DOI: 10.61782/fa.2023.0158

forum acusticum 2023

EFFECTS OF FACE-MASKED SPEECH ON SHORT-TERM MEMORY

Cleopatra Christina Moshona[1*] **André Fiebig**[1]

[1] Engineering Acoustics, Institute of Fluid Dynamics and Technical Acoustics, Technische Universität Berlin, Berlin, Germany

ABSTRACT

Recent findings suggest that face masks may have a negative impact on memory encoding. However, these findings have not been sufficiently validated and it remains unclear whether the reason for the decrease in encoding performance is the lack of visual or auditory cues. The present study examines the effect of face masks on short-term memory performance by disentangling the role of auditory cues and examining the effects of speech style. For this purpose, 33 German native listeners were presented with audio recordings of a native speaker uttering sentences with and without a face mask in conversational and naturally elicited Lombard speech and completed a cued-recall task. Results show that recall performance in the face-masked conditions tended to be lower than in the baseline, though this effect was small and only significant for the "doubly adverse" Lombard speech mask condition.

Keywords: *face masks, Lombard clear speech, adverse conditions, short-term memory, speech acoustics*

1. INTRODUCTION

In everyday situations, interpersonal communication often takes place in adverse acoustic conditions caused by extrinsic factors that are not directly controllable by interlocutors. This is the case with face masks, which anecdotally have been known to hamper conversational exchange. Indeed, recent studies provide evidence that face masks act as low-pass acoustic filters, dampening frequencies in the range of 2-8 kHz [1, 2]. Furthermore, there is a consensus in literature that face masks increase communication effort between interlocutors, irrespective of noise levels. In particular, face-masked speech results in increased listening effort [1, 3, 4]. In addition, it can cause vocal fatigue for the speaker [5, 6], especially when clear-speech mechanisms are triggered in an attempt to compensate against the adversity of a face mask and ensure message transmittance. Next to increased, immediate processing loads as a result of decreased intelligibility [1, 3, 7, 8], the effects of face-masked speech seem to extend to memory processing as well. In a study analyzing the effects of masks on listeners' cued recall of audio-visually presented sentences, results showed a significant decrease of listeners' recall performance in sentences spoken with a mask [9]. The authors postulated an increase of processing demands, which in turn reduces the resources available for encoding speech in memory. Another study, employing audio-visually presented read material, assessed memory performance in quiet and competing face-masked speech for instructed conversational and clear speaking styles [10]. In contrast to [9], masks did not impair recall for neither native nor non-native speech in quiet. However, in the presence of noise, conversational speech produced with a mask was more difficult to remember than without a mask, especially for non-native speech. Clear speech compensated for the communicative barrier imposed by the mask, improving recall significantly, even in the highest noise condition. Given that both studies presented the test material audio-visually, it is difficult to say whether the drop in performance should be attributed to missing auditory cues, visual cues, or both. In our study we aimed to disentangle this aspect by focusing solely on the auditory channel. In addition, we explored how face-

**Corresponding author*: c.moshona@tu-berlin.de.

masked speech in different, naturally elicited speaking styles affects recall performance and acoustic measures. To this extent, we formulated the following research questions: (1) Can previous findings on the detrimental effects of face-masked speech on recall performance be replicated, if the to be recalled material is presented over the auditory channel only? (2) How, if at all, does speaking style (conversational versus Lombard) affect recall performance? (3) What is the impact of face masks and speaking styles on acoustic measures of speech?

2. METHODS

2.1 Test material

Adopting similar methodology as in [9] for better comparability, the test material consisted of 90 semantically coherent German sentences, modeled after the Oldenburger Satztest (OLSA) [11]. The sentences were conceptualized to contain interchangeable parts of sentence. The last two words were the keywords to be recalled and consisted of a total of four or five syllables to balance out difficulty and prevent word length effects. We opted for two keywords instead of one in order to facilitate mnemonic processing, encouraging strategies such as visualization or association. All sentences had the same syntactical structure and were made up of 5-6 words, beginning with a subject and followed by a verb, a numeral, an adjective and an object. The latter two words were in plural form, e g.*"Klaus zaubert sechs kurze Texte" ("Klaus conjures six short texts")*. Subjects were either common German names or a noun with its respective article. To avoid context-based bonuses in recall performance, the sentences were not highly predictable. Lexical and name frequency was high and each word, except for numerals, only appeared once. To ensure comparability between sentences and recording conditions and enable detailed acoustic analyses, the phonemic distribution of all sentences was controlled.

