Sigl, Anna-Lena; Fleischmann, Carolin; Cardon, Peter W.; Aritz, Jolanta; Koße, Tamara

Working Paper

Speech-to-text technology for global virtual teams: A SWOT analysis.

Rosenheim Papers in Applied Economics and Business Sciences, No. 9/2023

Provided in Cooperation with:
Rosenheim Technical University of Applied Sciences

Suggested Citation: Sigl, Anna-Lena; Fleischmann, Carolin; Cardon, Peter W.; Aritz, Jolanta; Koße, Tamara (2023) : Speech-to-text technology for global virtual teams: A SWOT analysis., Rosenheim Papers in Applied Economics and Business Sciences, No. 9/2023, Technische Hochschule Rosenheim, Rosenheim, https://nbn-resolving.de/urn:nbn:de:bvb:861-opus4-23270

This Version is available at:
https://hdl.handle.net/10419/279577

Technische Hochschule Rosenheim

Speech-to-text technology for global virtual teams: A SWOT analysis.

Anna-Lena Sigl[1]
Carolin Fleischmann[1]
Peter Cardon[2]
Jolanta Aritz[2]
Tamara Koße[1]

[1]Rosenheim Technical University of Applied Sciences
[2]University of Southern California

Rosenheim Papers in Applied Economics and Business Sciences

No. 9/2023

Speech-to-text technology for global virtual teams: A SWOT analysis.

Abstract

Virtual team meetings are increasingly supported with advanced technology. This study investigates the extent to which AI-enabled speech technology can be useful for global virtual teams (GVT). A survey was conducted in GVTs in 2020 and 2021, when people's lives were primarily dominated by the pandemic. A transcription software was used to support the collaboration. A total of 530 survey responses were analyzed using a structured approach- qualitative content analysis. The data was structured using the SWOT framework that aimed at comprehensively answering the research question "To what extent is the use of AI-supported speech technology in GVTs useful?". AI-generated transcripts are helping to overcome language and time zone barriers in GVT. Yet, they also cause misunderstandings and impact openness of communication. Further results and implications for GVT are discussed.

Keywords:

- Speech-to-text technology
- global virtual teams
- NLP
- SWOT analysis

1. Introduction

Real-life meetings, where face-to-face exchanges and lively interactions take place, were drastically reduced, or even eliminated over the last few years. The COVID-19 pandemic forced employees to work remotely and switch to digital forms of collaboration (Durakovic et al, 2022).

This means that there is still the possibility of sharing know-how and experiences, which is what good teamwork thrives on. Digital meetings make it easier for people to get together, both in their private and working lives.

An increasing number of teams are culturally different and spread across geographic borders. These teams are mostly limited to digital collaboration.

To be able to support GVTs, there are a variety of technologies, including AI-powered technologies. This study takes a closer look at an AI-enabled speech technology that is applied in GVTs. In this regard, the research question "To what extent is the use of AI-enabled speech technology useful in GVTs?" is addressed. Data from 530 members of virtual teams is used for a SWOT analysis of speech-to-text technology, specifically automated transcription of meeting recordings. We were investigating whether the use of the speech technology tool is a facilitator for GVTs, how reliable the technology is in use, and which opportunities and threats team members anticipate with the future use of speech-to-text technology in GVT.

2. Theoretical foundations and state of research

In recent years, various forms of AI have become ubiquitous in society and business. One area of AI is natural language processing (NLP), including applications that realize speech recognition for dictation systems or voice-controlled transcription (Chowdhary, 2020).

The hybrid field of computational linguistics (CL) and its subdomain of NLP have seen rising interest. They combine the fields of AI and linguistics (Clark et al., 2010). Both fields have long investigated language processing in human brains and by computers. CL laid the groundwork NLP by translating rules of human language into code and developing algorithms to process language in its context (Tsujii, 2021).

In this context, CL helps in the exploitation of information sources, as well as in the overcoming of language barriers, and represents a facilitation in dealing with machines. For example, CL can be involved in language teaching or in translating from one language to another (Shaykhislamov & Makhmudov, 2020).

Particularly since public interest in Large Language Models (LLMs) has skyrocketed, natural language processing (NLP) has been established as one of the key areas of AI. Vast amounts of information in the form of news, books or reports recorded in natural language are continuously accessible digitally. Therefore, there is an increasing need to enable computers to retrieve and process recorded natural language (Chopra et al., 2013; Nadkarni et al., 2011). NLP is a field of research that represents and automatically analyzes human languages using theoretically motivated computer techniques (Chowdhary, 2020).

Before LLMs, NLP was mostly known for grammar checkers or a conceptual search (Joseph et al., 2016). NLP is integrated into web and mobile applications, enabling natural interaction between humans and computers.

Two fields of activities can be distinguished in this interaction: 'Natural Language Understanding' (NLU) and 'Natural Language Generation' (NLG) (Liu et al., 2017). Due to the ambiguity of language and the multiple perceptions of word or sentence meanings, understanding natural

language is a particularly difficult task for machines. NLU pursues the task of recognizing and inferring natural language when it is input. Whereas NLG focuses on computer systems producing human speech into an understandable text.

