Investigating and comparing remote recording methods

Preprint · December 2022

CITATIONS
0

READS
175

2 authors:

Stephanie Berger
Christian-Albrechts-Universität zu Kiel
17 PUBLICATIONS **53** CITATIONS
SEE PROFILE

Jana Neitsch
27 PUBLICATIONS **183** CITATIONS
SEE PROFILE

Investigating and comparing remote recording methods

Stephanie Berger[1], *Jana Neitsch*[2]

[1]ISFAS, Kiel University, Germany
[2]IfLA, University of Stuttgart, Germany
sberger@isfas.uni-kiel.de, jana.neitsch@ling.uni-stuttgart.de

ABSTRACT

Since the beginning of the COVID-19 pandemic in 2020, many researchers in laboratory speech sciences were forced to find new ways to collect their data with recording procedures functioning remotely, even with social distancing and travel restrictions in place. In this paper, we report a freeware remote method that offers several advantages, such as the direct storage on the experimenter's computer without data transformation or compression. We compare remote recordings via Zoom and Audacity created by the experimenters, as well as participants recording themselves with their own smartphones and the speech analysis program Praat. All methods are recorded simultaneously. The phonetic analyses investigate acoustic voice features such as pitch, formants and voice quality, and identifies durational differences between the methods. While remote recordings will likely never reach the quality of laboratory recordings, we propose to frame remote phonetic research as "digital fieldwork" in future publications, with legitimate results, always considering the requirements of the specific study at hand.

Key words: remote recording, duration, formants, voice quality, zoom, audacity, praat, smartphone.

1. Introduction

Since 2020, researchers, reviewers, and students of speech sciences all over the world frequently encounter phrases stating that the researchers could not complete their data collection due to the global COVID-19 pandemic or that they had to adapt their recording procedure and experimental setting to a non-face-to-face remote experiment due to hygiene restrictions. Since then, it has been difficult to conduct speech production experiments for the purpose of acoustic analyses (e.g., [1]). While the restrictions are now easing up, the development of various virtual methods is still relevant in order to build a toolbox with many options that researchers can choose from – depending on the requirements of the study in question – while at the same time offering opportunities for implementing more sustainable and eco-friendly research in the future. However, it is also important to be familiar with the settings and characteristics of such programs and apps, in order to be able to accurately assess the advantages and disadvantages of remote recording procedures.

During the pandemic between 2020 and 2022, speech scientists were almost forced to find alternative recording methods for their experimental research, relying less on laboratory recordings, and developing a new kind of digital fieldwork. These new methods involve both hardware (e.g., smartphone, handy recorder) and software (e.g., recording apps or programs for smartphones and notebooks). In particular (and even before the global pandemic), an increasing amount of reported production studies are conducted via the participants' smartphones (e.g., [1], [2], [3], [4], [5]) or by using particular online conferencing platforms (e.g., [6], [7]) – sometimes even both simultaneously. Another method used in fieldwork prior to the pandemic was recording participants with Praat ([8]).

However, all of these recording procedures strongly depend on the type and quality of the respective microphone, which may be integrated in the device itself or used externally (e.g., headset microphones). In addition, external microphones can have a Bluetooth function or a standard wire connection, which again substantially influences the recorded speech signal since the connection might share a bandwidth with a wifi

network ([9], [10]). Previous literature has also shown that remote speech production experiments are influenced by the quality of the internet connection ([6], [11]). Additionally, the resulting speech signal can be influenced if several programs simultaneously access the notebook's microphone (Thiele, p.c.). Furthermore, cloud-based systems (e.g., used in [12]) can be an issue because data security policies differ depending on the researchers' country and institution. For example, many European institutions, like the universities of the authors – need programs to adhere to European data security policies, which is not necessarily the case for platforms like Zoom or other programs, whose data are stored outside the EU by default. Institutions also may have different policies on programs that are allowed to be used for mentioning personal information for the same reason. Additionally, some institutions (for example, Kiel University, whose Zoom accounts were used for the present study) have disabled cloud storage for recordings, which is the recording method used in other studies (see [12]).

However, these circumstances make it more important than ever to report the experimental setup precisely by specifying the type of smartphone, the recording app, and the recording environment. A detailed report is not only necessary for experimental replications but also useful with respect to the statistical analysis. For instance, [12] provide their readers with seven different devices ranging from notebooks (e.g., MacBook Pro) to smartphones (e.g., Oneplus Nord N10) to include various brands. Additionally, the authors make use of cloud-based Zoom recordings, which can be difficult in some countries (see explanation above). These influencing factors indicate the necessity for reporting a detailed account of the experimental procedure to ensure the replicability of experiments and statistical analyses. Hence, we will not directly raise the question of the reliability of remotely collected data (e.g., [12]). We instead suggest an alternative recording solution and support the idea of finding new options, sharing, testing, and optimizing them, while gathering as much experimental data about the recording systems as possible.

1.1. Previous research

Several studies have addressed recording methods outside of laboratories in the years before and during the global pandemic. According to [1], previous remote production studies tend to report two main findings: first, f0 was usually observed to be a stable factor that was affected neither by the file format of the recorded speech data nor by the respective device that was used (cf. [1], [13],], [14],], [15],], [16]).

Second, various studies have observed that the F1-F2 vowel space was often distorted ([13],], [17],], [18]). For instance, regarding the formant analysis, [1] report concerning F1, F2, and F3 that data recorded via Zoom consistently had significantly lower formant values, mainly when extracted in Praat, than data recorded with their baseline method (Zoom H6 Handy Recorder) or a smartphone recording. Furthermore, [6] found that there was less variation for male speakers than for female speakers across devices and that, in general, the "back-mid area of the vowel space is particularly vulnerable to variations between devices" (p. 1216).

Other studies have also reported durational differences for durations overall and segment durations in particular regarding cloud-based data storage compared to local storage recording methods. This was found to be the case especially for fricatives ([12]). Durational differences were also found by [19], who reported on irregular filtering algorithm artifacts causing duration differences of about 119 ms for Zoom compared to smartphones. Duration differences in smartphone recordings have been reported to be more reliable (e.g., [12]).

Another measurement that has been reported in previous literature is intensity, and it has been reported to vary based on recording settings ([1]). For example, [1] report significant drops in intensity in recordings via Zoom. Intensity is a very sensitive factor that highly depends on the distance between the source and the microphone ([20], p. 21), which cannot be guaranteed during a whole experimental session if not done in a laboratory setting. Therefore, from our point of view, intensity still poses a challenge for future investigations. In the current work we report intensity results for recording conditions that were made with the same headset at the same time, which reduces the effect of the microphone. The other recording condition is excluded for intensity measures as they were made simultaneously but with a different microphone.

Finally, concerning voice quality, [2] reported no differences across different devices for HNR (harmonics-to-noise ratio), jitter, or shimmer. In contrast, [12] reported an influence of the recording environment that they used on voice quality. For instance, their findings suggest that jitter was greater when recordings were made in a conference room compared to those made in the laboratory. However, voice quality measurements are also strongly dependent on the quality of the microphone. They are more reliable if a headset microphone is used to

minimize differences in the distance between the speaker and the microphone. This study focuses only on HNR as it is seen as more stable overall than jitter and shimmer ([21]).

1.2 Premise and predictions

In this study, we tested a recording method for Windows-based computers that has several advantages: It a) respects data security issues since the sensitive data do not have to be exchanged via the internet but are directly created and stored on the experimenter's computer, b) supervision of the experiment session is possible, and c) participants will not have to download any tools like Praat. Additionally, the method uses a freeware program and is compatible with any video conferencing app.

