Are Acoustic Markers of Voice and Speech Signals Affected by Nose-and-Mouth-Covering Respiratory Protective Masks?

[*,†,‡,§,‖] **Youri Maryn, PhD,** [*,¶] **Floris L. Wuyts, and** [*] **Andrzej Zarowski, MD,** *¶*Wilrijk,* †*Ghent,* ‡*Gent,* §*, and* ‖*Lokeren, Belgium*

Summary: Background. Worldwide use of nose-and-mouth-covering respiratory protective mask (RPM) has become ubiquitous during COVID19 pandemic. Consequences of wearing RPMs, especially regarding perception and production of spoken communication, are gradually emerging. The present study explored how three prevalent RPMs affect various speech and voice sound properties.

Methods. Pre-recorded sustained [a] vowels and read sentences from 47 subjects were played by a speech production model (‘Voice Emitted by Spare Parts’, or ‘VESPA’) in four conditions: without RPM (C1), with disposable surgical mask (C2), with FFP2 mask (C3), and with transparent plastic mask (C4). Differences between C1 and masked conditions were assessed with Dunnett’s t test in 26 speech sound properties related to voice production (fundamental frequency, sound intensity level), voice quality (jitter percent, shimmer percent, harmonics-to-noise ratio, smoothed cepstral peak prominence, Acoustic Voice Quality Index), articulation and resonance (first and second formant frequencies, first and second formant bandwidths, spectral center of gravity, spectral standard deviation, spectral skewness, spectral kurtosis, spectral slope, and spectral energy in ten 1-kHz bands from 0 to 10 kHz).

Results. C2, C3, and C4 significantly affected 10, 15, and 19 of the acoustic speech markers, respectively. Furthermore, absolute differences between unmasked and masked conditions were largest for C4 and smallest for C2.

Conclusions. All RPMs influenced more or less speech sound properties. However, this influence was least for surgical RPMs and most for plastic RPMs. Surgical RPMs are therefore preferred when spoken communication is priority next to respiratory protection.

Key Words: Respiratory protection masks−Speech−Voice−Acoustics.

INTRODUCTION

COVID pandemic forced us all across the planet to take sanitary and personal protection measures in an attempt to control coronavirus-related disease and mortality. Among measures like hand hygiene and social distancing, nose-and-mouth-covering respiratory protective mask (RPM) may prevent airborne transmission of infectious disease through respiratory droplets produced when infected persons cough, sneeze, talk, shout, or sing. Wearing such protective devices, however, may come with various physiological and psychological burdens,[1] including difficulties in spoken communication[2]

Only few studies investigated effects of RPMs on spoken communication. Regarding speech perception, it appears that mouth-and-nose-covering RPMs have only little influence on how speakers are perceived in terms of speech intelligibility, especially in relatively quiet environments.[3-7] But what happens with speech sound properties when the speaker is wearing a RPM? In other words: are acoustic voice and/or speech measures influenced by RPM, and if yes, to what degree? These questions are hypothesized to be relevant to people across society (eg, teachers) and especially in health care (eg, health care providers having to explain, speech-disordered patients undergoing acoustic voice and speech assessment, hearing-impaired patients undergoing hearing aid, or cochlear implant optimization, etc.). Maryn[8] recorded four subjects twice with and without disposable surgical mask and found that various acoustic voice markers (ie, sound intensity level, fundamental frequency, jitter local, shimmer local dB, smoothed cepstral peak prominence and Acoustic Voice Quality Index) did not uniformly differed between these two conditions. Corey et al[9] also investigated the spectral effects of various masks on speech signals. They found, in general, that most masks had limited influence below 1 kHz but attenuated higher frequencies (especially above 4 kHz) by differing quantities. This was least for surgical masks.

To our knowledge, however, how strong findings regarding RPMs also pertain to other acoustic properties that are relevant to and commonly determined in clinical speech and voice assessment −i.c., fundamental frequency, perturbation, harmonics-to-noise ratio, smoothed cepstral peak prominence, Acoustic Voice Quality Index, formant properties, spectral moments− has not been investigated before. To fill this hiatus and to isolate influences of RPMs without having to take head and articulatory movements nor friction-related noise into account, a sound-producing head and torso imitation was self-built for the present study to compare a set of clinically relevant acoustic voice and

---

Accepted for publication January 13, 2021.

From the *Department of Otorhinolaryngology & Head and Neck Surgery, European Institute for ORL-HNS, Wilrijk, Belgium; †Department of Rehabilitation Sciences, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium; ‡Faculty of Education, Health & Social Work, University College Ghent, Gent, Belgium; §Faculty of Psychology and Educational Sciences, School of Logopedics, Belgium; ‖Phonanium, Lokeren, Belgium; and the ¶Lab for Equilibrium Investigations and Aerospace (LEIA), University of Antwerp, Wilrijk, Belgium.