People are often unaware of the power behind understanding what is spoken or written, which seems so simple and effortless (Tsujii, 2021). In order to be able to communicate with each other, language is the primary as well as the most natural form that can be chosen for an exchange of information. Natural language has countless exceptions to rules, which means that language cannot be narrowed down to a few elementary rules, as is the case with formal languages such as programming language. Background information is necessary to infer the meaning, otherwise the context will be inferred incorrectly. Also, the ambiguity of words is not helpful for a correct language recognition. Speech recognition means that machines and programs are capable of recognizing words and phrases in spoken language and, furthermore, converting them into a format that is readable by machines (Trivedi et al., 2018).

A critical and important factor for speech recognition is the surrounding acoustics, since it makes an enormous difference whether the speech is recorded via a high-quality microphone or via a cell phone with a poor connection (Erzin, 2009).

Certain criteria can be distinguished that make the speech recognition process easier and harder. On the one hand, it is a matter of the speech style. If it is a completely free way of speaking, special systems that can, for example, break off sentences or make improvements must be used. In general, free speech is a complication in speech recognition because of its high complexity. Single words, on the other hand, can be easily recognized by the system. On the other hand, it is more difficult for machines and programs to perform speech recognition if a large vocabulary (1,000 words or more) is used. If a smaller vocabulary is used, the words are recognized more reliably by the machine.

The user group is another distinguishing criterion of speech recognition that includes the persons whose speech is to be recognized. The program can easily adapt to the way a person speaks. However, the combination of several speakers make speech recognition more difficult since several individual ways of speaking and conversational overlaps occur (Watanabe et al., 2020). Automatic speech recognition aims at analyzing and characterizing the speaker identity in order to extract information (Gaikwad et al., 2010).

Due to the rapid development of information technology and computer hardware and software, speech recognition technology has become a key technology. Transcription, the transfer of and audio or video recording to a written text, is used for documentation in business and research (Cardon et al., 2021).

However, a lot of information is still lost in the creation of a transcript, since the conversational situation cannot be reproduced completely. When using transcription systems that cannot record non-verbal communication, some behaviors are completely disregarded. Context and conversational characteristics such as non-verbal communication are lost (Fleischmann et al., 2021).

A transcript needs to reproduce the discussion contents as detailed as possible and exactly. However, this leads in turn to a poor readability of the transcript due to the inclusion of vast amounts of information. For this reason, transcription rules exist that enable data reduction through defined guidelines. For a uniform creation of a transcript, the transcription rules specify how certain linguistic phenomena are to be represented in writing. Transcripts can be differentiated with regard to the degree of their documentation accuracy into summarizing, scientific and journalistic transcripts.

Summary transcripts are reduced to the most important points of the conversation and reproduce them in a meaningful form without reproducing the exact wording. In contrast, the

scientific transcript offers a higher degree of detail. The entire content of the conversation is transcribed word for word, including sentence or word breaks. Journalistic transcripts, on the other hand, reproduce the conversation in a reader-friendly form by smoothing out the colloquial language and translating it grammatically into standard language.

Different transcription systems exist, each targeting different areas in social situations (Davidson, 2009). On the one hand, the content-semantic transcript smoothes the colloquial language and provides a transcript that resembles written communication. This makes the transcript easier to read and thus provides quick access to the content of the conversation. Furthermore, there is the semi-interpretative working transcription, which captures and transcribes non-verbal communication in addition to verbal communication. Non-verbal behaviors include, for example, eye contact, body movements, facial expression which will only be possible if video in addition to audio files are used by the transcription software. Verbal transcription, on the other hand, records para-verbal communication, such as laughter and pausing.

The semi-interpretative working transcription is a very rich way of transcribing suitable for catching (intercultural) nuances and transcribing conversations with more than two speakers since the speaker statements can be mapped synchronously as well as successively. Yet, most speech-to-text technology focusses solely on the spoken words.

For the successful collaboration of a team, in a virtual working environment, communication and interaction is very crucial (Cagiltay et al., 2015). The most important factor for virtual team collaboration is trust, which is built up in particular through communication between the team members (Duran & Popescu, 2014). Communication problems can arise due to a lack of feedback or due to the chosen communication channel. These difficulties in communication occur more often with people who have different cultures and speak different languages. Therefore, a speech-to-text technology may particularly aid these types of teams (Fleischmann et al., 2021). Effective use of communication technology that GVTs employ depends on the team task and the context (Tenzer & Pudelko, 2016).

