Journal of the Acoustical Society of America, 154, 152–166. https://doi.org/10.1121/10.0020064

Voice onset time and vowel formant measures in online testing and laboratory-based testing with(out) surgical face masks

Antje Stoehr,[1,a] Christoforos Souganidis,[1,2] Trisha B. Thomas,[1,3] Jessi Jacobsen,[1] and Clara D. Martin[1,4]

[1] *Basque Center on Cognition, Brain and Language, Donostia–San Sebastián, 20009, Spain*

[2] *University of the Basque Country, Vitoria–Gasteiz, 01006, Spain*

[3] *University of the Basque Country, Leioa, 48940, Spain*

[4] *Ikerbasque Basque Foundation for Science, Bilbao, 48009, Spain*

Since the COVID-19 pandemic started, conducting experiments online is increasingly common, and face masks are often used in everyday life. It remains unclear whether phonetic detail in speech production is captured adequately when speech is recorded in internet-based experiments or in experiments conducted with face masks. We tested 55 Spanish–Basque–English trilinguals in picture naming tasks in three conditions: online, laboratory-based with surgical face masks, and laboratory-based without face masks (control). We measured plosive voice onset time (VOT) in each language, the formants and duration of English vowels /iː/ and /ɪ/, and the Spanish/Basque vowel space. Across conditions, there were differences between English and Spanish/Basque VOT and in formants and duration between English /iː/–/ɪ/; between conditions, small differences emerged. Relative to the control condition, the Spanish/Basque vowel space was larger in online testing and smaller in the face mask condition. We conclude that testing online or with face masks is suitable for investigating phonetic detail in within-participant designs although the precise measurements may differ from those in traditional laboratory-based research.

[a] Email: a.stoehr@bcbl.eu

I. INTRODUCTION

The COVID-19 pandemic has posed challenges to conducting laboratory-based psycholinguistic research and caused research studies to move online. In 2020, many psycholinguistic research laboratories around the world remained closed and internet-based studies were the only option for collecting data. When laboratories reopened, participants were often obliged to wear face masks during experiments. As of the time this article was written, many countries have lifted official mask mandates, and people infected with COVID-19 are generally no longer required to isolate themselves. However, it is still strongly advised to wear face masks in several countries if you have symptoms consistent with COVID-19 or have been in contact with infected people to reduce airborne disease transmission. Given the proven efficacy of face masks in reducing the spread of respiratory diseases, it is likely that their use may be mandated again in the future to ensure public health and safety. Both online testing and the use of face masks in on-site research challenge phonetically oriented research, as the phonetic properties of speech may be altered. In this paper, we present a systematic comparison of a set of phonetic properties in Spanish–Basque–English trilinguals' speech production elicited in three conditions: online, in the laboratory with surgical face masks, and the laboratory without face masks. We examine the phonetic detail through measures of voice onset time (VOT) and vowel formants, the most widely used measures in bi-/multilingual phonetic research (Cabrelli Amaro and Wrembel, 2016; Hansen Edwards and Zampini, 2008).

A. Face masks in speech production studies

To date, only two published studies have investigated the direct influence of face masks on the phonetic properties of vowel production (Bond et al., 1989; Georgiou, 2022a) and none have investigated the consequences on plosive production. Long before the COVID-19 pandemic, Bond et al. (1989) found that wearing oxygen masks reduced the (American English) vowel space, especially along the first formant (F1) dimension. This finding may be explained by the physical barrier face masks constitute. Face masks may restrict jaw and lip movements, which may limit the F1 range of vowels and affect the articulation of labial consonants, respectively (Saedi et al., 2015). Thus, a restricted jaw movement can explain the Bond et al. finding of a reduced vowel space along the F1 dimension. However, it is important to note that the oxygen masks used by Bond et al. were quite distinctive from types of masks that are commonly used today to prevent disease transmission. Therefore, their findings might not be generalizable to other (more flexible) types of masks. More recently, Georgiou (2022a) studied the effect of the now commonly used surgical and cotton face masks on phonetic detail in the production of the Cypriot Greek vowels /i e a o u/. Unlike Bond et al., Georgiou (2022a) did not find that F1 was detectably altered by wearing either face mask. This could mean that the more flexible face masks used in Georgiou's study did not restrict jaw movement as did the more static oxygen masks in Bond et al.

Though Georgiou (2022a) did not find evidence of altered F1, he did find that wearing face masks affected the production of vowels along the second formant (F2) dimension. However, not all vowels were affected similarly: surgical face masks were associated with increased F2 in /e/ and /u/ and decreased F2 in /a/, but cotton face masks were associated with decreased F2 in /e/ and /a/. This could be the result of face masks filtering certain frequencies, and allowing others to pass. A systematic comparison of the acoustic attenuation caused by various types of surgical, respirator, cloth, transparent, and shield face masks showed that face masks generally attenuate frequencies above 1 kHz (2 kHz for surgical face masks) with the strongest attenuation above 4 kHz (Corey et al., 2020). This attenuation may affect F1 as well as the higher frequency F2. Across all face masks, surgical face masks provided the best acoustic performance, which may explain why Georgiou (2022a) did not observe effects on F1. However, in Georgiou (2022a), alterations in F2 were not limited to high frequency as reported by Corey et al. (2020). Instead, also the relatively low F2 of

/u/ (surgical face masks) and intermediate F2 of /a/ (surgical and cotton face masks) were affected, which means that the filtering properties reported in Corey et al. (2020) cannot fully explain the Georgiou (2022a) results.

In contrast to the sparse research into the effect of face masks on speech production, the consequences of face masks on speech perception have received considerable attention. This perception research supports the idea that wearing face masks during speech production impacts phonetic speech properties, which may lead to either impaired speech intelligibility (Atcherson et al., 2017; Corey et al., 2020; Goldin et al., 2020; Magee et al., 2020) or enhanced speech intelligibility, at least under certain conditions (Cohn et al., 2021 for enhanced intelligibility with face masks during clear speech production; Pycha et al., 2022 under noise (but see Toscano and Toscano, 2021); Zellou et al., 2023 for enhanced intelligibility with face masks for coarticulatory vowel nasalization in lax vowels). Impaired speech perception can be explained by the filtering and attenuating properties of face masks (Bond et al., 1989; Corey et al., 2020; Georgiou, 2022) and, if applicable, the lack of visual information (Lalonde and Werner, 2019). Improved intelligibility of face-masked speech can be explained by speakers' compensation for the face masks' restricting properties. They may be speaking louder (Asadi et al., 2020) producing Lombard speech (Bond et al., 1989)—which is characterized by speaking more loudly, with higher fundamental frequency and longer vowel durations—and by articulating more clearly than when not wearing face masks, at least under certain conditions (Cohn et al., 2021; Pycha et al., 2022; Zellou et al., 2023). In this sense, an altered acoustic signal in face-masked speech may not only result from the physical properties of face masks themselves, but also from the psychological state of the speaker who may change their speaking style in order to adapt to the face mask.

In sum, wearing face masks likely affects the phonetic properties of speech due to the physical characteristics of the masks themselves and/or their psychological effect on the speaker. This may lead to unrepresentative phonetic measurements during data analysis, which can cause researchers to reach faulty conclusions about the speech system. Moreover, we are not aware of any studies that have systematically investigated the effect of face masks on the production of various speech sounds in a multilingual's languages. The present study investigates the effect of commonly used surgical face masks on trilinguals' vowel and plosive production.

B. Online testing in speech production studies

Online testing is an attractive option for speech production research. First, speech data can be safely acquired in an environment where no face masks are required. Second, independent of safety concerns, larger populations can be accessed, leading to better-powered research studies. The latter would be especially beneficial for bi-/multilingualism research, which is generally restricted by the local availability of participants, often resulting in underpowered studies (Brysbaert, 2021). However, testing participants online means that speech is recorded using diverse devices in diverse environments, which may cause variability, especially between participants. There is growing evidence that online studies relying on reaction time and speech onset time measures can reliably detect well known psycholinguistic effects, such as the word frequency effect (e.g., Anwyl-Irvine et al., 2020; de Leeuw and Motz, 2016; Fairs and Strijkers, 2021; Hilbig, 2016; Vogt et al., 2022).

To our knowledge, only a small number of studies investigated the suitability of online audio recordings for phonetic research. Four relatively small-scale studies with English-speaking participants suggest that the use of different remote recording devices affects phonetic analyses of vowel production (Bulgin et al., 2010; Calder et al., 2022; Freeman and De Decker, 2021; Zhang et al., 2021). In all studies, participants recorded themselves with several devices or applications simultaneously, meaning any differences in vowel measurements could be attributed to the recording device or application rather than the characteristics of the specific production. Zhang et al. had participants produce isolated vowels and record themselves at home with three devices in parallel:

(1) the built-in microphones on their laptops running the Zoom cloud meeting application (32 kHz .m4a files), (2) the built-in microphones of their mobile phones using the Awesome Voice Recorder application (256-bps, 44.1 kHz .wav files) and the Recorder application (256-bps, 44.1 kHz .ogg files), and (3) a high-quality Zoom H6 Handy Recorder (24-bit, 44.1 kHz .wav files). Vowel formants were extracted using Praat (version 6.1.08; Boersma and Weenink, 2019) and VoiceSauce (Shue, 2010). In laptop recordings using the Zoom cloud meeting application, F1, F2, and F3 were lower than when recorded with the high-quality H6 recorder (F3 was only lower when extracted with Praat). In mobile phone recordings, only F2 was lower than in recordings made with the high-quality H6 recorder when extracted with VoiceSauce but not when extracted with Praat. The participants in the Freeman and De Decker (2021) study (one female and one male American English speaker) recorded themselves in the laboratory using five devices while reading a word list: a high-quality H4n Pro field recorder (16-bit, 44.1 kHz .wav files) and four identical iPads using different recording applications: (1) the offline Voice Memos application (32-bit, 44.1 kHz .m4a files), (2) the Zoom cloud meeting application (32-bit, 32 kHz .m4a files), (3) the Microsoft Skype application (32-bit, 16 kHz .mp4 files), or (4) the Microsoft Teams application (32-bit, 16 kHz .mp4 files). The Zoom recordings were also saved by two receivers using either a high-quality, medium-quality, or low-quality internet connection. Critical measures were F1 and F2 to assess vowel space shape and vowel overlap patterns, as well as spectral tilt to assess vowel nasalization. The measurements varied by recording device/application and transmission most strongly for the female speaker and in frequencies between 750 and 1500 Hz. However, these differences were generally small, and a larger sample size would be needed to draw generalizable conclusions. Just recently, Calder et al. (2022) tested 18 English-speaking participants with various language backgrounds who recorded conversations and isolated words elicited in a reading task in their homes using the Zoom cloud meeting application (32 kHz .m4a files) and also portable audio recorders (15/18 Olympus, 1/18

TASCAMDR-100MKII, 1/18 VoicetracerDVT 2050, 1/18 Philips VTR8060; all 16-bit, 44.1 kHz .wav files) and a SLINT omnidirectional condenser lavalier lapel microphone. Critical measures were vowel F1 and F2. Vowel formants were analyzed either raw, normalized using the Lobanov method or the Watt and Fabricius modified method. In Zoom recordings, raw F1 was lower and raw F2 was higher compared to recordings done with the portable recorder. Using Lobanov normalization, no effect of the recording condition was detected. Using the Watt and Fabricius modified normalization, both F1 and F2 were higher when recorded with Zoom.