Address correspondence and reprint request to: Youri Maryn, GZA Sint-Augustinus, European Institute for ORL-HNS, Oosterveldlaan 24, 2610 Wilrijk, Belgium. E-mail: youri.maryn@gza.be

Journal of Voice, Vol. ■■, No. ■■, pp. ■■−■■
0892-1997

https://doi.org/10.1016/j.jvoice.2021.01.013

speech measures with and without commonly used mouth-and-nose-covering RPMs. Because only spare parts were used for this model, it was called the 'Voice-Emitted-by-Spare-PArts' or 'VESPA'. Based on Corey et al,[9] the hypothesis in this VESPA study was that acoustic voice and speech markers are influenced by mouth-and-nose-covering RPMs.

METHODS

Initial voice recordings

The same voice samples from 50 Flemish Dutch-speaking subjects as in the study of Maryn et al[10] were employed in this investigation. Their voices were recorded at the beginning of the standard voice assessment as part of routine clinical practice. Primary laryngological diagnoses included in the sample, using an Olympus ENF-V flexible transnasal chip-on-tip laryngostroboscope (Olympus Corporation, Tokyo, Japan), were the following: 12 normal vocal folds, 13 vocal fold nodules, nine unilateral vocal fold paralysis, five post-head and neck cancer treatment, four muscle tension dysphonia, three laryngitis, two polypoid mucosa, one presbylarynx, and one leukoplakia. This group consisted of 29 women and 21 men, with ages ranging from 10 to 77 years (mean = 44.9 years, standard deviation = 19.2 years). Because of (quasi)aphonic sustained vowels, however, the recording of three subjects were rejected from post-hoc analysis. The final sample was considered to be adequately representative of a voice clinic population, reflecting different ages, genders, different types and degrees of voice quality, and voice-related disability.

At the start of a standard voice assessment, every participant was asked to sustain vowel [a] for at least 5 seconds and to read a phonetically balanced text at comfortable pitch and loudness. Both voice samples were recorded in an anechoic audiometric booth using an AKG C420 head-mounted condenser microphone and digitized at a sampling rate of 44.1 kHz and 16 bits of resolution using the Computerized Speech Lab (CSL model 4500; KayPENTAX Corp., Lincoln Park, New Jersey). All samples were saved as WAV files. The vowel samples used in this study were trimmed to include only the middle 3 seconds. The continuous speech (ie, read text) samples were formatted to include only the first two sentences. Further editing of these original voice recordings consisted of the following two steps.

1. Per subject, extracted vowel and speech segments were chained in the following order using Praat: pause of 1 second, two sentences, pause of 2 seconds, 3-second sustained [a], and pause of 1 second. Except for the continuous speech segments, all these parts had constant duration.
2. All these 50 chained sound signals were concatenated to a single long sound file to enable a single audio presentation of all 50 concatenations after one another. To indicate boundaries between two chained sound files for later segmentation of the long sound files, however, short acoustic markers were interjected. Every in-between acoustic token to demarcate the margins between two concatenated sound files consisted of two sinusoidal cycles of 0.001 seconds and between -1 and $+1$ Pascal.

The resulting long sound signal is the same as in Maryn et al[11] This is demonstrated in Figure 1.

VESPA sound recording setup

With no commercial torso and head with mouth simulator available, spare parts were assembled as following to construct a convenient physical model, as demonstrated in Figure 2 (top). This approximate head model could then serve as a stand for the application of RPM's and consequently to test differences in airborne speech signals with and without RPM's.

1. Body: hollow female fashion mannequin doll with stylistic head, torso and arms from coated/polished fiberglass.
2. Mouth: round hole with 3.5 cm diameter.
3. Auricles: two bolts of 2.5 cm sticking out at 3.5 cm from each other in the coronal plane (representing superior and inferior auricle attachment, around which the straps of the RPM run, and upon which the microphone headset was mounted).
4. Sound source: loudspeaker with 6.5 cm diameter affixed immediately behind the mouth opening and connected via minijack to the computer.

Although it is not assumed that this model equals the anatomical/physiological properties of an average human body, its setup was considered appropriate for standardized presentation and evaluation of speech sound signals that is consistent across recording conditions.

Mouth-and-nose-covering RPMs

In this study, the influence of the following three mouth-and-nose-covering RPMs on acoustic voice and speech markers was assessed. These RPMs were chosen because they are commonly used, also by patients undergoing clinical voice and speech assessment.

1. Disposable medical/surgical mask with nonwoven three layers (SunginCare, Hangzhou Sunten Textile Co., Zhejiang, China).
2. Sanbang 9051A FFP2 respirator (Foshan Nanhai Weijian Sanbang Protective Products Technology Co., Foshan, China).
3. Translucent mask with a transparent plastic window located at the mouth and knit in a cloth frame. This kind of mask was produced by a knitting team of GZA Sint-Augustinus personnel (established at the beginning of COVID19 lockdown when RPM supply was low) for healthcare professionals to facilitate oral