3. Methodology

3.1. Survey design and sample characteristics

In this study, we collected data on the use of a speech-to-text tool in GVT, specifically a transcription tool. In order to answer the research question "To what extent is the use of AI-enabled speech technology useful in GVTs?", students participants of a 7-week global virtual team project were surveyed. The project runs every year at about 17 universities across the globe. The data for this research was collected in 2020 and 2021, with a total of 530 students participating. Specifically, 177 students were recorded as responding fully to the survey in 2020 and 353 students participated in the 2021 survey. The students were enrolled in the business program and had an average age in their early 20s. In 2020, the gender distribution of respondents was: 46.1% female, 43.7% male, and 0.3% other. In 2021, we recorded 40.7% female, 39.3% male participants, and 0.2% other.

Teams consisting of four to six students worked together on a real company project. The team members were distributed globally and worked together virtually. As a result, there was no face-to-face meeting at any time.

The GVT had a structured transcription activity where they were asked to use the AI-enabled speech-to text tool in one of their team meetings and to share and review the transcript with the team after the meeting. The very short survey was open for one week after the activity. The first question, "How accurate was the meeting transcript?" was aimed at respondents' satisfaction with the extent to which the transcript produced by the transcription software was

perceived as accurate. It was measured on a numerical rating scale from 0 (completely dissatisfied) to 5 (very satisfied). Next, participants replied to three open-ended questions. The questions were about the aspects under which the accuracy of the transcript was decreased and whether using this technology would be helpful, especially if used regularly. We also asked about any concerns that respondents might have. The exact questions were: "What aspects of the meeting transcript were not accurate?", "Do you think this type of meeting tool would be useful if you used it regularly? How might this meeting tool be useful?" and "Do you have any concerns about using this type of meeting tool?".

3.2. Qualitative content analysis

In this study, we collected data on the use of a speech-to-text tool in GVT, specifically a transcription tool. A qualitative content analysis was carried out to map and cluster data using the SWOT framework. Specifically, the process model of inductive category formation was used for the data analysis (Mayring, 2016). The formation of inductive categories is also referred to as open coding, which is a principle of grounded theory (Glaser & Strauss, 1967). In this research, selected literature is also consulted to enrich the results with existing research in the field.

The survey results from 2020 and 2021 were first imported into an Excel file. Each individual quote was given a unique designation (e.g., A1, A2, A3, etc.). Separate tabs were created for the survey results of each question and each year. The evaluation was carried out sequentially by first evaluating the year 2020 for each question, and then the year 2021. This two-step process was designed to be able to catch any systematic differences in responses between the two year. After no pattern emerged, the years were summarized.

The data evaluation of the open questions was carried out according to the process model of inductive category formation, which represents the methodological procedure (Mayring, 2016). The research question "To what extent is the use of AI-supported speech technology in GVTs useful?" was answered by analyzing the strengths, weaknesses, opportunities, and threats. The data material was coded for each table sheet using the Grounded Theory Methodology (GTM) developed by Glaser and Strauss (1967). Through GTM, object-related theories were formed based on the data, which were thus inductive. The data material was analyzed line by line and the 'open coding' according to GTM could be applied. When a suitable quote was discovered for the first time, a category was formed, whose name briefly and comprehensively describes the content. In the further course of the analysis suitable data material could be assigned to already existing categories. Likewise, new categories were inductively formed from the data material if the data material could not be assigned to the already existing classifications. The assignment of the individual data records to the category designation was done by color marking the cells. Each category was given its own color.

When the material throughput was about 40 %, a review of the already formed category system took place. This made it possible to avoid content overlaps and generalizing categories, which resulted in a more precise segmentation.

The data material of the formed categories was checked again for their category definition. The qualitative content analysis (Mayring, 2016) enabled a clear classification of the research results in the SWOT framework.

4. Research results

Generally, survey participants tended to be satisfied with the accuracy of the meeting transcript. 73% were at least somehow satisfied with transcript accuracy. Yet only 17% were very completely satisfied and 16% were even completely dissatisfied.

Open-ended comments offer more detailed insights into the strengths, weaknesses, opportunities, and threats that are perceived with the AI-enabled speech-to-text technology (refer to figure 1 for a summary).

4.1. Strengths of speech to text technology in GVTs

One major strength of speech-to-text technology for GVTs is the *variety of possible uses*. Participants found the use of transcription software helpful for important meetings ("If there is a very important meeting, where all small details mean a lot" (B59)), in short meetings ("It could be useful but mostly for short encounters" (B43)) as well as in long meetings ("It would be better if you leave a long conversation as text" (B77)). In addition, its use in conferences specifically was deemed viable by one of the respondents: "It could be useful in conferences that needed to be referenced." (B176). Also, the use of transcription software in interviews "This would be useful when interviewing people for a paper." (B42) and online meetings "It could be useful for online meetings where not everyone speaks the same language." (B62) is considered helpful. In addition, participants suggested that the technology "could be useful for professors recording lectures" (B112) and presentations: "would be best used in a set-stage presentation where the pacing is scripted" (B111).

The *user perception* when operating the tool represents the second category (Figure 1). In general, the transcript is perceived as easy to read and to be interpreted, as one team member stated: "each group members were able to read and understand the transcript" (A35). Good recording takes place through clear pronunciation of the speakers, which one of the interviewees expressed by saying "If one person spoke clearly for a few seconds, it would pick it up well." (A12).

