Zhang, X. (2022). Multimodal Human Interaction in Videoconference Interpreting. *Current Trends in Translation Teaching and Learning E*, *9*, 121 – 148. **https://doi.org/10.51287/cttl20224**

MULTIMODAL HUMAN INTERACTION IN VIDEOCONFERENCE INTERPRETING

Xiaojun Zhang

Xi'an Jiaotong-Liverpool University

Abstract

The evolution of communication technologies such as video conferencing and remote meeting has created ample opportunities for distance communication in real time and has led to alternative ways for delivering interpreting services. Videoconference interpreting, either spoken-language or sign-language interpreting, is best described as a 'multimodal' way to deliver interpreting remotely which has been used for simultaneous, consecutive and dialogue interpreting. This paper focuses on the technical issues of integrating multimodal information of videoconference into interpreting and multimodal human interaction in videoconference interpreting. A prototype computer-aided videoconference interpreting system, CACIS, is introduced as well.

Keywords: Videoconference interpreting (VCI), multimodal

human interaction, meeting content analysis, computer-aided interpreting

1. INTRODUCTION

ISO/FDIS 24019 Simultaneous interpreting delivery platforms — Requirements and recommendations says "The pandemic public health situation affecting the entire world completely changed the meeting and conference market. It led to an unprecedented surge in the demand for distance meeting and interpreting facilities, in which simultaneous interpreting delivery platforms, amongst other technologies, play an important part." (2021, iv) Like other business processes, conferences and meetings of all kinds are also going digital. Increasingly, people are using computer technology both offline and online to support their meeting objectives. Since the outbreak of the Covid-19 virus and its subsequent spread across the world in 2020, video conferencing platforms have enjoyed a windfall. The trend for video conferencing established during the lockdown continued even after this, with companies and organisations often having it as a preferred form of communication for both financial and convenience reasons. In this scenario, remote meeting systems appear to be the mainstream of distant communication, since they play a crucial role in the generation of ideas, documents, relationships, and actions within an organisation.

Nowadays, video conferencing has established itself as a tool for verbal and visual interaction in real time, also between two or more sites. Collaborative workspaces in corporate networks and on the Internet offer geographically distributed collaborators a virtual repository for documents related to a project or a conference. Electronic meeting support systems (i.e., interactive, network-connected whiteboards and videoconferencing appliances[1]) are available for the benefit of those who share the same room as well as those who are in remote locations.

Following these trends, conference interpreting seems to have fully embraced displaced, multimodal, and technology-mediated communication. According to Constable (2015), remote delivery of interpretations has already moved from the public services context to the conference sphere. The evolution of information and communication technologies (ICT) has created ample opportunities for distance communication in real time and has led to alternative ways of delivering interpreting services. Remote interpreting (RI), i.e., oral interpretation of a remote meeting, refers to the use of communication technologies to gain access to an interpreter in another room, building, town, city, or country (Braun, 2015: 352). In this paper we use

[1] Some examples of videoconferencing appliances are Cisco (Webex Room Kit, Room Kit Mini and Room Kit Plus), Poly (Studio X30 and X50), Yealink (VC200) and Neat (in partnership with Zoom), among others.

videoconference interpreting (VCI) as a cover term for remote interpreting via videoconference and interpreter-mediated video conferencing. VCI, either spoken-language or sign-language interpreting, is best described as a way to deliver interpreting services remotely in which the parties are connected through a video meeting. Both consecutive and simultaneous interpreting (including dialogue interpreting) can be implemented for VCI.

2. VIDEOCONFERENCE INTERPRETING (VCI)

VCI shares features of both face-to-face and mediated, screen-based communication. Having video-based access to an interpreter or to speakers and hearers influences the kind of input data available to interpreters (context, situational knowledge, multimodality, cognition, etc.) and its impact on the unfolding interaction (cf. Vranjes and Brône, 2020). In other words, the focus has shifted to the complex socio-institutional framework that encompasses interpreting events and the way this affects the interpreting process and the participants involved. While most studies have explored the implications and applications of video technology for interpreting, offering in-depth discussions about potential advantages and shortcomings (see, for instance, the overviews by Braun, 2020, and Pöchhaker, 2020), very few have actually analysed the multimodal nature of VCI or

its opportunities for research.

