frontiers | Frontiers in Communication

ORIGINAL RESEARCH
published: 09 May 2022
doi: 10.3389/fcomm.2022.874215

Face-Masked Speech Intelligibility: The Influence of Speaking Style, Visual Information, and Background Noise

Anne Pycha[1]*, *Michelle Cohn*[2] and *Georgia Zellou*[2]

[1] Department of Linguistics, University of Wisconsin, Milwaukee, WI, United States, [2] Department of Linguistics, University of California, Davis, Davis, CA, United States

The current study investigates the intelligibility of face-masked speech while manipulating speaking style, presence of visual information about the speaker, and level of background noise. Speakers produced sentences while in both face-masked and non-face-masked conditions in clear and casual speaking styles. Two online experiments presented the sentences to listeners in multi-talker babble at different signal-to-noise ratios: −6 dB SNR and −3 dB SNR. Listeners completed a word identification task accompanied by either no visual information or visual information indicating whether the speaker was wearing a face mask or not (congruent with the actual face-masking condition). Across both studies, intelligibility is higher for clear speech. Intelligibility is also higher for face-masked speech, suggesting that speakers adapt their productions to be more intelligible in the presence of a physical barrier, namely a face mask. In addition, intelligibility is boosted when listeners are given visual cues that the speaker is wearing a face mask, but only at higher noise levels. We discuss these findings in terms of theories of speech production and perception.

Keywords: speech production, speech perception, speech intelligibility, face mask, background noise

OPEN ACCESS

Edited by:
Sónia Frota,
University of Lisbon, Portugal

Reviewed by:
Luis Jesus,
University of Aveiro, Portugal
Tim Ziemer,
University of Bremen, Germany
Stefanie Shattuck-Hufnagel,
Massachusetts Institute of Technology, United States

***Correspondence:**
Anne Pycha
pycha@uwm.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *11 February 2022*
Accepted: *19 April 2022*
Published: *09 May 2022*

Citation:
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. doi: 10.3389/fcomm.2022.874215

INTRODUCTION

During the COVID-19 pandemic, face masks became commonplace throughout the world. Despite their efficacy in helping to prevent virus transmission, face masks present an obstacle for speech communication (Bottalico et al., 2020; Hampton et al., 2020; Saunders et al., 2021). To begin with, masks obscure speakers' mouths and therefore deprive listeners of visual cues that can be used to support comprehension (Giovanelli et al., 2021; Truong and Weber, 2021). Even for the audio signal, face masks act as a physical barrier for sound waves and have been shown to reduce signal transmission from the mouth (specifically, a "simulated" mouth consisting of a loudspeaker in a dummy head; Palmiero et al., 2016). In overcoming this communicative challenge, both speakers and listeners might play a role. Speakers, for example, can modulate their speaking style to enhance intelligibility. Listeners, for their part, can make use of additional cues, such as visual information about the face-masked status of the speaker, and they may also adjust their listening strategies in response to signal degradation. In the current study, the goal is to pinpoint the ways in which these speaker and listener adaptations interact during speech communication while wearing a face mask. To that end, the current study investigates the intelligibility of face-masked speech

while manipulating speaking style, availability of visual information about the speaker, and level of background noise. In doing so, this work evaluates adaptation theories of speech production, as well as social and cognitive accounts of speech perception.

Face Masks and Speakers

In everyday conversations, people often speak casually. But when listening conditions are difficult, speakers may adapt by shifting to a "clear" speech style (Lindblom, 1990). In the presence of background noise, for example, speakers' productions become louder, slower, and higher-pitched (the Lombard effect; Lombard, 1911; Brumm and Zollinger, 2011). Clear speech produces intelligibility benefits across a wide range of situations (for review, see Smiljanić and Bradlow, 2009), including face-mask situations. For example, Smiljanić et al. (2021) found that clear speech produced with a face mask increased intelligibility, compared to casual speech produced with or without a face mask. In a similar vein, Yi et al. (2021) found that, across both face-masked and non-face-masked conditions in speech-shaped noise (SSN) and multitalker babble, clear speech was better understood than conversational speech. Furthermore, in an audio-only condition, they found similar word identification accuracy in SSN for clear face-masked speech and conversational non-face-masked speech, suggesting that the clear speech style compensated for the signal degradation from the face mask.

In related work, the current authors have also shown that clear speech style boosts intelligibility in face-masked situations (Cohn et al., 2021), although the pattern of results differed from those of other studies. Crucially, these findings showed that listeners' comprehension accuracy was actually *greater* in a face-masked clear condition than in a non-face-masked clear condition. No such boost occurred for the casual style, which does not demand that the speaker produce clarity; nor did it occur for a positive-emotional speaking style, which does not demand clarity either, but has nevertheless been shown to produce intelligibility benefits for listeners (Dupuis and Pichora-Fuller, 2008). Note that this pattern is inconsistent with *automatic adaptation accounts* of speech production (e.g., Junqua, 1993), which claim that, in the presence of a communication challenge (such as noise, or a face mask), speakers will adapt their productions automatically regardless of speech style. However, this pattern is consistent with *targeted adaptation accounts* (Hazan et al., 2015; Garnier et al., 2018), which claim that speakers adapt to challenges by actively tailoring their productions to specific communicative needs of a given situation; here, the need to speak clearly while also overcoming the physical barrier of the mask.

The current study attempts to replicate the clear vs. casual pattern of speech style results reported by Cohn et al. (2021), but also extend this line of research to investigate how the pattern changes when different demands are made of the listener.

Face Masks and Listeners

While several studies have addressed the role of the speaker in face-masked communication, less is known about the role of the listener. In general, previous research has demonstrated that listener beliefs and behaviors affect their interpretation of the speech signal, and the same can be expected to hold true in face-masked situations. Here, the focus is on two different features that have been shown to influence the listener: their use of visual cues about the speaker, and their response to different levels of signal degradation.

Integrating Cues About the Speaker

Listeners' experiences of speech are shaped by their beliefs about the identity or origin of the speaker. Many studies investigating this issue have asked participants to listen to an audio signal accompanied by pictures of talkers with different apparent ethnic or racial identities. Results have shown that listeners interpret the same speech signal differently, depending upon whether they believe the speaker is foreign-born or native (e.g., Rubin, 1992; McGowan, 2015; Ingvalson et al., 2017).

Two different *social perception* models have been proposed to account for these effects. According to a *bias account*, bias against non-dominant groups reduces attention to the speech signal (Rubin and Smith, 1990; Rubin, 1992; Kang and Rubin, 2009; Lippi-Green, 2011). This model predicts reduced intelligibility for non-dominant speaker groups, correlated with the degree to which they are the object of bias within a particular societal context. In contrast, an *alignment account* proposes that the modulating factor is not bias per se, but rather the fit between social expectations and the signal (Babel and Russell, 2015; McGowan, 2015). This model predicts reduced intelligibility when listeners' expectations about a speaker do not match the speech that they produce, and enhanced intelligibility when they do match, regardless of whether the expectations concern a dominant or a non-dominant group.

The literature contains empirical support for both *bias* and *alignment* theories. Rubin (1992), for example, examined the perception of native-accented American English speech that was accompanied either by a photo of a person with Asian facial features, or by a photo of a person with Caucasian facial features. Despite the fact that the speech samples were the same across conditions, American English listeners showed better comprehension in the Caucasian photo condition, in line with the predictions of the *bias account*. Other studies have also reported reduced intelligibility or increased accentedness ratings for non-dominant social groups, including a Syrian identity presented alongside German speech (Fiedler et al., 2019), an image of a person from Morocco accompanying Dutch speech (Hanulíková, 2018), and an image of a person from South Asia accompanying English speech (Kutlu, 2020). Applying these results to the current study, one potential bias against face-masked speakers is that they are difficult to understand. One would therefore predict speech intelligibility to decrease whenever listeners are presented with an image of a face-masked speaker, compared to an image of non-face-masked speaker.

