COPiL CAMBRIDGE OCCASIONAL PAPERS *in* LINGUISTICS
Volume **18**, Article 2: pp. 24–56, 2026 | ISSN: 2050-5949

Audio Visual Data Show that Face Coverings Have Minimal Impact on Speech Comprehension and Do Not Affect Speaker Social Evaluations*

Chloe Patman
University of Cambridge

Kirsty McDougall
University of Cambridge

Paul Foulkes
University of York

Abstract This study explored the perceptual consequences of speaking with a face covering. Our experiment tested listeners' perceptions of audio-visual stimuli presenting female English speakers in niqāb, cotton mask, and no face covering conditions. Part 1 assessed minimal pair word identification via a binary choice test where stimuli were mixed with background noise. Stimuli were either matched (audio and video components from the same recording), or mismatched (video components with no face covering matched with niqāb or cotton mask audio, or vice versa). Part 2 evaluated perceived social traits like friendliness, trustworthiness and intelligibility. Participants with varying levels of exposure to face coverings were compared. The word identification results revealed a surprising interaction whereby combining niqāb audio and niqāb visual components yielded higher performance accuracy than niqāb visual and no mask audio components. Cotton mask visuals gave the lowest minimal pair word identification accuracy, and cotton mask audio resulted in the slowest reaction times. Face coverings did not significantly affect speakers' social evaluations. However, listeners did evaluate the niqāb-wearing speakers more slowly. Finally, prolonged exposure, assessed via daily exposure to face coverings, had no effect on the results.

1 Introduction

The COVID-19 pandemic of 2019-2021 significantly altered the way people communicated, particularly with the widespread adoption of face coverings for routine interaction. Early in the pandemic, face coverings received considerable negative media attention. Concerns were raised about their potential impact on communication and development – for example, fears that infants might struggle to recognise

* The first author is supported by a Harding Distinguished Postgraduate Scholarship at the University of Cambridge. Thanks are given to Vincent Hughes for co-supervising the first author's MA thesis, which was a preliminary investigation of this larger scale project. Additionally, thanks are also given to Marju Kaps and Katrina Kechun Li for technical assistance, and to Julia Schwarz for methodological and statistical advice.

emotions (Glaze 2022), or that face coverings could hinder language learning and social interaction (Hamzelou 2020).

Issues of speech clarity were also prominent, particularly for individuals who rely on lip-reading or visual speech cues (Gilbody-Dickerson 2020). Moreover, the loss of visual information can have a pronounced impact on speech understanding in individuals with hearing loss, who already face challenges due to degraded auditory input (Atcherson, Mendel, Baltimore, Patro, Lee, Pousson & Spann 2017, Lansing & McConkie 2003, Sumby & Pollack 1954). These concerns are especially relevant in healthcare settings, where face coverings – such as cotton or surgical masks – remain in regular use and where effective communication is critical for patient care and safety (Ford 2020). Understanding how face coverings affect communication remains important, not only in the context of the pandemic, but also in ongoing scenarios where face coverings continue to be widely used.

Face coverings were in use before the pandemic, of course, for example in medical contexts (e.g. surgical masks) and for religious purposes (e.g. Muslim women wearing various forms of veil). Even prior to the pandemic, however, face coverings were often criticised. For instance, a Muslim teaching assistant was suspended from work in a Yorkshire school as she refused to remove her veil when requested to do so by a male teacher (Simpson 2006).

Although face covering usage has decreased markedly since the pandemic's peak, it is still fairly common to encounter an individual wearing a face covering, e.g. at large-scale events like conferences or on public transport. Moreover, post-pandemic there have been reports of criminals exploiting the social acceptance of face coverings using them as disguises (Rawlinson 2021). Face coverings are therefore still a physical barrier through which people interact on a regular basis, and still the source of much emotive discourse and even calls for litigation (e.g. Washington Times 2024). Ongoing research continues to explore various social and practical implications of face coverings, including for example their effect on voice quality (Fiorella, Cavallaro, Nicola & Quaranta 2023, Jiang, Ng, Song & Chen 2024) and acoustic properties of speech (Nguyen, McCabe, Thomas, Purcell, Doble, Novakovic, Chacon & Madill 2021, Zhang, He, Li, Zhang & Hu 2022).

The present study was designed to test empirically some of the negative social comments about face coverings, such as those outlined above. We conducted a two-part experiment to assess the perception of speech produced through face coverings, considering both visual and auditory components.

Our study aimed to address concerns regarding the impact of face coverings on communication clarity and the social evaluations of speakers. Specifically, how social information (e.g. the visual presence of a face covering) and phonetic cues (e.g. the acoustic effect of a face covering) interact to influence speech and speaker perception (Campbell-Kibler 2008, Drager 2010, Hay, Nolan & Drager 2006a, Hay, Warren & Drager 2006b, Plichta & Preston 2005). A substantial amount of research in phonetics ignores the multimodal nature of speech which is comprised of both auditory and visual components (Rosenblum 2005). The speech perception literature, however, highlights the importance of visual cues, showing that seeing the speaker's mouth and the movement of their articulators can improve speech com-

prehension (Sumby & Pollack 1954). Previous research has also shown that when visual or auditory cues are disrupted, missing cues from one modality can often be recovered from the other (Grant & Seitz 2000). We therefore investigate the relative contributions of audio and visual information in processing naturalistic speech (e.g. with background noise) when produced through a face covering (Schwarz, Li, Sim, Zhang, Buchanan-Worster, Post, Gibson & McDougall 2022).

Finally, we also considered whether regular exposure to face coverings affect speech comprehension or speaker social evaluations. This was achieved by comparing the results from listeners who have regular and ongoing exposure to face coverings (e.g. via working in the healthcare sector) and those who do not. We ask if a listener's social perceptions can change with increased exposure thereby adopting a growth mindset where beliefs towards a face covering can be developed through learning (Dweck 2006).

This article is structured as follows. The remainder of the introduction provides a summary of previous research on face coverings, including their effect on speech acoustics, perception and processing. Following this review, Section 2 outlines the design of the two-part experiment conducted for the present study. The results are presented in Section 3 with the implications of these findings outlined in Section 4. Concluding comments are summarised in Section 5.

1.1 Acoustic effects of face coverings on speech

A number of studies have investigated the acoustic impact of face coverings on the speech signal (see Badh & Knowles 2023: for a review). This work mainly considers the effect of face coverings on spectral measures such as amplitude (loudness) and frequency (particular ranges of the speech spectrum, which can affect specific types of sounds).

With respect to amplitude, since face coverings present a physical barrier to the transmission of the speech signal, we might expect them to reduce sound pressure levels (Zhang et al. 2022). However, a range of amplitude effects have been found for different face coverings. Transparent face coverings made from plastic resulted in the most substantial amplitude reduction. For example, Corey, Jones & Singer (2020) reported a transmission loss of between 8 and 14 decibels (dB). The results for cotton masks are mixed. Balamurali, Enyi, Clarke, Harn & Chen (2021) found relatively large effects, with amplitude reduction of between 6.9 and 9.9 dB. In contrast, for Corey et al. (2020) the effect of the cotton mask was highly variable, depending on the material and the weave. In a comparative analysis of niqāb, balaclava and surgical mask, Llamas, Harrison, Donnelly & Watt (2008) found transmission loss to be frequency-dependent, with a greater or lesser degree of sound attenuation observed in different parts of the frequency spectrum (see further below). The variation between studies likely reflects differences in specific face coverings and experimental methods used. For example, in the experiments by Balamurali et al. (2021) and Corey et al. (2020), masks were mounted on a head-shaped loudspeaker to filter pre-existing recordings, whereas Llamas et al. (2008) recorded the speaker directly while wearing face coverings. The speech signal is likely to differ when

recorded directly rather than simply filtered, because human talkers are likely to hyper-articulate as a compensation measure when wearing a face covering (Traunmüller & Eriksson 2000; see further below).

Turning to the effect of face coverings on different frequency ranges, previous work has indicated that these acoustic changes are likely brought about by the constraints wearing a face covering places on the movement of the articulators. Such constraints can hinder the full lowering of the jaw, for instance, leading to a reduction in the first vowel formant frequency, or F_1. When testing eight mask types (motorcycle helmet, balaclava with and without mouth hole, strip of adhesive tape, niqāb, surgical mask, hoodie and scarf combination, and full-head rubber mask), Fecher (2014) found significant effects on the acoustic properties of various fricative sounds, including sound intensity and energy distribution across the frequency range. The work is, however, is limited in that it only analysed isolated sounds and syllables rather than spontaneous speech.

In recent years, more work has begun to investigate the effect of a face covering on other spectral measures including f_0 (the production correlate of vocal pitch) and harmonics-to-noise-ratio (HNR, a measure of additive noise relative to speech in the signal). To compensate for intelligibility issues when wearing a face covering, speakers have been found to use strategies including a higher f_0 and higher HNR (Geng, Lu, Guo & Zeng 2023). The increase in f_0 and HNR is presumably a consequence of hyper-articulation (Traunmüller & Eriksson 2000) in which speakers instinctively compensate for the impaired transmission of speech via the face covering by speaking more loudly and clearly. Increasing loudness is achieved by raising the pressure of air being expelled from the lungs. This in turn increases the pressure behind the vocal folds and makes them vibrate more quickly, which leads to a rise in f_0. Other work conducting an analysis of f_0, formant frequencies (F_1 and F_2), and smooth cepstral peak prominence across different face covering conditions (no face covering, cloth mask, surgical mask, KN95 mask, and a surgical mask over a KN95 with and without a face shield) found a significant effect of mask type for f_0, intensity and smoothed cepstral peak prominence (Joshi, Procter & Kulesz 2021).

1.2 The impact of face coverings on identification of phonemes and words

Several studies show that in quiet controlled listening conditions, phoneme and word identification remains largely unaffected by face coverings. For example, Brown, Van Engen & Peelle (2021) observed minimal listening errors in a sentence transcription task. Fecher (2014) similarly found participants were able to identify isolated sounds and controlled syllables to a very high degree of accuracy (overall 92% correct). Work by Llamas et al. (2008) also found minimal effects of a surgical mask and niqāb on speech comprehension using monosyllabic CVC English words like tat and pat, with misperceptions occurring in just 2% of the trials overall. The results from these studies, however, likely indicate a ceiling effect in tasks that were insufficiently challenging to establish the threshold at which face coverings might become problematic for word/phoneme identification in more naturalistic, routine communication. In particular, previous work has focused on relatively simple tasks

under optimised recording conditions, for example minimal pair discrimination in studio quality. Future work needs to test more naturalistic and challenging speech styles like spontaneous or continuous speech heard in background noise or multi-talker babble.