To summarize, online speech recordings may fundamentally improve access to multilingual communities, but it is possible that phonetic measurements taken from speech recorded online differ from recordings made under controlled conditions in the laboratory. These differences may arise from differences in sampling rate, internet connection, and from differences in the recording environment. The present study tests whether online studies are suited to investigating phonetic detail in trilinguals' speech production through measures of VOT in plosives and formants in vowels.

C. VOT

VOT is the time interval between a plosive's burst release and voicing onset and the most important acoustic differentiator of phonologically voiced from phonologically voiceless plosives (Lisker and Abramson, 1964). The VOT continuum (Figure 1) can generally be divided into three phonetic ranges: prevoicing (negative VOT), short lag (short positive VOT, usually <30 ms) and aspiration (long positive VOT >30 ms, usually around 70 ms).

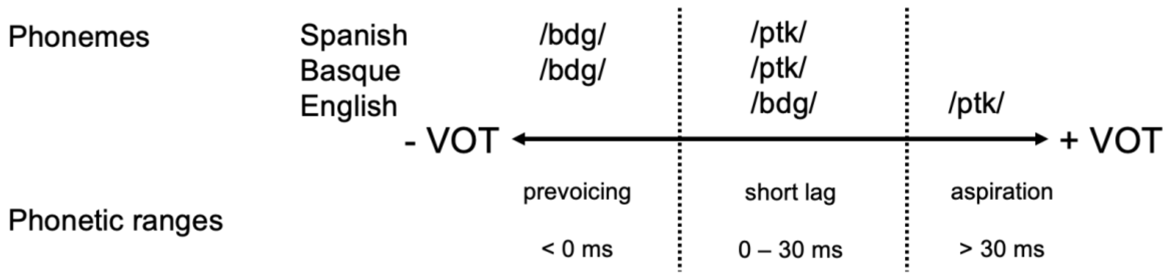


FIG. 1. VOT in Spanish, Basque, and English.

Spanish, (Standard)[1] Basque, and English have a voicing contrast between voiceless /p t k/ and voiced /b d ɡ/. English, however, differs from Spanish and Basque in the phonetic implementation of the voicing contrast (e.g., Lisker and Abramson, 1964; Souganidis et al., 2022). English /p t k/ are produced with aspirated VOT, but both Spanish and Basque /p t k/ are produced with short lag VOT. Bi-/multilingual speakers are generally known to produce language-specific VOT for voiceless plosives if their languages differ in the phonetic implementation of voicing (among many others: Flege, 1987, 1991; Stoehr et al., 2017). Importantly, although Spanish–Basque–English trilinguals often do not produce monolingual-like aspiration in English, they generally produce longer VOT in English than in Spanish and Basque (Stoehr et al., 2023). English /b d ɡ/ fall within the short lag range, but Spanish and Basque /b d ɡ/ are produced with prevoicing, that is, negative VOT. Native speakers of true-voicing languages like Spanish or Basque often carry over prevoicing to voiced plosives in their aspirating nonnative language, but they may differ in the proportion of voiced plosives produced with prevoicing (for Dutch–German bilinguals: Stoehr et al., 2017; for German–Italian–English trilinguals: Geiss et al., 2022) or they may produce distinct prevoicing durations in their true-voicing and aspirating languages (for Portuguese–English bilinguals: Osborne and Simonet, 2021).

In this study, we test whether the expected crosslinguistic VOT production differences between English and Spanish/Basque are detectable when speech production is elicited online or while participants wear surgical face masks in the laboratory. We hypothesize that Spanish–Basque–English trilinguals produce language-specific VOT in each condition. If this hypothesis is true, we predict that Spanish–Basque–English trilinguals produce voiceless plosives with longer VOT in English than in Spanish and Basque in each condition and a smaller proportion of voiced plosives with prevoicing in English than in Spanish and Basque in all conditions. However, it is possible that face masks constitute a physical barrier that shortens the duration of aspiration in English voiceless plosives, which may reduce the VOT production difference between English and Spanish/Basque when participants wear surgical face masks. We further hypothesize that prevoicing in voiced plosives cannot always be measured in online testing because it is a subtle acoustic signal that may not be captured by all recording devices and in uncontrolled environments, which may lead to a lower proportion of prevoiced plosives in the online condition.

D. Vowel formants

Formants refer to resonant frequencies of the vocal tract and are one of the primary acoustic cues for distinguishing vowels (Peterson and Barney, 1952). The first formant (F1) corresponds to vowel height and is correlated with tongue height and jaw position, such that vowels produced with a higher tongue and a more closed jaw position have smaller F1. The second formant (F2) corresponds to vowel backness and is correlated with the length of the vocal tract, such that vowels produced further back in the mouth have smaller F2. Formants are usually defined by automatic tracking algorithms like the one used by Praat (Boersma and Weenink, 2021). Formants are traditionally measured in Hertz (Hz) or on the psychoacoustical Bark scale (Zwicker, 1961).

The Spanish and Basque vowel inventories comprise the same five vowels /i e a o u/ (Hualde, 1991; Ladefoged and Johnson, 2010). The vowels /i a u/ form the vowel space, which is delimited by the distance between /i/–/a/, /a/–/u/, and /u/–/i/. The vowel space size is an important measure in various disciplines, including speech development (Flipsen and Lee, 2012; Pettinato et al., 2016), speech directed to infants (Rattanasone et al., 2013) and foreigners (for a review: Piazza et al., 2022), clinical linguistics (Sapir et al., 2010; Skodda et al., 2012), and sociolinguistics (Fox and Jacewicz, 2017; Pierrehumbert et al., 2004). Among measures for determining vowel space size, a particularly promising measure is the Formant Centralization Ratio (FCR), which maximizes sensitivity to vowel centralization and reduces inter-speaker variability (Sapir et al., 2010).

In the present study, we use FCR measures to test whether the size of the Spanish and Basque vowel space differs when participants' speech is recorded online or while they wear surgical face masks in the laboratory. As such, this research question focuses on a general influence of the testing condition on speech production in the (near) native languages and does not address multilinguals' speech production per se. As Spanish and Basque have the same vowel inventory, we consider these two languages together. We hypothesize that the size of the Spanish/Basque vowel space in Spanish–Basque–English trilinguals differs by condition. If this hypothesis is true, we predict that surgical face masks restrict the jaw to some extent, thereby resulting in a smaller vowel space when participants wear face masks compared to when they do not (Bond et al., 1989). Since the online recordings are made with various recording devices, we predict differences in vowel space size in speech elicited online versus in the laboratory without face masks, but it is unclear whether the different recording devices result in a smaller or larger vowel space.

The English vowel inventory is considerably larger than the Spanish and Basque vowel inventories, but the number and type of vowels differ by variety and dialect. The production of certain English vowel contrasts, such as the contrast between tense /iː/ (e.g., in "sheep") and lax /ɪ/ (e.g., in "ship") are reportedly difficult for native speakers of Spanish and other languages lacking this contrast (e.g., Cebrian, 2007; Cebrian et al., 2021; Georgiou, 2022b). The vowels /iː/ and /ɪ/ differ in three dimensions: vowel height (/iː/ is higher/has smaller F1), vowel backness (/iː/ is more frontal/has larger F2), and duration (/iː/ is longer). Importantly, although speakers of languages lacking the

/iː/–/ɪ/ contrast differ from native English speakers' production of /ɪ/ (e.g., Cebrian et al., 2021), they generally distinguish /iː/ and /ɪ/ in production to some extent, either producing distinct formants (Georgiou, 2022b) and/or duration (Cebrian, 2007; Cebrian et al., 2021; Georgiou, 2022b).

In this study, we test whether the /iː/–/ɪ/ contrast is measurable in Spanish–Basque–English trilinguals' speech elicited online or while they wear surgical face masks in the laboratory. We hypothesize that Spanish–Basque–English trilinguals produce English /iː/ and /ɪ/ distinctly in each condition. If this hypothesis is true, we predict that Spanish–Basque–English trilinguals' production of English /iː/ and /ɪ/ differs in at least one of the following three measures in each condition: F1, F2, or duration. However, we expect the exact formant values to differ by condition. If face masks reduce the vowel space size (Bond et al., 1989) as predicted above, it may reduce the distance between /iː/ and /ɪ/ in the F1–F2 space. If this hypothesis is true, we predict that the formant differences between /iː/ and /ɪ/will be smaller in the face mask condition than in the control condition.

In online testing, F1 is expected to be smaller than in the laboratory without face masks (Calder et al., 2022; Zhang et al., 2021) and F2 may be smaller (Zhang et al., 2021) or larger (Calder et al., 2022); in laboratory-based testing with surgical face masks, F2 is expected to be larger than in laboratory-based testing without face masks (Georgiou, 2022a).

II. METHODS

A. Participants

Fifty-five Spanish–Basque–English trilinguals participated (41 women, M_{age} = 25.15 years, SD_{age} = 5.90 years, range 18–39 years; see Section *Statistical analyses* for sample size determination). Participants reported their ages of acquisition for each language to research assistants trained to obtain this information (Table I). Forty-eight participants acquired Spanish from birth and Basque during childhood. Five participants acquired Spanish and Basque within their first year of life, and two acquired Basque from birth and Spanish at age 1 or 2 years, respectively. All participants learned English as a foreign language through formal instruction in school and reported no active knowledge and use of other languages. Participants lived in the vicinity of Donostia–San Sebastián in the Basque Autonomous Community in Spain at the time of testing.

Participants were recruited from the Basque Center on Cognition, Brain and Language's (BCBL) participant pool. As part of the BCBL's participant pool registration process, participants complete the Basque–English–Spanish Test (BEST; de Bruin et al., 2017), which measures three proficiency components, namely vocabulary knowledge through picture naming, word recognition through lexical decisions in line with the original LexTALE (Lemhöfer and Broersma, 2012), and general language proficiency through semi-structured interviews guided by a multilingual linguist and scored on a Likert-like scale from 1 ("lowest level") to 5 ("native-like level"). Across measures, the recruited participants had ceiling proficiency in Spanish, intermediate to high proficiency in Basque and intermediate proficiency in English. Their self-reported exposure to Spanish was highest, followed by Basque and then by English (Table 1).

TABLE I. Participant characteristics.

	Spanish			Basque			English		
	M	SD	Range	M	SD	Range	M	SD	Range
AoA [a] (years)	0.05	0.30	0–2	2.83	1.80	0–9	5.75	1.90	2–10
Vocabulary (0–65)	64.84	0.54	62–65	52.24	9.56	24–65	44.38	11.12	11–63

Word recognition (% correct)	93.75	6.20	73–100	86.11	7.98	49–97	67.16	8.47	47.5–88.75
Interview (1–5)	5	0	5–5	4.04	0.74	3–5	3.13	0.55	2–4
Self-reported exposure (%)	63.64	14.45	30–90	25.82	13.57	0–60	10.38	7.13	0–30

[a] Age of acquisition.

E. Apparatus and general procedure

Participants completed four sessions. They first completed an online familiarization session in which they saw all pictures paired with their written and auditory forms to enhance naming congruence during the test phases. Next, participants completed Spanish, Basque, and English picture naming tasks (PNT; blocked by language) in three conditions, each administered in different sessions: online, on-site in the laboratory with surgical face masks, and on-site in the laboratory without face masks (hereafter, control condition). The order of sessions was counterbalanced. Within each condition, Spanish and Basque blocks were counterbalanced and the English block was always administered last. We chose this order to reflect our participants' use of Spanish and Basque (but not English) in their day-to-day interactions. We argue that the influence of Spanish and Basque on English would persist even if the English block were presented first and furthermore presenting the English block first would unnaturally influence the next languages.

