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

Yale University

7 July 2021

Fieldwork in the pandemic

- In-person research has been restricted by the Covid-19 pandemic, resulting in rapid shifts to data collection using remote recordings
- How are acoustic measurements impacted by different recording devices and recording software?

Potential sources of variation

Potential sources of variation include:

- Compression – potentially altering frequency and duration (van Son 2005; de Decker & Nycz 2015; Nash 2001; Liu, Hsu, & Lee 2008)
- Filtering – potentially altering aperiodic noise, overall intensity, and relative intensity at different frequencies Dreiseitel & Schmidt 2006)
- And familiar concerns from prior work: sampling rate, ambient noise, shielding, and microphone placement and sensitivity (Barwick 2006; Bowern 2015; Seyfeddinipur & Rau 2020)

Methods overview

Two phases of data gathering:

1. Simultaneous recordings on six different devices
2. Transmission of recorded speech over four (video-)conferencing applications

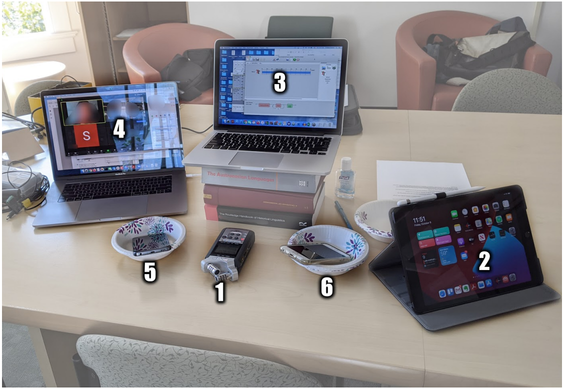
All recordings were compared against a ‘gold standard’ solid state Zoom H4n recorder

Stimuli

Stimuli were 94 target words embedded in the carrier sentence ‘we say [word] again’, elicited in randomized order from three native speakers of English

- Designed to test some parameters of different types (duration, frequency, aperiodic noise)
- Both broadly and in comparisons where they are part of phonological distinctions in English (e.g. f0 as related to stress and onset voicing)

Phase 1 set-up: Devices



Phase 1 recording setup. (1) Zoom H4n; (2) ipad; (3) computer with internal microphone; (4) computer with external headset microphone; (5) android phone; (6) iphone

Phase 2 set-up: Programs

- The recordings from the Zoom H4n (‘H4n’) solid state recorder were played through the sound card of a computer, using each program
- While not equivalent to recording live speech, this ensures identical signals transmitted in each remote recording condition
- The software programs which were tested were:
  - Zoom
  - Skype
  - Cleanfeed, a commonly used podcast interview platform
  - Facebook Messenger (using Audacity to make the recording, since Messenger does not have built-in recording capabilities – for comparison, an additional condition tested Audacity locally)

Acoustic Analysis

- Audio files were converted to 16,000 Hz uncompressed mono wav files, so they all had the same sampling rate and file type
- Force-aligned to the segmental level using the Penn Forced Aligner (Yuan & Liberman 2008)
- Measurements from the target words were extracted with scripts in Praat
- Results from mixed effects models, with Device (Phase 1) or Program (Phase 2) as the one fixed effect; the reference condition was the H4n recorder. Random intercept for speaker. Formants and COG also included a random intercept for segment.

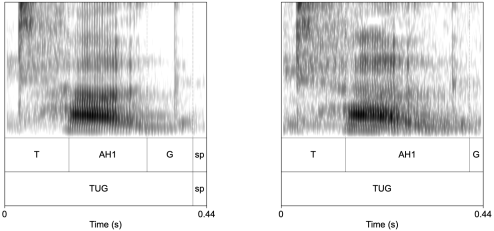
Summary of results

- While our set of measurements is not exhaustive, we cover several types of measurements
- These can be broadly grouped into measurements of duration, measurements of intensity of aperiodic noise, and measurements of frequency; all three categories exhibit effects

Duration

In many conditions, consonant duration was underestimated and vowel duration was overestimated

- In part due to lossy compression effects on timing
- Also in part due to boundaries being obscured by noise and lowered intensity of the signal



The word *tug* recorded by the H4n recorder (left) and ipad (right)

Aperiodic noise

- Some effects are caused by differences in capturing the signal and filtering meant to remove background noise or boost the speech signal
- This variation can directly impact measurements like Harmonics-to-Noise ratio (HNR) and center of gravity (COG)
- This can also have indirect effects on how well target characteristics can be identified

Frequency

- Measurements of frequency (particularly formants) are impacted in several conditions, likely due to a combination of lossy compression, filtering, and noise
- Depending on how the compression system handles repeating waves, these could be over-regularized or obscured
- Some spectral issues may be caused by changes in intensity, as lower sensitivity to a particular frequency ranges can shift frequency measurements
- Changes in intensity of different frequencies is also directly reflected in spectral tilt for all programs

Device Comparisons (Phase 1)

Table: Effects of device on acoustic measures. Each value is the estimate for that factor in the model predicting the given acoustic measure; stars indicate significance. For readability, only significant results are included.