A handful of studies have recently begun to investigate the impact of face coverings in more ecologically valid conditions. For example, Smiljanic, Keerstock, Meemann & Ransom (2021) tested the effect of face coverings in noise and when speech is produced by an L1 or L2 English speaker. Their results mirrored other studies, finding that word recognition was just as accurate for speech with and without a face covering in quiet listening conditions. However, face coverings impaired word recognition of L2 speech at a lower noise level than L1 speech. This suggests that listeners are more tolerant of face covering speech in noise when produced by an L1 speaker as opposed to an L2 speaker. Truong & Weber (2021) investigated how face coverings affect sentence intelligibility and recall for German native listeners when presented with speech from an adult and a child talker. The results indicate that face coverings significantly impair both intelligibility and recall, with similar effects observed for both talkers.

Bottalico, Murgia, Puglisi, Astolfi & Kirk (2020) and Brown et al. (2021) both conducted experiments in which speech recorded via face coverings was adapted artificially to simulate noisy conditions. Predictably, intelligibility lowered as conditions worsened. Interestingly, Bottalico et al. (2020) found that speech perception was more severely impacted by added background noise for a fabric mask compared with either a surgical or N95 mask. From this, suggestions were made that both surgical and N95 masks were better suited than fabric masks to noisy environments such as classrooms. Similarly, Giovanelli, Valzolgher, Gessa, Todeschini & Pavani (2021) examined the impact of face coverings on speech comprehension in a speech-in-noise task, where participants showed reduced performance, lower confidence, and increased effort when talkers wore a face covering.

1.3 Visual cues and their role in speech perception

Face coverings not only interfere with the acoustic signal but also affect a speaker's visual appearance, obstructing around 50% of the face (Fecher 2014). Llamas et al. (2008) hypothesised that communication challenges associated with face coverings stem from visual obstruction to the lips and mouth, rather than severe degradation to the acoustic signal. In speech perception more broadly, substantial research has focused on the impact of visual cues on speech intelligibility (Grant & Seitz 2000, Massaro & Cohen 1983, Yuan, Lleo, Daniel, White & Oh 2021). This work highlights the strength of visual cues, finding that visual access to the speaker's mouth and the movement of the articulators can enhance speech comprehension (Sumby & Pollack 1954). Additionally, the McGurk Effect (McGurk & MacDonald 1976) famously demonstrates that when audio and visual cues are mismatched, perception is strongly influenced by what is seen as well as what is heard.

With face coverings blocking the lower part of a speaker's face, visual cues from the lips and other articulators are obstructed. Listeners therefore pay more atten-

tion to the upper parts of the speaker's face, in particular the eyes and eyebrows (Vatikiotis-Bateson, Eigsti, Yano & Munhall 1998). Previous work has indicated the usefulness of different aspects of the face for speech processing. For example, the lower half provides information useful for the comprehension of segmental features (Llamas et al. 2008: 84), while the top half is useful for recovering signs related to prosodic structure (Lansing & McConkie 2003), such as movement of the eyebrows and eyes helping to encode information relating to intonation. Moreover, sociophonetic research has shown that socio-indexical information from the visual appearance of the speaker can influence listeners' perception of phonetic variables (Hay et al. 2006a,b, Niedzielski 1999).

The general consensus is that visual cues from the mouth and lips become more important for speech processing when speech is heard in challenging listening conditions (Atcherson et al. 2017, Lansing & McConkie 2003, Sumby & Pollack 1954, Yi, Pingsterhaus & Song 2021). However, it has also been shown that limited or distorted visual information is better than audio information only (Erber 1975, Atcherson et al. 2017). In a small scale study, Schwarz et al. (2022) considered the separate contributions of visual and acoustic cues to speech processing for speech produced through a face covering. A Standard Southern British English (SSBE) speaker was recorded in two conditions (cotton mask versus no face covering), reading both high predictability and low predictability sentences (e.g. 'For your birthday I baked this cake' versus 'Tom wants to know about this cake'). Two types of stimuli (matched or mismatched audio-visual information) were generated for a listening experiment. In the matched stimuli condition, participants saw and heard a speaker wearing a cotton mask. In the mismatched stimuli conditions, participants saw a speaker wearing no face covering whilst hearing cotton mask speech, or vice versa. Listeners were required to repeat verbally the last word in the sentence. Results revealed that when both the acoustic and visual signal were degraded by a face covering, adults and children (aged 8-12 years) responded with similarly reduced accuracy than when responding to speech produced without a face covering, but adults were slower to respond in the face covering visual condition. Children relied less on the visual information, but were affected similarly to the adults by the acoustic degradation. For both adults and children, the semantic predictability of the sentence significantly affected the results. For the highly predictable stimuli, the audio and visual effects of a face covering were smaller, such that the predictable context fully compensated for the acoustic and visual degradation caused by a face covering. These findings therefore highlight the significance of semantic predictability in interpreting speech produced through a face covering.

1.4 The effect of face coverings on social evaluations of a speaker

Research investigating the relationship between face coverings and the social evaluations of a speaker has predominantly concentrated on the identifications of emotions. For instance, Kret & De Gelder (2012) investigated participants' ability to recognise emotions when they had restricted access to the face. The study involved six females wearing different types of Muslim veils: a burqa covering the

whole face with a mesh layer over the eyes, a niqāb, covering the entire face apart from the eyes, and a hijab covering only the hair. When participants described the emotion felt by the veil wearer, participants most frequently associated fear with those wearing a burqa or niqāb, as opposed to a hijab. Noyes, Davis, Petrov, Gray & Ritchie (2021) compared the identification of positive and negative emotions, finding that positive emotions are generally easier to recognise than negative ones. Both Kret & De Gelder (2012) and Noyes et al. (2021) used acted emotions, however, which are potentially more extreme than the subtle or variable emotions encountered in real life situations. Whilst there is limited research on the effects of face coverings on speaker social evaluations, there is a sizeable sociophonetic literature examining how stereotypes and social categories influence phoneme identification and classification (see Drager 2010, Plichta & Preston 2005). Based on this, we hypothesize that the stereotypes associated with face coverings will affect both word identification and the social evaluations of a speaker. For example, we predict that listeners will perceive the speech signal differently depending on whether they associate a face covering with certain social or cultural stereotypes. Additionally, we expect listeners to evaluate the speech signal more negatively if they align with the negative stereotypes associated with face coverings.

1.5 The effect of prolonged exposure to speech through a face covering

During the COVID-19 pandemic, questions were raised regarding the impact of prolonged exposure to face coverings on our ability to process speech. Do speech comprehension and social evaluation of speakers improve the more we interact with people who are wearing a face covering? To our knowledge, Crinnion, Toscano & Toscano (2022) is the only study to have investigated this relationship thus far. Their experiment was conducted in the US between 2020 and 2021, where participants were assessed on their ability to transcribe speech produced through a face covering presented with added multi-talker babble. With 2021 being mid-pandemic, the aim of the study was to determine if increased exposure to face covering speech improved transcription accuracy. While the study found no significant effects, the researchers acknowledged its limitations, noting that during the initial data collection, face coverings were mandatory and therefore participants may have already adapted to perceiving speech produced through a face covering. Consequently, it remains unclear whether increased exposure affected responses. We therefore ask if a listener's social perceptions can change with increased exposure thereby adopting a growth mindset where abilities and beliefs towards a face covering can be developed through effort and learning (Dweck 2006).

1.6 Summary

While previous work has begun to investigate the contexts where speech produced through a face covering might be problematic, these studies have used highly controlled stimuli. For example, stimuli consisted of monosyllabic words, and were filtered for a face covering by playback through a loudspeaker as opposed to a hu-

man talker (Bottalico et al. 2020). Other limitations include the use of meaningful sentences where the target word is easily predicted based on the other words in the sentence (Brown et al. 2021). More recently, work has begun to investigate the impact of different face coverings (Bottalico et al. 2020), background noise (Brown et al. 2021), semantic predictability (Schwarz et al. 2022) and the individual contributions of the audio and visual signal (Schwarz et al. 2022). Independently these studies have shown an effect of each variable on speech comprehension. However, to date, little work has focused on the combined effect of continuous unpredictable speech heard in background noise and produced through a face covering.

To our knowledge few studies have investigated the effect of different face coverings on minimal pair word identification in semantically unpredictable sentences heard in background noise, i.e., the conditions that apply to much of speech communication in everyday circumstances (Cohn, Pycha & Zellou 2021, Pycha, Cohn & Zellou 2022). The work investigating this found that speech produced through a face covering was often more intelligible than no face covering speech in clear speaking conditions. This suggests that speakers actively adapt their articulation in response to wearing a mask. Furthermore, little work has investigated three further issues: (1) the separate contributions of the auditory and visual components associated with speech produced through a face covering and how they work together (cf. Schwarz et al. 2022), (2) the impact of face coverings on how listeners socially evaluate a speaker, and (3) the role of prolonged exposure to face coverings. The present study addresses these issues, serving to test empirically the sorts of concerns raised in the media about the effects of face coverings on spoken communication. Overall, we are interested in the relationship between social knowledge and speech perception (Campbell-Kibler 2008, Drager 2010), the interaction between auditory and visual components in speech perception (Hay et al. 2006a,b), and finally an individual's ability to develop more positive attitudes towards face coverings with regular exposure (Dweck 2006), assessed via the amount of daily exposure to face coverings.

Our research questions are therefore as follows:

i. Do the acoustic and visual effects of wearing a face covering affect the comprehension of minimal pair words in unpredictable connected speech heard in background noise?

ii. Do face coverings impact listeners' social evaluations of a speaker?

iii. Does increased exposure to face coverings affect listeners' minimal pair word comprehension and social evaluations of a speaker?

2 Methods and Materials

A two-part experiment was conducted. Part 1 investigated the relationship between the visual and acoustic components of face coverings on correct minimal pair word identifications in background noise. Part 2 addressed the effect of face

coverings on the social evaluations of a speaker. In both parts, the impact of listeners' experience of exposure to different types of face covering was also investigated.