2.2 Recordings

A native female non-professional speaker of German produced all 90 sentences in conversational and Lombard clear-speaking style, with and without a mask in the following order: Lombard speech mask, conversational speech mask, conversational speech no mask (baseline). The face mask used was an unvalved class 2 filtering face piece (FFP2), type 3M 9320+. This particular model was chosen, because its transfer function had been previously determined and its exact acoustic properties were therefor known [2]. Mono recordings were made in a sound-attenuated booth with a Sennheiser MD421-II cardioid studio microphone, at a sampling rate of $48\,\mathrm{kHz}$ in Audacity. The microphone was positioned at a $15\,\mathrm{cm}$ distance and $45°$ angle from the speaker's mouth. The Lombard-speech condition without a mask was used to adjust the microphone gain level at the beginning of the recording session. The sound pressure level (SPL) of the produced speech was tracked with a NTi Audio XL2 sound level meter, positioned next to the microphone. Communication with the experimentator, who was seated outside the booth took place over headphones. All recording and playback devices were routed via a RME Fireface UCX-II audio interface.

To ensure ecological validity and elicit naturally produced speech, while maintaining a controlled laboratory environment, the sentences were not read, but produced as spontaneously as possible, using a method common in language learning. To this extent, the speaker was cued by seeing the last three words of each sentence in their uninflected form on a screen, underneath a question meant to trigger the full sentence, in this case: *"Was zaubert Klaus?" ("What does Klaus conjure?")*, *"sechs, kurz, Text" ("six, short, text")*. To avoid hesitations during sentence production, the speaker first mentally constructed the sentence and then uttered it out loud. This way we intended to minimize potential recall boosts produced through read speech, which is characterized by reduced speech rate and clearer articulation [12].

To elicit clear-speech adaptations while simulating a noise environment, the speaker heard multitalker babble noise over circumaural, acoustically closed headphones (Beyerdynamic, DT 1770 Pro). This was expected to trigger Lombard-speech by disturbing the speaker's own auditory feedback loop. Lombard speech is not identical to clear speech, but as noted by [13] shares many of its characteristics, such as higher root mean square energy levels, higher mean F_0 and a shift of the spectral centre of gravity to higher frequencies. The multitalker babble consisted of mixed-gender six-talker babble, which was created by superimposing concatenated, read sentences of six individual speakers. Prior to superimposing the speakers, the chains of sentences were trimmed or filled with silence to all have the same length and were then normalized. This was done to minimize the effect of single voices standing out and distracting the speaker during sentence production. The babble noise was looped and mixed into the audio communication channel. Playback level was calibrated at approximately $75\,\mathrm{dB(A)}$ using a

HMS III HEAD acoustics artificial head. This level was deemed optimal to trigger adaptations, while at the same time avoiding leakage during recordings.

2.3 Participants

A total of 33 participants (13 female, 20 male) took part in the listening experiment (mean age = 31.6 years, SD = 8.5 years, range = 23-56 years). Participants were mainly university students and academic staff who were naive in regards to the research topic. All participants were native speakers of German, reported normal hearing and vision and no reading or spelling disability. All participants were provided written information about the study and written consent was obtained from all participants. Compensation was offered in the form of trial participant credit.

2.4 Experimental design and procedure

The experiment was implemented in Matlab (R2023a) in a within-subjects design with speech condition (conversational speech mask, conversational speech no mask, Lombard speech mask) as the independent variable and memory performance, quantified as percentage of correctly recalled keywords, as the dependent variable. The 90 sentences were divided into fifteen blocks of six sentences. This meant that each of the three speech conditions included 30 sentences, divided into 6×5 blocks. Speech condition was blocked and blocks alternated so that the same speech condition did not appear twice in a row. The order of the blocks was counterbalanced across participants and the sentence order was randomized for each session. The sentences had a mean duration of approximately $2\,\mathrm{s}$ and were presented with an inter-sentence-interval of $2500\,\mathrm{ms}$.