In addition, the use of the speech technology tool can strengthen *team collaboration*, as one team member described, for example, "Allows us to communicate effectively and collaboratively. Each member can engage evenly while others listen and take their own perspectives as meetings continue." (B3).

Finally, the technology turns out to be helpful for overcoming *language barriers*: "It can be useful for students that have a language barrier." (B84).

Another strength is the generation of *minutes* that capture the talking points of the meeting. The minute taker uses the transcript to produce detailed minutes; "It allows for members to go back and check over their notes for the meeting minutes." (B153). The transcript is also used as an aid in maintaining a chronological order of the meeting: "it helps the group members follow the chronological order" (B126) and creates a meeting structure "it helps set up a good meeting structure [...] can help lead us to well-constructed conversations." (B82).

Similarly, the transcript serves team members as a supplement to their personal notes that one of the respondents stated as follows "it makes the task of taking notes much easier" (B217). As indicated from the survey results, the attention of the meeting participants can be focused on sharing information "this type of meeting tool [...] be [...] useful, since it means the group can forego a notetaker, thus allowing everyone to participate with their full attention" (B130).

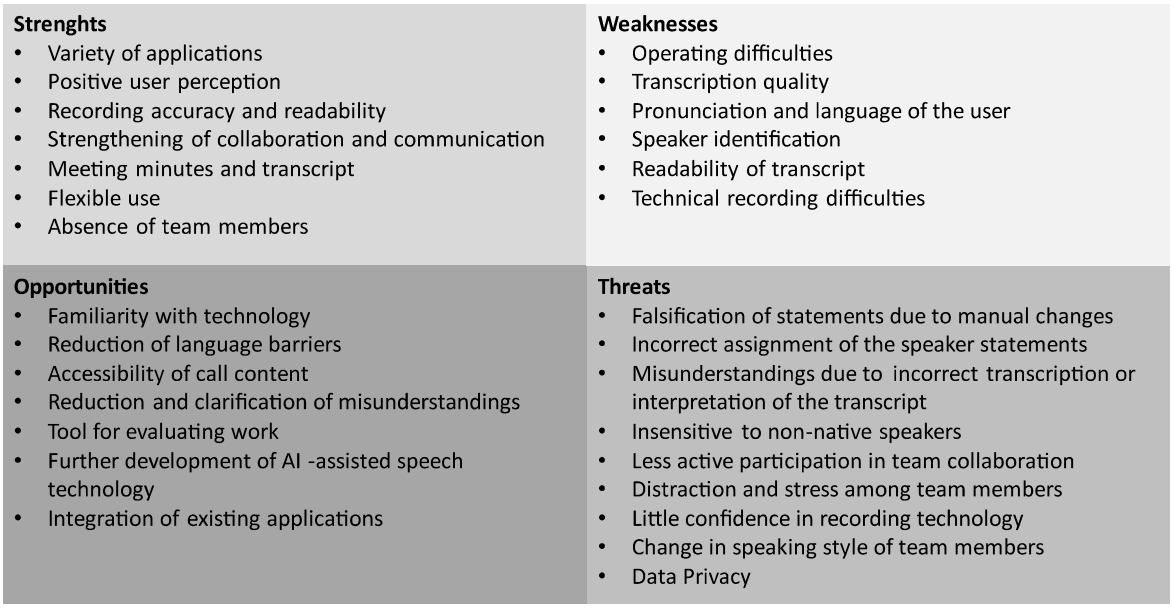

Figure 1. SWOT analysis of speech-to-text technology in GVT.

One participant stated that the transcript is useful for brainstorming, idea generation and strategy development. "The transcript could be used for forming ideas." (B108).

Also, it is worth mentioning that misunderstandings can be cleared up more easily by checking the content of the conversation using the transcript and communication becomes more effective "It may help avoid misunderstanding and increase communication efficiency." (B133). Using the technology and thus communicating *independent of place and time* is another advantage: "It is an easier way to communicate despite location and time barriers." (B11).

Finally, one frequently mentioned benefit of automated meeting transcripts was the *absence* of a meeting member in the meeting (see Figure 1): "This meeting tool can be very useful for someone who was not present, or to look back on after the meeting for specific words said or topics." (B304). It was described that absent team members can catch up on the discussed meeting contents by means of the transcript. In addition, the content of the meeting can be easily understood by reading the transcript: "It is [...] helpful for people who were absent or for reviewing past discussions" (B312), and can be used as a reference at any time: "[it's] useful for the persons who were absent during the meeting and for the persons who were there to remind themselves of what was said" (B179).

The described categories of strengths were emerging from systematic content analysis. These findings have been placed in a common category system (Figure 1).

4.2. Weaknesses of AI-enabled speech technology in GVTs

When using the transcription software in virtual teams whose members are globally distributed, not only strengths but also a number of weaknesses emerged. Figure 1 also provides a summary of identified weaknesses.

