PROJECT MUSE

(Don't) try this at home! The effects of recording devices and software on phonetic analysis: Supplementary material



Chelsea Sanker, Sarah Babinski, Roslyn Burns, Marisha Evans, Jeremy Johns, Juhyae Kim, Slater Smith, Natalie Weber, Claire Bowern

Language, Volume 97, Number 4, December 2021, (Article)

Published by Linguistic Society of America
DOI: *https://doi.org/10.1353/lan.2021.0079*

For additional information about this article
https://muse.jhu.edu/article/842383/summary

[34.227.112.145] Project MUSE (2024-03-28 10:16 GMT)

Supplementary materials for '(Don't) try this at home! The effects of recording devices and software on phonetic analysis', by Chelsea Sanker, Sarah Babinski, Roslyn Burns, Marisha Evans, Jeremy Johns, Juhyae Kim, Slater Smith, Natalie Weber, and Claire Bowern.
Language 97(4).e360–e382, 2021.

Supplementary Materials: (Don't) try this at home! The effects of recording devices and software on phonetic analysis

CHELSEA SANKER, SARAH BABINSKI, ROSLYN BURNS, MARISHA EVANS, JEREMY JOHNS, JUHYAE KIM, SLATER SMITH, NATALIE WEBER, CLAIRE BOWERN

The supplementary materials presented here contain further information about the statistical models used to test each effect and further discussion of the results for individual measurements, including effects which did not reach significance. See Section 3 in the main text for an overview of the results. Some of the points that are included in the main text are also presented in the supplementary materials, in order to provide a complete presentation of the results here.

1. FURTHER INFORMATION ABOUT RECORDING DEVICES AND SOFTWARE

Table S1 below provides further information about the devices used in Phase 1 of recording. The numbers refer to the photograph of the setup in Figure 2 of the main text. For clarity, as in the main text, the Zoom H4n recorder will be referred to as H4n.

Number	Device	Specifications	Output
1	Zoom H4n	uncompressed, 44,100 Hz sampling rate, internal microphone; recorder is approximately 3 years old	wav
2	iPad	8th generation, iOS 14, on airplane mode, using VoiceMemos, internal microphone, 'compressed' setting	m4a
3	Macbook Pro	running OS 10.15 (Catalina), using internal microphone recording to Audacity, running PsychoPy to present the stimuli	wav
4	Macbook Pro	running OS 10.15 (Catalina), using external microphone recording to Audacity, recording with mid 2015 Audio Technica headset microphone using iXr external sound card	wav
5	Android phone	model LM-X320TA, running Android version 9, recording with the built-in application Audio Recorder (the settings do not give options for compression)	m4a
6	iPhone	iPhone 6s, iOS 14, recording with internal microphone using VoiceMemos, uncompressed format	m4a

TABLE S1. Specifications for recording devices used.

[34.227.112.145] Project MUSE (2024-03-28 10:16 GMT)

For the software conditions (Phase 2 of the recording), we chose to test software that, we believe, is commonly being used in remote field recordings. The software tested included Zoom, Skype, Facebook Messenger (recorded through Audacity, because it does not have an in-app recording option), the web-based podcast program Cleanfeed, and Audacity (without any virtual transmission, to distinguish between effects of Messenger and effects of Audacity). We chose only free recording programs. Since the settings of some of these software programs can vary substantially, we specify our recording setup below. All settings and program versions were up to date as of November 2020.

- **Zoom (v 5.3.1)**: We tested three configurations: remote recording versus locally recorded; in the remote condition, compressed versus 'original sound'[1] (without echo cancellation); and extracted from video versus audio only. The two remote recordings were done on a Mac and a Windows PC, with the former being set to 'original sound' and the latter recording with the default, compressed settings. The local recording was also done on a Mac. Files were output as wav (audio only) or mp4 (audio and video)
- **Skype:** We recorded the call using Skype's built-in recording feature that captures audio and video. The local recording was done on a Mac running 10.14, and the remote recording on a PC with Windows 10 (Skype v 8.65.0.78.). Files were saved as mp4.
- **Messenger/Audacity:** Facebook Messenger is a widely used application for linguistic fieldwork. Although there is no built-in recording system, we used Audacity (version 2.4.2) running in the background of the remote recorder's PC to record the call's audio. Audacity is widely used by fieldworkers as a way to record audio directly from a computer sound card (e.g. Mihas 2012, Johnson et al. 2018, Purnell et al. 2013). Files were saved in Audacity as uncompressed 16bit wav. To distinguish between effects of Messenger and effects of Audacity, a second condition used Audacity alone; as in the other condition, the sound card was treated as audio input to the Audacity program.
- **CleanFeed:** This is an online platform (https://cleanfeed.net/) that allows the user who initiates the call to manage the settings and make audio recordings. In our case, the 'remote' recorder (in the role of fieldworker) initiated and recorded the call, and this was done on a PC running Windows 10. Cleanfeed also has options of muting speakers and selecting which channel to record. Our settings were such that the remote recorder was muted and only the audio stream playing the stimuli was recorded. Files are saved as wav.

The recordings from CleanFeed and Messenger (through Audacity) did not include videos. Software such as Zoom and Skype provide the option to extract audio tracks, but given that a) the quality of the audio file is not altered by the presence or absence of video and b) remote fieldworkers may find videos useful to see certain articulatory features (such as rounding), facilitate general communication with the linguistic consultant, or for sign language research, we included video recording where possible. However, we did not further analyze the video recordings except to extract the audio signal. Similar issues raised in this paper for audio

[1] According to Zoom's settings, the 'original sound' option 'disables noise suppression, removes high pass filtering, and automatic gain control.'

fieldwork probably also apply to fieldwork with sign languages, particularly the horizontal compression identified in Section 2.5 below. See Lucas et al. 2013 for further discussions of sign language fieldwork, and Hou et al. 2020 on strategies for web-based sign language data collection and annotation.

Since consultants may use their phones during remote elicitation sessions, we also considered the inclusion of phone apps in our remote recording conditions (that is, where the audio signal is played through the cellphone or tablet and recorded remotely). However, logistical issues with recording and the already ballooning number of testing configurations led us to exclude this condition from Phase 2. For example, Facebook Messenger's mobile app also does not seem to allow recording apps to run in the background and record the call. Some other apps on iOS devices are allowed to run in the background while recording, but they use the Voice Memos app, which was already tested in our 'device' condition in Phase 1. Most crucially, our method of treating the H4n as external input and using it to play our recordings was unreliable on phones and tablets, where the external source did not reliably select the device as the microphone input. An external sound card would have perhaps allowed this, at the expense of testing the device audio itself.

Some additional comments about the file processing pipeline are in order. As briefly discussed in the main text, we converted all files to wav format and downsampled them to 16kHz for processing with the p2fa forced alignment algorithm. The effects of downsampling on digital audio files are well-known (cf. Johnson 2012). The only measurement for which downsampling is likely to affect our measurements is COG (center of gravity, a measurement used to characterize fricatives). We did find measurement differences in COG, but note that they are not due to the downsampling method; we would probably find even larger differences if we compared the non-downsampled recordings, due to their different sampling rates. Note that our aim in this experiment is **not** to compare recorded speech to live speech; rather, we are primarily comparing different forms of recorded speech to one another. Therefore, while for a research project where the aim is to represent speech as accurately as possible, we would probably use a higher sampling rate (as permitted by the recording device) and not downsample, in our case, we wish to treat the sound files as similarly to one another as possible, to be sure that any differences we see are due to the type of recording.

2. RESULTS

2.1. EFFECTS OF DEVICE. Here we present the full results for the effect of each device on the acoustic measurements. All statistical results are from mixed effects models calculated with the lme4 package in R (Bates et al. 2015). The p-values were calculated by the lmerTest package (Kuznetsova et al. 2015). The reference condition, which the other conditions were compared against, was always the H4n recorder. Because of a technical issue, one of the recordings for one of the speakers was lost, so the analyses of effects by device only include two speakers instead of three.

OVERALL DEVICE EFFECTS. This section presents results for the main measurements of each phonetic characteristic by device; the following section will examine interactions between device and phonological predictors.

Table S2 presents the summary of a linear mixed effects model for consonant duration (in ms) as predicted by the device. There was a random intercept for speaker.

Several of the conditions found significantly different consonant duration than the baseline H4n recorder, as is discussed in the main text. Possible sources of these differences in measured duration are discussed below.

	Estimate	SE	t-value	p
(Intercept)	116.6	2.5	45.8	< 0.001
Device Android	-2.8	3.6	-0.79	0.43
Device ExternalComputerMic	4.7	3.6	1.3	0.19
Device InternalComputerMic	-9.6	3.6	-2.7	0.008
Device iPad	-9.0	3.6	-2.5	0.012
Device iPhone	-4.9	3.6	-1.4	0.18

TABLE S2. Linear mixed-effects model for consonant duration (in milliseconds). *Reference level Program = H4n.*

Table S3 presents the summary of a linear mixed effects model for vowel duration (in ms) as predicted by the device. There was a random intercept for speaker.

Vowels were significantly longer than the baseline standard in the iPad condition. The differences in consonant duration seen above appear to be largely offset by the differences in vowel duration. That is, those conditions where the vowels are shorter are the same ones where the consonants are longer. Note, however, that the magnitude of the effects is overall quite small; less than 10 ms for most cases (which is the level of resolution of the forced aligner).

	Estimate	SE	t-value	p
(Intercept)	164.3	12.8	12.9	0.025
Device Android	4.3	7.0	0.61	0.54
Device ExternalComputerMic	-4.9	7.0	-0.7	0.48
Device InternalComputerMic	8.6	7.0	1.2	0.24
Device iPad	15.5	7.0	2.2	0.027
Device iPhone	6.1	7.0	0.87	0.39

TABLE S3. Linear mixed-effects model for vowel duration (in milliseconds). *Reference level Program = H4n.*

As noted in the main text and repeated here, there are two possible (not mutually exclusive) causes of segment differences. One is differences in boundary identification. In this case, properties of the digitization affect the performance of the forced aligner, such that segment boundaries are placed in different positions. A second source of difference is variation in the timing of segments which is introduced by compression. In this case, the segments do *actually* have different durations in the recording file (though not, of course, in the original speech). To illustrate the problem, consider the sets of alignments in Figure S1. The figure shows the spectrogram and two sets of alignments. The file is CS's speech recorded by Skype. The upper tier is the alignment as run on the actual file. The bottom tier is the alignment as run on the 'gold standard' recording. They begin close to identical (compare differences for the first phrase 'we

say *latch* again'), as this is the first utterance in the recording. However, the second phrase ('we say *sheep* again') shows a fairly consistent offset, starting from 'we', and the utterance clearly begins before the given boundary. This is most readily explained by compression affecting the length of the silence in the pause (marked by 'sp').

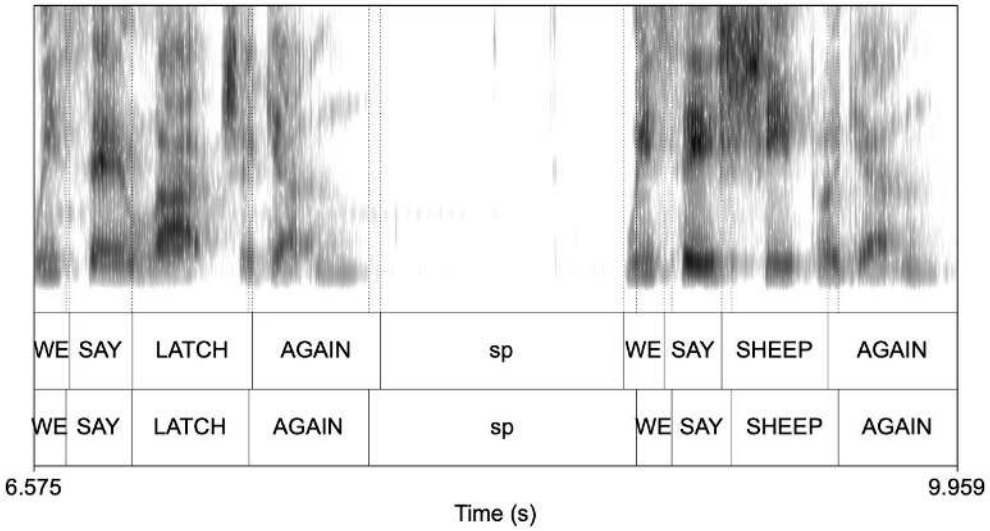


FIGURE S1. Comparison of word-level alignments from the Skype condition (top tier) and the gold standard H4n (bottom tier) for speaker CS. Audio from the Skype condition (= Figure 3 of main text).

It is likely that differences in measured duration *by device* mostly reflect differences in boundary identification rather than alterations to the actual timing of segments. However, as we discuss below in Section 2.5, the timing is directly influenced in some conditions, particularly in comparisons across programs. A lower signal-to-noise ratio makes boundaries more difficult to identify. This issue is not specific to the forced aligner. Humans also depend on segmentation cues that are obscured by low intensity or high background noise; indeed, automatic segmentation in such cases is likely to be preferable for comparisons, because segmentation biases will be consistent, while manual segmentation is likely to be more variable.

