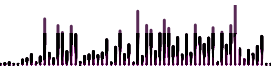
Live-Transkription für Zoom-Meetings - Ein Praxisbericht der Universität Bern

Jessica Hohermuth, Sibylle Reichel (Universität Bern)

DOI: https://doi.org/10.21248/gups.69152



aus dem Sammelband

Digitale Barrierefreiheit in der Bildung weiter denken
Innovative Impulse aus Praxis, Technik und Didaktik

Herausgeber*innen

Dr. Sarah Voß-Nakkour, Linda Rustemeier, Prof. Dr. Monika M. Möhring,
Andreas Deitmer, Sanja Grimminger

Verlag

Universitätsbibliothek Johann Christian Senckenberg

1. Auflage 2023

DOI: https://doi.org/10.21248/gups.62773
ISBN 978-3-88131-102-1

Live-Transkription für Zoom-Meetings - Ein Praxisbericht der Universität Bern

Jessica Hohermuth, Sibylle Reichel (Universität Bern)

Abstract:
Personen mit Hörbeeinträchtigung haben bei Online-Meetings oft Schwierigkeiten, wenn durch eine schlechte Videoübertragung oder bei Präsentationen mit verkleinertem oder gar keinem Sprecher*innenvideo das Lippenlesen erschwert oder unmöglich ist. Eine Live-Transkription des gesprochenen Vortrags kann hilfreich sein. Eine solche ist in Zoom und Teams zwar integriert, jedoch mit verschiedenen Einschränkungen. Um ein deutsches Transkript zu erhalten, wurde daher in Zusammenarbeit dreier Stellen (SWITCH, aiconix, Unibe) eine Lösung entwickelt, die den Ton eines Zoom-Meetings via Stream auf eine Website umleitet, auf der der Input transkribiert und das Ergebnis dann via API wieder in das Zoom-Meeting eingespeist wird. Seit November 2021 ist diese Lösung an der Universität Bern im Einsatz. Dabei sind verschiedene Schwierigkeiten zu bewältigen, wie etwa der Umgang mit konkurrierenden Zielen, Qualität und Geschwindigkeit des Transkripts sowie die Berücksichtigung passender Lexika, um eine gute Transkription auch wenig frequenter Fachbegriffe zu ermöglichen sowie die Erkennung schweizerdeutsch gesprochener Texte. Im Beitrag wird aus Anwender*innen- und Entwickler*innensicht die technische Lösung vorgestellt, technische und systematische Herausforderungen werden beleuchtet, Erfahrungen mit der Anwendung beschrieben und ein Ausblick gegeben, wie die Möglichkeit der Speech-to-Text-Konvertierung in verschiedenen Szenarien noch eingesetzt werden kann (Live-Transkription vs. nachträgliche Transkription, Verbesserung der Transkription durch Erweiterung der Lexika, Integration weiterer Sprachen).

Schlüsselbegriffe: Online-Meeting, Live-Transkription, Zoom, Speech-to-Text, Hörbeeinträchtigungen, Lippenlesen

1. Einleitung

Videokonferenzsysteme ermöglichen eine örtlich unabhängige Kommunikation und ersetzen damit die Face-to-Face-Kommunikation (Blokland et al., 1998: 97) vor Ort. Dass die Online-Kommunikation jedoch sowohl andere Voraussetzungen für eine gelingende Informationsvermittlung erfordert (vgl. Körschen et al., 2002: 19 oder Friebel et al., 2003: 10) als auch neue Herausforderungen – insbesondere für Personen mit Beeinträchtigung – mit sich bringt, sind Aspekte, die bei der Umstellung von „on site" zu „online" oft nicht bedacht werden. Aber auch wenn diese Aspekte berücksichtigt werden, ist die Gestaltung eines barrierearmen Meetings nicht nur technisch komplex, sondern muss auch unterschiedlichen Bedürfnissen und Ausprägungen eines Menschen mit körperlicher oder geistiger Behinderung gerecht werden. In diesem Beitrag wird die Live-Transkription als eine Lösung von unterschiedlichen „medialen Strategien" (Rink, 2020: 54) beleuchtet, die die Teilnahme an Zoom-Meetings für gehörlose und schwerhörige Personen erleichtern soll. Seit Herbst 2021 bietet die Universität Bern eine Live-Transkription in deutscher Sprache an, die in Zoom integriert werden kann. Diese Lösung wurde in Zusammenarbeit mit SWITCH und aiconix entwickelt und implementiert. Welche systematischen und technischen Dilemmata sich in der Entwicklung und im Einsatz der Lösung ergaben, wird im Folgenden kritisch dargelegt.