Difficulties in use and operation are mentioned as a major weakness. Some respondents perceived the setup of the transcription software on the computer as a complicated and time-consuming process. "Having it set up was definitely super difficult, my team wasted a good amount of meeting time just trying to figure it out." (C193). This is particularly relevant when using the tool for the first time: "I am not very tech savvy so I struggle making sure I am using it right or running the program right." (C210) and "we still don't really know how to use it well" (C205). In addition, navigating and using the transcription software is cumbersome: "This [...]

tool was difficult to navigate and use since we couldn't find a way for everyone to see the transcript live and share our screen at the same time." (C208).

Another weakness is the *quality* of the transcription. First and foremost, a lack of accuracy in the actual wording can be noted. For example, one team member reported that "the transcription is not accurate which makes it less useful for the group" (C264). The accuracy of the transcribed conversation content is impaired by omissions of individual words as well as parts of sentences. This was evident from the statements "It didn't catch every single word from the conversation." (A69) and "sentences were not transcribed fully or correctly" (A59). Likewise, words that were never used in the conversation are added in the transcribed conversational dialogue: "There was a [...] amount of words that were replaced with different words that were not said." (A188). In addition, words are often confused with similar sounding words :"Some linking words were wrong and other words were not understood and replaced by words with close pronunciation." (A112). Furthermore, the interviewees noted that the transcription software had difficulty in correctly capturing filler words, brand names, location names, and personal names "Some parts of the sentence were inaccurately transcribed including [...] names and brands" (A129), as well as culturally specific expressions "Some phrases that are culturally specific were not recorded accurately." (A132). Furthermore, grammar, punctuation, and spelling posed obstacles: "There were some grammar mistakes from time to time" (A197), "Punctuation and spelling were the only parts of the transcript that were not completely accurate, which made it hard to understand what was being said at some points" (A104).

Consequently, one of the students expressed concern that these inaccuracies in the transcription could lead to *misinterpretations*: "The inaccuracy of the text could lead to misinterpretations of what was actually said." (C332).

Another weakness of the transcription software is evident in the *pronunciation*, language, and para-verbal communication, i.e., style of speaking, which represents an additional category: "The wording can be off if the speaker has a thicker accent or speaks hastily." (A274). If someone speaks too fast this may lead to transcribing incorrectly. Particularly with accents, the transcript has trouble accurately capturing the wording "The accuracy of the transcript varies with the accent of the speaker." (A272). Inaccuracies also occur with slow "If the speaker is slowing his/her talking pace, the meeting transcript would appear incorrect words." (A231) or quiet "At times, when others spoke quietly it didn't get what they said either." (A237) pronunciation by team members. Ultimately, the technology also has difficulty "when words were mumbled" (A287) to transcribe correctly.

Another weakness is *speaker identification*. The AI-enabled speech technology has a hard time distinguishing the speakers in the team "Had a hard time distinguishing between speakers" (A299). It also cannot properly assign speakers when speaking at the same time "If several people spoke at the same time, [it] put this all to one 'Speaker' and wasn't able to differentiate the people." (A302). "When several people talked at the same time, the program got confused." (A363). This research took place in a virtual team environment. Speaker identification becomes even more difficult for in-person or hybrid team meetings.

Further, the transcript is perceived as confusing and *difficult to read*: "the transcript is not [...] user friendly and looks like lines of code. It would be [...] easier to read in a word document or another simpler format." (C194). In particular, long transcripts were perceived as difficult to keep track of and unstructured, containing many irrelevant talking points, which two of the participants described: "It captures a lot of [...] little conversations alongside the actual agenda, which may make the transcript too lengthy and irrelevant" (C371) and "the transcript was confusing & difficult to read concisely" (C385).

Finally, *technical difficulties* are a concern. Some of the speakers are only captured fragmentarily in the recording: "It didn't capture all of what we said. It only recorded two of the speakers conversations" (A375) or only one of the speakers is recorded during the recording "The

transcript only recorded one person's audio and not the rest of the members." (A391). In addition, when using the transcription software, recording glitches occur in that the record button does not always work reliably or the recording stops. Specifically, one of the interviewees described the situation as "we had to record the whole meeting in video and then upload it to the program as pressing the button to record it didn't work properly" (C228). A solid network connection and good audio quality also contribute to the quality of the transcription: "The transcript depends on the wifi strength so if someone had weaker wifi, their audio would cut out and therefore, the [...] transcription was less accurate" (A435) and "Also, if someone had poor audio quality, the software had a much more difficult time transcribing what they were saying." (A434). Ultimately, compatibility issues occur in that the transcription software does not work properly with some video conferencing tools. This could be inferred from the statements "there are some compatibility issues" (C225) and "did not work directly with Skype" (C217).

4.3. Opportunities from using AI-enabled speech technology in GVTs

The use of speech-to-text technology presents some opportunities to GVT, which are also summarized in Figure 1.