The experiment took place in a sound-attenuated booth, under controlled laboratory conditions and ran on a laptop, equipped with an external keyboard and a mouse. Participants were told that they would be listening to stimuli which had been partly produced in adverse conditions, but were not aware that this included face-masked or Lombard speech. Stimuli were presented over the same headphones used for recording at a fixed playback level of $68-73\,\mathrm{dB(A)}$ for the conversational speech no mask and respectively a Δ_L of +7,5$\mathrm{dB(A)}$ for the Lombard speech mask and a Δ_L of $-2\,\mathrm{dB(A)}$ for the conversational speech mask condition. These values were slightly lower than the actual sound pressure levels determined with the XL2 meter, but were chosen to maintain a comfortable playback level throughout the experiment. Participants were instructed by a single experimenter to listen carefully to the six sentences presented in each block and memorize the last two words. A self-paced cued-recall task followed each block. For this task, the first three words of the sentence, up to the numeral, were presented on the screen (e. g. *"Klaus zaubert sechs ..."*) with the last two final words left blank. Participants were asked to fill out these two words by typing them on their keyword. All sentence beginnings of a block were available at once in the order they were presented, and participants were allowed to choose the order in which they typed their responses. This resulted in a total of 180 keywords for each participant to recall, two per sentence and thus twelve per block. This number was chosen to account for the short-term memory's limited storage capacity of six to seven items, while allowing a wide enough range that would prevent ceiling or floor effects. Following the recall task, participants were asked to elaborate on what strategies they used to memorize the words from a set of multiple choice options.

2.5 Acoustic measures

Acoustic measures were specifically selected to estimate speech intensity loss or gain, spectral attenuation caused by the low-pass characteristics of face masks, changes in spectral distribution, vocal load and speech intelligibility. In addition, established psychoacoustic measures were calculated. To estimate overall speech intensity, the energy-equivalent continuous sound level (L_eq) was determined in ArtemiS SUITE, using A-weighting. Psychoacoustic measures included loudness (N_5) as per DIN 45631/A1, sharpness (S) as per DIN 45692, fluctuation strength (F) and specific tonality (tuHMS) according to the ECMA-74(17th) standard. Using the Relative Approach Method (RAM), changes in the time and frequency patterns were quantified by extrapolating the signal history [14]. In addition, four spectral moments weighted by the power spectrum (center of gravity, standard deviation, skewness and kurtosis), as well as measures of timing and fundamental frequency (F_0, F_range) within a speaker-optimized, restricted range of 120-500 Hz, using autocorrelation and a Gaussian window, were automatically calculated with the software Praat [15]. Finally, the long term average spectrum (LTAS), using a bandwidth of $100\,\mathrm{Hz}$ was calculated to determine the mean spectral level in low (0-1 kHz), mid (1-3 kHz) and high (1-8 kHz) frequency ranges, as well as the low-high spectral energy ratio (LH) between the same bands. The low-high energy ratio provides information on spectral slope/spectral tilt.

An increase of the LH-ratio (in absolute values), resulting from less energy in the 1-8 kHz frequency range, implies steeper slopes, which in turn are an indicator of lower intelligibility [16]. An increase of mean energy in the (1-3 kHz) frequency band is associated with greater intelligibility, but also increased vocal load [17].

2.6 Analysis

Each keyword to be recalled was automatically scored by comparing the typed in string to the target keywords stored in the experiment code in Matlab. Correctly memorized keywords received a score of 1, while false or missing keywords received a score of 0. String matching was optimized in R, using the 'stringDist' function of the 'MKmisc' package [18], which computes distance values between strings. This allowed for identification of common typographical errors such as deletions, insertions or other mismatches, which were then manually reviewed and corrected. A mixed model binary logistic regression was calculated in Jamovi, using the module 'GAMLj' [19] and a logit link function for the outcome. Recall performance was entered as the dependent, binary variable, speech condition as a fixed effect and serial sentence position within a block (1-6), as well as word position within a sentence (ultimate, penultimate) as covariates. Up to third degree polynomials of sentence position were included to test for effects of serial position. The model included a random intercept for items as well as a random intercept for participants and a random slope for the participant-specific recall curve within a block.