2. Live-Transkription in Online-Meetings

Obwohl die Universität Bern eine Präsenzuniversität ist, sind Veranstaltungen vor Ort seit 2020 – je nach epidemiologischer Lage und den jeweiligen Anordnungen des Bundes – nur eingeschränkt oder mit entsprechenden Maßnahmen möglich (bspw. Maskenpflicht oder hybride Veranstaltungen mit nur der Hälfte der Studierenden in Präsenz). Diese Situation stellte neue Anforderungen an Lehrveranstaltungssettings und führte zur vermehrten Nutzung der Videokonferenzsysteme „Teams" und „Zoom". Seit Frühjahr 2020 sind beide Systeme an der Universität Bern für alle Angehörigen verfügbar, wobei insbesondere Zoom für Lehrzwecke rege genutzt wird. Im Jahr 2021 waren es insgesamt 229.263 Meetings und 244 Webinare, die von 8.498 aktiven User*innen eingerichtet wurden und an denen unzählige Zuhörer*innen teilgenommen haben. Der eigens eingerichtete Zoom-Support kümmert sich dabei um Anliegen der Zoom-Nutzenden. Da Zoom selbst derzeit nur suboptimale Features zur Transkription bietet, wurde dank einiger Anfragen die Dringlichkeit

nach einer besseren Lösung erkannt und umgesetzt. Die Umsetzung gestaltete sich aber schwieriger als erwartet und benötigte den Einbezug unterschiedlicher Überlegungen – von technischen Anforderungen über die zoomeigenen Voraussetzungen bis hin zu einem Verständnis für Barrieren, mit denen Personen mit Hörbeeinträchtigung in Online-Meetings konfrontiert sind.

2.1 Barrieren für gehörlose und schwerhörige Personen

Wenn Personen mit Gehörlosigkeit oder Schwerhörigkeit an Online-Meetings teilnehmen, ist das Lippenlesen oftmals die einzige Möglichkeit, der Kommunikation zu folgen. Um die Mundbewegungen überhaupt erkennen zu können, muss mit Videoansicht kommuniziert werden. Daraus ergeben sich erhebliche Schwierigkeiten. Zuerst einmal muss bei allen am Gespräch beteiligten Personen eine Kamera installiert und – was oft nicht der Fall ist – im Meeting eingeschaltet sein. Wird Ton und Bild aufgrund von Verbindungsproblemen oder schwacher Internetleistung nicht synchron übertragen, dann ruckelt oder stockt das Videobild, was die Erkennung der Lippen erschwert und die Mitverfolgung des Gesagten verunmöglicht. Dies ist auch der Fall, wenn technische Störungen zu einer schlechten Bildqualität (z.B. verpixeltes Bild) oder zu einem kompletten Bildausfall führen. Nicht zuletzt kommt es auch vor, dass die Lippen durch Hygienemasken oder Vollbärte verdeckt sind, die das Lippenlesen verhindern.

In einem Online-Meeting bringen Teilnehmende aus unterschiedlichen Umgebungen verschiedene Voraussetzungen in die Kommunikationssituation mit. Eine Kamera, die nicht frontal, sondern seitlich zum Gesicht ausgerichtet ist, und wenig Licht im Raum, sind schlechte Voraussetzungen, die zu einer großen Herausforderung werden. Weiter sind auch laute Hintergrundgeräusche für Personen mit Hörschwäche, die einzelne Laute noch wahrnehmen können, irritierend. Selbst wenn die Lippen sichtbar sind, das Videobild in guter Qualität vorhanden, die Übertragung ruckelfrei, die Umgebungsvoraussetzungen geeignet und ein Headset mit gutem Mikrofon genutzt wird, kann auch das Ansichtssetting im Online-Meeting selbst Schwierigkeiten bereiten. Obwohl Zoom für die Ansicht der Videobilder einige Einstellungen anbietet (Video pinnen, Galerie- und Sprecheransicht und Aufmerksamkeitsmodus) und auch einzelne Videobilder vergrößert werden können, sind diese Funktionen nur bedingt geeignet. Mit dem Aufmerksamkeitsmodus übersteuert der/die Besitzer*in des Meetings die Videoansicht der Teilnehmenden,

was im schlimmsten Fall betroffene Personen ausschließen kann, wenn sie die Videobilder anderer Teilnehmenden nicht mehr sehen können. Die manuelle Skalierung einzelner Videobilder wiederum ist besonders mühsam, wenn mehrere Personen am Gespräch beteiligt sind.