However, our review of previous studies suggests that this solution will still cause major issues for phonetic measures and can by no means be considered an ideal recording method. As with other remote recording options, this also depends on the requirements of the type of investigation that researchers want to run, and it might not be suitable for every type of study or desired measurement. Since the speech signal is – as in other studies – transmitted via (an online) video conferencing app, we still expect effects of compression and noise cancellation algorithms. We propose framing research based on remotely recorded speech materials as "digital fieldwork" to be seen as distinct from lab recordings.

In this study, we suggest using Audacity ([22]) as a way to record the incoming sound from a video conferencing app, in this case, Zoom ([23]). (Unfortunately, recording the incoming sound is not possible without installing a virtual microphone on iOS systems.) To test our method, we compare the Audacity recordings with simultaneous recordings made by the participants themselves via Praat ([8]) and their own smartphones. Praat constitutes the baseline condition in our investigation because there is no direct signal transmission via the internet and thus possible effects from such signal transmission are avoided. We also made a local recording of the Zoom call, and the participants recorded themselves on their smartphones. All recordings took place simultaneously; see Section 2.4 below for a detailed description of the experiment design.

We make the following predictions in this study:

1. In line with previous research (cf. [1], [13], [14], [15], [16]), we predict that f0-related measurements are stable across the recording methods.
2. Formant values will be lower in the Audacity and Zoom recordings, as the signal was transmitted via the internet, compared to the baseline recordings in Praat and the smartphone recordings (cf. [1]).
3. The mid-back area of the vowel space will be more variable between devices and recording methods ([6]).
4. The vowel space of male speakers will be less variable than the vowel space of female speakers ([6]).
5. The duration of the target utterance will be shorter in the Audacity recordings because the speech signal traveled through the internet and is affected by Zoom's algorithms (cf. [1]), compared to the Praat and smartphone recordings.
6. The fricative onset of the stimuli in this study will be shorter in the Zoom and Audacity recordings than the Praat and smartphone recordings because of the reaction time of Zoom's noise cancellation algorithm (see also [12] and [19]).
7. The Zoom and Audacity recordings will be misaligned and delayed compared to the Praat and smartphone recordings because of the internet transmission.
8. There will be a drop in mean intensity in the Audacity and Zoom recording conditions as the signal had to travel through the internet and experienced some compression that way (cf. [1]).
9. Finally, voice quality (i.e., HNR and Hammarberg Index) will not differ significantly between the Audacity, Zoom and Praat recording methods since room size and microphone are identical (cf. [12], [2]). The smartphone recording, however, will differ because it was produced with the internal microphone rather than the same headset as the other three recordings.

2. Data and Methods

Making recordings via a video conferencing app has major advantages over other methods – such as, for example, giving participants materials to record themselves on a smartphone. Virtual in-person experiments via video conferencing can be supervised by the experimenters. Participants can be asked to repeat a part of the material,

for example, in case they made a reading mistake in read speech tasks, or if the internet connection cut out part of the spoken material. At the same time, the number of repetitions can be logged by the experimenters. If participants record themselves and simply hand in one finished recording at the end, it is not possible to reconstruct how many times the recording was made, which can undermine spontaneity.

However, recordings via video conferencing app also have issues like internet connectivity, data compression, and in-built noise reduction algorithms that affect the speech signal. The method we are proposing can therefore not be considered ideal or without its problems, but it offers an opportunity for remote research, depending on the requirements it has to fulfil for the specific research project.

2.1 The choice of recording apps

In this study, we compare the recordings made with four methods simultaneously: Zoom ([23]), smartphone, Audacity ([22], a method that – to our knowledge – was not yet investigated in the previous literature), and a local Praat recording ([8]). The participants recorded themselves simultaneously in Praat and on their smartphone, and saved the files to a university cloud folder to which only each participant and the experimenters had access. The Praat recording is saved as a WAV file, but depending on the app, there are different file types from the smartphones. The experimenters simultaneously made a recording in Zoom as well as a recording of the incoming sound in Audacity.

Zoom was chosen as the video conferencing platform because it is widely distributed. It also has an in-built recording function which must be activated specifically in the account. Zoom not only saves a complete audio recording of the entire conversation, but also speaker-separated audio files which are saved as stereo m4a files. They were converted into mono WAV files by the experimenters before analysis, meaning that additional data compression is used. Zoom also saves a video file. Because the video is automatically saved, this must be included in consent and data security information distributed to participants before the experiment, and – depending on the study – deleted immediately. The experimenter can disable saving the recordings on a Zoom cloud server and can enable only local storage, which means that once a file is deleted on the recording computer it will no longer appear in the Zoom account. For this study, only local storage was enabled as a requirement of Kiel University, which provided the Zoom account running the experiment sessions and making the recordings. Newer versions of Zoom (from April 2021 onwards) also have the option to transmit and, ultimately, record the original in-coming sound without noise reduction. This option was not available when the recording process was started, but should be tested in future studies. Since noise cancellation was still enabled in the current study, the recordings are made with more technical influence on the speech signal. That also means that we can – in theory – investigate the most problematic conditions and most severe differences from the control Praat recording.

In order to circumvent data security issues with Zoom usage (depending on the research institution), it is worth considering the use of recording methods that are not automatically saved on a platform's server or listed and accessible from an account. **Audacity** offers an opportunity to create a recording of the incoming sound directly on the experimenter's computer without saving it online. The recordings can be saved directly as WAV files and do not have to undergo additional data compression. Although the sound for the Audacity recording is transmitted from the speaker through Zoom and therefore has to travel through the internet and experiences compression to some extent, there is no need for further file and data compression caused by re-formatting. Another advantage is that this method should work with other common video conferencing platforms like Big Blue Button, Skype, WebEx, etc. Screen sharing can be used in Zoom to show the tasks to the participants, and the experimenter creates the recordings in the background. Audacity is already widely used in the phonetics community and is free to use. That means it is also a feasible recording program for students working on class projects or theses. One caveat is that recording the incoming sound is only easily possible on a Windows computer. Mac users would have to invest in a virtual microphone set-up for their device, as MacBooks do not support the recording of incoming sound without additional software (this is not an Audacity, but an Apple caveat).

Participants were asked to record themselves simultaneously on their own smartphone and in a professional audio recording tool (Praat, in most cases, Audacity when technical issues arose with the Praat recording feature). The recording in **Praat** was chosen so that a professional recording with the same headset as the Zoom call without compression on its way through Zoom from speaker to experimenter was available for comparison. The participants were instructed to record a Mono sound with a 44.1 kHz sampling rate and to save the file in WAV format.

The **smartphone** recording (referred to also as Phone recording or Phone from here on) was made as another comparison recording because the initial idea for the project stemmed from discussions in the linguistic colloquium at Kiel University. Students had to instruct others to make recordings for their class projects or theses because the recordings could not be made in person. The participants in the students' experiments made recordings on their personal cell phones and sent the recordings to the students running the experiments. However, this meant that the students were not able to control how many times the recordings were made before one version was sent to them. On top of that, the suitability of cell phone recordings has been in constant discussion (see, e.g., [1], [2]), so it seemed advantageous to take the opportunity and also compare their quality with that of other methods. The smartphones recorded in different file types, mostly in m4a format, but also WAV, mp3, and AAC. These were converted into mono recordings and re-formatted as WAV files to allow analysis in Praat.