Several studies have made observations which challenge the *bias account*. McGowan (2015) conducted a study similar to that of Rubin (1992), except that the speech samples consisted of Chinese-accented (specifically, Mandarin-accented) English, rather than native-accented English. Some listener participants had very limited exposure to Chinese-accented English, while other participants were of Chinese-American heritage. Results

for both groups showed that accuracy was higher when speech was accompanied by a photo of a person with Asian facial features, compared to a person with Caucasian facial features. This finding is not compatible with a bias account: if bias against a non-dominant social group reduces attention to the signal, one would not expect better accuracy in the Asian photo condition. Instead, this finding is compatible with an *alignment account*, whereby consistency, or alignment between visual information (here, a photo), and the speech signal leads to better language comprehension. Yi et al. (2013), Babel and Russell (2015), and Gnevsheva (2018) also report findings that are compatible with an *alignment account*. Relatedly, a study by McLaughlin et al. (2022) finds no evidence for implicit racial biases in audio-visual benefits for accented vs. unaccented speech, further challenging a *bias account*. Applying these results to the current study, people plausibly have certain expectations about face-masked speakers (e.g., they produce speech that is sometimes altered by a physical barrier). Under the *alignment account*, one expects enhanced intelligibility whenever listeners are given information about the speaker that supports their expectations.

In many of the studies in this literature, the accompanying images relied upon phenotypical traits determined in large part by genetic factors, such as hair color and facial features, or on apparent region-of-origin (e.g., Niedzielski, 1999; Hay et al., 2006). The images used in the current study are of a different nature, because face masks constitute a transient, non-phenotypical, non-regional characteristic of a speaker. It remains an open question whether such characteristics can also affect speech intelligibility, but at least one study suggests that they might. D'Onofrio (2019) presented participants with audio recordings accompanied by photos of the same individual with different clothing, hairstyle, and facial expressions, and reported that these different stylistic presentations (or "personae") affected lexical recall. In the current study, line drawings of the same individual either with or without a face mask are presented to listeners in order to test whether this affects intelligibility.

Listener Responses to Signal Degradation

In everyday communication, listeners confront many factors that potentially make the speech signal more difficult to understand, such as foreign accents and background noise, as well as face masks. In theory, one might expect each of these factors to affect listener behavior in a simple linear fashion. In reality, the existing literature suggests more complex scenarios. To begin with, the impact of degraded signals extends beyond intelligibility and affects other cognitive variables, such as listener effort. Complicating the picture further, different sources of degradation do not always combine in an additive fashion.

Research on listener effort has focused on speech signals presented in the presence of background noise at different signal-to-noise ratios (SNR). As SNR becomes lower, listeners generally do worse on listening tasks, as expected (e.g., Pichora-Fuller et al., 1995; Fallon et al., 2000). This is true for face-masked speech as well: Toscano and Toscano (2021) found that comprehension accuracy was at ceiling across face-mask conditions at SNR +13 dB, but accuracy was significantly lower for masked speech conditions at −3 dB SNR. Less conspicuously, SNR also affects effort: as SNR becomes lower, listeners give higher ratings of their listening effort (Rudner et al., 2012). Again, the same holds true for face-masked speech: Brown et al. (2021) reported higher effort ratings for face-masked conditions, compared to non-face-masked conditions. In addition to subjective effort ratings, SNR has been shown to modulate pupil responses (Zekveld et al., 2010), recall tasks (Rabbitt, 1966, 1968), and performance on simultaneous non-speech tasks (e.g., Broadbent, 1958; Sarampalis et al., 2009; for an overview, see Strand et al., 2018). These results highlight the fact that listening is not a passive activity, but a complex cognitive behavior, as proposed by *cognitive accounts* (Heald and Nusbaum, 2014).

Research on different sources of degradation underscores a similar point. For example, Smiljanić et al. (2021) examined two such sources: face masks worn by a speaker, and background noise (six-talker babble). Their results showed that in quiet conditions, face-masked speech was just as intelligible as non-face-masked speech (see also Magee et al., 2020). In noisy conditions, however, the presence of a face mask decreased intelligibility compared to the no-mask condition. This suggests that the listeners' experience of signal degradation may have emerged from the specific combination of face-mask plus background noise, rather than by each factor independently.

Complex interactions have also been reported for other types of challenging signals. For example, Adank et al. (2009) asked participants to do a sentence verification task with audio recordings in two different English accents (Southeastern Britain vs. Glasgow) accompanied by three different levels of background noise. Their results show a significant interaction between accent and noise level, suggesting that each accent-plus-noise combination may have placed a unique demand on the listener. van Wijngaarden et al. (2002) and Rogers et al. (2006) report related results. More broadly, Adank (2012) found that while background noise and a non-native accent both led to increased difficulty for listeners, these two sources of degradation correlated with increased activity in different regions of the cortex, suggesting that listeners apply different strategies for comprehending speech-in-noise and foreign accents (see also Van Engen and Peelle, 2014). The takeaway message from this line of work is that each different degradation combination may have the potential to elicit a distinct pattern of listener behavior.

In addition to these considerations, it is also established that SNR interacts with visual information. For example, the audio-visual benefit derived from observing a speaker's lip and face movements varies according to the degree of intelligibility (Ross et al., 2007) and level of background noise (Sumby and Pollack, 1954). Given this previous work using dynamic information as portrayed in video clips, we might also expect that SNR would interact with static visual images of a speaker. The current study pursued these questions of listener behavior by presenting face-masked and non-face-masked speech at two different SNRs. In Experiment 1, we presented stimuli in noise at −6 dB SNR; in Experiment 2, we presented them at −3 dB SNR. We manipulated SNR across experiments, rather than within a single experiment, so that the no-image condition of Experiment 1 could stand alone as a replication of our previous study (Cohn et al., 2021), which was conducted at −6 dB

SNR. From a simple perspective, one might expect the highest levels of comprehension to occur for non-face-masked speech at the higher, potentially easier SNR, and the lowest levels of comprehension for face-masked speech at the lower, potentially more difficult SNR. One might also expect that any advantages conferred by the presence of a visual image would decrease at the easier SNR. However, given the results discussed above, as well as recent findings on speech-style interactions (Cohn et al., 2021), more complex results are anticipated. These findings will speak to theories of speech production and perception with the overarching goal to elucidate the impact of face masks on comprehension during everyday communication.

Current Study and Predictions

Two online experiments reported here investigate intelligibility of American English target words in sentences produced with or without a fabric face mask, across two speaking styles (casual and clear), accompanied by either no image or an image of the speaker (presented as a line drawing). Thus, each experiment crossed three factors, with two levels each: 2 face-mask conditions * 2 speaking styles * 2 image conditions. Sentences were presented in multi-talker babble, at −6 dB SNR (noisier) in Experiment 1 and −3 dB SNR (less noisy) in Experiment 2.

In both experiments, an effect of speech style is predicted, such that sentences produced in clear speech will exhibit higher target-word accuracy rates than those produced in casual speech, in line with prior work (Smiljanić and Bradlow, 2009). Crucially, speech style is also predicted to interact with face-mask conditions. In Experiment 1 at −6 dB SNR, identical to the SNR used in the authors' previous work (Cohn et al., 2021), a replication of the prior finding is expected: that is, face-masked speech should be *more* intelligible than non-face-masked speech in the clear style, with no such effect in the casual style. This pattern would support a *targeted adaptation account* of production (Lindblom, 1990). According to this account, speakers balance production-oriented and listener-oriented factors in order to tune the speech signal to the communication needs of a particular situation. Our previous and currently expected findings support this idea because they suggest that, while speakers do tune their speech for the specific situation of trying to speak clearly while wearing a face mask, they do not make changes in the absence of a defined communicative goal, even when wearing a face mask. In Experiment 2, at −3 dB SNR, an interaction between style and face-masking is also predicted. However, in accordance with *cognitive accounts* (Heald and Nusbaum, 2014), the reduced demands on the listener might allow participants to behave differently toward the speech signal, resulting in a different interaction with speech style than in Experiment 1. For example, given the reduced importance of clear speech in quieter conditions, it is possible that the advantage for clear face-masked speech may be reduced or disappear entirely in Experiment 2.

Also in both experiments, an effect of image is predicted. As proposed by an *alignment account*, overall greater intelligibility for face-masked speech is predicted when the participants also see an image of a masked speaker, because listeners receive visual information about the speaker which is consistent (or "matched") with the signal. Alternatively, the *bias account* would predict overall lower intelligibility when participants see the face-masked image, because listeners may hold a bias against face-masked speakers that they are more difficult to understand.

EXPERIMENT 1: −6 DB SNR

Experiment 1, conducted online, tests the intelligibility of spoken sentences in a 2 (face-mask vs. no-face-mask) * 2 (clear vs. casual speech) * 2 (no image vs. image) design. Sentences were presented in multi-talker babble at −6 dB SNR.

Methods

Participants

Listener participants ($n = 112$) were native English speakers from the United States and undergraduates from University of California, Davis, recruited from the Psychology subjects pool (mean age = 19.45 years, sd = 1.46 years; 86 female, 23 male, 3 non-binary). All participants reported no hearing difficulty.