3. RESULTS

3.1 Recall performance

The model explained $R^2_{\text{cond.}} = 46.3\,\%$ ($R^2_{\text{marg.}} = 29.4\,\%$) of the variance in the data structure. The Omnibus test confirmed that each modeled parameter contributed significantly to R^2, see Table 1. All random model components were successfully tested for significance using likelihood ratio tests, ($p < 0.05$). The observed average recall performance results per condition are shown in Figure 1. The odds of correctly recalling a keyword decreased in the Lombard speech mask condition ($exp_{\text{B}} = 0.770, p = 0.002$) and in the conversational speech mask condition ($exp_{\text{B}} = 0.894, p = 0.141$) as compared to conversational speech no mask baseline, though these effects were rather small and only significant in the first case, see Table 2. Recall accuracy was lower in the Lombard speech mask condition, as compared to the conversational speech mask condition, though the odds were not significantly different. The estimated marginal means (EMM) of the speech conditions are shown in Table 3.

Word position and sentence position notably affected recall accuracy. Figure 2 shows a serial position effect for recall performance for all conditions with the first sentence being memorized better (primacy effect) than the middle sentences and the last sentences having the highest recall accuracy (recency effect). Performance variance was greater for sentences 1-3, with performance converging to the same level for sentences 4-6. The figure also shows that the ultimate word was easier to memorize than the penultimate word, across all sentence positions and conditions. The EMM of serial position were 0.318 (SE = 0.030) at the mean and increased to 0.370 (SE = 0.033) at one standard deviation (SD) below the mean and 0.731 (SE = 0.035) at one SD above the mean respectively. Thus, the probability of a correct response tended to increase as serial position decreased or increased. The EMM of word position were 0.318 (SE = 0.030) at the mean, decreased to 0.245 (SE = 0.028) at one SD below

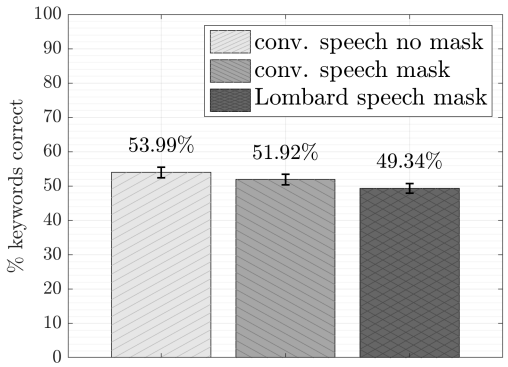

Figure 1. Average recall performance per condition. The vertical bars represent standard mean errors.

Table 1. Fixed Effect Omnibus tests

	χ^2	df	p
condition	11.773	2.000	0.003
sentence position	21.094	1.000	0.001
word position	43.044	1.000	0.001
sentence position^2	494.127	1.000	0.001
sentence position^3	25.647	1.000	0.001

Table 2. Post-hoc comparisons - condition

conditions	exp(B)	SE	z	p_{holm}
L_m-$conv_m$	0.861	0.066	-1.954	0.101
L_m-$conv_{nm}$	0.770	0.059	-3.421	0.002
$conv_m$-$conv_{nm}$	0.894	0.068	-1.473	0.141

L_m = Lombard speech mask, $conv_m$ = conv. speech mask, $conv_{nm}$ = conv. speech no mask.

Table 3. Estimated marginal means

				95% CI	
condition	**Prob.**	**SE**	**df**	**Low.**	**Up.**
L_m	0.289	0.030	Inf	0.234	0.351
$conv_m$	0.321	0.032	Inf	0.262	0.386
$conv_{nm}$	0.345	0.033	Inf	0.284	0.413

Note: Estimated means while keeping constant other effects in the model to the mean.

the mean and increased to 0.401 (SE $= 0.036$) at one SD above the mean. Therefor, the probability of a correct response tended to increase for ultimate words. During the course of the experiment participants displayed a learning effect with recall performance increasing as a function of time (not depicted).

