Intelligibility of face-masked speech depends on speaking style: Comparing casual, clear, and emotional speech

Michelle Cohn, Anne Pycha, Georgia Zellou

Abstract
This study investigates the impact of wearing a fabric face mask on speech comprehension, an underexplored topic that can inform theories of speech production. Speakers produced sentences in three speech styles (casual, clear, positive-emotional) while in both face-masked and non-face-masked conditions. Listeners were most accurate at word identification in multi-talker babble for sentences produced in clear speech, and less accurate for casual speech (with emotional speech accuracy numerically in between). In the clear speaking style, face-masked speech was actually more intelligible than non-face-masked speech, suggesting that speakers make clarity adjustments specifically for face masks. In contrast, in the emotional condition, face-masked speech was less intelligible than non-face-masked speech, and in the casual condition, no difference was observed, suggesting that ‘emotional’ and ‘casual’ speech are not styles produced with the explicit intent to be intelligible to listeners. These findings are discussed in terms of automatic and targeted speech adaptation accounts.

Keywords: face-masked speech, models of speech production, speech-in-noise word comprehension

1. Introduction
Due to the rapid spread of the novel coronavirus (SARS-CoV-2), wearing face masks in public has become increasingly commonplace throughout the world (Matusiak et al., 2020). Despite their health advantages, face masks have the potential to make everyday communication significantly more difficult. Most obviously, face masks obscure the talker’s mouth, depriving listeners of important visual cues. Less conspicuously to the casual observer, face masks also alter the acoustic signal (Bond et al., 1989; Corey et al., 2020; Fecher & Watt, 2011; Saeidi et al., 2016; Saigusa, 2017), reducing speech transmission by an estimated 3-4% (Palmiero et al., 2016). Given these findings, one straightforward prediction is that listeners should experience reduced intelligibility of speech produced with a face mask, relative to speech produced without one. Yet previous studies exploring this issue have yielded decidedly mixed results, raising theoretical questions about what face-masked speakers do to ensure that they are understood.

Only a handful of studies, with very small groups of participants, have previously investigated the impact of face masks on speech intelligibility. In some studies, accuracy for speech produced with a face mask was lower than for non-face-masked speech (Winch et al., 2013; Wittum et al., 2013); in another, intelligibility was lower for speech produced with surgical masks only when it was presented with multi-talker babble, a more difficult listening condition (Fecher & Watt, 2013). Yet, other studies report no differences in intelligibility for similar types of masks (e.g., surgical masks in Atcherson et al., 2017; Radonovich et al., 2009; Thomas et al., 2011), or even an improvement in intelligibility for speech produced with a face mask (Mendel et al., 2008). At first glance, this is surprising: shouldn’t a face mask make speech comprehension more difficult for the listener? Yet, almost none of those studies instructed talkers about how to speak; critically, it is possible that without explicit instruction, talkers might have adjusted their speech for the face-masked condition (e.g., Mendel et al., 2008). Understanding such adjustments is important for theories of cognition because human speech is a remarkably durable system of communication that, despite the wide range of environmental conditions present in everyday life, generally succeeds (Assmann & Summerfield, 2004). Yet, pinpointing exactly how and why it manages to succeed — particularly when confronted with a relatively novel barrier to communication, such as face masks — remains an ongoing challenge for language researchers.

The current study addresses this issue by explicitly manipulating speakers’ style of speech while wearing a fabric face mask, common in the COVID-19 pandemic (MacIntyre & Hasanain, 2020). Our focus is solely on the auditory domain. In particular, we compare intelligibility for face-masked speech produced
in three explicit styles: ‘casual’, ‘clear’, and ‘positive-emotional’. Both clear and emotional speech contain acoustic features which suggest they require greater articulatory effort (e.g., higher intensity and pitch variation in emotional speech; Laukka et al., 2005), relative to ‘casual’ speech. Also, both ‘clear’ and ‘emotional’ styles have been shown to improve listeners’ comprehension relative to less effortful, ‘casual’ styles (Bradlow & Bent, 2002; Dupuis & Pichora-Fuller, 2008; Gordon & Ancheta, 2017; Picheny et al., 1985; Smiljanić & Bradlow, 2009). Despite this similarity, the two styles nevertheless have different goals: ‘positive emotional’ styles are used to convey speaker sentiment, while ‘clear’ styles are produced with the specific intent to be more intelligible.