2.1 Stimuli production

2.1.1 Stimuli content

Part 1: To test the relative contributions of acoustic and visual components of face covering on minimal pair word identification, stimuli were designed to assess participants' ability to distinguish between minimal pairs, i.e., pairs of words differing in a single phoneme.[1] The target words were placed at the end of semantically unpredictable carrier sentences adapted from Kalikow, Stevens & Elliott (1977), e.g. 'Bob heard Paul talk about the CELL/SHELL' (/sel/ versus /ʃel/). The sentences were modified to (i) include specific target words and the selected minimal pairs, and (ii) to ensure the carrier sentence did not contain other instances of the phoneme by which these pairs of words differed, avoiding possible priming. The following minimal pair categories were chosen as they are known to be the source of regular confusion in ordinary speech: /f/ and /s/, /f/ and /h/, /p/ and /k/, /p/ and /h/, and /s/ and /ʃ/ (Miller & Nicely 1955, Redford & Diehl 1999, Wang & Bilger 1973). All participants were exposed to the same 120 sentences which were derived from each of the five phoneme pairs (i.e., 12 per phoneme). A lexical frequency analysis was also conducted, using the British National Corpus (Davies 2007), to ensure the minimal pair word pairs were well matched for frequency. Potential word pairings were rejected if one word occurred at least five times more frequently than the other. This threshold led to 30% of the original stimuli being rejected. A stricter threshold would have led to the rejection of too many stimuli.

Part 2: Stimuli were designed to elicit a portion of continuous speech from which social evaluations of a speaker could be judged. The texts chosen for the stimuli were nine stories, outlining the influence of face coverings on our ability to recognise emotions, for example, 'When people wear a face covering it can really feel like they are staring at you, but in reality, they are probably smiling behind the face covering' (Boone 2020).[2] The stories were taken from personal experiences, tweets and news articles illustrating the impact of face coverings on emotion recognition. Each story was adapted to first person, allowing it to be narrated as a personal account of events. Although the theme of each story was consistent, each speaker narrated a different story to prevent content priming effects and reduce the likelihood of participants making direct comparisons between speakers based on repeated material. This approach also minimised the risk of listener fatigue and helped maintain engagement across trials. Moreover, varying the content allowed for a more naturalistic presentation of speech.

[1] The stimulus sentences are available at https://osf.io/bwtph/files/osfstorage/66cc8de228ea3e5e6a8c940e.

[2] The texts of the stories are available at https://osf.io/bwtph/files/osfstorage/66cc85ac66c6782080b69a73.

Visual condition	Audio condition	Speakers
Niqāb	Niqāb, No mask	4, 5, 6
Cotton	Cotton, No mask	1, 2, 3
No mask	Cotton, Niqāb, No mask	7, 8, 9

Table 1 A summary of the speakers assigned to each audio and visual condition.

2.1.2 Speakers

Nine young adult female speakers of SSBE were selected to produce the stimuli. As seen in Figure 1, three speakers were assigned to one visual condition (cotton mask, niqāb and no face covering). Speakers were assigned to only one visual condition to preserve the authenticity of each visual condition while still allowing us to compare across audio stimuli (e.g. it would have been inauthentic to have speaker 4 wear a niqāb for half the experiment and then no face covering for the second half). As seen from Table 1, there were three speakers assigned per visual condition. This helped to control for speaker-specific effects. For instance, if speakers 4, 5 and 6 were less intelligible this would equally affect the niqāb and no mask audio conditions. Therefore, we can attribute differences in results between these two conditions to the face covering rather than the speaker. Additional steps were also implemented to reduce speaker-related variability. These included matching the speakers for accent, age and sex; accounting for individual speaker variability in our statistical model; and finally, although no formal testing of fluency was undertaken, the individuals were chosen and monitored by a trained phonetician to ensure there were no marked differences in the fluency of the speakers. In the judgement of the authors, they were all standard and clear with little elision. Finally, the multi-speaker design also aimed to minimise listener fatigue and ensure that observed effects could be attributed to the face covering rather than the individual speaker.

To ensure consistency across speakers without introducing potential ethical biases, all speakers were white. However, for the niqāb condition, individuals with dark hair and brown eyes were selected, as is statistically more typical for those who traditionally wear a veil. To ensure authenticity, a regular veil wearer assisted with the veil dressings. We chose a cotton mask and niqāb because they are associated with different indexicalities. The cotton mask is frequently associated with the public health contexts, especially since the COVID-19 pandemic. Alternatively, the niqāb is more commonly associated with cultural, religious and political identities.

2.1.3 Recording procedure

To capture the audio signal a Sennheiser microphone and MixPre recorder was used. The video components were captured using a video tripod and iPhone 13 Pro camera. All recordings were conducted in the sound-treated room of the Phonetics

Figure 1 Static visual representation of the nine speakers presented to listeners.

Laboratory at the University of Cambridge. The sentence stimuli were captured using a two-phase set-up. Phase 1 captured audio-only recordings. Here the speaker sat in front of a monitor, reading all the sentence stimuli with no face covering. The aim of this phase was to collect a baseline set of recordings, which then allowed speakers to imitate their own productions in subsequent recordings, enabling for the cross-dubbing of audio and visual components. Phase 2 captured combined audio-video recordings. As shown in Figure 1, each speaker sat in front of a black backdrop, focusing her gaze on a fixation cross attached to the video tripod. Each speaker was played back her own baseline recordings from phase 1 using a Yamaha monitor speaker and was asked to imitate her productions. The speakers did phase 2 twice: firstly, wearing no face covering and secondly wearing either a cotton mask or niqāb. As the stimuli were short sentences, speakers were able to imitate their own productions with a similar rhythm and prosody such that these repeated utterances could be used for cross-dubbing different audio and video components. For the story passage, speakers only completed stage 2 with audio-video capture. That is, they were asked to read the story passage in a natural story-telling manner, firstly wearing no face covering and secondly wearing either a cotton mask or niqāb.

2.1.4 Creation of stimuli

To account for variations in vocal effort across speakers, all stimuli were normalised for intensity and set to 70 dB using Praat (Boersma & Weenink 2023) before being mixed with background noise. We chose to normalize the intensity as our aim was to reduce variability stemming from individual differences in vocal effort or speaking volume, which could have otherwise confounded comparisons across face covering conditions. While we acknowledge that this procedure may have removed a

natural compensatory effect associated with face coverings, we deemed it a necessary trade-off to ensure consistency and comparability across stimuli produced by different speakers. All stimuli were mixed with pub noise from freesounds.org (Islabonita 2013). Pub noise has high temporal variation in the spectrum, which can lead to increased masking of the speech content. The speech and noise were mixed at a sound-to-noise ratio (SNR) of 7 dB using a Praat script (Harrison 2022) with both the background noise and speech samples having matching sampling rates (48 kHz) and bit depths (16). An SNR of 7 dB was chosen as it reflected a naturalistic level of background noise, similar to that heard in everyday communication (Wu, Stangl, Chipara, Hasan, Welhaven & Oleson 2018). Stimuli were initially downloaded in a high-quality format, compressed using HandBrake (Petit 2023) and the 2-pass encoding setting. This ensured efficient loading times for participants without significantly degrading the quality of the speech or video signal.

Part 1

For the minimal pair comprehension test, all sentence stimuli were edited to include a 30ms gap after the last word in the utterance. This ensured every recording had an identical lapse of time between the offset of the target word and when participants were able to select a response.

From the 120 sentences, two types of stimuli were manually grafted for the listening test using iMovie (Apple 2023). Grafting, in this sense, involved a two-part process: (i) independent segmentation of the audio and visual recordings at the sentence level, and (ii) overlapping the segmented audio and video to synchronise them. First, control (matched) stimuli were created using 60 sentences. The control (matched) stimuli had matching audio and video components. For example, the speaker spoke with a niqāb and was also visually presented wearing a niqāb. The other 60 sentences were mismatched stimuli, i.e., the speaker spoke with her allocated face covering (niqāb or cotton mask) but was visually presented wearing no face covering (or vice versa). Whilst grafting the stimuli in this way could potentially introduce a confound related to audio-visual asynchrony, we took multiple steps to minimise this risk. First, the audio was overlaid onto the visual for all stimuli, including those in both the matched and mismatched conditions. Second, the grafting method was adapted from (Schwarz et al. 2022), who validated its effectiveness in producing naturalistic, rhythmically aligned audio-visual stimuli. Thirdly, the stimuli were short sentences meaning speakers were able to reproduce their own utterances with closely matched rhythm and prosody. This allowed us to graft different audio onto the video components without introducing perceptible asynchrony. Finally, all three authors – trained phoneticians – carefully reviewed the stimuli to ensure that audio-visual synchrony was preserved. Distribution of sentences across audio and visual combination ensured two things: (i) an equal number of sentences for each audio and visual condition, and (ii) each speaker produced each phoneme from the minimal pairs at least once.

All participants were exposed to the 120 sentences in blocks of 20, where they would receive a short break at the end of the block. The stimuli were randomised

per block, with each block having samples from every speaker. Participants were exposed to stimuli in both the matched (60 sentences) and mismatched conditions (60 sentences) and saw all 9 speakers and the three face covering conditions. No participant was ever exposed to the same sentence twice as the sentences were not repeated. However, participants were exposed to both minimal pairs (e.g. they heard 'Bob heard Paul talk about the cell' and 'Bob heard Paul talk about the shell'). While it is possible that exposure to cell possible affected their subsequent exposure to shell, this was offset by the size of testing and the fact minimal pairs were most likely separated by several stimuli.

Part 2

For the stories, only matched (control) stimuli were used. For example, the speaker was recorded speaking with a cotton mask and visually presented wearing a cotton mask. Grafting of the matched stimuli was still required, however, as the audio signal was taken from the MixPre recorder and the video from the iPhone 13 Pro. Only matched stimuli were created because the passages were longer stretches of continuous speech, meaning it was too challenging for speakers to imitate their recording, remembering the content, rhythm and prosody of the passage. Therefore, as previously mentioned, speakers only completed phase 2 when recording the stories, reading them in a natural story-telling manner.

2.2 Listeners

This experiment aimed to understand better if increased exposure to face coverings affects listeners' comprehension and social evaluations of a speaker. Participants were therefore recruited from two distinct demographic profiles. The first group had daily exposure to face coverings (e.g. UK NHS [National Health Service] workers) while the second group had infrequent or no daily exposure to face coverings. The extent of exposure was established via Prolific (Damer & Bradley 2014) , where participants were asked to report the number of hours per day they spend interacting with individuals wearing face coverings. Participants who reported spending more than six hours daily in such interactions were classified as having 'daily exposure'. Alternatively, those who reported exposure of one hour or less were categorised as having 'infrequent or no daily exposure'. Other criteria implemented through Prolific included the following: (i) English as a first language, (ii) no learning/hearing or language disabilities, (iii) normal vision, and (iv) UK residency. We acknowledge that because hearing status was self-reported, this does not necessarily indicate normal hearing. To help address this limitation, all participants were required to complete a sound check prior to beginning the experiment, and individuals who did not pass this check were excluded from participation. Biographical information (including age, sex, linguistic expertise, ethnicity and years of residency in the UK) was also collected. A power analysis (Brysbaert & Stevens 2018) was conducted to determine the recommended number of participants. When accounting for all independent variables, the recommended number of participants

per demographic group was 40. Therefore, 40 participants with daily exposure to face coverings and 40 with no exposure to face coverings were recruited.

