UNMASKING THE TRUF: IMPACT OF COMMUNITY MASKS ON THE PERCEPTION OF VOICELESS FRICATIVES IN ENGLISH

Massimiliano Canzi & Tamara Rathcke

University of Konstanz
{massimiliano.canzi, tamara.rathcke}@uni-konstanz.de

ABSTRACT

The current study aims at quantifying the effects of wearing a face mask on speech perception, by investigating performance of native English listeners in a phoneme monitoring task with monosyllabic words containing voiceless fricatives. Previous experimental work on the topic has mainly focussed on the effects of acoustic filtering caused by the use of face coverings with mixed results and weak effects of mask wearing on speech perception. In this experiment, we explore the interplay of acoustic filtering with other potentially relevant factors such as the presence of visual cues, lexical frequency and listener-specific background. We provide evidence that suggests the impact of face coverings (esp. FFP-2 face mask) on speech perception is not directly moderated by the acoustic properties of masked speech. Rather, it is inked to an interplay of audio-visual integration, the absence of visual cues for (some) target fricatives, and the listener-specific sociolinguistic background.

Keywords: face mask, fricative perception, th-fronting, lexical frequency, COG, intensity

1. INTRODUCTION

Following the initial outbreak of the SARS-CoV-2 pandemic in March 2020, wearing face masks had become obligatory in public places around the world. There exist many anecdotal reports of speech perception being negatively affected by the presence of a face mask cf. [1]. In theory, such negative impact of a face mask might arise for several reasons. Speech perception is multimodal and it relies on a number of cues [2], including visual cues which are lacking in masked speech. The visual modality is particularly important for the perception of labial and dental sounds whose articulation has more readily visible articulatory movements (e.g. [3]). Such visual cues are arguably less crucial for the perception of alveolar or post-alveolar sounds [4]. Moreover, a face mask can be considered a filter to the acoustic properties of speech, due to the air stream passing through the fabric [5] [4].

There has been a long-standing interest in the impact of culturally motivated face coverings on speech acoustics and perception in forensic contexts [4] [6] [7]. For example, the study by [7] analysed recordings of 9 speakers (F = 5) with seven types of face coverings, including surgical masks, motorcycle helmets, ski and cycling masks. The speakers produced words of the CVC structure, with either onsets or codas containing /f/, /θ/ or /ʃ/. Presence of a face covering was shown to affect the centre of gravity (COG) in all fricatives, especially (labio)dentals. Spectral skewness and curtosis were also affected in all fricatives while only dentals additionally showed a large modification of the spectral peak.

Other studies focused on the perceptual effects of face coverings [6] [8]. For example, the impact of of niqāb, balaclava and a surgical mask was examined with regards to the perception of voicing, manner and place of articulation of stops [6]. Some subtle misperceptions were observed in place of articulation of fricatives (especially /f/ vs. /θ/) and nasals (esp. /n/ and /ŋ/), though overall, a surprisingly small number of misperceptions was attested. In contrast to the study of the isolated sound perception [6], the experiment in [8] examined listeners' transcriptions of whole words and sentences produced by speakers wearing different types of face masks (N95, surgical, cloth). The impact on intelligibility was found to be rather negligible and also comparable across all face masks (3-5%).

A combined effect of the presence of face masks or other face coverings and adverse listening conditions (like an impoverished signal-to-noise ratio, SNR) has also been examined [7] [9]. Overall, the use of surgical or FFP-2 masks shows little negative impact on perception in quiet listening conditions. In lower SNR-levels, an audio-visual presentation can give a 10% boost in intelligibility as compared to the same signal presented as audio-only [7]. Speech perception can be similarly affected in older and younger adults, though older listeners tend to perform slightly worse on the intelligibility tasks in noise and find listening to masked speech more strenuous and effortful [9]. Surprisingly though, a face mask with a transparent window to expose

speaker mouth does not seem to generally promote intelligibility [9].

