ideas, enabling discussions and sharing of information, and leading to clear action plans and decisions [CRAL11, ABSR14, MAVS18, NE17].

At the same time, studies have also shown that meetings are often perceived as negative or unproductive by employees [ASM<sup>+</sup>12, MFMZ14] and can have many negative effects, especially if the participants were not satisfied with the meetings. These negative effects range from emotional exhaustion and decreased employee engagement [LWAB16] to an increase in stress and fatigue levels [LR05]. Additionally, a negative impact on their job satisfaction [RAS<sup>+</sup>10].

Several studies have been performed to examine the possible factors that influence the quality of a meeting in order to identify areas for improvement. They have found various aspects impacting participants' meeting quality, such as participants' behavior in meetings [LWAB16], interactions between group members [SK16], or design characteristics (e.g. the use of an agenda) [ALWR18, CRAL11].

More recently, researchers have also looked into a more automatic analysis of meeting quality based on the extraction from features of human speech. Studies have shown, that the turn-taking patterns might be an indicator for the perceived meeting quality of participants [CSR16, LCR13, LM18]. More specifically, Lai *et al.* [LCR13] showed that higher meeting satisfaction is not related to equal participation among participants, but to the possibility that the participants can take the floor freely and not in a strict order. In another publication, Lai and Murray [LM18] automatically predicted meeting satisfaction using a multi-modal approach containing acoustic (e.g. pitch, loudness), lexical (e.g. word embeddings) and turn-taking features (e.g. participation equality). Overall, they found that different facets of meeting quality can best be predicted using individually trained models with selected features of all three modalities.

Yet, all of these studies are limited as they often depend on human coding of events from recorded audio data rather than automatic extraction, or focus on meetings in a highly controlled environment in which participants are given a specific role and task.

The objective of our work is to examine whether it is possible to measure participants' quality of their time spent in meetings using automatically extractable features of audio recordings from meetings. In particular, we want to extend previous work by analyzing data of real-life meetings and by providing an approach to extract features from raw audio files automatically. Therefore, we are investigating the following three research questions:

- RQ 1:** What are measurable characteristics that contribute to the perceived quality of a meeting according to related work?
- RQ 2:** Which features from audio recordings can be captured and how do they correlate with the quality of the meeting?
- RQ 3:** Is it possible to automatically provide awareness of the meeting quality based on speech data?

To answer these research questions, we first performed a literature analysis of the existing work in this area. Second, we designed and conducted a multi-day field study with professionals in a company based in Germany.

In total, we collected data from 25 meetings, including 78 survey responses from their attendees. The gathered data consists of the raw audio recordings from the meetings as well as a post-meeting survey that assesses the perception of meeting quality. We developed an approach to extract speech-related features of a raw audio file which we used to analyze the collected audio recordings. Among others, our approach can extract general features (e.g. meeting duration), turn-taking features (e.g. number of speaker changes), centrality features (e.g. participation equality) and prosodic features (e.g. speech rate). Based on the collected data and our approach, we performed an empirical analysis. One of our main findings is that the more open and positive the meeting atmosphere and the more lively and active a meeting is, the higher the perceived quality of the meeting. This can also be seen in the correlation between the speech-feature number of speaking turns and the meeting quality. At the same time, no other automatically captured feature related to participants' meeting quality.

The following are the main contributions of this thesis:

1. A week-long field study in a German company and the thereby collected data.
2. An approach to automatically analyze an audio recording and extract a set of speech features that are possibly related to participants' perceived meeting quality.
3. The results of an empirical analysis of factors contributing to participants' perceived quality of time spent in 25 meetings which include the survey answers and extracted speech features.
4. Discussion of opportunities to raise awareness of meeting information, such as group behavior, in order to foster the meeting quality of participants.

This thesis is structured as follows. Chapter 2 starts with a detailed overview of related work. Chapter 3 describes our approach to extract features of a raw audio file automatically. The description of our methodology, the participants and the study procedure are given in Chapter 4. Chapter 5 presents the findings of the empirical data analysis. Chapter 6 discusses our findings and presents future work and our ideas to extend this research. Chapter 7 concludes this work.



# Related Work

Related work will be discussed in the following with respect to six categories. We start by defining meeting quality and provide an overview of measurable factors influencing meeting quality. Next, we introduce the human speech and its analysis. Then, we show how features extractable from human speech relate to the meeting quality. Afterwards, we outline existing meeting corpora and conclude this chapter with a discussion of several approaches to provide awareness of the meeting information and their impacts.

## 2.1 Definition of Meeting Quality

Whether a meeting was successful or not depends on several aspects of quality. Based on a literature review, the quality of a meeting is often captured by three terms: satisfaction, effectiveness and efficiency. Related work has focused primarily on the analysis of meeting satisfaction and effectiveness.

**Meeting satisfaction.** Meeting satisfaction describes a subjective assessment of individuals on how well their time was spent during the meeting [Kle10]. According to Briggs *et al.* [BRdV06] meeting satisfaction consists of two sub-components: satisfaction with meeting process and satisfaction with meeting outcome. Meeting process satisfaction refers to participants' satisfaction of procedures and tools used in a meeting and how well the group works together during the meeting. Outcome satisfaction occurs when a participant feels positive about what is created in a meeting and how well the group is achieving the meeting goals.

Meeting satisfaction is a complex variable and difficult to quantify. So far, researches used different questionnaires for this purpose. Briggs *et al.* [BRdV06] assessed meeting satisfaction using a questionnaire of 12 questions participants answered by indicating their agreement. Two example questions are 'I feel satisfied with the way in which today's meeting was conducted' and 'I feel good about today's meeting process'. Rogelberg *et al.* [RAS<sup>+</sup>10] assessed participants' meeting satisfaction using a six-item scale. They presented six adjectives (stimulating, boring, unpleasant, satisfying, enjoyable and annoying) to the participants. Then, the participants were asked to indicate how these words described their meetings [RAS<sup>+</sup>10]. Next, a summed aggregation of the responses was calculated for each participant, representing the meeting satisfaction score. Cohen *et al.* [CRAL11] measured participants perception of meeting quality similarly, but using a 13-item scale and slightly different adjectives, including worthwhile, efficient and useful.

**Meeting effectiveness.** Meeting effectiveness describes participants' perceived degree of goal achievement of a meeting [PKS18]. Nixon and Littlepage [NL92] quantified meeting effectiveness

by two items: participants' degree of decision satisfaction and goal attainment of the group. Leach *et al.* [LRWB09] used three items to measure meeting efficiency by asking participants to rate the effectiveness of the meeting in terms of goal achievement of their own goals, colleagues goals and the department's goals. Rogelberg *et al.* [RLWB06] assessed meeting effectiveness in the same way as Leach *et al.* [LRWB09] except for using a six-item scale. Odermatt *et al.* [OKK16] extended this six-item measure by questions assessing whether the meeting provided the opportunity to acquire useful information and to network with people and if the meeting promoted commitment.

**Meeting efficiency.** Meeting efficiency describes the relationship between the invested time and meetings' outcome [Dav97]. Davison [Dav97] defined meeting efficiency as a meeting outcome measurable by the number of in-depth discussions of issues, the efficient use of time in a meeting (e. g. if participants spent time on serious discussion), the percentage of time spent in discussions of agenda-based items and the degree of result-orientation in the meeting.

## 2.2 Measurable Factors Contributing to Meeting Quality

Since we are interested in automatically measuring the meeting quality, we focus on the measurable ones. In general, these factors either address the design of a meeting or are focused on the interaction and contributions of participants to meetings.

**Meeting design characteristics.** Meeting quality can be influenced by the meeting design characteristics which can be categorized into temporal, physical, procedural and attendee aspects of the meeting. Temporal characteristics include the start and end times of the meeting. If they are not met, it can significantly lower meeting satisfaction and effectiveness [ALWR18, CRAL11]. The physical natures of meeting characteristics are qualities of the meeting room and its equipment, which also impact participants perceived meeting quality [LRWB09, CRAL11]. Procedural factors which are related to participants perceived meeting quality include the use of an agenda either a circulating written agenda before the meeting or a verbal agenda at start of the meeting [LRWB09, CRAL11]. Lastly, the attendee aspects are determined by the influence of the meeting size on participants' perceived quality. If a meeting has too many or unnecessary attendees, participants have the feeling that their time is not used in an effective way [CRAL11].

**Meeting interactions and behaviors.** The structure of meeting interactions and participants' behavior in meetings play a crucial role in perceived meeting quality. Meetings can be characterized by a centralized structure in which the interaction revolves around a main actor, or in a decentralized structure in which all participants are equally involved in the interaction [SK16]. Moreover, interactions can be in a functional or dysfunctional structure. The functional structure includes proactive statements or statements that structure the meeting, such as the delegation of tasks [SK16]. In contrast, the dysfunctional structure contains criticizing statements or complaints [SK16]. Kauffeld and Sauer [SK16] found that functional networks have a significantly more decentralized interaction structure than dysfunctional networks. Previous research has shown that functional meeting actions promote meeting satisfaction and team productivity, while these are negatively affected by dysfunctional meeting actions [KLW12].

Functional meeting behavior corresponds to meeting citizenship and good meeting behavior whereas dysfunctional meeting behaviors relate to counterproductive meeting behavior [LWAB16]. The main findings of the work of Lehmann *et al.* [LWAB16] were that counterproductive meeting behaviors lead to decreased employee engagement and increased emotional exhaustion. In

addition, participants' perceptions of uncivil meeting behaviors in a meeting, e.g. offending or intimidating of other attendees, has negative consequences on the meeting satisfaction [OKK<sup>+</sup>18]. In contrast, good meeting behavior is linked to an increase of time well spent of participants and a higher meeting satisfaction [LWAB16]. Moreover, participants of meetings with good behaviors feel more energized, which could boost their work engagement [LWAB16].

**Participation and involvement.** Individual participation and involvement influence the meeting quality reasonably. Participation refers to the degree of employees contributions in a meeting containing ideas, opinions, thoughts and feelings [YCA15]. Leach *et al.* [LRWB09] measured participants' involvement by assessing the extent to which the participants work hard and to which the participation is widespread among attendees. According to Leach *et al.* [LRWB09] higher involvement leads to a higher perception of meeting effectiveness. It should be noted that the involvement level negatively correlates with the meeting size [LRWB09]. Moreover, Nixon *et al.* [NL92] highlighted the positive impact of open communication towards meeting effectiveness. They assessed open communication with a variety of factors, e.g. if all participants participated, or whether participants feel comfortable to work together and sharing their own ideas within the meeting.

**Laughter.** Laughing is an important social nonverbal signal and an essential part of human social interactions [SGS<sup>+</sup>12]. Moreover, it is an indicator for the positive affective state of the human being and for a positive perception of discourse element [SGS<sup>+</sup>12]. Laughter measures not only the quality of the interaction but also to which extent a person is engaged in the interaction [SGS<sup>+</sup>12]. Additionally, laughter and humor unleashed positive socio-emotional communication and have effects on group performance [LWA14]

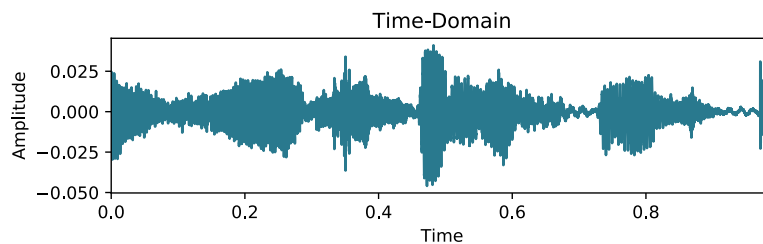
## 2.3 Extracting Audio Features - Speech Processing

To automatically extract any of the features from the speech, the digital representation of the speech needs to be analyzed. The study of these speech signals and their processing methods are called speech processing. To show how this can be done and to provide valuable background knowledge, we give a brief introduction to speech processing in general, the digital representation of speech, how features can be extracted and conclude with common extracted low-level features, such as the prosodic features or Mel-frequency cepstral coefficients.

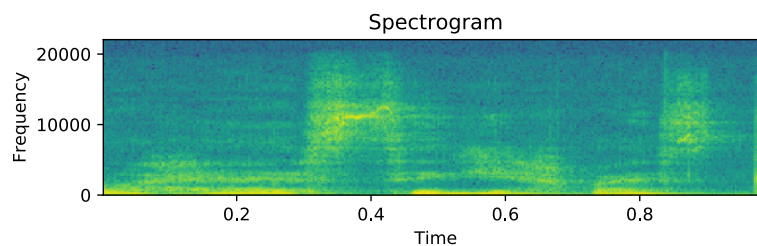
**Speech processing.** Human speech contains much information, such as speaker's identity, gender, emotion, and also about what a person is talking. Speech processing deals with the automatic extraction of information transmitted in a speech signal [Rey02]. Among others, some speech processing tasks are speaker recognition, speech recognition, gender identification and language recognition. For computer systems, it is quite challenging to solve these complex tasks automatically because computers have to mimic human behavior and external factors (e.g. noise) are very common. Thus, they are usually implemented using supervised and unsupervised machine learning algorithms.

**Sound.** Vibrations of objects create sound [RH10]. In human speech, it is the vibration of vocal folds which produces sound by setting an airstream in motion. The vibrating vocal folds push the normally uniformly distributed air molecules apart and together, which leads to a sound [RH10]. The sound wave travels through a medium, e.g. air molecules, until our ears perceive it. Among

others, the digital representation of human speech could be in the time-domain or time-frequency-domain, depicted in Figures 2.1 and 2.2, respectively. The time-domain represents the sound wave with the time on the x-axis and the amplitude on the y-axis. The fundamental characteristics of the speech signal, such as amplitude and frequency, are represented in a simple sine wave in Figure 2.3. The spectrogram, a frequency-time-domain representation of the sound, represents on the x-axis the time and on the y-axis the frequency. The color is proportional to the energy of the sound wave. With a spectrogram, one can observe speech sounds and their properties.

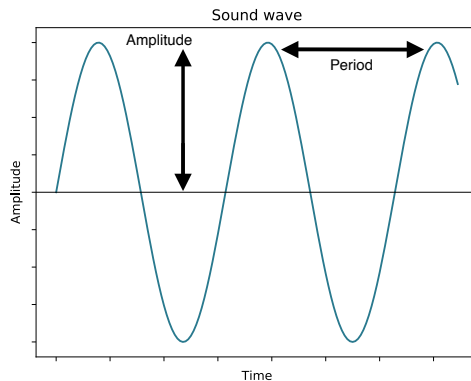


**Figure 2.1:** Sound Wave

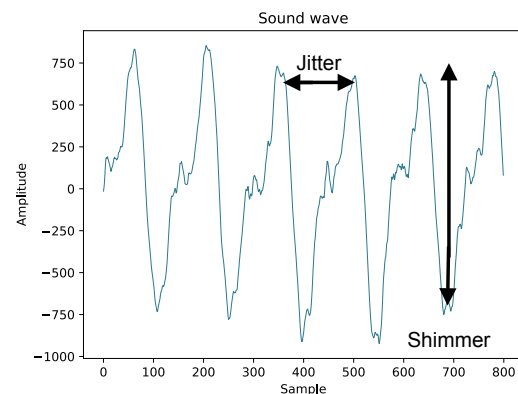


**Figure 2.2:** Spectrogram

**Windowing.** The audio is recorded with a sampling frequency, e.g. 16kHz implies 16'000 samples for each second. This equals the resolution of the audio signal: the higher the sampling frequency, the higher its resolution. The features are usually extracted on a windowed speech signal. Because it contains information about the characteristics of the speaker's vocal tract and the speech signal's temporal characteristics change quickly [Yel15]. Windowing divides the signal into smaller segments (called frames) of a specific size (called window size or frame size). On each frame, a certain feature will be extracted, e.g. fundamental frequency. After the extraction on a frame, the window shifts by its length (called window length) and extracts the next features of the currently underlying frames. Depending on the window size and window length the frames can be overlapping this is usually preferred to not miss out any sudden fluctuations. Typically, the features are extracted every 10ms on a short-term frame size of 25ms [Yel15]. For a 16kHz signal, this implies that 400 samples per frame ( $0.025s \cdot 16kHz$ ) are considered on which the features are extracted. After extraction, the window shifts by 160 samples ( $0.010s \cdot 16kHz$ ) and extracts the features based upon these underlying samples.



**Figure 2.3:** Sine Wave



**Figure 2.4:** Representation of Jitter and Shimmer values [TOL13]

**Prosodic features.** Prosodic features, also known as vocal or acoustic features, are a class of non-verbal cues describing ‘how someone speaks’, including pitch, energy and jitter/shimmer [Jay11].

Fundamental Frequency (F0) represents the number of vibration cycles (periods) repeated per seconds, in Hertz (Hz) [Ger03]. Human’s ear perceives the fundamental frequency as pitch: the higher the frequency, the higher the perceived pitch. Healthy young people can hear sound frequencies ranging from 20 to 20’000 Hz [Gra00]. The fundamental frequency greatly varies between several factors, such as age, gender or speaker’s mood [TE95]. As an example, females usually have a fundamental frequency of 210 Hz and males around 120 Hz [TE95].

Intensity refers to the amplitude of the vibrations, which equals the size of oscillation of vocal folds. It is measured in decibel (dB) and perceived as loudness by the human ear. The higher the intensity, the greater the amount of energy carried by a wave, leading to a louder sound perceived by the human ear. The smaller the intensity, the quieter humans perceive the sound. Normal conversation ranges around 50 - 70 dB whereas shouting is around 80 - 90 dB [Hac96].

Jitter/Shimmer are vocal characteristics describing the cycle-to-cycle variations of the fundamental frequency and amplitude of a signal because pitch and intensity slightly change over consecutive cycles [FH09]. Jitter represents the frequency variation from cycle to cycle and shimmer represents the amplitude variation of a sound wave. In Figure 2.4, these parameters are graphically represented. This variation of vibration cycles depends on several factors, such as the intonation or the emotional or mental state of the speaker [KPW15]. For example, jitter/shimmer features can be indicators of underlying stress in the speech [LTJ<sup>+</sup>07].

**Mel-frequency cepstral coefficients.** Mel-frequency cepstral coefficients (MFCC) is an extractable feature of the sound signal which is commonly used in speaker and speech recognition technologies [FH09]. MFCCs popularity in speech analysis tasks is attributable to the facts that MFCCs carry discriminative speaker information and are relatively efficient to compute.

## 2.4 Speech Features Contributing to Meeting Quality

Not only the content and the spoken words of human speech, but also non-verbal elements can provide information about social relationships, human behavior, personality or even about the

quality of meetings. In this section, we provide an overview of existing work analyzing the relationships between features extractable from human speech within meetings and participants perceived meeting quality. Moreover, other work that goes beyond the prediction of the meeting quality and focuses on the prediction of meeting productivity or social-psychological group variables such as a dominant speaker.

**Prediction of meeting satisfaction.** Chowdhury *et al.* [CSR16] created a classification system to predict user-satisfaction using features of different modalities: turn-taking features (e.g. amount of turns, pause between turns or overlapping speech), prosodic features (e.g. pitch or loudness) and lexical features (e.g. words transformed to vectors). Their system predicts user-satisfaction, which is either neutral, positive or negative, of real-life call-center dialogues using a trained model. The best performance was achieved if the model was trained on turn-taking features.

Lai *et al.* [LCR13] analyzed the relationship between turn-taking features and satisfaction, leadership and cohesion. Their research is based on meetings of the AMI corpus and its participants' post-meeting ratings. More information about this corpus can be found in Chapter 2.5. Turn-taking freedom (participants speak out in a relatively free manner and not in strict order) and amount of individuals very short utterances (speaker statements with a shorter duration than 500 ms) are good predictors for satisfaction and cohesiveness ratings. According to their findings, meeting satisfaction relies less on the equality of speaking time between participants, more on turn-taking-freedom and, consequently, on how well the information flows and how democratic the decision-making process was [LCR13].

In another research, Lai and Murray [LM18] focus on the prediction of three questions of the AMI post-meeting ratings about the overall satisfaction, participant attention satisfaction and information overload. They used acoustic, turn-taking and lexical features extracted of AMI corpus meetings. Their findings are that the three questions cannot uniformly be predicted using the same classifier. Moreover, feature removal significantly impacts the performance of some models. For example, adding prosodic features into the model improved the prediction of attention satisfaction but lower accuracy for the prediction of general satisfaction [LM18]. In summary, the best results are achieved by selecting different features from all three modalities to predict each question separately. This implies that different aspects of group satisfaction are expressed through different modalities [LM18].

**Prediction of meeting productivity.** Murray [Mur14,Mur15] analyzed productivity shifts within a meeting itself and the differences between a productive and unproductive meeting using linguistic and structural features. He measured productivity within a meeting by counting the number of important, summary-worthy sentences. His rationale behind this calculation was that a higher amount of important sentences represent a higher amount of productivity. To summarize, his main findings were that meetings having a large number of sentiment words tend to be less productive and meetings having a dominant participant tend to be more productive.

**Prediction of social-psychological group variables** Further work focused on the prediction of social-psychological components of groups in meetings, such as perceived dominance, extroversion or leadership. These are all important components, but their implication on meeting quality is yet unclear.