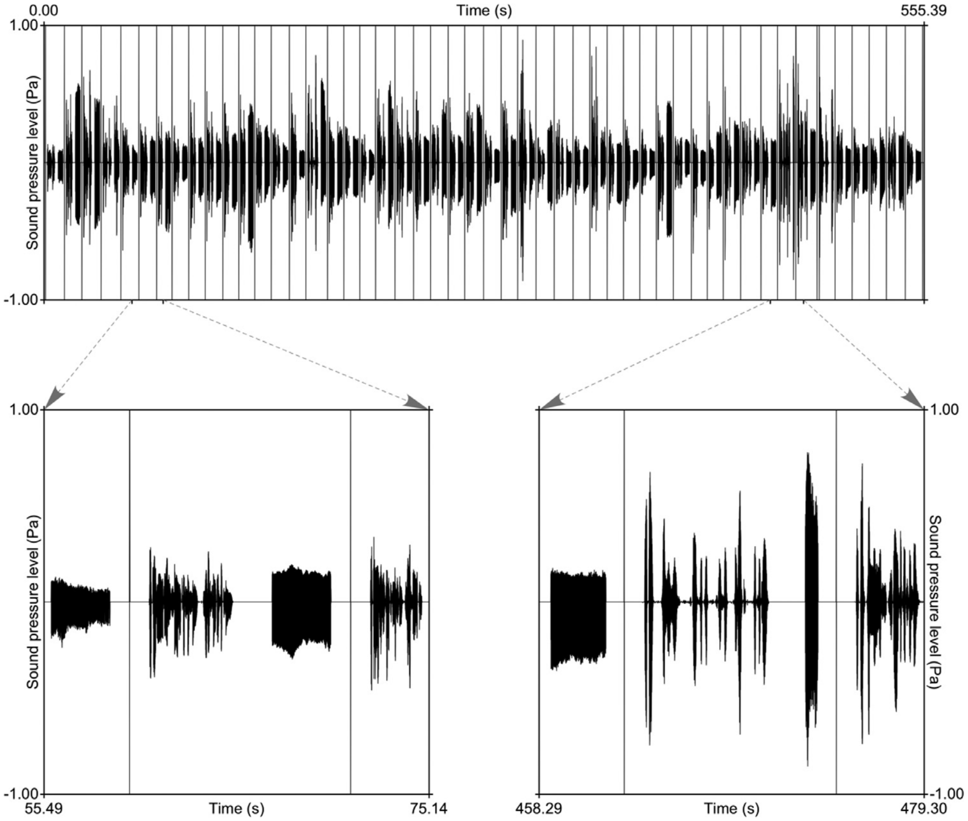

FIGURE 1. Illustration of the chaining of the original voice recordings in this study. Top oscillogram: sequence of extracted sound signal segments (3-second sustained [a] vowel, and 2 sentences of read text) of the fifty subjects to one long sound chain of 555.39 seconds. Bottom two oscillograms: sixth (left) and forty-third (right) concatenated sound files (pause, text segment, pause, vowel segment, and pause) with their boundaries designated by an imprinted acoustic mark.

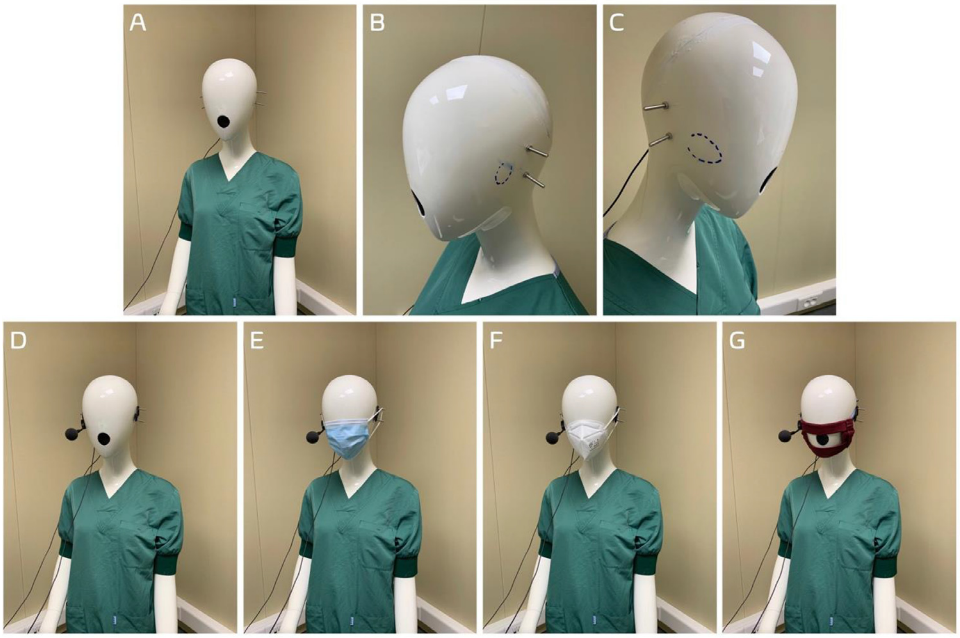

FIGURE 2. Photographs of the VESPA model, as it was situated in the sound treated room. Top (A, B, C): VESPA model without microphone or RPM. To ensure consistent microphone placement relative to VESPA's sound source, blue dashed elliptic markings where applied to indicate the left (B) and right (C) spots where the microphone's behind-the-neck headband should make contact with the model's head. Bottom (D, E, F, G): VESPA with only microphone as control condition (D) and with microphone plus surgical mask (E), FFP2 mask (F) or transparent mask (G).

communication and lipreading, especially when talking to patients with hearing difficulties.