Our study delves into this 'dark side' of the interpreting 'moon'. We will adopt a theoretical strand along the postulates of multimodal (interaction) studies by Müller et al. (2013, 2014), Jewitt (2014) and Mondada (2016), among others. Multi-sensory integration enables effective meaning construction, which explains why interpreters usually find it easier when they can have a direct view of the speaker's face and the meeting room (cf. Moser-Mercer, 2005). Inspired by human bimodal perception (cf. Besle et al., 2004) in which both sight and sound are used to improve the comprehension of speech, our proposal underlines the interplay of verbal (spoken and written) and nonverbal data (e.g., gestures, detected emotions, etc.) for successful videoconference communication/interpreting. Our proposal also adheres to the subfield of meeting content analysis, as a convenient way to help interpreters prepare for a given meeting and provide a better user experience. Our main aim is to come up with a core list of key features and resources that may be used to inform the development of VCI technology and multilingual conference support applications in the future. Our suggestions could be also used to complement or revise current standards and guidelines, such as *AIIC Guidelines for Distance Interpreting* (2019, v1.0) or ISO/DIS 24019:2021 *Simultaneous interpreting delivery platforms — Requirements and recommendations*.

3. MULTIMODAL HUMAN INTERACTION

As stated above, communication technologies are playing an increasingly important role in interpreting practices. Processing of audio-visual data provides interpreters with basic, factual information on participants such as who is presenting (speaker, audience, etc.) and what words are being said. Combining different sources of information into a meeting record of who said what, when and to whom, is often also useful for interpreting studies. For instance, group-dynamics-based models can reveal how a group interacts, or they can be applied to abstracting and summarising the meeting overall content. Finding ways to integrate the varying analyses required for a particular meeting support application has been a major challenge. In addition, modelling and analysing multimodal human-to-human communication scenes to real-world applications has required careful design of both interface and system, so that they are user-centric and demand-captured at the same time. Furthermore, deciding how to evaluate such systems breaks new ground, whether it is done intrinsically or from an end user's point of view. In this section, we offer a brief overview of multimodal human interaction, including meeting content analysis and human-like models.

3.1. Meeting content analysis

In the 1990s, new information technology (IT) tools and research paradigms enhanced understanding of human communication, as larger and larger amounts of audio-visual recordings were available in digital formats. During this decade, separate advances in the audio and video analysis of recordings led to the first implemented systems for interaction capture, analysis, and retrieval. The early Filochat system (Whittaker et al., 1994) took advantage of handwritten notes to provide access to recordings of conversations, while BBN's Rough'n'Ready system (Kubala et al., 1999) enhanced audio recordings with structured information from speech transcription supplemented with speaker and topic identification. Video indexing of conferences was also considered in early work by Kazman et al. (1996). Multi-channel audio recording and transcription of business or research meetings was applied on a considerably larger scale in the Meeting Recorder project at ICSI, Berkeley (Morgan et al., 2003), which produced a landmark corpus that was reused in many subsequent projects.

In the early 2000s, it quickly became obvious that meeting analysis technologies needed to address a significant subtask of the modalities used for all human communication. This in turn required appropriate capture devices, which needed to be placed in instrumented meeting rooms, due to constraints on their position, size, and connection

to recording devices, as exemplified by the MIT Intelligent Room with its multiple sensors (Coen, 1999). For instance, Classroom 2000 (Abowd, 1999) was an instrumented classroom intended to capture and render all aspects of the teaching activities that constitute a lecture. The Microsoft Distributed Meetings system (Cutler et al., 2002) supported live broadcast of audio and video meeting data, along with recording and subsequent browsing. Experiments with lectures in this setting, e.g., for distance learning, highlighted the importance of video editing based on multimodal cues (Rui et al., 2003). Instrumented meeting or conference rooms were also developed by Ricoh Corporation, along with a browser for audio-visual recordings (Lee et al., 2002), and by Fuji Xerox at FXPAL, where the semi-automatic production of meeting minutes, including summaries, was investigated (Chiu et al., 2001). By then, the technology seemed mature enough, however, for corporate research centres to engage in the design of such rooms and accompanying software, with potential end-user applications seeming not far from reach.