A person within a meeting is perceived as dominant by others if this person has the ability to control a conversation or influence a social interaction partner [Mas02]. Jayagopi *et al.* [JBOGP08, Jay11] predicted the most-dominant person of the meeting using the person's total speaking length relative to the speaking time of other participants. Other research in social-psychology leads to similar results in which speaking time is a strong predictor for the dominance perception by humans [JBOGP08,Mas02]. Moreover, Jayagopi *et al.* [Jay11] achieved higher accuracy in predicting

a dominant person using feature combination of the total speaking energy (how much and how loud a person speaks), the total speaking turns and total number of successful interruptions (e.g. speaker A starts speaking before speaker B finishes and speaker A continues speaking). According to Wang *et al.* [WKKO18] participants perceived a person as dominant if this person spoke with a high intensity and low fundamental frequency.

Beyan *et al.* used acoustic features to identify emergent leadership [BCBM17] and to detect a democratic or autocratic leadership style of an emergent leader [BCBM18]. An emergent leader is a person naturally having the characteristics of a leader than from a higher authority [BCC<sup>+</sup>16]. Bhattacharya *et al.* [BFZ<sup>+</sup>18] automatically processed the multimodal sensor outputs (visual, non-verbal and verbal metrics) in correlation with user-reported rankings of emergent group leaders. Their analysis resulted in a non significant correlation between the extracted nonverbal acoustic features and the perceived leadership. However, a combination of visual, non-verbal and verbal features or even verbal metrics alone predicted an emergent leadership with an accuracy of 90 % and 100 %, respectively.

So far, most of the existing related work heavily depends on human annotations and codings from recorded audio data. An approach towards automatic detection was published by Bhattacharya *et al.* [BFZ<sup>+</sup>18]. They created a smart meeting room capturing audio and video recordings in an unobtrusive manner. However, they did not analyze the meeting affect of participants and used data of meetings discussing Lunar Survival Tasks, which is a hypothetical survival scenario [BFZ<sup>+</sup>18]. Lai *et al.* [LCR13, LM18] used for their analysis not only automatically extracted features but also manual transcriptions. Moreover, their analysis depends on mainly scenario meetings in which attendees are given a particular role and a task to accomplish. In summary, related work showed two main characteristics, towards which we want to contribute: a) the automatically extraction of speech features of an audio recording, b) the analysis of naturally occurring meetings. Additionally, we want to integrate the laughs occurring in meetings into our analysis because related work does not yet include this metric.

## 2.5 Meeting Corpora

There are already several meeting corpora out there that contain audio recordings of multi-party meetings. In particular, the AMI Meeting Corpus is widely used in previous work. Below we provide a selection of publicly available meeting corpora.

**AMI Meeting Corpus.** The AMI Meeting Corpus [Car07] consists of around 100 hours of audio and video recordings of 171 meetings, which are captured in three different meeting rooms. Besides the recordings and its orthographic transcriptions, AMI Meeting Corpus provides a variety of annotations, e.g. meeting summaries, dialog acts and head movement. The audio recordings are captured with a microphone array and microphones around participants' necks. All meetings have four attendees. Around two-thirds of the meetings are scenario meetings, meaning that the attendees are assigned to a role which they represent during the meeting. The remaining meetings are naturally occurring meetings.

**ICSI Meeting Corpus.** The ICSI Meeting Corpus [DBCS04] consists of 70 hours of audio recordings of meetings having between three and ten attendees. All meetings are recorded using close-talking microphones. ICSI comprises around 75 naturally occurring meetings, in a sense that they would have taken place even without the recordings. They were usually regular weekly meetings of ICSI working teams. Moreover, the corpus provides transcriptions and a set of annotations, such as dialogue acts or topic changes.

**Emergent LEADER corpus.** The Emergent LEADER corpus (ELEA) [SCAMGP12] contains 40 meetings corresponding to approximately 10 hours of audio and video recordings. The meetings were recorded with a microphone array and consisted of three to four participants. ELEA comprises scenario meetings where participants had to solve a particular task, which is a hypothetical survival scenario (winter survival task). However, participants were not assigned to a specific role. Moreover, this corpus provides orthographic transcriptions and participants' answers to a questionnaire about personality traits, cognitive ability and emergent leadership.

**Group Affect and Performance Corpus.** The Group Affect and Performance (GAP) Corpus [BM18] consists of 13 audio recordings of small group meetings, corresponding to approximately 100 minutes of recording. The participants had to solve the winter survival task. Moreover, the GAP corpus contains transcriptions, annotations and post-meeting ratings of satisfaction. To capture the audio, a portable audio recording device (i.e. Zoom H1 Handy Recorder) was used.

## 2.6 Meeting Feedback Systems

Existing research implemented several systems with the purpose to provide meeting attendees feedback about group dynamics or information discussed in meetings. They can be categorized in participant-centered which visualize collaboration structures or content-centered which focus on the meeting content itself, e.g. what information has been discussed within the meetings. Some of the systems integrate audio and video recordings. Since we focus on audio recordings, we present in this section existing work which uses primarily audio signals.

**Participant-centered systems.** Graphical representations of group dynamics within a meeting are called group, or social mirrors [KK09]. They reflect the collaborations between participants in real-time which are dynamically updated. Various devices are used as a medium to share the visualizations across all meeting attendees, such as the table, a display at the end of the table, projections on the wall or the participant's mobile phone.

A popular choice among researchers is the representation of how equal the contribution currently is by displaying the speaking time of each participant [KWT05, SHET07, RMM<sup>+</sup>12, KCHP08]. Sometimes this visualization is supported through additional information, such as interruptions [ZBL<sup>+</sup>11]. An example of a simple feedback system is the representation of the distribution of speech time by a dynamically adjusting histogram, in which each participants' speaking time is represented by a bar [DPB04].

Reading about these systems, we noticed that group mirrors are characterized by multiple factors which we describe in the following:

- **Metaphorical vs. diagram representations:**  
Metaphorical representations visualize the information implicitly by changing specific properties of a pictorial scene [SSH09]. In contrast, the diagram visualization represents the information in charts or graphs [SSH09].
- **Visualization of qualitative vs. quantitative information:**  
Quantitative data aims to be solely objective and without judgment about the quality of the represented data, as done in [RMM<sup>+</sup>12]. In contrast, visualizations of qualitative data provide an interpretation of the represented information and therefore a standard for desirable behavior, e.g. quality of arguments or quality of the meeting. Obviously, quantitative data can more easily be automatically captured than qualitative data [SSH09].



- Real-time only vs. historical information:  
Besides providing visual feedback on the current situation, some visualization include an information history, e.g. of the last 30 seconds [DPB04] or five minutes [BK07b, BK07a].
- Individual vs group information:  
The systems could provide either for all participants the same visualization (e.g. in [SSHF09, DPB04]) or a personalized representation of each participants' own meeting behaviors (e.g. in [AMS17, LRS<sup>+</sup>18, SCM<sup>+</sup>14]).
- Active vs. passive visualizations:  
Conversation feedback systems can be passive by only visualizing the information or active by inviting participants to adjust their behavior [LRS<sup>+</sup>18, AMS17, SCM<sup>+</sup>14]. On the one hand, passive feedback could be provided by showing the participants their own current collaboration score, which is representing how much they have contributed so far [AMS17]. On the other hand, active feedback could be given by individual facilitation messages with guidance depending on the participant's current collaboration state, e.g. for an over-participant the message could be 'How about to be a listener?' [AMS17].

Despite the fact that existing work used different graphical representations, their evaluations showed similar results. Using group mirrors in meetings the participants regulated their behavior. This resulted in a more balanced collaboration [KCHP08, DPB04, BK07a, BK07b, KK09, BKD08, RMM<sup>+</sup>12, KWT05]. According to Di Micco *et al.* [DPB04], especially over-participants changed their behavior and reduced their speaking time. However, Di Micco *et al.* [DPB04] did not find a contribution increase for under-participants, whereas Sturm *et al.* [SHET07] measured an increase. Moreover, Di Micco *et al.* [DPB04] noticed that participants holding non-critical information decreased their speaking time, whereas participants with critical information were not impacted. Streng *et al.* [SSHF09] noticed a greater impact on behavior adjustments using metaphoric visualizations.

**Content-centered systems.** Content-centered systems enable the participants to review and edit the shared content of the meeting by structuring and visualizing it. An approach to provide a thematic overview of the discussions is word clouds [KGF02, SWQ<sup>+</sup>17, CBS<sup>+</sup>19, ZLPCT18]. These represent relevant and related topics graphically which are clustered into groups. Usually, these clusters consist of keywords which are extracted from utterances instead of the entire utterances. The use of these keywords supports the user in information extraction and reduces distraction [CBS<sup>+</sup>19].

This visualization can be extended by providing temporal data, which allows the participant to extract historical and temporal trends. A possibility could be to visualize thematic content over time, e.g. which topic was discussed at what point in time within the meeting [SBB<sup>+</sup>18]. A line for each speaker could complement this visualization, which indicates whether the attendees are talking about similar topics, represented through the proximity of these lines [SWQ<sup>+</sup>17]. Moreover, the distribution of speaking time could be visualized using the line's thickness. These tools could be interactive, so that the user could access more information such as the summarized content [SBB<sup>+</sup>18].

The advantages of using such systems are to inspire participants with new ideas and additional perspectives, to reduce information distortion or to provide participants with the ability to check facts and reduce communication ambiguities quickly [KGF02, SWQ<sup>+</sup>17, ZLPCT18]. Moreover, these systems act as a memory aid and participants remembered new tasks and details more easily [SBB<sup>+</sup>18]. Other systems assist keeping the focus on meeting relevant issues in order to avoid talking about irrelevant topics [CBS<sup>+</sup>19].



# Approach to Automatically Extract Meeting Quality Features

To extract relevant data of the gathered raw audio recordings of the meetings, we developed an approach for this purpose. In this chapter, we first discuss existing open-source libraries for different speech analysis tasks. Then, we provide deeper insights into the task of identifying who spoke when during a meeting which is called speaker diarization. Lastly, we present the approach we developed and its extractable speech features.

## 3.1 Open-Source Libraries

In order to solve our task to extract various speech features of an audio recording, we searched for existing tools and libraries in the field of speech processing. We were not only looking for systems extracting 'low-level' features such as MFCCs, but also on the extraction of 'high-level' features such as the detection of speaker turns. Since we noticed that many speech processing software is implemented in Python, we provide here an overview of mostly packages in this language. The libraries presented by us in this chapter are only a selection according to our needs and feasibility. Depending on the speech processing task there are many more projects, e.g. for feature extraction. The following sections are grouped by the speech processing task and contain a small overview of some systems.

### 3.1.1 Low-Level Feature Extraction

Low-level features usually contain information relevant for pattern recognition and speech processing tasks because extracted features are related to characteristics of the vocal tract of the speaker, such as MFCC or intensity. Therefore, the extraction of low-level features usually forms the basis for other speech processing tasks, such as speech recognition. The extraction of audio features takes an audio file as an input and returns a list or file containing the specified features. In order to extract these features, the algorithm splits the audio signal into a specified number of frames, and then, computes the features for each frame.

**Parselmouth.** Parselmouth [JTDB18] is a Python interface for the core functionality of Praat. Praat [Boe02] is a software to analyze, synthesize and manipulate speech. It is developed by the

Institute and Phonetic Sciences of University of Amsterdam released in 1992 and is still maintained and extended. Praat is widely used among phonetics scientists and much research in their field relies on this tool. Praat provides different functions for acoustic analysis, like pitch extraction, which can be accessed in Python using Parselmouth. Parselmouth is available on Github and is actively maintained.

**OpenSmile.** OpenSmile [EWGS13] is a toolkit for audio feature extraction and pattern recognition written in C++. OpenSmile runs by using a configuration file which defines the features to extract. This configuration file is written in its own language, not C++. However, some configuration files are directly included when downloading OpenSmile. They provide functions for the extraction of common features, such as MFCC.

**Librosa, Yaafé and SpeechPy.** Librosa [MRL<sup>+</sup>15], Speechpy [Tor18] and Yaafé [MEF<sup>+</sup>10] are Python packages to extract audio features in Python for automatic speech recognition or speaker recognition applications. These libraries are independent of each other but offer the same basic functionality to extract features of an audio file. They differ in various additional functions, different performances or the scope of their documentation.

### 3.1.2 Voice Activity Detection

Voice activity detection, also known as speech activity detection, splits the audio file into segments containing speech and non-speech. Non-speech segments could contain, for example background noise, silence or coughing. The main approaches to solve this task can be divided into three categories: model-based, energy-based and hybrid [Yel15]. The model-based detectors classify segments as speech or non-speech using machine learning and a model that is trained on labelled data. Energy-based detectors decide based on a threshold of the intensity of the signal if the segment contains speech. The hybrid approach is a mixture of the two other approaches. Several libraries are available to solve this task.

**AUDIo TOKEnizer.** AUDIo TOKEnizer [Ami19] is an energy-based audio activity detection tool that can process data from a raw audio file or in real-time from microphone input. It can be used as a command line application or as a Python library. The user must set the detection threshold (in dB). This value can vary depending on the recording. AUDIo TOKEnizer provides several adjustable parameters, including the minimum or maximum length a voiced signal.

**Py-webrtcvad.** Py-webrtcvad [Wis19] is a Python interface to Google's WebRTC Voice Activity Detector. Googles' VAD [BBC<sup>+</sup>19] is a current state-of-the-art voice activity detection based on a pre-trained model to classify speech and non-speech. Therefore, it is a model-based approach.

### 3.1.3 Gender Detection

Gender detection of speech streams seems to be an easily solvable problem because female and male speech have different characteristics: Female speech is generally associated to a higher pitch, to be more breathy and to have vowel formants located in higher frequencies than male speech [DCV<sup>+</sup>18]. Therefore, a basic approach could be to identify the gender for each segment based on its pitch values [DCV<sup>+</sup>18]. However, this represents not the most robust approach, as speakers could have strong accents or extreme pitch ranges [DCV<sup>+</sup>18].

**Ina Speech Segmenter.** Doukhan *et al.* [DCV<sup>+</sup>18] open-sourced their gender detection system which splits the audio signal in either segments containing music or speech. The speech segments are tagged with speaker's gender (male or female). Their toolkit is based on a Convolutional Neural Network (CNN). The published model was trained on 2'284 French speakers which achieved an accuracy of 97.42%.

**LIUM SpkDiarization.** The primary purpose of this toolkit is to identify who speaks when, so-called speaker diarization, which is further described in Section 3.2. In addition to its primary task, the tool identifies the gender of the speaker. This toolkit's gender detection uses a Gaussian Mixture Model trained with a French broadcast news corpus.

### 3.1.4 Overlapping Speech Detection

Overlapping speech detection refers to the task of identifying the segments in a recording where at least two speakers are talking at the same time. The automatic detection of overlapping speech in real conversational speech, e.g. multi-party conversations like meetings, is a tough task, especially using unsupervised learning is one of the hardest problems in automatic speech processing [CDL<sup>+</sup>18]. Despite the efforts to solve and improve this task, its accuracy is still below an acceptable level [SGS<sup>+</sup>18].

It is a challenge to use naturally occurring overlapping speech as training data because for supervised learning the overlaps have to be labelled very precisely which is not possible manually. Therefore, models are often trained on artificially created overlaps by mixtures of two speech sources. Moreover, pre-trained models might not work in another domain or on a different language, especially not if it is trained on artificial data.

Detection of overlapping speech would not only be very interesting in terms of group dynamics within meetings, but it can also significantly improve the accuracy of other speech processing tasks. For example, speaker diarization showed a significantly better performance if overlapping speech segments were excluded [CBL13]. After all, we could not find a working open-source solution. Therefore, we implemented a simple energy-based algorithm, with the assumption that segments containing the overlapping speech are louder than the rest of the meeting. However, we did not achieve the wanted results and decided to not include overlap detection in our final pipeline.

An alternative to overlapping speech detection, could be blind source separation. Blind source separation detects segments with more than one speaker and tries to separate the audio segments in multiple audio recordings each containing the speech of one person. We could use the first step of a blind source separation algorithm in which the overlapping speech segments are identified and extract these segments. An implementation of such a system can be found on Github [DeW19] with a publication of a pre-trained model using five speakers. The pre-trained model was not able to produce usable results with our data and we were not able to train our own model.

### 3.1.5 Laughter Detection

Ryokai *et al.* [RDLH<sup>+</sup>18] open-sourced their machine learning algorithms to detect naturally occurring laughter within a real conversational speech. They implemented a neural network in Python which predicts the probability of an audio frame containing laughter. Aside from the possibility to train an own model, a pre-trained model using the switchboard corpus<sup>1</sup> is available. The neural network takes a raw audio file as an input and returns a list of audio files containing laughter. With the algorithm, the user can set two different parameters: a threshold for probability

---

<sup>1</sup>Telephone Speech Corpus of approx. 260 h of speech

of classifying a frame as laughter, the minimum length in seconds a laugh needs to be in order to be identified as laughter.

### 3.1.6 Emotion Recognition

The human voice can carry a lot of information. It is not only conveyed by the literal content of the speech but also by the speakers tone or emotional state [GGPL17]. These two information channels do not necessarily overlap, for example, if someone is sarcastic and utters a positive statement with a negative stress (e.g. angry tone). This is the reason why a tool to extract the emotions of a meeting should integrate both channels. The literal content of the speech, i.e. the syntactic-semantics, can be extracted using sentiment-analysis [GGPL17]. The other information channel can be analyzed using paralinguistic information such as the extraction of intonation, pitch or loudness [GGPL17]. Koolagudi *et al.* [KMB18] present an overview of currently existing research in this area, the accuracy of emotion detection varies between around 60 % to 90%, whereas the highest accuracy was achieved on a train and test set of actors speaking a sentence with emotional pronunciation according to the specification. These systems typically work for the trained speakers and usually show a lower accuracy if the model would be applied to a new speaker. Therefore, the generalizability of a model highly depends on the audio files the model is trained on, e.g. if a model is trained using only male voices, it is harder to predict emotions for female speech, the same goes for languages. Garcia-Garcia *et al.* [GGPL17] tested two different services for automatic emotion detection, one of which was an open-source solution, but produced poor results with many miss-classified audio files. In the end, we could not find a solution with reasonably good results in real-life scenarios and it seems like this task has not been solved yet. Therefore, we were not able to detect emotions of the speech. We found a possibility to extract sentiment of the recognized German speech [RQH10], e.g. syntactic-semantics of what the speaker is literally saying, but due to time constraints, we could not implement it.

**Speech Emotion Analyzer.** Speech Emotion Analyzer [Mit] is a neural network which is capable of detecting five different emotions (angry, calm, fearful, happy and sad) from speech recordings. The developers published a pre-trained model using around 2000 audio files which are recordings of actors speaking a particular emotion in North American English or British English. They claimed accuracy of more than 70% on their test files. We decided to use this trained model and give it a try even though it is trained in a different language. We recorded short audio files from us speaking emotional, e.g. angry, happy. Unfortunately, the emotions of all our test cases were wrongly identified. Therefore, we decided not to use this implementation for our thesis.

### 3.1.7 Speech Recognition

Speech recognition is the translation of spoken language to text from an audio signal, also known as speech to text. There exist some commercial (online) services for speech recognition. However, online speech recognition raises security and privacy concerns. Therefore, we focused on offline speech recognition tools which are running locally. In our research, we found three open-source APIs which we present in this section. The three presented tools use state-of-the-art machine learning algorithms based on at least an acoustic model. The acoustic model maps audio signals and phonemes (the speech sounds in a language). The freely available, pretrained models for these APIs were usually trained with clean data, a sparse vocabulary or on English data sets. Therefore, their use is limited and the user might train an own model. To train a model two main things need to be available. Firstly, a large amount of (labelled) language data needs to be accessible. Secondly, the training should run on a server with enough memory and several CPU cores, preferably with a GPU, otherwise, the training takes forever or is not even possible.

**DeepSpeech.** DeepSpeech [Moz19] is an automatic speech recognition engine implemented by Mozilla. Mozilla recognized the issue that a large amount of labelled data is not accessible to everyone and this hampers the development of open-source software in this area. Therefore, Mozilla published its speech recognition algorithm and pre-trained model on Github to the open-source community. The pre-trained model can be used for the inference of English audio files. DeepSpeech is based on Python.

**Sphinx 4 and PocketSphinx.** Sphinx 4 and PocketSphinx are speech recognition systems written in Java and C, respectively. Both systems support US English, German and many other languages. Sphinx 4 and PocketSphinx are based on two essential models: the acoustic model and the language model. The language model describes the grammar, e.g. probability of word X being followed by word Y. Both models can be generated using other applications from CMUSphinx which is an open-source group who are the developers of Sphinx 4, PocketSphinx and more speech recognition tools. PocketSphinx is a lightweight version of Sphinx 4 with the goal to run on a smartphone and comes with default models trained in English.

**Kaldi.** Kaldi is another tool for speaker-independent automatic speech recognition. Kaldi uses an acoustic model in addition to a language model and a phoneme lexicon. Milde and Köhn [MK18, RAML<sup>+</sup>15] provide trained models in German. They steadily re-train them using more data and different microphones, leading to more generic use cases. Their latest model, published on 5 March 2019 on Github, is trained on 630h of speech consisting of three distinct corpora. Unfortunately, we were only able to get Kaldi up and running with a German solution just before we analyzed the data. Therefore, we were not able to integrate Kaldi into our pipeline in time.