Investigating face-masked speech with explicit speech-style comparisons can serve as a test for different theories of speech production by examining whether speakers’ adaptations — and subsequent listener intelligibility — are shaped by automatic versus clarity-targeted adjustments in difficult listening situations.

On the one hand, automatic adaptation accounts propose that speakers automatically produce more effortful speech in response to communication barriers. For example, speakers produce overall louder, slower, and higher-pitched speech in the presence of background noise (the ‘Lombard’ effect; Lombard, 1911; for a review see Brumm & Zolinger, 2011). Many findings suggest that Lombard speech is an automatic reflex in response to the speaker’s inability to hear themselves (Junqua, 1993): speakers have difficulty suppressing the effect (Pick et al., 1989) and exhibit it even in non-interactive contexts (Egan, 1972); furthermore, preschool children (presumably less attuned to listener needs) exhibit it (Siegel et al., 1976), as do monkeys (Sinnott et al., 1975). Lombard adjustments also appear to benefit the listener: in quiet, Lombard speech is more intelligible, relative to non-Lombard speech (Lu & Cooke, 2008). Parallel to Lombard effects, recent work found that speakers spoke more loudly when wearing any type of mask (surgical, fabric, etc.) relative to when they were non-face-masked (Asadi et al., 2020), reflecting more effortful speech. In the current study, an automatic account predicts that face-mask-wearing would lead speakers to increase articulatory effort while speaking, relative to non-face-masked speech, producing gains in intelligibility across all three styles (‘casual’, ‘clear’, ‘positive-emotional’).

On the other hand, targeted adaptation accounts propose that speakers actively control their productions based on the situation-specific needs for clarity. For example, some propose that Lombard adjustments are targeted (for discussion, see Garnier et al., 2018) evidenced by differences in speakers’ ‘clear’ speech based on type of interference (Garnier et al., 2006; Hazan et al., 2012; Hazan & Baker, 2011). Speakers also adjust their ‘clear’ speech according to (apparent) perceptual needs of the interlocutor (Zellou & Scarborough, 2015) in multidimensional and tailored ways (Bradlow, 2002; Scarborough & Zellou, 2013). Broadly, these findings are in line with the ‘Hypo- and Hyper-Articulation’ (H&H) theory (Lindblom, 1990), which proposes that adaptations to improve clarity are under the speaker’s active control, weighing intelligibility for the listener in real-time against the articulatory effort required. If the speaker judges that the communicative context is difficult, they might exert greater articulatory effort by producing ‘clearer’, hyper-articulated speech. Otherwise, speakers preserve articulatory effort, producing segmental variants that are less enhanced for intelligibility.

The current study examines how intelligibility is shaped by both a communicative barrier (face mask) and explicit instructions to be clear. The targeted adaptation account predicts that face-masked speech contains adjustments for intelligibility, relative to non-face-masked speech, that differ depending upon speaking style. For one, we would not predict intelligibility enhancements for ‘casual’ speech styles in either masking condition. Additionally, we expect that in the ‘clear’ speech style alone, face-masked speech should be equally — or even more — intelligible than non-face-masked speech. To make this claim, however, the inclusion of ‘emotional’ speech is critical to evaluate whether speaker adaptations are indeed actively targeted for clarity, or whether they are merely an automatic consequence of more effortful speech produced in certain styles (e.g., Laukka et al., 2005). Here, we predict that ‘positive-emotional’ speech will be more intelligible than ‘casual’ speech (in line with Dupuis & Pichora-Fuller, 2008) but, unlike ‘clear’ speech, it will not show targeted clarity enhancements in face-masked conditions.