Thanks to the success of deep neural networks in computer science in the 2010s, a new series of approaches capable of fusing information from different modalities in hidden space at the intermediate level has successfully entered multimodal analysis. Very soon it became clear that a finer-grained level of meeting content

analysis and abstraction, i.e., technology for remote audio-visual conferencing, was required to provide intelligent access to multimedia and multimodal human interaction. Nowadays, there are methodologies and techniques stemming from Artificial Intelligence (AI) and Natural Language Processing (NLP) suitable for handling the integration of multimodal data and domain knowledge. These methods can fully utilise the multimodal data through learning correlational representations (i.e., multimodal fusion of data across different modalities) and achieving multimodal data and knowledge fusion (multimodal fusion of data with domain knowledge). However, quite a few methods, including deep learning, can be used to learn hidden representations, while further correlational mining techniques are necessary for data-driven correlational representations (Zhu et al., 2020).

Question answering, video summarisation, visual pattern mining and recommendation are just some examples of applications that need diverse domain knowledge for multimodal fusion of data with knowledge. Some of them could potentially benefit VCI. For instance, video summarisation, a technology that creates a concise and complete synopsis by selecting the most informative parts of video content, could be integrated in current interpreting delivery platforms, so as to facilitate the work of interpreters and booth mates when resuming work after a break (as a sort of quick

update) or even as an aid to relay (as a kind of monolingual or multilingual summary).[2]

3.2. Multimodal human interaction analysis

The need for advanced multimodal signal processing for content abstraction and access has been addressed in the past decade, but large, collaborative enterprises were needed in order to address the full complexity of human interaction in meetings. In addition, training powerful machine learning algorithms has required large amounts of data and appropriate reference annotations in several modalities.

Quite often the public nature of most of the funding involved in such initiatives has ensured the public availability of the data. For instance, two projects at Carnegie Mellon University (CMU), the Informedia project (Wactlar et al., 1996) and its Interactive Systems Laboratory (ISL) project (Waibel et al., 2001), were among the first to receive public funding to study multimodal capture, indexing and retrieval, with a focus on meetings. This was directly concerned with recording and browsing meetings based on audio and video information, emphasising the role of speech transcription and summarisation for information access (Burger et al., 2002). In Europe, the FAME

[2] For an overview on video summarisation technologies, see Apostolidis et al. (2021) and Haopeng et al. (2022).

project developed a system prototype that used multimodal information streams from an instrumented room to facilitate cross-cultural human-human conversation (Rogina and Schaaf, 2002). A second prototype, the FAME Interactive Space (Metze et al., 2006), provided access to recordings of lectures via a tabletop interface that recognised voice commands from a user. The M4 European project introduced a framework for the integration of multimodal data streams and for the detection of group actions (McCowan et al., 2005), and proposed solutions for multimodal tracking of the focus of attention of meeting participants, multimodal summarisation, and multimodal information retrieval. The IM2 National Center of Competence in Research (Switzerland) is a large, long-term initiative in the field of interactive multimodal information management, which focused on multimodal meeting processing and access. The CHIL European project has explored the use of computers to enhance human communication in smart environments, especially within lectures and post-lecture discussions, following several innovations from the CMU/ISL and FAME projects mentioned above (Waibel and Stiefelhagen, 2009). The US CALO project developed, among other things, a meeting assistant focused on advanced analysis of spoken meeting recordings, along with related documents, including e-mails (Tür et al., 2010). Its major goal was to learn to detect high-level aspects of human interaction which could serve to create summaries based on action items.

A multimodal human-computer interaction system facilitates human-like interaction. The human-like interaction between the computer and the user supported by multiple-modality technology has been applied in educational technology for a long time, especially in the field of pedagogical agent and intelligent tutoring systems (Jia, 2015). Typical scenarios of multimodal human-computer interaction in translation and interpreting fields are audiovisual translation (where dubbers or subtitlers transcribe and simultaneously translate audio texts from a video clip into a different language) and respeaking (where interpreters produce live subtitles via speech recognition software).[3] In terms of multimodality, dubbing, subtitling and respeaking transcend the mere transposing of written words or speech in a screenplay, bringing into the equation other multimodal elements that make up a video product. The images, the accompanying music, the sound effects and all the multimodal components together contribute to screen-based communication.