One major opportunity is the possibility to *overcome language barriers* through the use of transcription software in the team. Regular use of the tool is seen as helpful, as two team members explained: "I think this would be useful regularly. Even if it is not completely accurate, it had enough information for me to remember what we talked about in the conversation. Especially when people are hard to understand due to language barriers and accents, the transcription is nice because it helps break down that barrier a little bit." (B208) and "It could be useful for online meeting where not everyone speaks the same language". (B62). AI-enabled language technology can help make the content of the conversation accessible to those who are less fluent in the language or to those with hearing impairments: "This could be helpful for international students who aren't fluent in a specific language that the group is speaking and would benefit from a written transcript of the discussion" (B110). "This could be a [...] handy tool for anyone who works with the hearing impaired, or for people who want a more exact record of what was said at a sensitive meeting." (B38).

This statement indicates another opportunity, namely that by recording and transcribing the conversation with the help of the transcription software, emerging *misunderstandings* can be clarified. Thus, the transcript can act as evidence and the recording can eliminate misinterpretations: "I can see how using these types of tools can work well in today's global workspace. Communicating with colleagues and vendors face to face helps to limit the number of misconceptions that can come via written communication. Having that conversation recorded also helps minimize misunderstandings and errors." (B6).

Also, AI-enabled speech technology can be used as a tool to *evaluate the work* of individuals on the team. This can shed light on the engagement and performance of individual team members: "I can also see this tool being useful for interdepartmental purposes, assessing employee engagement/ participation, and measuring meeting productivity" (B5). Yet, the automated evaluation of work may also be seen as a threat rather than an opportunity as outlined in the next section.

Another opportunity is to use the produced transcript for *further AI-enabled analysis*. "It could be useful especially with the search & tag features, being able to search back into conversations to find important details you might forget. Also, in the future could become very useful if it was able to update your calendar with deadlines or commitments (if it heard 'Oh yes I will do the report by Monday' then it could add that to your schedule/reminders." (B388). This requires the integration with tools that are already used on a daily basis: "It needs to be integrated into one of the hundreds of apps that we already use like Skype or Webex." (C409).

4.4. Threats from using AI-enabled speech technology in GVTs

One mentioned risk is the *falsification* of transcripts by allowing manual changes in the written transcript. A participant worries "that my words in the transcription could be manually changed by someone else" (C191). In this context, a false attribution of the speaker's statements is possible, since the statements of the persons cannot be clearly separated and correctly attributed "(it) didn't say who said what so anything could be blamed on someone because things aren't accurate and it doesn't have names attached to comments." (C322). One strength of transcription software is that it can help clarify misunderstandings within the team. However, it is through an incorrect transcript that the risk of *misunderstanding* first arises: "It can lead to some misunderstanding if the transcription isn't accurate." (C319). Here, the correct interpretation, which is only done by reading the transcript, is very risky.

In addition, there is a risk that *non-native speakers* of the team language may feel disadvantaged in their language use, as the AI-supported language technology often recognizes people's accents incorrectly or not at all. Participants note that "it could [...] be offensive to some cultures if the wrong words were mistakenly added in the transcripts." (C242) and "that [...] it might cause certain biases because it fails to account for cultural differences and speech variations." (C235).

Another fear is the negative impact on *teamwork*. Active participation and engagement in group work can be reduced among team members. It is possible that "people [are] not fully engaging because they think they can access the information later" (C397) and "This tool may allow users to have a strong dependence and inertia. [...] The convenience of this tool will make some group members less active in participating in the meeting because, after the meeting, they can review the content of the meeting" (C400).

Furthermore, the technology poses a risk of *distraction:* "it serves as a distraction because none of us like to be recorded" (B396), and causes stress: "It is helpful if you want to go back and see main discussed points, but otherwise causes side stress" (B394). There is also some general wariness regarding the recording of meetings: "I am skeptical about any recording software." (C156). Thus, there is a risk that users will have little trust for AI-enabled speech technology.

At the same time, the speech style of the people may change from informal to formal. This situation was stated only by one of the respondents "if I used this with people knowing, it changes there tone and language from informal to formal" (B10).

In addition, some of the respondents worry about *data security*. They would be concerned "if it was used when people didn’t know and was used against them." (C308). This may be the case if promotion decisions were based on automated evaluation of team member performance. Another respondent states: "I have concerns that it pretty much makes security or privacy nonexistent when it comes to a meeting. Information that may want to be kept private could be accidentally shared with the wrong parties. Usually what is put in writing is permanent, this tool makes everything permanent." (C163). Thus, privacy is important so that data cannot be published or shared.

5. Discussion on the usefulness of AI-enabled speech-to-text technology

Considering the factors outlined in the previous chapter, the technology can create a transcript that is useful for GVTs in certain points.

Many challenges in GVTs occur due to language and cultural differences. By transcribing the content of the conversation, which is performed by speech recognition, language barriers in the team can be alleviated and the learning of new language can be facilitated. This is especially helpful in GVTs. The option of replaying and reconstructing conversation content that is not understood with the help of the transcript can support overcoming language barriers.