Our approach uses listener comprehension as a method of probing speaker behavior. An alternative approach would be to focus on acoustic analyses of speaker output, but such analyses severely limit the
amount of data that can be evaluated. Furthermore, as Goldinger (1998) points out, acoustic characteristics do not always correlate with perceptual processes, so the psychological validity of such analyses remains unknown.

2. Methods
2.1. Stimuli
A set of 156 low-predictability sentences were selected from the Speech Perception in Noise (SPIN) corpus (Kalikow et al., 1977). Two native English (US) speaking adults, one male and one female (one of whom is a trained research assistant in the UC Davis Phonetics Lab), produced the sentences. Due to COVID-19 social distancing measures, speakers were from the same household and recorded the stimuli in a quiet room in their home using a head-mounted microphone (Shure WH20XLR) and USB audio mixer (Steinberg UR12). Speakers produced sentences to a real listener (the other speaker, who wrote down each final target word as it was produced) while following an online Qualtrics survey, which presented instructions for whether they would be wearing a fabric mask or not (e.g., “Now take off your mask”) and for each speaking style (Speech style instructions provided in Table 1).

<table>
<thead>
<tr>
<th>Table 1. Speech style instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clear</td>
</tr>
<tr>
<td>Casual</td>
</tr>
<tr>
<td>Positive-Emotional</td>
</tr>
</tbody>
</table>

Speaking styles were collected on different days (clear, casual, then positive-emotional). In each style, speakers began with the masked condition, followed by the non-face-masked condition, which allowed them to keep the microphone in the same location. Speakers were recorded reading each sentence (in the same order, presented one at a time; 44.1 kHz sampling rate). Each speaker produced each sentence six times: 2 face mask conditions x 3 speaking styles. Acoustic measurements (by style and face mask condition) are shown in Table 2. As seen, there is no across-the-board pattern for the three styles that distinguishes between face-masked versus non-face-masked speech, which suggests that if speakers did indeed make compensations for the presence of the mask, they did not do so uniformly for every speaking style.

<table>
<thead>
<tr>
<th>Table 2. Means (and standard deviations) of sentence acoustics by speech style and face-masking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Casual</td>
</tr>
</tbody>
</table>


In order to reduce ceiling effects that might obscure differences across face-masked versus non-face-masked speech, sentences were presented in a difficult listening condition: in multitalker babble (MTB) at a challenging signal to noise ratio (SNR) (Cohn & Zellou, 2020). MTB was generated with 2 female and 2 male Amazon Polly voices (US-English: Joanna, Salli, Joey, Matthew) producing the Rainbow Passage (Fairbanks, 1960) (normalized to 60 dB SPL and resampled to 44.1 kHz in Praat). For each target sentence, a 5-second sample from each Polly voice was randomly selected and mixed into a mono channel. Each target sentence was gated into the unique 4-talker babble recording (SNR = -6 dB), starting 500 ms after noise onset and ending 500 ms before noise offset. Finally, the overall intensity of the sentence (in noise) was amplitude-normalized (60 dB).

2.2. Participants
Participants consisted of 63 native speakers of American English (mean age: 20±1.4 years, range 18-25), with no reported hearing impairments, recruited from the UC Davis Psychology Subject Pool, who received course credit for participation. The study was approved by the UC Davis Institutional Review Board (IRB) and subjects completed informed consent before participating.

2.3. Procedure
The experiment, conducted online using Qualtrics, began with a sound calibration procedure: participants heard one sentence produced by each speaker (not used in experimental trials), presented in silence at 60 dB, and were asked to identify the sentence from three multiple choice options, each containing a phonologically close target word. After, they were instructed to not adjust their sound levels again during the experiment. Next, participants completed the speech-in-noise word identification task. On each trial, participants heard a stimulus sentence (presented once only) and then typed the final word. Assignment of sentences to a Speaker, Face Mask Condition, and Speaking Style was pseudo-randomized across 4 lists.