Auditory Stimuli

A set of 154[1] low-predictability sentences from the Speech-Perception-in-Noise (SPIN) corpus was selected (Kalikow et al., 1977). The full set of the sentences were produced by both a female and male speaker using a head-mounted microphone (Shure WH20XLR)[2], audio mixer (Steinberg UR12), and face masks made of fabric. Speakers produced the same set of sentences (in the same order), first face-masked and then non-face-masked across three styles: in clear and casual speech styles, as well as a third style, positive-emotional, which is not analyzed here in order to constrain the scope of the present work. Each speaker produced the sentences for a real interlocutor (the other speaker), who wrote down the final word of each sentence as it was produced, in light of prior work showing that speakers naturally produce more intelligible speech in the presence of a real interlocutor, vs. an imagined one (Scarborough and Zellou, 2013). Speakers were given explicit instructions about how to produce each style. For clear speech, the instructions were: "In this condition, speak clearly to someone who may have trouble understanding you." For casual speech, the instructions were: "In this condition, say the sentences in a natural, casual manner." The recordings used in the current study are identical to those used in the authors' previous investigation of face-masked speech (Cohn et al., 2021).

Because each style and masking condition was recorded in one long sound file, we force-aligned the productions with the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) to determine consistent boundaries to segment each sentence. **Figure 1** plots the long-term average spectra (LTAS) of the 154 recorded sentences across the four production conditions (2 face-masking conditions * 2 speech styles), calculated (Quené and van Delft, 2010) and plotted with Praat (Boersma and Weenink, 2021) (relative to $2e^{-05}$ Pascal, the default in Praat[3]). Note

[1] Excluding problematic sentences with the keywords "slave" and "clan".

[2] Microphone was located outside of the mask and equidistant from the mouth for the face-masked / non-face-masked conditions.

[3] "The normative auditory threshold for a 1000-Hz sine wave", per Praat documentation. Therefore, values lower than $2e^{-05}$ will have negative values.

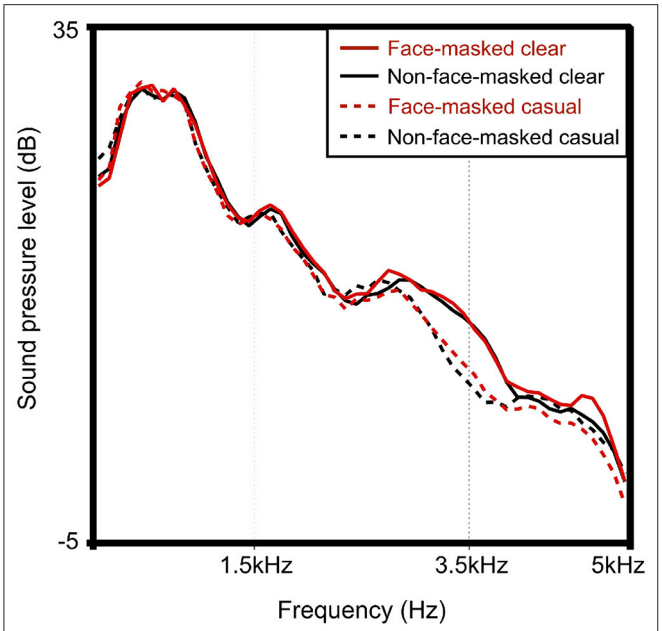

FIGURE 1 | (Color online) Long term average spectra (LTAS) of originally recorded sentences, by face-mask and speaking style condition. The y-axis shows relative sound pressure level (SPL) to $2e^{-05}$ Pascal (the default in Praat). Note this is for sentences prior to the other pre-processing steps (e.g., normalizing intensity to an average of 60 dB and mixing with noise).

that the LTAS was calculated for unmodified sentences (i.e., not intensity normalized). As seen, both clear speech conditions exhibit greater intensity than casual conditions, particularly above 2.5 kHz. Furthermore, within both clear and casual styles, the masked condition exhibits slightly higher intensity at some higher frequencies (2.5–5 kHz) than the unmasked condition.

After each sentence had been segmented from the recording, we normalized the intensities to an average of 60 dB (relative to $2e^{-05}$ Pascal) in Praat. Multi-talker babble (MTB) was created using American English voices generated from Amazon Polly (Joanna, Salli, Joey, Matthew) producing the "Rainbow Passage" (Fairbanks, 1960) [normalized intensity to an average 60 dB (relative to $2e^{-05}$ Pascal) and resampled to 44.1 kHz in Praat]. For each stimulus sentence, a 5-s sample from each Polly voice was randomly selected and mixed into a mono channel. Each sentence was mixed with the unique 4-talker babble recording at −6 dB SNR; the sentence started 500 ms after MTB onset and ended 500 ms before MTB offset. The intensity of each sentence-plus-MTB stimulus was then normalized to 60 dB (relative to $2e^{-05}$ Pascal) in Praat. Additionally, two sound calibration sentences ("Bill heard we asked about the host", "I'm talking about the bench") produced by the two speakers but not included in the SPIN trials, were also normalized in intensity to 60 dB. Normalizing the intensity of all sound files ensured that they would be at a consistent volume throughout the experiment, although it does not reflect the actual SPL (which would vary based on each participants' playback hardware).

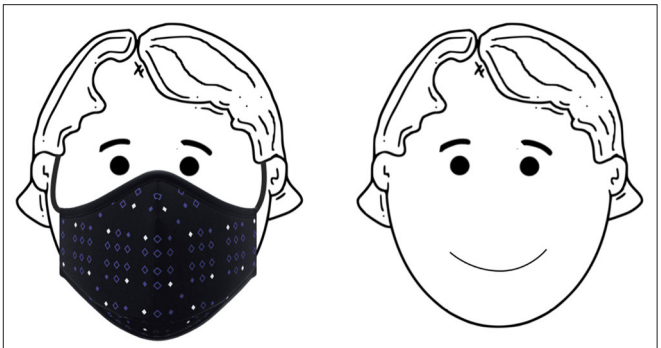

FIGURE 2 | Face-masked and non-face-masked images used as visual information about the speaker.

Picture Stimuli

An open-source line drawing formed the basis of the speaker images (**Figure 2**). In selecting the drawing, the goal was to choose a relatively abstract image, devoid of many specific cues to speaker identity, that could realistically accompany either a male or a female voice. To create the face-masked version of the speaker, an adapted image of a fabric face-mask was pasted onto the drawing.

Procedure

Participants completed the experiment online via Qualtrics. In order to ensure that participants could hear the stimuli properly, the study began with two sound calibration questions. They heard two sentences presented ("Bill heard we asked about the host", "I'm talking about the bench") and were asked to select the correct sentence from a set of options containing phonological competitors of the final word (e.g., "Bill heard we asked about the coast", "Bill heard we asked about the toast"). If they did not select the correct sentence, they were asked to complete the sound calibration again. Once participants passed the calibration procedure, they instructed not to change the volume until the experiment ended[4].

Next, participants were familiarized with the stimuli and the experimental task. A series of instructions introduced them to the noisy background of other talkers, the two target talkers, and the task of typing the final word of each sentence. In situations where the participants were unsure about the final word, they were encouraged to guess.

Two pseudorandomized lists of the SPIN sentences were generated. The first half of the list was randomly presented in either the No-Image block (no picture, 52 trials), or in the Image block (with a picture, 52 trials). In the Image block, listeners were presented with an image of a face (**Figure 2**) that was always congruent with the actual face-masking condition of the recording (i.e., a face-masked picture for face-masked recordings, and a non-face-masked picture for non-face-masked recordings). The second half of the list was randomly presented

[4]Note that while we normalized the intensity of all sound files to 60 dB relative to Praat's default reference level ($2e^{-05}$ Pascal), the actual volume levels varied across participants' machines as this was an at-home, online experiment.