The online familiarization phase and the online condition were programmed in jsPsych (de Leeuw, 2015) using the open-source study management system JATOS (Lange et al., 2015). Fifty participants reported completing the online condition on a laptop and five on a desktop computer. Forty-three reported using the microphone integrated into their laptop (as instructed) and 12 reported using an external microphone. In the on-site conditions, participants were tested individually in sound-attenuating chambers at the BCBL's satellite laboratory at the University of the Basque Country in Donostia–San Sebastián. The PNT was run on a laptop computer (HP EliteBook Folio 1040 G3) using OpenSesame software (version 3.2.7 Kafkaesque Koffka; Mathôt et al., 2012). Voice recordings were made with a Samson C01U PRO professional USB condenser microphone (Samson Technologies, Hicksville, NY) (set to 100%). In the face mask condition, participants were provided with a standard surgical face mask type IIR with bacterial filtration efficiency ≥ 98%, which they wore fully covering the mouth and nose. In all test conditions, participants' responses were saved as .wav files with a 44.1 kHz sampling rate. At the beginning of each session, participants gave informed consent. At the beginning of the online familiarization session, they performed a microphone check; at the end of the online condition, they completed a questionnaire. After the control condition, they also completed a reading aloud task for a different project. Participants were compensated with €36 paid via bank transfer or PayPal and three stamps on their fidelity card (ten stamps merit an additional gift). The BCBL's Ethics Committee approved the study.

F. Materials

The Spanish and Basque PNT included 43 words each and the English PNT included 44 words (see supplementary material[2]). The Spanish and Basque word lists comprised 24 items for VOT analysis and 30 items for vowel analysis (11 items were used for both analyses). The English word list comprised 24 items for VOT analysis and another 20 items for vowel analysis. All words were repeated once, resulting in a total of 86 Spanish and Basque productions each and 88 English productions per condition.

In each language, the 24 VOT items were composed of four items per plosive (/b/, /d/, /ɡ/, /p/, /t/, /k/). These items had a plosive–vowel onset, were one or two syllables long, and had first-syllable stress. Across languages, items were matched for the number of phonemes and syllables, and for the vowel following the plosive. As the English vowel inventory differs from the Spanish and

Basque vowel inventories and since nonnative speakers are often influenced by orthography when producing nonnative vowels (for a review: Hayes-Harb and Barrios, 2021), we mostly matched the words on the orthographic vowel (e.g., Basque “porru” /poru/ leak, Spanish “pollo” /poʎo/ chicken and English “pocket” /pɑkɪt/ were considered matched on the vowel).

The 30 Spanish and Basque vowel items were composed of ten items per corner vowel (/i/, /a/, /u/). These corner vowels appeared in the stressed position of the word. Since we did not expect the vowel space size to differ between Spanish and Basque, we measured the vowel space size for both languages combined; therefore, we did not match the vowel stimuli on any variables across languages. The 20 English vowel items consisted of 10 (near) minimal pairs between /iː/ and /ɪ/, such as “sheep” and “ship”.

Throughout, items with the lowest cognate rate possible between Spanish and English ($M_{\text{SpanishItems}} = 0.17$; $M_{\text{EnglishItems}} = 0.15$) and between Basque and English ($M_{\text{BasqueItems}} = 0.10$; $M_{\text{EnglishItems}} = 0.16$) were selected (0 = no orthographic overlap; 1 = full orthographic overlap; supplementary material[2]). Items were represented by color drawings selected from the MultiPic database (Duñabeitia et al., 2018) when they were available (Spanish 32/43; Basque 26/43; English 24/44), and the remaining pictures were selected from open content online sources.

G. Procedure

Each trial began with a fixation cross in the center of the screen for 500 ms. Afterwards, a picture appeared for 4000 ms, during which the recorder was active. There were two naming cycles in each language. Within each, the pictures were presented in random order. Participants were offered breaks in between the different language blocks. The PNT took around 25 min.

H. Analyses

A Acoustic analyses

Phonetic measurements were taken in Praat Software (version 6.1.08; Boersma and Weenink, 2021). VOT of voiced plosives was measured as the (negative) interval in milliseconds between the onset of prevoicing and the release of the burst; VOT of voiceless plosives was measured as the (positive) interval in milliseconds between the release of the burst and the onset of the following vowel. VOT measurements were determined through visual inspection of the waveform and the spectrogram viewed at 0–5000 Hz. For vowel formant analysis, the critical vowels were labeled by hand. Vowel onset and offset were defined as the first and last reliable glottal pulses with visible formants in the spectrogram viewed at 0–10 000 Hz. Afterwards, F1 and F2 were extracted at vowel midpoint using ceilings of 5500 Hz and 5000 Hz for female and male participants, respectively. Vowel duration was extracted in parallel.

B Statistical analyses

Statistical data analyses were conducted in RStudio (version IDE 2022.02.2+485; RStudio Team, 2022) run on R (version 4.2.0; R Core Team, 2022) and using the lme4 package (version 1.1-29; Bates et al., 2015). We obtained p values for t statistics through the lmerTest package (version 3.1-3; Kuznetsova et al., 2017). We used the performance package (version 0.9.0; Lüdecke et al., 2021) to check model assumptions. In linear mixed-effects models, data points with standardized residuals more than 2.5 standard deviations from 0 were removed using the LMERConvenienceFunctions package (version 3.0; Tremblay and Ransijn, 2020). Significant interactions were investigated with Bonferroni-corrected pairwise comparisons using the emmeans package (version 1.7.4-1; Lenth, 2022). Vowel formants were converted to barks with the barktools package (version 0.2.0; Stanley, 2022), which uses Traunmüller (1990) formula. Data were visualized using the ggplot2 package (version 3.3.6; Wickham, 2016). The result tables for all analyses are provided in the supplementary material[2].

A power calculation using the pwr package (version 1.3-0; Champely et al., 2020) showed that a sample of 55 participants was needed to reach 80% power for a medium effect size in the linear regression investigating changes in the Spanish/Basque vowel space by condition. A databased power calculation using the MixedPower package (version 0.1.0; Kumle et al., 2021) and data for the first 15 participants showed that a sample size of 55 provided high power for the dependent variables of the remaining research questions (i.e., 90% power for the fixed effect *Language* in the linear mixed-effects model on voiceless plosives; 99% power for the fixed effect *Language* in the logistic regression on voiced plosives; 84% power for the fixed effect *Vowel* in the linear mixed-effects model on English vowels; all based on 1000 simulations). Adding more participants did not increase power for the fixed effect *condition*, the interactions between *Language* and *condition* or *Vowel* and *condition*, which is why 55 was selected as the sample size.

III. RESULTS

I. Plosive production

There were 23 760 possible productions (55 participants × 3 languages × 3 conditions × 48 productions). Recording problems for one participant in the Spanish and English online conditions resulted in the loss of 96 trials (0.4% of the data). Another 1471 trials (453 trials in the control condition, 479 trials in the face mask condition, and 539 trials in the online condition; 6.22% of the data) were excluded from the analyses because of an incorrect response (wrong word or no response) or because VOT could not be measured reliably, for example due to background noise masking the onset of the burst and/or voicing or due to coarticulation (e.g., "um basket"), which does not allow for determining the voicing onset.

C Voiceless plosives

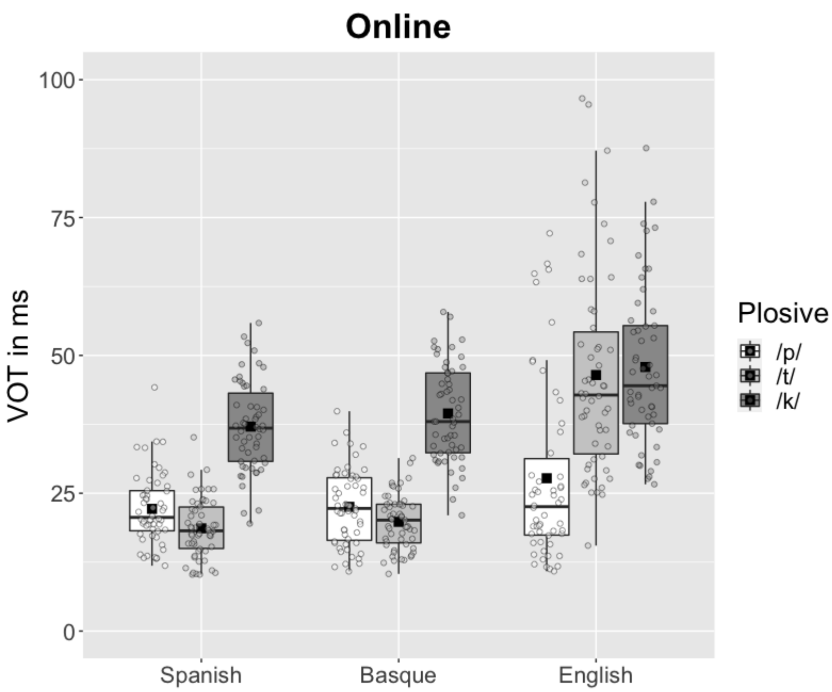
The analysis tested whether participants produced /p/, /t/, /k/ with longer VOT in English than in Spanish and Basque in all conditions. Because VOT reportedly differs by plosive (/p/</t/</k/; e.g., Volaitis and Miller, 1992; Stoehr et al., 2023 for a similar population), the factor *Plosive* was included in the model. The linear mixed-effects model had *VOT in ms* as the dependent variable with fixed effects for *Language*, *condition*, and *Plosive*, as well as an interaction between *Language* and *condition*. The model included random intercepts for *Participant* and *Item*, as well as by-*Participant* random slopes for *Language* and *condition*. No by-*Item* random slope for *condition* was included, as it caused a singularity warning [lmer formula: VOT~Language*condition+Plosive +(1+Language+condition|Participant)+(1|Item)]. 334 outliers (2.90% of the data) were removed (see Section *Statistical analyses*). We used Helmert contrast coding for the three-level categorical variable *Language* to create two contrasts of interest. The first contrast (hereafter, Language_ESB) compared the difference between English and the mean of Spanish and Basque. The second contrast (hereafter, Language_SB) compared the difference between the means of Spanish and Basque. This coding scheme provided the maximal power to test for a difference between English versus Spanish and Basque (Schad et al., 2020). We used deviation coding for the three-level variable *Plosive* to create two contrasts of interest. The first contrast (hereafter, Plosive_pt) compared /p/ [0.5] to /t/ [–0.5], and the second contrast (hereafter, Plosive_tk) compared /t/ [–0.5] to /k/ [0.5]. These two contrasts allowed us to capture the predicted VOT increase from /p/ to /t/ and from /t/ to /k/ (Volaitis and Miller, 1992). The same coding scheme was used for *condition* to compare the face mask condition [0.5] to the control condition [–0.5] (condition_Mask) and the online condition [0.5] to the control condition [–0.5] (condition_Online).