2.2 Materials

The materials were presented to the participants in a PowerPoint presentation via screen sharing. The participants were shown a situational context, which they were instructed to read silently for themselves. Afterwards, an interrogative (referred to as target utterance from here on) was presented, which the participants had to produce as naturally as possible, but fitting in the previous context. In total, each speaker produced 24 target utterances that contain a representative set of German vowels, fricatives, and nasals to allow for a respective analysis. The target utterances were pre-evaluated in several studies (e.g., [24], [25], [26]). Since this paper focuses on the recording methods, the effect of the different situational contexts on the analyzed interrogatives is not investigated. Only the comparison of the different recording conditions is reported. Figure 1 shows an example of a situational context and the corresponding interrogative, both in German and translated.

Situational context:	Translation:
Du gehst mit deinen Freunden in einem asiatischen Restaurant essen und siehst, dass Wasabi dazu bestellt werden kann. Dich hat schon immer interessiert, wie diese Paste schmeckt und du fragst dich, ob deine Freunde gemeinsam mit dir probieren wollen. Du sagst:	You and your friends go out for dinner to an Asian restaurant and you see that you can order wasabi with your meal. You have always been interested in how this paste tastes and are wondering if your friends want to try it with you. You say:
Wer isst denn Wasabi? *trsl.: Who eats PRT wasabi?*	

Figure 1: *An example of a situational context presented to the participants (in German, with English translation, abbreviated as trsl.) and the target utterance that participants produce (PRT means a particle and is the English placeholder for the German particle 'denn').*

2.3 Participants

The sample for this study includes six female and six male speakers (ages 21 to 52, average age = 35.1 years, *SD* = 9.8 years). None of the speakers reported learning a language other than German before the age of six, and all grew up in Germany. All participants were coached in how to use Praat for the recordings.

2.4 Set-up and procedure

Audacity version 2.1.3 ([22]) was used throughout the recordings. The Zoom version could not be kept constant as updates needed to be installed to ensure that the screen sharing in particular would continue working, especially because the recordings were made over the course of roughly two months (versions 5.6.1 to 5.7.1).

The physical set-up from the experimenters' side was kept constant for all recordings. All recordings were made on a Lenovo IdeaPad 110-17ACL (from ca. 2017). The computer was placed on a desk and connected to the internet router via a cable. The distance between the router and the computer and the cable placement were kept as constant as possible.

Figure 2 below depicts the recording set-up of the experimenters. The computer ran the PowerPoint presentation with the stimuli, the Zoom call, and the Audacity recording. All other programs were closed. Video

on the call was enabled on both sides unless the internet connection on one end could not handle the video and negatively influenced the audio quality. The experimenter did not watch the participants and was muted the entire time to avoid interfering with the audio. If the audio was not transmitted correctly, the experimenter briefly activated the microphone to request the participant to produce the target utterance again. Only the last target utterance that was produced was included in the analyses. The computer's loudness was set to 50% for all experiment trials.

Audacity has a variety of recording settings available. The audio host "Windows WASAPI" (= Windows Audio Session API) was chosen because it allowed for the recording of the incoming sound. Most importantly, the microphone is set to the channel with the sound from the Zoom call, so either a headset or the integrated loudspeakers (see Figure 3).

The participants were sitting in front of their computers with their headsets on. Six participants had in-ear headphones with the microphone on the cable, six participants had on-ear headphones with cable connection and integrated microphones. In order to have a realistic remote recording session, the only restrictions were that the microphone was part of the headphone and – most importantly – that they were cable-connected to avoid compressed sound transmission over Bluetooth (see [9]). Detailed information on participants' headsets and computers is available in Table 1.

The first slide of the PowerPoint presentation detailed the set-up the participants should have: the phone on the table near the edge between them and the computer, so they could still comfortably reach either the return key, the left, and right arrow keys, or their mouse to click through the experiment, see Figure 4 for a schematic overview. Once the participants had set up their end of the recording, the Zoom, and Audacity recordings were started from the experimenters' side. The resulting additional recordings before the experiment began were not included in the analyses. Afterwards, the participants were able to control the screen and click on the next slide, which was an instruction page. They were told to start their Praat recording with default settings (44.1 kHz sampling rate). In order to allow for later alignment of the recordings, participants were instructed to knock on the table once all recordings were running. They then read a situational context silently before producing the target utterance out loud. At the end of the experiment, they were instructed to stop the Praat recording and save it as a WAV file. A link to a folder on Kiel University's cloud server was provided, where the WAV file and the smartphone recording were uploaded. The cloud server adheres to European data security protocols, and participants were made aware of this in a consent form before the experiment started.

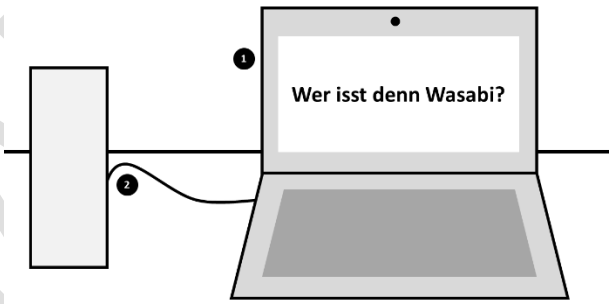

Figure 2: *Physical experiment set-up for experimenters: 1) laptop with webcam running Zoom call and recording, Audacity recording, experiment slides via PowerPoint, 2) LAN connection*.

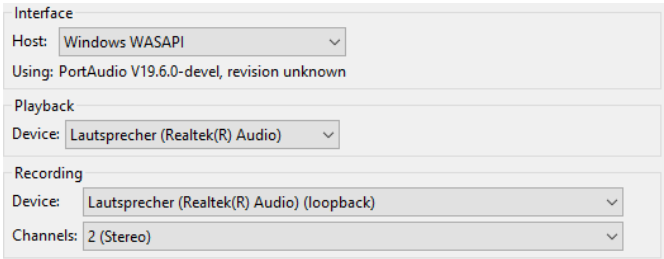

Figure 3: *The settings of the recording devices in Audacity (via "edit", then "preferences" under the topic "device"). The audio host is "Windows WASAPI", and the recording device uses the loopback function with a headset or loudspeakers.*

Table 1: *The devices and programs used by the speakers. Speakers s01 to s06 are female, speakers s07 to s12 are male.*

S	Computer (year)	Headset	Phone	Phone app	Praat
s01	MacbookPro, 13 inch (2020)	Apple iPhone Headset (in-ear)	iPhone 11 Pro	Garage Band	6.1.42
s02	Lenovo ideapad 110-17ACL (2017)	SilverCrest SKH 50 B1 (on-ear)	Xiaomi Mi A2	AudioRec	6.1.48
s03	Lenovo ideapad 330S (2018)	Soundcore Life Q30 (in-ear)	iPhone 8	Sprachmemos	6.1.12
s04	Asus X551MA (2014)	JBL TUNE 110 (in-ear)	Samsung Galaxy A51	Diktierfunktion	6.0.17
s05	hp 17-x037cl (2016)	ISY Wired Earbuds Headset IIE 3700-GY (in-ear)	Huawei Mate 10 lite	Diktierfunktion	6.1.38
s06	MacbookPro (2013)	Apple iPhone Headset (in-ear)	iPhone 11	Sprachmemos	6.1.41
s07	Lenovo T450p	Sennheiser PC 8.2 (on-ear)	Samsung Galaxy S7	Diktierfunktion	6.1.09
s08	Apple iMac 18,2 (2018)	Sony MDR-1RNC (on-ear)	iPhone 11 Pro	Sprachmemos	6.1.42
s09	MacbookPro, 15 inch (2018)	Apple iPhone Headset (in-ear)	iPhone 12 Pro	Sprachmemos	6.0.14
s10	Acer Spin 5 (2019)	JBL T290 (in-ear)	Asus Zenfone 6	Audiorecoreder	6.0.52
s11	Lenovo Ideapad S340 (2019)	Speedlink Orios RGB 7.1 Gaming Headset (on-ear)	Samsung Galaxy A20	Diktierfunktion	6.1.12
s12	MacbookPro, 11.4 inch (2019)	Equip Chat Headset (model nr. 245302) (on-ear)	iPhone SE	Sprachmemos	6.1.35

As mentioned above, the computers and headsets differed across participants in order to have a realistic situation when working with data collected remotely. Since the participants used the same devices for both the Praat recording and the transmission to Audacity, the factor device should not have an effect.