2.3 Listening test

The listening test was built using Gorilla (Anwyl-Irvine, Massonnié, Flitton, Kirkham & Evershed 2020) and presented to listeners using the following structure: consent form, sound check (Milne, Bianco, Poole, Zhao, Oxenham, Billig & Chait 2021), demographic questionnaire, word identification test, social evaluation test, and review of any technical difficulties. Participants completed the study at home using their own laptops. While we could not control the home environment, participants were explicitly instructed to only complete the study if they were able to do so in a quiet space where there was minimal background noise. They were also required to complete the study using either Firefox or Chrome as these browsers caused the least lag for both Macbooks and PCs (Bridges, Pitiot, MacAskill & Peirce 2020). Participants were instructed to use their laptop or PC speakers to play the videos, rather than headphones. This was to ensure the stimuli more closely reflected everyday communication rather than optimal listening conditions. Finally, participants were also instructed to set their speaker volume to an appropriate level during the sound check and not to adjust the volume after this point. An appropriate level was described to participants as a level where the sentences could be clearly heard without causing discomfort. On average, the experiment took 25 minutes.

Part 1

The first part of the experiment was the minimal pair word identification task, consisting of four practice stimuli and 120 test stimuli. For every trial, participants were initially presented with a fixation cross, followed by a video of a speaker producing a semantically unpredictable sentence, and finally two images corresponding to the minimal pair words (see Figure 2). For example, for the sentence 'Bob heard Paul talk about the CELL/SHELL', participants would be presented with the image of a cell and a shell. Their task was to pick (via keyboard response) the image which is most relevant to the sentence. They were instructed to respond as quickly and as accurately as possible. The correct image was randomly assigned to either the right or left side of the screen. While participants were not explicitly instructed to identify a specific word, the task was framed this way to encourage them to listen to the entire sentence rather than focus solely on a single word.

To minimize the potential confounding effect of using images – specifically, the possibility that some words are more straightforward to represent with an image than others – we took several precautions during the design phase. These included several rounds of planning and selecting target words. Minimal pair words that were notably difficult to depict visually or were reported as hard to interpret by participants in the pilot study were excluded. We retained only those items that were judged to be clear and interpretable. While some variation is inevitable, we

believe we have constrained its impact by avoiding visually ambiguous items and by balancing lexical frequency across stimuli.

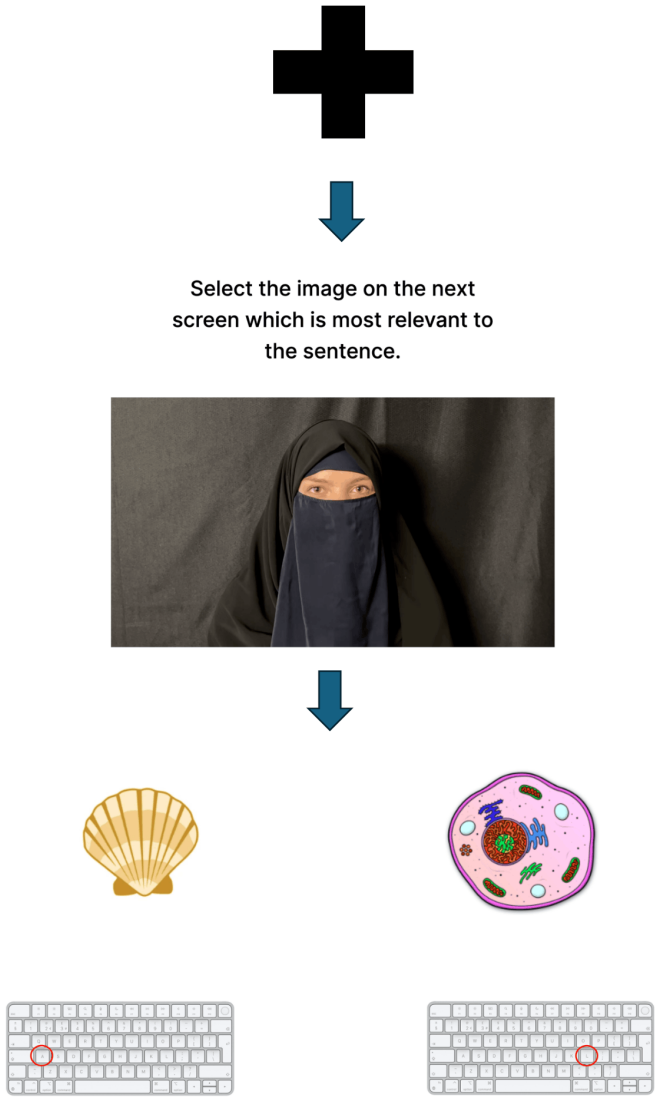

Figure 2 A visual representation of the trial design used for the comprehension task.

Part 2

In part 2 of the experiment, listeners completed the social evaluation task. The participant watched the speaker tell the anecdotal story and then rated the voice with respect to five attributes: friendliness, intelligence, intelligibility, trustworthiness, and British nationality. The attributes were selected for the following reasons: Friendliness and Trustworthiness were included due to concerns about babies' ability to recognize emotions when adults wear face coverings (Glaze 2022). Intelligibility was chosen in response to concerns regarding clarity of communication (Brown et al. 2021). Finally, Intelligence and Nationality were selected because, even before the pandemic, face coverings were frequently the target of social criticism. Each evaluation was graded on a 7-point Likert scale (see Figure 3). The three niqāb wearing speakers were always presented first to ensure the niqāb ratings were independent and not made relative to another face covering. This approach also aimed to minimise any social-desirability bias. Thus the participants were initially shown the three niqāb wearing speakers in a random order, followed by the remaining 6 speakers wearing either a cotton mask or no face covering presented in a random order. In part 2, participants were also explicitly told to answer as quickly as possible.

Ethical approval for this study was granted by the University of Cambridge, Faculty of Modern and Medieval Languages and Linguistics Research Ethics Committee. Informed consent was obtained from both the speakers (who were recruited to record the stimuli) and the participants (who completed the online experiment). Speakers consented to the use of their audio and video recordings in the online listening test, as well as to the inclusion of individual videos and still frames in conference presentations and academic publications. However, in accordance with the University of Cambridge's ethics policy, these recordings cannot be uploaded to a public repository. Listener participants consented to the public sharing of their anonymised response data, which have been made available on OSF.[3]

2.4 Data analysis

2.4.1 Response scores and reaction times

In both parts of the experiment, response scores and reaction times were recorded. Responses to the minimal pair word identification task were marked using a binary scale (correct/incorrect). Reaction times were captured from the end of the target word to the onset of participants pressing the keyboard response. Responses to the social evaluations task, given on a 7-point Likert scale, were z-scored on a by-participant basis to account for individual variation in use of the scale. Reaction times were recorded as the length of time participants took to rate all five attributes. This was measured from the offset of the anecdotal story to the onset of participants submitting their responses for all five attributes.

[3] https://osf.io/bwtph/overview?view_only=ff25417980434291aafd5411bc8c5aae.

Figure 3 A visual representation of the trial design used for the social evaluations task.

2.4.2 Statistical analysis

All statistical analysis was carried out in RStudio (Ihaka & Gentleman 2023) using the following packages: `tidyverse` (Wickham, Averick, Bryan, Chang, McGowan,

François, Grolemund, Hayes, Henry, Hester, Kuhn, Pedersen, Miller, Bache, Müller, Ooms, Robinson, Seidel, Spinu & Yutani 2019), `ggplot2` (Wickham 2016), and `lme4` (Bates, Mächler, Bolker & Walker 2015). Data were only analysed if participants had completed the entire experiment. Twelve participants were also excluded due to reporting that their video or sound did not play for more than 20% of the experiment. Additional participants were recruited to reach the target of 80 responses.

Part 1

The minimal pair word identification test results were marked on a binary scale (1 = correct, 0 = incorrect), then analysed using a logistic mixed effects regression model. To evaluate which fixed effects had a significant effect on correct minimal pair word identifications, a null model was fitted including only the random effects (speaker, listener, phoneme). A model for each fixed effect was then created including the fixed effect and the random effects. Fixed effects were added for the level of exposure to face coverings (daily exposure, no exposure) and the audio/visual components (no face covering, niqāb, cotton). The ANOVA function in R was used to compare the null model to a model with each fixed effect, then results used to indicate whether the model significantly improved with the fixed effect added. Interactions between the fixed effects were also tested if the fixed effect improved the model. Results of the model outputs, containing an estimate, standard error rate and a p-value, were then examined to evaluate the relationship between the variables. Treatment coding was used to set the contrast between the exposure levels and the audio/visual components. For the face covering exposure analysis, the no face covering exposure was used as the baseline condition (0) with daily face covering exposure set to level 2. Similarly, the no face covering audio and visual condition was used as the baseline (0), with the cotton mask set to level 2 and the niqāb to level 3.

Part 2

The z-scored social evaluation test results were analysed using a linear mixed effects regression model. To test which fixed effects had a significant effect on the social evaluation ratings, a null model was fitted including only the random effects (speaker, passage, social evaluation). Exposure to face coverings (daily exposure vs no exposure) and face covering type (no face covering, cotton and niqāb) were included as fixed effects. The ANOVA comparison and setting up of variable contrasts was done using the same method as described for part 1.

A significance threshold of < 0.05 was used in all analyses.

3 Results

3.1 Part 1. Minimal pair word identification test results

Raw results and full detail of statistical analyses from Part 1 are provided at https://osf.io/bwtph/files/osfstorage/66e1bfe283dad68c1d2e0910.

3.1.1 Word identifications

The results of the ANOVA model comparisons revealed that the rate of correct word identifications was best modelled with an interaction between the audio and visual components [χ^2= 52.96, $p < 0.001$]. Figure 4 plots the effects of the audio and visual components on the percentage of correct minimal pair word identifications for each face covering (FC) separately. One interaction between the audio and visual components was significant, with a second reaching near significance. With respect to the niqāb, participants had the highest score in the matched condition (niqāb audio + niqāb visual, data point E (87.2)) (β = 0.76, SE = 0.34, $p < 0.03$). Unexpectedly, this condition outperformed the two mismatched conditions: (i) niqāb audio + no FC visual, data point B (85.4%) and (ii) no FC audio + niqāb visual, data point D (80.3%). Opposite to the effect seen for the niqāb, participants had the lowest score in the matched cotton mask condition (cotton audio + cotton visual, data point G (72.6%)) compared to the cotton mask mismatched conditions: (i) cotton audio + no FC visual, data point A (87.6%), and (ii) no FC audio + cotton visual, data point F (78.8%). This interaction, however, narrowly missed statistical significance (β = -0.65, SE = 0.34, $p = 0.059$).