## 3.2 Speaker Diarization

Based on existing literature, we found that the analysis of turn-taking behavior might be related to the meeting quality. Therefore, we want to automatically identify in an audio file who spoke when. To address this task, Speaker Diarization Systems can be used. These systems answer the question "who spoke when" by splitting an audio file into segments containing speech and determining which segments are spoken by the same speaker (represented in Figure 3.1) [Yel15].

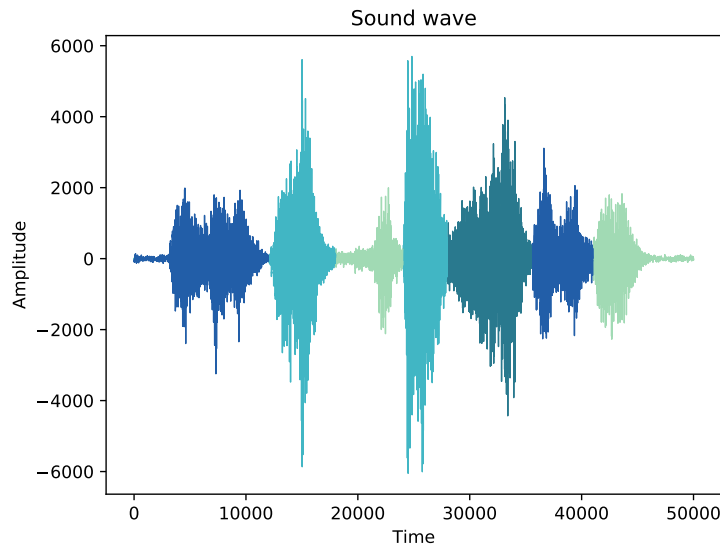
In this section, we provide an overview of how speaker diarization works, its challenges, as well as a short description of six open-source speaker diarization systems. We conclude this section with a comparison of the six systems.

### Diarization Pipeline

In general, a speaker diarization system needs solely the raw audio recording as an input. Besides, some clustering algorithms which take the number of speakers as an additional input.

Speaker diarization consists of five main steps:

1. Audio Preprocessing
2. Feature Extraction
3. Voice Activity Detection
4. Speaker Change Detection
5. Speaker Clustering



**Figure 3.1:** Segmentation of audio recording in its speakers; same speakers are identified by the same color

**Audio Preprocessing.** As a first step, the raw audio signal is filtered to reduce background noise or humming. For this, either a wiener filter, band-pass filter or another noise reduction method can be applied [Yel15]. The output of this step is a clean, filtered audio file.

**Feature Extraction.** The goal of the second step is to extract a set of features from the audio signal. These features enable the system to differentiate between speakers and all subsequent steps depend on these extracted features. There are different features which could be extracted, e.g. Mel-frequency cepstral coefficients (MFCC). Therefore, the output of this step is a set of extracted features which are used as inputs for the subsequent steps.

**Voice Activity Detection.** Voice activity detection splits the audio file into speech and non-speech segments. The non-speech segments are not further analyzed because they do not contain valuable information. If they are not removed, the error rate could increase [Yel15]. The output of this step is a text file containing speech segments, e.g. segment id, start time and length.

**Speaker Change Detection.** In this step, the previously identified speech segments are further split into homogeneous segments with the goal to identify speaker change points. These are timestamps in which the speaker before is different from the speaker after the timestamp. A common approach to identify segments containing speaker changes is hypothesis testing which analyzes the probability of two segments corresponding to the same speaker. To test this, a window shifts over all segments and calculates on each frame a criterion based similarity/distance metric which is calculated on an estimated model over the speech data [Yel15]. Then, the metric is compared to a threshold indicating whether it contains a change point or not [Yel15]. A widely used criterion is the Bayesian Information Criterion. The output of this step is a text file containing speech segments split into smaller segments which correspond to timestamps of speaker changes.



**Speaker Clustering.** This step clusters segments which contain similar information. There are two different approaches for this: bottom-up or top-down. Most of the current state-of-the-art speaker diarization systems use bottom-up approaches [Yel15] which are also known as agglomerative clustering. They start with too many clusters which are initialized randomly, e.g. each segment is initialized to its own cluster. Then, clusters are iteratively compared to each other and the closest pair is merged until one cluster per speaker is remaining [ABE<sup>+</sup>12]. In contrast, top-down or hierarchical clustering works the other way around. It starts with having only one cluster which is recursively split and merged until a threshold (e.g. number of speakers, similarity measure) is reached [ABE<sup>+</sup>12]. [ABE<sup>+</sup>12]. The output of this step is a text file of speech segments, each labelled with the speaker id of the corresponding speaker uttering it. An example extract is shown in Table 3.1.

Meeting id	Start time (in centiseconds)	Length (in centiseconds)	Speaker id
10001	1	299	S55
10001	305	59	S55

**Table 3.1:** Example of Speaker Diarization Output: Speech-Segments labelled with Speaker Id

## Challenges of Speaker Diarization

There are some difficulties for diarization systems that are not trivial to overcome and lower its accuracy. Meetings are characterized by fully-spontaneous and highly interactive conversations between different participants which might result in short segments of speech or in segments containing overlapping speech. For the diarization system it is difficult to identify short or overlapping segments correctly. Additionally, existing diarization systems are often trained and evaluated on broadcast news data. However, this type of data shows characteristics that are different to a meeting room environment which is usually exhibit to background noise, reverberation or speakers changing their position [IFM<sup>+</sup>05]. Another difficulty might be if speakers show emotional changes within a meeting because changing the tone of voice is mediated by variation of prosodic features. Due to the fact, that speaker diarization systems work through the analysis of extracted features it might be difficult to link various emotions to the same speaker. For example, a speaker becoming angry and raising his voice, it could lead to a different representation of his extracted prosodic features. Lastly, we noticed that it is challenging to distinguish between similar voices, e.g. having three young female attendees within one audio file.

## Overview Open-Source Speaker Diarization Systems

To perform speaker diarization we looked at six open-source libraries, which we will shortly introduce. The systems perform all steps of the speaker diarization pipeline resulting in an output file which contains the segments of speech labelled with a speaker id.

**PyAudioAnalysis.** PyAudioAnalysis [Gia15] is a Python library with a wide range of functionality for audio analysis, such as feature extraction, speech recognition, audio segmentation or speaker diarization. PyAudioAnalysis solves speaker diarization in an unsupervised manner. Their implementation needs a trained model, which is already published. However, there is no information on the underlying data used for training the model nor how one could retrain it. The

user can perform speaker diarization by merely calling one method which automatically performs all necessary steps, e.g. feature extraction, voice activity detection. We used the version 0.2.5 which is published on Python Package Index (PyPi). The user is able to define the number of clusters by giving the number of speakers as an input.

**Pyannote.Audio.** Pyannote.Audio is a project in Python providing modules to perform feature extraction, voice activity detection, speaker change detection and speaker embeddings. The user can run the modules separately or sequentially. By running the modules successively a full speaker diarization can be performed. We used the most recently released version 1.0 of Pyannote.Audio [YBB18]. This system uses supervised machine learning algorithms to perform speaker diarization. Therefore, users need to train their own model using labelled data. Pyannote.Audio already provides the necessary implementations to train its own model but not a pre-trained model. The above mentioned modules, except of the feature extraction, need a model to run and perform its task. We trained them using audio recordings of the AMI meeting corpus. At the time of writing this thesis, the developers of Pyannote.Audio were implementing a new release with promising results. However, this new version is still under development and not officially released as of now. Therefore, we were not able to use this new version. However, it seems to be a very good solution for the future, since it is highly maintained and the developers are renowned researchers in this field.

**LIUM Speaker Diarization.** LIUM Speaker Diarization (LIUM) [MM10] is a toolkit to perform speaker diarization. It is written in Java and runs as a command line tool. We used the version 8.4.1. LIUM uses trained models for speech detection and speaker clustering. They are trained on French broadcast data. LIUM was released in 2010 and was the first open-source software for this purpose with reasonable accuracy (10.8% of DER on broadcast news data [MM10]). It has been seen as a state-of-the-art system for a few years and is often used as a baseline for newer algorithms [CTS17].

**Sidekit for Diarization.** Sidekit for Diarization (S4D) [BDL<sup>+</sup>18] is a collection of state-of-the-art components to develop a diarization system. It is a python package and we used version 0.0.2 published on PyPi. In comparison to the other diarization systems we used, S4D needs much more programming to perform the speaker diarization task. Our implementation used the following tools provided by S4D: First, we used the S4D's voice activity detection to identify speech segments. Then, these segments are split into smaller segments using Gaussian Divergence criterion. Then, linear clustering merges the adjacent segments assuming to be of the same speaker (1), followed by a global hierarchical agglomerative clustering to merge speakers over the entire file (2). The resulting clusters are resegmented using I-Vectors (3) and Viterbi (4). However, steps (1), (2), (3) and (4) require a threshold as a stopping criterion which the user should define.

**IBDiarization.** IBDiarization [Idi19a] is a toolkit running on Linux and implemented in C++. There exists a Python wrapper [Idi19b] around IBDiarization which enables us to call functions of the toolkit in a Python application. Unfortunately, there is no paper or documentation that explains this toolkit in more detail. However, it is published on GitHub by Idiap Research Institute, a semi-private non-profit research institute focusing, among others, on speech processing and speaker diarization. IBDiarization enables the user to input the number of speakers of the audio file.

**Aalto Speaker Diarization.** Aalto speaker diarization (Aalto) is implemented in Python as part of a master thesis [MO14]. Aalto Speaker Diarization is built on top of Aalto's Automatic Speech

recognition system (AaltoASR) and uses AaltoASR's feature extraction as well as its voice activity system. Aalto provides the possibility to give the number of speakers as an input.

### Evaluation Metric: Diarization Error Rate (DER)

To compare different speaker diarization systems and to evaluate their performance, the de facto standard metric, called diarization error rate (DER), can be used. DER is calculated with the following formula [Bre17]:

$$DER = \frac{\text{false alarm} + \text{missed detection} + \text{confusion}}{\text{total}} \quad (3.1)$$

In formula 3.1 *false alarm* is the duration of non-speech incorrectly classified as speech, *missed detection* is the duration of speech incorrectly classified as non-speech, *confusion* is the duration of miss-identified speakers, and *total* is the total duration of speech in the reference. The *reference* is the labelled data towards which we compare our diarization result. In our case, the reference was the timestamps and speaker ids of the annotations provided by the corpora. The *hypothesis* is the output of the speaker diarization system. Usually, it is typical to tolerate an error of 250ms before and after a speaker transition for the calculation of the DER [Bre17].

The Python library Pyannote.Metrics [Bre17] provides efficient implementations of several metrics, including DER, to compare speaker diarization systems.

### Comparison of Speaker Diarization Systems

To decide which of the six open-source diarization systems we want to use for our approach, we compared their performance. For this comparison, we used the raw audio recordings from the AMI and ICSI Meeting Corpus. All proposed systems take an audio file of 16kHz as input. In this section, we start by reporting our experiences using the frameworks followed by our results of the evaluation.

Unfortunately, we had some issues with S4D because the framework returned an error message while processing 19 ICSI audio files. We were not able to solve the error and were forced to exclude these files from the analysis. Moreover, S4D relies on the specification of four parameters which the user needs to define before diarization. Since we did not know which threshold would produce the best results, we implemented a range of thresholds and ran the data for every combination. Then, we chose thresholds resulting in a low average diarization error rate over all inputs. Moreover, we noticed that Aalto took an unusual amount of time to process some files, sometimes it took more than three times as long as the recording. This performance issue could depend on our used hardware, however, the differences in processing time between some files were considerable. For Pyannote.Audio we were not able to diarize other files than the data of the AMI corpus. That is the reason why we left out the diarization of the ICSI data.

For all systems, we calculated the diarization error rate with a tolerance of 250ms using the Python library Pyannote.Metrics. We excluded the files containing an error rate exceeding 100% since they would skew our results. Table 3.2 contains an overview of our resulting DER rates. One can see, that on average the error rates are quite high for both corpora. The top three systems with the lowest average diarization error rates were S4D, Pyannote.Audio and IBDiarization. Therefore, we decided to have a closer look at IBDiarization and S4D. As a next step, we performed diarization with both systems on our test data. We used recordings which we captured with our

audio recording devices used for the field study as test data. S4D's results of this processing step, were not consistent because for some files it worked perfectly, and for others, it did not work at all so we could not use the result. IBDiarization worked uniformly very well on our test data. Due to these results, we decided to choose IBDiarization as our framework to perform the task of who was speaking when.

System	DER on AMI Corpus				DER on ICSI Corpus			
	<i>M</i>	<i>SD</i>	<i>MIN</i>	<i>MAX</i>	<i>M</i>	<i>SD</i>	<i>MIN</i>	<i>MAX</i>
PyAudioAnalysis	73.2	13.6	42.3	98.2	69.9	12.0	44.8	98.5
Pyannote.Audio	42.6	12.3	11.6	77.1	—	—	—	—
LIUM Speaker Diarization	55.0	20.0	20.5	99.6	55.5	19.8	19.7	99.1
S4D <sup>a</sup>	44.2	20.7	9.1	95.9	42.2	19.1	17.6	96.7
IBDiarization	60.1	20.9	19.2	99.5	49.6	17.7	27.6	93.9
Aalto	59.7	19.4	20.3	99.5	57.0	16.9	27.0	99.8

<sup>a</sup>. Used Thresholds: Linear Clustering=1.5, Agglomerative Clustering=2.73, I-Vectors=20, Viterbi=-250

**Table 3.2:** Diarization Error Rates of Speaker Diarization Systems (*M*=Mean, *SD*=Standard deviation, *Min*=Minimum, *Max*=Maximum of DER values)

### 3.3 Our Approach to Automatically Extract Features

To analyze the relationship between participants perceived meeting quality and features extractable from speech, we build a tool to automatically extract a set of features of an audio recording. This approach integrates audio preparation, speaker diarization, laughter detection, and prosodic feature extraction. Its functionality and the extracted features are described in more detail in this section.

#### 3.3.1 Implementation

This approach is intended for processing one audio file at a time. For each recording, all the steps described below need to be performed in order to extract and later analyze the speech features.

**Audio Preprocessing.** The first step consists of the audio preprocessing. The tool takes a raw audio file as an input, regardless of its sampling size. In case the audio file is not already 16kHz, it will be downsampled to 16kHz using the library Librosa. The reason for this downsampling is that our used signal processing approaches expect an audio file of this sample size. Moreover, a stereo recording is converted to mono. Then, we removed constant background noise such as hum. To perform noise reduction, an audio file which contains only this noise is necessary. We created a noise file for each meeting room. In order to automatically remove background noise, we implemented an automatic approach using SoX [eXc19]. SoX is a command line tool to apply effects on sound files, such as a noise filter. As a next step, we applied a low-pass filter with a cut-off of 700 Hz to remove high-frequency noises, and a high-pass filter with a cut-off of 50 Hz to remove low-frequency noises.

The output of this audio preprocessing step is a filtered and noise reduced audio file, downsampled to 16kHz.

**Laughter Detection.** To capture the laughter within a meeting, we used the laughter detection described in 3.1.5. We set the probability threshold to 0.8 and the minimum length to 0.1 which the laugh segments need to fulfill in order to be identified. We processed each downsampled, filtered audio recording of a meeting. The system returned short audio segments containing the detected laughs. We listened to the audio files and decided whether they contained real laughs or trash. Some meetings contained many segments labelled as laughs, but were actually trash. Therefore, we manually changed these segments as non-laughs. So, we cleaned the laughs to not have any false positives, i.e. labelled as laugh but containing trash. However, we have no metric about the amount of false negatives, i.e. amount of laughs in a meeting which are not detected as laughs. We performed this manual cleaning step, in order to have a more accurate input for our empirical data analysis.

The output of this step, is the number of laughs for each meeting.

**Speaker Diarization.** To detect for each meeting 'who was speaking when', we used the Python wrapper for IBDiarization as a diarization system. Again, the input was a 16kHz, filtered audio file for each meeting.

The output for each meeting is a textfile containing segments of speech labelled with a speaker id. An example is shown in Table 3.1.

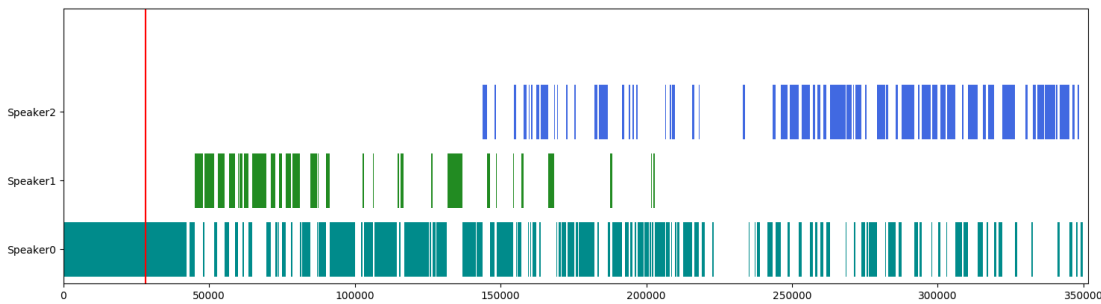
**Voice Activity Detection.** We run pyWebRTC for all meetings to detect speech and non-speech segments. Then, the speech segments are further split into smaller segments using a segmentation approach of the package S4D. Technically, the used speaker diarization system has a voice activity detection (VAD) system integrated. However, the diarization with this system on our data resulted in too much non-speech classified as speech. Therefore, we used pyWebRTC as our VAD. Since the VAD system of IBDiarization is not easily replaceable, we run both independently and merge the outputs of these two systems automatically. We achieve this by resegmenting the diarization output according to the identified segments of our VAD system. The resulting diarization output is more accurate using this approach which improves the performance of all the succeeding steps in our pipeline. The output of this step is a diarization file, which is resegmented and contains less non-speech segments.

**Diarization Editor.** As shown in Chapter 3.2, diarization systems are not very precise and show a fairly high diarization error rate. Thus, we did not want to rely on the output of the diarization system entirely, because it could be too erroneous, and skew the results of the data analysis. Therefore, we decided to review the results and adjust them manually if necessary. Although we aim to implement an automatic approach to extract features of the audio file, we still wanted to make sure that the diarization output was of satisfactory quality. If future systems are more precise, this step can be omitted.

For each meeting, we performed the following steps. First, we created an individual audio file for each speaker which contains only his/her speech. After listening to multiple short segments of each speaker, we decided whether most of them are classified correctly. If not, we decided to change some speaker labels manually. For this purpose, a tool to visualize the diarization output and modify speaker ids while listening to the recording was used. A screenshot of this tool is shown in Figure 3.2. Originally, this tool was implemented by Max Hollmann [Max19] and we slightly adjusted it to fulfill our needs. The tool requires two inputs: the output of the diarization step and the corresponding audio recording of the entire meeting. The diarization is visualized

by having the time on the x-axis and the different speaker ids on the y-axis (refer to Figure 3.2). Additionally, the audio can be played, stopped, fast-forwarded or rewound. The current position of the recording is dynamically represented using a red line. Therefore, we always knew the speaker id of the speaker we are currently hearing. The tool allows changing the speaker label of the current segment (i.e. current position in the audio recording) or of all segments labelled with this speaker id. At any point in time, the new diarization can be saved in a text file.

For most meetings, the labels were already correct, and we did not apply this step. For some meetings, especially if there were many attendees (more than four) or similar voices (all females), this step helped us to increase the accuracy of the diarization output. Nevertheless, we did not listen to the entire meeting recordings. Instead, we skimmed the audio files and changed what we noticed.



**Figure 3.2:** Screenshot of Diarization Editor

**Prosodic Feature Extraction.** We extracted the prosodic features using the Python interface Parselmouth to the software Praat, as described in Section 3.1.1. We extracted all prosodic features on utterance level. For this we merged all consecutive segments of the same speaker with a shorter silence than 500 ms [SSB01, LM18]. We split the filtered audio file of each meeting into files corresponding to speakers' utterances. Then, over all audio files which contain only the utterance, we extracted the prosodic features. The intensity and pitch values are extracted for every 10ms (frame length) with a window size of 40ms. Pitch values are extracted using the auto-correlation method, which is one of best algorithms for this purpose and balances complexity, accuracy and robustness [Boe93]. To extract the speech rate of each speaker, we used an existing script executable in Praat. More information about its algorithms and a validation can be found in [DJW09].

### 3.3.2 Extracted Speech Features

Using the approaches described above we were able to compute several speech features, which we categorized into individual-level and group-level features. The former is extracted for each individual speaker, whereas the latter is an aggregated value over all speakers within one meeting and is separately calculated for each meeting. All features are based upon the outputs of our approach which is described above. To calculate them we used the libraries Pandas and NetworkX in Python [McK11, HSS08]. Related work calculated the features on so-called spurts, i.e. continuous speech segments of the same speaker separated by at least 500ms silence [LCR13]. Therefore, we merged continuous segments of the same speaker which are not separated by non-speech segments

longer than 500ms on which we calculated the speech features. The values depending on the meeting duration are normalized by dividing them by the meeting duration. If we refer to speech features in the remaining thesis we relate to these extracted features describe in the following.