2.5 Data treatment

The four different recordings for each speaker were time-aligned as best as possible in Audacity. In most cases, the knock on the table before the realization of the first utterance was used to align the recordings. Both the waveform as well as the spectrogram were used for alignment. Sometimes, the knock was not picked up loudly enough by the recordings. In those cases, a clear (spectrogram) distinction in the first utterance was used. The utterance onset could not be used for alignment because it was the fricative [v], which was frequently shorter in the Zoom and Audacity recordings, likely because of the influence of Zoom's noise reduction algorithm, see Section 3.3 below for an analysis.

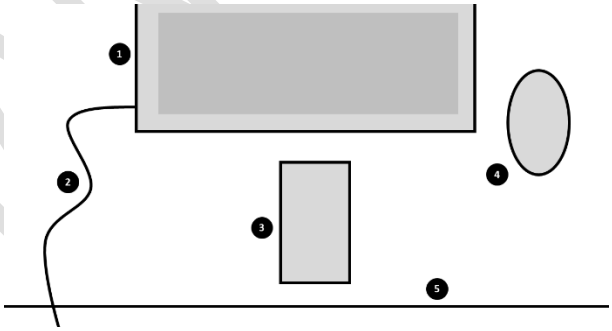

Figure 4: *Physical experiment set-up for participants: 1) keyboard, 2) mouse (return keys, left/right arrow keys, and mouse options for clicking through the experiment), 3) cable-connected headset, 4) smartphone, 5) edge of the table.*

The isolated target utterances in the recordings were then cut and saved to new files. A Praat script was used to cut the sound files based on labeled TextGrid intervals. The intervals had the same length and place around the target utterance for all recording conditions so that the resulting audio files were all the same length and duration, and timing measurements were possible. The cut sound files were then run through WebMAUS ([27]) to obtain TextGrids with automatically generated word- and segment level annotations and segmentations. These TextGrids were then manually corrected. Afterwards, a Praat script labeled vowel centers with the vowel SAMPA symbol from the MAUS annotation.

The process of alignment highlighted an issue with the recording via Zoom which made manual corrections of the TextGrids necessary. This issue occurred for all speakers, though to differing degrees. Even though all four recordings were exactly aligned at the beginning of the recording, they were audibly misaligned by the end of the recording. The Zoom and Audacity recordings remained aligned with one another, as did the Praat and Phone recordings. The MAUS annotations were manually corrected for the online recordings (Zoom and Audacity) and the offline recordings (Praat and Phone). In some rare cases, the Phone and Praat recordings were slightly misaligned as well. In those cases, the segment boundaries were also adjusted. It follows that the boundaries of the segments could not be aligned perfectly over all conditions, and therefore slight measurement differences are to be expected. The manual corrections were based on auditory impressions as well as visual observations in the Praat spectrogram.

2.6 Data analyses

The recorded data were analyzed for each of the target utterances evaluating pitch measurements, including the mean, the maximum and minimum of f0 and median pitch (all in Hz), excursion size (semitones, st), and standard deviation of pitch (st) as dependent variables. Additionally, we measured the duration (ms) of the whole target utterance, mean intensity (dB), and the voice quality parameter HNR (Harmonics-to-Noise Ratio in dB). The spectral slope (related to a speaker's vocal effort) is reflected in the information described by the Hammarberg Index ([28]). The Hammarberg Index, which is associated with expressivity ([29]) and emotions ([30]), is usually defined as the intensity difference between the maximum intensity in a lower frequency band (0–2000 Hz) versus a higher frequency band (2000–5000 Hz).

Intensity could be measured because the conditions were identical in the three recording methods; the Phone recording was excluded here. The sound was recorded in or – in the case of Audacity and Zoom – transmitted simultaneously from the same room with the same microphone and computer on the speakers' side. The experimenter always made the Audacity and Zoom recordings on the same device. That means, if participants moved and changed their position and distance between mouth and microphone, it was identical for all three recordings.

All aforementioned measurements were made with the ProsodyPro script ([31], version 5.7.8.1), except for the standard deviation of pitch, which was added by the authors. The ProsodyPro measurements were made for the entire utterance except for the voice quality features HNR and Hammarberg Index, which were measured on the word level, focusing only on the words *Wer* ('who') and *denn* and looking at the measurements for each word separately.

Additionally, formants were measured in the center of the vowel using the Praat-internal "To Formant (burg)..." function, following the method in [32]. In this study, only monophthongs were investigated, and a length distinction, e.g., between long and short [a], was also not included. Other studies, like [6], only compared the results of one speaker per gender. This study includes six male and six female speakers. For comparison, the formant values F1 and F2 were therefore normalized using the Lobanov transformation included as the `normLobanov()` function in the R package `phonR` ([33]) and therefore used one of the standard formant transformations available.

Furthermore, the time difference between the start of each phrase to the start of the cut sound file (see Section 2.5 for an account of the data preparation) was calculated with a script by the authors. This was possible because the recordings from the different methods were time-aligned in the beginning of the recording, and a script cut the recordings up at precisely the same times around the phrases before analyses. This measure was chosen to investigate the misalignment between the different recording methods that was noticed while aligning the recordings. Finally, the duration of the fricative [v] at the stimulus onset was measured to investigate the effect of Zoom's noise cancellation algorithm.

For the statistical analyses, we calculated linear mixed effects regression models using the `lmer()` function in R Studio ([34], version 2021.09.0, R version 4.1.1), with *Recording method* (Praat, Audacity, Zoom, Phone) and *Gender* (male vs. female) as fixed factors, and *Speakers* (S), and *Items* as crossed random factors for the adjustments of intercepts (cf. [35], [36]). Random slopes were added for the fixed factors to the random-effects structure. They were only kept if they improved the model fit ([37], [38]). For a comparison of the models, we used the `anova()` function to be able to compare the LogLikelihood and to arrive at the p-values.

3. Results

The following sections present the results obtained from the measurements and subsequent statistical analyses. First, the pitch-related features will be presented, followed by the formant results. Then, durational aspects, voice quality, and intensity-related measures are reported.

3.1 Pitch-related features

Six different pitch-related features were analyzed. There was no significant main effect on the five f0 measurements mean, maximum, minimum f0, standard deviation of pitch, and median pitch (all *p*-values >= .06). Only the gender of the speakers had a significant main effect on the pitch features, with female speakers having higher values, as can be expected (mean f0: β = 105.44, *SE* = 11.21, *df* = 10.11, *t* = 9.41, *p* < .001; maximum f0: β = 146.89, *SE* = 21.33, *df* = 10.22, *t* = 6.89, *p* < .001; minimum f0: β = 61.92, *SE* = 7.15, *df* = 10.02, *t* = 8.67, *p* < .001; standard deviation of pitch: β = 9.19, *SE* = 2.28, *df* = 10.05, *t* = 4.02, *p* = .002; median pitch: β = 104.59, *SE* = 10.51, *df* = 10.01, *t* = 9.96, *p* < .001).