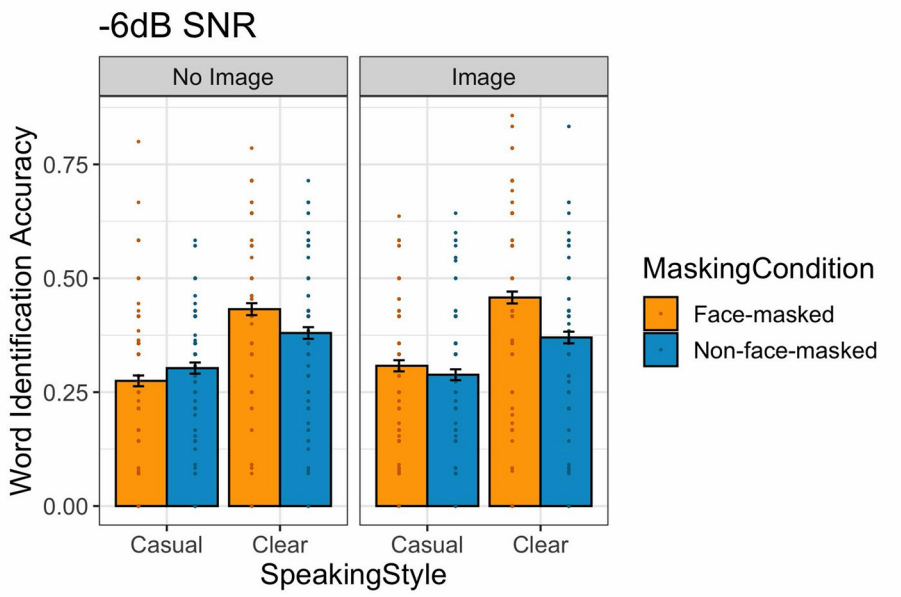

FIGURE 3 | (Color online) Target word identification accuracy for Experiment 1, −6 dB SNR. The bars show the mean for each speech style, face-masking, and image condition. The error bars indicate standard errors of the mean. Individual points show mean accuracy for each participant across conditions.

in the other block. Ordering of blocks (No-Image, Image) were counterbalanced across participants, and list correspondence to the block was counterbalanced across subjects. All subjects heard each sentence once (balanced across speaker, condition, and speaking style). Note that participants were also exposed to a positive-emotional speaking style, not analyzed here.

Thus, for this experiment, each participant heard 104 sentences with MTB at −6 dB SNR. For each trial, participants typed the final word of the sentence.

Analysis

Participants' typed responses for the target words were converted to lowercase and stripped of punctuation and extra spacing, using regex in R (version 4.1.2). Accuracy in target word identification was scored as binomial data (1 = correct, 0 = incorrect), and modeled with a mixed effects logistic regression using the *lme4* R package (Bates et al., 2015). Fixed effects included Face-Masking Condition (face-masked, non-face-masked), Speech Style (clear, casual), Visual Information (no image, image) and all possible interactions. Random effects included by-Participant and by-Speaker random intercepts, as well as by-Participant random slopes for Visual Information, and by-Participant and by-Speaker random slopes for Speaking Style and Face-Masking Condition[5]. Models including by-Listener and/or by-Speaker random slopes for Speaking Style and/or Face-Masking Condition resulted in singularity errors, thus they were dropped from the final model. The retained model lmer syntax is: Accuracy ∼ Face-Masking Condition*Visual Information*Speaking Style + (1+ Visual Information | Listener) + (1 | Speaker).

[5]Note that by-Sentence random intercepts were not included, because the sentences were pseudorandomized. Each sentence was always associated with a particular Visual Information, Speaking Style, and Face-masking Condition across the versions, such that they are not random.

Results

Figure 3 displays word identification accuracy across conditions, and **Table 1** provides the output of the statistical model. The model showed an effect of Face-Masking Condition wherein listeners were more accurate for face-masked speech. There was also an effect of Speaking Style, such that listeners were more accurate at identifying target words for clear speech than for casual speech. Face-Masking Condition also interacted with Visual Information: face-masked speech was *more* intelligible when presented with an image. Face-Masking Condition also interacted with Speaking Style, revealing higher accuracy for face-masked clear speech than the other conditions. No other interactions were observed.

Discussion of Experiment 1

The results of Experiment 1 show that intelligibility is higher for face-masked speech than for non-face-masked speech. On the face of it, this result would seem unexpected, given that face masks act as a physical barrier which reduces speech transmission from the mouth (Palmiero et al., 2016). However, this result is less surprising in light of findings showing that Lombard adjustments result in more intelligible speech in noisy conditions (Junqua, 1993; Lu and Cooke, 2008), which suggests that the speakers who recorded the stimulus sentences made adjustments to overcome the face-mask barrier, and that these adjustments were advantageous for listeners with competing background noise.

The results of Experiment 1 also indicate that intelligibility is higher for clear speech than for casual speech. This finding was

TABLE 1 | Summary statistics for the linear mixed effects model for Experiment 1, −6 dB SNR.

	Coef	SE	z	*p*
(Intercept)	−0.71	0.33	−2.14	0.03
Face-masking condition (face-masked)	0.08	0.02	3.71	<0.001
Visual information (image)	0.02	0.03	0.63	0.53
Speaking style (clear)	0.29	0.02	13.69	<0.001
Face-masking condition (face-masked) * Visual information (image)	0.05	0.02	2.47	0.01
Face-masking condition (face-masked) * Speaking style (clear)	0.08	0.02	3.86	<0.001
Visual information (image) * Speaking style (clear)	−3.2e-03	0.02	−0.15	0.88
Face-masking condition (face-masked) * Visual information (image) * Speaking style (clear)	−0.01	0.02	−0.54	0.59

Num. observations = 11,455, Num. listeners = 112, Num. speakers = 2.

expected, given the clear speech intelligibility benefit (Smiljanić and Bradlow, 2009). Furthermore, intelligibility was higher for face-masked clear speech than for the other conditions. This finding replicates the results of previous work that presented identical stimuli at the same noise level, namely −6 dB SNR (Cohn et al., 2021). This pattern of results supports a *targeted adaptation account* of speech production (e.g., Lindblom, 1990), and suggests speakers actively tailor their productions in response to the communicative situation (here, the need to overcome the barrier of the mask while also following the instructions to speak clearly).

Finally, Experiment 1 shows that intelligibility is higher for face-masked speech in the visual information condition, compared to other conditions. Thus, participants were more accurate when they knew that the speaker was wearing a face mask. This finding provides support for *alignment accounts* (e.g., McGowan, 2015), which claim that listeners benefit from information about speakers, as long as it is consistent with information in the speech signal. Such a finding is difficult to reconcile with *bias accounts* (e.g., Rubin, 1992), which claim that intelligibility decreases when listeners are biased against a speaker (e.g., "people with face masks are hard to understand").

As discussed above, listening is a complex behavior that is actively shaped by the communicative context, and previous work has provided support for this idea by showing that listeners respond to face-masked speech differently at different SNRs (Toscano and Toscano, 2021). Therefore, Experiment 2 tested the factors of speech style, face-masking, and visual information at a higher, less noisy SNR, −3 dB.

EXPERIMENT 2: −3 DB SNR

The design of Experiment 2, also conducted online, was identical to that of Experiment 1. The only difference was that MTB was mixed with the target sentences at −3 dB SNR.

Methods

Participants

One hundred sixteen native English speakers from the United States participated in Experiment 2 (mean age = 19.86 years, sd = 1.94 years; 86 female, 28 male, 2 non-binary). They were recruited through the University of California, Davis Psychology subjects pool. All participants reported no hearing difficulty. None of the participants for Experiment 2 had previously participated in Experiment 1.

Stimuli

Stimuli consisted of the same 154 SPIN recorded sentences in the face-masked and speech style conditions used in Experiment 1. Randomly selected clips of Amazon Polly talkers were generated to create a novel production of 4-talker babble for each sentence (full method described in **Section Auditory Stimuli**). The SPIN sentences were mixed with 4-talker babble at −3 dB SNR and normalized in intensity to 60 dB (relative to $2e^{-05}$ Pascal).

Procedure

The procedure was identical to that in Experiment 1.

Analysis

Accuracy was scored with the same methods as in Experiment 1. A model including by-Listener and/or by-Speaker random slopes for Face-Masking Condition and/or Speaking Style resulted in singularity errors. The retained model lmer syntax is: Face-Masking Condition*Visual Information*Speaking Style + (1+ Visual Information| Listener) + (1 | Speaker).