We extracted this set of features, based on findings of related work, i.e. in [LCR13, LM18, JBOGP08, Jay11, JSCO<sup>+</sup>12]. The features are categorized into general features, turn-taking features, centrality features and prosodic features. General features include universal meeting characteristics, including meeting duration. Turn-taking features encompass features related to group dynamics and participants' turn-taking behaviour, such as the number of turns or total speaking length. The centrality features reflect to which extent the meeting was centered on a single attendee. At least one of these centrality features, i.e. total speaking length per person, is related to the dominance status of the speaker within the meeting [Mas02, Jay11]. To calculate the eigenvector and degree centrality features, we represented the speaker changes as a multi-edged graph while the nodes are the participants and the directed edges represent the changes from one speaker to another. The prosodic features refer to properties of the speech, e.g. the fundamental frequency, which are only calculated on the individual-level.

## Group-Level Features

The following list represents a description of the extracted features and the calculations behind the values.

### General Features:

- Duration of Meeting (in centiseconds)
- Total Amount of Laughs:  
This feature equals the sum of all laughs within a meeting divided by the meeting duration.

### Turn-Taking Features:

- Total Speaking Length:  
Speaking length is the duration that a person is speaking, total speaking length defines the sum of speaking length of all participants divided by the meeting duration.
- Total Speaking Turns:  
A speaking turn is the continuous period of time a person is speaking separated by at least 500ms silence [SSB01, LM18]. Total speaking turn is the sum of speaking turns of all participants divided by the meeting duration.
- Total Silence Length:  
This feature equals the length of non-speech within a meeting, and is calculated by subtracting speaking length from meeting duration and dividing by the meeting duration.
- Average Turn Duration:  
Average duration of speaking turns is the ratio of total speaking length and total speaking turns.
- Total Number of Speaker Changes:  
A speaker change occurs if adjacent speech-segments are uttered by different speakers. Total number of speaker changes equals the sum of all speaker changes divided by meeting duration.

- **Turn-Taking Freedom:**  
Turn-taking freedom  $F_{\text{cond}}$  is calculated using this formula [LCR13]:

$$F_{\text{cond}} = 1 - \frac{H_{\text{max}}(Y|X) - H(Y|X)}{H_{\text{max}}(Y|X)} \quad (3.2)$$

where  $H(Y|X)$  is the conditional entropy of speaker Y is the next participant, given speaker X is currently speaking, and  $H_{\text{max}}(Y|X)$  is the maximum possible value for groups of a certain size. Values closer to 1 indicate a free speaking order, whereas values closer to 0 represent a strict turn-taking order, e.g. if speaker X always follows after speaker Y.

#### Centrality Features:

- **Participation Equality:**  
This feature is calculated using the following formula [LCR13]:

$$P_{\text{eq}} = 1 - \frac{\sum_i^N (T_i - T)^2 / T}{E} \quad (3.3)$$

where N equals the number of attendees,  $T_i$  the speaking length for participant i, T the equal participation which is defined as  $\sum_i^N (T_i) / N$ , and E the maximum possible value of the term under the sum (i.e. when for the entire meeting only one participant is speaking). Values closer to 0 show imbalanced participation, and values closer to 1 show greater equality.

- **Maximum Percentage of Speaking Length of Speaker with Longest Speaking Duration:**  
We calculated each participant's total speaking length divided by the total speaking length of all participants combined. The speaker with the highest ratio was chosen. This feature measures participant's speaking length relative to the speaking length of the other attendees.
- **Maximum Percentage of Speaking Turns of Speaker with Highest Amount of Speaking Turns:**  
We calculated each participant's total speaking turns divided by the total speaking turns of all participants combined. The speaker with the highest ratio was chosen. This feature measures participant's speaking turns relative to the speaking turns of the other attendees.
- **Maximum Degree Centrality:**  
Equals the degree centrality of the participant having the highest degree centrality. Degree centrality measures how many times an attendee responds or addresses other attendees [SK13].
- **Maximum Eigenvector Centrality:**  
Equals the eigenvector centrality of the participant having the highest eigenvector centrality. Eigenvector centrality measures not only how many times an attendee responds or addresses other attendees but also take into account the degrees of the addressees or respondents.

#### Individual-Level Features

For the individual level we extracted the following features for each participant per meeting.

##### Turn-Taking Features:

- **Total Speaking Turns:**  
Total speaking turns of the participant divided by the meeting duration



- Speaking Length:  
Speaking length of the participant divided by the meeting duration
  - Total Speaking Length
  - Average Speaking Length
  - Standard Deviation Speaking Length
  - Minimum Speaking Length
  - Maximum Speaking Length

**Centrality Features:**

- Participant's Percentage of Speaking Length of the Total Speaking Length of all Participants:  
Participant's total speaking length divided by the total speaking length of all participants.
- Participant's Percentage of Speaking Turns of the Total Speaking Turns of all Participants:  
Participants' total of speaking turns divided by the total speaking turns of all participants.
- Degree centrality of the participant
- Eigenvector centrality of the participant

**Prosodic Features:**

- Mean fundamental frequency (in Hz)
- Standard deviation of fundamental frequency (in Hz)
- Range of fundamental frequency (in Hz):  
We subtracted minimum from maximum frequency
- Mean intensity (in dB)
- Standard deviation of intensity (in dB)
- Range of intensity (in dB):  
We subtracted minimum from maximum intensity to assess degree of variation in speech loudness
- Local Jitter (in %)
- Local Shimmer (in %)
- Speech rate:  
Number of syllables in each segment divided by total segment duration.
- Articulation rate:  
Number of syllables in each segment divided by total speaking time of this segment.



# Study Method: Field Study

To examine which factors extractable from audio recordings are related to the participant's perception of the quality of the meeting, we designed and conducted a field study with professionals in a company based in Germany and collected the necessary data. On the one hand, we gathered the raw audio recording of meetings and on the other hand the responses of post-meeting surveys of meeting participants attending the recorded meetings. When we use the term recording within this thesis, we refer to audio recordings and not to video recordings. We will describe the participants, our procedure, collected data and insights of the empirical analysis in the following sections.

## 4.1 Participants

Through personal contacts, we were able to conduct the study in a company based in Germany having around 60 employees. The participants of our study were approximately 30 professionals working in this company. The employees could attend multiple times in this study if they had more than one meeting during the study period. We set a limit of a maximum of 20 survey submissions per participant to prevent skewed data. The participants received a reimbursement in the form of an Amazon gift card between EUR 10 and EUR 25 depending on their number of submitted surveys. The study participants average professional work experience within their role was 7.6 ( $\pm 6.0$ ) years, ranging from 0 to 20 years. The employee role of 36 attendees (46.2%) corresponds to an individual contributor. They showed an average work experience of 4.5 ( $\pm 4.2$ ) years, ranging from 0 to 20 years. The remaining 42 attendees (53.8%) are best described as leaders or managers with an average work experience of 10.3 ( $\pm 6.0$ ) years, ranging between 0.5 and 20 years.

## 4.2 Method

Our study was conducted for five working days. During this period, all meetings held in the company and which were allowed to be recorded were part of the study. At the beginning of the study week, we shortly introduced all employees about our procedure to conduct the study. We told them that we would place audio recording devices in three meeting rooms. We asked them, to manually record the audio of the meeting and to fill out a short survey when the meeting is over. They could decide freely whether they want to participate and to answer the survey. We designed this self-administered survey concise with a completion time of approximately 1 - 3 minutes to increase the submission rate. The participants' answers were anonymized entirely and we cannot

attribute the answers to individual employees. Additionally, we informed them that a researcher of us would be in the office for the entire study week to monitor the study and data collection.

After this introduction, we placed the recording devices and the surveys in three meeting rooms and left them ready for the entire week. During the first day of study conduct, we briefed all attendees before each meeting to manually start and stop the recording at the beginning and end of the meeting. When the meeting was over, we asked them to fill out the surveys. Additionally, next to the recording devices we have put a small note on the table with the request to record the meeting and to hand in the questionnaires. After the first working day, the employees autonomously recorded the meeting and submitted the surveys at a high rate. Only meetings where all participants permitted to be recorded, were part of the study. Each attendee was allowed to pause or stop the recording at any time. We used two different devices to record the audio depending on the meeting room size. For the smallest meeting room (usually used for one-on-one meetings) we used the recorder *Zoom H5* and for the two larger meeting rooms the recorder *Zoom H1n* with two extended omnidirectional microphones of the type *Olympus ME-33*. We decided to use an omnidirectional microphone since we wanted to capture the audio unobtrusively.

To understand how the participants perceived the quality of the just attended meeting, we created a survey of 14 questions: two questions about participant's perceived meeting quality, three questions about presence of meeting aspects and their contribution to the meeting quality, one question about participant's emotional state, six questions about aspects of the meeting and its quality, five questions about the meeting structure, and three questions about the background of the participant. The participants could answer the survey either using the online or paper version. We implemented the online survey on LimeSurvey [Gmb] which was hosted on a server of the Department of Informatics at University of Zurich. An example question is shown in Table 4.1 and the complete survey can be found in Appendix B. All questions were voluntary and the participant was allowed to drop out at any time or not to answer all questions. Out of the 14 questions, five were open-ended, while nine questions had a closed set of answers. We decided to use the Likert items as response format for questions 1, 2 and 3. Because it is generally regarded as a construct to measure opinions, attitudes or feelings [AS07]. Question 6 is related to participants emotions. Since participants' emotional state not only could influence his perception of satisfaction but also on acoustic features, we decided to assess it using the standardized Self-Assessment Manikin (SAM) [BL94] on a pictographic five-point scale. SAM can be used to assess emotions in three independent dimensions: valence, arousal and dominance. The images used in the survey were taken from [BV16].

Question-Nr	Question	Subquestion
Q1	In general, how satisfied have you been with this meeting? ( <i>not satisfied at all - not satisfied - neutral - satisfied - extremely satisfied</i> )	
Q4	Did any of the following aspects affect your perception of your time spent in this meeting? ( <i>no, did not affect - yes, negatively affected - yes, positively affected - not applicable</i> )	2.2 Interruptions between members.  2.5 Positive and open meeting atmosphere.

**Table 4.1:** Sample Survey Questions

## 4.3 Data

During this week we recorded around 17 hours and 40 minutes of raw audio meeting data. We captured 25 meetings, with an average duration of approximately 40 ( $\pm 25$ ) minutes. The shortest meeting had a duration of 00:15:18 and the longest 01:40:53. We did not capture a meeting in which the recording was paused and restarted again, which would have resulted in an incomplete recording of the meeting. We collected data of meetings having between two and eight attendees: 28% (N=7) of the meetings have two attendees, 32% (N=8) three attendees, 24% (N=6) four attendees and 16% (N=4) more than four attendees. In total, the 25 meetings consisted of 90 meeting participants. Out of these 90 meeting participants 86.67% (N=78) submitted the post-meeting survey. On average, 88.60% ( $\pm 17.0$ ) of all attendees per meeting filled out the survey. The submission rate of 16 meetings was 100%, whereas we had only two meetings with 50% of meeting attendees submitting the survey (both on one-on-one meetings). In Appendix C we provide an overview of the captured meetings. After one week of data collection, we ended up with 78 valid responses from meeting attendees. We had to exclude eight responses because the stated meetings were not recorded or the recording was only a tiny fraction of the entire meeting, and therefore, useless.

We recorded 7 one-on-one meetings (28%) and 18 many-to-many meetings (72%), e.g. brainstorming, discussions. Within three recordings of the many-to-many meetings one participant, respectively, stated to have a one-to-many meeting, e.g. a presentation. Nevertheless, we treated these meetings as many-to-many as well since only the minority mentioned to have another type of meeting.

## 4.4 Analysis

Statistical analysis was performed using Pandas in Python [McK11], dplyr in R [WFHM18] and SPSS Statistics [Cor]. The use of parametric or non-parametric tests was always chosen based on the normality distribution of the underlying variables. We tested for normality using the Shapiro-Wilk test. An alpha level of .05 is chosen for all statistical tests. Before data analysis, we preprocessed the survey responses and extracted speech features, respectively, which is described in the following paragraphs. After the data preprocessing, we mapped the two data sources based on the combination of meeting time, meeting date and meeting room. Then, we analyzed the data on either a individual-level which means the analysis of the answers of individual participants, or on group-level which refers to an aggregated level of the responses of all attendees within a meeting.

**Preprocessing of Survey Data.** We prepared the survey data such that it can be analyzed on either a individual-level or a group-level. Therefore, we coded the answers of questions in the Likert item response format to numbers. The used coding is shown in Table 4.2. Then, to calculate the group-level we summed up the points of all the answers from the participants belonging to the same meeting. Since not all meetings had the same amount of attendees, we divided the aggregated value by the number of attendees, resulting in an end score which we used to compare the answers between the meetings. Therefore, if we relate to the group-level of participants' responses of question 2 we use the term agreement-score for each factor, for answers of question 1 as meeting-satisfaction-score and for answers of question 3 as time-well-spent-score.

To calculate the group-level value for the remaining answers, we summed up the answers of all participants for each category.

For the analysis, we created a mapping between question 2 and 4, for which we refer to using the stated keyword, depicted in Table 4.3. During the next chapters we will refer to these meeting

aspects or meeting factors using the stated keywords in Table 4.3.

**Preprocessing of Audio Data.** The collected raw audio recordings are given as an input to our tool (see Section 3.3.1), and we extracted a set of speech features on a meeting- and individual-level. For an overview of the extracted features, please refer to Section 3.3.2. In the following chapters, we will refer to these extracted values using the term speech features. To increase the quality of the prosodic features, we performed a data cleansing step. First, we removed the rows containing no value for specific features, e.g. null or missing values. Then, we identified data points as outliers if its value is more than  $1.5 \cdot IQR$  above the third quartile or below the first quartile. Due to a misclassified segment, the outlier might contain non-speech instead of speech. Therefore, we decided to remove the outliers. The remaining speech features did not have to be cleaned.

Likert Item	Code
Not satisfied at all	-2
Not satisfied	-1
Neutral	0
Satisfied	1
Extremely satisfied	2
Not well at all	-2
Not well	-1
Neutral	0
Well	1
Extremely well	2
Strongly disagree	-3
Disagree	-2
Slightly disagree	-1
Neither agree nor disagree	0
Slightly agree	1
Agree	2
Strongly agree	3

**Table 4.2:** Coding of Likert Items

Subquestions of Question 2	Subquestions of Question 4	Keyword
Meeting Aspects or Meeting Factors ( <i>Strongly disagree - Strongly agree</i> )	Aspects affecting perception of quality of time spent in the meeting ( <i>No affect, positive affect, negative affect</i> )	
At least one meeting participant was dominant.	Distribution of speech time across members.	Distribution Speech time
The meeting members interrupted each other.	Interruptions between members.	Interruptions
The meeting was useful for me.	The usefulness of this meeting.	Usefulness
My attendance in this meeting was relevant.	Your relevance in this meeting.	Participants' Relevance
The atmosphere in this meeting was positive and open.	Positive and open meeting atmosphere.	Open Atmosphere
The exchange in this meeting was lively and active.	Lively and active exchange in this meeting.	Lively Exchange
There was one or more key people in this meeting.	Presence of one or more key people.	Key people

**Table 4.3:** Mapping of Subquestions Between Question 2 and 4





# Results

We analyzed the collected data of our multi-day study in order to understand participant's meeting affect better and if it could be measured automatically. We present an overview of variables of the captured meetings, such as that the participants were overall satisfied with the meetings.

Then, we describe our findings of factors contributing to the meeting quality, which we categorize in four steps. First, we start by demonstrating which meeting factors influenced participants' meeting quality based on their own statement. Second, since we asked the participants to rate the occurrence of a meeting factor within the meeting, we performed a correlation between the reported meeting factors and the meeting quality. Third, using the extracted speech features, we examined if we can measure the participant's reported meeting factors. Lastly, we performed correlations between the automatically extracted speech features and the participants' perceived meeting quality.

Table 5.1 highlights the main findings of this empirical data analysis and the corresponding sections.

Main finding	Section
Having a positive and open atmosphere affected the quality of the time spent in meetings of most participants positively.	5.2
Meetings having a chairperson resulted in higher perceived meeting satisfaction than meetings without a chairperson.	5.3.2
The relevance of the attendance of the participants or an open, positive meeting atmosphere increases as the meeting quality increases.	5.3.1
The only significant relationship between speech features and meeting quality was measured with a positive correlation between total number of speaking turns and meeting satisfaction.	5.5

**Table 5.1:** Main Findings of Empirical Data Analysis

## 5.1 Overview of Captured Meetings

Overall we captured 25 meetings. Most meetings were considered positive with a group-level satisfaction ranging from satisfied to highly satisfied and a quality or the time spent of the meeting

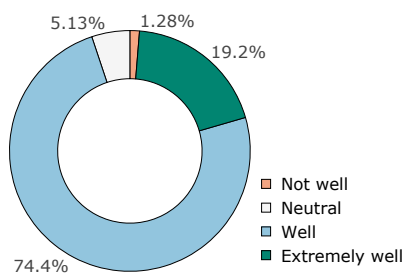
ranging from not well to extremely well while the most mentioned their time was spent well. A summary of means, standard deviations, minimum and maximum values of all group-level variables can be found in Appendix D.

## Meeting Quality

*Individual-Level.* We analyzed the meeting quality regarding two aspects: meeting satisfaction and quality of spent time in meetings. In question 3 of the survey, we asked the participants about the quality of their time spent in meetings. The responses ranged from *not well* to *extremely well*. In general, participants felt their time was spent *well* ( $Median = 1.0$ ,  $Mean = 1.115$ ,  $SD \pm 0.534$ ). Participants answers about their general satisfaction with the meeting (question 1 in survey) showed similar results but in a narrower range without negatively affected meeting satisfactions. Participants mentioned to have *neutral*, *satisfied* or *extremely satisfied* meetings. On average, the participants were *satisfied* with the attended meetings ( $Median = 1.0$ ,  $Mean = 0.923$ ,  $SD \pm 0.449$ ). The Figures 5.1 and 5.2 show the distribution of participants' answers of their perceived quality of their time spent in meetings and their general satisfaction with the meeting, respectively. We performed a Spearman's rank correlation to evaluate the relationship between these two variables which resulted in a statistically significant positive relationship,  $r_s(76) = 0.441$ ,  $p < .001$ . This indicates that the quality of time spent increases as the satisfaction increases.

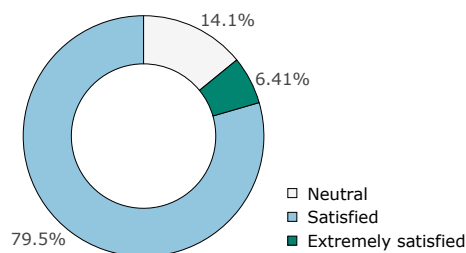
*Group-Level.* As mentioned earlier, we aggregated the individual responses for all attendees belonging to the same meeting to one value representing the group-level. The lowest group-level meeting satisfaction was 0.5 and the highest 1.667. On average, the meetings showed a mean satisfaction score of 0.957 indicating that the mean meeting satisfaction was slightly below *satisfied* ( $Median = 1.0$ ,  $SD \pm 0.260$ ). The time-well-spent score for meetings ranged from 0.5 to 2.0. The meetings showed a slightly higher mean time-well-spent score than *well* ( $Median = 1.0$ ,  $Mean = 1.156$ ,  $SD \pm 0.331$ ). Spearman's rank correlation was run to assess the relationship between the perception of the time spent and meeting satisfaction on the group-level. There was a statistically significant, fairly substantial correlation between these two group-level variables,  $r_s(23) = 0.516$ ,  $p = .008$ .

Distribution of quality of the time spent in the meeting



**Figure 5.1:** Distribution of Participant's Quality of the Time Spent in Meetings (Individual-Level)

Distribution of meeting satisfaction



**Figure 5.2:** Distribution of Participant's Meeting Satisfaction (Individual-Level)

## General Speech Features

*Meeting Duration.* Meetings lasted, on average, around 40 ( $\pm 25$ ) minutes. The shortest meeting had a duration of approximately 15 minutes and the longest of around one hour and 41 minutes.

*Laughs.* Some meetings seem to have a very pleasant atmosphere and there were up to 131 laugh sequences in a meeting. On the other side, there were some meetings without any detected laughter. The laughs found with the laughter detection algorithm resulted in a mean of  $26 \pm 35$  laughter per meeting. Since meetings had various lengths, we normalized this value to the meeting duration (in seconds), resulting in a mean of 0.009 laughs ( $Median = 0.005$ ,  $SD \pm 0.011$ ,  $Min = 0$ ,  $Max = 0.049$ ).

## Turn-Taking Features

*Total Speaking Length.* On average, the meetings consisted of 85.0% speaking time of the entire meeting duration ( $Median = 89.8\%$ ,  $SD \pm 10.2$ ). The lowest value was 55.2 % total speaking time because during these meetings attendees watched meeting related videos which were excluded as speech. The maximum speaking time was 96.5 % of the meeting duration.