Figure S2 illustrates an item for which segmentation is notably different in different conditions, *tug* as produced by speaker CS. The final consonant /g/ has formant structure due to incomplete closure, which seems to result in it being segmented differently in the two conditions. In the baseline condition, the drop in intensity and lack of clear higher formants results in a relatively early boundary between the vowel and the final consonant. In the iPad condition, the divide is not so sharp, due to background noise, and the boundary between the vowel and final consonant is put much later. Note that in both recordings, the clear boundary between the initial consonant and the vowel is identified nearly identically.

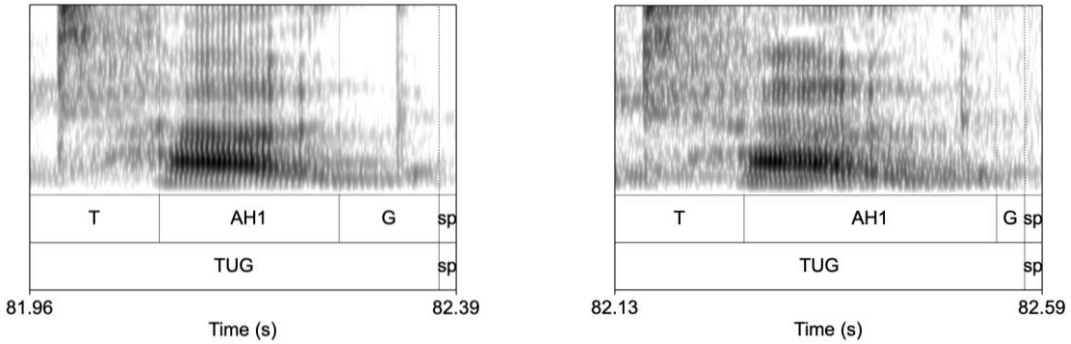


FIGURE S2. The word *tug* as produced by speaker CS and recorded by the H4n recorder (left) and iPad (right) (= Figure 4 of main text).

Table S4 presents the summary of a linear mixed effects model for mean f0 (in Hz) in vowels as predicted by the device. There was a random intercept for speaker.

There were no significant effects of Device on mean f0, though f0 was marginally lower in the iPad and iPhone conditions. In a larger dataset, the effect might reach significance. However, it is worth noting that the differences in measured f0 are small relative to the expected size of phonological f0 patterns.

	Estimate	SE	t-value	p
(Intercept)	180.4	5.1	35.1	0.011
Device Android	1.0	2.2	0.47	0.63
Device ExternalComputerMic	0.98	2.2	0.45	0.65
Device InternalComputerMic	-0.98	2.2	-0.45	0.65
Device iPad	-3.3	2.2	-1.5	0.13
Device iPhone	-3.6	2.2	-1.7	0.098

TABLE S4. Linear mixed-effects model for mean f0 (in Hz) in vowels. *Reference level Program = H4n.*

Figure S3 presents the distribution of f0 measurements for each speaker in each condition. Given the similar distributions across conditions, the different results are unlikely to be the result of pitch tracking errors; effects are not driven by a small number of major measurement differences. None of the conditions excluded more than 5 tokens as unmeasurable, so the results are also not the result of different exclusions. The differences might be related to the differing boundary assignments in each condition, as also reflected in the duration differences. Different boundaries could reduce extrinsic f0 effects of voicing of the neighboring consonants.

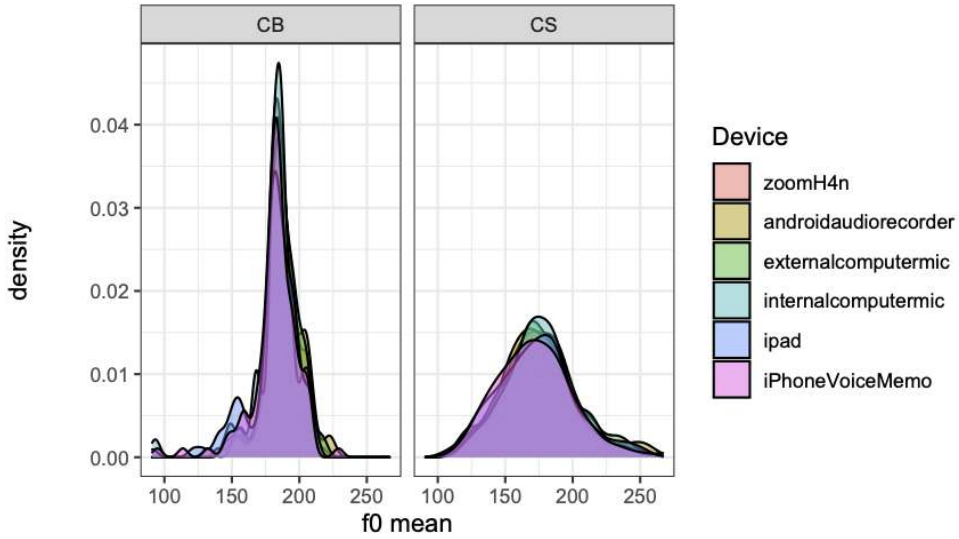


FIGURE S3. Density plots for the mean f0.

Table S5 presents the summary of a linear mixed-effects model for peak timing (in ms) -- the position of the maximum f0 relative to the beginning of the vowel, as predicted by the recording program. There was a random intercept for speaker.

There were no significant effects of Device on peak timing, but there were suggestive trends for Android and ExternalComputerMic which could be expected as a side effect of differences in vowel duration, because when the beginning of the vowel is put earlier, then the peak occurs later relative to that boundary. The size of the differences is small, though the differences are large enough relative to the size of actual peak timing effects that they could alter results. Many of the differences are due to how many items identify the peak f0 as occurring at the beginning of the vowel, which could be a result of the differences in the boundary identified for the beginning of the vowel.

	Estimate	SE	t-value	p
(Intercept)	27.5	3.2	8.6	< 0.001
Device Android	6.6	4.5	1.5	0.15
Device ExternalComputerMic	7.4	4.5	1.6	0.1
Device InternalComputerMic	-2.9	4.5	-0.64	0.53
Device iPad	-0.6	4.5	-0.13	0.9
Device iPhone	-4.4	4.5	-0.98	0.33

TABLE S5. Linear mixed-effects model for f0 peak timing (in milliseconds). *Reference level Program = H4n.*

Table S6 presents the summary of a linear mixed-effects model for jitter in vowels as predicted by the device, i.e. the cycle-to-cycle variation in f0. There was a random intercept for speaker.

There were no significant effects of Device on jitter measurements, which is consistent with the generally reliable f0 measurements.

[34.227.112.145] Project MUSE (2024-03-28 10:16 GMT)

	Estimate	SE	t-value	p
(Intercept)	0.021	0.0017	12.3	0.0025
Device Android	0.00031	0.0016	0.19	0.85
Device ExternalComputerMic	-0.0012	0.0016	-0.72	0.47
Device InternalComputerMic	-0.0009	0.0016	-0.56	0.58
Device iPad	0.0022	0.0016	1.3	0.18
Device iPhone	0.0019	0.0016	1.2	0.24

TABLE S6. Linear mixed-effects model for jitter in vowels. *Reference level Program = H4n.*

Table S7 presents the summary of a linear mixed-effects model for spectral tilt (H1-H2) in vowels as predicted by the device. There was a random intercept for speaker.

Spectral tilt was significantly lower in the Android condition than in the baseline H4n condition, and marginally higher in the iPhone condition. Even the differences that were not significant are rather large relative to the size of meaningful spectral tilt differences. The differences might indicate variation in how well the devices record higher and lower frequencies. The differences do not seem to be the result of distance from the speaker; the phones and the baseline H4n device were similarly close to the speaker, and the phones have opposite effects. (See Figure 2 in the main paper for a photograph that shows the physical arrangement of the recording devices.)

	Estimate	SE	t-value	p
(Intercept)	-2.0	2.0	-1.0	0.48
Device Android	-1.5	0.59	-2.5	0.013
Device ExternalComputerMic	-0.93	0.59	-1.6	0.11
Device InternalComputerMic	-0.57	0.59	-0.97	0.33
Device iPad	0.41	0.59	0.69	0.49
Device iPhone	1.0	0.59	1.7	0.084

TABLE S7. Linear mixed-effects model for spectral tilt in vowels. *Reference level Program = H4n.*

Table S8 presents the summary of a linear mixed-effects model for Harmonics-to-Noise Ratio (HNR) in vowels as predicted by the device. There was a random intercept for speaker.

HNR was significantly lower in the InternalComputerMic condition than in the baseline H4n condition, indicating more noise relative to the periodic components of the vowel in this condition than the baseline condition. This might reflect differences in sensitivity of the device's microphone to the periodic frequencies present in the signal. However, this result might also be due to distance from the speaker; this microphone was the furthest from the speaker. Impressionistically, internal computer microphones also pick up more noise from computer fans. See Section 2.4 below for measurements of signal-to-noise ratio.

	Estimate	SE	t-value	p
(Intercept)	6.4	1.3	4.9	0.12
Device Android	0.59	0.37	1.6	0.11
Device ExternalComputerMic	0.039	0.37	0.11	0.92
Device InternalComputerMic	-1.5	0.37	-4.2	<0.001
Device iPad	-0.34	0.37	-0.92	0.36
Device iPhone	-0.23	0.37	-0.63	0.53

TABLE S8. Linear mixed-effects model for HNR in vowels. *Reference level Program = H4n.*

Table S9 presents the summary of a linear mixed-effects model for F1 in vowels as predicted by the device. There was a random intercept for speaker and for vowel. All formant analyses used the measurements in Hertz. Lobanov normalization did not substantially change the results, so those analyses are not included here.

F1 was significantly lower in the InternalComputerMic, iPad, and iPhone conditions than in the baseline H4n condition. These results might be related to the trends found in spectral tilt measurements. As has been demonstrated previously, formant measurements are influenced by how the formants align with the harmonics (Chen et al. 2019). The effects of device on formant measurements vary by vowel; it is important to keep in mind that a lack of consistent overall effect across vowels does not mean that a device condition had no impact on formant measurements. Differences in how the formant measurements for each vowel are impacted by device are presented at the end of this section.

	Estimate	SE	t-value	p
(Intercept)	613.5	55.1	11.1	< 0.001
Device Android	-7.3	7.5	-0.98	0.33
Device ExternalComputerMic	-8.7	7.5	-1.2	0.24
Device InternalComputerMic	-19.8	7.5	-2.7	0.008
Device iPad	-15.2	7.5	-2.0	0.042
Device iPhone	-25.7	7.5	-3.4	< 0.001

TABLE S9. Linear mixed-effects model for F1 in vowels. *Reference level Program = H4n.*

Table S10 presents the summary of a linear mixed-effects model for F2 in vowels as predicted by the device. There was a random intercept for speaker and for vowel.

F2 was significantly lower than the baseline measurement in the InternalComputerMic condition, and significantly higher in the iPad and iPhone conditions. The results vary substantially for different vowels, as is presented at the end of this section. One of the major effects seems to be attributable to diphthongization of high and mid-high tense vowels, so failure to capture the full trajectory of the formants within the vowel results in altered estimation of the mean F2.

	Estimate	SE	t-value	p
(Intercept)	1897.5	115.7	16.4	< 0.001
Device Android	56.8	25.7	2.2	0.027
Device ExternalComputerMic	-33.5	25.7	-1.3	0.19
Device InternalComputerMic	-77.4	25.7	-3.0	0.0026
Device iPad	145.0	25.7	5.7	< 0.001
Device iPhone	70.6	25.7	2.7	0.0061

TABLE S10. Linear mixed-effects model for F2 in vowels. *Reference level Program = H4n.*

Table S11 presents the summary of a linear mixed-effects model for center of gravity (COG) in fricatives as predicted by the device. There was a random intercept for speaker and for segment.

The overall measurements were far higher in the ExternalComputerMic and InternalComputerMic conditions than in the baseline condition. This was largely due to the sibilants; COG measurements by fricative will be addressed near the end of this section. Measurements were also significantly higher in the Android condition.

	Estimate	SE	t-value	p
(Intercept)	2078.6	892.1	2.3	0.058
Device Android	440.3	132.5	3.3	0.00095
Device ExternalComputerMic	1172.5	132.5	8.9	< 0.001
Device InternalComputerMic	1115.2	132.5	8.4	< 0.001
Device iPad	-196.7	132.5	-1.5	0.14
Device iPhone	125.3	132.5	0.95	0.34

TABLE S11. Linear mixed-effects model for COG for fricatives. *Reference level Program = H4n.*

IMPACT ON CONTRASTS. Effects of device in measurements of these characteristics are primarily a concern if they alter our ability to find contrasts. In this section, we test whether measurements of acoustic correlates of phonological contrasts are altered by the recording device. These selected contrasts are known to exist in English and should be reflected by the measurements that we are using. When the regression models found no significant interaction between Device and the phonological categories, the results are illustrated just with a figure.

Figure S4 illustrates vowel duration as influenced by stress. The device did not have any substantial impact on these measurements. The effect of stress is significant or marginally significant in all conditions, and of a similar size.