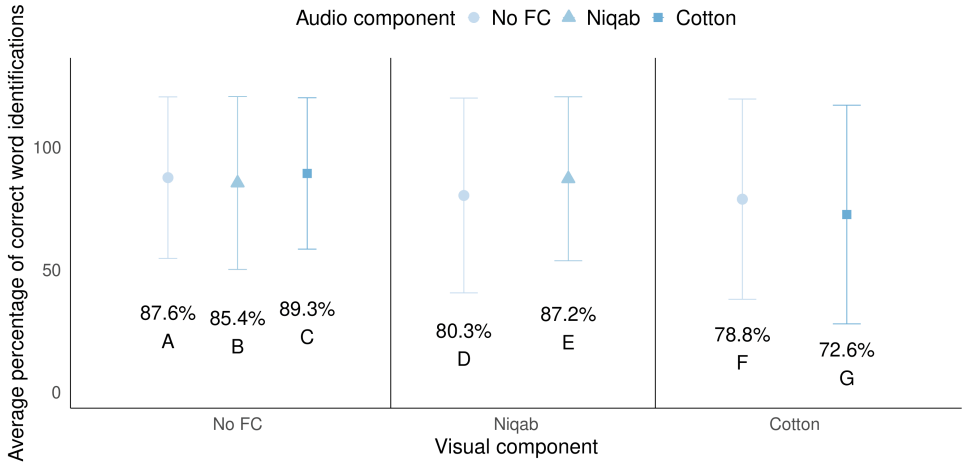

Figure 4 Correct word identifications per audio and visual components. The visual components are presented on the x-axis. The audio components are presented via shape (No FC = circle, Niqāb = triangle, Cotton = square). Error bars represent ±1 standard deviation from the mean. The percentages below each data point indicate the average percentage for that condition and the alphabetic labelling is used in the text to help explain interactions between the different conditions.

The model output also summarises the independent effects of audio and visual components. Since the model was fitted with an interaction between audio and visual components, the results of these fixed effects are arguably not fully independent. As seen from the circular data points in Figure 4, when holding the no FC audio constant a significant difference was observed between all three logical combinations: the no FC visual (87.6%, A in Figure 4) and the niqāb visual (80.3%, D in Figure 4) (β = -0.57, SE = 0.28, $p < 0.04$), no FC visual (87.6%, A in Figure 4) and cotton mask visual (78.8%, F in Figure 4) (β = -0.70, SE = 0.28, $p < 0.01$) and the niqāb (80.3%, D in Figure 4) and cotton mask visual (78.8%, F in Figure 4) (β = -0.271, SE = 0.13, $p < 0.05$). In other words, participants were best at interpreting the no FC speech when shown no FC visual. Out of the two FC conditions, participants were worst at interpreting no FC speech when shown a cotton mask as opposed to a niqāb. Moving on to the independent effect of audio components (see the leftmost panel), when holding the no FC visual component constant, no significant differences were observed between the no FC audio (87.6%, A in Figure 4) and niqāb audio (85.4%, B in Figure 4) (β = -0.27, SE = 0.33, $p = 0.49$). Similarly, no differences were found between the no FC audio (87.6%, A in Figure 4) and the cotton mask audio (89.3%, C in Figure 4) (β = 0.24, SE = 0.33, $p = 0.47$). Finally, no differences were also found between the niqāb audio (85.4%, B in Figure 4) and the cotton mask audio (89.3%, C in Figure 4) (β = -0.22, SE = 0.12, $p \geq 0.06$).

Finally, as mentioned in Subsection 2.2, data were collected from two listener groups: one with daily exposure to FC's and the other with no regular exposure to FCs. Exposure to FCs had no significant impact on the model [χ^2 = 0.60, $p = 0.44$].

3.1.2 Reaction times

The variability in reaction time results was best accounted for using a model with the audio components only [χ^2 = 52.96, $p < 0.001$]. On average participants took 1252ms (SD = 602 ms) to respond to the cotton mask audio stimuli. They were significantly faster when responding to both the niqāb (mean RT = 1160ms, SD = 613ms) and no FC (mean RT = 1186ms, SD = 605 ms) audio components (i.e., on average they were faster by 92ms and 66ms respectively). The regression output confirmed that the cotton mask audio yielded a significant increase in reaction times compared to both the niqāb audio (β = -91.81, SE = 19.37, $p < 0.001$) and no FC audio (β = 65.58, SE = 19.37, $p < 0.001$). However, the difference between no FC and niqāb audio was not statistically significant (β = -26.23, SE = 19.32, $p = 0.2$). Conversely, the visual component had no significant effect on reaction time responses [χ^2 = 3.49, $p = 0.17$]. Finally, no effects were observed between the FC exposure variable and reaction times [$\chi^2(1)$ = 1.44, $p = 0.23$].

3.1.3 Post hoc phoneme analysis

Raw results and full detail of statistical analyses from the post hoc phoneme analysis are found at https://osf.io/bwtph/files/osfstorage/66e1bfe2b49632e985a3cfa9. To assess whether the fixed effects (audio components, visual compo-

nents, and amount of exposure to FCs) influenced word identification accuracy per phoneme, a post hoc analysis was performed. This involved subsetting the data per phoneme (10 phonemes). 10 independent mixed-effects logistic regression models were then run to test if the fixed effects accounted for the word identification accuracy scores for each phoneme. For this analysis, audio and visual components were assessed using a two-level distinction: FC vs no FC. Analysis according to the different FC types was not possible as the sample size was too small after subsetting. The results for this section are presented as follows: firstly, the phonemes where the visual components affected performance accuracy are discussed, then the phonemes where the audio components affected accuracy are discussed. Finally, the phonemes where both audio and visual components interacted to affect accuracy are presented. /f, s/ = [f], /f, h/ = [h], /p, h/ = [h], /p, h/ = [p], /p, k/ = [p] and /s, ʃ/ = [ʃ] will not be further discussed as neither audio nor visual components had a significant effect on performance accuracy. The following notation /f, s/ = [f] is used to refer to the minimal pair phoneme categories and their realisation. Specifically, the minimal pair phonemes are represented in slashes with the speaker's realisation indicated in square brackets. Minimal pair word identification accuracy was significantly affected by visual components for /f, h/ = [f]. Figure 5 shows that participants performed significantly worse when presented with the FC visual component (β = -1.82, SE = 0.43, $p < 0.001$). In other words, [f] was identified more accurately when presented with the no FC visual.

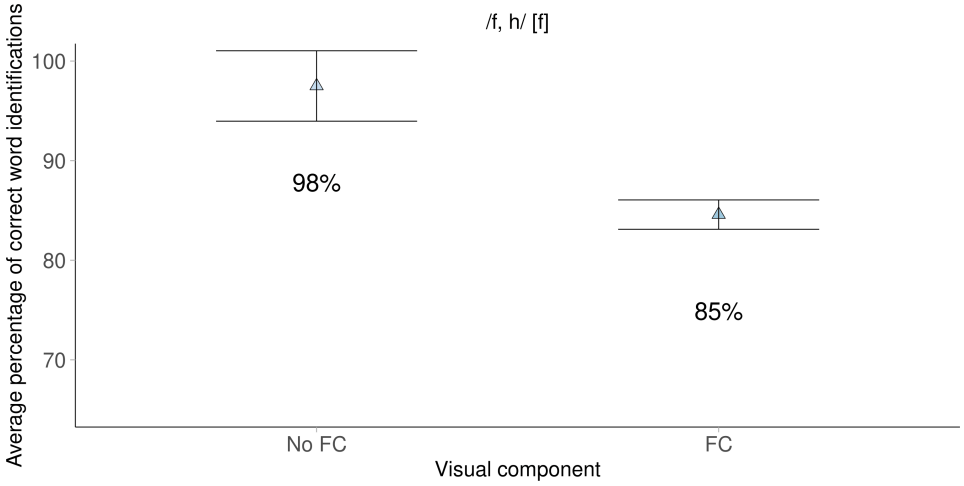

Figure 5 A plot illustrating the results for /f, h/ = [f], the phoneme realisation where the visual components had a significant effect on correct word identification. The visual components are presented on the x-axis (No FC = left and FC = right). Error bars represent ±1 standard deviation.

The phoneme realisations where word identifications were significantly affected by the audio component include: /f, s/ = [s] and /p, k/ = [k]. As seen in Figure 6, [s] was identified more accurately when participants were presented with the No FC audio (β = -0.94, SE = 0.39, $p < 0.016$). However, unexpectedly, for [k], participants

performed more accurately when presented with FC audio (β = -1.9354, SE = 0.3927, $p < 0.001$) (see Figure 7).

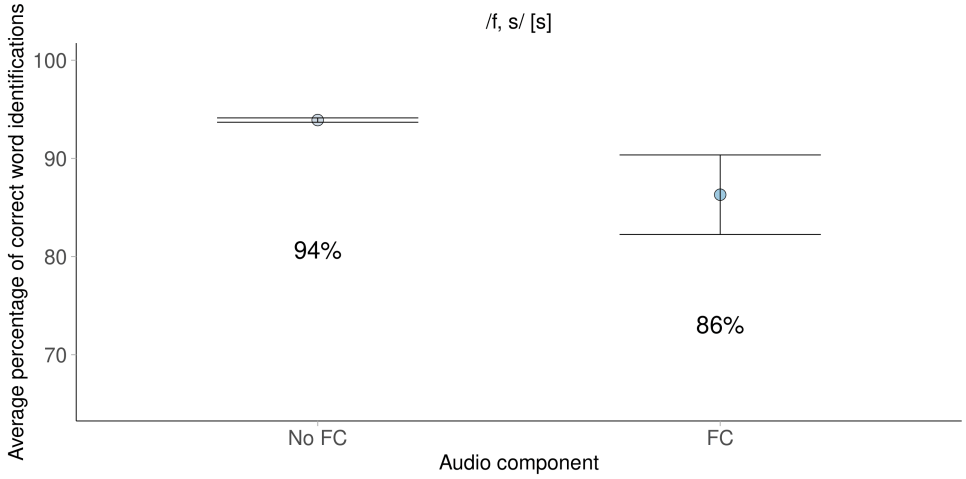

Figure 6 A plot illustrating results for the phoneme realisations where the audio component had a significant effect on the percent of correct word identification. The audio components are presented on the x-axis (No FC = left and FC = right). Error bars represent ±1 standard deviation.