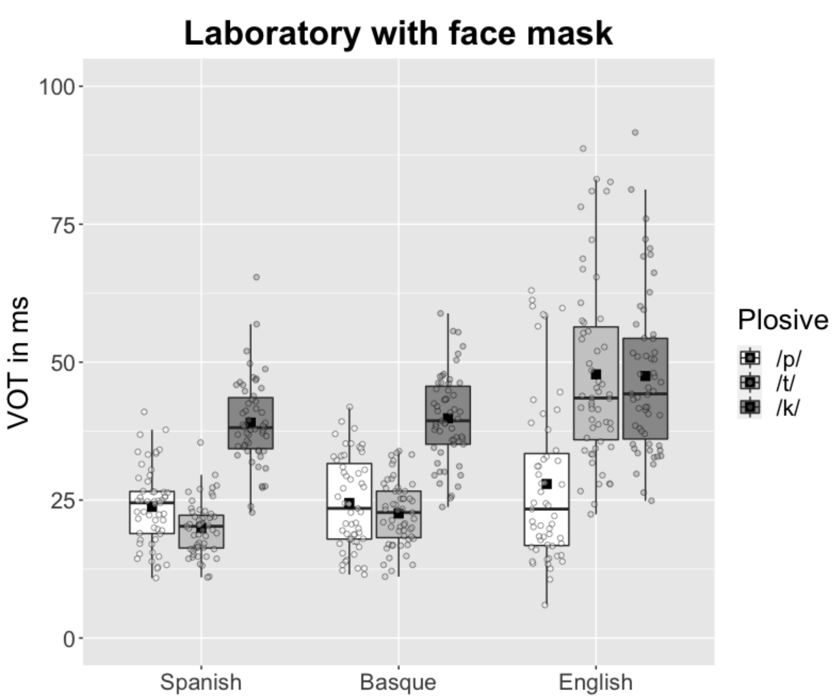
Participants produced shorter VOT for /p/ than /t/ ($\beta = -14.609$; standard error, $SE = 3.347$; $t = -4.365$; $p<0.001$) and shorter VOT for /t/ than /k/ ($\beta = 20.175$; $SE = 3.347$; $t = 6.028$; $p<0.001$).

A significant effect of *Language_ESB* showed that participants produced longer VOT in English than in Spanish and Basque (β = 13.419; *SE* = 3.198; t = 4.196; p<0.001). No significant effect of *Language_SB* was observed, suggesting that Spanish and Basque VOT were not detectably different (β = –1.476; *SE* = 2.903; t = –0.508; p = 0.615). A significant effect of *condition_Online* showed that VOT recorded online was shorter than control (β = –2.483; *SE* = 0.743; t = –3.341; p = 0.002). There was no detectable effect of *condition_Mask* (β = –0.173; *SE* = 0.507; t = –0.341; p = 0.735). The model detected significant interactions between *Language_ESB* and *condition_Mask* (β = –1.532; *SE* = 0.584; t = –2.625; p = 0.009) and *Language_ESB* and *condition_Online* (β = 1.157; *SE* = 0.589; t = 1.966; p = 0.049). No other significant interactions were observed. The results are visualized in Figure 2.

Pairwise comparisons by *condition* confirmed that English VOT was longer than Spanish and Basque VOT in all conditions (online: English vs. Spanish: β = 14.67; *SE* = 3.54; t = 4.141; p<0.001; English vs. Basque: β = 13.33; *SE* = 3.52; t = 3.791; p = 0.001; laboratory-based with face mask: English vs. Spanish: β = 13.43; *SE* = 3.54; t = 3.792; p = 0.001; English vs. Basque: β = 11.88; *SE* = 3.52; t = 3.379; p = 0.004; control: English vs. Spanish: β = 14.38; *SE* = 3.54; t = 4.060; p<0.001; English vs. Basque: β = 12.84; *SE* = 3.52; t = 3.652; p = 0.002). Pairwise comparisons by *Language* showed that English VOT was affected differently than Spanish and Basque VOT by condition. Compared to control, we found a) English VOT was shorter in both the laboratory-based condition with face mask (β = –2.050; *SE* = 0.549; t = –3.734; p<0.001) and the online condition (β = 2.308; *SE* = 0.725; t = 3.184; p = 0.006), b) Spanish and Basque VOT were shorter only in the online condition (Spanish: β = 2.600; *SE* = 0.714; t = 3.642; p = 0.001; Basque: β = 2.799; *SE* = 0.713; t = 3.927; p<0.001).

The results demonstrate that experiments conducted online and in the laboratory with surgical face masks are suitable for detecting VOT differences between aspirating and true-voicing languages in voiceless plosives. However, VOT duration is shorter when recorded online across languages compared to control. Furthermore, English aspiration is shorter when participants wear surgical face masks compared to control.





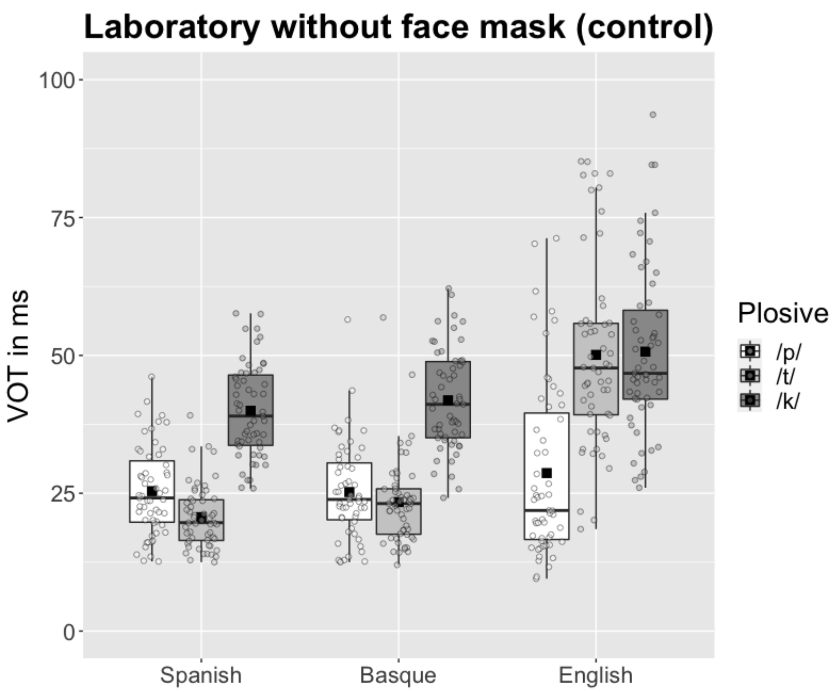


FIG. 2. VOT by language and plosive in the online (top), laboratory-based with face mask (middle), and laboratory-based without face mask (control; bottom) conditions. Each dot shows an individual participant; the black square shows the mean; the horizontal line shows the median.

D Voiced plosives

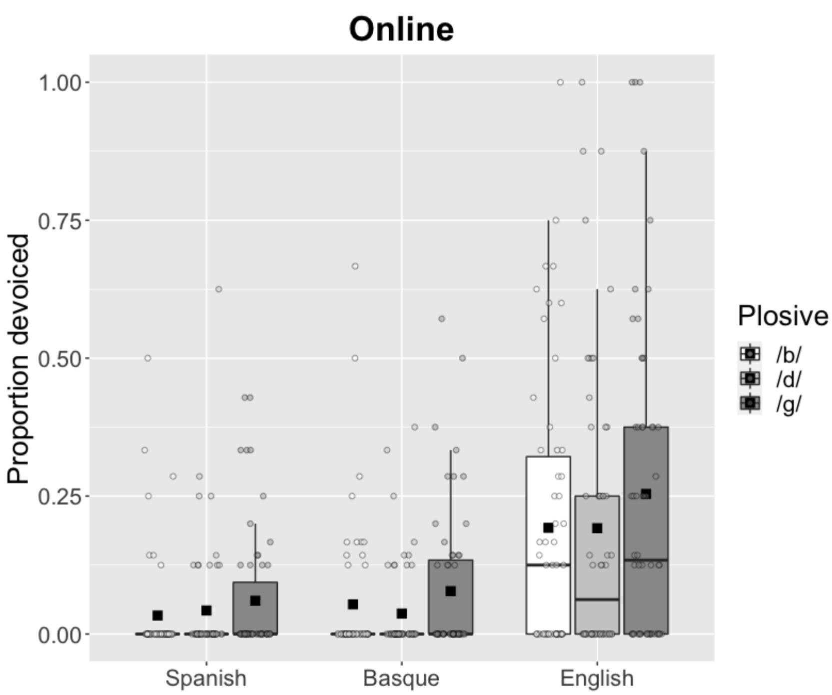
The analysis tested whether participants produced a larger proportion of /b/, /d/, /ɡ/ with positive VOT in English than in Spanish and Basque in all conditions[3]. The logistic mixed-effects model had the proportion of devoiced productions (positive VOT coded as 1; negative VOT coded as 0) as a dependent variable. The three-level variable *Plosive* was coded as two contrasts of interest. The first contrast (hereafter, Plosive_bd) compared /b/ [0.5] to /d/ [–0.5]; and the second (hereafter, Plosive_dɡ) compared /d/ [–0.5] to /ɡ/ [0.5]). The remainder of the model and the coding schemes were the same as in the model on voiceless plosives reported above [glmer formula:

[3] We also ran a linear mixed-effects model testing for prevoicing duration differences across languages and conditions. This model did not detect any significant differences in prevoicing duration (supplementary material[4]).

Proportion Devoiced~Language*condition+Plosive+(1+Language+condition|Participant)+(1|Item)].

Participants produced a larger proportion of /b/, /d/, /g/ with positive VOT in English than in Spanish and Basque (β = 2.113; *SE* = 0.274; *z* = 7.703; *p*<0.001). In addition, participants devoiced /g/ more frequently than /d/ (β = 0.694; *SE* = 0.253; *z* = 2.742; *p* = 0.006). No other significant main effects or interactions were observed. Results are visualized in Figure 3.

These results suggest that experiments conducted online and in the laboratory with surgical face masks are suitable for detecting differences in the proportion of voiced plosives produced with positive VOT between aspirating and true-voicing languages.