Excursion size, on the other hand, showed an interaction between all levels of the variables *Recording method* and *Gender* (all *p*-values < .001). A Tukey post-hoc pairwise comparison revealed that the excursion size measurements of the male speakers in the sample were significantly lower in the Phone recordings than the other three recording methods (Praat: β = -4.63, *SE* = 0.79, *df* = 1107.0, *t ratio* = -5.86, *p* < .001; Zoom: β = -4.84, *SE* = 0.79, *df* = 1107.0, *t ratio* = -6.13, *p* < .001; Audacity: β = -5.57, *SE* = 0.79, *df* = 1107.0, *t ratio* = -7.06, *p* < .001). There were no such effects for the other recording methods and no significant effects overall for the female speakers (all *p*-values > .98). Figure 5A shows the mean excursion size for all speakers, while Figure 5B shows the excursion size split by speaker. There is extreme variation in the recordings with Audacity, Praat, and Zoom, especially for three of the male speakers. This pattern also occurs with other measurements (e.g., maximum f0, plus the other pitch measurements for speaker s08).

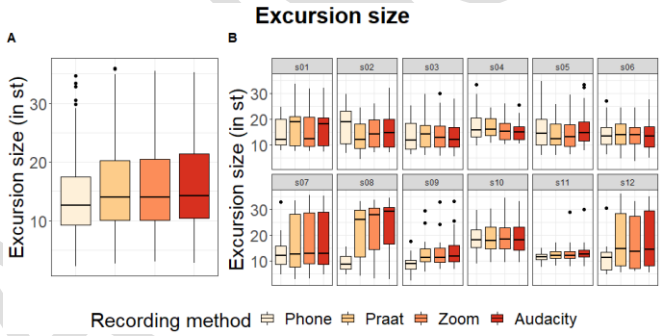

Figure 5: *Excursion size for A) all speakers together and B) the individual speakers. Speakers s01 to s06 are the female speakers in the sample, and s07 to s12 are the male speakers.*

3.2 Formants

We analyzed the normalized formants F1 and F2 for this study. For both formants, there were neither significant main effects nor interactions of *Recording method* and/or *Gender* on the formant measures (all *p*-values > .60). However, Figure 6 shows that male speakers tend to have a smaller vowel space than female speakers. When looking only at the mean values of the formants per speaker group and recording method (represented by the SAMPA symbols in the figure), the measurements of the vowels [e ɛ ɪ a ə ɔ] were very similar between the recording methods. The measurements of [i] differ, though, and there was a lot of variability and overlap with the back closed/half-closed vowels [u ʊ o].

3.3 Duration-related features

Three duration-related measurements were included: the duration of the target utterance as a whole, the duration of the first segment of the target utterance – [v] as the consonant onset of the *wh*-element *Wer* ('who') – as well as the time difference between the start of each recording chunk (time-aligned between the recordings) and the

beginning of the phrase to test whether there are significant misalignments happening because of delay in internet transmission.

There were no significant *Gender* effects for any of the three duration measurements (all *p*-values > .3). Furthermore, there were no significant effects of *Recording method* on the target utterance duration (all *p*-values > .57).

However, there were significant main effects of *Recording method* on the duration of the initial [v] and the potential misalignment. The duration of [v] does not differ significantly between Audacity and Zoom (β < 0.001, *SE* = 0.002, *df* = 1110.0, *t* = 0.03, *p* = .98), but the differences between Audacity and Praat (β = 0.007, *SE* = 0.002, *df* = 1110.0, *t* = 3.56, *p* < .001) and Audacity and Phone (β = 0.007, *SE* = 0.002, *df* = 1110.0, *t* = 3.73, *p* < .001) are significant. The Praat and Phone recordings do not differ from each other (β = < 0.001, *SE* = 0.002, *df* = 1110.0, *t* = 0.173, *p* = .86). A post-hoc pairwise comparison revealed no significant difference between Zoom and Audacity on the one hand, and Praat and Phone on the other hand. Both the Audacity and the Zoom recordings had significantly shorter [v] durations than the segments in the Praat and Phone recordings, see also Figure 7.

Finally, there was also an effect of the *Recording method* on the misalignment of the recordings. Again, there was no significant difference between Audacity and Zoom (β = -0.0008, *SE* = 0.007, *df* = 562.1, *t* = -0.131, *p* = .90). There were, however, significant differences between Audacity and Praat (β = -0.02, *SE* = 0.009, *df* = 17.02, *t* = -2.78, *p* = .01), as well as Audacity and Phone (β = -0.04, *SE* = 0.02, *df* = 11.95, *t* = -2.3, *p* = .04). There was no significant difference between the Praat and Phone recordings (β = -0.01, *SE* = 0.01, *df* = 12.66, *t* = -0.93, *p* = .37). Pairwise comparisons in the post-hoc analysis were not significant, but a visual inspection of the data (see Figure 8) suggests that the Audacity and Zoom recordings lag behind the Praat and Phone recordings. This result fits the impression of the authors during the data treatment that, at some point, the recordings were no longer aligned, even though the beginning of the recordings were. This was the case for most speakers, but not all of them.

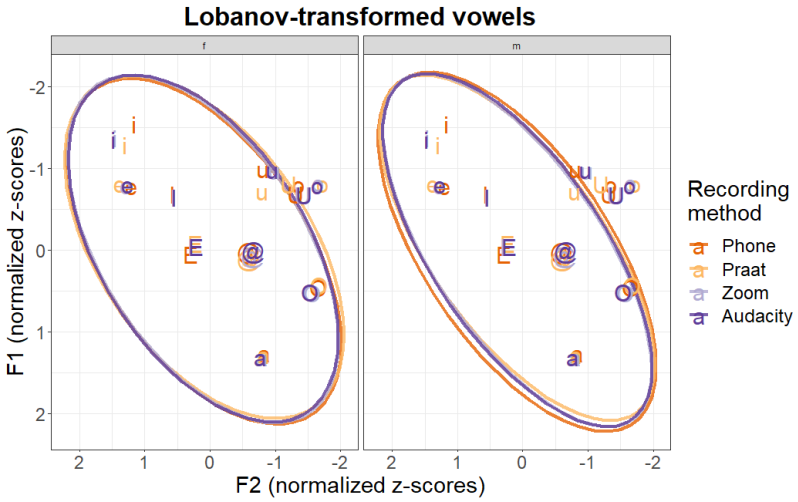

Figure 6: *Vowel spaces of the female (left) and male (right) speakers in the sample. The formant values are Lobanov-transformed.*

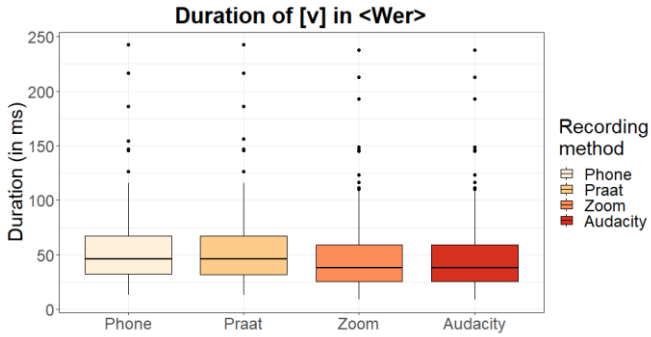

Figure 7: *The duration of the initial [v] in <Wer> ('who') depending on the recording method.*