*Total Speaking Turns.* The meetings showed an average number of turns of 257 ( $Median = 224.0$ ,  $SD \pm 171.5$ ). There were some meetings having only 44 turns and others up to 812 turns. However, this number largely depends on the duration of the meeting. Therefore, if we normalize the number of turns by the meeting duration (in seconds) we receive the following values:  $Median = 0.096$   $Mean = 0.102$ ,  $SD \pm 0.031$ ,  $Min = 0.047$ ,  $Max = 0.166$ .

*Turn-taking Freedom.* There was a slight tendency of a rather free than strict speaking order of the meeting attendees, indicating that they spoke in random order. This is expressed with the mean of the turn-taking freedom of 0.578 ( $Median = 0.591$ ,  $SD \pm 0.150$ ). There were some meetings having a rather strict speaking order expressed through the minimum turn-taking freedom of 0.308. Other meetings showed a rather free speaking order (maximum turn-taking freedom of 0.805). The turn-taking freedom could only be calculated on meetings having a size of at least three participants.

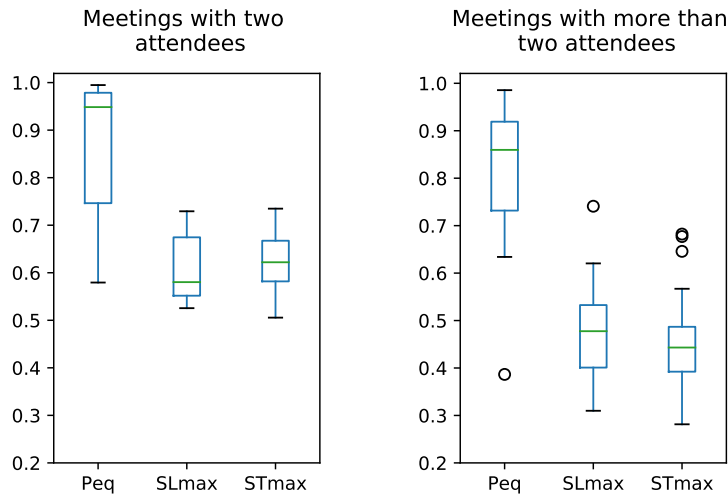
## Centrality Features

We decided to split the description of the centrality features for meetings below two attendees (28%,  $N=7$ ) and the remaining ones having more than two attendees (72%,  $N=18$ ). The box plots in Figure 5.3 visualize these differences. The degree and eigenvector centrality measures were only calculated for meetings with more than two attendees.

*Participation Equality.* On average, meetings of two attendees showed a slightly higher level of equality in the participation of attendees ( $Mean = 0.853$ ,  $Median = 0.948$ ,  $SD \pm 0.172$ ,  $Min = 0.579$ ,  $Max = 0.995$ ) than meetings of more than two attendees ( $Mean = 0.821$ ,  $Median = 0.860$ ,  $SD \pm 0.151$ ,  $Min = 0.387$ ,  $Max = 0.986$ ). This indicates that the distribution of speech time was more balanced in two-party meetings than multi-party meetings.

*Maximum Percentage of Speaking Turns.* Meetings of more than two participants had a mean of 45.7% of all turns within a meeting that are attributed to one speaker ( $Median = 44.3\%$ ,  $\pm 11.9\%$ ,  $Min = 28.1\%$ ,  $Max = 68.2\%$ ). In contrast, in meetings with two attendees, on average, 62.3 % of all turns belong to one speaker ( $Median = 62.2\%$ ,  $SD \pm 7.8\%$ ,  $Min = 50.6\%$ ,  $Max = 73.5\%$ ).

*Maximum Percentage of Speaking Length.* For meetings having more than two attendees, we quantified a maximum percentage of speaking length between 31.0% and 74%. On average, meetings have a speaker talking more than  $48.0\% \pm 11.4$  (*Median* = 47.8%) of the entire speaking time. For meetings having only two attendees, this data looks slightly different: Meetings have on average an attendee talking for 61.2% of the entire speaking length (*Median* = 58.0%, *SD*  $\pm 8.1\%$ , *Min* = 52.5%, *Max* = 72.9%).



**Figure 5.3:** Distribution of Centrality Features (*Peq*=Participation Equality, *SLmax*=Maximum Percentage of Speaking Length, *STmax*=Maximum Percentage of Speaking Turns)

## 5.2 Stated Meeting Factors Contributing to Meeting Quality

We explicitly asked the participants which meeting factors affected their quality of the time spent in meetings. Additionally, we asked the participants to rate each factor according to its presence in the meeting, e.g. to which extent they agree with having a lively and active exchange in the meeting.

Based on these two answers, we were able to not only identify which factors impacted participants' quality but also whether this impact depended on the presence of the factor within the meeting.

**Analysis** Table 5.2 presents the distribution of participants' responses on the meeting factors effect and the occurrence of factors in the meeting. The column meeting factor refers to the keywords resulting from the mapping between questions 2 and 4 (see Table 4.3). The percentages and frequencies belong to participant's response of the impacts of the factors on their quality of time spent in the meeting (question 4) whereas the means and standard deviations correspond to participants' agreement of the occurrence of the factor for the corresponding group (question 2). With the term group, we refer in this section to the participant's answer choice of question 4, e.g. *no*

*effect, affected positively or affected negatively.* For example, to calculate the mean value for the factor 'Distribution of Speech Time', we first grouped the participants into three groups according to their answer to the factor's impact on their quality. Then for each group, we calculated the mean score of participant's agreement of question 2 representing to which extent this factor characterized the meeting. The columns Kruskal-Wallis H contain the statistical test results of the Kruskal-Wallis H test which were performed on the mean agreement scores for each factor and group. Our rationale behind this test is to examine whether the impact of the factors of the quality depends on the occurrence of the factor which can be calculated by comparing the mean agreement scores between the groups.

**Dunn's Post Hoc Test: Differences Between Groups.** The results of the Kruskal-Wallis H test shows that for five out of seven meeting factors the groups show significantly different means of the agreement scores which represent the occurrence of the factor within the meeting. This indicates, that the influence not only depends on the factor itself but also on the occurrence of it. However, Kruskal-Wallis H does not provide information between which groups the differences occur only provides evidence for a difference between mean ranks of at least one pair of groups. Therefore, for the factors showing statistically significant different means, we performed the Dunn's post hoc test. The following list presents the results:

- Interruptions showed a statistically significant different mean agreement score between the groups *no effect* and *negative affect* ( $p = .02$ ).
- The usefulness of the meeting showed a significant difference in the mean agreement scores between the groups *positive effect* and *negative effect* ( $p = .004$ ).
- The relevance of participants' attendance showed a significant different mean between *no effect* and *positive effect* ( $p = .01$ ) and between *positive effect* and *negative effect* ( $p = .009$ ).
- Lively and active exchange showed different mean agreement scores between *no effect* and *positive affect* ( $P < .001$ ) and between *positive effect* and *negative affect* ( $p = .04$ ).

**Results.** *Lively and active Exchange.* We want to know whether lively and active exchange influenced the participant's perceived quality of time spent in the meeting. Additionally, if it depends on the occurrence of the factor, meaning whether an active exchange characterized the meeting. First, we can see that 82.9% (N=63) of the attendees mentioned to be positively affected by an active exchange, whereas 15.8% (N=12) were not affected and 1.3% (N=1) negatively affected. The Kruskal-Wallis H test showed a significant difference between the means of the factors of each group. As we can see the agreement for an active exchange for the group who mentioned the meeting quality is positively impacted is 2.349 whereas the mean of the negative impact is 1.417. The results of the Dunn's post hoc test showed that this difference between these two groups is significant. Moreover, the difference between means the groups *positive affect* and *no effect* is significant as well. The mean score for the group *positive affect* is significantly higher than the mean agreement for the other two groups, indicating that if a lively exchange positively affected the quality, the meetings generally tend to be more characterized by lively exchanges. Moreover, if a lively and active exchange did not characterize the meeting, it either had no effect or a negative effect on the meeting quality.

*Interruptions.* Most of the attendees mentioned that their meeting quality was not affected by the interruptions between members. However, we can see that if the interruptions showed a negative or no impact it might depend on its occurrence because the mean agreement scores significantly differ between the groups *no effect* and *negative affect*. This indicates that if the interruptions

were perceived as negative the participants perceived to have more interruptions than if the interruptions did not affect.

*Presence of Key People.* No one mentioned being negatively affected by the presence of key people within a meeting. Compared to the other factors, this one has a somewhat similar distribution of the impact on the meeting quality. Moreover, the mean agreement score of the presence of key people in meetings is significantly higher if this factor resulted in a positive effect on participants' time-well-spent than if it did not affect the quality. This points to the fact that when key people are present in meetings, the participant's meeting quality is positively affected. However, if no key people are present in a meeting, this does not affect the quality of the meeting.

*Meeting Usefulness.* Interestingly, but not surprisingly, if meetings had a negative effect, they were considered less useful than meetings with a positive effect.

*Participants' Relevance.* For 63.0 % of the attendees, the relevance of their attendance positively affected their meeting quality, for 32.9% it showed no effect and for 4.1% a negative effect. Meetings, where participants indicated that they had a positive impact on their time spent, showed a higher mean agreement score of their relevance in the meeting than the other two groups.

**Summary.** In summary, the main factors positively contribute to participants' perceived meeting quality is an open, active meeting atmosphere (83.1%), closely followed by an active, lively exchange (82.9%). Moreover, the occurrence of interruptions between members has the highest amount of answers for not impacting participants' meeting quality (89.1%). The meeting usefulness showed the highest percentage of negative impact (6.7%), especially if the participants felt the meeting was less useful for them.

## 5.3 Correlation Between Reported Meeting Factors and Meeting Quality

As mentioned before, we asked the participants to rate their agreement of the factor's occurrence in the meeting. We set this reported meeting factor in correlation with their perceived meeting quality. We want to examine if the meeting quality depends on their perceived meeting characteristics. We refer to this reported meeting factors as inherent factors.

In addition, we included background variables of the participants, such as his role in the company, or variables about the meeting, such as the meeting size, in the analysis and tested whether the meeting quality depends on them. We will refer to these two variables as external factors in this section.

### 5.3.1 Inherent Factors

We analyzed if there exists a relationship between the answers of the subjects about the meeting quality and their perception of the occurrence of various meeting factors. To analyze this, we run for each meeting factor a Spearman's rank correlation with the meeting quality, on an individual-level as well as group-level.

**Perceived quality of time spent in meetings.** From this analysis, it is particularly noticeable that the two meeting factors for a relevant attendance of the participant and for an open, positive meeting atmosphere correlate with the participants' perceived quality of time spent in meetings, both levels (group and individual). Interestingly, the relationship between the quality of time spent

Meeting Factor		No effect	Positive effect	Negative effect	Kruskal-Wallis	
					$\chi^c(2)$	<i>p</i>
Distribution of Speech Time	%	76.1	21.1	2.8	3.202	.202
	N	54	15	2		
		<i>M</i> = −0.019 <i>SD</i> = 1.985	<i>M</i> = 0.933 <i>SD</i> = 1.486	<i>M</i> = 1.500 <i>SD</i> = 0.707		
Interruptions	%	89.1	7.8	3.1	6.900	.032*
	N	57	5	2		
		<i>M</i> = −1.554 <i>SD</i> = 1.451	<i>M</i> = −0.800 <i>SD</i> = 1.304	<i>M</i> = 1.500 <i>SD</i> = 0.707		
Meeting Usefulness	%	20.3	73.0	6.7	8.895	.012*
	N	15	54	5		
		<i>M</i> = 1.933 <i>SD</i> = 0.961	<i>M</i> = 2.296 <i>SD</i> = 0.690	<i>M</i> = 0.400 <i>SD</i> = 1.817		
Participant's Relevance	%	32.9	63.0	4.1	10.452	.005*
	N	24	46	3		
		<i>M</i> = 1.833 <i>SD</i> = 1.167	<i>M</i> = 2.435 <i>SD</i> = 0.720	<i>M</i> = 0.333 <i>SD</i> = 2.082		
Open Atmosphere	%	14.3	83.1	2.6	0.634	.728
	N	11	64	2		
		<i>M</i> = 2.363 <i>SD</i> = 0.674	<i>M</i> = 2.422 <i>SD</i> = 0.708	<i>M</i> = 1.000 <i>SD</i> = 2.828		
Lively Exchange	%	15.8	82.9	1.3	13.624	.001*
	N	12	63	1		
		<i>M</i> = 1.417 <i>SD</i> = 1.240	<i>M</i> = 2.349 <i>SD</i> = 0.544	<i>M</i> = 1.000 —		
Presence of Key People	%	57.7	42.3	—	350.5 <sup>a</sup>	.0038 <sup>a*</sup>
	N	41	30	—		
		<i>M</i> = 0.540 <i>SD</i> = 2.129	<i>M</i> = 1.933 <i>SD</i> = 1.202	—		

a. Mann-Whitney U Test

\*. Correlation significant at  $p < 0.05$ , \*\*. Correlation significant at  $p < 0.001$  level

**Table 5.2:** Distribution of Meeting Factor's Effect on Meeting Quality and Occurrence of Factor Expressed by the Means (Individual-Level) (%=Percentage of Responses of Factor's Impact on Quality 4, N=Number of Responses of Factor's Impact on Quality, M=Mean of Agreement Score of Factor's Occurrence, SD=Standard Deviation of Agreement Score of Factor's Occurrence,  $\chi^c(2)$ = Result of Kruskal-Wallis H Test with Degree of Freedom,  $p$ =P-Value)

in meetings and the meeting usefulness is statistically significant on individual-level, but not on the group-level. The exact values are depicted in Table 5.3.

Additionally, we performed a stepwise linear regression procedure to measure the power of a meeting factor's occurrence to predict the time-well-spent score of a meeting (on a group-level). Resulting in a model in which an open atmosphere and participant's relevance statistically significantly predict the time-well-spent score,  $F(1, 24) = 6.311, p = .007$ . The corresponding model summary and coefficient table can be found in Appendix E.

**Meeting Satisfaction.** If we perform correlation for this meeting factors with the meeting satisfaction we noticed on a individual-level a correlation between meeting satisfaction and three variables: (1) meeting usefulness ( $r_s(76), p = .003$ ), (2) participants' relevance ( $r_s(76), p < .001$ ), (3) lively exchange ( $r_s(76), p = .01$ ). On the group-levels, the positive relationship between meeting satisfaction and the relevance of the attendance of the participant was statistically significant ( $r_s(23) = 0.580, p = .002$ ). Besides this, meeting satisfaction correlated positively with a lively and active exchange within the meeting ( $r_s(23) = 0.426, p = .034$ ).

Factor	Group-Level				Individual-Level				
	<i>M</i>	<i>SD</i>	$r_s(23)$	<i>p</i>	<i>N</i>	<i>M</i>	<i>SD</i>	$r_s$	<i>p</i>
Distribution of Speech Time	-0.101	1.415	0.396	.177	77	0.169	1.867	0.103	.372
Interruptions	-1.488	1.023	0.646	.097	76	-1.461	1.483	0.135	.244
Usefulness	2.218	0.569	0.331	.106	78	2.128	0.958	0.469	.000*
Participant's Relevance	2.318	0.501	0.422	.036*	78	2.167	1.025	0.430	.000*
Open Atmosphere	2.397	0.517	0.491	.013*	78	2.372	0.791	0.287	.011*
Lively Exchange	2.236	0.510	0.146	.486	78	2.141	0.817	0.212	.062
Attendance of Key People	0.901	1.531	-0.036	.865	74	1.095	1.874	0.100	.398

\*. Correlation significant at 0.05, \*\*. Correlation significant at 0.001 level

**Table 5.3:** Spearman's Rank Correlations Between Meeting Factors and Participant's Quality of the Time Spent (*N*=Number of Answers, *M*=Mean of Agreement, *SD*=Standard Deviation of Agreement,  $r_s$ =Spearman's Rank Correlation Coefficient, *p*=*P*-Value)

### 5.3.2 External Factors

Different external factors, such as the gender, the role of the meeting participant within the meeting or within the company, or the size or start time of the meeting might affect the perceived quality of a meeting. Therefore, we examined each of these categories.

**Gender.** A Mann-Whitney U test indicated *no* statistically significant difference between genders and their perceived quality of the time spent in meetings ( $U = 725.5, p = .326$ ) or their meeting satisfaction ( $U = 722.5, p = .297$ ).



**Attendee's Role Within the Meeting.** We analyzed whether attendee having different meeting roles perceived the meeting quality differently. Within eight meetings, all attendees mentioned to be a *Participant* and no one stated to be a *Chairperson* controlling the meeting. Therefore, we excluded these meetings which do not have an explicit chairperson for this analysis. The following analysis is based upon 14 *Chairpersons* (26.42%) and 39 *Participants* (73.58%). The performance of a Mann-Whitney U test showed *no* statistically significant difference ( $U = 244.5, p = .237$ ) of perceived quality of the time spent between *Participants* ( $Median = 1.0, Mean = 1.076, SD \pm 0.580$ ) and *Chairpersons* ( $Median = 1.0, Mean = 1.214, SD \pm 0.579$ ). Likewise, meeting satisfaction showed *no* significant difference either ( $Mann - Whitney U = 266.5, p = .426$ ), with *Participants* having a mean of 1.0 ( $Median = 1.0, SD \pm 0.392$ ) and *Chairpersons* a mean of 0.906 ( $Median = 1.0, SD \pm 0.462$ ).

**Attendee's Role Within the Company.** We examined whether participants perceived the meeting quality differently depending on their role within the company. Therefore, we performed a Mann-Whitney U test which resulted in *no* statistically significant difference in perceived quality of time spent between contributors ( $Median = 1.0, Mean = 1.11, SD \pm 0.575$ ) and *Leader/Managers* ( $Median = 1.0, Mean = 1.12, SD \pm 0.504$ ),  $U = 745.0, p = 0.445$ . In contrast, the *Contributors* ( $Median = 1.0, Mean = 1.083, SD \pm 0.368$ ) and *Leader/Managers* ( $Median = 1.0, Mean = 0.786, SD \pm 0.470$ ) showed a statistically significant different meeting satisfaction, according to a Mann-Whitney U test ( $U = 550.5, p = .002$ ). This shows that *Leader/Managers* show generally a lower meeting satisfaction than *Contributors*.

**Presence of a Chairperson.** We analyzed if there exists a difference in perceived meeting quality between meetings having a chairperson and meetings without having someone with the responsibility to clearly lead or manage the meeting. A Mann-Whitney U test showed that there was *not* a statistically significant difference in perceived time-well-spent score between meetings with and without chairperson ( $U = 67.0, p = 0.487$ ), with a mean time-well-spent score of 1.145 for meetings having no chairperson ( $N = 8, Median = 1.0, SD \pm 0.208$ ), a mean of 1.161 for meetings with a chairperson ( $N = 17, Median = 1.0, SD \pm 0.382$ ). However, there was a statistically significant difference in meeting satisfaction score between meetings that had been held with a chairperson ( $Median = 1.0, Mean = 1.049, SD \pm 0.230$ ) and without chairperson ( $Median = 0.708, Mean = 0.760, SD \pm 0.216$ ), according to a Mann Whitney U test ( $U = 26.0, p = .003$ ). This indicates that meetings without a chairperson are perceived as less satisfying than meetings having a chairperson.

**Meeting Size.** We conducted a Kruskal-Wallis H test to determine if there were differences in the time-well-spent scores between various meeting sizes. There was *no* statistically significant difference between meeting sizes ( $\chi^2(2) = 1.992, p = .574$ ), with a mean score of 1.286 ( $Median = 1.0, SD \pm 0.393$ ) for meetings having two attendees, 1.083 ( $Median = 1.0, SD \pm 0.236$ ) for three attendees, 1.222 ( $Median = 1.0, SD \pm 0.344$ ) for four attendees, and 0.975 ( $Median = 1.0, SD \pm 0.369$ ) for more than four attendees. We performed the above mentioned analysis not only on the time-well-spent scores, also on the group-level scores of the meeting satisfaction. Likewise, this Kruskal-Wallis H test resulted in *no* statistically significant differences between meeting size either, ( $\chi^2(2) = 2.712, p = .438$ ), meetings with two attendees have a mean satisfaction of 1.071 ( $Median = 1.0, SD \pm 0.189$ ), three attendees of 0.917 ( $Median = 1.0, SD \pm 0.154$ ), for four attendees of 0.903 ( $Median = 0.875, SD \pm 0.436$ ), and more than four attendees of 0.917 ( $Median = 0.9, SD \pm 0.233$ ).

**Meeting Start Time.** Then, we analyzed for differences of the time-well-spent scores between three meeting start times: Morning meetings before noon (52%,  $N=13$ ), early afternoon meetings

before 3:30 pm (28%, N=7), late afternoon meetings after 3:30 pm (20%, N=5). A Kruskal-Wallis H test showed that there was *not* a statistically significant difference in perceived time-well-spent score of various meeting times,  $\chi^2(2) = 0.2119, p = 0.8995$ , with a mean time well spent score of 1.185 ( $Median = 1.0, SD \pm 0.384$ ) for meetings in the morning, mean of 1.143 ( $Median = 1.0, SD \pm 0.325$ ) for early afternoon meetings, and mean of 1.1 ( $Median = 1.0, SD \pm 0.224$ ) for late afternoon meetings. Furthermore, we applied the Kruskal-Wallis H test to check the difference between meeting satisfaction and meeting time, and this difference was *not* statistically significant either,  $\chi^2(2) = 3.5699, p = 0.1678$ , with a mean satisfaction of 0.964 ( $Median = 1.0, SD \pm 0.249$ ) for meetings in the morning, mean of 1.067 ( $Median = 1.0, SD \pm 0.275$ ) for early afternoon meetings, and mean of 0.783 ( $Median = 0.75, SD \pm 0.217$ ) for late afternoon meetings.