[3] Respeaking is a fairly recent notion. A more accurate definition is provided by Romero-Fresco (2011: 1): "a technique in which a respeaker listens to the original sound of a (live) programme or event and respeaks it, including punctuation marks and some specific features for the deaf and hard of hearing audience, to a speech recognition software, which turns the recognised utterances into subtitles displayed on the screen with the shortest possible delay". For further information, the interested reader is referred to Romero-Fresco (2018), Moores (2020) and Sandrelli (2020).

Multimodal human-computer interaction could also be applied to dialogue systems, virtual assistants, and machine interpreting. This model of interaction could also be successfully applied to videoconference interpreting, as it involves (i) processing of different input modalities: speech recognition, lip-reading, eye-tracking, gesture recognition and hand-writing recognition (cf. Oviatt et al., 2017); as well as (ii) various levels of integrations: leveraging visual modality for speech recognition (Choe et al, 2019), integrating simultaneous lip movement sequences into speech recognition (Lin et al., 2021), isolating target speech from a multi-speaker mixture signal with voice and face references (Qu et al., 2020), and grounding speech recognition with visual objects and scene information (Gupta et al. 2017).

4. SOME SUGGESTIONS AND CONCLUDING REMARKS

In VCI, like any other traditional modes of interpretation, interpreters tend to write down important pieces of information. For instance, in consecutive interpreting via video, interpreters will take notes in any style of their choice, or even use clues from minutes of previous meetings. Whatever the form of written record, it will be subjective and incomplete. Even with the best note-taking skills/procedures, interpreting errors often occur, which can only be possibly resolved by going back to what actually happened. The technology now

exists to capture the entire meeting process, keeping the text and graphics generated during a meeting together with the audio and video signals. If only interpreters could use the multimedia recordings of meetings to find out or remember what they need to know for the post-process debriefing stage, to recall main agreements taken, the more important issues considered, etc., then using these recordings would become an attractive addition to their note-taking skills. This can only happen when it is possible to recognise, structure, index, and summarise meeting recordings automatically so that they can be searched efficiently. One of the long-term goals of VCI support technology is to make it possible to capture and analyse what the meeting interlocutors are doing together using portable equipment, and to put together a wide range of applications supporting the group, using different systems or a unique platform or web services for tasks like speech recognition, summarising, and analysing the group's interaction. This will also enable other interpreters to make use of archives of conferences (akin to a multimodal corpus).

As can be easily inferred from the previous sections, successful delivery of interpretations remotely (via video link, videoconferencing system or fully-fledged simultaneous interpreting system) needs to take multimodality into account. Some key elements have been identified so far: multimodal input for the construction of meaning,

multimodal meeting interaction and various technologies for multimodal data capture. In this final section we advance a draft prototype that includes core components mentioned above. Drawing on previous work (Zhang, 2015), we propose a prototype of a computer-assisted VCI system which integrates domain knowledge and multimodal data (video, image, audio, text), as well as AI and NLP data capture and processing tools (e.g., MT, TM, ASR, AI-informed recognition of faces, signs, emotions, etc.). A plot diagram of the core components of the multimodal VCI system is provided in Fig. 1.

In the scenario of VCI, it is necessary to develop new models and algorithms to enable computers to support interpreters' interactions and enable them to extract both explicit and implicit information present in conferences and meetings. Human communication is about information, but the verbal form (words) of what people say is never the whole story: nonverbal cues also play a key role. Indeed, when people talk to each other, facial expressions, posture, and voice quality, among others, convey invaluable information for meaning construction, including people's reactions and attitude to what is actually being 'said'. Multimodal input provides helpful social cues to interpreters.