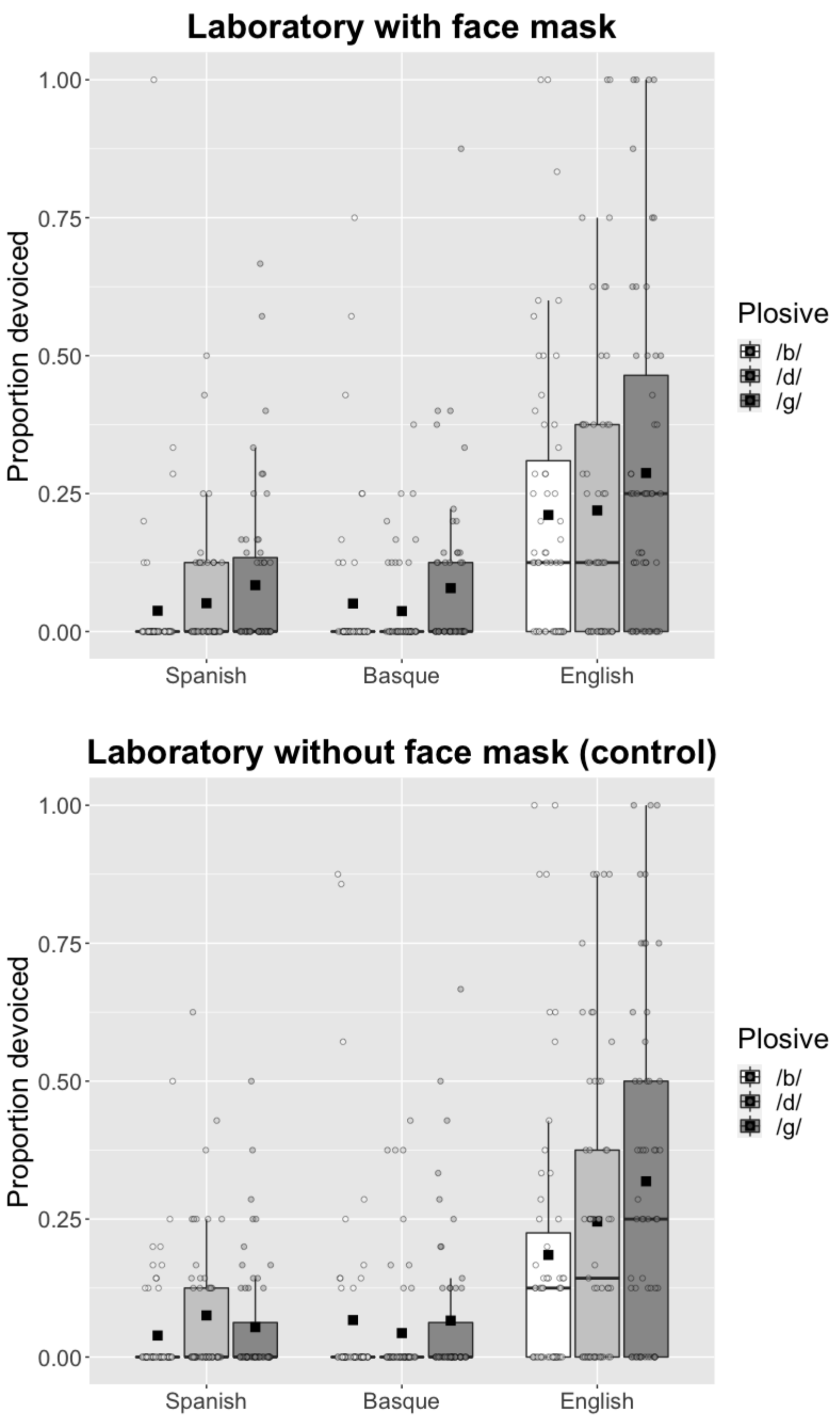


FIG. 3. Proportion of devoiced productions in the online (top), laboratory-based with face mask (middle), and laboratory-based without face mask (control; bottom) conditions. Each dot shows an individual participant; the black square shows the mean; the horizontal line shows the median.

J. Vowel production

In Spanish and Basque, there were 19 800 possible productions (55 participants × 2 languages × 3 conditions × 60 productions). In English, there were 6600 possible productions (55 participants × 2 vowels × 3 conditions × 20 productions). Recording problems for one participant in the Spanish and English online conditions resulted in loss of all 60 Spanish and all 40 English trials (1.52% of the data). Across conditions, 79 Spanish (0.80% of the data), 168 Basque (1.70% of the data), and 138 English trials (2.10% of the data) were excluded from the analyses because a participant failed to respond within the time limit, produced a wrong word or there was interference from background noise.

To exclude any formant tracking errors, we removed productions with z-scored formant values larger than 3 or smaller than –3 (Kirkham and McCarthy, 2021). This resulted in removal of 338 Spanish (3.46% of the data) and 361 Basque (3.71% of the data) trials in the vowel space analysis and 76 English trials in the F1 analysis (1.18% of the data) and 108 English trials in the F2 analysis (1.68% of the data). For all languages combined, F1 tracking errors amounted to 1.16% of the data in the online condition, 1.39% of the data in the control condition, and 1.59% of the data in the face mask condition. F2 formant tracking errors amounted to 1.80% of the data in the control condition, 2.16% of the data in the face mask condition, and 2.36% of the data in the online condition.

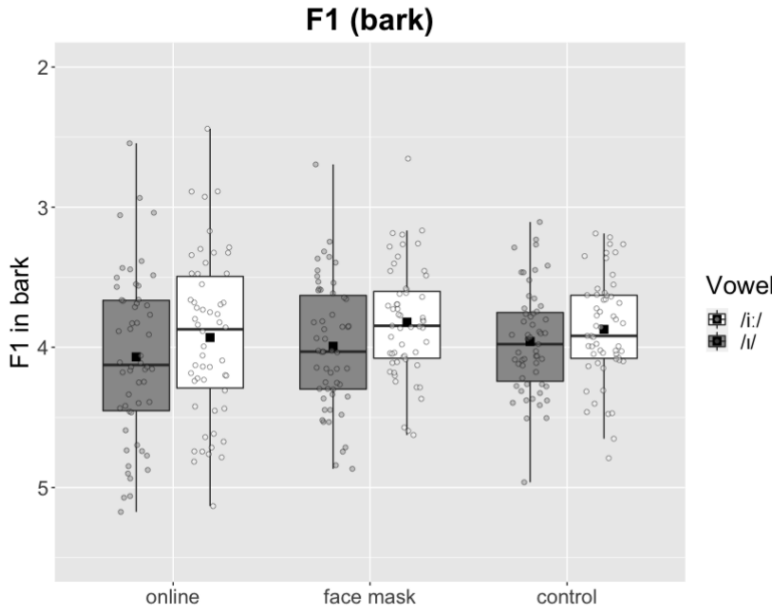
E Production of the English /iː/–/ɪ/ contrast across conditions

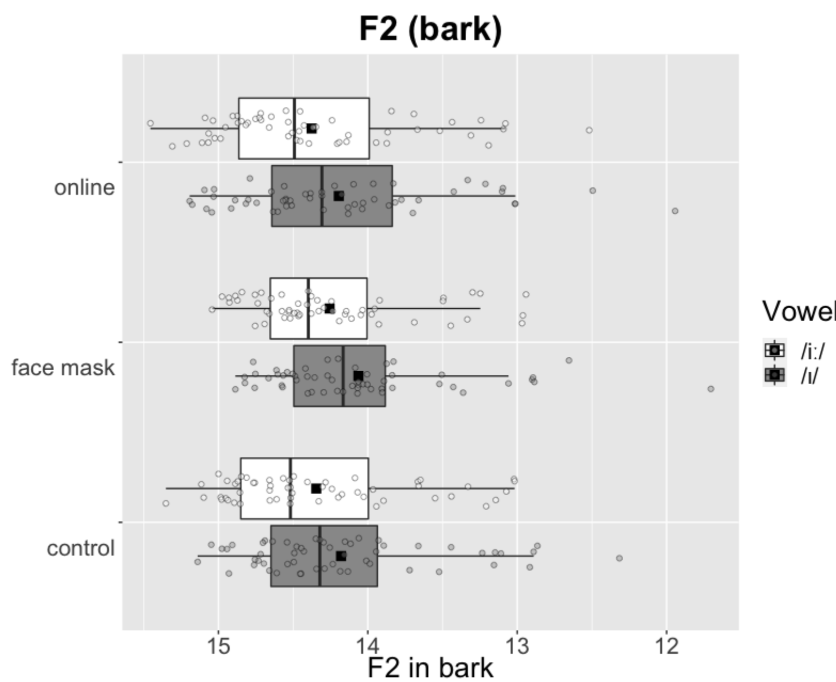
The analyses tested whether participants produced the English vowels /iː/ and /ɪ/ distinctly in all conditions. We fitted three linear mixed-effects models with *F1 (bark)*, *F2 (bark)*, and *duration (ms)* as dependent variables. All models had fixed effects for *Vowel* and *condition* as well as an interaction term between *Vowel* and *condition*. The models included random intercepts for *Participant* and *Item*, as well as by-*Participant* random slopes for *Vowel* and *condition* [lmer formula: F1/F2/Duration~Vowel*condition+(1+Vowel+condition|Participant)+(1|Item)]. We used deviation coding for the binary variable *Vowel* (/ɪ/ coded as –0.5; /iː/ coded as 0.5). As in the VOT models, we used deviation coding for *condition* to compare the laboratory-based condition with face mask [0.5] to control [–0.5] (condition_Mask) and the online condition [0.5] to control [–0.5] (condition_Online). 195 outliers (3.07% of the data) were removed in the F1 analysis, 177 (2.80% of the data) in the F2 analysis, and 114 (1.78% of the data) in the duration analysis (see Section *Statistical analyses*).

Across measures, participants produced /iː/ and /ɪ/ distinctly (Figure 4). Significant results are presented below. The F1 model detected that participants produced /iː/ with smaller F1 than /ɪ/ ($\beta = –0.133$; $SE = 0.054$; $t = –2.473$; $p = 0.018$). There was a significant interaction between *Vowel* and *condition_Mask* ($\beta = –0.084$; $SE = 0.028$; $t = –2.993$; $p = 0.003$). Pairwise comparisons by *condition* showed that the F1 difference between /iː/ and /ɪ/ was significant when participants wore a surgical face mask in the laboratory and when they were tested online but not in the control condition (laboratory-based with face mask: $\beta = 0.175$; $SE = 0.055$; $t = 3.150$; $p = 0.003$; online: $\beta = 0.140$; $SE = 0.056$; $t = 2.510$; $p = 0.016$; control: $\beta = 0.084$; $SE = 0.055$; $t = 1.518$; $p = 0.137$). The F2 model detected that participants produced /iː/ with larger F2 than /ɪ/ ($\beta = 1.797e\text{-}01$; $SE = 8.278e\text{-}02$; $t = 2.171$; $p = 0.038$). Moreover, participants produced smaller F2 in the laboratory-based condition with face mask than in control ($\beta = –1.501e\text{-}01$; $SE = 3.912e\text{-}02$; $t = –3.836$; $p<0.001$). The duration model detected that /iː/ productions were longer than /ɪ/ productions ($\beta = 20.481$; $SE = 4.675$; $t = 4.381$; $p<0.001$). In addition, vowel duration was longer in the laboratory-based condition with face mask compared to control ($\beta = 6.065$; $SE = 2.541$; $t = 2.387$; $p = 0.021$) and shorter in the online condition compared to control ($\beta = –14.953$; $SE = 3.401$; $t = –4.397$; $p<0.001$).

Overall, experiments conducted online and in the laboratory with surgical face masks are suitable for detecting small formant and duration differences in Spanish–Basque–English trilinguals' production of the /iː/–/ɪ/ contrast.

Journal of the Acoustical Society of America, 154, 152–166. https://doi.org/10.1121/10.0020064





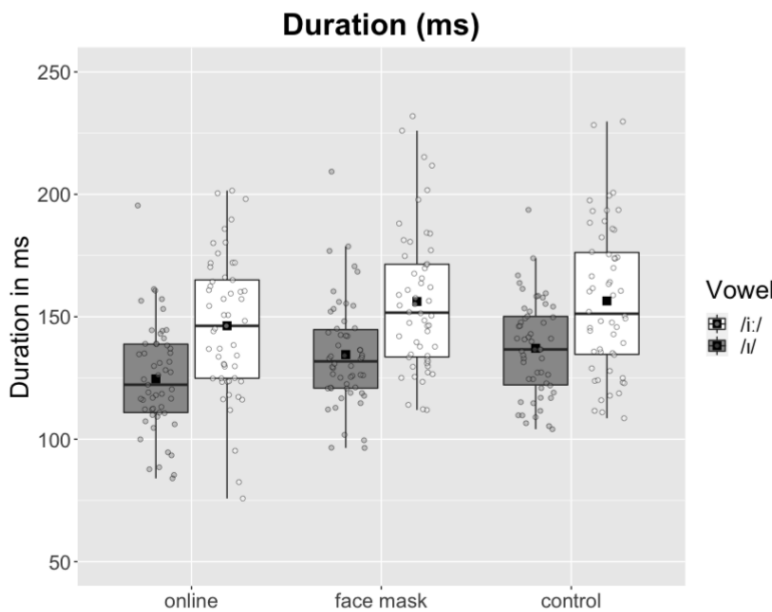


FIG. 4. F1 (top), F2 (middle), and duration (bottom) of English /iː/ and /ɪ/ by condition. Each dot shows an individual participant; the black square shows the mean; the horizontal line shows the median.