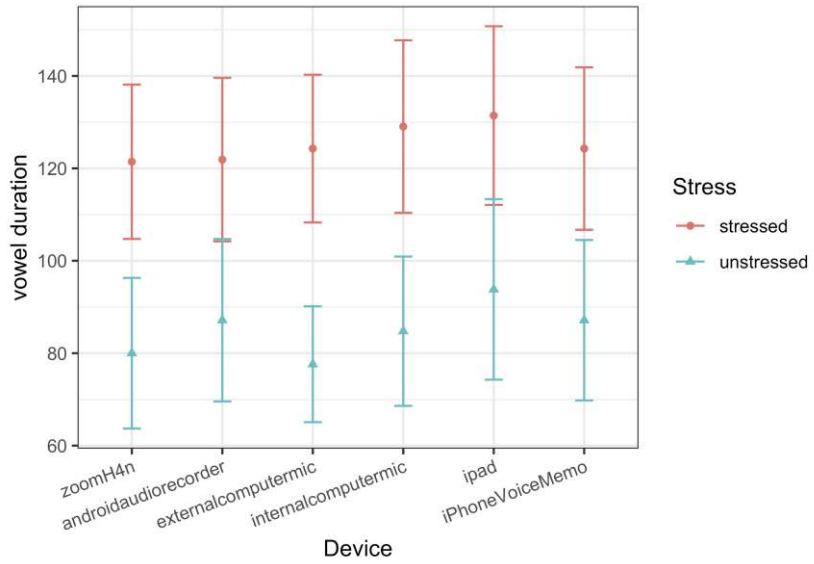


FIGURE S4. Measured vowel duration as predicted by device and stress. Pooled raw data, not the model results. Whiskers indicate the standard error (= Figure 5 of main text).

Figure S5 illustrates maximum f0 in vowels as influenced by stress. There are no substantial effects; all of the conditions find a significant effect, of a similar size.

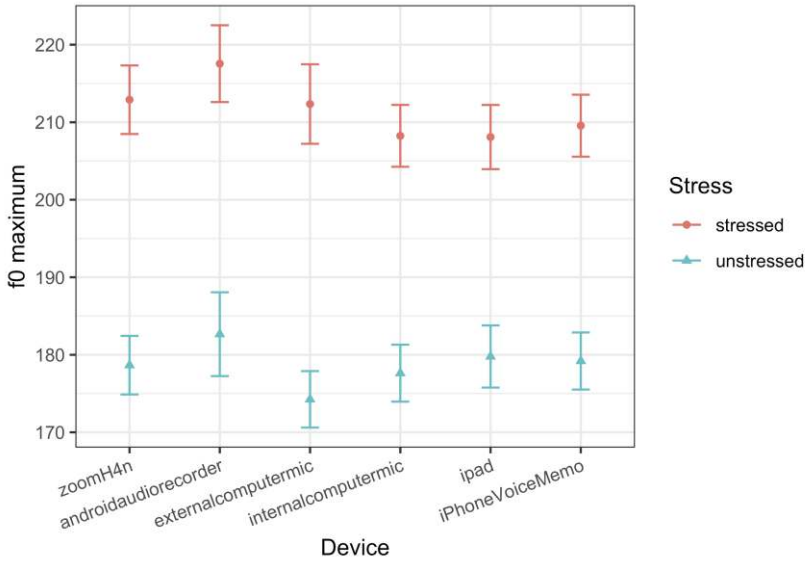


FIGURE S5. Measured F0 maximum as predicted by device and stress. Pooled raw data, not the model results. Whiskers indicate the standard error.

Figure S6 illustrates vowel duration as influenced by coda voicing. There are no substantial effects, although overall vowel duration differs across devices; all of the conditions find a significant effect, though some of them seem to be overestimating the effect, which could be a concern.

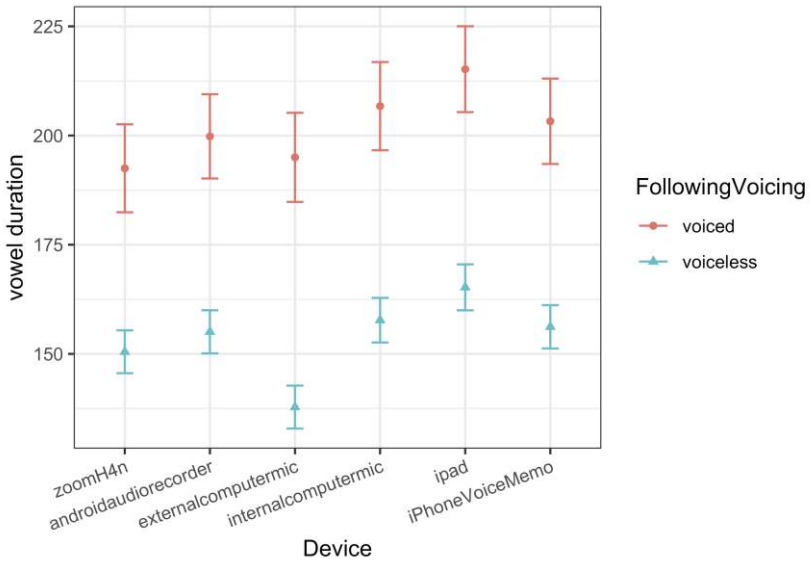


FIGURE S6. Measured vowel duration as predicted by device and coda voicing. Pooled raw data, not the model results. Whiskers indicate the standard error.

Table S12 presents the summary of a linear mixed-effects model for HNR in vowels as predicted by the device and the coda voicing. There was a random intercept for speaker.

Vowels followed by voiceless codas generally have a lower HNR than vowels before voiced codas. None of the interactions reach significance, though several of the conditions seem to be underestimating the size of the effect, which is consistent with those conditions overall having more noise and thus lower HNR. Figure S7 illustrates HNR in vowels as influenced by coda voicing.

	Estimate	SE	t-value	p
(Intercept)	8.7	1.2	7.5	0.047
Device Android	0.55	0.61	0.89	0.37
Device ExternalComputerMic	0.38	0.61	0.62	0.54
Device InternalComputerMic	-2.0	0.61	-3.3	0.0012
Device iPad	-0.9	0.61	-1.5	0.14
Device iPhone	-0.55	0.61	-0.9	0.37
FollowingVoicing Voiceless	-4.1	0.53	-7.8	< 0.001
Device Android:FollowingVoicing Voiceless	0.034	0.75	0.046	0.96
Device ExternalComputerMic:FollowingVoicing Voiceless	-0.66	0.75	-0.89	0.38
Device InternalComputerMic:FollowingVoicing Voiceless	0.61	0.75	0.82	0.41
Device iPad:FollowingVoicing Voiceless	0.83	0.75	1.1	0.27
Device iPhone:FollowingVoicing Voiceless	0.43	0.75	0.58	0.56

TABLE S12. Linear mixed-effects model for HNR in vowels, with coda voicing as a factor. *Reference level Program = H4n, FollowingVoicing = voiced.*

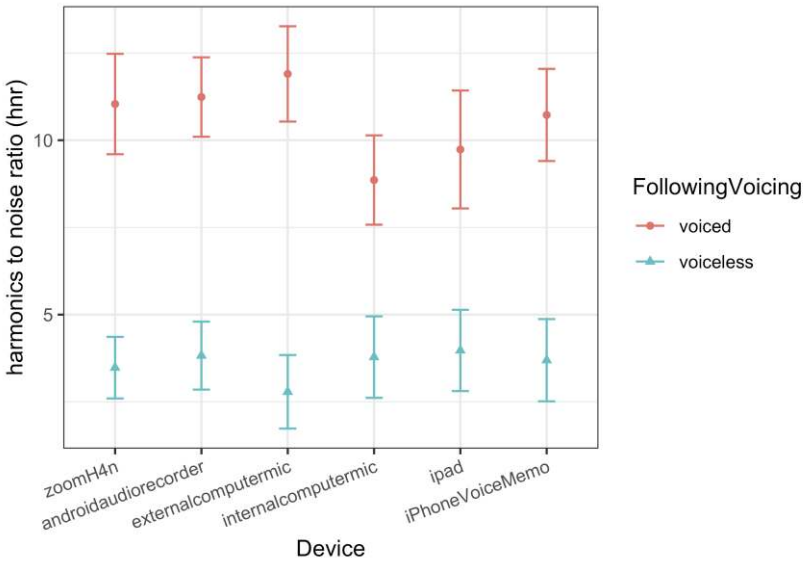


FIGURE S7. Measured HNR as predicted by device and coda voicing. Pooled raw data, not the model results. Whiskers indicate the standard error.

Table S13 presents the summary of a linear mixed-effects model for HNR in vowels as predicted by the device and the onset voicing. There was a random intercept for speaker.

HNR differences between voiced and voiceless onsets are slightly decreased for the InternalComputerMic, iPad and iPhone devices, which is a concern primarily because the difference is small even in the baseline condition. Differences may be due to boundary assignment, as modal voicing is a cue used to identify vowels. Figure S8 illustrates HNR in vowels as influenced by onset voicing.

	Estimate	SE	t-value	p
(Intercept)	6.9	1.38	5.0	0.086
Device Android	0.7	0.68	1.0	0.3
Device ExternalComputerMic	0.18	0.68	0.27	0.79
Device InternalComputerMic	-1.8	0.68	-2.6	0.0096
Device iPad	-0.54	0.68	-0.79	0.43
Device iPhone	-0.47	0.68	-0.69	0.49
PrecedingVoicing Voiceless	-1.3	0.62	-2.2	0.033
Device Android:PrecedingVoicing Voiceless	-0.2	0.88	-0.23	0.82
Device ExternalComputerMic:PrecedingVoicing Voiceless	-0.35	0.88	-0.4	0.69
Device InternalComputerMic:PrecedingVoicing Voiceless	0.53	0.88	0.6	0.55
Device iPad:PrecedingVoicing Voiceless	0.35	0.88	0.4	0.69
Device iPhone:PrecedingVoicing Voiceless	0.41	0.88	0.47	0.64

TABLE S13. Linear mixed-effects model for HNR in vowels, with onset voicing as a factor. *Reference level Program = H4n, PrecedingVoicing = voiced.*

[34.227.112.145] Project MUSE (2024-03-28 10:16 GMT)

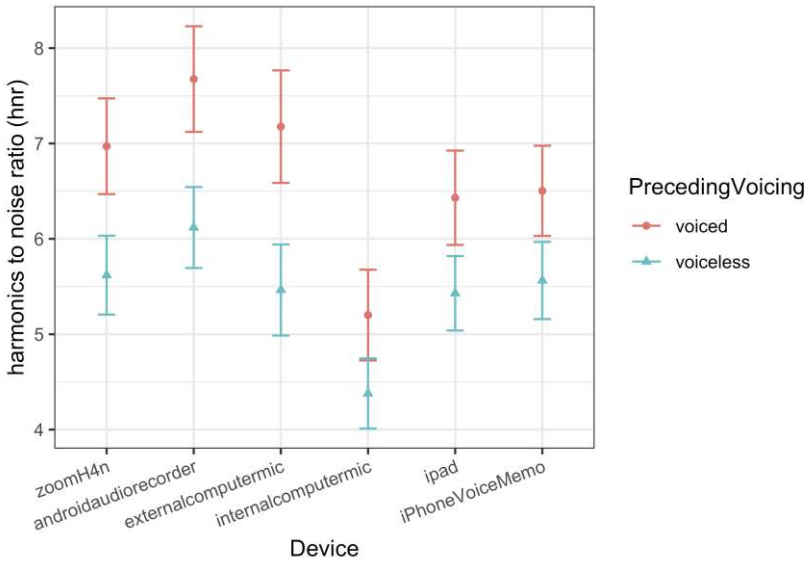


FIGURE S8. Measured HNR as predicted by device and onset voicing. Pooled raw data, not the model results. Whiskers indicate the standard error.

Table S14 presents the summary of a linear mixed-effects model for spectral tilt in vowels as predicted by the device and the onset voicing. There was a random intercept for speaker.

The effect is only marginally significant in the baseline condition; it is only a small effect, but has been established elsewhere (e.g. Kong et al. 2012). Though most of the differences are not significant, it is important to note that they are large relative to the size of the actual effect; there are clear distortions of spectral tilt, which are likely to obscure measurements. Figure S9 illustrates spectral tilt in vowels as influenced by onset voicing.