3.2 Mnemonic strategies and error assessment

All participants except one person reported to have used at least one mnemonic strategy to facilitate keyword recall. Out of the multiple choices listed, visualizations (n=21) and associations (n=15) were used most frequently, followed by loud rehearsal (n=12), story creation (n=8) and alphabetical strategies involving memorizing specific letters (n=8). Two participants reported loci-like techniques, using spatial association/information to memorize keywords. In addition, participants reported to have prioritized which keywords they chose to focus on or type in first, concentrating on the last two sentences, the first and last sentence or the last four sentences. One participant reported to have ignored the beginning of the sentences. Participants whose recall performance was above average almost always used visualizations. Omissions were by far the most frequent reason to categorize a response as incorrect, followed by misplacements (recalling keywords correctly, but attributing them to the wrong sentence), semantic similarity (e. g. "*Schweine*" / "*pigs*" for "*Ferkel*" / "*piglets*") and association errors (e. g. "*saftige Erdbeeren*" / "*juicy strawberries*" for "*kalte Erdbeeren*" / "*cold strawberries*"). In very few cases phonetic errors were noted (e. g. "Gräser" / "grasses" for "Gläser" / "glasses').

3.3 Acoustic measures

Table 4 summarizes the acoustic measures for all speech conditions. Compared to the baseline condition without a mask and as a result of the mask's low-pass filtering effects, the conversational speech mask condition was characterized by both lower sound levels ($-2.2\,\mathrm{dB}$) and lower perceived loudness ($-7.95\,\mathrm{sone}$), as well as lower sharpness ($-0.14\,\mathrm{acum}$). The mean levels for all analyzed frequency bands were lower in this condition. This held particularly true for the 1-8 kHz band with a drop of $-5.11\,\mathrm{dB}$. Consequently, the low-high energy ratio was higher, which implies lower intelligibility. The mask's attenuating effects in the higher frequency regions are also reflected in the spectral moments with the center of gravity being lower in the conversational speech mask condition as compared to the baseline ($-370.87\,\mathrm{Hz}$), the dispersion of spectral energy covering a more narrow frequency range, the distribution being more peaked and more positively skewed. The RAM revealed less variance in the spectral and temporal patterns ($-8.13\,\mathrm{cPa}$). Duration was slightly higher, but fundamental frequency and tonality measures showed no notable differences. The speech style change from conversational to Lombard in the mask condition accounted for diametrically opposite results. Compared to the baseline, the Lombard speech mask condition was characterized by an increase of sound level ($+7.67\,\mathrm{dB}$), loudness ($+33.11\,\mathrm{sone}$), tonality ($+0.48\,\mathrm{tuHMS}$), fluctuation strength ($+0.05\,\mathrm{vacil}$), mean levels in all analyzed bands, center of gravity ($+83.33\,\mathrm{Hz}$) and spectro-temporal variance ($+9.13\,\mathrm{cPa}$). However, the dispersion of spectral energy around the mean was lower ($-226.24\,\mathrm{Hz}$) than in the baseline. The increased mean spectral levels are indicative of higher intelligibility, but also imply an increase of vocal load, especially due to the boost of $+8.75\,\mathrm{dB}$ at the 1-3 kHz band. Compared to the baseline, the Lombard speech mask condition also exhibited higher duration and an increase of fundamental frequency by $+24.98\,\mathrm{Hz}$. There were no notable differences regarding sharpness. These differences between conditions are visualized in Figure 3, which also confirms attenuation peaks at 3 kHz and 6 kHz, as shown in [2].

4. DISCUSSION

In our study, we critically examined the effects of face masks and speech style on recall performance and acoustic measures by simulating conversational and Lombard