F The Spanish/Basque vowel space across conditions

This analysis tested whether the size of the Spanish/Basque vowel space differed by condition. We calculated the vowel space for each participant in each language and condition as the FCR (Sapir et al., 2010) in bark, expressed as (F2/u/+F2/a/+F1/i/+F1/u/)/(F2/i/+F1/a/). A larger FCR means that the vowel space is smaller (more centralized), and a smaller FCR means that the vowel space is larger (less centralized). We fitted a linear regression model with the *FCR in bark* as the dependent variable. The model had *condition* as fixed effect. *Condition* was deviation coded to compare the laboratory-based condition with face mask [0.5] to control [–0.5] (condition_Mask) and the online condition [0.5] to control [–0.5] (condition_Online; lm formula: FCR~condition). As noted in the Section *Vowel formants*, we did not predict the Spanish and Basque vowel spaces—which are composed of the same vowels—to be differently affected by the testing conditions. Therefore, we averaged the results across Spanish and Basque and did not include *Language* as a factor in the model. The vowel space was smaller in the laboratory-based condition with surgical face mask (β = 0.027; SE = 0.010; t = 2.683; p = 0.008) and larger in the online condition (β = –0.045; SE = 0.010; t = –4.506; p<0.001) compared to control (Figure 5).

These results show that compared to control, experiments conducted online are associated with a larger vowel space, and experiments conducted while participants wear surgical face masks are associated with a smaller vowel space.

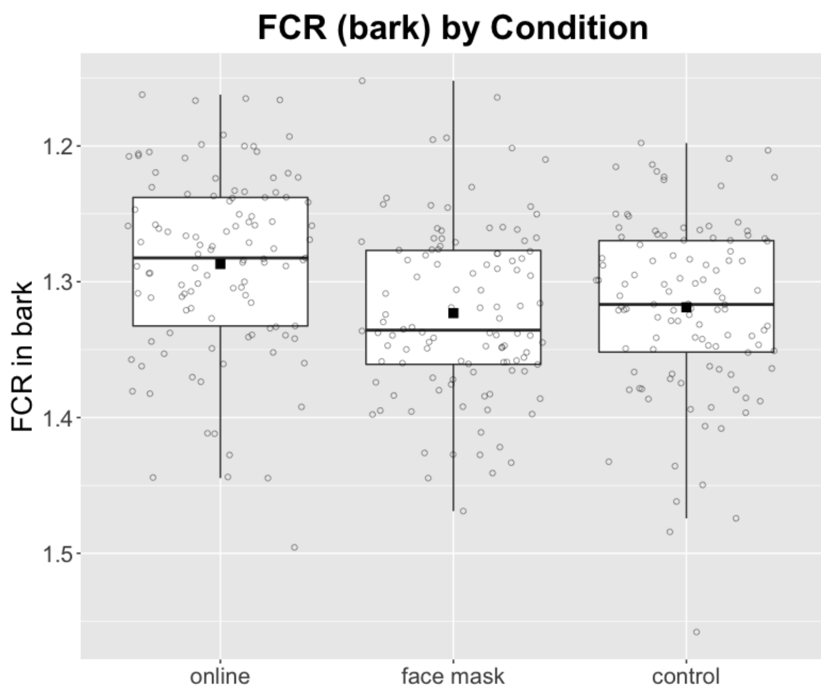


FIG. 5. FCR by condition as a measure of the Spanish/Basque vowel space size. Each dot shows an individual participant; the black square shows the mean; the horizontal line shows the median. The smaller the FCR the larger the vowel space.

IV. DISCUSSION

The present study investigated whether recording participants' speech while they wear surgical face masks in the laboratory and recording their speech online using jsPsych (de Leeuw, 2015) and

JATOS (Lange et al., 2015) are reliable options when investigating phonetic detail in speech production. To that end, we compared these two methods to speech production elicited on-site in the laboratory without face masks. We focused on phonetic detail through measures of VOT in voiceless and voiced plosives, and vowel formants in isolated words produced by Spanish–Basque–English trilingual adults.

K. Plosive production across conditions

Production differences between English versus Spanish and Basque were present in all conditions: participants produced voiceless plosives (/p/, /t/, /k/) with longer VOT in English than in Spanish and Basque and produced voiced plosives (/b/, /d/, /ɡ/) more frequently with positive VOT in English than in Spanish and Basque, thus confirming our predictions. As such, testing participants in the laboratory when surgical face masks are required or testing them online are suitable options for investigating crosslinguistic differences in plosive production.

However, the exact VOT duration of voiceless plosives differed by condition. In the present study, wearing surgical face masks reduced VOT duration of English—but not Spanish and Basque—voiceless plosives by 2 ms on average compared to control. This finding shows that surgical face masks specifically affect the duration of aspiration (in English voiceless plosives) but not the duration of voiceless plosives in general, as Spanish and Basque short lag voiceless plosives were not affected by participants wearing face masks. This is likely because surgical face masks are positioned close to the lips and act like a physical barrier to the aspiration air stream passing through the lips. Short lag VOT as common in Spanish and Basque is likely too short to be affected by this physical barrier. Therefore, the finding that surgical face masks reduce aspiration duration in English but not short lag VOT in Spanish and Basque appears to be reflective of the phonetic characteristics of English as an aspirating language and Spanish/Basque as true-voicing languages rather than being the result of differences in language proficiency between languages. Importantly, this shortening of

English aspirated VOT in the face mask condition did not affect the crosslinguistic VOT difference between English and Spanish/Basque, which remained significant.

In online testing, across languages and relative to the control condition (in the laboratory without a face mask) VOT in voiceless plosives was on average 3 ms shorter. Participants were likely more relaxed during the online session conducted in their homes, which may have led to more natural—and thus more representative—VOT production. The formal laboratory environment may have imposed more pronunciation effort, thus leading to longer VOT production than in the online condition. Previous research partly supports this assumption (Robb et al., 2005): native English speakers produced longer syllable durations in speech recorded in the laboratory compared to speech recorded outside the laboratory; however, VOT duration for these speakers did not statistically differ by environment. It is possible that the formal laboratory setting affected syllable and VOT durations but that the relatively small sample size of 20 was not sufficient to detect small VOT differences by environment (3 ms in the present study) in Robb et al.

We did not find evidence for our prediction that the proportion of voiced plosives produced with prevoicing would differ between the online and control conditions. We predicted that since prevoicing is a subtle acoustic signal, the uncontrolled environment and recording devices in the online condition would not capture the presence of prevoicing as reliably as the professional recorder and environment in the control condition. As we did not detect any differences between the online and control conditions in the proportion of voiced plosives produced with prevoicing, the present results are encouraging for online testing, showing that this uncontrolled environment is suitable for recording subtle acoustic signal differences.

In sum, our data show that speech recordings made online and in the laboratory when participants wear surgical face masks are suitable when investigating VOT production in voiceless and voiced plosives of multilinguals speaking true-voicing and aspirating languages.

L. Production of the English /iː/–/ɪ/ vowel contrast across conditions

Participants produced the English vowels /iː/ and /ɪ/ with distinct F1, F2, and duration, with /iː/ having a higher (smaller F1) and more frontal (larger F2) position and longer duration than /ɪ/, thus confirming our predictions and supporting previous findings that native speakers of languages lacking the /iː/–/ɪ/ contrast produce these vowels with distinct F2 (Georgiou, 2022b) and/or duration (Cebrian, 2007; Cebrian et al., 2021; Georgiou, 2022b). Unlike previous studies, the present study—with its larger participant sample (55 in the present study, 10 in Georgiou, 2022b; 30 in Cebrian, 2007; 43 in Cebrian et al., 2021)—detected production differences between /iː/ and /ɪ/ across all three measures. This may be attributed to the present study being well-powered and thus able to detect small spectral and temporal differences in vowel production.

Against our prediction, the F1 difference between /iː/ and /ɪ/ was larger when participants wore a face mask (mean difference 0.176 bark) compared to control (mean difference 0.096 bark). In fact, a post hoc test failed to find an F1 difference between /iː/ and /ɪ/ in the control condition. We speculate that participants compensated for the communicative restrictions imposed by the face mask by hyperarticulating, which may have enhanced the F1 difference between /iː/ and /ɪ/ when participants wore face masks. Our finding that participants produced both vowels with longer duration when they wore face masks compared to control supports the hyperarticulation assumption. Previous work reporting enhanced intelligibility of face-masked speech (Cohn et al., 2021; Pycha et al., 2022; Zellou et al, 2023) further supports that people may be hyperarticulating when wearing face masks. Overall, when wearing surgical face masks, participants produced both /iː/ and /ɪ/ with smaller F2, corresponding to a more posterior position compared to control. This finding is against our prediction, which was based on Georgiou's (2022a) finding that participants produce /iː/ with numerically larger F2 when they wear surgical face masks. However, although Georgiou showed that wearing surgical face masks affects vowel production, his results varied by vowel. Some Cypriot Greek vowels were produced with larger F2 (significant difference for /e/ & /u/; numerical difference for /i/ & /o/) and others with smaller F2 (/a/) relative to a control measure without face masks. The more posterior position observed in the present study may result from the surgical face mask acting as a physical barrier at the front of the mouth, thus pushing the position of the front vowels /iː/ and /ɪ/ to a slightly posterior position. This is in line with research reporting a reduced vowel space size when participants wear oxygen face masks (Bond et al., 1989). A reduced vowel space size means that the F2 of front vowels, such as /iː/ and /ɪ/, becomes smaller, which is what we observed in the present study.

Finally, we observed a shorter vowel duration in the online condition compared to control but neither F1 nor F2 differed between the online and control conditions. The shorter vowel duration in online testing is in line with the shorter VOT duration of voiceless plosives in all languages in online testing discussed above and provides further support for our argument that the formal laboratory environment imposed more pronunciation effort than online testing, which may affect temporal properties of speech production. The lack of detectable F1 or F2 differences between the online and control conditions was unpredicted given the Calder et al. (2022) and Zhang et al. (2021) findings of smaller F1 and smaller F2 (Zhang et al., 2021) or larger F2 (Calder et al., 2022) in vowels recorded online using the Zoom cloud meeting application. These differences between the present study and the Calder et al. and Zhang et al. studies may be related to the different online testing tools used in the present study (jsPsych/JATOS) and in Calder et al. and Zhang et al. (Zoom cloud meeting application). Importantly, the Zoom cloud meeting application as used by Calder et al. and Zhang et al. had a different sampling rate (32 kHz) than their in-person recording devices (44.1 kHz), which may have contributed to their observed differences between recording conditions. In the present study, both online and laboratory-based recordings were made at 44.1 kHz, and it is possible that these identical recording settings minimized between-condition differences.