Auch wenn eine Übersetzung in die Gebärdensprache mittels Videokommunikation möglich ist, ergeben sich durch die Videoübertragung dieselben Schwierigkeiten wie beschrieben. Die integrierte Zoomfunktion „Verdolmetschung" ist ebenfalls nicht geeignet, da lediglich der auditive Kanal zur Verfügung steht – nicht aber das Bild der dolmetschenden Person.

2.2 Zoomeigene Möglichkeiten zur Live-Transkription

Um die Schwierigkeiten, die sich beim Lippenlesen mittels Videobild ergeben, zu umgehen, sind Untertitel mit synchroner Übersetzung oder sogenannte Live-Transkriptionen (engl. „Closed Captioning") geeigneter. Zoom bietet die Funktion „Closed Captioning" mit einer manuellen (eine Person im Meeting tippt das Gesprochene über ein separates Fenster ein) und einer automatischen Variante an. Bei der automatischen Variante „Auto-Transkription" handelt es sich um eine nahezu simultane, automatische Übertragung von Speech-to-Text in derselben Sprache. Non- oder paraverbale Signale, wie diese oft bei SDH im Fernsehen angezeigt werden (Mälzer et al., 2020: 327), werden nicht übersetzt. Bislang bietet Zoom eine Auto-Transkription nur in englischer Sprache an (Stand: Januar 2022), die für die vielen Lehrveranstaltungen, die online auf Deutsch (manchmal sogar auf Schweizerdeutsch) durchgeführt werden, nicht geeignet ist. Auch haben wir uns gegen die Einbindung zahlreicher Apps, die Zoom zur Installation über den Marketplace anbietet, entschieden. Diese bieten ebenfalls nur eine englische Transkription an (Stand: Januar 2022), bräuchten zusätzlichen Support unsererseits und sind auch hinsichtlich ihrer Qualität fragwürdig. Nützlich ist hingegen die zoomeigene Einstellung zur Schriftgröße der Untertitel, die individuell reguliert werden kann und damit die Sichtbarkeit des Untertitels unterstützt.

Die Abklärung von zoomeigenen Möglichkeiten ergaben wichtige Anforderung an die neue Lösung: eine möglichst einfache Handhabung, Personalunabhängigkeit (Dolmetscher*in, App-Support) und Anpassungen sollten möglich bleiben, um individuelle Bedürfnisse berücksichtigen zu können.

3. Technische Umsetzung und Herausforderungen

Restriktionen von Software und Schnittstellen stellen generell ein Hindernis dar. Bei großen Playern wie etwa Zoom, Apple oder Microsoft ist die Entwicklung eines Angebots in der Regel nicht beeinflussbar und für Anwender*innen können unangekündigte Systemänderungen dazu führen, dass Schnittstellen und Funktionen nicht mehr oder plötzlich in anderer Form verfügbar sind (anders als beim Einsatz von Open-Source-Software; vgl. Ganten, 2021). Rechtliche Aspekte spielen ebenfalls eine wichtige Rolle: Nutzt man den Transkriptionsdienst von Zoom (oder auch Teams) selbst, werden die Daten nicht kennwortgesichert und der Serverstandort ist nicht eindeutig definierbar. Zoom sagt zwar zu, dass die Daten in den Rechenzentren des Landes eines Meeting Hosts verarbeitet werden, eine eindeutige Zuordnung ist aber nicht möglich. Die integrierte, einfach zu bedienende Lösung steht also in Konkurrenz zum Datenschutz.

SWITCH ist als Reseller für Hochschullösungen von Zoom in der Schweiz unser Ansprechpartner für die Anforderung an unsere Lösung. Schließlich kam es durch die entsprechenden Kontakte zu einer Zusammenarbeit der Universität Bern, SWITCH und aiconix. Als Anbieter KI-basierter, automatisierter Speech-to-Text-Lösungen hat aiconix unter anderem das Ziel, individuelle, in Zoom integrierbare Lösungen für die automatische Live-Transkription in deutscher Sprache zu entwickeln.

Schwierigkeiten, die sich zu Beginn stellten, waren, dass die Schnittstelle von Zoom zunächst nicht frei zugänglich war und daher zum Beispiel keine einfache Möglichkeit bestand, via API-Call den Ton abzugreifen und Text auf gleichem Weg als Untertitel einzubinden. Je nachdem, welcher Zoom-Client (Windows, MAC, iOS) genutzt wird, gibt es Unterschiede in der Aktualisierung der Untertitelanzeige, was zum Teil auch zu Problemen führen kann. Obwohl seit Frühjahr 2021 die API von Zoom offen ist, können die verschiedenen Clients (d.h. Apps) nicht über diese Schnittstelle beeinflusst werden.