Results

Figure 4 displays word identification accuracy across conditions, and **Table 2** provides the output of the statistical model. The model revealed an effect of Face-Masking Condition wherein listeners were more accurate for face-masked speech than non-face-masked speech. Additionally, there was an effect of Speaking Style, indicating that listeners were better at identifying words produced in clear speech than causal speech. There were also several interactions. First, Face-Masking Condition interacted with Visual Information, such that face-masked speech was less intelligible in the image condition than in the no-image condition. Additionally, there was an interaction between Face-Masking Condition and Speech Style, where there was less of an increase for clear face-masked speech than for *casual* face-masked speech (recall that the factors were sum coded), seen in **Figure 4**. Finally, Visual Information and Speaking Style

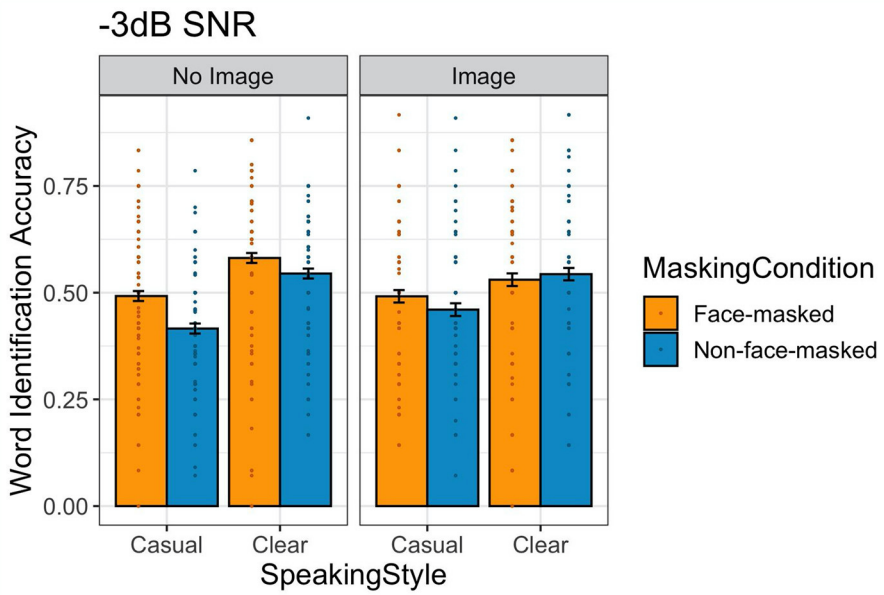

FIGURE 4 | (Color online) Target word identification accuracy for Experiment 2, −3 dB SNR. The bars show the mean for each speech style, face-masking, and image condition. The error bars indicate standard errors of the mean. Individual points show mean accuracy for each participant across conditions.

TABLE 2 | Summary statistics for the linear mixed effects model of Experiment 2, −3 dB SNR.

	Coef	SE	df	z
(Intercept)	0.03	0.29	0.09	0.93
Face-masking condition (face-masked)	0.07	0.02	3.66	<0.001
Visual information (image)	0.01	0.04	0.22	0.83
Speaking style (clear)	0.19	0.02	9.4	<0.001
Face-masking condition (face-masked) * Visual information (image)	−0.05	0.02	−2.6	<0.01
Face-masking condition (face-masked) * Speaking style (clear)	−0.06	0.02	−2.94	<0.01
Visual information (image) * Speaking style (clear)	−0.05	0.02	−2.61	<0.01
Face-masking condition (face-masked) * Visual information (image) * Speaking style (clear)	−0.01	0.02	−0.62	0.53

Num. observations = 11,868, Num. listeners = 116, Num. speakers = 2.

interacted, where accuracy was lower in the image condition for clear speech.

Discussion of Experiment 2

Some of the basic findings of Experiment 2 were similar to those of Experiment 1: intelligibility was higher for face-masked speech compared to non-face-masked speech, as part of an automatic, highly generalized response to a barrier, as proposed by Junqua (1993). Additionally, we see the clear speech intelligibility effect (Smiljanić and Bradlow, 2009), with higher accuracy for clear speech compared to casual speech.

Other results from Experiment 2, however, differ from those of Experiment 1. As revealed by a comparison of **Figures 3**, **4**, overall accuracy was higher in Experiment 2 (noisy at −3 dB SNR), compared to Experiment 1 (noiser at −6 dB SNR). This is an expected outcome which is consistent with previous work examining SNRs (e.g., Pichora-Fuller et al., 1995; Fallon et al., 2000).

In addition to this across-the-board change, the results of Experiment 2 also differ in their patterning. Whereas, in Experiment 1, face-masked clear speech was more intelligible than other conditions, this effect is less apparent in Experiment 2. One finding is that we see a more consistent increase in intelligibility in −3 dB SNR for the face-masked *casual* conditions. It is not immediately apparent why this should be the case. One speculation is that different levels of background noise set up different expectations for listeners. With more background noise, listeners might come to expect a clearer style, because they are aware that the speaker must make adjustments in order to be understood. With less background noise, listeners might come to expect a less clear, potentially more casual style, because they are aware that the conditions are easier for the speaker. Another finding from Experiment 2 concerns the role of visual information, where having additional visual cues that the speakers were masked actually reduced intelligibility, in line with *bias accounts* (e.g., Rubin, 1992). Overall,

Experiment 2 findings suggest that reliance on visual information about speakers decreases in less noisy listening conditions, with weaker intelligibility benefits for both face-masking and clear speech.

POST-HOC ANALYSIS

To directly compare across the two SNRs, we fit a combined model to the accuracy data for both Experiment 1 and 2 data. The model structure was: Accuracy ∼ Face-Masking Condition*Visual Information*Speaking Style*SNR + (1+ Visual Information + Speaking Style | Listener) + (1 | Speaker) (note that a model including by-Listener random slopes for Face-Masking Condition resulted in a singularity error).

A combined plot, showing accuracy across both SNRs, is shown in **Figure 5** and the output of the statistical model is provided in **Table 3**. Results confirmed some of the general findings observed: higher accuracy for clear speech, as well as face-masked speech. Furthermore, as expected, we observe a sizable decrease in intelligibility at the lower SNR, −6 dB. There was also an interaction between Speaking Style and SNR, wherein there was a larger increase for clear speech in the more difficult SNR (-6 dB). This was further mediated by a 3-way interaction with Face-Masking Condition: accuracy was even higher for face-masked clear speech in the −6 dB SNR. Finally, we observed a 3-way interaction between Face-Masking Condition, Visual Information, and SNR, showing higher accuracy with visual information for face-masked speech in the more difficult SNR. No other effects or interactions were observed.

GENERAL DISCUSSION

The current study investigated the interaction of speaker- and listener-related factors in the comprehension of face-masked speech. The general findings, observed across the two experiments, are that intelligibility is higher when speakers wear a face mask and also when speakers use a clear speaking style. Furthermore, intelligibility can be boosted when listeners know that the speaker is wearing a face mask. Together, these observations reveal that speakers and listeners are remarkably flexible in the way they adjust their planning and comprehension processes to fit the real-time communicative context. Notably, these general findings manifested themselves in different patterns depending upon the extent of signal degradation, yielding distinct sets of interactions in a noisier situation (Experiment 1) compared to a less noisy situation (Experiment 2).

In both experiments, participants exhibited better overall performance when listening to face-masked speech. On the face of it, this is a surprising finding, particularly since face masks have been shown to reduce speech signal transmission from a (simulated) mouth by 3–4% (Palmiero et al., 2016). And yet, it is well-established that speakers make adjustments to overcome communication barriers. For example, the Lombard effect demonstrates that speakers change their productions in the presence of background noise (Lombard, 1911; Brumm and Zollinger, 2011) and Lombard-speech is more intelligible to listeners when it is mixed with noise (e.g., Lu and Cooke, 2008). The current results suggest that speakers also adjust their productions in the presence of a different type of barrier, namely a face mask. Furthermore, the fact that the advantage for face-masked speech occurs in both clear and casual speech styles suggests that these adjustments occur regardless of the speaking goal and are therefore, to a certain extent, automatic (Junqua, 1993). This finding is consistent with previous studies on how speakers behave in the presence of noise (Pick et al., 1989) as well as more recent studies of face-masked productions (Asadi et al., 2020).

In both experiments, participants also exhibited better overall performance when listening to clear speech. This is not a surprising finding, since dozens of studies have reported clear-speech advantages across a wide range of experimental conditions (Smiljanić and Bradlow, 2009). Also as expected, overall accuracy was lower in Experiment 1 (noiser at −6 dB SNR) than Experiment 2 (less noisy at −3 dB SNR), consistent with previous work examining SNRs (e.g., Pichora-Fuller et al., 1995; Fallon et al., 2000). However, looking at the interactions, it is also apparent that the intelligibility benefit of clear speech depends upon the listening context. To begin with, the differences between clear and casual speech styles are more apparent in a noisier situation at −6 dB SNR (Experiment 1) compared to a less noisy situation at −3 dB SNR (Experiment 2). The clear speech advantage seems to be stronger depending on the difficulty of the listening condition.