Our data suggest that speech recordings made online and in the laboratory when participants wear surgical face masks are suitable when investigating the production of the nonnative /iː/–/ɪ/ contrast.

M. The Spanish/Basque vowel space size across conditions

Relative to the control condition, participants' vowel space was smaller when tested in the face mask condition and larger when tested in the online condition. The reduced vowel space in the face mask condition was predicted given the previous research finding that wearing oxygen face masks was associated with a smaller vowel space (Bond et al., 1989). The reason for this smaller vowel space may be due to the face mask restricting the jaw (and consequently F1) and the length of the vocal tract (and consequently F2). The assumption of face masks being associated with decreased F2 is also in line with our finding that participants produced the English front vowels /iː/ and /ɪ/ with smaller F2 in the face mask condition, indicating a more posterior place of articulation. A reduced vowel space is associated with less clear and less intelligible speech (Bradlow and Bent, 2002). Our finding of a smaller vowel space when participants wear surgical face masks, therefore, directly relates to previous research, which found that speech produced with face masks may be less intelligible than speech produced without face masks (Atcherson et al., 2017; Corey et al., 2020; Goldin et al., 2020; Magee et al., 2020).

The larger vowel space in the online condition was not unexpected, as we assumed that the vowel space size differs between speech recorded online and control. However, given the lack of previous research on this topic, we were unable to predict whether online testing would result in a smaller or larger vowel space. We assumed that the driving force behind differences in vowel space size between online testing and control may be related to the use of various recording devices in online testing. When examining Figure 4, which shows the production of English /iː/ and /ɪ/, there appears to be considerably more variability between participants' F1 production recorded online compared to control. In the vowel space size analysis, however, we observe similar between-participant variability in the online and control conditions (Figure 5). This may be because we measured the vowel space size as the FCR, a measure which reduces between-participant variability (Sapir et al., 2010). Therefore, the larger vowel space in the online condition does not appear to result from greater between-participant variability. To test whether the larger vowel space in online testing may result from larger within-participant variability, which may have pushed the formant means to more extreme positions, we computed compactness scores for each vowel by condition and participant. These compactness scores were computed as the standard deviation of the mean of F1 multiplied by the standard deviation of the mean of F2 multiplied by π, assuming that vowel categories are elliptical (Kartushina and Frauenfelder, 2014). Surprisingly, vowels in the online condition were the most compact, followed by the control condition and face mask condition ($M_{\text{Online\_Compactness}} = 1.20$; $M_{\text{Control\_Compactness}} = 1.35$; $M_{\text{FaceMask\_Compactness}} = 1.44$). It appears, then, that the larger vowel space size in the online condition (relative to control) does not emerge from between- or within-participant variability. As an alternative explanation, we propose that participants may have experienced more psychological stress in the formal laboratory environment than when they performed the online experiment in their homes. Psychological stress has been found to be associated with a smaller vowel space size (Karlsson et al., 2000). The present study did not include measures of the L3-English vowel space and future research can investigate if the vowel space in a language with relatively low proficiency is similarly or even more strongly affected by differences in the testing environment as the native language(s). If psychological stress is the driving force behind a reduced vowel space in laboratory-based testing, it is possible that the reduction is even larger in a low(er) proficiency language because any stress level is likely enhanced by having to speak in a less proficient language. However, there is no evidence that the hypothesized psychological stress affects the temporal property VOT differently across languages, as we observed longer VOT in the control condition than in the online condition in Spanish, Basque, and English alike.

In summary, online testing in the home environment without an experimenter may have led to more natural speech production resulting in shorter VOT duration in all languages, shorter vowel duration of English /iː/ and /ɪ/, and a larger but more representative vowel space size than observed in the formal control condition conducted in the laboratory.

V. CONCLUSIONS

Testing participants in the laboratory while they wear surgical face masks or recording their speech online appear to be valid options when investigating phonetic detail in trilinguals' speech production. Across conditions, we observed the predicted phonetic differences between trilinguals' languages or within trilinguals' least proficient language. However, small phonetic differences emerged between conditions. Wearing surgical face masks was associated with shorter aspiration in English voiceless plosives, a larger F1 difference between English /iː/ and /ɪ/, smaller F2 and longer duration in English /iː/ and /ɪ/, and a smaller Spanish/Basque vowel space, all compared to control. When participants wear surgical face masks, two competing forces appear to be at play. On the one hand, surgical face masks shorten the vocal tract and restrict the articulators. A shortened vocal tract can explain the observed shorter VOT in English (aspirated) voiceless plosives and the lower F2 in the English vowels /iː/ and /ɪ/. The combination of a shortened vocal tract and restriction of the articulators can also explain the smaller vowel space in Spanish/Basque. On the other hand, participants seem to compensate for the limitations imposed by surgical face masks by hyperarticulating, which can explain the larger F1 difference and longer duration in English vowels when participants wore face masks.

Online testing was associated with shorter VOT in voiceless plosives in Spanish, Basque, and English, shorter vowel duration in English /iː/ and /ɪ/, and a larger Spanish/Basque vowel space, all compared to control. Overall, online testing may make participants feel more at ease, resulting in a more natural—and more ecologically valid—speaking style, which may have led to shorter VOT in voiceless plosives, shorter vowel duration, and a larger vowel space compared to control. Future studies still need to investigate how masked and online studies might differentially affect languages with different properties (e.g., different vowel space density). Nevertheless, we conclude that testing trilinguals' production of isolated words while they wear surgical face masks in the laboratory or record their speech online using jsPsych (de Leeuw, 2015) and JATOS (Lange et al., 2015) are suitable options for within-participant designs.

ACKNOWLEDGMENTS

This work was supported by institutional grants from the Basque Government [BERC 2022–2025 program] and the Spanish State Research Agency [BCBL Severo Ochoa excellence accreditation CEX2020-001010/AEI/10.13039/501100011033] awarded to the BCBL. This project has also received funding from the European Union's H2020 research and innovation program [Marie Skłodowska-Curie grant agreement No 843533 awarded to AS]; the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program [grant agreement No 819093 to CDM]; the Spanish State Research Agency [BES-2017-082500 to CS; PID2020-113926GB-I00 to CDM; PID2021-123578NA-I00/AEI/10.13039/501100011033/FEDER, UE, & FJC2020-044978-I to AS]; and by the Basque Government's Department of Education [Predoctoral training program for research staff PRE_2021_2_0006 awarded to TT].

AUTHOR DECLARATIONS

Conflict of Interest

We have no conflicts to disclose.

Ethics Approval

The present study has been approved by the Basque Center on Cognition, Brain and Language's Ethics Committee prior to data collection (approval code 230222ML). Informed consent for participation and collection of speech data were obtained from all participants.

DATA AVAILABILITY

The data that support the findings of this study are openly available in the Open Science Framework at http://doi.org/10.17605/OSF.IO/XYH3K.

[1]At least two Basque varieties spoken in France employ aspiration (Zuberoan: Gaminde et al., 2002; Mounole, 2004; Mixean: Egurtzegi and Carignan, 2020). Here, we focus on Standard Basque spoken in Gipuzkoa/Spain, for which no aspiration has been found (Souganidis et al., 2022).

[2]See supplementary material at https://doi.org/10.1121/10.0020064 for stimulus materials; for cognate rate measures; for results tables; and for linear mixed-effects model on prevoicing duration.

[3]We also ran a linear mixed-effects model testing for prevoicing duration differences across languages and conditions. This model did not detect any differences in prevoicing duration (supplementary material2).

REFERENCES

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., and Evershed, J. K. (**2020**). "Gorilla in our midst: An online behavioral experiment builder," Behav. Res. **52**, 388–407.

Asadi, S., Cappa, C. D., Barreda, S., Wexler, A. S., Bouvier, N. M., and Ristenpart, W. D. (**2020**). "Efficacy of masks and face coverings in controlling outward aerosol particle emission from respiratory activities." Sci. Rep. **10**, 15665.

Journal of the Acoustical Society of America, 154, 152–166. https://doi.org/10.1121/10.0020064

Atcherson, S. R, Mendel, L. L., Baltimore, W. J., Patro, C., Lee, S., Pousson, M., and Spann, M. J. (**2017**). "The effect of conventional and transparent surgical masks on speech understanding in individuals with and without hearing loss," J. Am. Acad. Audiol. **28**, 58–67.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Softw. **67**, 1–48.

Boersma, P., and Weenink, D. (**2019**). "Praat: Doing phonetics by computer (Version 6.1.08) [computer program]," http://www.praat.org/.

Boersma, P., and Weenink, D. (**2021**). "Praat: Doing phonetics by computer (Version 6.1.40) [computer program]," http://www.praat.org/.

Bond, Z. S., Moore, T. J., and Gable, B. (**1989**). "Acoustic-phonetic characteristics of speech production in noise and while wearing an oxygen mask," J. Acoust. Soc. Am. **85**, 907–912.

Bradlow, A. R., and Bent, T. (**2002**). "The clear speech effect for non-native listeners," J. Acoust. Soc. Am. **112**, 272–284.

Brysbaert, M. (**2021**). "Power considerations in bilingualism research: Time to step up our game," Bilingualism **24**, 813–818.

Bulgin, J., De Decker, P., & Nycz, J. (**2010**). "Reliability of formant measurements from lossy compressed audio," in British Association of Academic Phoneticians Colloquium, March 29–31, London, UK. Available at https://research.library.mun.ca/684/1/Bulgin_De_Decker_Nycz_2010.pdf

Cabrelli Amaro, J., and Wrembel, M. (**2016**). "Investigating the acquisition of phonology in a third language - a state of the science and an outlook for the future," Int. J. Multiling. **13**, 395–409.

Calder, J., Wheeler, R., Adams, S., Amarelo, D., Arnold-Murray, K., Bai, J., Church, M., Daniels, J., Gomez, S., Henry, J., Jia, Y., Johnson-Morris, B., Lee, K., Miller, K., Powell, D., Ramsey-Smith, C., Rayl, S., Rosenau, S., and Salvador, N. (**2022**). "Is Zoom viable for sociophonetic research? A comparison of in-person and online recordings for vocalic analysis. Ling. Vanguard 20200148.

Cebrian, J. (**2007**). "Old sounds in a new contrast: L2 production of the English tense-lax vowel distinction," Proc. 16th Int. Congress Phonetic Sci. (pp. 1637–1640).

Cebrian, J., Gorba, C., and Gavaldà, N. (**2021**). "When the easy becomes difficult: Factors affecting the acquisition of the English /iː/-/ɪ/ contrast," Front. Commun. **6**, 660917.

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., and De Rosario, H. (**2020**). "Package 'pwr: Basic functions for power analysis (Version 1.3-0)," https://CRAN.R-project.org/package = pwr

Cohn, M., Pycha, A., and Zellou, G. (**2021**). "Intelligibility of face-masked speech depends on speaking style: Comparing casual, clear, and emotional speech," Cognition **210**, 104570.

Corey, R. M., Jones, U., and Singer, A. C. (**2020**). "Acoustic effects of medical, cloth, and transparent face masks on speech signals," J. Acoust. Soc. Am. **148**, 2371–2375.

de Bruin, A., Carreiras, M., and Duñabeitia, J. A. (**2017**). "The BEST dataset of language proficiency," Front. Psy. **8**, 522.

de Leeuw, J. R. (**2015**). "jsPsych: A JavaScript library for creating behavioral experiments in a web browser," Behav. Res. **47**, 1–12.

de Leeuw, J. R., and Motz, B. A. (**2016**). "Psychophysics in a web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task," Behav. Res. **48**, 1–12.