Likewise, the transcript can be a relief for the GVT in long meetings during non-working hours (due to time zone differences), because details can be forgotten. By transcribing, it can be guaranteed that no information of the conversation is lost. The transcript can be used by individual team members to refresh their memory. Further, the team can review and comprehend discussions of the conversation. By reflecting on the transcript, conflicts can be resolved on a constructive level. The generated transcript can additionally be used for brainstorming and enables the team to draw ideas and strategies from the conversation statements afterwards. The density of information and knowledge is given in GVTs by the many different points of view, which is very useful in strategy development. In addition, this can enable the team to make a well-informed decision.

Finally, flexibility in team collaboration can be observed through the use of the language technology tool. It is often the case that it is not possible for a team member to attend the meeting due to time differences and geographical dispersion. With the help of the transcript, collaboration can be more flexible and both the absent person in the team and the team members who attend the meeting benefit. The absent person can independently follow up on the course of the conversation by means of the transcript. At the same time, the time required by the team to bring the absent person up to date is minimized.

The use of NLP continues to advance due to the increasingly rapid access to knowledge and information as well as more sophisticated models and computing capacity. Therefore, AI-enabled speech technology can be applied in many different situations that can support GVTs in their collaboration, such as in important meetings or interviews. Among other things, the transcription software can relieve the team since all participating persons in the meeting can focus on the conversation without having to take notes themselves. A large majority of the respondents are satisfied with the existing transcript accuracy, and the software continues to improve.

However, there are also opposing voices, as their expectations regarding the accuracy of the transcript were not satisfied. Automatically generated transcripts still need to be checked for accuracy and completeness and certain risks persist.

The actual conversational situation can never be reproduced comprehensively, since most transcription software does not record non-verbal communication. This makes the development of the context more difficult. However, linguistic specifics can be represented in writing by defined transcription rules. This can help bring the mood of the conversation closer to the reader because it becomes clearer in which way something was said. However, the use of transcription rules can make reading the transcript even more complicated. In addition, the ambiguity of the words challenges the team to interpret the transcript correctly. Therefore, misunderstandings can often arise in the team.

Such misunderstandings are not only caused by an individual interpretation of the transcript but can also result from an incorrect transcript. These misunderstandings can be decreased with continuous feedback cycles in the team. Speech recognition faces the challenge of fully

capturing the open speech of individual team members who speak simultaneously during conversational discussions. Due to the global distribution of team members, there is a large user base and the vocabulary varies greatly, making speech recognition more difficult. Especially names of people and brands, and some accents and dialects tend to be poorly transcribed. For the most part, the GVTs are non-native speakers of the team language. These persons usually do not have accent-free pronunciation, which is still a hindrance during the use of transcription software.

Simultaneous speech, which often occurs in discourse, leads to an incomplete transcript. Speech recognition has difficulty identifying the voices of the speakers. Also, speaker statements are often attributed to the wrong person because speakers are not listed by name in the transcript. Team members can assign the transcribed statements to another person who did not make the statement, which can lead to additional potential for conflict. In addition, the transcript can be changed manually, which results in a manipulation of the speaker's statements, which can also lead to disputes and ambiguities in the team. It is important to ensure that the team is able to handle the data in the generated transcript securely. Data protection in particular is elementary for the team, so that no sensitive information can be released to the public.

In conclusion, GVTs faces the general challenge that team collaboration may be weakened by the use of transcription software. There can be a reduced engagement and less active participation in the discussion, because individual team members feel secure that they can extract the important information from the transcript afterwards. This in turn does not promote team cooperation and further conflicts can arise.

6. Conclusion

In summary, AI-enabled speech technology can be very useful for GVTs. The GVTs benefit from having a transcript of the conversation generated by machine. This helps the team to consolidate conversation content from which ideas and strategies can be further extracted and developed. Likewise, due to the cultural and geographical diversity in GVTs, communication is facilitated by the transcription of the conversation statements, and the digital networking enables a rapid exchange of information. In this respect, this technology is readily used by GVTs because it is helpful for collaboration and when communication problems arise.

However, in order to realize a successful use of the tool at all, certain factors have to be considered by the GVTs when using it. These relate to speech behavior and pronunciation as well as the ambient acoustics during recording. Among other things, the team must be trained to be able to safely use the AI-supported speech technology. In addition, GVTs should be educated in advance on what risks and challenges may arise when using the tool. After all, language is very complex, making it difficult to correctly interpret written statements. Misunderstandings can often occur, weakening team collaboration. Due to the fact that the transcript is sometimes unreliable and error-prone, it always requires reflection. Nevertheless, conflicts in the team can be resolved constructively through the differently expressed viewpoints that have been written down.

Ultimately, due to the low age average of the participants in the study and the controlled setting of the virtual team project, no generally valid conclusions can be drawn for society as a whole. However, as this is a qualitative research that explores the different attitudes of the respondents, this research still provides meaningful and valuable results.