3.1 Beschreibung der Lösung

Der gewählte Weg zweigt den Ton aus einem Zoom-Meeting via Live-Stream ab, ein Algorithmus bei aiconix transkribiert die Audioinformationen extern und führt das Ergebnis als plain text unter Verwendung der Zoom-Funktion „CC-Dienst eines

Drittanbieters“ via API-Token wieder in das Meeting zurück. Diese Lösung impliziert die Verwendung externer Rechenkapazität, wobei die Daten DSGVO-konform verarbeitet werden und mit einem persönlich gesetzten Passwort gesichert sind.

3.2 Dilemmata (technisch und systematisch)

Technisch konnte die Lösung zwar umgesetzt werden. Die Schwierigkeit lag jedoch darin, die einzelnen Parameter (Qualität, Geschwindigkeit, Einrichtung und Anzeige) gegeneinander abzuwägen.

- Qualität vs. Geschwindigkeit: Je schneller das Transkript angezeigt wird, d.h. je kürzer die externe Verarbeitung ist, desto schlechter wird die Qualität der Transkription. Bei einer Verzögerung von ca. 10 Sekunden kann eine sehr hohe Transkriptionsqualität erreicht werden. Jedoch ist die Anzeige von Untertiteln mit einer derartigen Verzögerung nicht mehr hilfreich für eine direkte Kommunikation. Nach mehrfachem Testen und der Rückmeldung von Personen mit Hörbeeinträchtigung wurde die maximale Zeit, nach der ein Transkriptionsergebnis angezeigt wird, auf 4 Sekunden gesetzt. Das heißt, dass bei fortlaufendem Sprechen spätestens nach 4 Sekunden die entsprechenden Wörter bzw. (Teil-)Sätze angezeigt werden, der Algorithmus also nicht unbedingt auf eine Sprechpause oder die Vervollständigung eines Satzes wartet.

- Einfaches Handling vs. Integration im Meetingfenster: Je integrierter die Anzeige der Transkription sein soll (Untertitel im Meetingfenster), desto umständlicher ist die Einrichtung vor dem Meeting. Der Stream muss konfiguriert und gestartet werden. Das bei jedem Meeting individuell erzeugte Token muss manuell in Zoom eingefügt und die Untertitel müssen aktiviert werden. Wir haben uns trotzdem für diese etwas aufwändigere Einrichtung entschieden; Der Vorteil ist, dass die Untertitel direkt im Meeting angezeigt werden. In einem vorgängigen Prototyp erschienen die Untertitel in einem separaten Fenster, was aber betroffenen Personen Schwierigkeiten bereitete. Es war auf diese Weise schwieriger, gleichzeitig Mimik und Gestik zu erfassen.

- Je grösser die Untertitel (Schriftgröße, Zeilenanzahl), desto mehr werden die übertragenen Inhalte aus dem Meeting davon überlagert. Besonders bei kleinen Bildschirmen kann das dazu führen, dass beispielsweise geteilte Folien nicht

mehr sichtbar sind. In der ersten Version wurden die Untertitel so an Zoom gesendet, dass vollständige Sätze erschienen sind. Das hatte zur Folge, dass bei langen Passagen nur die letzten drei Zeilen des Untertitels sichtbar waren. Insbesondere bei einer groß eingestellten Schriftgröße, die die Nutzenden selbst in Zoom anpassen können, war nur ein Teil des Transkripts auf dem Bildschirm sichtbar. Dieses Verhalten der Software wurde dahingehend angepasst, dass der Zeilenvorschub immer nur eine Zeile beträgt. Auch wenn viel Text auf einmal transkribiert wird, kann so das vollständige Transkript verfolgt werden.

4. Fazit - Perspektive

Die entwickelte Lösung für deutsche Untertitel in Zoom-Meetings muss noch feinjustiert werden, ist aber bereits im Einsatz und für Personen mit Hörbeeinträchtigung hilfreich. Wichtig war uns, dass die Lösung anpassbar und nachhaltig ist. Neben der Optimierung der vorgestellten Lösung gilt es auch, die rasanten Entwicklungen anderer Anbieter im Auge zu behalten. So verfügt Teams seit Oktober 2021 über eine deutsche Live-Transkription und die Möglichkeit automatische Transkriptionen in zahlreichen Sprachen zu nutzen. Auch Zoom hat angekündigt, bald mehr Sprachauswahl bei der Live-Transkription anzubieten (vgl. Zoom-Blog: 2021).