Overall, existing studies demonstrate that despite subjective beliefs that masked speech is more difficult to understand [1], the documented effects on speech perception seem to be rather subtle (if at all present). However, few studies examined how speech intelligibility may be affected by an interplay of the acoustic filtering effects and the absence of visual cues arising due to face masks, lexical frequency and listener-specific background. The present study was based on the following three hypotheses:

- Acoustic degradation (due to filtering [4]-[5]) along with the absence of visual cues may be jointly contributing to the lower level of intelligibility in masked speech.

- Lexical frequency and lexical competition may moderate the intelligibility effect, with high-frequency words and no minimal pairs causing less difficulties than low-frequency words and the presence of minimal pairs.

- Listener-specific background (specifically the presence of sound mergers in their variety, e.g. due to th-fronting [10]) may further mediate the perception of masked speech.

2. METHODS

2.1. Participants

Seventy native speakers of English (F = 44; Age: Mo = 20-29, R = 18-65) volunteered to participate. The participants were hailing from nine different English-speaking countries, including Australia, Canada, England, India, Ireland, New Zealand, Scotland, Singapore and the United States. The participants from England and Scotland were coded as th-fronters due to the exposure to /θ/-/f/ merger in these varieties [10]. The participants were recruited via social media and could decide to enter a prize draw in recognition of their time and efforts. They were screened for visual and hearing impairments and reported normal hearing and normal or corrected-to-normal vision.

2.2. Stimuli

The target fricatives of the present study included /f/, /θ/, /s/ and /ʃ/. They were chosen because of the presence (/f/, /θ/) or the absence (/s/, /ʃ/) of visual cues during their production. Moreover, the discriminability of /θ/ and /f/ is reduced in some varieties of English, due to a fricative merger [10].

A total of 100 monosyllabic English word stimuli were devised for the purposes of the experiment. Five minimal quadruplets contained target fricatives in onset positions (e.g. *fie/thigh/shy/sigh*). Ten minimal pairs were created with targets in coda positions (e.g. *deaf/death* and *mess/mesh*). Forty monosyllabic words contained the target fricatives without a lexical competitor (e.g. *fact,shame*). In addition, 20 monosyllabic distractors without a target fricative were created. Lexical frequency measures were obtained for all stimuli from the SUBTLEX-UK database [11]. A total of 100 experimental items were recorded audio-visually in two experimental conditions (mask, no mask) by one white, male, native speaker of Southern British English in his thirties. For the masked speech condition, the speaker was wearing a FFP-2 mask that was commonly used in Germany at the time of the recording. The mask prevents listeners from observing any lip movements of the speaker.

2.3. Procedure

The study utilized an online platform called Gorilla [12]. Participants were instructed to perform the experiment in a quiet room using only wired or built-in mouse and keyboard, headphones, or speakers. All stimuli were presented audio-visually and twice, once in the no mask condition and once in the mask condition. Each participant was presented with two distinct blocks of stimuli, one without a mask and one with a mask. The order of presentation of the two blocks was counterbalanced across participants. Within each block, the order of all stimuli was fully randomized. Participants were tasked with monitoring the stimuli for the occurrence of one of the four fricatives and were prompted to respond as quickly as possible by clicking on the corresponding button (F, TH, S, and SH) shown on the experimental screen. If none of the four fricatives were detected, participants were instructed to click on the X button located in the center of the screen. Following the phoneme monitoring task, participants completed a questionnaire that included general demographic questions and queries about their experience wearing face masks in public.

2.4. Data treatment

Data collected in the phoneme monitoring task were analysed with R (v4.2.1) [13] using RStudio [14]. Two main outcome variables were selected: (1) accuracy (correct, incorrect) which refers to whether or not the participant correctly selected the right phoneme in each trial and (2) reaction times (measured in ms and logarithmically transformed to assume a log-normal distribution [15]). In compari-

son to the binary variable of accuracy, reaction times might provide a more nuanced look at the data, by providing an indication of changes in cognitive load and processing ease across conditions [16]. For the statistical analysis, only reaction times of correct answers were selected.

A stepwise backwards model fitting procedure was deployed, by using model comparisons (through functions such as `anova` and `drop1`) [13] [17] [18] to obtain best-fit models. Post-hoc comparisons of the relevant factor levels were carried out using the `emmeans` (v1.8.0) package [19].