	Estimate	SE	t-value	p
(Intercept)	-2.4	2.2	-1.1	0.42
Device Android	-1.3	1.1	-1.1	0.25
Device ExternalComputerMic	-1.3	1.1	-1.2	0.25
Device InternalComputerMic	-0.42	1.1	-0.38	0.7
Device iPad	0.23	1.1	0.2	0.83
Device iPhone	0.86	1.1	0.78	0.43
PrecedingVoicing Voiceless	1.8	1.0	1.8	0.074
Device Android:PrecedingVoicing Voiceless	-0.98	1.4	-0.7	0.49
Device ExternalComputerMic:PrecedingVoicing Voiceless	0.14	1.4	0.1	0.92
Device InternalComputerMic:PrecedingVoicing Voiceless	-0.97	1.4	-0.69	0.49
Device iPad:PrecedingVoicing Voiceless	-0.21	1.4	-0.15	0.88
Device iPhone:PrecedingVoicing Voiceless	-0.24	1.4	-0.17	0.86

TABLE S14. Linear mixed-effects model for spectral tilt, with onset voicing as a factor. *Reference level Program = H4n, PrecedingVoicing = voiced.*

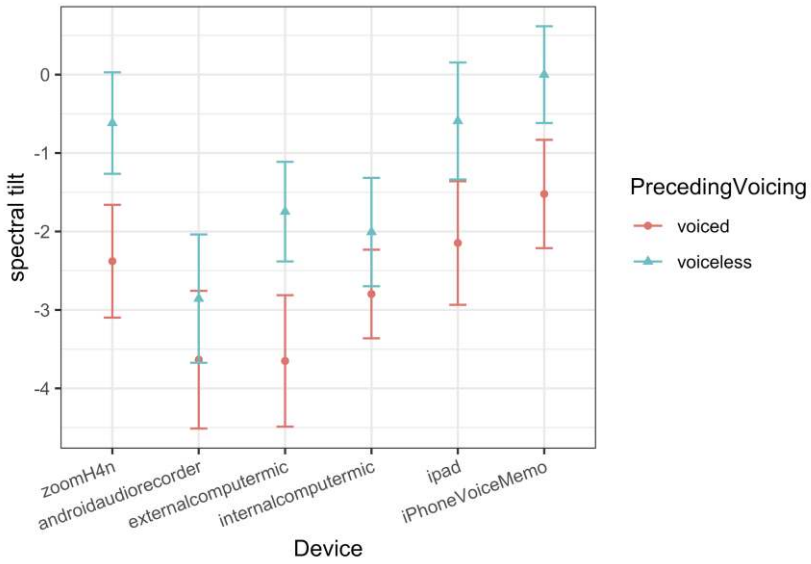


FIGURE S9. Measured spectral tilt as predicted by device and onset voicing. Pooled raw data, not the model results. Whiskers indicate the standard error.

Figure S10 illustrates maximum f0 in vowels as influenced by onset voicing. None of the effects are significant, but there is variation in how large the effect is estimated to be.

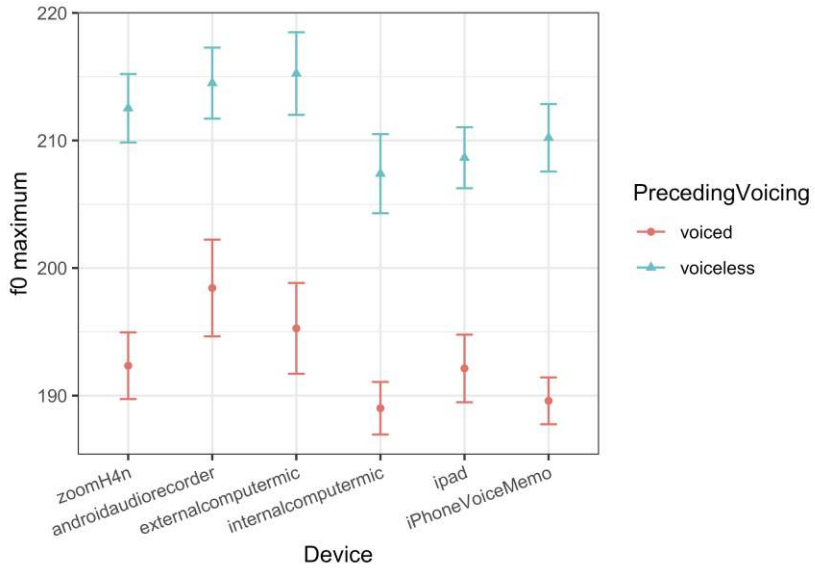


FIGURE S10. Measured f0 maximum as predicted by device and onset voicing. Pooled raw data, not the model results. Whiskers indicate the standard error.

Table S15 presents the summary of a linear mixed-effects model for COG in /s/ vs. /ʃ/ as predicted by the device. There was a random intercept for speaker.

	Estimate	SE	t-value	p
(Intercept)	5053.7	559.0	9.0	0.058
Device Android	632.7	181.6	3.5	0.00058
Device ExternalComputerMic	1723.1	181.6	9.5	< 0.001
Device InternalComputerMic	1359.7	181.6	7.5	< 0.001
Device iPad	-460.4	181.6	-2.5	0.011810
Device iPhone	226.6	181.6	1.2	0.21
Segment /ʃ/	-1689.3	254.2	-6.6	< 0.001
Device Android:Segment /ʃ/	-496.5	359.4	-1.4	0.17
Device ExternalComputerMic:Segment /ʃ/	-1587.6	359.4	-4.4	< 0.001
Device InternalComputerMic:Segment /ʃ/	-763.1	359.4	-2.1	0.035
Device iPad:Segment /ʃ/	301.0	359.4	0.84	0.4
Device iPhone:Segment /ʃ/	-253.8	359.4	-0.71	0.48

TABLE S15. Linear mixed-effects model for COG in sibilant fricatives, with particular fricative as a factor. *Reference level Program = H4n, Segment = /s/.*

The model finds the same effect noted above for overall COG measurements: The COG for /s/ is overestimated in the ExternalComputerMic condition and the InternalComputerMic condition. The interactions show that /ʃ/ is not as affected. Figure S11 illustrates COG by fricative. These results seem to be a combination of how well the microphones pick up low-frequency noise and how much background noise they pick up. The effects of this problem would likely be smaller for recordings with a higher sampling rate.

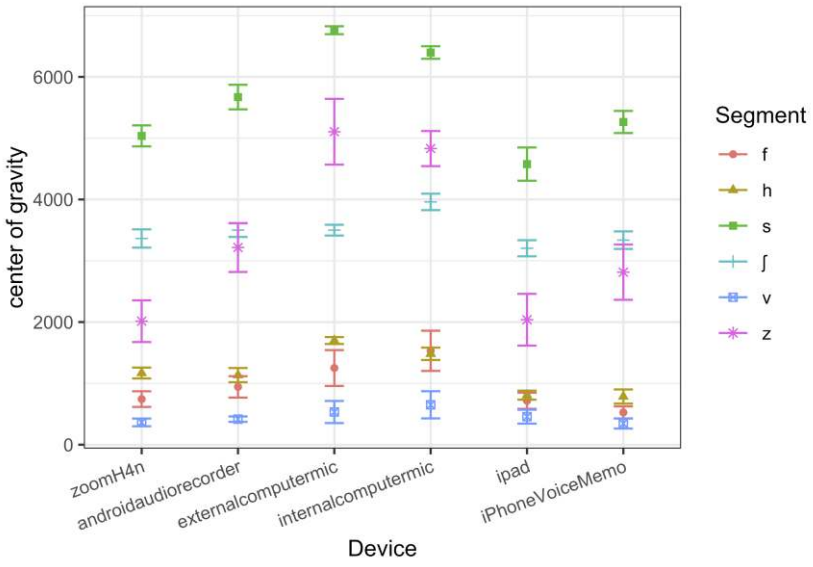


FIGURE S11. Measured center of gravity as predicted by device and segment, among fricatives. Pooled raw data, not the model results. Whiskers indicate the standard error.

Figure S12 illustrates F1 and F2 as influenced by vowel quality and device. Adding the interaction between vowel quality and device marginally improves the model for F1 (χ^2 = 105.0, df = 84, p = 0.06). The interaction between vowel quality and device significantly improves the model for F2 (χ^2 = 186.4, df = 84, p < 0.0001); the measured F2 varies considerably across conditions for some vowels.

The device conditions in Phase 1 all clearly pick out a recognizable vowel space. However, some of the vowels are shifted enough that they would be likely to cause problems for analysis. In particular, F2 measurements for /u/ and /ou/ were very high in many of the conditions; this is in part due to issues in identifying boundaries or tracking low-intensity formants, which altered which part of the diphthong were measured. Many of the words with /u/ lacked codas, so failure to capture the back portion of the offglide of the vowel resulted in only measuring the fronter beginning portion. While other vowels did not all exhibit systematic effects, there are several vowels that have strikingly variable measurements across conditions.

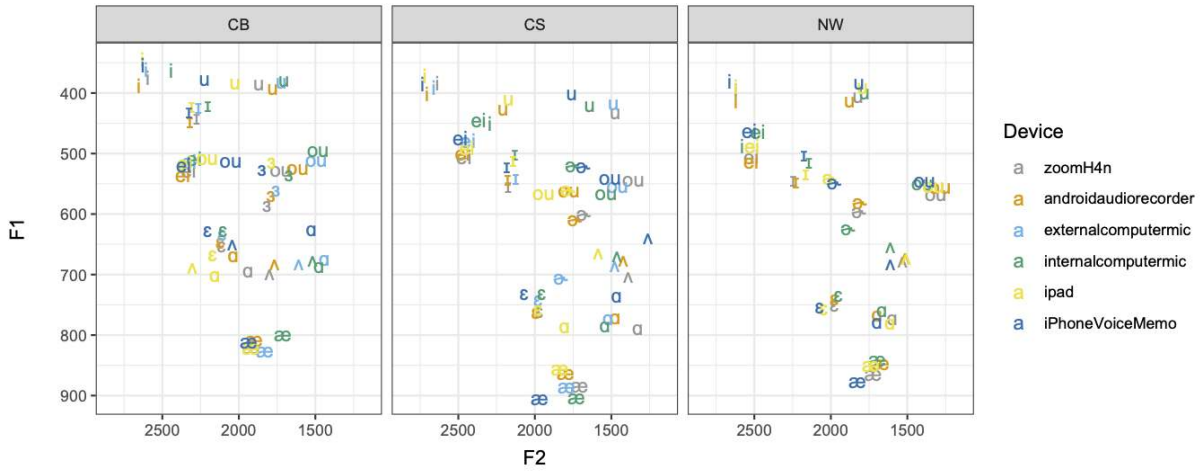


FIGURE S12. Vowel spaces for each speaker as measured in Phase 1, by-Device (= Figure 7 of main text).

2.2 EFFECTS OF PROGRAM. Here we present the full results for the effect of each program on the acoustic measurements. There were five software conditions compared to the H4n reference condition: Zoom, Skype, Cleanfeed, Facebook Messenger (recorded through Audacity, because it does not have an in-app recording option), and AudacityAlone. Note that four of these are testing applications for online transmission, while AudacityAlone is present to test whether the Audacity program causes effects in itself, to clarify how to interpret the results of the Messenger condition.

For these comparisons, we used a single Zoom condition, recording locally with the default audio settings. Although we tested several different Zoom conditions, there were no differences between any of them: Local vs. remote, operating system, conversion from mp4, or the 'Original Audio' setting. None of the characteristics measured exhibited significant effects of recording condition. The models comparing Zoom conditions to each other are presented in Section 2.3 below.

OVERALL EFFECTS. This section presents results for the main measurements of each phonetic characteristic by program; the following section will examine interactions between program and phonological predictors.

Table S16 presents the summary of a linear mixed-effects model for consonant duration as predicted by the recording program. There was a random intercept for speaker.

There were no significant consonant duration differences between the baseline recording and the recording made through Cleanfeed or Audacity alone. However, consonant durations were significantly shorter in all of the other conditions. Some of the effects on duration may be due to differences in intensity or background noise, which could alter the boundaries identified by forced alignment and would also be likely to produce similar effects in manual alignment, as discussed in Section 2.1 above. Some of the differences in duration might also reflect actual duration differences created by compression algorithms; see Section 2.5 for more discussion.

	Estimate	SE	t-value	p
(Intercept)	106.8	10.3	10.4	0.0072
Program AudacityAlone	1.1	2.9	0.38	0.7
Program Cleanfeed	-0.44	2.9	-0.15	0.88
Program Messenger	-11.6	2.9	-4.0	< 0.001
Program Skype	-8.5	2.9	-3.0	0.003
Program Zoom	-11.2	2.9	-3.9	< 0.001

TABLE S16. Linear mixed-effects model for consonant duration (in milliseconds). *Reference level Program = H4n.*

Table S17 presents the summary of a linear mixed-effects model for vowel duration as predicted by the recording program. There was a random intercept for speaker.

As for consonant duration, there were no significant vowel duration differences between the baseline recording and the recording made through Cleanfeed or AudacityAlone. However, vowel durations were significantly longer in all of the other conditions.

	Estimate	SE	t-value	p
(Intercept)	157.2	12.3	12.8	0.0025
Program AudacityAlone	-0.84	6.0	-0.14	0.89
Program Cleanfeed	0.37	6.0	0.061	0.95
Program Messenger	17.5	6.0	2.9	0.0039
Program Skype	19.8	6.0	3.3	0.0011
Program Zoom	31.5	6.0	5.2	< 0.001

TABLE S17. Linear mixed-effects model for vowel duration (in milliseconds). *Reference level Program = H4n.*

Table S18 presents the summary of a linear mixed-effects model for the mean f0 in vowels, as predicted by the recording program. There was a random intercept for speaker.

There was no significant effect on mean f0; none of the conditions differed significantly from the baseline H4n recorder.

	Estimate	SE	t-value	p
(Intercept)	181.1	3.8	48.2	< 0.001
Program AudacityAlone	0.17	1.5	0.12	0.91
Program Cleanfeed	-0.14	1.5	-0.095	0.92
Program Messenger	-1.3	1.5	-0.87	0.38
Program Skype	0.33	1.5	0.23	0.82
Program Zoom	0.63	1.5	0.43	0.67

TABLE S18. Linear mixed-effects model for mean f0 (in Hz) in vowels. *Reference level Program = H4n.*

Table S19 presents the summary of a linear mixed-effects model for peak timing -- the position of the maximum f0 relative to the beginning of the vowel, as predicted by the recording program. There was a random intercept for speaker.