Furthermore, clear and casual speech styles also interact differently with face-masking conditions. This is most apparent in Experiment 1, where the advantage for face-masked speech is strongest in the clear style, but not in the casual style, a pattern which replicates previous results at −6 dB SNR (Cohn et al., 2021). This pattern also suggests that, in response to a given situation, speakers may actually combine automatic adaptations with targeted adaptations. For example, a face mask is a barrier that is present regardless of how the speaker wishes to talk, and regardless of whether the speaker really wishes to be understood. A barrier that exhibits such across-the-board effects may conceivably give rise to automatic adaptations on the part of the speaker, which are not tailored to any particular communicative need, but simply serve to help overcome the barrier. The requirement to be clear, on the other hand, is a specific goal. To accomplish it, the speaker must take into account many situation-specific factors, including not just the speaker's own status (e.g., face-masked or not), but also the status of the listener and the surrounding environment. Such considerations conceivably give rise to targeted adaptations, tailored to the specific needs of the communicative situation. In a face-masked, clear-speech situation then, automatic and targeted adaptations may both be present. If this is the case, it may offer one explanation for why these effects have been so difficult to disentangle in previous work (e.g., Garnier et al., 2018). Indeed, this interpretation is supported by the LTAS of the sentences: face-masked clear speech shows increased amplitude of some of the higher frequencies, suggesting that some targeting was at play. At the same time, the LTAS shows that face-masked casual speech is boosted relative to non-face-masked casual speech.

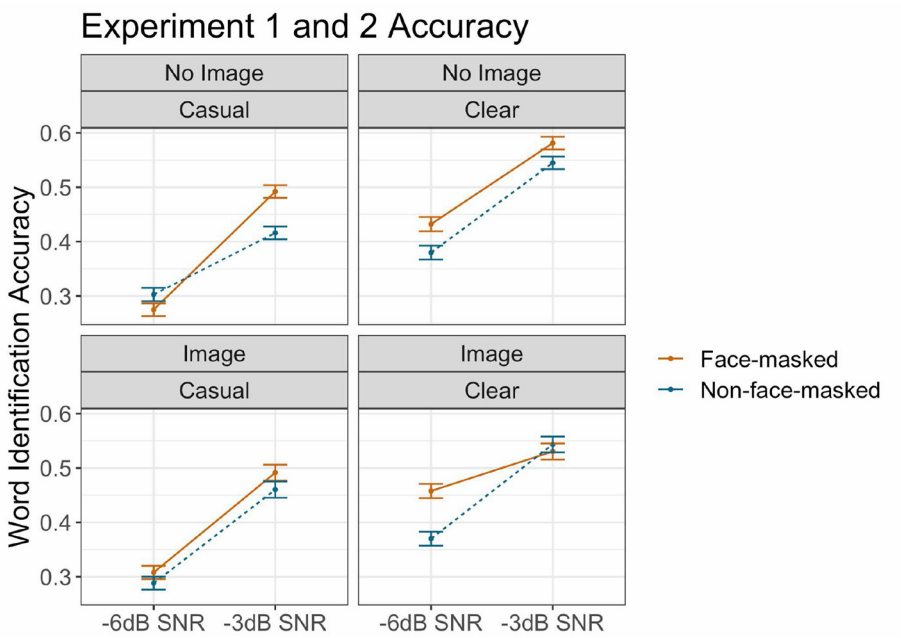

FIGURE 5 | (Color online) Target word identification accuracy for both experiments. The x-axis shows signal-to-noise SNR (−6 dB, −3 dB). The points show the grand mean for each SNR, with the lines indicating differences for SNR for face-masking condition (solid orange line = face-masked; dotted blue line = non-face-masked). Image and speech style conditions are faceted. The error bars indicate standard errors of the mean.

TABLE 3 | Summary statistics for the linear mixed effects model of the combined −3 and −6 dB SNR.

	Coef	SE	df	z
(Intercept)	−0.34	0.31	−1.1	0.27
Face-masking condition (face-masked)	0.07	0.01	5.21	<0.001
Visual information (image)	0.01	0.03	0.58	0.56
Speaking style (clear)	0.24	0.01	16.38	<0.001
SNR (−6 dB)	−0.37	0.03	−10.64	<0.001
Face-masking condition (face-masked) * Visual information (image)	−3.7e-04	0.01	−0.03	0.98
Face-masking condition (face-masked) * Speaking style (clear)	0.01	0.01	0.73	0.46
Visual information (image) * Speaking style (clear)	−0.03	0.01	−1.91	0.06
Face-masking condition (face-masked) * SNR (−6 dB)	1.6e-03	0.01	0.11	0.91
Visual information (image) * SNR (−6 dB)	2.3e-03	0.03	0.09	0.93
Speaking style (clear) * SNR (−6 dB)	0.05	0.01	3.32	<0.001
Face-masking condition (face-masked) * Visual information (image) * Speaking style (clear)	−0.01	0.01	−0.81	0.42
Face-masking condition (face-masked) * Visual information (image) * SNR (−6 dB)	0.05	0.01	3.59	<0.001
Face-masking condition (face-masked) * Speaking style (clear) * SNR (−6 dB)	0.07	0.01	4.75	<0.001
Visual information (image) * Speaking style (clear) * SNR (−6 dB)	0.03	0.01	1.74	0.08
Face-masking condition (face-masked) * Visual information (image) * Speaking style (clear) * SNR (−6 dB)	4.4e-04	0.01	0.03	0.98

Num. observations = 23,323, Num. listeners = 228, Num. speakers = 2.

These boosts appear to occur in the frequency range that tends to be attenuated by the presence of face-masks (above 1 kHz in Corey et al., 2020), suggesting that speakers actively compensate for the barrier.

From the perspective of the listener, the fact that clear and casual speech styles interact differently with face-masking conditions across Experiments 1 and 2 is consistent with the general notion, outlined in the Introduction, that each

combination of signal degradation has the potential to elicit a distinct pattern of behavior (Adank, 2012). For example, the findings of Smiljanić et al. (2021) suggested that face-masking conditions do not necessarily exhibit effects independently of noise conditions. Rather, particular *combinations* of these conditions gave rise to unique patterns of listener behavior. The current study supports this scenario, and, furthermore, shows that it also holds true when we combine different sources of degradation with different speech styles.

The current results also show that intelligibility can be boosted when listeners know that the speaker is wearing a face mask. Specifically, in the noisier situation of −6 dB SNR (Experiment 1), face-masked speech was more intelligible with the visual presentation of a face-masked image. Regardless of the theoretical framework that we adopt, this finding suggests that listeners possess some knowledge about what face masks do to the speech signal, and apply their knowledge ("the speaker is wearing a mask") in their interpretation of the signal. Given the timing of our study and the people who participated in it, this is not surprising. We recruited participants in Fall of 2021, well over a year into the COVID-19 pandemic. Our participants resided in California, a state with some of the strictest masking mandates in the United States. By the time that they listened to the stimuli in the current study, then, they had presumably been listening to masked speech for over a year and a half, and had familiarity with it.

Given this, the findings of Experiment 1 would be difficult to interpret within a bias account, in which listeners' knowledge about face-masked speech gets incorporated into a bias (e.g., "it is too hard to understand"). Under this scenario, knowledge that the speaker is wearing a face mask should lead to lower, not higher, intelligibility. Instead, our result provides support for *alignment accounts* (McGowan, 2015), which predict that comprehension should be easier whenever the characteristics of the speech signal align with social expectations about the speaker. Here, speech signals produced with a face mask aligned with participants' expectations, built up over at least 18 months of listening, about a speaker wearing a face mask.

In contrast, the intelligibility boost did not occur in the less noisy situation of −3 dB SNR (Experiment 2), where the visual image condition exhibited reduced accuracy, compared to the no-image condition. Here, one possibility is that in a relatively less noisy listening task (i.e., −3 dB, relative to −6 dB) in which listeners do not need to exert as much effort, bias effects could emerge. This possibility is consistent with prior work reporting a bias effect in the absence of background noise (Rubin, 1992) but similar intelligibility in more difficult conditions (e.g., −4 dB SNR in McLaughlin et al., 2022). Yet, other work has shown bias effects to persist at more challenging listening conditions (e.g., −10 dB SNR in Fiedler et al., 2019; −4 dB SNR in Yi et al., 2013), suggesting that other factors are also at play (e.g., speaking style). Future work using within-subject comparisons, particularly varying listening difficulty (e.g., *via* SNR levels, types of noise), can further test the reliability of bias effects.