For accuracy, a mixed logistic regression model was devised using the lme4 (v1.1.3) package [17]. Participants and stimuli were added as random intercepts. Slopes were allowed over the random intercept of participant where appropriate. Reaction time data were analysed using linear mixed-effects regression, again with package lme4 [17]. Varying intercepts were allowed for participant, item and browser. Slopes were allowed over participant where appropriate.

3. RESULTS

3.1. Acoustics of masked speech

We tested for acoustic differences in the spoken stimuli across all four fricatives as well as across conditions to see whether acoustic properties such as COG and intensity should be added to our main, perceptual analysis. We ran two linear models and we found a significant interaction of experimental condition and fricative for both COG and intensity. Post-hoc comparisons [19] show that mask wearing affects COG the most in /f/ [t(76) = -8.91, $p < .0001$] followed by /θ/ [t(76) = -5.18, $p < .0001$]. However, COG is not affected by mask wearing for both /ʃ/ [t(76) = -0.79, n.s.] and /s/ [t(76) = -1.77 , n.s.]. Conversely, /s/ appears to be the one fricative where intensity is the most affected by mask wearing [t(76) = -9.68, $p < .0001$], followed by /f/ [t(76) = -7.03, $p < .0001$], /ʃ/ [t(76) = -5.16, $p < .0001$] and /θ/ [t(76) = -3.117, $p < .001$]. All the p-values are corrected for multiple comparisons within the `emmeans` package.

3.2. Accuracy

To test the effects of mask wearing on phoneme monitoring accuracy, we devised a generalised linear logistic regression model. The best-fitting model included the interaction of experimental condition (mask, no mask) and target (F, TH, S, SH) [χ^2 = 53.51, $p < .0001$]. COG [χ^2 = 8.72, $p < .005$] and lexical frequency [χ^2 = 8.00, $p < .005$] entered as

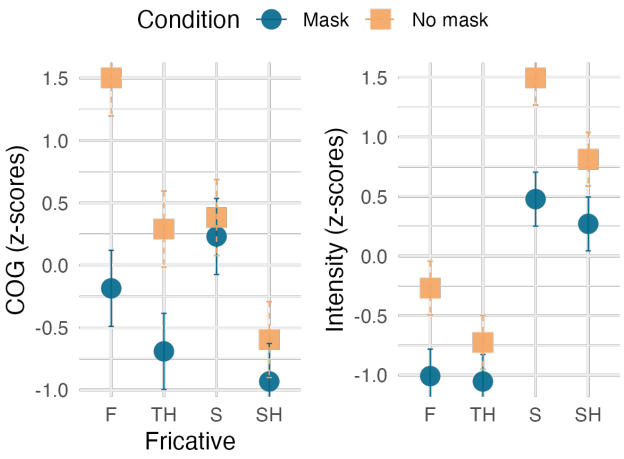

Figure 1: Model estimates of COG and intensity by experimental condition across all fricatives

main effects. Interactions of these factors with the experimental condition were not significant. Pairwise post-hoc comparisons for the interaction of interest reveal that the effect of condition is only significant for non-sibilants that normally have visual cues in the no-mask condition. The difference is particularly large for /θ/ (z = -10.01, $p < .0001$) compared to /f/ (z = -6.838, $p < .0001$). However, /f/ has the lowest accuracy rate across all conditions. The main effect of COG shows a negative relationship between accuracy and COG. The main effect of lexical frequency suggests that higher frequency goes hand-in-hand with higher accuracy scores.

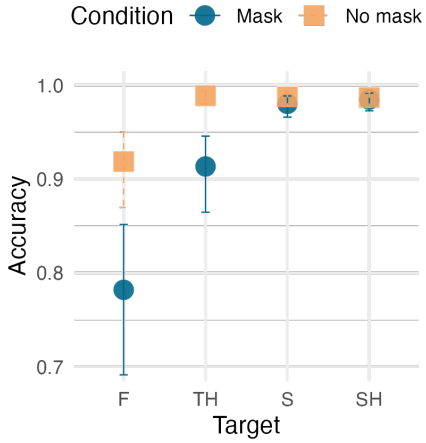

Figure 2: Model estimates of accuracy by experimental condition across all target fricatives