The f0 peak timing was significantly later for Zoom than the baseline H4n condition. This result is probably related to the overestimated vowel duration in the Zoom condition, as described above. Because the beginnings of the vowels were placed earlier, the peak f0 was later relative to that starting point. However, it is worth considering why none of the other conditions have effects on peak timing, when several of them did have duration effects. The different results might be due to the size of the duration effect; Messenger and Skype had smaller effects on duration than Zoom did, so the corresponding differences in peak timing are smaller and do not reach significance.

	Estimate	SE	t-value	p
(Intercept)	33.7	6.5	5.2	0.014
Program AudacityAlone	-0.93	4.3	-0.21	0.83
Program Cleanfeed	-0.073	4.3	-0.017	0.99
Program Messenger	6.5	4.3	1.5	0.13
Program Skype	5.0	4.3	1.1	0.25
Program Zoom	14.2	4.3	3.3	0.001

TABLE S19. Linear mixed-effects model for f0 peak timing (in milliseconds). *Reference level Program = H4n.*

Table S20 presents the summary of a linear mixed-effects model for jitter in vowels, as predicted by the recording program. There was a random intercept for speaker.

There was no significant effect of recording condition on measurements of jitter, though there was a marginal effect of the Zoom condition, finding more jitter than the H4n recorder. That is, there was more cycle-to-cycle variation in f0 as measured in the Zoom recording than in the baseline condition.

[34.227.112.145] Project MUSE (2024-03-28 10:16 GMT)

	Estimate	SE	t-value	p
(Intercept)	0.019	0.0033	5.6	0.023
Program AudacityAlone	0.00075	0.0012	0.63	0.53
Program Cleanfeed	0.0011	0.0012	0.92	0.36
Program Messenger	0.0008	0.0012	0.67	0.5
Program Skype	0.00039	0.0012	0.32	0.75
Program Zoom	0.0023	0.0012	1.9	0.059

TABLE S20. Linear mixed-effects model for jitter in vowels. *Reference level Program = H4n.*

Table S21 presents the summary of a linear mixed-effects model for spectral tilt (H1-H2) in vowels, as predicted by the recording program. There was a random intercept for speaker.

All of the programs exhibited effects of spectral tilt. Most of them underestimated spectral tilt, while Messenger overestimated it. The effects suggest that transmission for many of these programs is worse for lower frequencies than for higher frequencies. Notably, this effect is even present in the AudacityAlone condition. On the other hand, the higher spectral tilt in the Messenger condition might suggest that Messenger is amplifying low frequencies, in addition to the effects of Audacity making the recording for the Messenger condition.

	Estimate	SE	t-value	p
(Intercept)	-1.6	1.6	-1.0	0.4
Program AudacityAlone	-1.4	0.5	-2.9	0.0041
Program Cleanfeed	-1.3	0.5	-2.6	0.009
Program Messenger	4.6	0.5	9.1	< 0.001
Program Skype	-1.7	0.5	-3.3	< 0.001
Program Zoom	-2.0	0.5	-3.9	< 0.001

TABLE S21. Linear mixed-effects model for spectral tilt in vowels. *Reference level Program = H4n.*

Table S22 presents the summary of a linear mixed-effects model for the Harmonics-to-Noise Ratio (HNR) in vowels, as predicted by the recording program. There was a random intercept for speaker.

Messenger exhibited a much higher HNR than the baseline H4n condition. No other effects were significant, but they all have the trend towards being lower than the baseline, suggesting more noise. The much higher value for the Messenger condition is likely to have a different explanation; it is unlikely that this condition was capturing the periodic signal more reliably than the source recording. This result is not due to excluding unmeasurable items; no conditions excluded more than 3 tokens. The effect might come from amplification of low frequencies, as also probably underlies some of the effects in spectral tilt; low frequencies include the clearest harmonics, so if these frequencies are amplified, the HNR would appear to be higher.

	Estimate	SE	t-value	p
(Intercept)	7.3	1.0	7.0	0.016
Program AudacityAlone	-0.34	0.29	-1.2	0.24
Program Cleanfeed	-0.24	0.29	-0.84	0.4
Program Messenger	1.2	0.29	4.2	< 0.001
Program Skype	-0.3	0.29	-1.0	0.3
Program Zoom	-0.4	0.29	-1.4	0.17

TABLE S22. Linear mixed-effects model for HNR in vowels. *Reference level Program = H4n.*

Table S23 presents the summary of a linear mixed-effects model for F1, as predicted by the recording program. There was a random intercept for speaker and for vowel.

F1 was significantly lower in the Messenger condition than in the baseline condition. The cause of this effect is not entirely clear. A discussion of formant effects separated by vowel is presented at the end of this section, and offers more detail about possible sources of differences in formant measurements.

	Estimate	SE	t-value	p
(Intercept)	613.7	48.4	12.7	< 0.001
Program AudacityAlone	1.5	8.8	0.17	0.87
Program Cleanfeed	10.8	7.9	1.4	0.17
Program Messenger	-29.7	7.9	-3.8	0.00018
Program Skype	-3.5	7.9	-0.45	0.66
Program Zoom	-11.1	7.9	-1.4	0.16

TABLE S23. Linear mixed-effects model for F1 in vowels. *Reference level Program = H4n.*

Table S24 presents the summary of a linear mixed-effects model for F2, as predicted by the recording program. There was a random intercept for speaker and for vowel.

F2 was overestimated in all of the conditions, to varying degrees; the largest effect was in the Messenger condition. The end of this section addresses formant effects in more detail, separated by vowel.

	Estimate	SE	t-value	p
(Intercept)	1898.4	119.6	15.9	< 0.001
Program AudacityAlone	36.1	21.2	1.7	0.088
Program Cleanfeed	46.0	19.0	2.4	0.016
Program Messenger	91.0	19.0	4.8	< 0.001
Program Skype	42.0	19.0	2.2	0.027
Program Zoom	31.4	19.0	1.7	0.099

TABLE S24. Linear mixed-effects model for F2 in vowels. *Reference level Program = H4n.*

Table S25 presents the summary of a linear mixed-effects model for center of gravity (COG) in fricatives, as predicted by the recording program. There was a random intercept for speaker and for segment.

COG was significantly lower in the Cleanfeed and Messenger conditions, and marginally higher in the Zoom condition. As in the Device comparisons, the largest effects are on /s/ and /z/. Further analysis of differences between fricatives are presented at the end of this section, along with a discussion of possible sources of these differences.

	Estimate	SE	t-value	p
(Intercept)	1923.9	549.3	3.5	0.0094
Program AudacityAlone	220.6	140.6	1.6	0.12
Program Cleanfeed	-653.3	126.1	-5.2	< 0.001
Program Messenger	-904.1	126.1	-7.2	< 0.001
Program Skype	-196.3	126.1	-1.6	0.12
Program Zoom	220.7	126.1	1.7	0.08

TABLE S25. Linear mixed-effects model for COG for fricatives. *Reference level Program = H4n.*

IMPACT ON CONTRASTS. As noted for the comparisons by device, effects in these characteristics are primarily a concern if they alter our ability to find contrasts. In this section, we test whether contrasts depending on these characteristics are altered by the recording device.

Figure S13 illustrates vowel duration as influenced by stress. The effect of stress on duration is significant or marginally significant in all program conditions, and of a similar size.

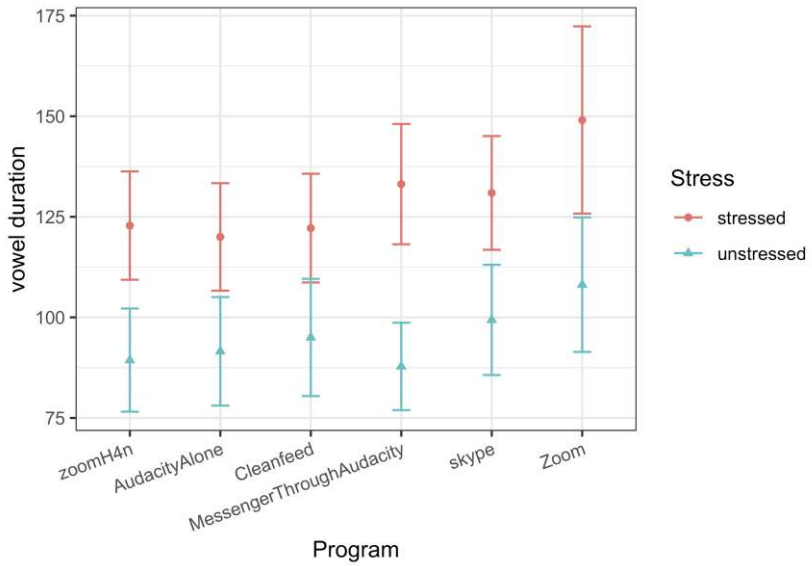


FIGURE S13. Measured vowel duration as predicted by program and stress. Pooled raw data, not the model results. Whiskers indicate the standard error (= Figure 6 of main text).

Figure S14 illustrates F0 in vowels as influenced by stress. The program did not have any substantial impact on these measurements; there was a clear separation between stressed and unstressed vowels in all conditions, though there is some variation in the size of the effect.

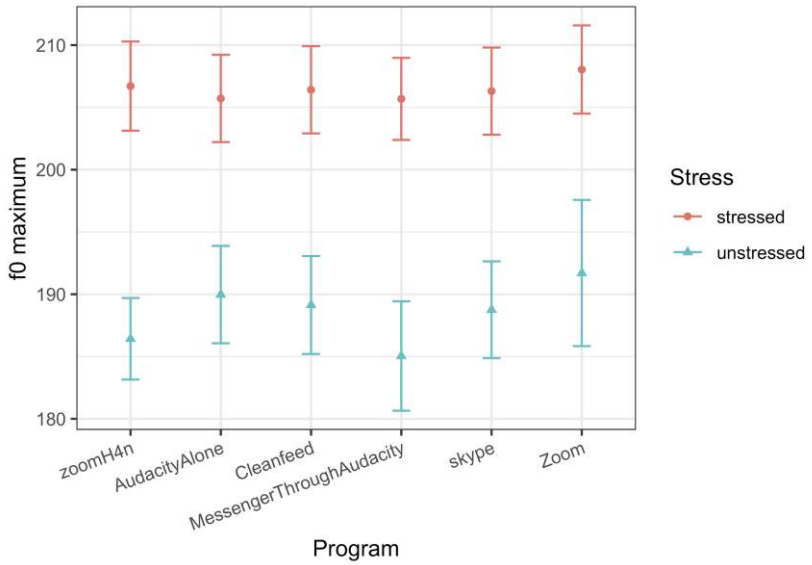


FIGURE S14. Measured f0 maximum as predicted by program and stress. Pooled raw data, not the model results. Whiskers indicate the standard error.

Figure S15 illustrates vowel duration as influenced by coda voicing. The program did not have any substantial impact on these measurements, though there was variation in the size of the effect, and some conditions were substantially overestimating overall vowel duration.

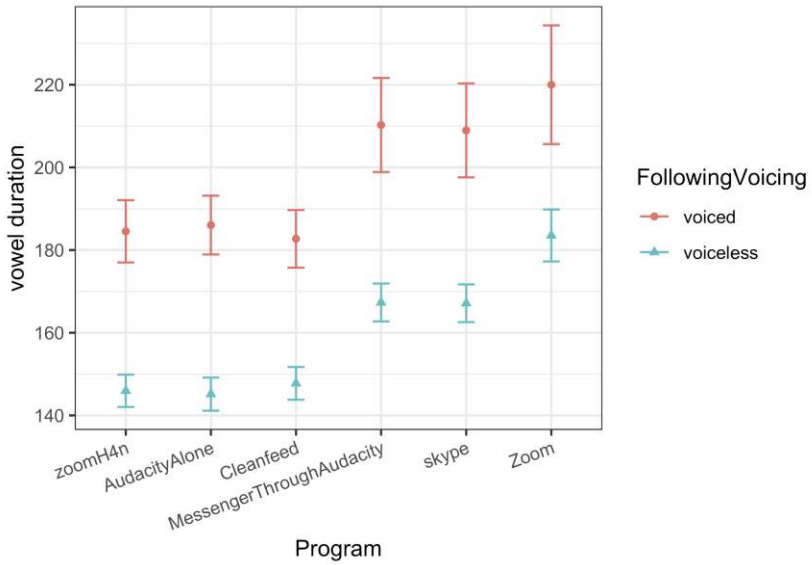


FIGURE S15. Measured vowel duration as predicted by program and coda voicing. Pooled raw data, not the model results. Whiskers indicate the standard error.

Figure S16 illustrates HNR in vowels as influenced by coda voicing. The program did not have any substantial impact on these measurements; the effect was a similar size in all conditions.