As noted in the Introduction, the visual information in the current study differed from most images used in the previous literature, which tend to highlight "phenotypical" characteristics of a speaker, such as ethnicity or region-of-origin. The current images differed only in the presence vs. absence of a face mask, thereby depicting different transient states of the same speaker, more similar to "personae" (D'Onofrio, 2019). The current results are therefore consistent with an emerging body of work which shows that transient, non-phenotypical information about a speaker also affects the process of speech comprehension (D'Onofrio, 2019). Note that in D'Onofrio (2019)'s work, images of the same individual differed in hair style, facial expression, and clothing, all of which can be chosen by a person to convey social meaning. For settings in which face masks are optional (e.g., at an outdoor concert), the decision to wear a face mask might convey social meaning in a similar manner. However, for settings in which they may be required by government or organizational mandates, (e.g., at a doctor's office, or in a school classroom during a pandemic), the social meaning of a face mask may be largely diminished or absent. The differences between these two kinds of transient characteristics are ripe areas for further investigation in future work.

In the current study, when visual information occurred, it was always congruent with the speech signal. That is, the image of a non-face-masked speaker always accompanied non-face-masked speech, and the image of a face-masked speaker always accompanied face-masked speech. This approach, which has been employed in previous studies (Gnevsheva, 2018), has the advantage of ecological validity, because participants are only exposed to scenarios that are possible in everyday life. Future work examining mismatched guise (e.g., face-masked speech with unmasked image) can further test the role of bias and alignment effects. The current study also used static, black and white line drawings to provide information about the speaker. Future work with photographic images or videos could further explore the role of visual cues to support intelligibility.

An additional limitation of the current study is that participants were all adults. Recent work (Schwarz et al., 2021) has shown that children also exhibit differences in the way they perceive face-masked speech—and might also make different clear speech adaptations to overcome the mask. Furthermore, the study included only one type of face mask, a fabric face-mask. Other types of masks are commonplace in medical environments (e.g., surgical masks) and they have shown to differentially affect speech-in-noise perception (Bottalico et al., 2020; Toscano and Toscano, 2021). Investigating the role of visual information about different types of masks is an avenue for future work.

While output from the mouth is the most important source of acoustic information for speech intelligibility, there is work showing that sound additionally radiates from other parts of a speaker's anatomy that would not be obstructed by a face mask (e.g., the lower eyelids in Abe, 2019). While we suspect any such effects would be negligible in high levels of background noise, as in the present study (−6 dB and −3 dB SNR), this raises interesting questions for future work. In particular, the extent to which speakers' adjustments specifically increase sound radiation from these uncovered areas could shed light on the dynamic types of adjustments speakers make in the presence of communication barriers.

This research also has practical implications for producing and perceiving speech in a face-masked world. In response

to degraded communication situations, face-masked speakers can actively modulate the way they talk, while listeners can adjust their listening strategies. Such findings are relevant for the COVID-19 pandemic, and, in settings such as hospitals and doctors' offices, they will remain relevant well into the future.

DATA AVAILABILITY STATEMENT

The de-identified raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board of University of California, Davis. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AP, MC, and GZ contributed to conception and design of the study and wrote sections of the manuscript. MC programmed the experiment and performed the statistical analysis. AP wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This material is based upon work supported by the National Science Foundation SBE Postdoctoral Research Fellowship to MC under Grant No. 1911855.

ACKNOWLEDGMENTS

The authors thank Melina Sarian for her help with stimulus collection.

REFERENCES

Abe, O. (2019). *Sound Radiation of Singing Voices* (PhD Thesis). University of Hamburg.

Adank, P. (2012). The neural bases of difficult speech comprehension and speech production: two activation likelihood estimation (ALE) meta-analyses. *Brain Lang.* 122, 42–54. doi: 10.1016/j.bandl.2012.04.014

Adank, P., Evans, B. G., Stuart-Smith, J., and Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 520–529. doi: 10.1037/a0013552

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 expiratory activities. *Sci. Rep.* 10, 15665. doi: 10.1038/s41598-020-72798-7

Babel, M., and Russell, J. (2015). Expectations and speech intelligibility. *J. Acoust. Soc. Am.* 137, 2823–2833. doi: 10.1121/1.4919317

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Boersma, P., and Weenink, D. (2021). *Praat: Doing phonetics by computer (version 6.1.40)*. Available online at: https://www.fon.hum.uva.nl/praat/

Bottalico, P., Murgia, S., Puglisi, G. E., Astolfi, A., and Kirk, K. I. (2020). Effect of masks on speech intelligibility in auralized classrooms. *J. Acoust. Soc. Am.* 148, 2878–2884. doi: 10.1121/10.0002450

Broadbent, D. E. (1958). *Perception and Communication*. New York, NY: Pergamon Press.

Brown, V. A., Van Engen, K. J., and Peelle, J. E. (2021). Face mask type affects audiovisual speech intelligibility and subjective listening effort in young and older adults. *Cogn. Res. Princip. Implicat.* 6, 49. doi: 10.1186/s41235-021-00314-0

Brumm, H., and Zollinger, S. A. (2011). The evolution of the Lombard effect: 100 years of psychoacoustic research. *Behaviour* 148, 1173–1198. doi: 10.1163/000579511X605759

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. doi: 10.1016/j.cognition.2020.104570

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. doi: 10.1121/10.0002279

D'Onofrio, A. (2019). Complicating categories: personae mediate racialized expectations of non-native speech. *J. Sociolinguistics* 23, 346–366. doi: 10.1111/josl.12368

Dupuis, K., and Pichora-Fuller, K. (2008). Effects of emotional content and emotional voice on speech intelligibility in younger and older adults. *Can. Acoustics* 36, 114–115. Available online at: https://jcaa.caa-aca.ca/index.php/jcaa/article/view/2064/1811

Fairbanks, G. (1960). "The rainbow passage," in *Voice and Articulation Drillbook*. Vol. 2 (Harper & Row New York), 127p.

Fallon, M., Trehub, S. E., and Schneider, B. A. (2000). Children's perception of speech in multitalker babble. *J. Acoust. Soc. Am.* 108, 3023–3029. doi: 10.1121/1.1323233

Fiedler, S., Keller, C., and Hanulíkov,á, A. (2019). "Social expectations and intelligibility of Arabic-accented speech in noise," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia*, 3085–3089.

Garnier, M., Ménard, L., and Alexandre, B. (2018). Hyper-articulation in Lombard speech: an active communicative strategy to enhance visible speech cues? *J. Acoust. Soc. Am.* 144, 1059–1074. doi: 10.1121/1.5051321

Giovanelli, E., Valzolgher, C., Gessa, E., Todeschini, M., and Pavani, F. (2021). Unmasking the difficulty of listening to talkers with masks: Lessons from the COVID-19 pandemic. *Iperception* 12, 1–11. doi: 10.1177/2041669521998393

Gnevsheva, K. (2018). The expectation mismatch effect in accentedness perception of Asian and Caucasian non-native speakers of English. *Linguistics* 56, 581–598. doi: 10.1515/ling-2018-0006

Hampton, T., Crunkhorn, R., Lowe, N., Bhat, J., Hogg, E., Afifi, W., et al. (2020). The negative impact of wearing personal protective equipment on communication during coronavirus disease 2019. *J. Laryngol. Otol.* 134, 577–581. doi: 10.1017/S0022215120001437

Hanulíková, A. (2018). The effect of perceived ethnicity on spoken text comprehension under clear and adverse listening conditions. *Linguistics Vanguard* 4, 20170029. doi: 10.1515/lingvan-2017-0029

Hay, J., Nolan, A., and Drager, K. (2006). From fush to feesh: exemplar priming in speech perception. *Linguistic Rev.* 23, 351–379. doi: 10.1515/TLR.2006.014

Hazan, V., Uther, M., and Grunland, S. (2015). "How does foreigner-directed speech differ from other forms of listener-directed clear speaking styles?," in *Proceedings of ICPhS 2015. 18th International Congress of Phonetic Sciences* (Glasgow: University of Glasgow).