Sound recording system

All experimental sound signals were recorded in an anechoic room with an ambient noise level of 13.5 $\mathrm{dB_{LAeq}}$, as measured with a CR:162B integrating averaging class II sound level meter (Cirrus Research plc, Hunmanby, North Yorkshire, United Kingdom).

The complete sound chain with all the original voice samples, pauses, and boundary markers (the upper oscillogram of 555.39 s in Figure 1) was radiated by VESPA's built-in loudspeaker (see Figure 2A). This airborne signal was recorded with the following equipment: AKG C544L head-mounted condenser microphone (AKG Acoustics, Vienna, Austria) positioned at 45° azimuth and 8 cm from VESPA's mouth, MPA V L mini-male-to-standard-female XLR connection with phantom power adapter (AKG Acoustics), RME Babyface Pro audio interface (RME, Haimhausen, Germany), and MacBook Air with OS High Sierra 10.13.6 (Apple Inc., Cupertino, California). Conform Švec & Granqvist[12] and Maryn & Zarowski,[13] this equipment was considered suitable for clinical speech assessment.

Experimental sound samples

The complete sound chain was played and captured by the sound recording system four times by VESPA's loudspeaker: without RPM (the control condition or C1), with surgical mask (C2), with FFP2 mask (C3), and with transparent mask (C4). This is illustrated in respectively Figures 2D, 2E, 2F, and 2G.

Praat software for MacIntosh version 6.0.31 (Institute of Phonetic Sciences, University of Amsterdam, The Netherlands) was applied for all acoustic signal editing, segment extraction and analysis in this VESPA study.

Acoustic markers

After extraction, there were 376 (two speech tasks × 47 subjects × four recording conditions) voice and/or speech samples available for determining the following 26 acoustic variables with Praat[a]:

- median fundamental frequency (f_O);
- median sound intensity level (IL);
- voice quality-related measures: jitter local (JL), shimmer local dB (SL), harmonics-to-noise ratio (HNR), smoothed cepstral peak prominence (CPPS), and Acoustic Voice Quality Index (AVQI);
- formant-related measures: first formant (F_1), bandwidth of first formant (B_{F1}), second formant (F_2), and bandwidth of second formant (B_{F2}),
- spectral moments 1 (SM1, or center of gravity), 2 (SM2, or standard deviation), 3 (SM3, or skewness), and 4 (SM4, or kurtosis);
- spectral slope between 0 Hz and 10,000 Hz (SS);
- mean energy in ten 1-kHz frequency bands between 0 and 10 kHz (FB1-FB10).

The AVQI was determined on concatenations of voiced segments of continuous speech with sustained [a], according to Maryn and Weenink.[14] The other markers were determined on the sustained [a] extracts.

Statistical analyses

All statistical analyses were completed using SPSS version 26.0 (SPSS Inc., Chicago, Illinois). Two-way ANOVA on four related samples was used to compare the 26 acoustic markers across the four recording conditions. Post-hoc Dunnett's tests were applied for this many-to-one comparison in which three pairs of related samples were juxtaposed: C1-C2, C1-C3, and C1-C4. Post-hoc results were considered statistically significant at $P \leq 0.05$. These methods were administered to answer the question: are the 26 acoustic metrics significantly different when wearing a RPM?

RESULTS

Descriptive statistics of the acoustic markers and their differences between C1 and C2/C3/C4 are provided in Table 1 and illustrated by the multiple line graphs in Figures 3, 4, and 5. Significance of these differences on the basis of Dunnett's t tests are provided in Table 2.

No mask (C1) versus surgical mask (C2). Sixteen of the 26 (ie, 61.5 %) acoustic markers were unaffected by C2: f_O, IL, JL, SL, HNR, F_1, B_{F1}, F_2, B_{F2}, SM1, SM2, SS, FB1, FB2, FB8, and FB9. For the remaining markers (CPPS, AVQI, SM3, FB3, FB4, FB5, FB6, FB7, and FB10), however, C2 differed significantly from C1. On average, CPPS decreased with 0.57 dB (with ΔCPPS between 0.08 dB and 1.32 dB) and AVQI increased with 0.25 (with ΔAVQI between -0.85 and 0.65). Spectral skewness increased with 2.9 (with ΔSM3 between -15.7 and 0.8), spectral kurtosis increased with 213 (with ΔSM4 between -1659 and 18), energy in spectral kHz bands 3, 4 and 5 decreased with 2.6 dB/Hz (with ΔFB3 between 2.1 dB/Hz and 3.3 dB/Hz), 2.2 dB/Hz (with ΔFB4 between 1.7 dB/Hz and 3.7 dB/Hz) and 1.2 dB/Hz (with ΔFB5 between -1.3 dB/Hz and 4.0 dB/Hz), respectively. Energy in spectral kHz bands 6, 7, and 10 increased with 3.4 dB/Hz (with ΔFB6 between -12.8 dB/Hz and 3.4 dB/Hz), 6.2 dB/Hz (with ΔFB7 between -13.9 dB/Hz and 1.9 dB/Hz) and 3.1 dB/Hz (with ΔFB10 between -9.5 dB/Hz and 5.4 dB/Hz), respectively. The differences in all the measures between C1 and C2 are illustrated in the left line graphs of Figures 3, 4, and 5.