---

1 As in previous studies of face-masked speech (e.g., Fecher & Watt, 2013) our stimuli were not recorded in noise. The absence of noise during production was crucial in order to prevent Lombard-like adjustments that would have confounded an investigation into masked-speech adjustments. In interpreting these results, the mismatch (i.e., noise present in perception, but not in production) should be borne in mind, although we note that the mismatch was identical for all experimental conditions.
and participants were randomly assigned to one of these lists. In total, each listener heard each of the 156 sentences once.

3. Analysis & Results

Keyword accuracy on each trial was coded binomially (1 = correct word identification, 0 = incorrect) (spelling errors were classified as incorrect) and modeled with a mixed-effects logistic regression with *lme4* R package (Bates et al., 2015). Estimates for *p*-values were computed using the *lmerTest* package (Kuznetsova et al., 2015). Fixed effects included Face Mask Condition (face-masked, non-face-masked), Speech Style (casual, emotional, clear), and their interaction. Random effects included by-Listener and by-Speaker random intercepts (models including more complex random effects structure, e.g. by-Listener random slopes for Style and Mask Condition, following Barr et al. (2013), resulted in singularity errors). Contrasts were sum coded.

Table 3 presents summary statistics for the model2. While there was no main effect of Face Mask Condition, there was an effect of Speech Style: overall, listeners showed higher accuracy for ‘clear’ speech. Face Mask Condition also interacted with Speech Style, which is illustrated in Figure 1: listeners were more accurate for face-masked ‘clear’ speech, relative to non-face-masked ‘clear’ speech. The opposite effect was seen for ‘emotional’ speech: lower accuracy for face-masked ‘emotional’ speech, relative to ‘emotional’ non-face-masked. The releveled model (reference level = ‘emotional’) showed an effect of Speech Style: lower accuracy overall for ‘casual’ speech [$\beta$=-2.7, $SE$=0.03, $z$=-8.4, $p<0.001$]; as seen in Figure 1, there was no effect of Speaking Style for ‘casual’ speech [$\beta$=-4.5, $SE$=3.3, $z$=-1.4, $p=0.17$].

<table>
<thead>
<tr>
<th>Table 3. Model outputs for keyword accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Coeff.</strong></td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>FaceMaskCondition(masked)</td>
</tr>
<tr>
<td>Style(emotional)</td>
</tr>
<tr>
<td>Style(clear)</td>
</tr>
<tr>
<td>FaceMaskCondition(masked)*Style(emotional)</td>
</tr>
<tr>
<td>FaceMaskCondition(masked)*Style(clear)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effects</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listener (Intercept)</td>
<td>0.34</td>
</tr>
<tr>
<td>Speaker (Intercept)</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Num. observations (n=9,819), listeners (n=63), speakers (n=2)

---

2 Overall intelligibility levels are relatively low; keyword identification in MTB was difficult, as intended.
4. Discussion
Results from the current study revealed that wearing a fabric face mask does not uniformly affect speech intelligibility across styles. This observation is consistent with targeted clarity adaptation accounts that speakers are dynamically assessing listener difficulty and adapting their ‘clear’ speech accordingly (e.g., Garnier et al., 2018; Hazan & Baker, 2011; Lindblom, 1990). Extending prior work (e.g., Atcherson et al., 2017; Radonovich et al., 2009), the current study more comprehensively examines the perception of face-masked speech, including a larger group of listeners and explicitly comparing intelligibility across speaking styles. The findings in the present study indicate that speakers make different speech style adaptations when they wear a face mask (compared to when they are non-face-masked) and these adjustments consequently affect intelligibility for listeners. In fact, when speakers produce clear speech while wearing a face mask, their utterances are more accurately understood by listeners than non-face-masked clear speech, consistent with the otherwise surprising results reported by Mendel et al. (2008).