Moreover, we combined the afternoon meetings into one category which is resulting in two meeting times: morning meetings (52%, N=13), afternoon meetings (48%, N=12). We applied a Mann-Whitney U test to check for a difference between these two meeting times and the meeting satisfaction score. It showed *no* statistically significant difference between morning meetings ( $Median = 1.0, Mean = 0.964, SD \pm 0.249$ ) and afternoon meetings ( $Median = 1.0, Mean = 0.949, SD \pm 0.282$ ) in terms of their meeting satisfaction score ( $U = 85.5, p = .667$ ).

## 5.4 Correlations between Extracted Speech Features and Meeting Factors

We examined if the participants' perception of the occurrence of the meeting factor can be assessed using the extracted speech features. Therefore, we run a correlation for each meeting factor and the 13 extracted speech features on the group-level (stated in Chapter 3.3.2).

The reported speech features for each meeting factors were chosen based on our beliefs that these might be connected or because they showed a significant relation.

**Distribution of Speech Time.** *None* of the following speech features has a significant association with the agreement score of the distribution of the speech time: Participation equality (Spearman,  $r_s(23) = -0.294, p = .154$ ), percentage of maximum speaking length (Pearson,  $r(23) = -0.045, p = .833$ ), and percentage of maximum speaking turns (Pearson,  $r(23) = -0.281, p = .331$ ). In general, we can not measure the participants' perceived distribution of speech time with an extracted speech feature.

**Lively and Active Exchange.** The turn-taking freedom was set in correlation with the lively and active exchange score of a meeting, resulting in a *not* statistically significant relationship (Pearson,  $r(16) = 0.403, p = .097$ ). However, there is a statistically significant, negative correlation between maximum eigenvector centrality and the exchange score of a meeting (Pearson,  $r(16) = -0.646, p = .004$ ).

In addition, the score of a lively and active exchange and (a) the mean speech rate (Spearman,  $r_s(23) = -0.459, p = .002$ ) and (b) the mean articulation rate (Spearman,  $r_s(23) = -0.402, p = .046$ ) are negatively correlated.

**Presence of Key People.** We investigated the relationship between centrality measures and the meeting's score of having more than one key people. The following relationships were *not* statistically significant: Maximum degree centrality (Spearman,  $r_s(16) = 0.400, p = .100$ ) or Maximum eigenvector centrality (Pearson,  $r(16) = 0.151, p = .551$ ). However, there exists a statistically significant, negative relationship between the score for a presence of a key people and the maximum speaking turns in percentage (Spearman,  $r_s(23) = -0.369, p = .05$ ). Maximum

speaking turns in percentage relates to the feature where for each meeting the speaker with the highest number of turns relative to the total speaking turns is extracted. Moreover, We noticed a significantly different mean of maximum percentage of speaking turns by various meeting sizes (One-way ANOVA:  $F(3, 21)=3.795, p=.026$ ). The Tukey post hoc test revealed a statistically significant difference in maximum percentage of speaking turns between meetings of size 2 and 4 ( $p=.040$ ), whereas between meetings of size 2 and 3, or 2 and more than 4 the difference was not significant ( $p=.060$  and  $p=.1$ , respectively). Therefore, we decided to exclude two-party meetings and performed the correlation again. The significant, negative relationship between the presence of a key person and the maximum speaking turns was even stronger (Pearson,  $r(16) = -0.613, p = .006$ ).

Moreover, the score of having one or more key people and (a) the mean speech rate ( $r_s(23) = 0.433, p = .031$ ) and (b) the mean articulation rate ( $r_r(23) = 0.50, p = .004$ ) are negatively correlated.

**Open and Positive Atmosphere.** Local jitter and local shimmer values are set in correlation with the open and positive atmosphere score and resulted in *no* statistically significant correlation (Spearman,  $r_s(23) = 0.017, p = .937, r_s(23) = 0.083, p = .692$ , respectively). The analysis generally showed no speech features related to this meeting factors.

### 5.4.1 Relationships Among Meeting Factors

Then, we examined if there exists some relationships between different meeting factors. By performing the correlations, our data showed four statistically significant, positive relationships at both group and individual levels. The perceived relevance of the participants' attendance depends on the meeting usefulness for the attendee. Moreover, if participants perceived to have a meeting with a dominant attendee, it is linked to having attendees interrupting each other. Moreover, in meetings where interruptions among attendees occurred, participants mentioned the presence of a key person which is also linked to the attendance of a dominant attendee. For the exact values, please refer to Table 5.4. The degree of freedoms for the individual-level correlation calculations varied because some participants did not answer the question. For group-levels we calculated the values on all meetings.

### 5.4.2 Relationships Among Speech Features

We analyzed if the extracted speech features show a significant relation to another speech features. For this, we run a correlation between the automatically extracted speech features among one another. We performed this analysis twice: once for the group-level features and once for the individual-level features. For an overview of the extracted features used in this section please refer to Chapter 3.3.2.

**Individual-Level.** The analysis of speech features on the individual-level revealed a statistically significant, positive relationship between participant's centrality measure and participant's total speaking length (degree centrality: Spearman,  $r_s(74) = 0.657, p < .001$ , eigenvector centrality: Spearman,  $r_s(74) = 0.862, p < .001$ ). This indicates that more central persons tend to speak for a longer time. Moreover, participants having a higher total speaking length tend to a higher amount of total turns (Spearman,  $r_s(88) = 0.869, p < .001$ ).

**Group-Level.** Meetings having a less strict speaking order showed a higher equal participation among group members. Moreover, both these variables are negatively linked to some centrality

Measure <sub>1</sub>	Measure <sub>2</sub>	Group-Level	Individual-Level	
		$r_s(23)$	$df$	$r_s$
Participants' Relevance	Usefulness	<b>0.614**</b>	76	<b>0.568**</b>
Participants' Relevance	Lively & Active Exchange	<b>0.483*</b>	76	0.047
Lively & Active Exchange	Atmosphere	0.216	76	<b>0.325*</b>
Lively & Active Exchange	Usefulness	0.188	76	<b>0.242*</b>
Presence Dominant Attendee	Interruptions	<b>0.572**<sup>a</sup></b>	73	<b>0.269*</b>
Presence Dominant Attendee	Key people	<b>0.477*</b>	72	<b>0.342*</b>
Interruptions	Key people	<b>0.664**</b>	70	<b>0.363*</b>

<sup>a</sup>. Pearson Correlation Coefficient (normally distributed data)

\*. Correlation significant at 0.05, \*\*. Correlation significant at 0.001 level

**Table 5.4:** Spearman's Rank Correlations Between Meeting Factors ( $r_s$ =Spearman's rank coefficient,  $df$ =Degree of Freedom)

features which might represent dominant persons (e.g. the maximum percentage of speaking length attributable to one person). The corresponding correlation coefficients are shown in Table 5.5. Rows with a degree of freedom of 16 indicate measurements on meetings with more than two participants, while the others are calculated on all meetings.

## 5.5 Correlation between Automatically Extracted Speech Features and Meeting Quality

Lastly, we analyzed if we can predict the meeting quality using the automatically extracted speech features. Therefore, We examined the correlation between the 13 extracted group-level features (refer Chapter 3.3.2) and the participants' meeting quality. This resulted in a moderately, positive relationship between meeting satisfaction and total amount of speaking turns within a meeting (Spearman,  $r_s(23) = 0.400, p = .048$ ). However, this statistically significant correlation is *not* existing for the relationship between the total number of speaking turns and the meeting's time-well-spent score (Spearman,  $r_s(23) = -.076, p = .717$ ). The other correlations between the extracted language characteristics and the meeting qualities did not lead to significant outcomes.

## 5.6 Threats to Validity

The following section discusses some points that threaten the validity of the results of this qualitative study.

### 5.6.1 External Validity

Our results are limited with regards to generalizability because of conducting the study only within one company and collecting a relatively small amount of samples. Furthermore, the company already had a very well established meeting culture which is supported by a meeting codex.

Measure <sub>1</sub>	Measure <sub>2</sub>	df	$r_s(23)$
Participation Equality	Percentage Max Speaking Length	23	-0.622**
Participation Equality	Percentage Max Speaking Length	16	-0.825**
Participation Equality	Percentage Max Speaking Turns	23	-0.444*
Participation Equality	Percentage Max Speaking Turns	16	-0.818**
Participation Equality	Max Eigenvector Centrality	16	-0.639*
Participation Equality	Turn-Taking Freedom	16 <sup>a</sup>	0.515 <sup>a</sup>
Turn-Taking Freedom	Percentage Max Speaking Length	16	-0.708**
Turn-Taking Freedom	Percentage Max Speaking Turns	16	-0.761**
Turn-Taking Freedom	Max Eigenvector Centrality	16	-0.880**
Turn-Taking Freedom	Total Speaking Length	16	-0.560*
Total Speaking Turns	Total Speaker Changes	23	0.715**

a. Pearson Correlation, both variables normally distributed

\*, Correlation significant at 0.05, \*\*, Correlation significant at 0.001 level

**Table 5.5:** Spearman's Rank Correlations Between Speech Features ( $r_s$ =Spearman's rank coefficient,  $df$ =Degree of freedom.)

In every meeting room was a list with the terms and ground rules that govern proper meeting behaviors and interactions that apply in the meetings. The study, therefore, leads to a first insight and could be extended and reproduced in several companies and for a more extended time period. Moreover, our meetings only consisted of one-on-one-meetings or many-to-many-meetings, such as brainstorming sessions, discussions, or status updates. Therefore, the generalizability might be limited to similar meeting types.

In addition, the presence of the researcher could have positively influenced the participants and their perception about the meeting quality. Next, the participants may have responded more positively than they actually felt, knowing that the results would be published. Although, everything is anonymized and neither a meeting nor a participant can be identified. Moreover, it is possible that the current mood of the participants influences the perceived meeting quality so that the same meeting would have been rated differently on a different day or time.

## 5.6.2 Internal Validity

As seen in Section 3.2 at the time of writing this thesis there exists no open source speaker diarization system with reasonably high accuracy. Therefore, after processing the audio files through the used diarization software, the output might not be entirely precise. Especially not when the meeting was characterized by many short speaking segments, overlapping speech, or attendees with similar voices. We tried to mitigate this, by skimming the audio files which we created based on the diarization output. If too many segments of a speaker were wrongly classified we manually changed the speaker labels, which is explained in Section 3.3.1.

### 5.6.3 Construct Validity

To ensure construct validity, we eliminated non-relevant questions and changed our wording of the questionnaire to be as clear, simple, and unambiguous as possible. However, a future post-meeting survey might use a 7-item Likert scale for question 1 and 3 to obtain a more detailed gradation of attendee's perception. Moreover, another small improvement of this survey for future studies would be to use a different labelling of the answer choices of the Likert item questions. So far, we used explicit wording for all answer options (e.g. *Agree*), which is an ordinal scale. Therefore, one can not make any assumptions about the spacing of the response options, e.g. if answer *Strongly agree* and *Agree* are equidistant to *Agree* and *Slightly Agree*. Thus, one might use explicit rating by presenting the numbers 1 - 7 as answer options which are supported with explicit wordings at both ends of the scale.

# Discussion and Future Work

Based on our literature research, we identified a set of automatically extractable features from audio recordings of a meeting. We examined whether it is possible to assess the participants' meeting quality using these features. In the following, we discuss our main findings.

It is important to mention that the analyzed meetings overall showed a high quality and hardly any employee rated the meetings as a waste of time. The reason for this could be a well-established meeting culture in the company in which we conducted the study. Therefore, future work should replicate this study and the analysis within a company having a broader variety of perceived meeting quality which could lead to new insights.

**Importance of open and positive atmosphere.** Using the post-meeting survey, we analyzed which factors contributed to participants' perceived quality of their time spent in meetings. This revealed that an open, positive atmosphere showed the highest percentage (83.1%) of subjects' answers to affect participants' quality of time spent in meetings positively. Hence, an atmosphere in which participants can contribute with ease is crucial for high meeting quality. On the contrary, participants' meeting quality was negatively affected by interruptions between group members. These two aspects, an open meeting atmosphere and interrupting members, might be mutually exclusive because in meetings where people often interrupt each other the participants might be reluctant and may not feel that they can contribute openly. In fact, participants' perception of uncivil meeting behaviours, e.g. contributions in a rude way or interruptions between participants, is linked to a lower meeting satisfaction [OKK<sup>+</sup>18].

Our finding that an open meeting atmosphere is linked to the meeting quality supports current research. Nixon *et al.* [NL92] highlighted the positive impact of open communication towards meeting effectiveness. Additionally, Germans prefer frank and open discussions among team members as a study has shown [KG15].

While we were not able to automatically measure the openness and positivity of the meeting atmosphere, we identified ways in which one might be able to capture them. In particular, features of voice quality, namely jitter and shimmer values, might be used to capture the openness and positivity to some extent<sup>1</sup>. In the past, jitter and shimmer have been found to be linked to the participant's emotional and mental state. Jitter and shimmer capture the openness/tightness of the participant's speech. Therefore, we might be able to identify whether participants speak in their normal voice or speak with a tight throat, which might be triggered by the presence of another person, e.g. an employee of a higher hierarchical level. However, we need a baseline for each individual person to accurately interpret the measured values and its changes within the same or between various meetings. Future research is needed to analyze the relationship between voice

---

<sup>1</sup>This information is based on our workshop with two members of the Institute of Computational Linguistics of the University of Zurich.

quality and an open atmosphere. Moreover, another possibility might be to automatically detect human emotion from the speech in combination with sentiment analysis to quantify the emotional expressions of the spoken content. The combination of these two emotional aspects would lead to more information because the meaning of a sentence or how others perceive the statement depends on the tone and use of emotional language. For example, if a person speaks in an angry tone, it is less likely for others to contribute with ease, especially for shy people.

Lai *et al.* [LCR13] highlighted that higher a turn-taking freedom, which represents that attendees are speaking in a relatively free manner, is linked to a higher meeting satisfaction. Although our analysis revealed no relation between the meeting openness and the turn-taking freedom, we could argue that if the participants spoke randomly successively it might be linked to an open atmosphere. Nevertheless, turn-taking freedom is just a representation of whether the attendees talked in strict order or freely; it does not imply that there is a strict order predefined in reality.

**Meaningfulness of a lively and active exchange** Another crucial factor contributing to a positive meeting quality is to have a lively and active exchange within the meeting. Around 82.9% of participants mentioned that their meeting quality was positively affected by lively and active exchanges in meetings. The impact of this factor depends on participant's perception of having an active and lively exchange within the meeting, e.g. if there were no active and lively exchanges in the meeting, it had no positive impact on participants meeting quality. Data analysis showed a negative correlation between meetings having an attendee with a high centrality score and participants' perception of having a lively and active exchange. In other words, meetings with a more central person were perceived as having fewer lively and active exchanges. In addition, the presence of a lively and active exchange is linked to higher ratings of meeting satisfaction. Moreover, related work found an impact on meeting satisfaction, if participants observe not actively participating attendees [OKK<sup>+</sup>18].

**Challenges to assess meeting quality using automatically extracted speech features.** The analysis of the relationship between various speech features and the perceived meeting quality resulted in only a moderately positive relationship between meeting satisfaction and the total amount of speaking turns within a meeting. All other features were not significantly related to the quality of the meeting. Speaking turns are speech sequences that are separated by a 500ms pause. Accordingly, it implies that speakers make a small pause while talking which might provides other participants with the possibility to take the chance to speak up. Therefore, a higher amount of speaking turns might be linked to a positive atmosphere.

Most of our results conform with the findings of Lai *et al.* [LCR13]: How well a meeting is going is neither depending on the dominance feature (i.e. total speaking time of a person relative to the others), nor equal participation among members, nor on the total speaking time of all attendees. However, in their study, a higher amount of speaking freely than in strict order showed significant correlation with meeting satisfaction. We were not able to show this effect. Possible reasons for discrepancies between our findings and the findings of Lai *et al.*'s [LM18] are that our meetings were perceived as mostly positive and there was not much variance between the meetings. However, another reason could be more profound. Since a group conversation depends on social and cultural factors, this affects the way people talk with each other and characterizes their interaction structures [LRS<sup>+</sup>18]. While Lai *et al.* [LCR13] findings are based on the AMI Meeting Corpus which captures meetings in England, Switzerland and the Netherlands, our results are based on meetings held in Germany. According to Köler *et al.* [KG15] meeting expectations and norms differ across cultures, even if underlying cultural values seem similar. Not only country dependent cultural differences but also organizations meeting culture influences participant's perception of meeting quality and how they prefer meetings to be held. This can hinder the generalizability of our findings to other cultures. Moreover, our meeting attendees know each



other rather well and are co-workers, whereas some meetings in AMI Meeting Corpus consisted of meetings where subjects did not know each other. This fact might have an impact on interactions among participants and the resulting meeting satisfaction.

**Implication of the role of attendees in the company or within the meeting on the meeting quality.** We analyzed the effect of the employee role and attendees meeting role on the perceived meeting quality. An interesting result is that meeting satisfaction is rated differently by employees of different roles: individual contributors showed a higher meeting satisfaction than leader/managers. Interestingly, there was no significant difference in perceived meeting quality between meeting participants and chairpersons who led the meeting. These findings contradict the results of existing research that the meeting facilitators or employees with higher positions rate the meeting quality higher than people who are not in those roles [CRAL11]. This effect is called self-serving bias and happens because people evaluate subjective situations in a way that they reflect their interests [CRAL11]. Moreover, employees in senior positions tend to show higher meeting effectiveness than juniors [LRWB09]. Both employee roles (participant and leader/manager) showed considerable variability among their years of experience. This fact implies that there might not be a clear distinction between senior and junior employees solely on this information. That being said, we can not originate the discrepancy entirely. Therefore, it should be further investigated in future work.

Another interesting finding was that meetings with a chairperson resulted in a higher meeting satisfaction than meetings without having an attendee in the role of being the leader of this meeting. This finding can be supported through existing research which highlights the importance of a leader within meetings who manages a group to achieve goals and stay on the agenda throughout the meeting [BCBM18, YCA15]. However, a leader only has a positive impact on meeting outcomes, if they do not lead poorly [Per09].

**Fostering meeting quality through awareness.** As shown in related work, meeting information systems can be divided into either participant-centered which promote self-reflection and behavior change, or content-centered which support the memorization of facts and generation of new ideas. We focus on the first, self-reflection and behavior change. Despite the general interest of researching group mirrors and their effects on social group behaviors such as balanced speaking time, there is no research about their implications on participants' perceived meeting quality. This analysis is part of future work and might reveal valuable insights, in particular, because participation equality which is capturable with the audio recording seems to be unrelated to participants' meeting satisfaction [LCR13].

So far, it is challenging to automatically assess the overall quality of the meeting using solely features extractable from the audio recording which applies for real-time and retrospective measurements of the quality. Lai and Murray [LM18] were able to automatically predict group satisfaction in retrospective using machine learning and a variety of features extractable from speech. However, we do not believe their approach is currently applicable in non-scenario meetings or on different meeting types because of two reasons. First, their models are trained on the AMI Meeting Corpus which contains scenario meetings where all participants' had to contribute to complete a given task whereas in naturally occurring meetings it might be the case that not everyone contributes or only contributes at the beginning. Second, they predict three aspects of satisfaction of the AMI post-meeting survey. The prediction achieved the best performance if each aspect is prediction uses an individual model with selected features. Therefore, we conclude that predicting meeting satisfaction is complex and the generalizability of their approach is yet not clear and researched. This indicates that currently, it is challenging to measure meeting satisfaction using speech features automatically.

However, we thought about another way to automatically assess the meeting quality and

to create awareness of it to increase it potentially. Our idea is an implementation of a meeting evaluation system which is in the scope for future work. This system measures the meeting quality using short meeting surveys which the attendees answer after each meeting. A possibility could be to use an already established post-meeting survey [PKS18]. Moreover, the system would be integrated into the companies calendar tool and would automatically present the survey to all employees who attended the meeting using a pop-up containing the questions. The primary use case of this system is to enable the employees reviewing the meeting qualities by observing various visualizations. For example, employees can assess the average meeting quality over time or the distribution of meeting quality levels. The visualized information is an aggregated group-level quality and calculated based on the participants' answers for each meeting. Moreover, the tool could be extended with selectable keywords which allows the meeting participants to indicate more precisely what influenced their perception of the meeting quality. This extension would increase the information content which could lead to a more specific intervention. This system enables the user to assess the overall quality of meetings. Therefore, meeting attendees, especially chairpersons, should review and reflect them periodically. This reflection helps to prevent meeting frustration [LWAB16] and is an useful technique to recognize meeting problems [ARS08]. Therefore, we see a potential for this system to foster meeting quality.