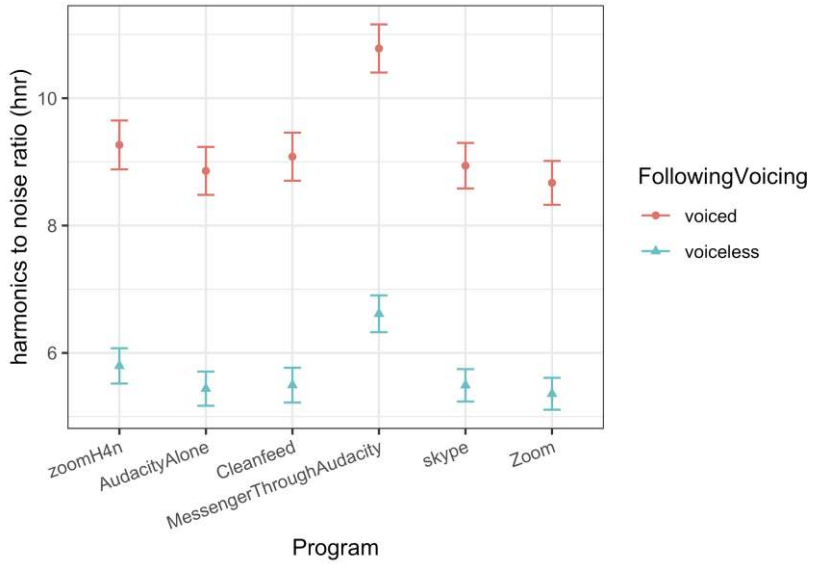


FIGURE S16. Measured HNR as predicted by program and coda voicing. Pooled raw data, not the model results. Whiskers indicate the standard error.

Figure S17 illustrates HNR in vowels as influenced by onset voicing. The program did not have any substantial impact on these measurements; the effect was a similar size in all conditions, even though Messenger substantially overestimated HNR for vowels in both environments.

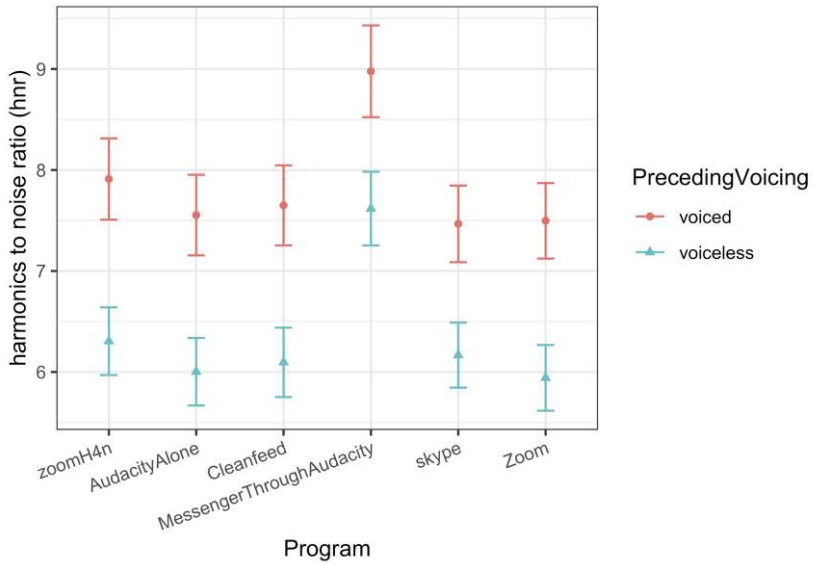


FIGURE S17. Measured HNR as predicted by program and onset voicing. Pooled raw data, not the model results. Whiskers indicate the standard error.

Figure S18 illustrates spectral tilt in vowels as influenced by onset voicing. The program did not have any substantial impact on these measurements; the effect was a similar size in all

conditions, even though Messenger substantially overestimated spectral tilt for vowels in both environments.

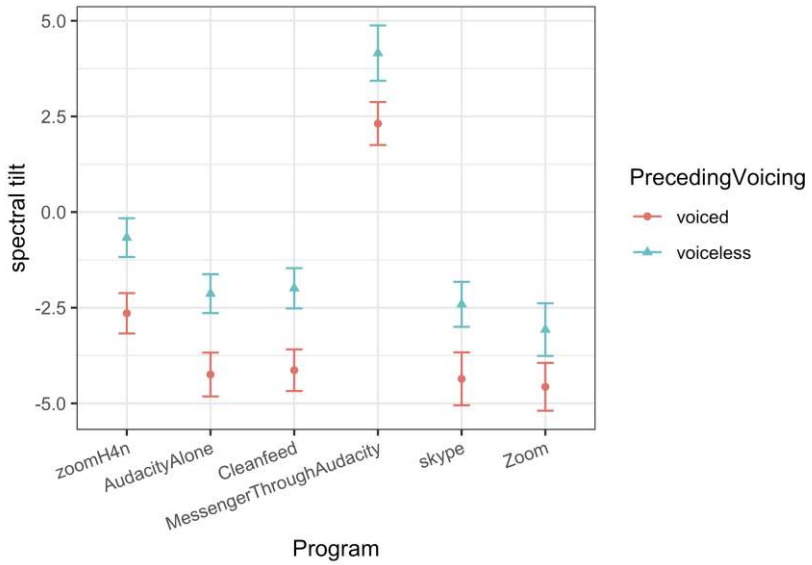


FIGURE S18. Measured spectral tilt as predicted by program and onset voicing. Pooled raw data, not the model results. Whiskers indicate the standard error.

Figure S19 illustrates maximum F0 as influenced by onset voicing. The program did not have any substantial impact on these measurements. The effect of onset voicing was significant and of a similar size in all conditions.

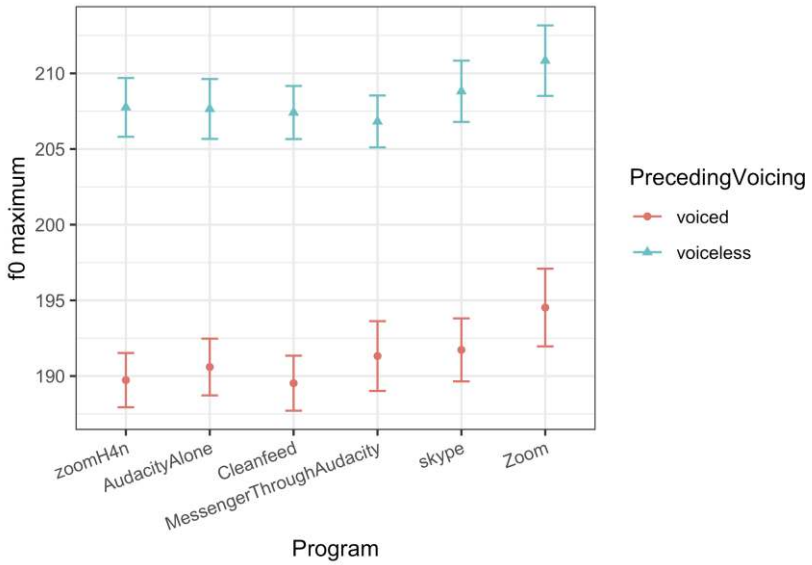


FIGURE S19. Measured f0 maximum as predicted by program and onset voicing. Pooled raw data, not the model results. Whiskers indicate the standard error.

Table S26 presents the summary of a linear mixed-effects model for COG in /s/ and /ʃ/, as predicted by the recording program and segment. There was a random intercept for speaker.

[34.227.112.145] Project MUSE (2024-03-28 10:16 GMT)

Messenger was substantially underestimating /s/, to the point where it has a slightly lower COG than /ʃ/, and they do not differ substantially.

	Estimate	SE	t-value	p
(Intercept)	4735.9	224.2	21.1	< 0.001
Program AudacityAlone	194.4	197.7	0.98	0.33
Program Cleanfeed	-301.3	197.7	-1.5	0.13
Program Messenger	-1676.7	197.7	-8.5	< 0.001
Program Skype	-380.4	197.7	-1.9	0.055
Program Zoom	509.3	197.7	2.6	0.01
Segment /ʃ/	-1544.7	279.5	-5.5	< 0.001
Program AudacityAlone:Segment /ʃ/	73.2	395.3	0.19	0.85
Program Cleanfeed:Segment /ʃ/	465.0	395.3	1.2	0.24
Program Messenger:Segment /ʃ/	1744.0	395.3	4.4	< 0.001
Program Skype:Segment /ʃ/	510.4	395.3	1.3	0.2
Program Zoom:Segment /ʃ/	-206.3	395.3	-0.52	0.6

TABLE S26. Linear mixed-effects model for COG in sibilant fricatives, with particular fricative as a factor. *Reference level Program = H4n, Segment = /s/.*

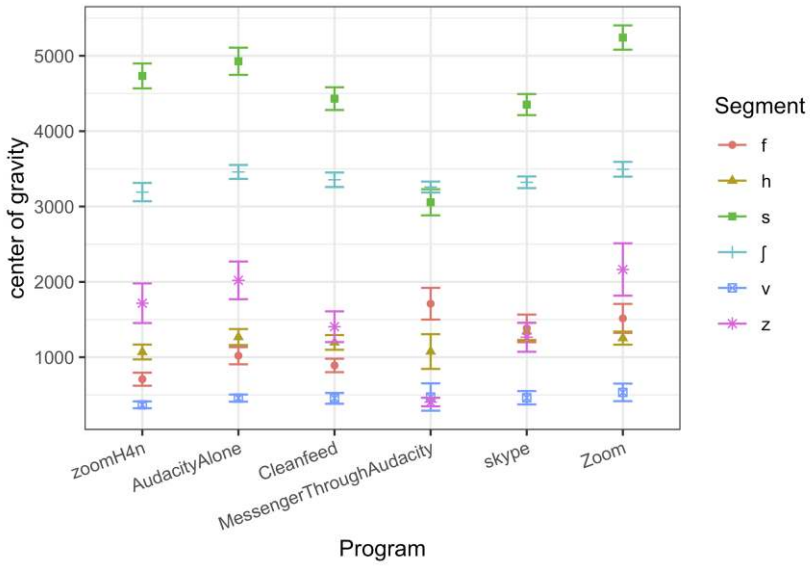


FIGURE S20. Measured center of gravity as predicted by program and segment, among fricatives. Pooled raw data, not the model results. Whiskers indicate the standard error.

Figure S20 illustrates COG across all fricatives. Zoom, Skype, and Messenger were also substantially overestimating the COG for /f/. Because the frication for /f/ is rather diffuse, this

could be the result of amplifying lower frequencies, or filtering out higher frequency aperiodic noise as 'background noise.'

Figure S21 illustrates F1 and F2 as influenced by vowel quality and device. Adding the interaction between vowel quality and device significantly improves the model for F1 (χ^2 = 208.3, df = 85, p = < 0.0001). The interaction between vowel quality and device also marginally improves the model for F2 (χ^2 = 102.3, df = 85, p = 0.097). There is substantial variation in the measurement of both formants in recordings made by different programs. These effects vary by vowel, which is why they did not show up as clearly in the some of the overall models for F1 and F2 above.

Many of the conditions produce measurements that substantially shift a vowel far into the region of a different vowel, which is likely to cause major problems in phonetic analysis and even in phonological categorization of tokens. While clusters for measurements of each vowel are mostly apparent, Messenger Through Audacity is a clear outlier for most of the vowels. The differences in formant measurements are likely to reflect a combination of factors. Some differences are directly due to compression algorithms changing spectral information. Other differences are indirect effects of differences caused by the recording program; background noise and filtering or amplifying certain frequencies can change the apparent center of a frequency band and might also lead to the wrong formants being identified.

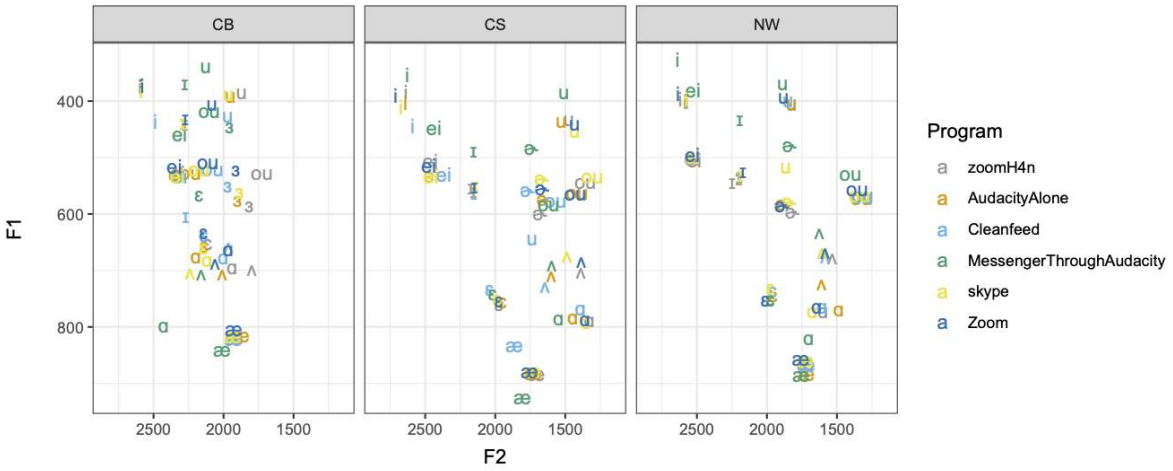


FIGURE S21. Vowel spaces for each speaker as measured in Phase 2, by-Program (= Figure 8 of main text).

2.3. COMPARING ZOOM CONDITIONS. This section provides the models comparing measurements across the Zoom conditions -- this varied based on whether the recording was local or remote, whether the computer was mac or windows, whether the files were converted from mp4 or not, and whether the recording used the 'Original Audio' setting in Zoom or not. Comparing these conditions makes it possible to narrow down which aspects of Zoom are causing the observed differences.