No mask (C1) versus FFP2 mask (C3). Eleven of the 26 (ie, 42.3 %) acoustic measures were not significantly influenced by C3: f_O, JL, SL, HNR, F_1, B_{F1}, B_{F2}, SM1, SM2, SS, and FB9. In the other 15 markers, however, C1 data

---

[a] Symbolic notation style of frequencies and formants are in accordance with the consensus report of Titze et al[30] and Titze[31] across this manuscript.

TABLE 1.
Mean (ie, M), Standard Deviation (ie, SD), Minimum (ie, Min) and Maximum (ie, Max) of the 26 Acoustic Measures per Recording Condition (C1: No Mask; C2: Surgical Mask; C3: FFP2 Mask; C4: Transparent Mask) and per Difference (ie, Δ) Between No-Mask and Mask Conditions

Acoustic marker	C1				C2				C3				C4			
	M	SD	Min	Max	M	SD	Min	Max	M	SD	Min	Max	M	SD	Min	Max
f_O	175	62	75	367	178	61	76	365	176	63	62	363	168	53	69	290
Δf_O	–	–	–	–	-2	16	−106	5	−1	18	−107	61	7	42	−107	197
IL	58,3	4,6	46,5	67,1	57,9	4,6	46,7	66,7	57,0	4,7	45,7	66,0	56,8	5,7	45,6	68,4
ΔIL	–	–	–	–	0,4	0,4	−0,6	1,7	1,3	0,6	0,0	2,5	1,5	3,0	−4,2	9,8
JL	1,12	1,62	0,16	7,34	1,14	1,56	0,17	7,03	1,14	1,60	0,16	6,86	1,14	1,49	0,13	7,11
ΔJL	–	–	–	–	−0,02	0,34	−1,17	1,63	−0,02	0,25	−1,07	0,68	−0,02	0,41	−0,96	1,83
SL	0,59	0,48	0,09	2,06	0,63	0,44	0,10	2,00	0,66	0,45	0,11	1,88	0,79	0,50	0,09	2,03
ΔSL	–	–	–	–	−0,03	0,14	−0,43	0,30	−0,06	0,17	−0,61	0,24	−0,20	0,29	−0,95	0,31
HNR	15,97	6,58	2,90	29,35	15,64	6,41	2,85	28,45	15,85	6,49	2,68	28,35	15,09	6,96	1,11	29,97
ΔHNR	–	–	–	–	0,33	0,71	−0,84	2,12	0,12	1,02	−1,71	2,74	0,88	3,15	−5,20	8,32
CPPS	11,94	3,81	4,31	21,24	11,37	3,70	3,77	19,92	11,18	3,64	3,87	19,61	10,93	3,57	3,84	18,81
ΔCPPS	–	–	–	–	0,57	0,24	0,08	1,32	0,76	0,29	0,00	1,63	1,01	0,71	−0,91	2,42
AVQI	4,01	1,77	0,80	8,02	4,26	1,77	1,28	8,12	4,37	1,74	1,31	8,17	4,42	1,84	1,10	8,87
ΔAVQI	–	–	–	–	−0,25	0,31	−0,85	0,65	−0,36	0,39	−1,66	0,56	−0,41	0,71	−2,40	1,14
F_1	602	158	330	1068	617	163	328	1071	617	164	322	1028	764	177	326	1027
ΔF_1	–	–	–	–	−15	25	−70	43	−15	42	−93	158	−162	204	−550	476
B_{F1}	330	169	52	859	349	181	61	961	318	162	62	842	302	189	48	995
ΔB_{F1}	–	–	–	–	−19	46	−172	98	12	64	−152	227	28	197	−371	544
F_2	1387	220	1000	1868	1377	208	992	1825	1331	190	981	1818	1274	156	932	1576
ΔF_2	–	–	–	–	10	42	−88	152	56	95	−75	463	113	195	−71	736
B_{F2}	310	162	34	741	323	260	28	1570	326	290	20	1659	159	168	9	1173