Device	Android	External Mic	Internal Mic	iPad (compressed)	iPhone (uncompr)
Consonant duration (ms)			-9.6**	-9.0*	
Vowel duration (ms)				15.5*	
Mean vowel f0 (Hz)					
Peak f0 timing (ms)					
Jitter					
Spectral tilt	-1.5*				
Harmonics-to-noise ratio			-1.5***		
F1 (Hz)			-19.8**	-15.2*	-25.7***
F2 (Hz)	56.8*		-77.4**	145.0***	70.6**
Center of gravity (frics)	440.3***		1172.5***	1115.2***	
Signal to noise ratio	10.2***	19.2***	-11.5***	-13.7***	-15.5***

Software Comparisons (Phase 2)

Table: Effects of program on acoustic measures. Each value is the estimate for that factor in the model predicting the given acoustic measure; stars indicate significance. For readability, only significant results are included.

Device	Audacity	Cleanfeed	Messenger	Skype	Zoom
Consonant duration (ms)			-11.6***	-8.5**	-11.2***
Vowel duration (ms)			17.5**	19.8**	31.5***
Mean vowel f0 (Hz)					
Peak f0 timing (ms)					14.2**
Jitter					
Spectral tilt	-1.4**	-1.3**	4.6***	-1.7***	-2.0***
Harmonics-to-noise ratio			1.2***		
F1 (Hz)			-29.7***		
F2 (Hz)		46.0*	91.0***	42.0*	
Center of gravity (frics)		-653.3***	-904.1***		
Signal to noise ratio	7.4***	8.5***	17.4***	21.7***	41.9***

Comparing Zoom conditions

We tested several combinations of settings in Zoom:

- whether the recording was local or remote
- whether the computer was mac or windows
- whether the files were converted from mp4 or not
- whether the recording used the ‘Original Audio’ setting in Zoom or not

There was very little variation between conditions

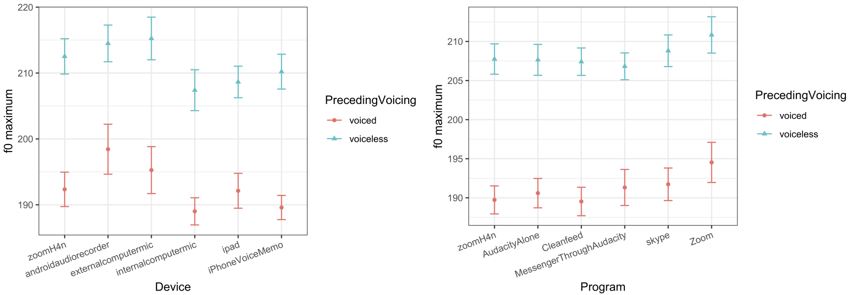
Relative measurements (Correlates of contrasts)

How clearly are the acoustic correlates of phonological contrasts preserved?

- Stress as reflected in vowel duration and f0 maximum
- Coda voicing reflected in vowel duration and HNR
- Onset voicing indicated in HNR, spectral tilt, and F0 maximum
- Vowel category indicated in F1 and F2
- Fricative identity indicated in COG

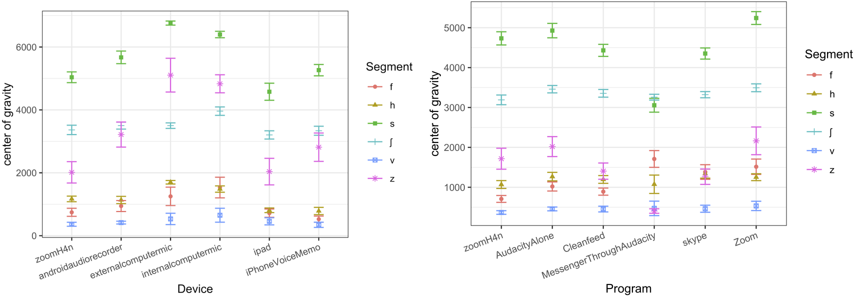
Most contrasts were captured by all conditions, even when the raw measurements differed by condition
But sometimes the size of the difference varied, and some contrasts were not captured