We tested the effect of th-fronting on non-sibilant monitoring accuracy, by devising a supplementary model with only /θ/ and /f/ as targets and with participants grouped into varieties that present th-fronting (England, Scotland) and those that do not (all other countries). The best-fitting model included

the non-significant interaction of experimental condition and merger [χ^2 = 0.57, n.s.], as well as the significant interaction of experimental condition and target [χ^2 = 22.01, p < .0001]. COG [χ^2 = 14.60, p < .001] and log-transformed lexical frequency [χ^2 = 5.7, p < .0001] also entered as main effects. Participant and item were allowed as varying intercepts. We found no effect of th-fronting on accuracy specifically. We then decided to not distinguish participants by the merger status in their variety for the main analysis.

3.3. Reaction Times

To determine whether reaction times were also affected by masked speech during phoneme monitoring, we fit a linear mixed regression model to the log-transformed reaction times of correct responses. The best-fitting reaction-time model included significant main effects of experimental condition [χ^2 = 9.14, p < .005], target fricative [χ^2 = 44.19, p < .0001], device (mobile, computer) [χ^2 = 17.52, p < .0001] and log frequency [χ^2 = 7.91, p < .005]. However, no significant interaction between condition and target was discovered, suggesting that differences in reaction times happen uniformly across all target fricatives. We ran post-hoc analyses using the `emmeans` package to determine marginal means and the direction of effects. In particular, reactions times appear to be higher for the mask (β = 6.45, SE = .04) than the no mask (β = 6.37, SE = .04) condition. Overall, sibilants appear to be identified significantly faster than non sibilants [z = -6.35, p = .0001].

We also tested the effects of th-fronting on reaction times of correct responses in a separate model that only included /θ/ and /f/ as target fricatives. The best-fitting model, among significant main effects of experimental condition, COG, log-transformed lexical frequency and target fricative, included a main effect of th-fronting [χ^2 = 3.85, p = .001]. However, no significant interaction of th-fronting and experimental condition was discovered. This finding suggests that participants whose English varieties present instances of th-fronting are generally characterised by slower reaction times. However, they are not impacted by masked speech any differently than participants who speak English varieties where th-fronting is not as prevalent.

4. DISCUSSION

The present study was designed to determine the effects of FFP-2 masks on the acoustics and the perception of voiceless fricatives in English. The acoustics of /f/ (COG) and /s/ (Intensity) are affected more

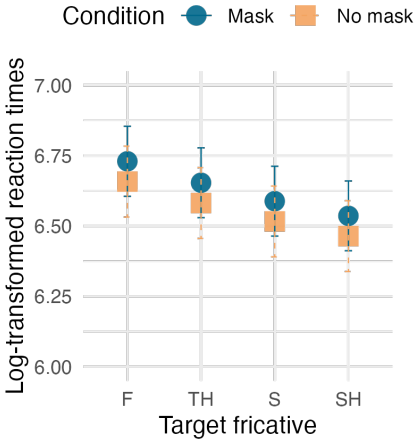

Figure 3: Model estimates of log-transformed reaction times by experimental condition across all target fricatives

than the acoustics of /θ/ and /ʃ/. However, perceptually, the fricatives most affected by a face mask are /θ/ and /f/ (not at all /s/, despite having the largest change in intensity overall). These findings suggest that the source of the effect is not (primarily) in the acoustic modification of masked speech (cf. [6], [8]). Rather, it is a combination of the lack of visual cues in (labio)dentals and a relatively low level of acoustic energy in non-sibilants. [20].

The experiment found that lexical frequency significantly predicted overall higher rates of phoneme monitoring accuracy and lower reaction times for correct responses. However, this is equally true for both experimental conditions. The presence of lexical competition did not significantly affect either the accuracy or reaction times. It is possible that the results were affected by non-uniform distribution of (competitor) stimuli across the lexical frequency range.

Among all fricatives, /θ/ is most affected by a face mask, while /f/ is also not recognised at ceiling in the unmasked audio-visual condition. The lower accuracy rates are not, however, a consequence of th-fronting in varieties in which /f/ maps onto both /f/ and /θ/. Rather, the result might be to do with the ambiguously encoded (labio)dental contrast in English [21].