Duñabeitia, J.A., Crepaldi, D., Meyer, A.S., New, B., Pliatsikas, C., Smolka, E., and Brysbaert, M. (**2018**). "MultiPic: A standardized set of 750 drawings with norms for six European languages," Quarterly J. Exp. Psy. **71**, 808–816.

Egurtzegi, A., and Carignan, C. (**2020**). "An acoustic description of Mixean Basque," J. Acoust. Soc. Am. **147**, 2791–2802.

Fairs, A., and Strijkers, K. (**2021**). "Can we use the internet to study speech production? Yes we can! Evidence contrasting online versus laboratory naming latencies and errors," PLoS One **16**, e0258908.

Flege, J. E. (**1987**). "The production of 'new' and 'similar' phones in a foreign language: Evidence for the effect of equivalence classification," J. Phon. **15**, 47–64.

Flege, J. E. (**1991**). "Age of learning affects the authenticity of voice-onset time (VOT) in stop consonants produced in a second language," J. Acoust. Soc. Am. **89**, 395–411.

Flipsen, P., and Lee, S. (**2012**). "Reference data for the American English acoustic vowel space," Clin. Ling. Phon. **26**, 926–933.

Fox, R. A., and Jacewicz, E. (**2017**). "Reconceptualizing the vowel space in analyzing regional dialect variation and sound change in American English," J. Acoust. Soc. Am. **142**, 444–459.

Freeman, V., and De Decker, P. (**2021**). "Remote sociophonetic data collection: Vowels and nasalization over video conferencing apps," J. Acoust. Soc. Am. **149**, 1211–1223.

Gaminde, I., Hualde, J. I., and Salaberria, J. (**2002**). "Zubereraren herskariak: Azterketa akustikoa," ("Zuberoa's plosives: An acoustic study"), Lapurdum **7**, 221–236.

Geiss, M., Gumbsheimer, S., Lloyd-Smith, A., Schmid, S., and Kupisch, T. (**2022**). "Voice onset time in multilingual speakers: Italian heritage speakers in Germany with L3 English," Stud. Second Lang. Acquis. **44**, 435–459.

Georgiou, G. P. (**2022a**). "Acoustic markers of vowels produced with different types of face masks," Appl. Acoust. **191**, 108691.

Georgiou, G. P. (**2022b**). "The acquisition of /ɪ/–/iː/ is challenging: Perceptual and production evidence from Cypriot Greek speakers of English," Behav. Sci. **12**, 469.

Journal of the Acoustical Society of America, 154, 152–166. https://doi.org/10.1121/10.0020064

Goldin, A., Weinstein, B., and Shiman, N. (**2020**). "How do medical masks degrade speech perception?," Hearing Rev. **27**, 8–9.

Hansen Edwards, J. H., and Zampini, M. L. (**2008**). Phonology and Second Language Acquisition (John Benjamins, Philadelphia).

Hayes-Harb, R., and Barrios, S. (**2021**). "The influence of orthography in second language phonological acquisition," Lang. Teach. **54**, 297–326.

Hilbig, B. B. (**2016**). "Reaction time effects in lab- versus web-based research: Experimental evidence," Behav. Res. **48**, 1718–1724.

Hualde, J. I. (**1991**). Basque Phonology (Routledge, London).

Karlsson, I., Banziger, T., Dankovicová, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F. and Scherer, K. (**2000**). "Speaker verification with elicited speaking styles in the VeriVox project," Speech Commun. **31**, 121–129.

Kartushina, N., and Frauenfelder, U. H. (**2014**). "On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation," Front. Psychol. **5**, 1246.

Kirkham, S., and McCarty, K. M. (**2021**). "Acquiring allophonic structure and phonetic detail in a bilingual community: The production of laterals by Sylheti-English bilingual children," Int. J. Bilingual. **25**, 531–547.

Kumle, L., Võ, M. L., and Draschkow, D. (**2021**). „Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R," Behav. Res. **53**, 2528–2543.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2017**). "lmerTest Package: Tests in linear mixed effects models," J. Stat. Softw. **82**, 1–26.

Ladefoged, P. and Johnson, K. (**2010**). A Course in Phonetics (Cengage Learning, Boston).

Lalonde, K., and Werner, L. A. (**2019**). "Infants and adults use visual cues to improve detection and discrimination of speech in noise," J. Speech Lang. Hear. Res. **62**, 3860–3875.

Lange, K., Kühn, S., and Filevich, E. (**2015**). "“Just another tool for online studies” (JATOS): An easy solution for setup and management of web servers supporting online studies," PLoS One **10**, e0130834.

Lemhöfer, K., and Broersma, M. (**2012**). "Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English," Behav. Res. **44**, 325–343.

Lenth, R. (**2022**). "emmeans: Estimated marginal means, aka least-squares means (Version 1.7.4-1)," https://CRAN.R-project.org/package = emmeans

Lisker, L., and Abramson, A. (**1964**). "A cross-language study of voicing in initial stops: Acoustical measurements," Word **20**, 384–422.

Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P., and Makowski, D. (**2021**). "Performance: An R package for assessment, comparison and testing of statistical models," JOSS **6**, 3139.

Magee, M., Lewis, C., Noffs, G., Reece, H., Chan, J. C. S., Zaga, C. J., Paynter, C., Birchall, O., Rojas Azocar, S., Ediriweera, A., Kenyon, K., Caverlé, M. W., Schultz, B. G., and Vogel, A. P. (**2020**). "Effects of face masks on acoustic analysis and speech perception: Implications for peri-pandemic protocols," J. Acoust. Soc. Am. **148**, 3562–3568.

Mathôt, S., Schreij, D., and Theeuwes, J. (**2012**). "OpenSesame: An open-source, graphical experiment builder for the social sciences," Behav. Res. **44**, 314–324.

Mounole, C. (**2004**). "Zubererazko herskarien azterketa akustikoa" ("The acoustic analysis of the plosives of Zuberoa"), Anuario Del Seminario De Filología Vasca "Julio De Urquijo" **38** 207–248.

Osborne, D. M., and Simonet, M. (**2021**). "Foreign-language phonetic development leads to first-language phonetic drift: Plosive consonants in native Portuguese speakers learning English as a foreign language in Brazil," Languages **6**, 112.

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

Pettinato, M., Tuomainen, O., Granlund, S., and Hazan, V. (**2016**). "Vowel space area in later childhood and adolescence: Effects of age, sex and ease of communication," J. Phon. **54**, 1–14.

Piazza, G., Martin, C. D., and Kalashnikova, M. (**2022**). "The acoustic features and didactic function of foreigner-directed speech: A scoping review," J. Speech Lang. Hear. Res. **65**, 2896–2918.

Pierrehumbert, J. B., Bent, T., Munson, B., Bradlow, A. R., and Bailey, J. M. (**2004**). "The influence of sexual orientation on vowel production," J. Acoust. Soc. Am. **116**, 1905–1908.

Pycha, A., Cohn, M., and Zellou, G. (**2022**). "Face-masked speech intelligibility: the influence of speaking style, visual information, and background noise," Front. Commun. **7**, 874215.

R Core Team (**2022**). R: A language and environment for statistical computing (Version 4.2.0) (R Foundation for Statistical Computing, Vienna, Austria).

Rattanasone, N. X., Burnham, D., and Reilly, R. G. (**2013**). "Tone and vowel enhancement in Cantonese infant-directed speech at 3, 6, 9, and 12 months of age," J. Phon. **41**, 332–343.

Robb, M., Gilbert, H., and Lerman, J. (**2005**). "Influence of gender and environmental setting on voice onset time," Folia Phon. Logopaed. **57**, 123–133.

RStudio Team (**2022**). RStudio: Integrated Development Environment for R (Version IDE 2022.02.2+485) (RStudio, Boston).

Saeidi, R., Niemi, T., Karppelin, H., Pohjalainen, J., Kinnunen, T., & Alku, P. (**2015**). "Speaker recognition for speech under face cover," in Proceedings 16th Annual Conference of the International Speech Communication Association, September 6–10, Dresden, Germany, pp. 1012–1016.

Journal of the Acoustical Society of America, 154, 152–166. https://doi.org/10.1121/10.0020064

Sapir, S., Ramig, L. O., Spielman, J. L., and Fox, C. (**2010**). “Formant Centralization Ratio (FCR): A proposal for a new acoustic measure of dysarthric speech,” J. Speech Lang. Hear. Res. **53**, 114–125.

Schad, D. J., Vasishth, S., Hohenstein, S., and Kliegl, R. (**2020**). “How to capitalize on a priori contrasts in linear (mixed) models: A tutorial,” J. Memory Lang. **110**, 104038.

Shue, Y.-L. (**2010**). “The voice source in speech production: Data, analysis and models,” PhD thesis, University of California, Los Angeles.

Skodda, S., Grönheit, W., and Schlegel, U. (**2012**). “Impairment of vowel articulation as a possible marker of disease progression in Parkinson’s Disease,” PLoS One **7**, e32132.

Souganidis, C., Molinaro, N., and Stoehr, A. (**2022**). “Bilinguals produce language-specific voice onset time in two true-voicing languages: The case of Basque-Spanish bilinguals,” Ling. Appr. Bilingualism (published online).

Stanley, J. (**2022**). “barktools: Functions to help when working with Barks (Version 0.2.0),” http://joeystanley.github.io/barktools

Stoehr, A., Benders, T., van Hell, J. G., and Fikkert, P. (**2017**). “Second language attainment and first language attrition: The case of VOT in immersed Dutch–German late bilinguals,” Second Lang. Res. **33**, 483–518.

Stoehr, A., Jevtović, M., de Bruin, A., and Martin, C. D. (**2023**). “Phonetic and lexical crosslinguistic influence in early Spanish-Basque-English trilinguals”, Lang. Learn. (published online).

Toscano, J. C., and Toscano, C. M. (**2021**). “Effects of face masks on speech recognition in multi-talker babble noise,” PLoS One **16**, e0246842.

Traunmüller, H. (**1990**). “Auditory scales of frequency representation,” J. Acoust. Soc. Am. **88**, 97–100.

Journal of the Acoustical Society of America, 154, 152–166. https://doi.org/10.1121/10.0020064

Tremblay, A., and Ransijn, J. (**2020**). "Package 'LMERConvenienceFunctions'. Model selection and post-hoc analysis for (G)LMER models (version 3.0)," https://CRAN.R-project.org/package = LMERConvenienceFunctions

Vogt, A., Hauber, R., Kuhlen, A. K., and Rahman, R. A. (**2022**). "Internet-based language production research with overt articulation: Proof of concept, challenges, and practical advice," Behav. Res. **54**, 1954–1975.

Volaitis, L. E., and Miller, J. L. (**1992**). "Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories," J. Acoust. Soc. Am. **92**, 723–735.

Wickham, H. (**2016**). ggplot2: Elegant Graphics for Data Analysis (Springer, New York).

Zellou, G., Pycha, A., and Cohn, M. (**2023**). "The perception of nasal coarticulatory variation in face-masked speech," J. Acoust. Soc. Am. **153**, 1084–1093.

Zhang, C., Jepson, K., Lohfink, G., and Arvaniti, A. (**2021**). "Comparing acoustic analyses of speech data collected remotely," J. Acoust. Soc. Am. **149**, 3910–3916.

Zwicker, E. (**1961**). "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," J. Acoust. Soc. Am. **33**, 248.