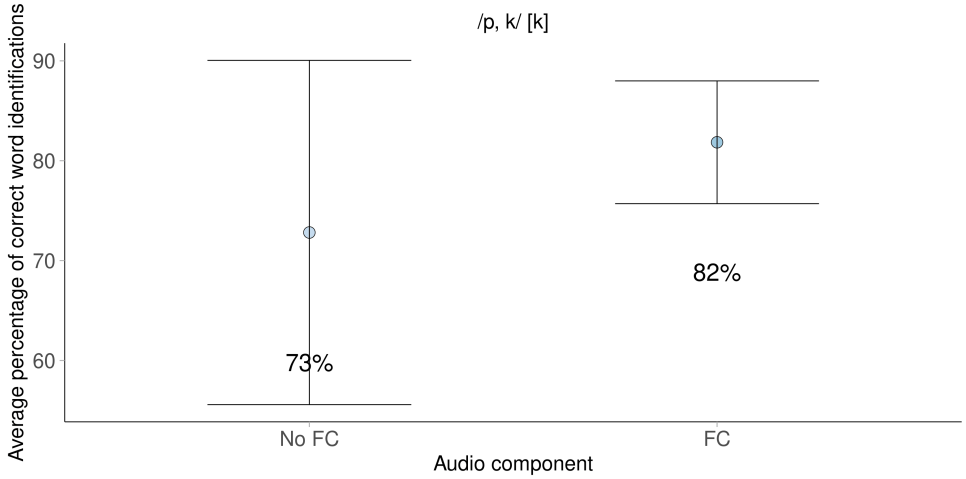

Figure 7 A plot illustrating results for the phoneme realisations where the audio component had a significant effect on the percent of correct word identification. The audio components are presented on the x-axis (No FC = left and FC = right). Error bars represent ±1 standard deviation.

Finally, for /s, ʃ/ = [s], minimal pair word identification accuracy was significantly affected by an interaction between audio and visual components. Unexpectedly, participants performed worse when shown no FC visual and played no FC audio (72.5%) (β = -2.73, SE = 0.80, $p < 0.001$) as opposed to (i) no FC visual and

FC audio (94.6%), (ii) FC visual and FC audio (84.4%), and (iii) FC visual and no FC audio (87%).

3.2 Part 2. Social evaluation results

Raw results and full detail of statistical analyses from Part 2 are provided at https://osf.io/bwtph/files/osfstorage/66e1bfe2def5abf645a3d558.

3.2.1 Social evaluation scores

As explained earlier, the stimuli for the social evaluations test were presented in the matched condition only, i.e., in each given story the audio signal was produced with the same FC as the one presented in the video.

Statistical analysis of the social evaluation scores found no significant difference between the two participant groups (i.e., those with daily exposure to FCs and those with no regular exposure) [$\chi^2 = 0$, $p = 1$]. The results of the two groups are therefore combined in Figure 8, which shows the distribution of z-scored Likert scale responses for each attribute according to the FC condition. A lower score reflects a lower rating for the labels of British, friendly, intelligent, intelligible and trustworthy. The most striking trends observed in Figure 8 are the lower friendliness and trustworthiness ratings for the cotton mask speakers, the lower intelligibility ratings for the niqāb speakers, and the higher British, friendly and intelligent ratings for the niqāb speakers. Despite these observed trends, no significant differences were found for any of the social evaluation ratings, although the scores for intelligibility narrowly missed significance [$\chi^2 = 5.85$, $p = 0.054$]. The coefficients for the other attributes were as follows: friendliness [$\chi^2 = 2.01$, $p = 0.36$], intelligence [$\chi^2 = 0.46$, $p = 0.79$], trustworthiness [$\chi^2 = 3.72$, $p = 0.16$], and British [$\chi^2 = 0.74$, $p = 0.69$].

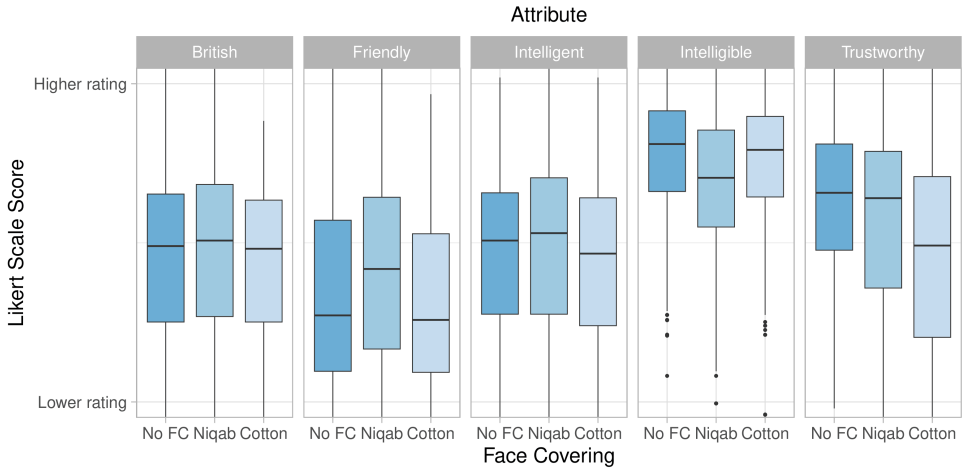

Figure 8 Box plot illustrating the distribution of Likert scale scores z-scored. The FC type is subsetted on the x-axis. The social evaluations are shown along the top of the Figure.

3.2.2 *Reaction times*

Figure 9 shows the distribution of reaction time results according to the three FC conditions. The results reveal that participants took significantly longer to attribute ratings for the niqāb-wearing speakers compared to both the No FC condition (β = 3536, SE = 1324, $p < 0.05$) and the cotton mask condition (β = 3956, SE = 1324, $p < 0.05$). On average, participants took 10,835ms to respond to the no FC videos, 10,415ms to respond to the cotton mask videos, and 14,371ms to respond to the niqāb videos. It therefore took just over 3.5 seconds (3,536ms) longer to respond to the niqāb videos than no FC videos, and almost 4 seconds (3,956ms) longer to respond for niqāb than cotton mask. The difference between No FC and cotton mask components was not significant (β = 421, SE = 1324, $p = 0.76$).

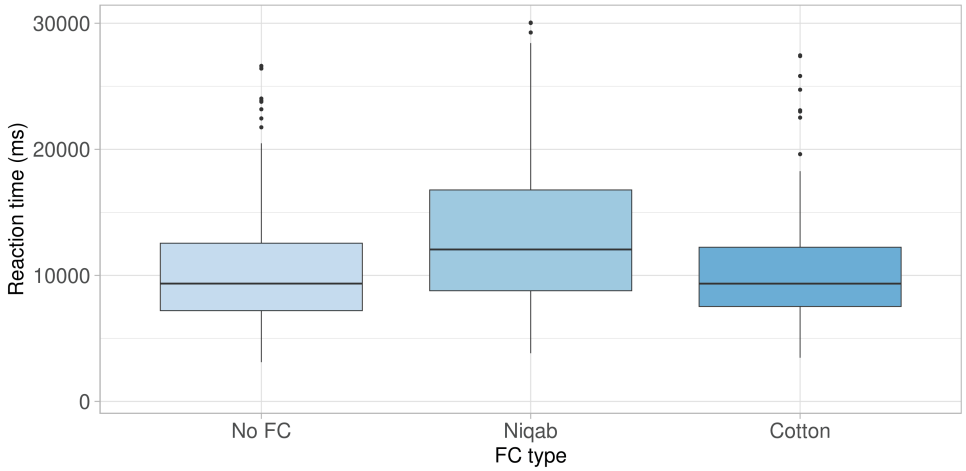

Figure 9 Box plot illustrating the distribution of reaction time results for the social evaluation test.

4 Discussion

We now discuss the results with respect to the three research questions.

4.1 *RQ1: Do the acoustic and visual effects of wearing a face covering affect the comprehension of minimal pair words in unpredictable connected speech heard in background noise?*

4.1.1 *Minimal pair word comprehension*

As indicated in Section 3.1.1, the minimal pair word comprehension results are complex, with one main interaction present and a second reaching near significance. In particular, the significant interaction illustrated that participants were best at comprehending niqāb audio when combined with niqāb visual. This finding is surprising as the niqāb visually blocks access to movement of the lips and mouth. However, prior research suggests that the chiffon material typically used

for niqābs does not cause significant acoustic reduction (Llamas et al. 2008). Furthermore, the niqab-wearing speakers might anticipate comprehension difficulties and therefore hyper-articulate (Traunmüller & Eriksson 2000), increasing the amplitude of the signal. Thus, from an auditory perspective, the niqāb is unlikely to impair the signal relative to the control (no face covering). Alternatively, from a visual perspective, it is reasonable to assume that when participants see a niqāb they expect listening conditions to be challenging as access to the lips and mouth are blocked. In anticipation of greater processing demands, participants therefore compensate by paying closer attention to the audio, which, in reality, is not significantly impaired. It is therefore a combination of these factors which could explain the high-performance accuracy observed for this interaction. This result supports the argument that missing visual cues from the niqāb can be compensated for by the auditory signal, which, in the context of phoneme identification, is unlikely to be impaired by the niqāb (Grant & Seitz 2000). What is surprising, however, is that the niqāb matched condition yields higher correct minimal pair word comprehensions (80.3%) than the cotton mask matched condition (72.6%), despite more of the face being visible in the latter. This difference might be due to greater acoustic impedance from the cotton mask, which has been shown to cause substantial amplitude reduction (Balamurali et al. 2021, Corey et al. 2020). However, this is unlikely due to the amplitude normalisation we applied to our stimuli. An alternative perspective is that listeners might compensate more with the niqāb than with the cotton mask, paying increasing attention to the audio as visual information is reduced. In other words, when presented with the niqāb visual the listener has less visual information and therefore less need/inclination to attend to it. By contrast, more visual information is present with the cotton mask, though it might be difficult to interpret, leading to difficulties reconciling the two channels. This result further develops Grant & Seitz (2000)'s argument about the compensation of auditory or visual cues. It does so by showing that when both auditory and visual cues are missing this hinders speech perception, making it more difficult for listeners to recover lost information from either channel.

4.1.2 Reaction times

Participants took significantly longer to respond to the comprehension sentences when they heard cotton mask speech as opposed to the no face covering and niqāb speech. As noted, a cotton mask causes greater acoustic reduction than a niqāb. The acoustic dampening caused by the cotton mask is likely to be the main factor slowing down listener responses, thereby enhancing our understanding of the relationship between acoustic degradation and speech comprehension.