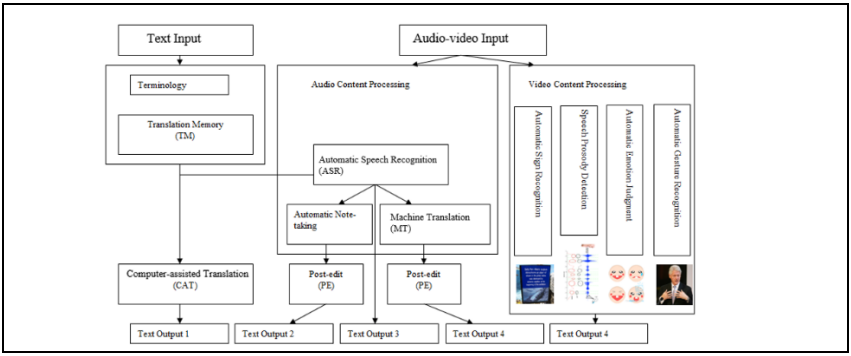

Figure 1. Computer-assisted VCI system architecture

The system proposed would be able to capture multimodal data generated by or supporting a multilingual meeting or conference (input related to the speech to be/being interpreted, including terminology and translation memory files) and to convert the data into structured and unstructured textual output, similarly to an augmented speech-to-text service. Users (interpreters) could input the terms they collected and the translation memory files they translated as the knowledge input. The audio-visual content of the living speech would also be input and processed in real time. In the audio content processing module, speech would be transcribed by an ASR tool and the transcription could be translated automatically by a machine translation (MT) system or by using a computer-aided translation (CAT) tool in different interpreting scenarios (see also Lewis and Niehues, in this volume). Key words and terms could also be

automatically extracted from the transcriptions by content analysis and keywords spotting technologies (akin to automatic note-taking). The automatic outputs of MT and note-taking could be post-edited by the other off-mike interpreter for the on-mike interpreter's information, for instance. In the video content processing module, the speaker's gestures and emotions could be recognised and grasped automatically, and his/her speech prosody and sign language on site could be transcribed as texts as well. All these core elements are also helpful clues for interpreters in the booth, remote or on-site.

Finally, VCI is directly linked to interpreting technologies in general, and more specifically to third-generation CAI (computer-assisted interpreting) tools (see Fantinuoli, this volume). In the future, we envisage further integration of CAI tools, NLP tools, RI platforms and multimodal core components of VCI systems.

Much research is still needed on natural (i.e., human-like) multimodal interaction and the automatic processing of communication scenes, in particular for mobile and ubiquitous settings, like remote interpreting via videoconferencing. Having to simultaneously manage multiple information channels, and diverse participatory structures will definitely (a) lead to novel forms of cross-modality interpreting, and (b) emphasise the organisational

and management issues of multimodal communication from interpreters' perspectives. In addition, user-centred assessment studies (performance evaluation, cognitive load, technology acceptance) are needed in order to measure the impact of computer-assisted VCI systems on professional and trainee interpreters. It will be also necessary to develop complex algorithms to analyse social dynamics and bimodal human communication automatically.

Understanding the complexity of technology in general (and computer-assisted VCI, in particular) requires careful study of how interpreters and other stakeholders adopt innovations and solutions, and how they interact and communicate with one another. By unpacking technology from the viewpoints of use and experience, computer-assisted VCI brings new perspectives to contemporary discussion about interpreters in change, their management of change, and their technology skills and needs for life-long learning.

The proposed architecture is an attempt to outline an 'ideal prototype' for a comprehensive support for interpreters delivering their service via videoconference and for interpreter-mediated video conferencing. When it comes to real implementation, we may of course choose to include only those high-performing and robust components which would be likely to offer reliable

and time-efficient assistance.

There is still a long way to go. But the silver lining is the growing interest in technology among developers, academics, and practitioners, which has already inspired novel tools and resources for interpreters. Computer-assisted VCI systems like the one suggested in this paper will be, no doubt, at the forefront of such developments.

ACKNOWLEDGEMENTS

This paper was carried out in the framework of the research projects PID2020-112818GB-I00 and the research network D5-2021_03, and the KSF project of XJTLU (Grant KSF-E-24).

REFERENCES

Abowd, G. D. (1999). Classroom 2000: An experiment with the instrumentation of a living educational environment. *IBM Systems Journal*, 38(4), 508-530.