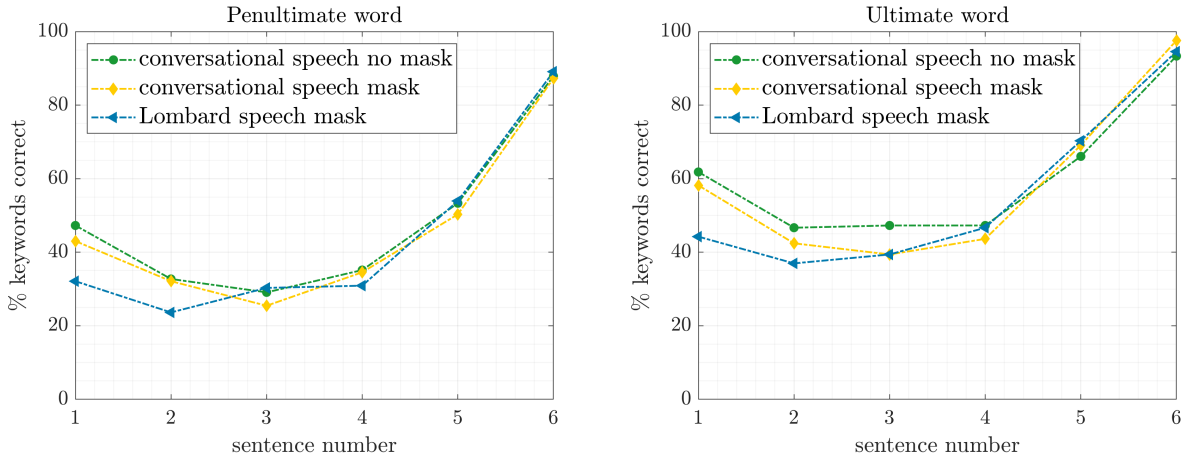

Figure 2. Proportion of correctly recalled keywords as a function of serial position within a block (1-6) and speech condition. Left: penultimate word, right: ultimate word.

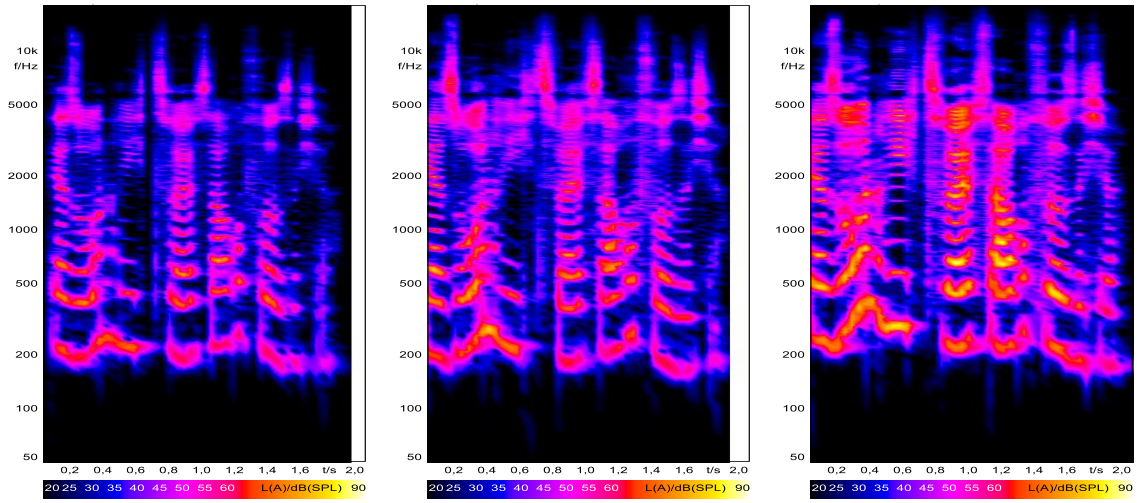

Figure 3. FFT spectra vs. time of the sentence "*Der Sohn riecht zwölf saure Pfirsiche*" in comparison: conversational speech mask (left), conversational speech no mask (middle) and Lombard speech mask (right).

speech that would naturally occur in an adverse, noisy environment. Our results are in line with previous findings by [9, 10], replicating the trend that face masks tend to reduce recall performance, even in quiet conditions. Recall accuracy was at a similar level as the results reported in [9], but lower than in [10], which could be owed to the different task used in the latter study. However, in our case the effect of the mask alone was not significant. This could be attributed to the fact that we only used auditory, instead of audio-visual presentation of spoken material, which implies that visual cues benefit memory processing more than auditory ones. Another explanation could lie in the fact both the speaker's produced speech intensity and the playback level were overall quite high in all conditions, that the speaker had high articulatory precision and that we used uninstructed spontaneous speech, which may