ΔB_{F2}	–	–	–	–	−13	225	−1290	410	−16	276	−1378	388	151	210	−893	535
SM1	588	193	322	1329	574	184	307	1228	557	160	313	1028	625	207	230	1023
ΔSM1	–	–	–	–	14	28	−36	103	31	54	−33	301	−36	183	−387	569
SM2	406	153	157	821	404	136	170	784	378	120	149	733	384	93	177	605
ΔSM2	–	–	–	–	2	32	−95	78	28	53	−140	159	22	131	−304	336
SM3	4,1	2,8	0,5	13,8	7,1	6,8	0,8	29,4	8,6	8,7	1,2	41,7	6,4	8,7	−0,7	51,1
ΔSM3	–	–	–	–	−2,9	4,2	−15,7	0,8	−4,5	6,1	−27,9	1,2	−2,2	6,8	−41,3	4,2
SM4	83	122	0	622	296	474	4	2281	424	735	8	4046	311	692	8	4520
ΔSM4	–	–	–	–	−213	356	−1659	18	−341	618	−3424	8	−228	625	−4219	36
SS	−19,1	5,4	−30,1	−8,2	−19,0	5,4	−30,2	−8,7	−19,6	5,3	−30,9	−8,5	−16,3	5,5	−31,2	−3,7
ΔSS	–	–	–	–	−0,1	0,5	−1,1	0,9	0,4	0,8	−0,9	2,9	−2,8	3,5	−9,6	5,2
FB1	27,8	4,4	16,2	36,4	27,3	4,4	16,4	35,9	26,5	4,5	15,4	35,2	25,7	5,3	14,1	35,2
ΔFB1	–	–	–	–	0,5	0,4	−0,6	1,6	1,3	0,6	−0,1	2,5	2,1	3,2	−4,0	9,9
FB2	17,5	7,6	−0,6	32,7	17,4	7,6	−0,7	32,7	16,1	7,7	−1,4	32,2	18,9	8,1	1,5	37,2
ΔFB2	–	–	–	–	0,1	0,1	−0,2	0,7	1,4	0,6	0,4	3,5	−1,4	1,9	−4,6	3,7
FB3	6,6	7,6	−7,9	24,3	3,9	7,6	−9,9	21,1	1,4	7,6	−12,6	18,4	−8,7	7,2	−20,8	6,6
ΔFB3	–	–	–	–	2,6	0,3	2,1	3,3	5,1	0,5	4,2	6,2	15,3	1,6	10,0	17,7
FB4	−0,4	8,2	−22,6	14,7	−2,5	8,3	−26,4	12,5	−6,0	8,2	−28,0	9,2	−12,9	7,3	−26,9	1,5
ΔFB4	–	–	–	–	2,2	0,3	1,7	3,7	5,6	0,4	4,7	6,7	12,5	1,7	4,3	14,8
FB5	−15,3	5,7	−27,1	1,1	−16,5	5,2	−26,5	−0,3	−19,2	4,9	−27,5	−4,6	−20,9	4,4	−28,0	−7,0
ΔFB5	–	–	–	–	1,2	1,1	−1,3	4,0	3,8	1,3	0,2	5,9	5,5	2,1	0,2	9,5
FB6	−19,4	7,2	−31,2	−1,2	−16,0	4,3	−21,5	−2,3	−16,7	3,6	−21,8	−5,3	−17,3	2,8	−22,5	−8,4
ΔFB6	–	–	–	–	−3,4	4,0	−12,8	3,4	−2,6	4,8	−13,2	5,1	−2,1	5,7	−14,5	7,8
FB7	−24,6	6,0	−32,7	−12,5	−18,4	2,5	−22,9	−12,5	−18,8	1,8	−21,8	−16,1	−19,3	2,0	−22,8	−15,5
ΔFB7	–	–	–	–	−6,2	4,3	−13,9	1,9	−5,8	5,5	−15,4	4,3	−5,3	5,8	−13,4	6,7
FB8	−27,7	4,9	−33,8	−11,7	−28,4	4,2	−33,8	−13,7	−29,7	3,6	−34,1	−17,2	−30,1	3,4	−33,9	−22,2
ΔFB8	–	–	–	–	0,8	3,3	−9,1	11,5	2,0	3,4	−6,3	13,8	2,5	4,3	−5,1	12,1
FB9	−29,9	4,3	−35,6	−17,3	−29,7	3,2	−33,6	−18,5	−30,4	2,7	−33,4	−22,1	−30,8	2,9	−33,8	−24,4
ΔFB9	–	–	–	–	−0,3	3,4	−6,1	12,1	0,5	3,5	−7,0	12,6	0,9	4,0	−5,9	11,3
FB10	−27,4	6,4	−35,9	−9,7	−24,3	3,8	−27,6	−11,0	−25,2	2,7	−27,9	−14,7	−26,3	1,2	−27,4	−22,8
ΔFB10	–	–	–	–	−3,1	3,5	−9,5	5,4	−2,2	4,4	−9,8	6,4	−1,1	5,8	−8,9	13,3