AIIC. (2019). AIIC Guidelines for Distance Interpreting. https://aiic.org/document/4418/AIIC%20Guidelines%20for%20Distance%20Interpreting%20(Version%201.0)%20-%20ENG.pdf

Apostolidis, A., Adamantidou, E., Metsai, A. I., Mezaris, V., & Patras, I. (2021). Video Summarization Using Deep Neural Networks: A Survey. In *Proceedings of the IEEE*, 109(11), 1838-1863.

Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: Early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, 20(8), 2225-2234.

Braun, S. (2015). Remote Interpreting. In H. Mikkelson & R. Jourdenais (Eds.), *The Routledge Handbook of Interpreting* (pp. 352-367). New York: Routledge.

Braun, S. (2020). "You are just a disembodied voice really". Perceptions of video remote interpreting by legal interpreters and police officers. In Salaets, H. & Brône, G. (Eds.), *Linking up with video: Perspectives on interpreting practice and research* (pp. 203-233). Amsterdam: John Benjamins.

Burger, S., MacLaren, V., & Yu, H. (2002). The ISL meeting corpus: the impact of meeting type on speech style. In *Proceedings of the 7th International Conference on Spoken Language Processing* (ICSLP2002), 301-304.

Chiu, P., Boreczky, J., Girgensohn, A., & Kimber, D. (2001). LiteMinutes: an Internet-based system for multimedia meeting minutes. In *Proceedings of the 10th international conference on World Wide Web* (WWW2001), 140-149.

Choe, S. K., Lu, Q., Raunak, V., Xu, Y., & Metze, F. (2019). On Leveraging Visual Modality for Speech Recognition Error Correction. In *Proceedings of the Thirty-sixth International Conference on Machine Learning* (ICML 2019). https://srvk.github.io/how2-challenge/assets/authors/TH2_paper_7.pdf

Coen, M. H. (1999). The future of human-computer interaction, or how I learned to stop worrying and love my intelligent room. *IEEE Intelligent Systems*, 14(5), 8-10.

Constable, A. (2015). Distance Interpreting: A Nuremberg Moment for our Time. *AIIC 2015 Assembly Day 3: Debate on Remote.*

Cutler, R., Rui, Y., Gupta, A., Cadiz, J. J., Tashev, I., He, L. W., Colburn, A., Zhang, Z., Liu, Z., & Silverberg, S. (2002). Distributed meetings: A meeting capture and broadcasting system. In *Proceedings of the tenth ACM international conference on Multimedia*, 503-512.

Gupta, A., Miao, Y., Neves, L., & Metze, F. (2017). Visual features for context-aware speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP 2017), 5020–5024.

Li, H., Ke, Q., Gong, M., & Drummond, T. (2022). Video Summarization Based on Video-text Modelling. *arXiv:2201.02494*

ISO. (n.d.). Simultaneous interpreting delivery platforms — Requirements and recommendations. https://www.iso.org/standard/80761.html

Jia, J. (2015). Intelligent Tutoring Systems. In Spector, M (Ed.), *Encyclopedia of Educational Technology* (pp. 411-413). Thousand Oaks, CA, USA: Sage.

Jewitt, C. (2014). An Introduction to Multimodality. In C. Jewitt (Ed.), *The Routledge Handbook of Multimodal Analysis* (pp. 15-30). London: Routledge.

Kazman, R., Al-Halimi, R., Hunt, W., & Mantei, M. (1996). Four paradigms for indexing video conferences. *IEEE multimedia*, 3(1), 63-73.

Kubala, F., Colbath, S., Liu, D., & Makhoul, J. (1999). Rough'n'Ready: a meeting recorder and browser. *ACM Computing Surveys* (CSUR), 31(2), 7.

Lee, D. S., Erol, B., Graham, J., Hull, J. J., & Murata, N. (2002). Portable meeting recorder. In *Proceedings of the 10th ACM International Conference on Multimedia*, 493-502.

Lin, Z., Zhao, Z., Li, H., Liu, J., Zhang, M., Zeng, X., & He, X. (2021). SimulLR: Simultaneous Lip-Reading Transducer with Attention-Guided Adaptive Memory.

arXiv:2108.13630

McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., & Zhang, D. (2005). Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3), 305-317.