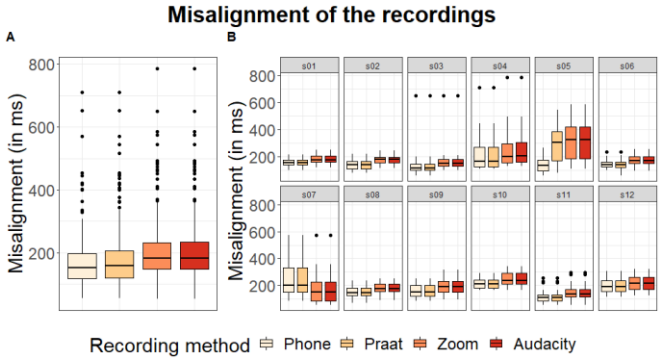

Figure 8: *The time difference between the beginning of the file and the target utterance start depending on the recording method. This shows the misalignment between "online" and "offline" recordings. A) shows all speakers combined, B) shows the results for each speaker*.

3.4 Mean intensity

For mean intensity (in dB), Phone was disregarded to assure comparable results. For *Gender*, we found no effect ($p > .1$). Regarding *Recording app*, we found that Praat ($\beta = 3.11$, $SE = 0.21$, $df = 780.29$, $t = 15.02$, $p < .001$) as well as Zoom recordings ($\beta = 3.68$, $SE = 1.18$, $df = 11.07$, $t = 3.10$, $p = .01$) differed significantly from the Audacity recordings. An additional visualization of the recordings of the single participants split by recording app indicates that while Audacity seems to show consistent results, the intensity of both the Praat and Zoom recordings vary across *Speakers*, see Figure 9.

3.5 Voice quality

Voice quality was measured on the word level for the sentence-initial *wh*-element *Wer* ('who') and the modal particle *denn* (lit: 'then') since they were identical across all target utterances. We used the harmonics-to-noise-ratio (HNR in dB) as well as the Hammarberg Index. The results for HNR show no interactions for the initial *Wer* (all *p*-values > .5). Our model showed a main effect of *Gender* for the initial *wh*-element ($\beta = 3.32$, $SE = 0.54$, $df = 10.03$, $t = 6.18$, $p < .001$). There was no effect of *Recording method* (all p-values > .49). Regarding *denn*, we found that HNR values were significantly higher in Zoom ($\beta = 0.81$, $SE = 0.30$, $df = 11.98$, $t = 2.68$, $p < .02$) and Audacity recordings ($\beta = 0.84$, $SE = 0.31$, $df = 11.53$, $t = 2.71$, $p < .02$) compared to Praat recordings. Furthermore, we found a main effect of *Gender* with higher HNR values for female speakers ($\beta = 6.32$, $SE = 0.66$, $df = 9.94$, $t = 9.62$, $p < .001$). Phone did not differ significantly from any of the other levels (all *p*-values > 0.40) with respect to HNR. This is depicted in Figure 10 for *denn*.

The results concerning the Hammarberg Index indicate no interactions for *Wer* (all *p*-values > .5) and our model showed no effect of *Gender* ($p = .81$). We found an effect of *Recording method* for 'who', with Zoom ($\beta = 0.96$, $SE = 0.29$, $df = 188.37$, $t = 3.40$, $p = .001$) and Audacity ($\beta = 0.98$, $SE = 0.29$, $df = 191.37$, $t = 3.40$, $p < .001$) having significantly higher values than Praat. Our results show similar findings for *denn*, since we found no interaction or effect of *Gender* (all *p*-values = .60), but an effect of *Recording method*. This effect again indicates higher values for Zoom ($\beta = 0.60$, $SE = 0.19$, $df = 823.00$, $t = 3.06$, $p = .002$) and Audacity ($\beta = 0.59$, $SE = 0.19$, $df = 823.00$, $t = 3.02$, $p < .003$) compared to Praat. Phone did not differ significantly from any of the other *Recording method* levels (all *p*-values > .33). Figure 11 shows the Hammarberg Index results for *Wer* as an example.

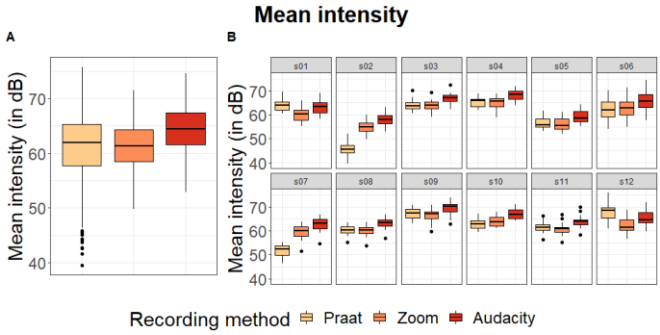

Figure 9: *A) The mean intensity depending on recording method across speakers. B) The mean intensity depending on recording method separated by speaker. Note that the Phone recordings are excluded to account for the difference in microphone*.

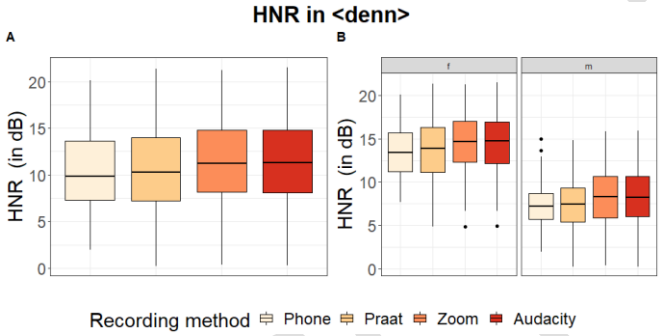

Figure 10: *The HNR results per recording method in <denn> (lit. 'then'), A) for all speakers combined, and B) split by Gender*.

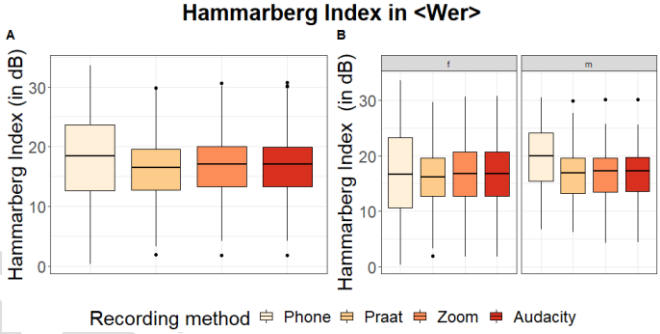

Figure 11: *The Hammarberg Index results per recording method in <Wer> ('who'), A) for all speakers combined, and B) split by Gender*.

4 Discussion

This study investigated different remote recording methods (Audacity, Zoom, Praat, and Phone) and compared them to see how they differ. For this purpose, several acoustic measures were taken from simultaneous recordings, and compared statistically. Audacity, as a remote recording option that is created and stored on the experimenters' devices while also allowing supervision of the participants, is mainly the focus of the discussion, compared against the other three options.