C1, condition without mask; C2, condition with surgical mask; C3, condition with FFP2 mask; C4, condition with transparent mask; IL, median sound intensity level; JL, jitter local; SL, shimmer local dB; HNR, harmonics-to-noise ratio; CPPS, smoothed cepstral peak prominence; AVQI, Acoustic Voice Quality Index; SM, spectral moment; SS, spectral slope; FB, mean energy in 1-kHz frequency bands.
Darker grey boxes indicate nonsignificant differences (corresponding with Wilcoxon test results in TABLE 2).

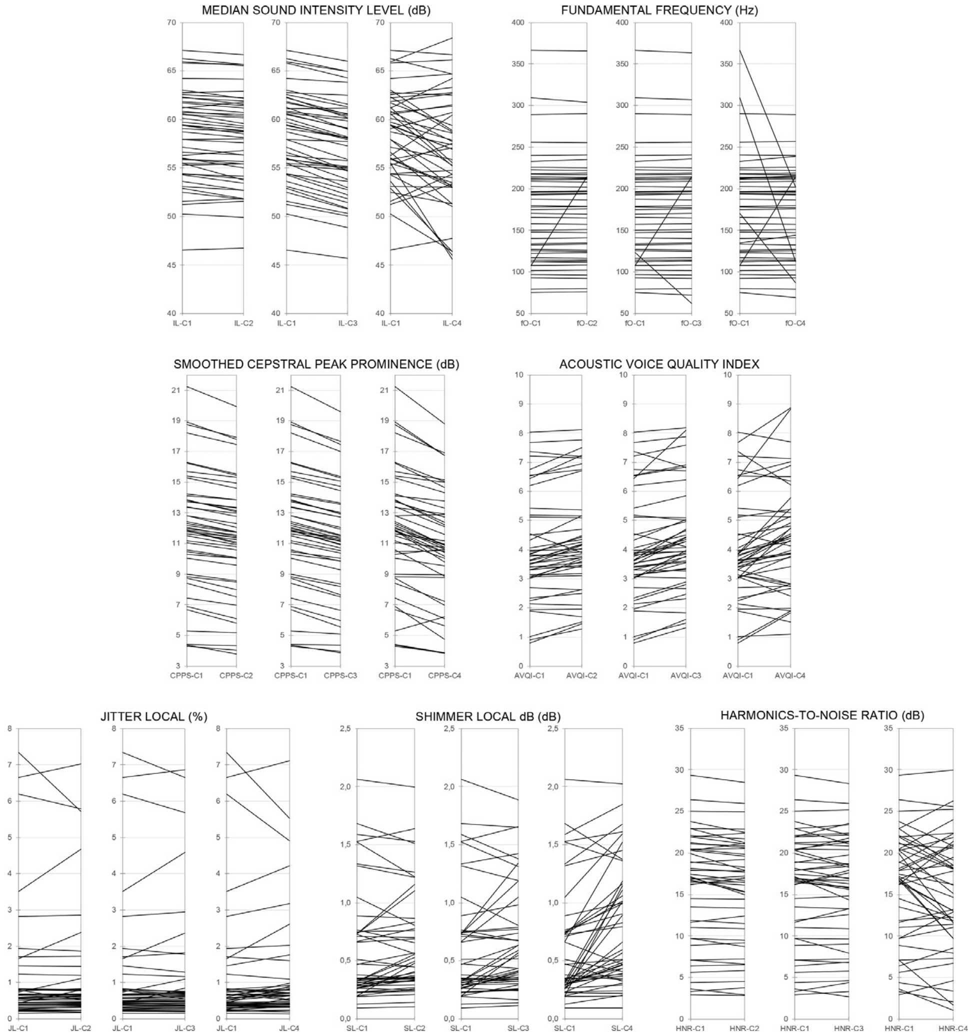

FIGURE 3. Multiple line plots illustrating differences per token in seven vocal physiology-related markers between without-mask condition (C1) and three with-mask conditions (C2, C3 and C4).

changed significantly from C3 data. IL decreased with 1.3 dB on average (with ΔIL between 0.0 and 2.5). The voice quality-related measures CPPS and AVQI, on average, decreased with 0.76 (with ΔCPPS between 0.00 dB and 1.63 dB) and increased with 0.36 (with ΔAVQI between -1.66 and 0.56), respectively. F_2 also decreased significantly with 56 Hz on average (with ΔF_2 between -75 Hz and 463 Hz). Spectral skewness and spectral kurtosis both increased significantly with 4.5 (with ΔSM3 between -27.9 and 1.2) and 341 (with ΔSM4 between -3424 and 8), respectively. Energy in all except the ninth spectral kHz bands changed significantly: mean decrease of 1.3 dB/Hz in FB1 (with ΔFB1 between -0.1 dB/Hz and 2.5 dB/Hz), of 1.4 dB/Hz in FB2 (with ΔFB2 between 0.4 dB/Hz and 3.5 dB/Hz), of 5.1 dB/Hz in FB3 (with ΔFB3 between 4.2 dB/Hz and 6.2 dB/Hz), of 5.6 dB/Hz in FB4 (with ΔFB4 between 4.7 dB/Hz and 6.7 dB/Hz), of 3.8 dB/Hz in FB5 (with ΔFB5 between 0.2 dB/Hz and 5.9 dB/Hz) and of 2.0 dB/Hz in FB8 (with ΔFB8 between -6.3 dB/Hz and 13.8 dB/Hz), and mean increase of 2.6 dB/Hz in FB6 (with ΔFB6 between -13.2 dB/Hz and 5.1 dB/Hz), of 5.8 dB/Hz in FB7 (with ΔFB7 between -15.4 dB/Hz and 4.3 dB/Hz) and of 2.2 dB/Hz in FB10 (with ΔFB10 between -9.8 dB/Hz and 6.4 dB/Hz). The differences between C1 and C3 are illustrated in the middle line graphs of Figures 3, 4 and 5.