7. References

Cagiltay, K., Bichelmeyer, B. A., & Akilli, G. K. (2015). Working with multicultural virtual teams: critical factors for facilitation, satisfaction and success. *Smart Learning Environments*, 2(1), 1–16. https://doi.org/10.1186/s40561-015-0018-7

Cardon, P., Ma, H., Fleischmann, C. (2021). Recorded Business Meetings and AI Algorithmic Tools. Negotiating Privacy Concerns, Psychological Safety, and Control. *International Journal of Business Communication,* https://doi.org/10.1177/23294884211037009

Clark, A. M., Fox, C., & Lappin, S. (2010). The Handbook of Computational Linguistics and Natural Language Processing. In Wiley-Blackwell eBooks. https://ndl.ethernet.edu.et/bit-stream/123456789/14228/1/28%20pdf.pdf

Chopra, A., Prashar, A., & Sain, C. (2013). Natural Language Processing. *International Journal of Technology Enhancements and Emerging Research*, 1(4), 131–134. https://doi.org/10.1109/inmic.2004.1492945

Chowdhary, K. R. (2020). *Fundamentals of Artificial Intelligence*. Springer eBooks. https://doi.org/10.1007/978-81-322-3972-7

Davidson, C. (2009). Transcription. Imperatives for Qualitative Research. *International Journal of Qualitative Methods, 8(2),* 35-52. https://doi.org/10.1177/160940690900800206.

Durakovic, I., Aznavoorian, L., & Candido, C. (2022). Togetherness and (work)Place: Insights from Workers and Managers during Australian COVID-Induced Lockdowns. *Sustainability*, 15(1), 94. https://doi.org/10.3390/su15010094

Duran, V., & Popescu, A. (2014). The Challenge of Multicultural Communication in Virtual Teams. *Procedia - Social and Behavioral Sciences*, 109, 365–369. https://doi.org/10.1016/j.sbspro.2013.12.473

Erzin, E. (2009). Improving Throat Microphone Speech Recognition by Joint Analysis of Throat and Acoustic Microphone Recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(7), 1316–1324. https://doi.org/10.1109/tasl.2009.2016733

Fleischmann, C., Aritz, J., & Cardon, P. (2021). Acceptance of Speech-to-Text Technology: Exploring Language Proficiency and Psychological Safety in Global Virtual Teams. *Proceedings of the 54th Hawaii International Conference on System Sciences*, 411-420. https://doi.org/10.24251/HICSS.2021.049

Gaikwad, S., Gawali, B. W., & Yannawar, P. L. (2010). A Review on Speech Recognition Technique. *International Journal of Computer Applications*, 10(3), 16–24. https://doi.org/10.5120/1462-1976

Glaser, B. & Strauss, A. (1967). The discovery of grounded theory: Strategies for qualitative research. Chicago, IL: Aldine.

Joseph, S. R., Hlomani, H., Letsholo, K., Kaniwa, F., & Sedimo, K. (2016). Natural Language Processing: A Review*. International Journal of Research in Engineering & Applied Sciences* 6(3), 207-210.

Liu, D., Li, Y., & Thomas, M. (2017). A Roadmap for Natural Language Processing Research in Information Systems. *Proceedings of the 50th Hawaii International Conference on System Sciences,* 1112-1121. https://doi.org/10.24251/hicss.2017.132

Mayring, P. (2016). *Einführung in die qualitative Sozialforschung*. (6th ed.). Beltz

Maznevski, M. L., & Chudoba, K. M. (2000). Bridging Space Over Time: Global Virtual Team Dynamics and Effectiveness. *Organization Science*, 11(5), 473–492. https://doi.org/10.1287/orsc.11.5.473.15200

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

Shaykhislamov, N. Z., & Makhmudov, K. S. (2020). Linguistics and its modern Types. *Academic Research in Educational Sciences*, 1(1), 358–361. https://doi.org/10.24411/2181-1385-2020-00049

Tenzer, H., & Pudelko, M. (2016). Media choice in multilingual virtual teams. *Journal of International Business Studies*, 47(4), 427-452. https://doi.org/10.1057/jibs.2016.13

Trivedi, A., Pant, N., Shah, P., Sonik, S., & Agrawal, S. (2018). Speech to text and text to speech recognition systems - A review. *IOSR Journal of Computer Engineering*, 20(2), 36–43. https://doi.org/ 10.9790/0661-2002013643

Tsujii, J. (2021). Natural Language Processing and Computational Linguistics. *Computational Linguistics*, 1–21. https://doi.org/10.1162/coli_a_00420

Watanabe, S., Mandel, M. I., Barker, J., Vincent, E., Arora, A., Chang, X., Khudanpur, S., Manohar, V., Povey, D., Raj, D., Snyder, D. S., Subramanian, A. S., Trmal, J., Yair, B. B., Boeddeker, C., Ni, Z., Fujita, Y., Horiguchi, S., Kanda, N., . . . Ryant, N. (2020). *CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings*. https://doi.org/10.21437/chime.2020-1