Table 4. Means and standard deviations of acoustic measures per condition

		$\text{conv}_{\text{nomask}}$		$\text{conv}_{\text{mask}}$		$\text{Lombard}_{\text{mask}}$	
Acoustic measure	**Unit**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Time	ms	1986	202	2035	193	2046	241
L_{eq}	dB(A)	79.15	1.92	76.93	1.91	86.81	1.70
Sharpness (S)	acum	1.55	0.11	1.41	0.09	1.57	0.09
Loudness (N_5)	sone	52.69	7.99	44.74	5.72	85.80	11.62
Specific Tonality	tuHMS	0.70	0.15	0.70	0.15	1.18	0.25
Fluctuation Strength (F)	vacil	0.16	0.04	0.16	0.04	0.22	0.06
F_0	Hz	221.85	14.06	225.24	13.46	246.83	12.88
F_{range}	Hz	171.00	61.21	167.58	55.01	188.51	56.31
$\text{Spectral Level}_{\text{low}}$ (0-1 kHz)	dB	26.44	1.45	25.38	1.67	32.18	1.41
$\text{Spectral Level}_{\text{mid}}$ (1-3 kHz)	dB	17.40	2.74	14.50	2.61	26.15	2.34
$\text{Spectral Level}_{\text{high}}$ (1-8 kHz)	dB	12.62	2.04	7.51	1.86	17.94	1.65
LH_{ratio}	dB	-13.82	2.08	-17.87	1.99	-14.25	1.77
Center of Gravity	Hz	1128.35	286.42	757.48	124.60	1211.68	202.63
$\text{Center of Gravity}_{\text{SD}}$	Hz	1512.20	341.78	974.70	183.66	1285.96	169.92
Skewness	–	3.04	0.82	4.45	1.00	2.63	0.91
Kurtosis	–	11.88	6.90	28.73	13.54	10.72	8.41
Relative Approach Method	cPa	61.47	5.46	53.34	4.87	70.59	5.62

have overshadowed potential effects. Interestingly, in our study, the modulatory effect of speech style was greater than that of the mask and had a negative impact. Lombard speech through a mask significantly reduced recall performance by almost 5 % on average throughout the experiment and by 8-18 % for the first two sentences of each block compared to the baseline. The results of the acoustic analyses support this finding, given that in the Lombard speech mask condition sound level, psychoacoustic loudness, mean spectral levels, sharpness, fluctuation strength, fundamental frequency, tonality and temporal-structural variance increased markedly and thereby compensated largely against the filtering effects of the mask. Though the increase of spectral energy especially in the higher frequency bands is associated with greater intelligibility, Lombard speech is also produced with higher vocal load, impacting phonation due to the increase of laryngeal tension and resulting in a more "stressed" voice quality. In unison, these acoustic parameters make a signal more noticeable and potentially also more disruptive, given that sharpness and loudness strongly correlate with annoyance [20]. In conclusion, these voice adaptations might have diverted attention from the cued-recall task, leaving less resources available for memory encoding. As further analyses showed, participants spent the least amount of time reflecting on solutions in the Lombard speech condition, which underpins the annoyance hypothesis. These findings differ from the clear speech benefits reported in [10]. Our study also showed evident serial position effects with the recency effect being larger than the primacy effect, resulting in a hook-shaped recall curve. Mnemonics and in particular visualizations and associations seemed to facilitate recall.

5. CONCLUSION

Answering the research questions formulated, our study (1) replicated detrimental effects of face masks on recall during auditory presentation, (2) showed significant, negative effects of Lombard speech for memory processing and (3) demonstrated an impact of face masks and speech style on acoustic measures. However, further validation of these findings is needed. In future experiments it would be interesting to analyze the effect of speaker individuality by including several speakers with different traits, quantifying the benefit of visual versus auditory cues in a mixed design utilizing both auditory-only and audiovisual presentation, comparing different types of clear speech and varying the background conditions for listeners to include noise. A more detailed, phonetic analysis of speech may also prove useful in assessing the effects of face masks and speaking styles.

6. ACKNOWLEDGMENTS

The authors extend their gratitude to Steffen Lepa for statistical guidance, Frederic Rudawski for technical support and the participants of the experiment for their time.