This work contributes to existing knowledge on masked-speech perception by integrating previous work on the effects of acoustic filtering with factors such as the absence of visual cues, lexical frequency, and listener-specific background. Future research should investigate the effects of face masks on a wider range of phonemes and cross-cultural aspects of audio-visual integration that may mediate speech perception.

5. ACKNOWLEDGEMENTS

We thank George Walkden for his contribution to the recording of all experimental stimuli. We are also grateful to the students of the *Speech Perception* class (Aikaterini Tsaroucha, Hanna Kim and Katharina Hölzl) for their help in collecting experimental data.

6. REFERENCES

[1] Goldin, A., Weinstein, B., Shiman, N., others, 2020. How do medical masks degrade speech perception. *Hearing review* 27(5), 8–9.

[2] Partan, S., Marler, P. 1999. Communication goes multimodal. *Science* 283(5406), 1272–1273.

[3] McGurk, H., MacDonald, J. 1976. Hearing lips and seeing voices. *Nature* 264(5588), 746–748.

[4] Fecher, N., Watt, D. 2011. Speaking under cover: The effect of face-concealing garments on spectral properties of fricatives. *ICPhs*, 663–666.

[5] Coniam, D. 2005. The impact of wearing a face mask in a high-stakes oral examination: An exploratory post-sars study in hong kong. *Language Assessment Quarterly: An International Journal* 2(4), 235–261.

[6] Llamas, C., Harrison, P., Donnelly, D., Watt, D. 2009. Effects of different types of face coverings on speech acoustics and intelligibility.

[7] Fecher, N. 2014. *Effects of forensically-relevant facial concealment on acoustic and perceptual properties of consonants*. PhD thesis University of York.

[8] Magee, M., Lewis, C., Noffs, G., Reece, H., Chan, J. C., Zaga, C. J., Paynter, C., Birchall, O., Rojas Azocar, S., Ediriweera, A., others, 2020. Effects of face masks on acoustic analysis and speech perception: Implications for peri-pandemic protocols. *The Journal of the Acoustical Society of America* 148(6), 3562–3568.

[9] Brown, V. A., Van Engen, K. J., Peelle, J. E. 2021. Face mask type affects audiovisual speech intelligibility and subjective listening effort in young and older adults. *Cognitive Research: Principles and Implications* 6(1), 1–12.

[10] Levon, E., Fox, S. 2014. Social salience and the sociolinguistic monitor: A case study of ing and th-fronting in britain. *Journal of English Linguistics* 42(3), 185–217.

[11] Van Heuven, W. J., Mandera, P., Keuleers, E., Brysbaert, M. 2014. Subtlex-uk: A new and improved word frequency database for british english. *Quarterly journal of experimental psychology* 67(6), 1176–1190.

[12] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., Evershed, J. K. 2020. Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods* 52(1), 388–407.

[13] Team, R. C., others, 2013. R: A language and environment for statistical computing.

[14] Racine, J. S. 2012. Rstudio: a platform-independent ide for r and sweave.

[15] Baayen, R. H., Milin, P. 2010. Analyzing reaction times. *International Journal of Psychological Research* 3(2), 12–28.

[16] Eggemeier, F. T. 1988. Properties of workload assessment techniques. In: *Advances in psychology* volume 52. Elsevier, 41–62.

[17] Bates, D. M. 2010. lme4: Mixed-effects modeling with r.

[18] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. 2017. lmertest package: tests in linear mixed effects models. *Journal of statistical software* 82, 1–26.

[19] Lenth, R., Singmann, H., Love, J., Buerkner, P., Herve, M. 2018. Emmeans: Estimated marginal means, aka least-squares means. *R package version* 1(1), 3.

[20] Jongman, A., Wayland, R., Wong, S. 2000. Acoustic characteristics of english fricatives. *The Journal of the Acoustical Society of America* 108(3), 1252–1263.

[21] Babel, M., McGuire, G. 2013. Listener expectations and gender bias in nonsibilant fricative perception. *Phonetica* 70(1-2), 117–151.