The different visual components had no effect on reaction times. These results contrast with those of Schwarz et al. (2022), who found that adult participants were significantly slower in responding to a visual face covering compared to no face covering visuals. This discrepancy could be attributed to task differences; in Schwarz et al. (2022), participants had to repeat orally the last word of a sentence,

imposing greater processing demands compared to the binary choice task used in the present study.

4.2 RQ2: Do face coverings impact listeners' social evaluations of a speaker?

4.2.1 Social evaluation scores

The presence of a face covering did not significantly impact the social evaluation scores given to a speaker for any of the five social attributes. These results are somewhat surprising when considering the negative media claims towards face coverings during the COVID-19 pandemic and findings from other research which suggest that visual cues can significantly influence how a speaker is perceived (Kret & De Gelder 2012). It was predicted that significant differences in social evaluations would arise between the niqāb-wearing speakers and those wearing no face covering or a cotton mask. The present study, however, found no such differences. These results can be viewed positively with regards to the use and integration of face coverings in everyday life: our participants did not judge speakers wearing face coverings negatively.

4.2.2 Reaction times

Participants took significantly longer to attribute social evaluations to the speakers wearing a niqāb. The longer reaction times show that listeners were more cautious and uncertain when evaluating niqāb-wearing speakers. This result can be explained by the complex interaction between social and phonetic factors in speech processing (Drager 2010). For example, pairing a standard SSBE accent with a niqāb (which has socially rich indexical cues) results in listeners making more deliberate judgements, as reflected in the longer response times. Campbell-Kibler (2008) further argues that social evaluations can change depending on phonetic cues and vice versa. In our experiment, the niqāb may have created uncertainty in how listeners perceive the SSBE speakers thereby increasing reaction times. An alternative suggestion, however, is that the increased reaction times are because of the experimental design which explicitly highlights the presence of the niqāb. Future research should explore this issue using more indirect methods.

4.3 RQ3: Does increased exposure to face coverings affect listeners' minimal pair word comprehension and social evaluations of a speaker?

The experiment showed that increased exposure to face coverings did not affect minimal pair word comprehension scores, social evaluations or response times. Very limited prior work has investigated the effect of increased exposure to face coverings on speech perception. We asked if a listener could adopt a growth mindset where beliefs towards a face covering could be developed through learning (Dweck 2006). Our results, however, are similar to those of Crinnion et al. (2022), confirming that increased exposure to speech produced through a face covering does not lead to improved accuracy of speech comprehension. The findings of the

present study therefore echo two possibilities suggested by Crinnion et al. (2022): (i) that listeners have already adapted to face covering speech, or (ii) they cannot adapt. With Crinnion et al. (2022) only testing audio stimuli, however, they suggest the possibility that listeners 'need additional information to show improvements in recognition of masked speech' (2022: 11) (e.g. a visual indicator that the talker is wearing a mask). The present study, however, confirms that visual information did not affect the results and therefore can be ruled out as an explanation. The online experimental set up is likely to have made listeners more self-conscious and sensitive in their evaluations meaning future research should test this using more indirect methods.

4.4 Practical implications

The results of the present study are applicable to various sectors. First, with respect to the social implications for veil wearers, the findings show that wearing a niqāb does not negatively affect listeners' social evaluations of a speaker, and nor does it affect word identification. The present findings challenge the stereotypes associated with veil wearers, such as the teaching assistant who was suspended from work because she refused to remove her veil in the classroom (Simpson 2006). It is worth noting, however, that these were SSBE speakers. Niqāb wearers in the UK have a range of regional and often ethnically marked accents. This is likely to impact how the niqāb wearing speakers are perceived with it having been found that phonetic cues can affect the attribution of social characteristics to a speaker (Campbell-Kibler 2008). In other words, the finding that a niqāb does not negatively influence social evaluations of SSBE speakers supports the argumentation that there is an interaction between social and linguistic factors, where listeners' perceptions are influenced by not only the niqāb but also the standard accent and the social meanings attributed to it. Based on these observations, further research is needed to explore potential interactions between face coverings and regional accents (Campbell-Kibler 2008, Drager 2010). Second, the results have direct implications for the healthcare sector, where concerns have been raised over how patients perceive healthcare professionals wearing a face covering (Ford 2020). Our results suggest that interlocutors do not perceive those wearing a face covering differently, although cotton masks do introduce some impedance to the acoustic signal, which can lower word recognition rates in certain conditions. Minimal pair word identification scores also showed that combining niqāb audio and visual components led to an increase in performance accuracy than niqāb visual and no mask audio components. This is promising for the education sector as it provides empirical evidence to suggest that wearing a niqāb is not necessarily problematic for teaching and communication as implied by the suspension of a niqāb wearing teaching assistant (Simpson 2006). It is worth highlighting that the present study was conducted in participants' homes using their personal computers, which may have led to variability in listening environments across participants. That said, steps were taken to mitigate these issues – for example, by including sound checks and providing instructions to complete the study in a quiet environment. Moreover, this

level of variability is still likely to be lower than what typically occurs in real-world communication settings, such as classrooms. A final important direction for future research could involve conducting similar experiments to the ones presented in this paper with participants who have both auditory and visual impairments. Including participants with sensory impairments in future studies would provide a more comprehensive understanding of the effects of face coverings on word identification and social evaluations for the wider population.

5 Conclusion

Our study addressed the following gaps in existing face covering research: (1) the separate contributions of the auditory and visual components associated with speech produced through a face covering and how they work together (cf. Schwarz et al. 2022), (2) how different coverings affect minimal pair word identification in semantically unpredictable sentences with background noise, (3) the impact of face coverings on social evaluations of speakers, and (4) the role of prolonged exposure to face coverings. We found that a face covering causes minimal acoustic degradation in terms of minimal pair word identification. Regarding the relative contributions of audio and visual information, cotton masks did introduce some impedance to the acoustic signal. However, minimal pair word identification accuracy was higher when audio from a niqāb was paired with niqāb visuals, compared to visuals of a speaker without a face covering. Social knowledge did not appear to interact with speech processing, as face coverings did not affect social evaluations. Finally, prolonged exposure to face coverings (as represented by daily versus infrequent exposure) had no impact on the results, suggesting that if speakers adapt to face coverings, this is likely to be a complex process which requires further research. These findings challenge negative media stereotypes and can be used to support the use of face coverings in education and the healthcare sector. Future research should explore how these effects apply in more diverse populations, including individuals with auditory or visual impairments.

Abbreviations

FC	Face covering
NHS	National Health Service
SNR	Sound-to-noise ratio
SSBE	Southern Standard British English

References

Anwyl-Irvine, A. L., J. Massonnié, A. Flitton, N. Kirkham & J. K. Evershed. 2020. Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods* 52(1). 388–407. doi:10.3758/s13428-019-01237-x.

Apple. 2023. iMovie. https://www.apple.com/uk/imovie/.

Atcherson, S. R., L. L. Mendel, W. J. Baltimore, C. Patro, S. Lee, M. Pousson & M. J. Spann. 2017. The Effect of Conventional and Transparent Surgical Masks on Speech Understanding in Individuals with and without Hearing Loss. *Journal of the American Academy of Audiology* 28(1). 58–67. doi:10.3766/jaaa.15151.

Badh, G. & T. Knowles. 2023. Acoustic and perceptual impact of face masks on speech: A scoping review. *PLOS ONE* 18(8). Article e0285009. doi:10.1371/journal.pone.0285009.

Balamurali, B. T., T. Enyi, C. J. Clarke, S. Y. Harn & J.-M. Chen. 2021. Acoustic Effect of Face Mask Design and Material Choice. *Acoustics Australia* 49(3). 505–512. doi:10.1007/s40857-021-00245-2.

Bates, D., M. Mächler, B. Bolker & S. Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1). 1–48. doi:10.18637/jss.v067.i01.

Boersma, P. & D. Weenink. 2023. Praat: Doing Phonetics by Computer. https://www.fon.hum.uva.nl/praat/.

Boone, J. 2020. Are they smiling or frowning behind that face mask? COVID-19 communication is tricky. https://news.vt.edu/content/news_vt_edu/en/articles/2020/10/unirel-covidcommunication.html.

Bottalico, P., S. Murgia, G. E. Puglisi, A. Astolfi & K. I. Kirk. 2020. Effect of masks on speech intelligibility in auralized classrooms. *The Journal of the Acoustical Society of America* 148(5). 2878–2884. doi:10.1121/10.0002450.

Bridges, D., A. Pitiot, M. R. MacAskill & J. W. Peirce. 2020. The timing mega-study: comparing a range of experiment generators, both lab-based and online. *PeerJ* 8. Article: e9414. doi:10.7717/peerj.9414.

Brown, V. A., K. J. Van Engen & J. E. Peelle. 2021. Face mask type affects audiovisual speech intelligibility and subjective listening effort in young and older adults. *Cognitive Research: Principles and Implications* 6(1). Article: 49. doi:10.1186/s41235-021-00314-0.

Brysbaert, M. & M. Stevens. 2018. Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition* 1(1). Article: 9. doi:10.5334/joc.10. Number: 1.

Campbell-Kibler, K. 2008. Accent, (ING), and the social logic of listener perceptions. *American Speech* 82(1). 32–64. doi:10.1215/00031283-2007-002.

Cohn, M., A. Pycha & G. Zellou. 2021. Intelligibility of face-masked speech depends on speaking style: Comparing casual, clear, and emotional speech. *Cognition* 210. Article: 104570. doi:10.1016/j.cognition.2020.104570.

Corey, R. M., U. Jones & A. C. Singer. 2020. Acoustic effects of medical, cloth, and transparent face masks on speech signals. *The Journal of the Acoustical Society of America* 148(4). 2371–2375. doi:10.1121/10.0002279. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7857499/.

Crinnion, A. M., J. C. Toscano & C. M. Toscano. 2022. Effects of experience on recognition of speech produced with a face mask. *Cognitive Research: Principles and Implications* 7. Article: 46. doi:10.1186/s41235-022-00388-4.

Damer, E. & P. Bradley. 2014. Prolific. https://app.prolific.co/researcher/workspaces/projects/6410963e498202e96155603a/draft.

Davies, M. 2007. British National Corpus. http://www.natcorp.ox.ac.uk/.

Drager, K. 2010. Sociophonetic Variation in Speech Perception. *Language and Linguistics Compass* 4(7). 473–480. doi:10.1111/j.1749-818X.2010.00210.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-818X.2010.00210.x.

Dweck, C. S. 2006. *Mindset: The New Psychology of Success*. New York, NY, US: Random House Publishing Group. Pages: x, 276.