In most of these measures, there were clearly no effects. The variation between conditions is very small and in most cases there was substantially more variation within each condition than across conditions. The models are provided anyway, to give a sense of how similar the different

Zoom conditions are. For duration, two comparisons were marginally significant, but would not withstand correction for multiple comparisons.

Table S27 presents the summary of a linear mixed-effects model for consonant duration as predicted by the recording condition. There was a random intercept for speaker.

There were no significant consonant duration differences between the different conditions, though there were some small differences that did not reach significance.

	Estimate	SE	t-value	p
(Intercept)	95.6	9.4	10.2	0.0064
Condition Mac Local mp4	-0.73	3.4	-0.21	0.83
Condition Mac Remote mp4	2.4	3.4	0.7	0.49
Condition Mac Remote wav	5.4	3.4	1.6	0.11
Condition Windows Remote wav	4.4	3.4	1.3	0.19
Condition Mac Remote mp4 OriginalAudio	1.5	3.4	0.44	0.66
Condition Mac Remote wav OriginalAudio	1.4	3.4	0.41	0.68

TABLE S27. Linear mixed-effects model for consonant duration. *Reference level Condition = Local, macOSX, not 'original audio', not from mp4.*

Table S28 presents the summary of a linear mixed-effects model for vowel duration as predicted by Zoom condition. There was a random intercept for speaker.

Vowel duration in the Mac Remote mp4 condition was shorter than in the reference set of conditions. This effect was below the threshold of significance; however, it is important to keep in mind the large number of tests being conducted. When correcting for multiple comparisons, this effect is no longer significant. Given the lack of other significant effects, it is likely that this is merely due to multiple comparisons, rather than being a true effect of this particular condition in measurement of vowel duration.

	Estimate	SE	t-value	p
(Intercept)	188.6	11.8	15.9	< 0.001
Condition Mac Local mp4	-3.2	7.0	-0.45	0.65
Condition Mac Remote mp4	-13.9	7.0	-2.0	0.047
Condition Mac Remote wav	-13.4	7.0	-1.9	0.057
Condition Windows Remote wav	-11.2	7.0	-1.6	0.11
Condition Mac Remote mp4 OriginalAudio	0.28	7.0	0.04	0.97
Condition Mac Remote wav OriginalAudio	-0.68	7.0	-0.097	0.92

TABLE S28. Linear mixed-effects model for vowel duration. *Reference level Condition = Local, macOSX, not 'original audio', not from mp4.*

Table S29 presents the summary of a linear mixed-effects model for mean f0 in vowels as predicted by Zoom condition. There was a random intercept for speaker.

There was no effect of recording condition on f0 mean in any of the conditions; all of the differences were very small.

	Estimate	SE	t-value	p
(Intercept)	181.7	3.8	48.1	< 0.001
Condition Mac Local mp4	-0.35	1.6	-0.21	0.83
Condition Mac Remote mp4	-0.46	1.6	-0.28	0.78
Condition Mac Remote wav	-0.19	1.6	-0.12	0.91
Condition Windows Remote wav	-0.51	1.6	-0.31	0.75
Condition Mac Remote mp4 OriginalAudio	0.1	1.6	0.061	0.95
Condition Mac Remote wav OriginalAudio	0.69	1.6	0.42	0.68

TABLE S29. Linear mixed-effects model for mean f0 in vowels. *Reference level Condition = Local, macOSX, not 'original audio', not from mp4.*

Table S30 presents the summary of a linear mixed-effects model for f0 peak timing -- the position of the maximum f0 relative to the beginning of the vowel, as predicted by Zoom condition. There was a random intercept for speaker.

There was no significant effect of recording condition on f0 peak timing; there were some differences, but they were relatively small compared to the degree of variation found within each condition.

	Estimate	SE	t-value	p
(Intercept)	48.0	7.2	6.6	0.0056
Condition Mac Local mp4	0..22	5.0	0.044	0.96
Condition Mac Remote mp4	-7.3	5.0	-1.4	0.15
Condition Mac Remote wav	-7.1	5.0	-1.4	0.16
Condition Windows Remote wav	-4.7	5.0	-0.94	0.35
Condition Mac Remote mp4 OriginalAudio	0.1	5.0	0.02	0.98
Condition Mac Remote wav OriginalAudio	1.2	5.0	0.23	0.81

TABLE S30. Linear mixed-effects model for f0 peak timing in vowels (in milliseconds). *Reference level Condition = Local, macOSX, not 'original audio', not from mp4.*

Table S31 presents the summary of a linear mixed-effects model for jitter in vowels as predicted by Zoom condition. There was a random intercept for speaker.

There was no effect of recording condition on jitter; all of the differences between conditions were very small.

	Estimate	SE	t-value	p
(Intercept)	0.021	0.0035	6.0	0.02
Condition Mac Local mp4	-0.000048	0.0013	-0.038	0.97
Condition Mac Remote mp4	0.000044	0.0013	0.034	0.97
Condition Mac Remote wav	0.0006	0.0013	0.47	0.64
Condition Windows Remote wav	-0.00083	0.0013	-0.65	0.52
Condition Mac Remote mp4 OriginalAudio	-0.00019	0.0013	-0.15	0.88
Condition Mac Remote wav OriginalAudio	0.00032	0.0013	0.25	0.8

TABLE S31. Linear mixed-effects model for jitter in vowels. *Reference level Condition = Local, macOSX, not 'original audio', not from mp4.*

Table S32 presents the summary of a linear mixed-effects model for spectral tilt (H1-H2) in vowels as predicted by Zoom condition. There was a random intercept for speaker.

There was no effect of recording condition on spectral tilt.

	Estimate	SE	t-value	p
(Intercept)	-3.6	1.9	-1.9	0.18
Condition Mac Local mp4	0.16	0.54	0.29	0.77
Condition Mac Remote mp4	-0.091	0.54	-0.17	0.87
Condition Mac Remote wav	-0.02	0.54	-0.036	0.97
Condition Windows Remote wav	-0.26	0.54	-0.47	0.64
Condition Mac Remote mp4 OriginalAudio	-0.31	0.54	-0.56	0.58
Condition Mac Remote wav OriginalAudio	-0.034	0.54	-0.063	0.95

TABLE S32. Linear mixed-effects model for spectral tilt in vowels. *Reference level Condition = Local, macOSX, not 'original audio', not from mp4.*

Table S33 presents the summary of a linear mixed-effects model for Harmonics-to-Noise Ratio in vowels as predicted by Zoom condition. There was a random intercept for speaker.

	Estimate	SE	t-value	p
(Intercept)	6.9	0.96	7.2	0.016
Condition Mac Local mp4	-0.0046	0.28	-0.017	0.99
Condition Mac Remote mp4	0.16	0.28	0.56	0.57
Condition Mac Remote wav	0.19	0.28	0.66	0.51
Condition Windows Remote wav	0.31	0.28	1.1	0.27
Condition Mac Remote mp4 OriginalAudio	0.0017	0.28	0.006	0.99
Condition Mac Remote wav OriginalAudio	0.00078	0.28	0.003	0.99

TABLE S33. Linear mixed-effects model for HNR in vowels. *Reference level Condition = Local, macOSX, not 'original audio', not from mp4.*

There was no effect of recording condition on the Harmonics-to-Noise Ratio.

Table S34 presents the summary of a linear mixed-effects model for F1 in vowels as predicted by the recording condition. There was a random intercept for speaker and for vowel.

There was no effect of recording condition on F1.

	Estimate	SE	t-value	p
(Intercept)	605.7	48.4	12.5	< 0.001
Condition Mac Local mp4	1.6	5.0	0.33	0.74
Condition Mac Remote mp4	2.9	5.0	0.57	0.57
Condition Mac Remote wav	4.1	5.0	0.81	0.42
Condition Windows Remote wav	2.8	5.0	0.55	0.58
Condition Mac Remote mp4 OriginalAudio	2.7	5.0	0.53	0.59
Condition Mac Remote wav OriginalAudio	8.9	5.0	1.8	0.079

TABLE S34. Linear mixed-effects model for F1 in vowels. *Reference level Condition = Local, macOSX, not 'original audio', not from mp4.*

Table S35 presents the summary of a linear mixed-effects model for F2 in vowels as predicted by the recording condition. There was a random intercept for speaker and for vowel.

There was no effect of recording condition on F2.

	Estimate	SE	t-value	p
(Intercept)	1934.0	128.0	15.1	< 0.001
Condition Mac Local mp4	6.3	18.0	0.35	0.73
Condition Mac Remote mp4	-2.1	18.0	-0.12	0.91
Condition Mac Remote wav	-1.1	18.0	-0.062	0.95
Condition Windows Remote wav	4.3	18.0	0.24	0.81
Condition Mac Remote mp4 OriginalAudio	10.0	18.0	0.56	0.58
Condition Mac Remote wav OriginalAudio	25.2	18.0	1.4	0.16

TABLE S35. Linear mixed-effects model for F2 in vowels. *Reference level Condition = Local, macOSX, not 'original audio', not from mp4.*

Table S36 presents the summary of a linear mixed-effects model for center of gravity (COG) in fricatives as predicted by the recording condition. There was a random intercept for speaker and for segment.

Most of the conditions had a slightly higher COG than in the recording made locally on a Mac, not using the 'original audio' setting and without conversion from mp4. The comparisons do not remain significant when adjusting for multiple comparisons.

[34.227.112.145] Project MUSE (2024-03-28 10:16 GMT)

	Estimate	SE	t-value	p
(Intercept)	1902.8	660.8	2.9	0.024
Condition Mac Local mp4	282.0	110.0	2.6	0.011
Condition Mac Remote mp4	244.1	110.0	2.2	0.027
Condition Mac Remote wav	196.9	110.0	1.8	0.074
Condition Windows Remote wav	256.8	110.0	2.3	0.02
Condition Mac Remote mp4 OriginalAudio	267.329	110.0	2.4	0.015
Condition Mac Remote wav OriginalAudio	242.739	110.0	2.2	0.028

TABLE S36. Linear mixed-effects model for COG in fricatives. *Reference level Condition = Local, macOSX, not 'original audio', not from mp4.*

2.4. SIGNAL-TO-NOISE RATIO. This section reports the measurements of signal-to-noise ratio (SNR) in each condition, both for Phase 1 (comparisons by device) and Phase 2 (comparisons by program). SNR was calculated following the formula given in 1.

(1) $SNR = 20log(P_{signal}/P_{noise})$

Figure S22 plots the average signal to noise ratio across each condition. It was calculated by measuring the mean energy in the 'signal' (that is, from the words used in the analysis) compared to the background noise, as measured in intervals labeled as silence, using the following formula.

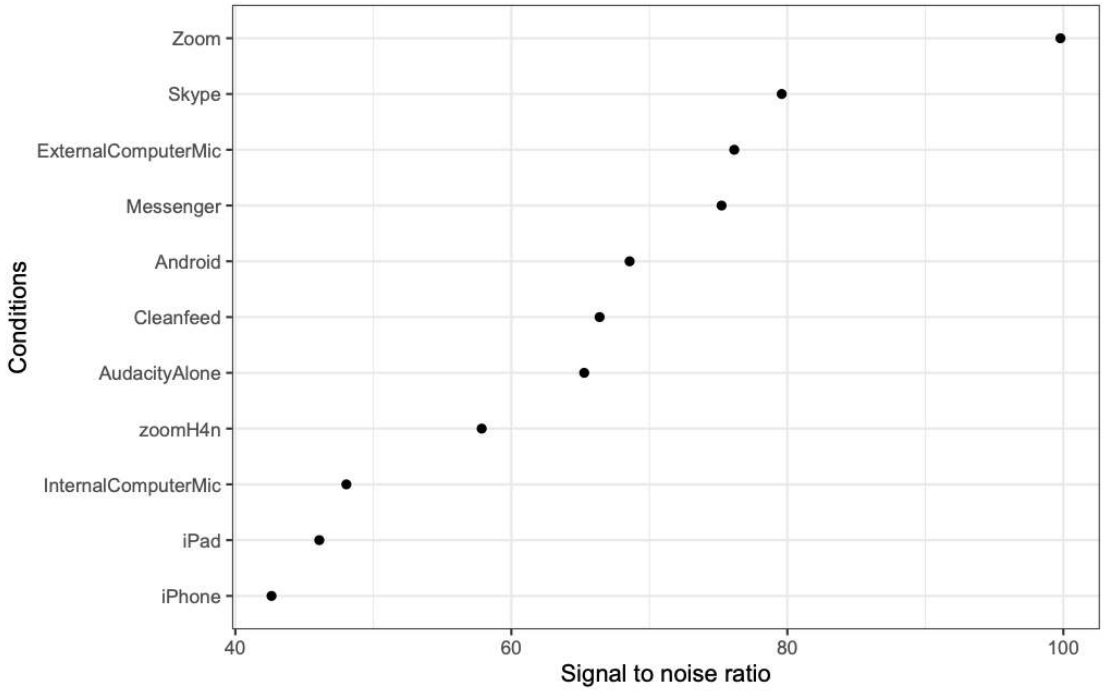


FIGURE S22. Signal to noise ratio by condition, across all devices and programs.