Nach heutigem Stand (Januar 2022) überwiegen die Vorteile der eigenen Lösung: Das Transkript ist speicherbar und für spätere Weiterverarbeitung verwendbar. aiconix beabsichtigt, neben den zahlreichen bereits vorhandenen Sprachen auch deutsche Dialekte (u.a. Schweizerdeutsch) zu implementieren. Des Weiteren ist vorgesehen, dass aufge zeichnete Meetings nachträglich mit eingebrannten Untertiteln versehen werden können. Diese erreichen durch verlängerte Rechenzeit, durch die Möglichkeit, Fachwortschatz über ein Lexikon einzubinden und nicht zuletzt durch eine manuelle Nachbearbeitung eine sehr gute Qualität. Sollte Zoom (wie auch schon Teams) bald eine Untertitelung von Meetings erlauben, wird der Schwerpunkt der aiconix-Lösung auf der Steigerung der Transkriptionsqualität und nicht mehr besonders der Transkriptionsgeschwindigkeit liegen, weitere Sprachen und Dialekte sollen berücksichtigt werden und der Datenschutz wird ausgebaut. Aus Sicht der Anwender*innen wollen wir die Stärken aller Möglichkeiten nutzen und den Nutzer*innen je nach Anwendungsszenario die beste Lösung näherbringen (Zoom, Teams, aiconix).

Bei allem Einsatz sollte aber ein Punkt nicht vernachlässigt werden: Die Entwicklung neuer automatischer Features führt auf der einen Seite nicht unbedingt dazu, dass Meetings barrierefreier werden, sondern kann auch ungewollte Nebeneffekte haben, wie z.B. falsche Transkriptionen (vor allem bei Zahlen) oder eine desinteressierte Wirkung auf Gesprächspartner*innen, wenn eine Person keinen Augenkontakt halten kann, weil sie Untertitel liest (vgl. Gonzalez, 2020, „possible harmful side effects“).

Auf der anderen Seite ist es in unserem vorgestellten Setting auch möglich, individuelle Bedürfnisse von Personen mit bestimmten Hörschwächen zu berücksichtigen und diese bei der Weiterentwicklung von Untertiteln mit einzubeziehen.

Quellen

Art Blokland, A. H. A. (1998): Effect of low frame-rate video on intelligibility of speech. Speech Communication 26, 97–103.

Ganten, P. (2021, 28.10.): Warum digitale Souveränität zwingende Grundlage für eine selbstbestimmte Zukunft ist, präsentiert auf der 20. Internationalen ILIAS-Konferenz. Online unter: https://www.youtube.com/watch?v=V5Y3V8NDm_M&list=PLJj1ocVfK3oaWambOklgJJMaWZ-yRUFGM&index=1 (zuletzt aufgerufen am 6.4.2022)

Keast, Q. (2020, 2.5.): Captions in video calls: better accessibility, but harmful side effects, UX collective, Online unter: https://uxdesign.cc/captions-in-video-calls-better-accessibility-but-harmful-side-effects-625d416f81af (zuletzt aufgerufen am 6.4.2022)

Körschen, M. et al. (2002): Neue Techniken der qualitativen Gesprächsforschung: Computergestützte Transkription von Videokonferenzen. Forum Qualitative Sozialforschung – FQS, Vol. 3(2) Art. 19. Online unter: https://www.qualitative-research.net/index.php/fqs/article/view/858/1865 (zuletzt aufgerufen am 6.4.2022)

Friebel, M. et al. (2003): „Siehst Du mich?" – „Hörst Du mich?" – Videokonferenzen als Gegenstand kommunikationswissenschaftlicher Forschung. kommunikation@gesellschaft, Jg. 4, 2003, Beitrag 1, 1–23. Online unter: https://duepublico2.uni-due.de/receive/duepublico_mods_00011055 (zuletzt aufgerufen am 6.4.2022)

Rink, I. (2020): Kommunikationsbarrieren. In C. Maaß & I. Rink (Hrsg.), Handbuch Barrierefreie Kommunikation. Kommunikation-Partizipation-Inklusion 3, 2020, S. 29–66.

Mälzer, N. und M. Wünsche (2020): Untertitelung für Hörgeschädigte (SDH). In C. Maaß & I. Rink (Hrsg.), Handbuch Barrierefreie Kommunikation. Kommunikation-Partizipation-Inklusion 3, 2020, S. 327–344.

Zoom-Blog (2021, 25.10.): Automatisch erstellte Untertitel von Zoom verfügbar für alle Benutzer mit kostenlosem Konto. Online unter: https://blog.zoom.us/de/zoom-auto-generated-captions (zuletzt aufgerufen am 6.4.2022)