7. REFERENCES

[1] P. Bottalico, S. Murgia, G. E. Puglisi, A. Astolfi, and K. I. Kirk, "Effect of masks on speech intelligibility in auralized classrooms," *The Journal of the Acoustical Society of America*, vol. 148, no. 5, pp. 2878–2884, 2020.

[2] C. Moshona, J. Hofmann, A. Fiebig, and E. Sarradj, "Bestimmung des Übertragungsverlustes von Atemschutzmasken mittels eines 3D-Kopfmodells unter Berücksichtigung des Ansatzrohres," in *Fortschritte der Akustik - DAGA 2023, 49. Jahrestagung für Akustik*, Deutsche Gesellschaft für Akustik e.V. (DEGA), 2023.

[3] V. A. Brown, K. J. V. Engen., and J. E. Peelle, "Face mask type affects audiovisual speech intelligibility and subjective listening effort in young and older adults," *Cognitive Research: Principles and Implications*, vol. 6, no. 1, p. 49, 2021.

[4] E. Giovanelli, C. Valzolgher, E. Gessa, M. Todeschini, and F. Pavani, "Unmasking the Difficulty of Listening to Talkers With Masks: lessons from the COVID-19 pandemic," *i-Perception*, vol. 12, no. 2, p. 204166952199839, 2021.

[5] V. S. McKenna, T. H. Patel, C. L. Kendall, R. J. Howell, and R. L. Gustin, "Voice Acoustics and Vocal Effort in Mask-Wearing Healthcare Professionals: A Comparison Pre- and Post-Workday," *Journal of Voice*, vol. 0, no. 0, 2021.

[6] V. V. Ribeiro, A. P. Dassie-Leite, E. C. Pereira, A. D. N. Santos, P. Martins, and R. de Alencar Irineu, "Effect of Wearing a Face Mask on Vocal Self-Perception during a Pandemic," *Journal of Voice*, 2020.

[7] M. Randazzo, L. L. Koenig, and R. Priefer, "The effect of face masks on the intelligibility of unpredictable sentences," in *Proc. of Meetings on Acoustics*, vol. 42, p. 032001, 2020.

[8] J. C. Toscano and C. M. Toscano, "Effects of face masks on speech recognition in multi-talker babble noise," *PLOS ONE*, vol. 16, no. 2, p. e0246842, 2021.

[9] T. L. Truong, S. D. Beck, and A. Weber, "The impact of face masks on the recall of spoken sentences," *The Journal of the Acoustical Society of America*, vol. 149, no. 1, pp. 142–144, 2021.

[10] R. Smiljanic, S. Keerstock, K. Meemann, and S. M. Ransom, "Face masks and speaking style affect audio-visual word recognition and memory of native and non-native speech," *The Journal of the Acoustical Society of America*, vol. 149, no. 6, pp. 4013–4023, 2021.

[11] O. Satztest, "Handbuch und hintergrundwissen (manual and background knowledge," 2000.

[12] M. Nakamura, K. Iwano, and S. Furui, "Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance," *Computer Speech and Language*, vol. 22, no. 2, pp. 171–184, 2008.

[13] V. Hazan and R. Baker, "Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. 2139–2152, 2011.

[14] R. Sottek and K. Genuit, "Models of signal processing in human hearing," *AEU - International Journal of Electronics and Communications*, vol. 59, no. 3, pp. 157–165, 2005.

[15] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2023. Version 6.3.09.

[16] Y. Lu and M. Cooke, "The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication*, vol. 51, no. 12, pp. 1253–1262, 2009.

[17] V. Hazan and D. Markham, "Acoustic-phonetic correlates of talker intelligibility for adults and children," *The Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3108–3118, 2004.

[18] M. Kohl, *MKmisc: Miscellaneous functions from M. Kohl*, 2022. R package version 1.9.

[19] M. Gallucci, "GAMLj: General analyses for linear models [jamovi module]," 2019. Retrieved from https://gamlj.github.io/.

[20] H. Fastl, "Psycho-acoustics and sound quality," in *Communication Acoustics* (J. Blauert, ed.), pp. 139–162, Springer Berlin Heidelberg, 2005.