Heald, S., and Nusbaum, H. (2014). Speech perception as an active cognitive process. *Front. Syst. Neurosci.* 8, 35. doi: 10.3389/fnsys.2014.00035

Ingvalson, E. M., Lansford, K. L., Federova, V., and Fernandez, G. (2017). Listeners' attitudes toward accented talkers uniquely predicts accented speech perception. *J. Acoustical Soc. Am.* 141, EL234–EL238. doi: 10.1121/1.4977583

Junqua, J. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. *J. Acoust. Soc. Am.* 93, 510–524. doi: 10.1121/1.405631

Kalikow, D. N., Stevens, K. N., and Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *J. Acoust. Soc. Am.* 61, 1337–1351. doi: 10.1121/1.381436

Kang, O., and Rubin, D. L. (2009). Reverse linguistic stereotyping: measuring the effect of listener expectations on speech evaluation. *J. Lang. Soc. Psychol.* 28, 441–456. doi: 10.1177/0261927X09341950

Kutlu, E. (2020). Now you see me, now you mishear me: Raciolinguistic accounts of speech perception in different English varieties. *J. Multilingual Multicult. Dev*. doi: 10.1080/01434632.2020.1835929

Lindblom, B. (1990). "Explaining phonetic variation: a sketch of the H&H theory," in *Speech Production and Speech Modelling*, Hardcastle, W. J. and Marchal, A., editors (Dordrecht: Springer), 403–439.

Lippi-Green, R. (2011). *English With an Accent: Language, Ideology, and Discrimination in the United States*. 2nd Edn. New York, NY Routledge.

Lombard, É. (1911). Le signe de l'élévation de la voix. *Annales Des Maladies de l'Oreille et Du Larynx* 37, 101–119.

Lu, Y., and Cooke, M. (2008). Speech production modifications produced by competing talkers, babble, and stationary noise. *J. Acoust. Soc. Am.* 124, 3261–3275. doi: 10.1121/1.2990705

Magee, M., Lewis, C., Noffs, G., Reece, H., Chan, J. C. S., Zaga, C. J., et al. (2020). Effects of face masks on acoustic analysis and speech perception: Implications for peri-pandemic protocols. *J. Acoust. Soc. Am.* 148, 3562–3568. doi: 10.1121/10.0002873

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal forced aligner: trainable text-speech alignment using Kaldi. *Interspeech* 2017, 498–502. doi: 10.21437/Interspeech.2017-1386

McGowan, K. B. (2015). Social expectation improves speech perception in noise. *Lang. Speech* 58, 502–521. doi: 10.1177/0023830914565191

McLaughlin, D. J., Brown, V. A., Carraturo, S., and Van Engen, K. J. (2022). Revisiting the relationship between implicit racial bias and audiovisual benefit for nonnative-accented speech. *Attenti. Percept. Psychophys*. doi: 10.3758/s13414-021-02423-w

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *J. Lang. Soc. Psychol.* 18, 62–85. doi: 10.1177/0261927X99018001005

Palmiero, A. J., Symons, D., Morgan, J. W., and Shaffer, R. E. (2016). Speech intelligibility assessment of protective facemasks and air-purifying respirators. *J. Occup. Environ. Hyg.* 13, 960–968. doi: 10.1080/15459624.2016.1200723

Pichora-Fuller, M. K., Schneider, B. A., and Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *J. Acoust. Soc. Am.* 97, 593–608. doi: 10.1121/1.412282

Pick, H. L., Siegel, G. M., Fox, P. W., Garber, S. R., and Kearney, J. K. (1989). Inhibiting the Lombard effect. *J. Acoust. Soc. Am.* 85, 894–900. doi: 10.1121/1.397561

Quené, H., and van Delft, L. E. (2010). Non-native durational patterns decrease speech intelligibility. *Speech Commun.* 52, 911–918. doi: 10.1016/j.specom.2010.03.005

Rabbitt, P. (1966). Recognition: Memory for words correctly heard in noise. *Psychon. Sci.* 6, 383–384. doi: 10.3758/BF03330948

Rabbitt, P. (1968). Channel-capacity, intelligibility and immediate memory. *Q. J. Exp. Psychol.* 20, 241–248. doi: 10.1080/14640746808400158

Rogers, C. L., Lister, J. J., Febo, D. M., Besing, J. M., and Abrams, H. B. (2006). Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. *Appl. Psycholinguist.* 27, 465–485. doi: 10.1017/S014271640606036X

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., and Foxe, J. J. (2007). Do you see what i am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex* 17, 1147–1153. doi: 10.1093/cercor/bhl024

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Res. High. Educ.* 33, 511–531. doi: 10.1007/BF00973770

Rubin, D. L., and Smith, K. A. (1990). Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative English-speaking teaching assistants. *Int. J. Intercult. Relat.* 14, 337–353. doi: 10.1016/0147-1767(90)90019-S

Rudner, M., Lunner, T., Behrens, T., Thorén, E. S., and Rönnberg, J. (2012). Working memory capacity may influence perceived effort during aided speech recognition in noise. *J. Am. Acad. Audiol.* 23, 577–589. doi: 10.3766/jaaa.23.7.7

Sarampalis, A., Kalluri, S., Edwards, B., and Hafter, E. (2009). Objective measures of listening effort: effects of background noise and noise reduction. *J. Speech Lang. Hearing Res.* 52, 1230–1240. doi: 10.1044/1092-4388(2009/08-0111)

Saunders, G. H., Jackson, I. R., and Visram, A. S. (2021). Impacts of face coverings on communication: an indirect impact of COVID-19. *Int. J. Audiol.* 60, 495–506. doi: 10.1080/14992027.2020.1851401

Scarborough, R., and Zellou, G. (2013). Clarity in communication: "Clear" speech authenticity and lexical neighborhood density effects in speech production and perception. *J. Acoust. Soc. Am.* 134, 3793–3807. doi: 10.1121/1.4824120

Schwarz, J., Li, K., Sim, J. H., Zhang, Y., Buchanan-Worster, E., Post, B., et al. (2021). Speech perception through face masks by children and adults. *Cambridge Language Sciences Annual Symposium*. doi: 10.33774/coe-2021-l88qk

Smiljanić, R., and Bradlow, A. R. (2009). Speaking and hearing clearly: talker and listener factors in speaking style changes. *Lang. Linguist. Compass* 3, 236–264. doi: 10.1111/j.1749-818X.2008.00112.x

Smiljanić, R., Keerstock, S., Meemann, K., and Ransom, S. M. (2021). Face masks and speaking style affect audio-visual word recognition and memory of native and non-native speech. *J. Acoust. Soc. Am.* 149, 4013–4023. doi: 10.1121/10.0005191

Strand, J. F., Brown, V. A., Merchant, M. B., Brown, H. E., and Smith, J. (2018). Measuring listening effort: Convergent validity, sensitivity, and links With cognitive and personality measures. *J. Speech Lang. Hear. Res.* 61, 1463–1486. doi: 10.1044/2018_JSLHR-H-17-0257

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Toscano, J. C., and Toscano, C. M. (2021). Effects of face masks on speech recognition in multi-talker babble noise. *PLoS ONE* 16, e0246842. doi: 10.1371/journal.pone.0246842

Truong, T. L., and Weber, A. (2021). Intelligibility and recall of sentences spoken by adult and child talkers wearing face masks. *J. Acoust. Soc. Am.* 150, 1674–1681. doi: 10.1121/10.0006098

Van Engen, K. J., and Peelle, J. E. (2014). Listening effort and accented speech. *Front. Hum. Neurosci.* 8, 577. doi: 10.3389/fnhum.2014.00577

van Wijngaarden, S. J., Steeneken, H. J. M., and Houtgast, T. (2002). Quantifying the intelligibility of speech in noise for non-native listeners. *J. Acoust. Soc. Am.* 111, 1906–1916. doi: 10.1121/1.1456928

Yi, H., Pingsterhaus, A., and Song, W. (2021). Effects of wearing face masks while using different speaking styles in noise on speech intelligibility during the COVID-19 pandemic. *Front. Psychol.* 12, 682677. doi: 10.3389/fpsyg.2021.682677

Yi, H.-G., Phelps, J. E. B., Smiljani,ć, R., and Chandrasekaran, B. (2013). Reduced efficiency of audiovisual integration for nonnative speech. *J. Acoust. Soc. Am.* 134, EL387–EL393. doi: 10.1121/1.4822320

Zekveld, A. A., Kramer, S. E., and Festen, J. M. (2010). Pupil response as an indication of effortful listening: the influence of sentence intelligibility. *Ear Hear.* 31, 480–490. doi: 10.1097/AUD.0b013e3181d4f251

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.