Besides this group-based awareness system, we propose an individual-based tool which dynamically and in real-time provides individual feedback depending on the participants' current meeting behavior which is shown on the participant's device (e.g. smartphone, tablet or laptop). Since subtle feedback is slightly more efficient and accepted than a text message [SCM<sup>+</sup>14], we suggest providing a pictographic representation of individual feedback instead of a textual facilitation message. Our tool would not only intervene to adjust the participant's contribution rate but also in case, the participant showed too much negative emotional speech, e.g. through an angry voice. Additionally, the system indicates if the attendee interrupts another member many times. Therefore, in contrast to existing work, we not only focus on providing awareness on the equal contribution but also the interruption-behaviour and emotional speech of participants. We chose to integrate the last two features since they might be related to an open meeting atmosphere. For example, the tool notifies the attendee if he uses a negative tone because a negative tone could lead to intimidated participants, which contradicts an open atmosphere.

Based on the existing experiences of behavioral changes triggered by feedback messages to balance the distribution of speech time among all members (e.g. [AMS17]), we believe that these two additional information channels could also lead to behavioral changes. Therefore, we think this system supports maintaining an open atmosphere and hence, might positively impact participants' perceived meeting quality. However, future work is necessary to analyze whether this individual feedback, especially the two additional information channels, would lead to an adjustment of participant's behaviour as well as its impact on the meeting quality.

For future work, we thought about a metaphorical representation of the participant's meeting behaviour using a flower. Since the visualization is intended to draw the participant's attention to his own behavior, we recommend that it is displayed on the participant's personal device and not on a shared dashboard. Thus each participant only sees his own behaviour and not that of the others. The size of the flower represents the contribution rate: the larger the flower the more is the participant contributing to the meeting. The flower's leaves will represent the interruptions. Every time the user interrupts another attendee, the flower will lose leaves which will grow again if the participant significantly changed his behavior and did not interrupt anyone for a certain amount of time. We will visualize the detected emotions from the speech by smoothly changing the flower's color from green (positive emotions) to red (negative emotions). By looking at the flower, the participant notices whether they should adjust their behavior in order to enable other participants to contribute more.

**Privacy** A major aspect of all meeting information systems is privacy concerns, which definitely should be considered and respected. Meeting attendees may not only feel discomfort while being recorded but also, showing too personal data could expose, insult or threaten the participants [ZBL<sup>+</sup>11]. Moreover, meetings are highly confidential data. Data storage, processing and access need to be accurately specified and securely implemented. Consequently, users should only have insights to the meetings in which they attended. Additionally, if the participants interact with the system, the information they receive should be entirely anonymized. For example, the proposed meeting evaluation system does not reveal the meeting satisfaction of each attendee but the aggregated group satisfaction.



# Conclusion

Meetings are ubiquitous in today's work environments and employees spend a significant amount of time in them. Unfortunately, they are often perceived as unproductive or a waste of time. Attending bad meetings can not only negatively affect the general mood of employees but also their well-being. Related work analyzed whether the participants' quality on an aggregated group-level value can be predicted. So far, their approaches either depend on the use of manual coding or on meetings in which participants are given a role and task. Therefore, this thesis explores whether it is possible to automatically measure participants' quality of time spent in meetings using extractable features of audio recordings from non-scenario meetings.

To analyze these objectives, we designed and conducted a multi-day field study with professionals in a company based in Germany. Overall, we collected the raw audio recording of 25 meetings and 78 post-meeting surveys answered from their attendees. Additionally, we developed an approach to extract speech-related features of a raw audio file which we used to analyze the collected audio recordings. Based on the collected data and our approach, we performed an empirical analysis. We analyzed meeting quality in two aspects: meeting satisfaction and quality of time spent. The results of our survey indicate that an open and positive meeting atmosphere positively affects participants' meeting quality. Moreover, satisfying meetings were characterized by a lively and active exchange. In addition, our findings show a higher meeting satisfaction for individual contributors than for leader or managers. Results further suggest that the number of speaking turns is linked to a higher meeting satisfaction but not to the quality of time spent. However, based on our analysis, this speech feature was the only one that was related to meeting quality. Due to our results and those from previous work, we conclude that it is currently challenging to assess meeting quality solely on the extracted features. We recommend concentrating on measuring an open atmosphere or active exchange, both of which correlate positively with the meeting quality and could therefore automatically capture it.

Nevertheless, we see a potential to increase the meeting quality through two independent systems which provide awareness about participant's meeting quality and meeting behavior by continuously prompted retrospection and an individual contribution dashboard, respectively. The first system is a meeting evaluation system that collects participant's quality using a post-meeting survey. This information will be aggregated on a group-level value in order to provide continuous feedback on participants' meeting quality. This retrospective analysis enables to periodically assess the ratings, identify trends and observe if intervention is necessary. The second, is a tool providing individual feedback on participant's device by graphically representing the participant's meeting behaviors, such as contribution rate, interruptions of other group members and emotional speech. Using this approach participants can in real-time assess whether they should change their behavior in order to create more open communication.



---

# Bibliography

- [ABE<sup>+</sup>12] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370, 2012.
- [ABSR14] Joseph A. Allen, Tammy Beck, Cliff W. Scott, and Steven G. Rogelberg. Understanding workplace meetings: A qualitative taxonomy of meeting purposes. *Management Research Review*, 37(9):791–814, 2014.
- [ALWR18] Joseph A. Allen, Nale Lehmann-Willenbrock, and Steven G. Rogelberg. Let’s get this meeting started: Meeting lateness and actual meeting outcomes. *Journal of Organizational Behavior*, 39(8):1008–1021, 2018.
- [Ami19] Sehili Amine. Audio tokenizer. <https://github.com/amsehili/auditok>, 2019. Retrieved April 17, 2019.
- [AMS17] Hiroyuki Adachi, Seiko Myojin, and Nobutaka Shimada. Activation of a co-located meeting by visualizing conversation and facilitating participant’s behavior. *Transactions of the Institute of Systems, Control and Information Engineers*, 30(11):427–438, 2017.
- [ARS08] Joseph A. Allen, Steven G. Rogelberg, and John C. Scott. Mind your meetings: Improve your organization’s effectiveness one meeting at a time. *Quality Progress*, 41:48–53, 2008.
- [AS07] I. Elaine Allen and Christopher A. Seaman. Likert scales and data analyses. *Quality progress*, 40(7):64–65, 2007.
- [ASM<sup>+</sup>12] Joseph A. Allen, Stephanie J. Sands, Stephanie L. Mueller, Katherine A. Frear, Mara Mudd, and Steven G. Rogelberg. Employees’ feelings about more meetings: An overt analysis and recommendations for improving meetings. *Management Research Review*, 35(5):405–418, 2012.
- [BBC<sup>+</sup>19] Adam Bergkvist, Daniel C. Burnett, Jennings Cullen, Narayanan Anant, Adoba Bernard, Brandstetter Taylor, Boström Henrik, and Bruaroey Jan-Ivar. Webrtc 1.0: Real-time communication between browsers. <http://w3c.github.io/webrtc-pc/>, 2019. Retrieved April 17, 2019.
- [BCBM17] Cigdem Beyan, Francesca Capozzi, Cristina Becchio, and Vittorio Murino. Multi-task learning of social psychology assessments and nonverbal features for automatic leadership identification. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI ’17*, pages 451–455, New York, NY, USA, 2017. ACM.

- [BCBM18] Cigdem Beyan, Francesca Capozzi, Cristina Becchio, and Vittorio Murino. Prediction of the leadership style of an emergent leader using audio and visual nonverbal features. *Trans. Multi.*, 20(2):441–456, 2018.
- [BCC<sup>+</sup>16] Cigdem Beyan, Nicolò Carissimi, Francesca Capozzi, Sebastiano Vascon, Matteo Bustreo, Antonio Pierro, Cristina Becchio, and Vittorio Murino. Detecting emergent leader in a meeting environment using nonverbal visual features only. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI '16*, pages 317–324, New York, NY, USA, 2016. ACM.
- [BDL<sup>+</sup>18] Pierre-Alexandre Broux, Florent Desnoux, Anthony Larcher, Simon Petitrenaud, Jean Carrire, and Sylvain Meignier. S4d: Speaker diarization toolkit in python. In *Interspeech 2018*, 2018.
- [BFZ<sup>+</sup>18] Indrani Bhattacharya, Michael Foley, Ni Zhang, Tongtao Zhang, Christine Ku, Cameron Mine, Heng Ji, Christoph Riedl, Brooke Foucault Welles, and Richard J Radke. A multimodal-sensor-enabled room for unobtrusive group meeting analysis. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 347–355. ACM, 2018.
- [BK07a] Tony Bergstrom and Karrie Karahalios. Conversation clock: Visualizing audio patterns in co-located groups. In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, pages 78–78. IEEE, 2007.
- [BK07b] Tony Bergstrom and Karrie Karahalios. Seeing more: visualizing audio cues. In Cécilia Baranauskas, Philippe Palanque, Julio Abascal, and Simone Diniz Junqueira Barbosa, editors, *Human-Computer Interaction – INTERACT 2007*, pages 29–42. Springer, Berlin, Heidelberg, 2007.
- [BKD08] Khaled Bachour, Frédéric Kaplan, and Pierre Dillenbourg. Reflect: An interactive table for regulating face-to-face collaborative learning. In *Times of Convergence. Technologies Across Learning Contexts*, pages 39–48. Springer, Berlin, Heidelberg, 2008.
- [BL94] Margaret M. Bradley and Peter J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.
- [BM18] McKenzie Braley and Gabriel Murray. The group affect and performance (gap) corpus. In *Proceedings of the Group Interaction Frontiers in Technology*, page 2. ACM, 2018.
- [Boe93] Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the institute of phonetic sciences*, volume 17, pages 97–110. Amsterdam, 1993.
- [Boe02] Paul Boersma. Praat, a system for doing phonetics by computer. *Glott international*, 5:341–245, 2002.
- [BRdV06] Robert O. Briggs, Bruce A. Reinig, and Gert-Jan de Vreede. Meefting satisfaction for technology-supported groups: An empirical validation of a goal-attainment model. *Small Group Research*, 37(6):585–611, 2006.



- [Bre17] Hervé Bredin. pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, 2017.
- [BV16] Alberto Betella and Paul F. M. J. Verschure. The affective slider: A digital self-assessment scale for the measurement of human emotions. *PLOS ONE*, 11(2):1–11, 2016.
- [Car07] Jean Carletta. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41(2):181–190, 2007.
- [CBL13] Delphine Charlet, Claude Barras, and Jean-Sylvain Liénard. Impact of overlapping speech detection on speaker diarization for broadcast news and debates. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7707–7711. IEEE, 2013.
- [CBS<sup>+</sup>19] Senthil Chandrasegaran, Chris Bryan, Hidekazu Shidara, Tung-Yen Chuang, and Kwan-Liu Ma. Talktraces: Real-time capture and visualization of verbal content in meetings. In *Proceedings of ACM CHI Conference on Human Factors in Computing Systems (CHI 2019)*. ACM, 2019.
- [CDL<sup>+</sup>18] Zhehuai Chen, Jasha Droppo, Jinyu Li, Wayne Xiong, Zhehuai Chen, Jasha Droppo, Jinyu Li, and Wayne Xiong. Progressive joint modeling in unsupervised single-channel overlapped speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 26(1):184–196, 2018.
- [Cor] IBM Corp. *IBM SPSS Statistics for Macintosh, Version 25.0*. Armonk, NY. Released 2017.
- [CRAL11] Melissa A. Cohen, Steven G. Rogelberg, Joseph A. Allen, and Alexandra Luong. Meeting design characteristics and attendee perceptions of staff/team meeting quality. *Group Dynamics: Theory, Research, and Practice*, 15(1):90, 2011.
- [CSR16] Shammur Absar Chowdhury, Evgeny A. Stepanov, and Giuseppe Riccardi. Predicting user satisfaction from turn-taking in spoken conversations. In *Interspeech 2016*, pages 2910–2914, San Francisco, USA, 2016.
- [CTS17] Pawel Cyrta, Tomasz Trzciński, and Wojciech Stokowiec. Speaker diarization using deep recurrent convolutional neural networks for speaker embeddings. In *International Conference on Information Systems Architecture and Technology*, pages 107–117. Springer, 2017.
- [Dav97] Robert Davison. An instrument for measuring meeting success. *Information & management*, 32(4):163–176, 1997.
- [DBCS04] Rajdip Dhillon, Sonali Bhagat, Hannah Carvey, and Elizabeth Shriberg. Meeting recorder project: Dialog act labeling guide. Technical report, International Computer Science Institute, 2004.
- [DCV<sup>+</sup>18] David Doukhan, Jean Carrière, Félicien Vallet, Anthony Larcher, and Sylvain Meignier. An open-source speaker gender detection framework for monitoring gender equality. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5214–5218. IEEE, 2018.

- [DeW19] DeWave. Single-channel blind source separation. <https://github.com/chaodengusc/DeWave>, 2019. Retrieved April 07, 2019.
- [DJW09] Nivja H. De Jong and Ton Wempe. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2):385–390, 2009.
- [DPB04] Joan Morris DiMicco, Anna Pandolfo, and Walter Bender. Influencing group participation with a shared display. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work, CSCW '04*, pages 614–623, New York, NY, USA, 2004. ACM.
- [EWGS13] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM, 2013.
- [eXc19] SoX Sound eXchange. <http://sox.sourceforge.net>, 2019. Retrieved April 05, 2019.
- [FH09] Mireia Farrús and Javier Hernando. Using jitter and shimmer in speaker verification. *IET Signal Processing*, 3(4):247–257, 2009.
- [Ger03] David Gerhard. *Pitch extraction and fundamental frequency: History and current techniques*. Department of Computer Science, University of Regina Regina, Canada, 2003.
- [GGPL17] Jose Maria Garcia-Garcia, Victor M. R. Penichet, and Maria D. Lozano. Emotion detection: A technology review. In *Proceedings of the XVIII International Conference on Human Computer Interaction, Interacción '17*, pages 8:1–8:8, New York, NY, USA, 2017. ACM.
- [Gia15] Theodoros Giannakopoulos. pyaudioanalysis: An open-source python library for audio signal analysis. *PLoS one*, 10(12):e0144610, 2015.
- [Gmb] Limesurvey GmbH. Limesurvey: An open source survey tool. <http://www.limesurvey.org>. Retrieved April 07, 2019.
- [Gra00] Lincoln Gray. Background science. *Journal of Perinatology*, 20:S5–S10, 2000.
- [Hac96] Tamas Hacki. Comparative speaking, shouting and singing voice range profile measurement: physiological and pathological aspects. *Logopedics Phoniatrics Vocology*, 21(3-4):123–129, 1996.
- [HSS08] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.
- [Idi19a] Idiap. Ibdiarization. <https://github.com/idiap/IBDiarization>, 2019. Retrieved April 19, 2019.
- [Idi19b] Idiap. pydiarization. <https://github.com/idiap/IBDiarization/tree/master/src/pydiarization>, 2019. Retrieved April 19, 2019.

- [IFM<sup>+</sup>05] Dan Istrate, Corinne Fredouille, Sylvain Meignier, Laurent Besacier, and Jean François Bonastre. Nist rt'05s evaluation: pre-processing techniques and speaker diarization on multiple microphone meetings. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 428–439. Springer, 2005.
- [Jay11] Dinesh Babu Jayagopi. *Computational modeling of face-to-face social interaction using nonverbal behavioral cues*. PhD thesis, École Polytechnique Fédérale De Lausanne, 2011.
- [JBOGP08] Dinesh Babu Jayagopi, Sileye Ba, Jean-Marc Odobez, and Daniel Gatica-Perez. Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues. In *Proceedings of the 10th International Conference on Multimodal Interfaces, ICMI '08*, pages 45–52, New York, NY, USA, 2008. ACM.
- [JSCO<sup>+</sup>12] Dineshbabu Jayagopi, Dairazalia Sanchez-Cortes, Kazuhiro Otsuka, Junji Yamato, and Daniel Gatica-Perez. Linking speaking and looking behavior patterns with group composition, perception, and performance. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 433–440. ACM, 2012.
- [JTDB18] Yannick Jadoul, Bill Thompson, and Bart De Boer. Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71:1–15, 2018.
- [KCHP08] Taemie Kim, Agnes Chang, Lindsey Holland, and Alex (Sandy) Pentland. Meeting mediator: Enhancing group collaboration with sociometric feedback. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems, CHI EA '08*, pages 3183–3188, New York, NY, USA, 2008. ACM.
- [KG15] Tine Köhler and Markus Gözl. *Meetings across Cultures*, page 119–150. Cambridge Handbooks in Psychology. Cambridge University Press, 2015.
- [KGF02] Robert E. Kraut, Darren Gergle, and Susan R. Fussell. The use of visual information in shared visual spaces: Informing the development of virtual co-presence. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work, CSCW '02*, pages 31–40, New York, NY, USA, 2002. ACM.
- [KK09] Tony Karahalios and Karrie Karahalios. Social mirrors as social signals: transforming audio into graphics. *IEEE Computer Graphics and Applications*, 29(5):22–32, 2009.
- [Kle10] Lisa Kleinman. *Physically present, mentally absent? Technology multitasking in organizational meetings*. PhD thesis, University of Texas, 2010.
- [KLW12] Simone Kauffeld and Nale Lehmann-Willenbrock. Meetings matter: Effects of team meetings on team and organizational success. *Small Group Research*, 43(2):130–158, 2012.
- [KMB18] Shashidhar G. Koolagudi, YV Srinivasa Murthy, and Siva P. Bhaskar. Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition. *International Journal of Speech Technology*, 21(1):167–183, 2018.
- [KPW15] Roi Kliper, Shirley Portuguese, and Daphna Weinshall. Prosodic analysis of speech and the underlying mental state. In *International Symposium on Pervasive Computing Paradigms for Mental Health*, pages 52–62. Springer, 2015.

- [KWT05] Olga Kulyk, Jimmy Wang, and Jacques Terken. Real-time feedback on nonverbal behaviour to enhance social dynamics in small group meetings. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 150–161. Springer, 2005.
- [LCR13] Catherine Lai, Jean Carletta, and Steve Renals. Modelling participant affect in meetings with turn-taking features. In *Proc. Workshop of Affective Social Speech Signals*, 2013.
- [LM18] Catherine Lai and Gabriel Murray. Predicting group satisfaction in meeting discussions. In *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data*, page 1. ACM, 2018.
- [LR05] Alexandra Luong and Steven G. Rogelberg. Meetings and more meetings: The relationship between meeting load and the daily well-being of employees. *Group Dynamics: Theory, Research, and Practice*, 9(1):58, 2005.
- [LRS<sup>+</sup>18] Moon-Hwan Lee, Yea-Kyung Row, Oosung Son, Uichin Lee, Jaejeung Kim, Jungi Jeong, Seungryoul Maeng, and Tek-Jin Nam. Flower-pop: Facilitating casual group conversations with multiple mobile devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(4):150:1–150:24, 2018.
- [LRWB09] Desmond J. Leach, Steven G. Rogelberg, Peter B. Warr, and Jennifer L. Burnfield. Perceived meeting effectiveness: The role of design characteristics. *Journal of Business and Psychology*, 24(1):65–76, 2009.
- [LTJ<sup>+</sup>07] Xi Li, Jidong Tao, Michael T Johnson, Joseph Soltis, Anne Savage, Kirsten M Leong, and John D Newman. Stress and emotion classification using jitter and shimmer features. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–1081. IEEE, 2007.
- [LWA14] Nale Lehmann-Willenbrock and Joseph A. Allen. How fun are your meetings? investigating the relationship between humor patterns in team interactions and team performance. *Journal of Applied Psychology*, 99(6):1278, 2014.
- [LWAB16] Nale Lehmann-Willenbrock, Joseph A. Allen, and Dain Belyeu. Our love/hate relationship with meetings: Relating good and bad meeting behaviors to meeting outcomes, engagement, and exhaustion. *Management Research Review*, 39(10):1293–1312, 2016.
- [Mas02] Marianne Schmid Mast. Dominance as expressed and inferred through speaking time: A meta-analysis. *Human Communication Research*, 28(3):420–450, 2002.
- [MAVS18] Joseph E. Mroz, Joseph A. Allen, Dana C. Verhoeven, and Marissa L. Shuffler. Do we really need another meeting? the science of workplace meetings. *Current Directions in Psychological Science*, 27(6):484–491, 2018.
- [Max19] Hollmann Max. Lium diarization editor. <https://github.com/maxhollmann/lium-diarization-editor>, 2019. Retrieved April 17, 2019.
- [McK11] Wes McKinney. pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, 14, 2011.
- [MEF<sup>+</sup>10] Benoit Mathieu, Slim Essid, Thomas Fillon, Jacques Prado, and Gaël Richard. Yaafe, an easy to use and efficient audio feature extraction software. In *ISMIR*, pages 441–446, 2010.