The results for the pitch-related features largely fit the previous literature (see also [1], [13], [14], [15], [16]): Most of the pitch features (mean, maximum, minimum, final f0, median pitch, and standard deviation of pitch) showed no significant effect of the recording method, which suggests that these features tend to be less affected by different recording apps, internet transmission, and data compression. The only significant results of the recording method occurred in the form of an interaction with speaker gender: the excursion size of the male speakers was significantly smaller in the Phone recordings than the other methods. Looking at speaker-specific results suggests that there are strong individual differences between the different recording methods – especially for excursion size, but also for the other pitch measures. It seems as though particularly excursion size as a

measurement is extremely sensitive to the influence of the headset. The three recording conditions with so much variation were made with the same headset, and the Phone recording was made with a different microphone. The substantial individual differences warrant further research, also in order to find out if a specific type of headset yields more consistent results on an individual level. Prediction 1 can therefore be supported for the majority of features for the current sample. For the analysis of the mentioned pitch features (with some reservedness towards excursion size), the use of Audacity (via Zoom) appears to be as viable as Phone, Praat or direct Zoom recordings. This is the case especially after future research finds more information on headsets, and the size of the speaker sample is increased.

Unlike previous research (cf. [1]), we found no significant differences between the recording methods in terms of formant values. That means that Prediction 2 is not supported by the current data, suggesting that remote recording methods – including Audacity – are viable for formant investigations. As suggested in [6], we could visually observe that there is more variability between the recording methods in the mid-back area of the vowel space, regardless of speaker gender, therefore partially supporting Prediction 3 for the current sample. We found no gender effects, which is likely because of the normalization procedures that were necessary for working with several speakers. Nevertheless, visually there is no difference in variation between male and female speakers concerning the different recording methods – unlike the results presented by [6] – so Prediction 4 is not supported by our data.

[19] suggest that Zoom recordings reduce the duration of target utterances overall compared to smartphone recordings. This was not the case in the current study: the statistical analysis of the target utterance duration yielded no significant differences between the four recording methods, meaning Prediction 5 is not supported by the current data. However, this should be revisited in future studies with connected speech, as it is likely that the utterance duration is affected by the internet connection or noise cancellation algorithms, but this does not arise with isolated utterances.

We did, however, find that the duration of the fricative onset [v] of the isolated target utterances was significantly shorter in the Zoom and Audacity recordings than the Praat and Phone recordings, as was predicted in Prediction 6, based on [12] and [19]. This result suggests an influence of the noise cancellation algorithm of Zoom, as the differences in fricative duration occur in the recordings that are created after internet transmission and transmission via Zoom, while the "offline" recordings are not affected. It is likely that the noise cancellation algorithm takes some time to register that the sound is not background noise that needs to be filtered out, but a speech segment that needs to be transmitted to the listener. The difference between the "online" and "offline" recordings is only slight, suggesting that there is very little latency of the algorithm, but the difference is significant nonetheless. With the fricative [v], this investigation used a segment that is inherently very similar to the background noise the algorithm is supposed to filter out. Additionally, the experiment procedure included only short target utterances between longer stretches of silence where the speakers read a context silently to themselves. This prolonged silence and sudden transition into speech probably increased the effect of the noise cancellation algorithm. Future studies should a) compare the different recording methods again, but look at effects on longer stretches of connected speech, as it is assumed that the effect of noise cancellation on segment duration is minimized in that context, and b) re-run the experiment without the noise cancellation algorithm active in Zoom, which is an available feature since version 5.6.3. (April 2021).

Prediction 7 is also supported by the current data: results show that the Zoom and Audacity recordings are significantly delayed compared to the Praat and Phone recordings, suggesting that misalignment takes place at some point in the recordings. The misalignment occurred for all speakers, but to varying degrees. That suggests that the reason for the misalignment may be that the internet connection was overloaded, either by the experiment and Zoom call itself, or because too many people in the vicinity of the individual network were online at the same time. The fluctuation in internet connection may cause additional buffer times in the Zoom transmission that are not perceivable in the conversation, but appear in the comparison of simultaneous recordings. This is important to note even though phonetic studies potentially using this recording method are unlikely to compare several recordings, which would be only relevant for specific methodological investigations like this one. However, this effect has to be kept in mind for studies that want to investigate pause duration in the future. Pause duration was not part of the current investigation since it was concerned with isolated utterances. It should be assumed, though, that this misalignment might appear but largely go unnoticed when pauses occur. While the difference was only slight in this study, it was significant, so an effect on features like pause duration cannot be ruled out.

For mean intensity, Prediction 8 suggested that there is an intensity drop in the "internet methods" Zoom and Audacity compared to the Praat recordings (the Phone recordings were excluded here to account for the difference in microphone). This was not the case in the current data. Rather, the results show that Audacity had significantly higher mean intensity than Zoom and Praat, a result that was also fairly consistent across speakers. We assume that other technological processes play a role in this result, perhaps relating to default settings of Audacity or priorities in the recording computer, that we cannot address at this point, but would be interesting to investigate further in the future.

Regarding voice quality, Prediction 9 stated that HNR and the Hammarberg Index would not differ significantly between Audacity, Zoom, and Praat, but that the Phone recording would be different because it used a different headset. This is only partly what we found. On the *wh*-element, there was no effect of *Recording method*, but the particle *denn* showed higher HNR values in the Zoom and Audacity recordings than the other two methods, suggesting less breathy voice quality when transmitted and recorded via the internet. Additionally, male speakers overall had lower HNR values than female speakers, suggesting a breathier voice quality which is considered to be atypical: a breathy voice quality tends to be connected to lower perceptions of dominance and increased introversion ([39], [40]). For the Hammarberg Index, we found no *Gender* effects on either of the tested words. However, the results show higher values for Zoom and Audacity compared to Praat and Phone, suggesting increased expressivity ([29]). This was found for both tested words which seems to point to a consistent difference between the recording methods. It seems that Zoom's transmission increases a phonetic feature that is associated with expressiveness, or that some processes in Praat and Phone decrease that same feature. This cannot be solved or interpreted further at this point, but voice quality measurements with remote recordings – created with any method – should be carefully interpreted in the context of remote recording.

5 Conclusion

Overall, the global pandemic from 2020 and onwards pushed us as speech scientists to develop alternative recording methods that allow phonetic research outside of the laboratory. While research and travel are no longer severely affected by the pandemic, it is nevertheless always helpful to develop alternative research methods that can complement each other. This is the case in light of sustainability as well: reducing traveling to collect data could be supported by working virtually.

The proposed method here – Audacity, recorded locally on the experimenter's device via video conference in Zoom – offers a chance to record participants remotely, while the trials can be supervised and repetitions tracked, and the participant's computer screen can be used to present materials. Audacity appears to show stable results, especially for pitch-related measures and formants. The method is not without its problems. There are effects inherited from the noise cancellation of Zoom, like segmental reductions after long silences, or possibly increases of pause durations due to increased buffer times. Mean intensity proves to be problematic, and there are differences in voice quality between the methods as well.

While no remote recording method that exists to date can approximate the quality recorded in a laboratory in a soundbooth, one should not disregard recording remotely. Rather, remote recording should be framed as "digital fieldwork". Traditional fieldwork rarely has the luxury of recording in ideal conditions, so phonetic recordings made in the context of fieldwork also do not have the quality of lab recordings. The major source of influence on "digital fieldwork" is the hardware – the computer of the experimenter, of the participant, and the participant's headset (see also [6]). The upside – similar to traditional fieldwork – is that the participants are recorded in their own spaces which can increase comfort and naturalness. Therefore, remote recordings are not ideal, but as usable and relevant for phonetic research as traditional fieldwork recordings, so long as a study is appropriately framed.

Acknowledgments

We thank Oliver Niebuhr, Margaret Zellers, Kathrin Feindt, Martina Rossi, Nele Kiupel, and Felix Thiele for the valuable and important discussions and comments on previous drafts, and all our participants.