Signal to noise ratio should be above 50 DH for adequate recordings. Here the highest signal to noise ratios come from the Zoom recordings, presumably as an effect of the Zoom software suppressing background noise. Our gold standard recording did not have a particularly high signal to noise ratio, compared to some of the other recording devices used in the live recording condition. This is probably due in part to the sensitivity of the H4n's microphone and picking up background noise from the air conditioning system and external traffic noise.[2] Some of the software programs include filters to amplify what is identified as speech or suppress sounds that are identified as background noise; while this may improve perceptual clarity, it is altering the acoustic signal and could potentially influence the results in misleading ways, so having a higher SNR is not necessarily indicative of a better recording.

Table S37 presents the summary of a linear mixed-effects model for SNR as predicted by device. SNR was calculated for each sentence, using the maximum amplitude of the target word and of the silence following the sentence. There was a random intercept for speaker.

	Estimate	SE	t-value	p
(Intercept)	57.0	4.4	12.8	0.045
Device Android	10.2	1.0	10.1	< 0.001
Device ExternalComputerMic	19.2	1.0	19.0	< 0.001
Device InternalComputerMic	-11.5	1.0	-11.4	< 0.001
Device iPad	-13.7	1.0	-13.5	< 0.001
Device iPhone	-15.5	1.0	-15.3	< 0.001

TABLE S37. Linear mixed-effects model for SNR. *Reference level Program = H4n.*

Table S38 presents the summary of a linear mixed-effects model for SNR as predicted by program. There was a random intercept for speaker.

	Estimate	SE	t-value	p
(Intercept)	57.9	1.8	31.6	< 0.001
Program AudacityAlone	7.4	1.4	5.5	< 0.001
Program Cleanfeed	8.5	1.4	6.3	< 0.001
Program Messenger	17.4	1.4	12.9	< 0.001
Program Skype	21.7	1.4	16.1	< 0.001
Program Zoom	41.9	1.4	31.0	< 0.001

TABLE S38. Linear mixed-effects model for SNR. *Reference level Program = H4n.*

Keep in mind that SNR is based on the amplitude of the target words and the amplitude of the background noise as measured in the pauses between utterances. Reduction of noise when

[2] As mentioned in the main article, we attempted to mimic a reasonable field situation in that we recorded in a 'quiet' room but did not attempt to remove all background noise. While the building was quiet, there was both noise from the building's air conditioning system and traffic noise from the street outside.

there is no speech does not necessarily mean that a program like Zoom was equally effective at reducing background noise during speech, or that it removed noise in a way that leaves crucial acoustic characteristics of the speech signal intact.

Differences in the amplitude of the signal and the background noise could directly impact some acoustic measures, including intensity, center of gravity, and the harmonics-to-noise ratio, which are each discussed above. Differences in amplitude are also likely to be part of the explanation for differences in identification of segment boundaries. However, as will be discussed in the following section, differences in segmentation are not all indirect effects of amplitude or other characteristics.

2.5. TIMING ISSUES. To account for whether duration differences and other measurement differences were due to how the forced aligner was placing boundaries or if they were the result of actual duration differences caused by the condition, we tried combining these recordings with the textgrids produced for the baseline condition. The recordings were made under identical conditions (either because they were made at the same time or recorded from the H4n recorder's output), so the intervals should be identical; if they indeed are the same, the boundaries identified in the solid state recording condition should be transferable across all recording conditions, i.e. the textgrids produced for the H4n condition could be used for analysis in all condition.

However, the textgrids from the baseline condition do not align with the other conditions. Because of the substantial timing differences, it was impossible to use the textgrids from the baseline condition to make measurements in the other recording conditions. The lack of alignment across conditions makes clear that the compression/decompression systems of these programs created differences in timing. While the changes for any individual word are small (about 10 ms at most), these small mis-alignments can combine to produce substantial misalignment between recordings. (Note that for analytical purposes all files were aligned individually, so these offsets are not driving the differences between results). Many of the changes in timing can be attributed to compression in the silences between utterances, but there are also likely to be effects of compression during the utterances.

The following figures plot the difference between the interval timestamps for the gold standard H4n versus three recording conditions (Messenger, Cleanfeed, and Zoom), to illustrate the extent of the timing differences. Because the order in which the stimuli were presented was randomized between speakers, measurements are done separately for individual speakers.

As can be seen from Figures S23 and S24, the Zoom condition (in black) produced boundaries close to the boundaries in recordings from the 'gold standard' H4n recorder. The Messenger and Cleanfeed conditions, however, can differ from the H4n recorder by several hundred milliseconds.

To see an example comparing alignments alongside a spectrogram, consider Figure S1 above, which shows a comparison of alignments for sample words.

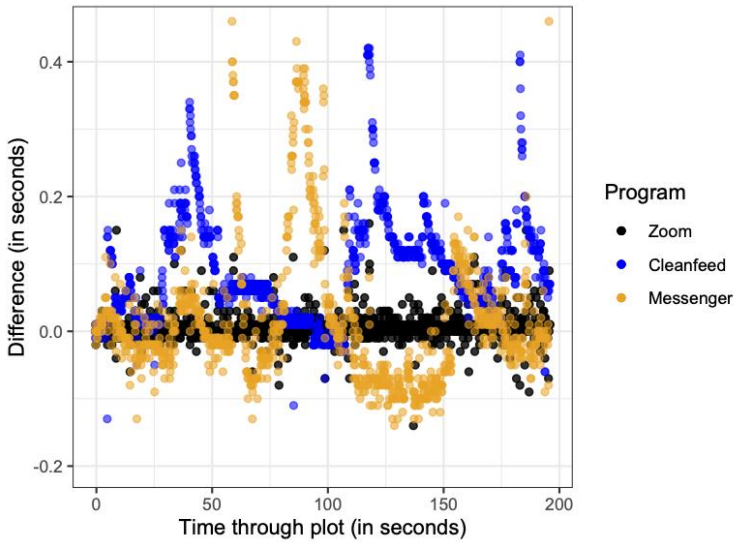


FIGURE S23. Difference in alignment between the H4n and three Program conditions (Messenger, Cleanfeed, and Zoom) for Speaker 1 (CS).

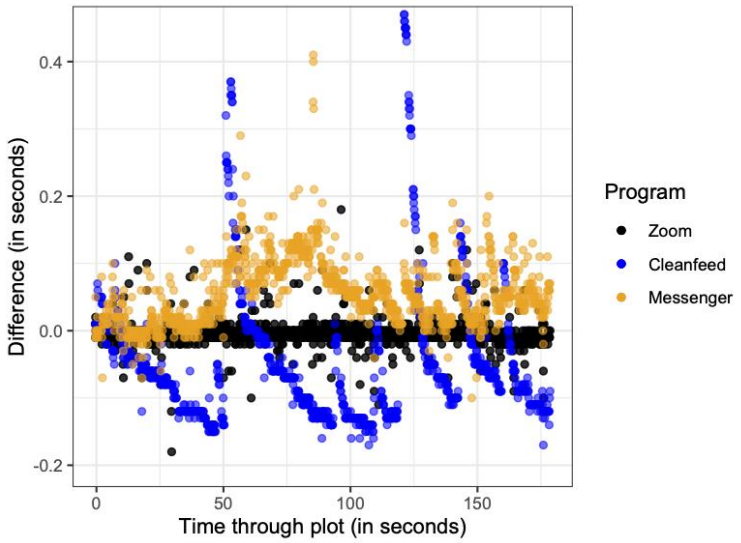


FIGURE S24. Difference in alignment between the H4n and three Program conditions (Messenger, Cleanfeed, and Zoom) for Speaker 2 (CB).

2.6. ADDITIONAL COMMENTS ON SOFTWARE. Here we offer some additional impressionistic summary comments about the software options and their relative reliability and ease of use, for researchers who are intending to make online recordings.

Cleanfeed was very user-friendly and performed well overall. It is probably the least well known of the set of software options tested here (but is used in podcasting interviews). It was straightforward to set up. The software allows the user to choose which speakers to record and how to record them (separate tracks, together, etc.). Individual participants can be muted. Muting individual participants was not particularly important for our tests but using this would allow a

way to have multiple remote participants while avoiding possible interference. However, it has a big drawback that video is not present, which limits its effectiveness.

Skype and **Zoom** are well known to participants; they are easy to set up and use. However, they exhibit extensive digital artefacts, so it is important to be careful when using these programs. Information about the conditions of recording (including any settings) should be included with recording metadata.

Facebook Messenger performed poorly in our tests, frequently giving outputs that differ from all the other conditions. Because Audacity alone behaves like the gold standard for almost all tests, the effects of Messenger recorded through Audacity, the effects cannot be attributed to Audacity itself. However, the effects might be due to how Messenger compresses the audio or in how Audacity interacts with audio input from Messenger. Messenger is widely available, but provides little control over recordings and produces unreliable results.

3. LIST OF STIMULI

Table S39 provides a list of the words elicited for analysis. The order of items was randomized for each speaker. Words occurred within the frame sentence 'We say ____ again.'

bad	cheap	fade	leave	rib	ten	insult (n.)
badge	chest	fan	mace	rich	tick	insult (v.)
base	chew	file	match	ridge	tongue	permit (n.)
bat	chip	fuss	maze	rim	tug	permit (v.)
batch	choke	fuzz	mob	rip	van	survey (n.)
batch	chug	gap	mop	roam	vase	survey (v.)
bead	clock	half	neck	robe	vote	suspect (n.)
bean	clog	have	paid	sap	wash	suspect (v.)
bet	deck	jest	pet	sheep	watch	torment (n.)
bid	den	joke	pick	ship	wish	torment (v.)
bit	dip	jug	pig	shoe	witch	
boat	do	lash	pile	sick	zap	
cab	edge	latch	plod	sue	zip	
cap	etch	leaf	plot	tap	zoo	

TABLE S39. Words used as stimuli.

4. SCRIPTS, RECORDINGS, and DATA

Scripts, stimuli, audio files, text grids, and raw result files have been uploaded to osf, at the following address: https://osf.io/yf9k8/?view_only=9458f75d3fdd4dadb98164e7d9f07560. In addition, the following Praat scripts were used.

Duration, jitter, mean f0, HNR: The script is included with the supplementary materials, DurationVoiceReportExtractor.

Peak timing: The script was modified from McCloy, Daniel. 2012. PRAAT SCRIPT "SEMI-AUTO PITCH EXTRACTOR". GitHub repository. https://github.com/drammock/praat-semiauto/blob/master/SemiAutoPitchAnalysis.praat

Spectral tilt: Vicenik, Chad. n.d. PraatVoiceSauceImitator. Praat Script Resources. http://phonetics.linguistics.ucla.edu/facilities/acoustic/PraatVoiceSauceImitator.txt

Formants: McCloy, Daniel & August McGrath. 2012. PRAAT SCRIPT "SEMI-AUTO FORMANT EXTRACTOR". GitHub repository. https://github.com/drammock/praat-semiauto/blob/master/SemiAutoFormantExtractor.praat

Center of Gravity (COG): DiCanio, Christian. 2013. Spectral moments of fricative spectra script in Praat. Scripts. https://www.acsu.buffalo.edu/~cdicanio/scripts/Time_averaging_for_fricatives_2.0.praat

REFERENCES

BATES, DOUGLAS; MARTIN MÄCHLER; BEN BOLKER; and STEVE WALKER. 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67(1). 1–48.

CHEN, WEI-RONG; DOUGLAS H. WHALEN; and CHRISTINE SHADLE. 2019. F0-induced formant measurement errors result in biased variabilities. Journal of the Acoustical Society of America 145(5). EL360-366.

HOU, LYNN; RYAN LEPIC; and ERIN WILKINSON. 2020. Working with ASL Internet Data. Sign Language Studies 21(1). 32–67.

JOHNSON, KEITH. 2012. Acoustic and auditory phonetics. 3rd ed. Oxford: Wiley-Blackwell.

JOHNSON, LISA M.; MARIANNA DI PAOLO; and ADRIAN BELL. 2018. Forced Alignment for Understudied Language Varieties: Testing Prosodylab-Aligner with Tongan Data. Language Documentation & Conservation 12. 80–123.

KONG, EUN JONG; MARY E. BECKMAN; and JAN EDWARDS. 2012. Voice onset time is necessary but not always sufficient to describe acquisition of voiced stops: The cases of Greek and Japanese. Journal of Phonetics 40. 725-744.

KUZNETSOVA, ALEXANDRA; PER BRUUN BROCKHOFF; and RUNE HAUBO BOJESEN CHRISTENSEN. 2015. lmerTest: Tests in Linear Mixed Effects Models. https://CRAN.R-project.org/package=lmerTest. R package version 2.0-29.

LUCAS, CEIL; GENE MIRUS; JEFFREY L. PALMER; NICHOLAS J. ROESSLER; and ADAM FROST. 2013. The effect of new technologies on sign language research. Sign Language Studies 13(4). 541–564.

MIHAS, ELENA. 2012. Subcontracting native speakers in linguistic fieldwork: A case study of the Ashéninka Perené (Arawak) research community from the Peruvian Amazon. Language Documentation and Conservation 6. 1–21.

PURNELL, THOMAS; ERIC RAIMY; and JOSEPH SALMONS. 2013. Making linguistics matter: Building on the public's interest in language. Language and Linguistics Compass 7(7). 398–407.