- [MFMZ14] André N. Meyer, Thomas Fritz, Gail C. Murphy, and Thomas Zimmermann. Software developers' perceptions of productivity. In *Proceedings of the 22Nd ACM SIGSOFT International Symposium on Foundations of Software Engineering, FSE 2014*, pages 19–29, New York, NY, USA, 2014. ACM.
- [Mit] Puthran Mitesh. Speech emotion analyzer. <https://github.com/MITESHPUTHRANNEU/Speech-Emotion-Analyzer>. Retrieved April 17, 2019.
- [MK18] Benjamin Milde and Arne Köhn. Open source automatic speech recognition for german. In *Proceedings of ITG 2018*, 2018.
- [MM10] Sylvain Meignier and Teva Merlin. Lium spkdiarization: an open source toolkit for diarization. In *CMU SPUD Workshop*, 2010.
- [MO14] Antonio Macías Ojeda. Speaker diarization. Master's thesis, Aalto University, 2014.
- [Moz19] Mozilla. Deep speech. <https://github.com/mozilla/DeepSpeech>, 2019. Retrieved April 17, 2019.
- [MRL<sup>+</sup>15] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25, 2015.
- [Mur14] Gabriel Murray. Learning how productive and unproductive meetings differ. In *Canadian Conference on Artificial Intelligence*, pages 191–202. Springer, 2014.
- [Mur15] Gabriel Murray. Analyzing productivity shifts in meetings. In *Canadian Conference on Artificial Intelligence*, pages 141–154. Springer, 2015.
- [NE17] KARIN Niemantsverdriet and Thomas Erickson. Recurring meetings: an experiential account of repeating meetings in a large organization. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1, 2017.
- [NL92] Carol T. Nixon and Glenn E. Littlepage. Impact of meeting procedures on meeting effectiveness. *Journal of Business and Psychology*, 6(3):361–369, 1992.
- [OKK16] Isabelle Odermatt, Cornelius J. König, and Martin Kleinmann. Development and validation of the zurich meeting questionnaire (zmq). *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, 66(5):219–232, 2016.
- [OKK<sup>+</sup>18] Isabelle Odermatt, Cornelius J. König, Martin Kleinmann, Maria Bachmann, Heiko Röder, and Patricia Schmitz. Incivility in meetings: Predictors and outcomes. *Journal of Business and Psychology*, 33(2):263–282, 2018.
- [Per09] Robert D. Perkins. How executive coaching can change leader behavior and improve meeting effectiveness: An exploratory study. *Consulting Psychology Journal: Practice and Research*, 61(4):298, 2009.
- [PKS18] Nils Prenner, Jil Klünder, and Kurt Schneider. Making meeting success measurable by participants' feedback. In *Proceedings of the 3rd International Workshop on Emotion Awareness in Software Engineering, SEmotion '18*, pages 25–31, New York, NY, USA, 2018. ACM.
- [RAML<sup>+</sup>15] Stephan Radeck-Arneth, Benjamin Milde, Arvid Lange, Evandro Gouvea, Stefan Radomski, Max Mühlhäuser, and Chris Biemann. Open Source German Distant Speech Recognition: Corpus and Acoustic Model. In *Proceedings Text, Speech and Dialogue (TSD)*, pages 480–488, Pilsen, Czech Republic, 2015.

- [RAS<sup>+</sup>10] Steven G. Rogelberg, Joseph A. Allen, Linda Shanock, Cliff Scott, and Marissa Shufler. Employee satisfaction with meetings: A contemporary facet of job satisfaction. *Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management*, 49(2):149–172, 2010.
- [RDLH<sup>+</sup>18] Kimiko Ryokai, Elena Durán López, Noura Howell, Jon Gillick, and David Bamman. Capturing, representing, and interacting with laughter. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 358:1–358:12, New York, NY, USA, 2018. ACM.
- [Rey02] Douglas A. Reynolds. An overview of automatic speaker recognition technology. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–4072. IEEE, 2002.
- [RH10] Stuart Rosen and Peter Howell. *Signals and systems for speech and hearing*, volume 29. Brill, 2010.
- [RLWB06] Steven G. Rogelberg, Desmond J. Leach, Peter B. Warr, and Jennifer L. Burnfield. "not another meeting!" are meeting time demands related to employee well-being? *Journal of Applied Psychology*, 91(1):83, 2006.
- [RMM<sup>+</sup>12] Flaviu Roman, Stefano Mastrogiacomio, Dyna Mlotkowski, Frédéric Kaplan, and Pierre Dillenbourg. Can a table regulate participation in top level managers' meetings? In *Proceedings of the 17th ACM International Conference on Supporting Group Work*, GROUP '12, pages 1–10, New York, NY, USA, 2012. ACM.
- [RQH10] Robert Remus, Uwe Quasthoff, and Gerhard Heyer. Sentiws-a publicly available german-language resource for sentiment analysis. In *Proceedings of the 7th International Language Ressources and Evaluation (LREC'10)*, pages 1168–1171, 2010.
- [RSK07] Steven G. Rogelberg, Cliff Scott, and John Kello. The science and fiction of meetings. *MIT Sloan management review*, 48(2):18–21, 2007.
- [SBB<sup>+</sup>18] Yang Shi, Chris Bryan, Sridatt Bhamidipati, Ying Zhao, Yaoxue Zhang, and Kwan-Liu Ma. Meetingvis: Visual narratives to assist in recalling meeting context and content. *IEEE transactions on visualization and computer graphics*, 24(6):1918–1929, 2018.
- [SCAMGP12] Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast, and Daniel Gatica-Perez. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia*, 14(3):816–832, 2012.
- [SCM<sup>+</sup>14] Gianluca Schiavo, Alessandro Cappelletti, Eleonora Mencarini, Oliviero Stock, and Massimo Zancanaro. Overt or subtle? supporting group conversations with automatically targeted directives. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 225–234. ACM, 2014.
- [SGS<sup>+</sup>12] Stefan Scherer, Michael Glodek, Friedhelm Schwenker, Nick Campbell, and Günther Palm. Spotting laughter in natural multiparty conversations: A comparison of automatic online and offline approaches using audiovisual data. *ACM Trans. Interact. Intell. Syst.*, 2(1):4:1–4:31, 2012.

- [SGS<sup>+</sup>18] Neeraj Sajjan, Shobhana Ganesh, Neeraj Sharma, Sriram Ganapathy, and Neville Ryant. Leveraging lstm models for overlap detection in multi-party meetings. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5249–5253. IEEE, 2018.
- [SHET07] Janienke Sturm, Olga Houben-van Herwijnen, Anke Eyck, and Jacques Terken. Influencing social dynamics in meetings through a peripheral display. In *Proceedings of the 9th International Conference on Multimodal Interfaces, ICMI '07*, pages 263–270, New York, NY, USA, 2007. ACM.
- [SK13] Nils Christian Sauer and Simone Kauffeld. Meetings as networks: Applying social network analysis to team interaction. *Communication Methods and Measures*, 7(1):26–47, 2013.
- [SK16] Nils Christian Sauer and Simone Kauffeld. The structure of interaction at meetings: A social network analysis. *Zeitschrift für Arbeits-und Organisationspsychologie A&O*, 2016.
- [SSB01] Elizabeth Shriberg, Andreas Stolcke, and Don Baron. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *EUROSPEECH-2001*, pages 1359–1362, 2001.
- [SSH09] Sara Streng, Karsten Stegmann, Heinrich Hußmann, and Frank Fischer. Metaphor or diagram?: comparing different representations for group mirrors. In *Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open 24/7*, pages 249–256. ACM, 2009.
- [SWQ<sup>+</sup>17] Yang Shi, Yang Wang, Ye Qi, John Chen, Xiaoyao Xu, and Kwan-Liu Ma. Ideawall: Improving creative collaboration through combinatorial visual stimuli. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 594–603. ACM, 2017.
- [TE95] Hartmut Traunmüller and Anders Eriksson. The frequency range of the voice fundamental in the speech of male and female adults. *Unpublished manuscript*, 1995.
- [TOL13] João Paulo Teixeira, Carla Oliveira, and Carla Lopes. Vocal acoustic analysis–jitter, shimmer and hnr parameters. *Procedia Technology*, 9:1112–1122, 2013.
- [Tor18] Amirsina Torfi. Speechpy-a library for speech processing and recognition. *arXiv preprint arXiv:1803.01094*, 2018.
- [WFHM18] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2018. R package version 3.5.2.
- [Wis19] Wiseman. py-webrtcvad, version 2.0.10. <https://github.com/wiseman/py-webrtcvad>, 2019. Retrieved April 17, 2019.
- [WKKO18] Tzu-Yang Wang, Ikkaku Kawaguchi, Hideaki Kuzuoka, and Mai Otsuki. Effect of manipulated amplitude and frequency of human voice on dominance and persuasiveness in audio conferences. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):177, 2018.
- [YBB18] Ruiqing Yin, Hervé Bredin, and Claude Barras. Neural Speech Turn Segmentation and Affinity Propagation for Speaker Diarization. In *19th Annual Conference of the International Speech Communication Association, Interspeech 2018, Hyderabad, India, 2018*.

- [YCA15] Michael Yoerger, John Crowe, and Joseph A. Allen. Participate or else!: The effect of participation in decision-making in meetings on employee engagement. *Consulting Psychology Journal: Practice and Research*, 67(1):65, 2015.
- [Yel15] Sree Harsha Yella. *Speaker diarization of spontaneous meeting room conversations*. PhD thesis, École Polytechnique Fédérale De Lausanne, 2015.
- [ZBL<sup>+</sup>11] Ying Zhang, Marshall Bern, Juan Liu, Kurt Partridge, Bo Begole, Bob Moore, Jim Reich, and Koji Kishimoto. Ubiquitous meeting facilitator with playful real-time user interface. In *International Conference on Ubiquitous Intelligence and Computing*, pages 3–11. Springer, Berlin, Heidelberg, 2011.
- [ZLPCT18] Chengwei Zhang, Marcelo López-Parra, Junyu Chen, and Ling Tian. Costorm: a term map system to aid in a collaborative ideation process. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, pages 1–12, 2018.



# Contents on CD-ROM

- **Zusfsg.txt**  
German version of the abstract of this thesis.
- **Abstract.txt**  
English version of the abstract of this thesis.
- **MasterThesis.pdf**  
Copy of this thesis.
- **ConsentForm.pdf**  
Consent Form for the Field-Study.
- **SourceCode.zip**  
Source Code of Approach to Automatically Extract Features
- **Analysis.zip**  
Scripts of the Analysis



# **Post-Meeting Survey**

## Questions about your perception of this meeting

### 1. In general, how satisfied have you been with this meeting?






☐ Not satisfied at all
 ☐ Not satisfied
 ☐ Neutral
 ☐ Satisfied
 ☐ Extremely satisfied

### 2. Please indicate for each statement about this meeting how strongly you agree or disagree.

	Strongly disagree	Dis-agree	Slightly disagree	Neither agree nor disagree	Slightly agree	Agree	Strongly agree	No answer
This meeting was useful for me.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My attendance in this meeting was relevant.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
At least one meeting participant was dominant.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The atmosphere in this meeting was positive and open, and I could contribute with ease.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The meeting members interrupted each other.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The exchange in this meeting was lively and active.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
There was one or more key people in this meeting.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

### 3. Please answer this question by completing the statement below:

I felt my time in this meeting was spent ...

☐  Not well at all
 ☐  Not well
 ☐  Neutral
 ☐  Well
 ☐  Extremely well

### 4. Did any of the following aspects affect your perception of your time spent in this meeting?

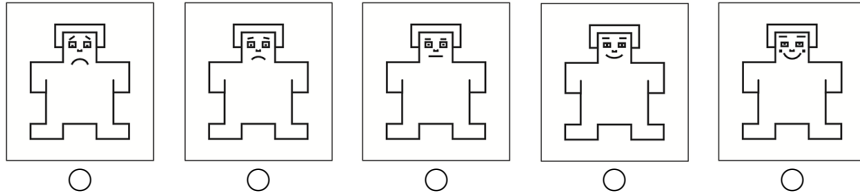
	No, did not affect	Yes, negatively affected	Yes, positively affected	Not applicable
Distribution of speech time across members (e.g. at least one speaker was dominant, all participated equally).	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Interruptions between members.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The usefulness of this meeting for you.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Your relevance in this meeting.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Positive and open meeting atmosphere.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Lively and active exchange in this meeting.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Presence of one or more key people.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

### 5. Were there any other factors that affected how well the meeting went?

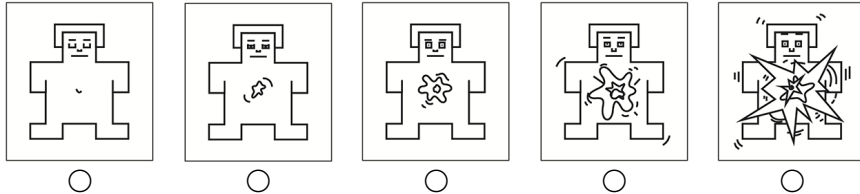
☐ No  
☐ Yes: \_\_\_\_\_  
 \_\_\_\_\_

**6. Please choose for each row the image representing you in the meeting the best.**

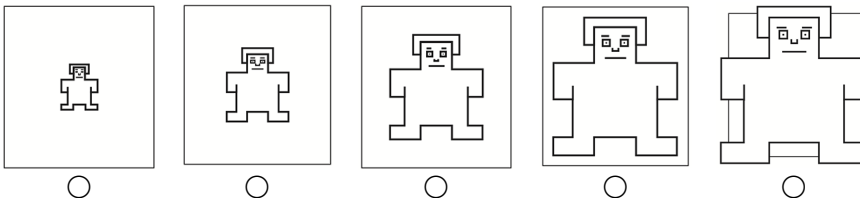
a) Valence / Negative & Positive



b) Arousal / Activation



c) Dominance



**Questions about the meeting structure**

7. Date of the meeting: \_\_\_\_\_
8. Start time of the meeting: \_\_\_\_\_
9. Number of meeting participants (including you): \_\_\_\_\_
10. What type of meeting was it?
  - ☐ One-On-One
  - ☐ One-To-Many (e.g. presentation)
  - ☐ Many-To-Many (e.g. discussion, brainstorming, status updates)
11. Which of the following best describes your role in this meeting?
  - ☐ Chairperson (controls the meeting)
  - ☐ Participant
  - ☐ Other: \_\_\_\_\_

**Questions about your background**

12. How many years of professional work experience do you have in your role? \_\_\_\_\_
13. Which of the following best describes your role in the company?
  - ☐ Contributor
  - ☐ Leader
  - ☐ Manager
  - ☐ Other: \_\_\_\_\_
14. What is your gender?
  - ☐ Female
  - ☐ Male
  - ☐ Other: \_\_\_\_\_
  - ☐ Prefer not to answer

Thank you very much for your time and valuable help!



# Overview Meetings

In Table C.1, an overview of the captured meetings is provided - along with the information about its size, duration, time-well-spent-score and meeting satisfaction-score on a group-level.

Id	Size	Duration ( <i>hh : mm : ss</i> )	Time-Well-Spent-Score	Satisfaction-Score
1	2	00:46:39	2.00	1.50
2	4	00:27:29	1.67	1.67
3	4	01:40:53	1.67	1.00
4	2	00:15:38	1.50	1.00
5	2	00:29:17	1.50	1.00
6	5	00:56:52	1.40	1.20
7	3	00:34:01	1.33	1.00
8	3	00:16:13	1.33	1.00
9	3	00:38:50	1.33	0.67
10	4	00:15:18	1.00	1.00
11	2	00:51:25	1.00	1.00
12	3	00:17:33	1.00	1.00
13	3	00:58:16	1.00	1.00
14	2	00:24:19	1.00	1.00
15	2	00:34:41	1.00	1.00
16	8	00:44:17	1.00	1.00
17	3	00:58:03	1.00	1.00
18	2	01:34:56	1.00	1.00
19	8	00:58:48	1.00	0.80
20	4	01:40:50	1.00	0.75
21	3	00:25:33	1.00	0.67
22	4	00:27:21	1.00	0.50
23	4	00:29:28	1.00	0.50
24	3	00:23:51	0.67	1.00
25	7	00:28:42	0.50	0.67

**Table C.1:** Overview of Meetings



## Appendix D

# Descriptive Statistics Group-Level

Table D.1 provides an overview of mean, standard deviations, medians, minimum and maximum values of all variables on a group-level. Moreover, the results of the normality tests performed with Shapiro-Wilk are provided as well.

Variable	Mean	Std	Min	Max	Median	Shapiro (N=25)	Shapiro (N=18)
Time-well-spent	1.156	0.331	0.5	2.0	1.0	0.864*	0.869*
Meeting Satisfaction	0.957	0.26	0.5	1.667	1.0	0.827**	0.861*
Distribution of Speech Time <sup>a</sup>	-0.101	1.415	-3.0	2.333	-0.292	0.96	0.924
Interruptions <sup>a</sup>	-1.488	1.023	-3.0	1.0	-1.667	0.94	0.928
Meeting Usefulness <sup>a</sup>	2.218	0.569	1.167	3.0	2.333	0.93	0.895*
Attendee's Relevance <sup>a</sup>	2.318	0.501	1.0	3.0	2.333	0.885*	0.813*
Open Atmosphere <sup>a</sup>	2.397	0.517	1.0	3.0	2.417	0.904*	0.869*
Lively Exchange <sup>a</sup>	2.236	0.51	0.833	3.0	2.292	0.904*	0.913
Key People <sup>a</sup>	0.901	1.531	-2.5	3.0	1.467	0.918*	0.91
Meeting Duration (in centiseconds)	254207.32	152473.158	91782.0	605278.0	190453.0	0.842*	0.837*
Number of Atten- dees	3.12	1.333	1.0	6.0	3.0	0.904*	0.805*
Total speaking length	0.85	0.102	0.552	0.965	0.898	0.831**	0.719**
Total speaking turns <sup>b</sup>	0.102	0.031	0.047	0.166	0.096	0.958	0.964
Total silence length	0.15	0.102	0.035	0.448	0.102	0.831**	0.719**

Average Turn Duration	929.529	367.556	497.217	2057.227	954.022	0.906*	0.973
Total number of speaker changes <sup>b</sup>	0.054	0.019	0.019	0.097	0.054	0.919*	0.855*
Turn-taking freedom	0.578	0.15	0.308	0.805	0.591	—	0.944
Participation equality	0.83	0.154	0.387	0.995	0.86	0.889*	0.88*
Maximum percentage of speaking length	0.517	0.121	0.31	0.741	0.478	0.97	0.958
Maximum percentage of speaking turns	0.504	0.132	0.281	0.735	0.443	0.954	0.926
Maximum eigenvector centrality	0.649	0.043	0.567	0.709	0.654	—	0.943
Maximum degree centrality	38.661	31.93	10.571	136.5	28.667	—	0.783**
Total sum of laughs <sup>b</sup>	0.009	0.011	0.0	0.049	0.005	0.721**	0.754**
Mean Speechrate	4.496	0.315	4.003	5.063	4.668	0.934	0.934
Mean Articulation-rate	4.988	0.288	4.419	5.442	5.041	0.962	0.958
Mean Intensity	55.135	5.791	44.443	63.676	57.778	0.911*	0.917
Mean Pitch	175.118	14.093	149.48	212.584	175.805	0.957	0.961
Mean Local Jitter	0.035	0.005	0.025	0.044	0.035	0.949	0.946
Mean Local Shimmer	0.19	0.012	0.167	0.207	0.191	0.921	0.907

a. Agreement-score of question 2 of the survey

b. Normalized by meeting duration in seconds instead of centiseconds to provide more information

\*, Correlation significant at 0.05, \*\*, Correlation significant at 0.001 level

**Table D.1:** Descriptive Statistics of Group-Level Variables (*Std=Standard deviation, Min=Minimum value, Max=Maximum value, Shaprio (N=25)=Shapiro-Wilk's effect size on all meetings, Shapiro (N=18)=Shapiro-Wilk's effect size on meetings with more than two attendees.*)

## Appendix E

# SPSS Output Stepwise Linear Regression Analysis

The Tables E.1, E.1 and E.3 contain the outputs of the step-wise linear regression which is performed in SPSS. The dependent variable is the group-level time-well-spent score and the independent variables the seven meeting factors.

ANOVA <sup>a</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.620	1	.620	7.079	.014 <sup>b</sup>
	Residual	2.015	23	.088		
	Total	2.635	24			
2	Regression	.961	2	.480	6.311	.007 <sup>c</sup>
	Residual	1.674	22	.076		
	Total	2.635	24			

a. Dependent Variable: Time well spent

b. Predictors: (Constant), Atmosphere

c. Predictors: (Constant), Atmosphere, Relevance

Table E.1: ANOVA

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.485 <sup>a</sup>	.235	.202	.295972081
2	.604 <sup>b</sup>	.365	.307	.275873754

a. Predictors: (Constant), Atmosphere

b. Predictors: (Constant), Atmosphere, Relevance

Table E.2: Model Summary

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.410	.287		1.430	.166
	Atmosphere	.311	.117	.485	2.661	.014
2	(Constant)	-.051	.345		-.148	.884
	Atmosphere	.270	.111	.421	2.438	.023
	Relevance	.241	.114	.365	2.115	.046

**Coefficients<sup>a</sup>**

Model		95.0% Confidence Interval for B		Collinearity Statistics	
		Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-.183	1.003		
	Atmosphere	.069	.553	1.000	1.000
2	(Constant)	-.766	.664		
	Atmosphere	.040	.500	.969	1.032
	Relevance	.005	.478	.969	1.032

a. Dependent Variable: Time well spent

**Table E.3:** Coefficients