Metze, F., Gieselman, P., Holzapfel, H., Kluge, T., Rogina, I., Waibel, A., & Wölfel, M. (2006). The 'fame' interactive space. In *Proceedings of Machine Learning for Multimodal Interaction* (MLMI2006), 285-296.

Mondada, L. (2016). Challenges of multimodality: Language and the body in social interaction. *Journal of Sociolinguistics*, 20(3), 336-366.

Morgan, N., Baron, D., Bhagat, S., Carvey, H., Dhillon, R., Edwards, J., Gelbart, D., Janin, A., Krupski, A., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., & Wooters, C. (2003). Meetings about meetings: research at ICSI on speech in multiparty conversations. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP2003), 740-743.

Moser-Mercer, B. (2005). Remote Interpreting: Issues of Multi-Sensory Integration in a Multilingual Task. *Meta*, 50(2), 727-738.

Müller, C., Cienki, A., Fricke, E., Ladewig, S.H., McNeill, D., & Tessendorf, S. (2013). *Body – Language – Communication: An International Handbook on Multimodality in Human Interaction.* Vol. 1. Berlin and Boston: De Gruyter Mouton.

Müller, C., Cienki, A., Fricke, E., Ladewig, S.H., McNeill, D., Tessendorf, S. (2014). *Body – Language – Communication: An International Handbook on Multimodality in Human Interaction*. Vol. 2. Berlin and Boston: De Gruyter Mouton.

Oviatt, S., Schuller, B., Cohen, P. R., Sonntag, D., Potamianos, G., & Antonio krüger, A. (Eds.). (2018). *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition* - Volume 2 October 2018. Association for Computing Machinery and Morgan & Claypool.

Pöchhacker, F. (2020). 'Going Video': Mediality and Multimodality in Interpreting Studies. In Salaets, H., & Brône, G. (Eds.), *Linking up with video: Perspectives on interpreting practice and research* (pp. 13-45). Amsterdam: John Benjamins.

Qu, L., Weber, C., & Wermter, S. (2020). Multimodal Target Speech Separation with Voice and Face References. *Interspeech*.

Rogina, I., & Schaaf, T. (2002). Lecture and presentation tracking in an intelligent meeting room. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, 47-52.

Rui, Y., Gupta, A., & Grudin, J. (2003). Videography for telepresentations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 457-464.

Tür, G., Stolcke, A., Voss, L., Peters, S., Hakkani-Tür, D., Dowding, J., Favre, B., Fernandez, R., Frampton, M., Frandsen, M., Frederickson, C., Graciarena, M., Kintzing, D., Leveque, K., Mason, S., Niekrasz, J., Purver, M., Riedhammer, K., Shriberg, E., Tien, J., … Yang, F. (2010). The CALO

meeting assistant system. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6), 1601-1611.

Vranjes, J., & Brône, G (2020). Eye-tracking in interpreter-mediated talk: From research to practice. In H. Salaets & Brône, G. (Eds.), *Linking up with video: Perspectives on interpreting practice and research* (pp. 203-233). Amsterdam: John Benjamins.

Wactlar, H. D., Kanade, T., Smith, M. A., & Stevens, S. M. (1996). Intelligent access to digital video: Informedia project. *Computer*, 29(5), 46-52.

Waibel, A., Bett, M., Metze, F., Ries, K., Schaaf, T., Schultz, T., Soltau, H., Yu, H., & Zechner, K. (2001). Advances in automatic meeting record creation and access. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP2001), 597-600.

Waibel, A., Steusloff, H., Stiefelhagen, R., & Watson, K. (2009). *Computers in the human interaction loop*. London: Springer.

Whittaker, S., Hyland, P., & Wiley, M. (1994). Filochat: Handwritten notes provide access to recorded conversations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 271-277.

Zhang, X. (2015). The Changing Face of Conference Interpreting. In *New Horizon of Translation and Interpreting Studies* (pp. 255-263). Editions Tradulex.

Zhu, W., Wang, X., & Li, H. (2020). Multi-modal Deep Analysis for Multimedia. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10),3740-3764.