Word comprehension accuracy for ‘emotional’ speech also varied by face mask condition. Yet, the effect is the reverse from that seen for clear speech: face-masked ‘emotional’ speech is less intelligible than non-face-masked ‘emotional’ speech. Here, speakers do not appear to compensate for face mask-wearing, leading to a reduction in word comprehension accuracy in the face-masked condition. Meanwhile, non-face-masked emotional speech is still more intelligible than non-face-masked casual speech. Taken together, this suggests that emotional speech probably does involve increased articulatory effort leading to increased intelligibility (consistent with previous work, Dupuis & Pichora-Fuller, 2014). However, intelligibility of face masked ‘emotional’ speech does not supersede that of non-face-masked ‘emotional’ speech, strongly suggesting that speaker productions in the ‘clear’ condition — where face masked speech does supersede non-face-masked speech — are not only more effortful, but also specifically targeted for clarity. Still, the effect of emotion on intelligibility (e.g., impact of smiling on speakers’ phoneme distinctions, variation by type of emotion) remains an open question for future work. Meanwhile, there are no significant differences in ‘casual’ speech for face-masked versus non-face-masked conditions observed in the present study. This pattern also supports predictions from targeted adaptation accounts (e.g., Lindblom, 1990; Garnier et al., 2018): when pressure to be intelligible is reduced, speakers do not make efforts to compensate for the effect of the face mask.

Our interpretation is that the presence of the face mask, a physical barrier between the speech production apparatus and the listener, leads speakers to produce speech that is even more intelligible than
in the absence of a face mask. These results have possible practical applications: by instructing talkers to “speak clearly” when wearing a face mask we find that speakers are even better understood, contra some public messaging (Goldin et al., 2020). Still, the extent to which this finding generalizes across different communicative scenarios (e.g., presence of noise during speech production) and the acoustic source of this effect remain open questions. Future studies can explore the extent to which increased vocal effort (indicated by increased intensity) might influence overall intelligibility (similar to spectral tilt improvements for Lombard speech in Lu & Cooke, 2008) in tandem with other acoustic differences.

As noted earlier, previous work suggests that face masks affect intelligibility of certain types of speech sounds more than others. For example, face masks attenuate high frequency information (Corey et al., 2020), which might disproportionately impact the acoustic realization of fricatives. Face masks could also alter certain types of articulations more than others, e.g. by preventing full lip protrusion for labials or reducing aspiration for voiceless stops. The current study took a holistic approach to understanding the impact of face-mask wearing and speech style on intelligibility, thus, it does not allow us to fully distinguish between these ‘physical impediment’ effects and targeted adjustments that speakers might make for clarity, an area open for future work. Note, however, that if physical impediments were the sole driver of our results, we would expect to see across-the-board effects for face-masked speech (relative to non-face-masked) in all three styles. Instead, our results show that the face-masked condition crucially interacts with style, supporting targeted adaptation accounts.

This paper examined solely the auditory domain, and we leave the question of audio-visual perception open for future work. From a theoretical perspective, visual information can exert independent effects on speech intelligibility (e.g., Babel & Russell, 2015), so it is important to first isolate effects which arise from acoustics alone, as we do here. From a practical perspective, life during a pandemic presents plenty of face-masked listening situations that are primarily auditory.

We have shown that, despite the acoustic distortion produced by fabric face masks, changes in speech style can produce marked improvements in intelligibility for listeners. Our study therefore provides a demonstration that it is speakers themselves, and the real-time adaptations they make for listeners, who make essential contributions to the robustness of human speech.

5. Acknowledgements
Thank you to Editor Sarah Brown-Schmidt and three anonymous reviewers for their thorough and insightful feedback on this paper. Thank you to Melina Sarian for her help on the stimulus collection. This material is based upon work supported by the National Science Foundation SBE Postdoctoral Research Fellowship to MC under Grant No. 1911855.

6. References


http://dx.doi.org/10.21437/Interspeech.2020-1336