Relative measurements: F0 maximum by onset voicing



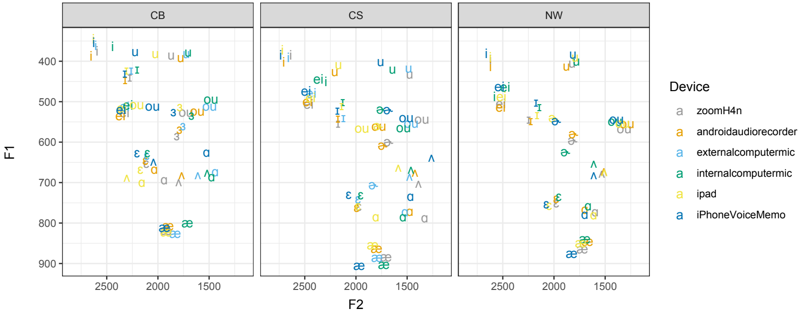
F0 maximum as predicted by condition and onset voicing, by device (left) and program (right). Whiskers indicate the standard error. No interactions between onset voicing and condition were significant.

Relative measurements: COG by fricative



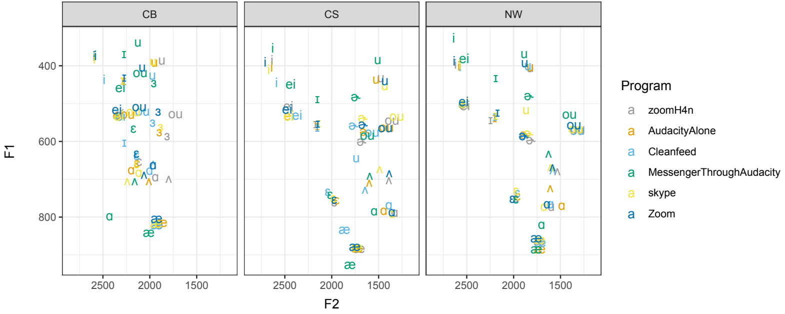
COG as predicted by condition and fricative, by device (left) and program (right). Whiskers indicate the standard error. Several of the interactions between condition and fricative are significant.

Relative measurements: Vowel spaces across Device



Vowel spaces by speaker in Phase 1 (comparisons by Device). The interaction between device and vowel significantly improved the models, but all conditions picked out a recognizable vowel space.

Relative measurements: Vowel spaces across Program



Vowel spaces by speaker in Phase 2 (comparisons by Program). The interaction between program and vowel significantly improved the models; measurements differed substantially between conditions, obscuring differences between categories.

Summary

- Both device and software affected phonetic measurements
- The acoustic correlates of contrasts generally remained clear, but some contrasts were exaggerated or underestimated
- A major concern for data gathered remotely or gathered in person in different ways (e.g. asking participants to record themselves)

Recommendations

- Documenting the recording setup is crucial: microphone used, program used, and any settings for programs that allow multiple settings
- Using different in-person devices is preferable to making recordings through video-conferencing software, particularly if devices are similar
- If using video-conferencing software, it is crucial to use the same program

Additional factors to consider

- We tested a sample of conditions, but this is far from covering all possibilities for devices and software
- We only examined English; noise-reduction algorithms may have a different impact on other languages
- All our virtual recordings were run on stable high-speed internet connections; slow connections would introduce additional issues not observed here

References

Barwick, Linda. 2006. A musicologist's wishlist: some issues, practices and practicalities in musical aspects of language documentation. *Language documentation and description 3(2005)*. 53–62.

Bowern, Claire. 2015. Linguistic fieldwork: A practical guide. 2nd edition. London: Palgrave Macmillan.

de Decker, Paul, and Jennifer Nycz. 2015. For the record: Which digital media can be used for sociophonetic analysis? *University of Pennsylvania Working Papers in Linguistics* 17(2). Article 7.

Dreiseitel, Pia, and Gerhard Schmidt. 2006. Evaluation of algorithms for speech enhancement. *Topics in acoustic echo and noise control: Selected methods for the cancellation of acoustical echoes, the reduction of background noise, and speech processing*, ed. by Eberhard Hänsler and Gerhard Schmidt, 431–484. Berlin: Springer.

Liu, Chi-Min; Han-Wen Hsu; and Wen-Chieh Lee. 2008. Compression artifacts in perceptual audio coding. *IEEE Transactions on Audio, Speech, and Language Processing* 16(4). 681–695.

Nash, Carlos Marcelo. 2001. Evaluating the use of adaptive transform acoustic coding (ATRAC) data compression in acoustic phonetics. Houston, TX: Rice University Master's Thesis.

Seyfeddinipur, Mandana, and Rau, Felix. 2020. Keeping it real: Video data in language documentation and language archiving. *Language Documentation & Conservation* 14. 503–519.

van Son, R.J.J.H. 2005. A study of pitch, formant, and spectral estimation errors introduced by three lossy speech compression algorithms. *Acta Acustica United with Acustica* 91. 771–778.

Yuan, Jiahong and Mark Liberman. 2008. Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics 2008.*