References

1. Zhang, C., Jepson, K., Lohfink, G., & Arvaniti, A. (2021). Comparing acoustic analyses of speech data collected remotely. *The Journal of the Acoustical Society of America*, 149(6), 3910-3916.
2. Grillo, E. U., Brosious, J. N., Sorrell, S. L., & Anand, S. (2016). Influence of smartphones and software on acoustic voice measures. *International journal of telerehabilitation*, 8(2), 9-14.
3. Manfredi, C., Lebacq, J., Cantarella, G., Schoentgen, J., Orlandi, S., Bandini, A., & DeJonckere, P. H. (2017). Smartphones offer new opportunities in clinical voice research. *Journal of Voice*, 31(1), 111-e1.
4. Uloza, V., Padervinskis, E., Vegiene, A., Pribuisiene, R., Saferis, V., Vaiciukynas, E., Gelzinis, A., & Verikas, A. (2015). Exploring the feasibility of smart phone microphone for measurement of acoustic voice parameters and voice pathology screening. *European Archives of Oto-rhino-laryngology*, 272(11), 3391-3399.
5. Vogel, A. P., Rosen, K. M., Morgan, A. T., & Reilly, S. (2014). Comparability of modern recording devices for speech analysis: smartphone, landline, laptop, and hard disc recorder. *Folia phoniatrica et logopaedica*, 66(6), 244-250.
6. Freeman, V., & De Decker, P. (2021). Remote sociophonetic data collection: Vowels and nasalization from self-recordings on personal devices. *Language and Linguistics Compass*, 15(7), e12435, 1211-1223.
7. Leemann, A., Jeszenszky, P., Steiner, C., Studerus, M., & Messerli, J. (2020). Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and videoconferencing. *Linguistics Vanguard*, *6*(s3), 1-16.
8. Boersma, P., & Weenink, D. 2018. *Praat: doing phonetics by computer*. Computer program.
9. Khalil, A. A., Elnaby, M. M. A., Saad, E. M., Al-Nahari, A. Y., Al-Zubi, N., El-Bendary, M. A., & El-Samie, F. E. A. (2014). Efficient speaker identification from speech transmitted over Bluetooth networks. *International Journal of Speech Technology*, 17(4), 409-416.
10. Al Bawab, Z., Locher, I., Xue, J., & Alwan, A. (2003). Speech recognition over bluetooth wireless channels. *Eighth European Conference on Speech Communication and Technology*, 1233-1236.
11. Siegert, I., & Niebuhr, O. (2021). Case report: Women, be aware that your vocal charisma can dwindle in remote meetings. *Frontiers in Communication*, 5, 611555.
12. Ge, C., Xiong, Y., & Mok, P. (2021). How Reliable Are Phonetic Data Collected Remotely? Comparison of Recording Devices and Environments on Acoustic Measurements. In *Interspeech*, 3984-3988.
13. Bulgin, J., De Decker, P., & Nycz, J. (2010). Reliability of formant measurements from lossy compressed audio. In: *British Association of Phoneticians Colloquium, March 29-31, 2010, London, UK*. (Unpublished)
14. Fuchs, R., & Maxwell, O. (2016). The effects of mp3 compression on acoustic measurements of fundamental frequency and pitch range. In *Speech Prosody 2016*, 523-527.
15. Jannetts, S., Schaeffler, F., Beck, J., & Cowen, S. (2019). Assessing voice health using smartphones: bias and random error of acoustic voice parameters captured by different smartphone types. *International journal of language & communication disorders*, 54(2), 292-305.
16. Maryn, Y., Morsomme, D., & De Bodt, M. (2017). Measuring the Dysphonia Severity Index (DSI) in the program praat. *Journal of Voice*, 31(5), 644-e29.
17. De Decker, P., & Nycz, J. (2011). For the record: Which digital media can be used for sociophonetic analysis?. *University of Pennsylvania Working Papers in Linguistics*, 17(2), 7.
18. Siegert, I., & Niebuhr, O. (2021). Speech signal compression deteriorates acoustic cues to perceived speaker charisma. *Elektronische Sprachsignalverarbeitung*, 1-10.
19. Zhang, C., Jepson, K., Lohfink, G., & Arvaniti, A. (2020). Speech data collection at a distance: Comparing the reliability of acoustic cues across homemade recordings. *The Journal of the Acoustical Society of America*, 148(4), 2717-2717.
20. Niebuhr, O., & Michaud, A. (2015). Speech data acquisition: the underestimated challenge. *KALIPHO-Kieler Arbeiten zur Linguistik und Phonetik*, 3, 1-42.
21. Ferrand, C. T. (2002). Harmonics-to-noise ratio: an index of vocal aging. *Journal of Voice*, 16(4), 480-487.
22. Audacity Team. (2019). *Audacity*, version 2.1.3. Computer program.
23. Zoom Video Communications Inc. (2021). *Zoom*. Computer program.

24. Neitsch, J. (2019). *Who cares about context and attitude?: Prosodic variation in the production and perception of rhetorical questions in German* (Doctoral dissertation).
25. Neitsch, J., & Niebuhr, O. (2019, August). Questions as prosodic configurations: How prosody and context shape the multiparametric acoustic nature of rhetorical questions in German. In *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia*, 2425-2429.
26. Neitsch, J., Barbosa, P. A., & Niebuhr, O. (2020, October). Prosody and Breathing: A Comparison Between Rhetorical and Information-Seeking Questions in German and Brazilian Portuguese. In *Interspeech*, 1863-1867.
27. Kisler, T., Reichel, U., & Schiel, F. 2017. Multilingual processing of speech via web services. *Computer Speech & Language,* 45, 326–347.
28. Hammarberg, B., Fritzell, B., Gaufin, J., Sundberg, J., & Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities. *Acta oto-laryngologica*, 90(1-6), 441-451.
29. Neitsch, J., & Niebuhr, O. (2020, October). Are Germans Better Haters Than Danes? Language-Specific Implicit Prosodies of Types of Hate Speech and How They Relate to Perceived Severity and Societal Rules. In *Interspeech*, 1843-1847.
30. Schmidt, J., Janse, E., & Scharenborg, O. (2016). Perception of emotion in conversational speech by younger and older listeners. *Frontiers in Psychology*, 7, 1-11.
31. Xu, Y. (2013). *ProsodyPro—A tool for large-scale systematic prosody analysis*. Laboratoire Parole et Langage, France.
32. Xu, Y., & Gao, H. (2018). FormantPro as a tool for speech analysis and segmentation. *Revista de Estudos da Linguagem*, 26(4), 1435-1454.
33. McCloy, D. R. (2012). Vowel normalization and plotting with the phonR package. *Technical Reports of the UW Linguistic Phonetics Laboratory*, 1, 1-8.
34. RStudio Team (2021). *RStudio: Integrated Development for R*. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.
35. Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics*. Cambridge, UK: Cambridge University Press.
36. Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390–412.
37. Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015). *Parsimonious mixed models*. Retrieved from https://arxiv.org/abs/1506.04967.
38. Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H. R., & Bates, D. M. (2017). Balancing type 1 error and power in linear mixed models. *Journal of Memory and Language*, 94, 305–315.
39. Moore, W. E. (1939). Personality traits and voice quality deficiencies. *Journal of Speech Disorders*, 4(1), 33-36.
40. Crystal, D. (1969). *Prosodic systems and intonation in English*. Cambridge: Cambridge University Press.