Erber, N. P. 1975. Auditory-Visual Perception of Speech. *Journal of Speech and Hearing Disorders* 40(4). 481–492. doi:10.1044/jshd.4004.481.

Fecher, N. 2014. *Effects of forensically-relevant facial concealment on acoustic and perceptual properties of consonants*. York: Doctoral dissertation, University of York dissertation. Https://etheses.whiterose.ac.uk/7397/1/Natalie_Fecher_PhD_2014.pdf.

Fiorella, M. L., G. Cavallaro, V. D. Nicola & N. Quaranta. 2023. Voice Differences When Wearing and Not Wearing a Surgical Mask. *Journal of Voice* 37(3). 467.e1–467.e7. doi:10.1016/j.jvoice.2021.01.026.

Ford, M. 2020. Nurses worry patients 'fear them more' when wearing mask. https://www.nursingtimes.net/news/coronavirus/nurses-worry-patients-fear-them-more-when-wearing-mask-19-05-2020/.

Geng, P., Q. Lu, H. Guo & J. Zeng. 2023. The effects of face mask on speech production and its implication for forensic speaker identification-A cross-linguistic study. *PLOS ONE* 18(3). Article: e0283724. doi:10.1371/journal.pone.0283724.

Gilbody-Dickerson, C. 2020. Coronavirus: Clear masks made to help lip-reading deaf people. https://www.bbc.com/news/uk-england-devon-53251962.

Giovanelli, E., C. Valzolgher, E. Gessa, M. Todeschini & F. Pavani. 2021. Unmasking the Difficulty of Listening to Talkers With Masks: lessons from the COVID-19 pandemic. *i-Perception* 12(2). 1–11. doi:10.1177/2041669521998393.

Glaze, B. 2022. Pandemic babies struggle to understand facial expressions, Ofsted warns. Section: Politics. https://www.mirror.co.uk/news/politics/pandemic-babies-struggle-understand-facial-26626506.

Grant, K. W. & P.-F. Seitz. 2000. The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America* 108(3). 1197–1208. doi:10.1121/1.1288668. Place: US.

Hamzelou, J. 2020. Coronavirus: Should children returning to school wear face coverings? https://www.newscientist.com/article/mg24732983-700-coronavirus-should-children-returning-to-school-wear-face-coverings/.

Harrison, P. 2022. batchCombineSpeechAndNoiseMatchedNoise.praat.

Hay, J., A. Nolan & K. Drager. 2006a. From fush to feesh: Exemplar priming in speech perception. *The Linguistic Review* 23(3). 351–379. doi:10.1515/TLR.2006.014. Place: Germany.

Hay, J., P. Warren & K. Drager. 2006b. Factors influencing speech perception in the context of a merger-in-process. *Journal of Phonetics* 4(34). 458–484. doi:10.1016/j.wocn.2005.10.001.

Ihaka, R. & R. Gentleman. 2023. R: A language and environment for statistical computing. https://www.r-project.org/.

Islabonita. 2013. Pub [.wav]. https://freesound.org/people/Islabonita/sounds/178525/.

Jiang, F., M. L. Ng, Y. Song & Y. Chen. 2024. Effect of Face Masks on Voice Quality Associated with Young and Older Chinese Adult Speakers. *Journal of Voice* Advance online publication. doi:10.1016/j.jvoice.2024.04.025.

Joshi, A., T. Procter & P. A. Kulesz. 2021. COVID-19: Acoustic Measures of Voice in Individuals Wearing Different Facemasks. *Journal of Voice* 37(6). 971.e1–971.e8. doi:10.1016/j.jvoice.2021.06.015.

Kalikow, D. N., K. N. Stevens & L. L. Elliott. 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America* 61(5). 1337–1351. doi:10.1121/1.381436.

Kret, M. & B. De Gelder. 2012. Islamic Headdress Influences How Emotion is Recognized from the Eyes. *Frontiers in Psychology* 3. Article: 110.

Lansing, C. R. & G. W. McConkie. 2003. Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception & Psychophysics* 65(4). 536–552. doi:10.3758/BF03194581. Place: US.

Llamas, C., P. Harrison, D. Donnelly & D. Watt. 2008. Effects of different types of face coverings on speech acoustics and intelligibility. *York Papers in Lingusitics* 2(9). 80–104.

Massaro, D. W. & M. M. Cohen. 1983. Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance* 9(5). 753–771. doi:10.1037/0096-1523.9.5.753. Place: US.

McGurk, H. & J. MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264(5588). 746–748. doi:10.1038/264746a0.

Miller, G. A. & P. E. Nicely. 1955. An Analysis of Perceptual Confusions Among Some English Consonants. *The Journal of the Acoustical Society of America* 27(2). 338–352. doi:10.1121/1.1907526.

Milne, A. E., R. Bianco, K. C. Poole, S. Zhao, A. J. Oxenham, A. J. Billig & M. Chait. 2021. An online headphone screening test based on dichotic pitch. *Behavior Research Methods* 53(4). 1551–1562. doi:10.3758/s13428-020-01514-0.

Nguyen, D. D., P. McCabe, D. Thomas, A. Purcell, M. Doble, D. Novakovic, A. Chacon & C. Madill. 2021. Acoustic voice characteristics with and without wearing a facemask. *Nature* 11(1). 1–11. doi:10.1038/s41598-021-85130-8.

Niedzielski, N. 1999. The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology* 18(1). 62–85. doi:https://doi.org/10.1177/0261927X99018001005.

Noyes, E., J. P. Davis, N. Petrov, K. L. H. Gray & K. L. Ritchie. 2021. The effect of face masks and sunglasses on identity and expression recognition with super-recognizers and typical observers. *Royal Society Open Science* 8(3). 1–18. doi:10.1098/rsos.201169.

Petit, E. 2023. HandBrake. https://handbrake.fr/.

Plichta, B. & D. Preston. 2005. The /ay/s have It: The perception of /ay/ as a north-south stereotype in United States English. *Acta Linguistica Hafniensia* 37(1).

107–130. doi:10.1080/03740463.2005.10416086.

Pycha, A., M. Cohn & G. Zellou. 2022. Face-Masked Speech Intelligibility: The Influence of Speaking Style, Visual Information, and Background Noise. *Frontiers in Communication* 7. Article: 874215. doi:10.3389/fcomm.2022.874215.

Rawlinson, K. 2021. Rise in suspects using face coverings to mask identity, say Kent police. https://www.theguardian.com/uk-news/2021/apr/16/rise-in-suspects-using-face-coverings-to-mask-identity-say-kent-police.

Redford, M. A. & R. L. Diehl. 1999. The relative perceptual distinctiveness of initial and final consonants in CVC syllables. *Journal of the Acoustical Society of America* 106(3). 1555–1565. doi:10.1121/1.427152. Place: US.

Rosenblum, L. 2005. Primacy of multimodal speech perception. In D. Pisoni & R. Remez (eds.), *The Handbook of Speech Perception*, 51–78. Blackwell.

Schwarz, J., K. K. Li, J. H. Sim, Y. Zhang, E. Buchanan-Worster, B. Post, J. L. Gibson & K. McDougall. 2022. Semantic Cues Modulate Children's and Adults' Processing of Audio-Visual Face Mask Speech. *Frontiers in Psychology* 13. Article: 879156. doi:10.3389/fpsyg.2022.879156.

Simpson, M. 2006. The woman at centre of veil case. http://news.bbc.co.uk/1/hi/uk/6068408.stm.

Smiljanic, R., S. Keerstock, K. Meemann & S. M. Ransom. 2021. Face masks and speaking style affect audio-visual word recognition and memory of native and non-native speech. *The Journal of the Acoustical Society of America* 149(6). 4013–4023. doi:10.1121/10.0005191.

Sumby, W. H. & I. Pollack. 1954. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America* 26(2). 212–215. doi:10.1121/1.1907309. Place: US.

Traunmüller, H. & A. Eriksson. 2000. Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America* 107(6). 3438–3451. doi:10.1121/1.429414.

Truong, T. L. & A. Weber. 2021. Intelligibility and recall of sentences spoken by adult and child talkers wearing face masks. *The Journal of the Acoustical Society of America* 150(3). 1674–1681. doi:10.1121/10.0006098.

Vatikiotis-Bateson, E., I.-M. Eigsti, S. Yano & K. G. Munhall. 1998. Eye movement of perceivers during audiovisualspeech perception. *Perception & Psychophysics* 60(6). 926–940. doi:10.3758/BF03211929.

Wang, M. D. & R. C. Bilger. 1973. Consonant confusions in noise: a study of perceptual features. *The Journal of the Acoustical Society of America* 54(5). 1248–1266. doi:10.1121/1.1914417.

Washington Times. 2024. North Carolina votes to ban face masks for medical reasons in public. https://www.washingtontimes.com/news/2024/may/16/north-carolina-votes-to-ban-face-masks-for-medical/.

Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. https://cran.r-project.org/web/packages/ggplot2/citation.html.

Wickham, H., M. Averick, J. Bryan, W. Chang, L. McGowan, R. François, G. Grolemund, A. Hayes, A. Henry, J. Hester, M. Kuhn, T. Pedersen, E. Miller, S. Bache, K. Müller, J. Ooms, D. Robinson, D. Seidel, V. Spinu & H. Yutani. 2019. Welcome

to the tidyverse. https://www.tidyverse.org/packages/.

Wu, Y.-H., E. Stangl, O. Chipara, S. S. Hasan, A. Welhaven & J. Oleson. 2018. Characteristics of Real-World Signal to Noise Ratios and Speech Listening Situations of Older Adults With Mild to Moderate Hearing Loss. *Ear and Hearing* 39(2). 293–304. doi:10.1097/AUD.0000000000000486.

Yi, H., A. Pingsterhaus & W. Song. 2021. Effects of Wearing Face Masks While Using Different Speaking Styles in Noise on Speech Intelligibility During the COVID-19 Pandemic. *Frontiers in Psychology* 12. Article: 682677. doi:10.3389/fpsyg.2021.682677.

Yuan, Y., Y. Lleo, R. Daniel, A. White & Y. Oh. 2021. The Impact of Temporally Coherent Visual Cues on Speech Perception in Complex Auditory Environments. *Frontiers in Neuroscience* 15. Article: 678029. doi:10.3389/fnins.2021.678029.

Zhang, T., M. He, B. Li, C. Zhang & J. Hu. 2022. Acoustic Characteristics of Cantonese Speech Through Protective Facial Coverings. *Journal of Voice* 22(1). 1–9. doi:10.1016/j.jvoice.2022.08.029.

Chloe Patman
University of Cambridge
cep72@cam.ac.uk