No mask (C1) versus transparent mask (C4). Only seven of the 26 (ie, 26.9 %) acoustic markers, were not significantly impacted by placement of the transparent mask on VESPA: f_O, JL, B_{F1}, SM1, SM2, FB9, and FB10. IL decreased with 1.5 dB on average (with ΔIL between -4.2 dB and 9.8 dB). On average, voice quality worsened significantly as measured by the four following acoustic indices: SL and AVQI increased respectively with 0.20 (with ΔSL between -0.95 dB and 0.31 dB) and 0.41 (with ΔAVQI between -2.40 and 1.14), whereas HNR and CPPS decreased with respectively 0.88 (with ΔHNR between -5.20 dB and 8.32 dB) and 1.01 (with ΔCPPS between -0.91 dB and 2.42 dB). The formants were also significantly affected by C4: F_1 increased with 162 Hz (with ΔF_1 between -550 Hz and 476 Hz), whereas F_2 decreased with 113 Hz (with ΔF_2 between -71 and 736) and its bandwidth also decreased with 115 Hz (with ΔB_{F2} between -893 Hz and 535 Hz). The following

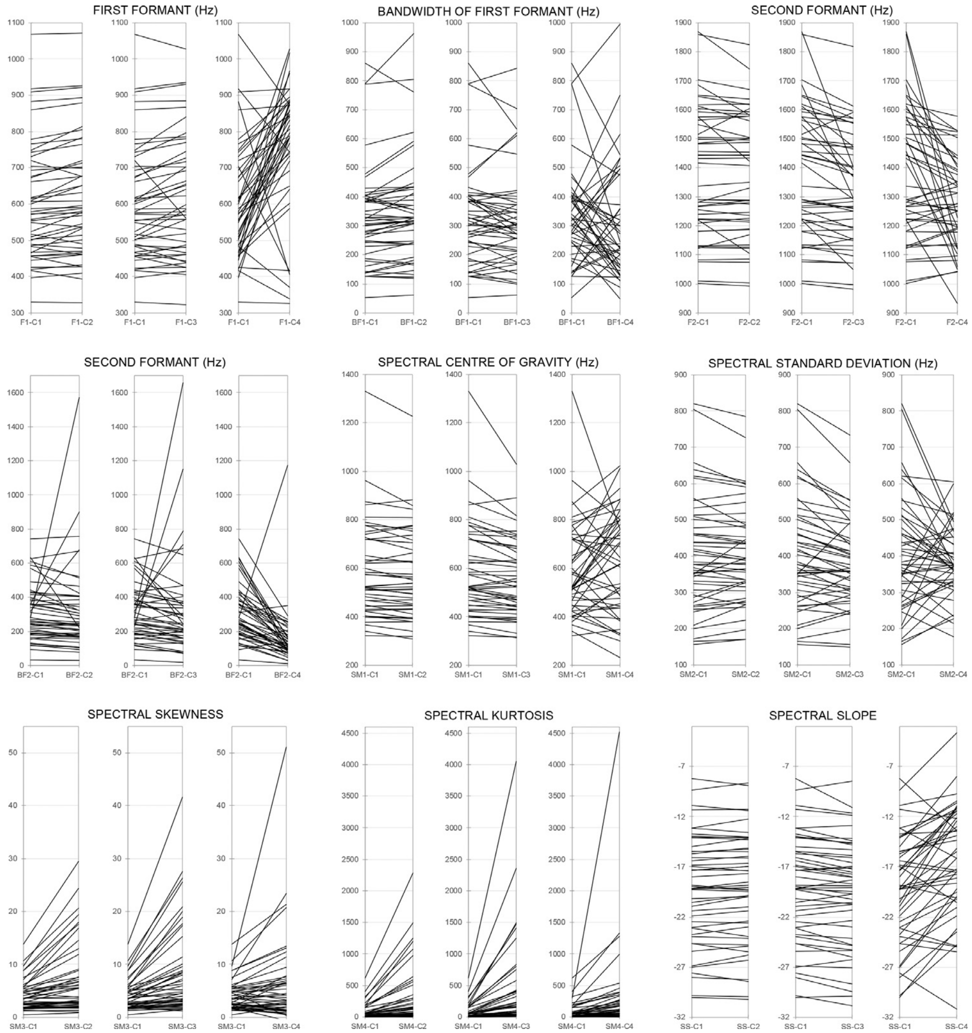

FIGURE 4. Multiple line plots illustrating differences per token in nine frequency-domain speech signal properties between without-mask condition (C1) and three with-mask conditions (C2, C3 and C4).

spectral energy distribution markers increased significantly: increase in skewness of 2.2 (with ΔSM3 between -41.3 and 4.2), kurtosis of 228 (with ΔSM4 between -4219 and 36), slope of 2.8 (with ΔSS between -9.6 and 5.2), energy in FB2 of 1.4 dB/Hz (with ΔFB2 between -4.6 dB/Hz and 3.7 dB/Hz), energy in FB6 of 2.1 dB/Hz (with ΔFB6 between -14.5 dB/Hz and 7.8 dB/Hz) and energy in FB7 of 5.3 dB/Hz (with ΔFB7 between -13.4 dB/Hz and 6.7 dB/Hz). Finally, the following spectral energy distribution markers decreased significantly: energy in FB1 of 2.1 dB/Hz (with ΔFB1 between -4.0 dB/Hz and 9.9 dB/Hz), energy in FB3 of 15.3 dB/Hz (with ΔFB3 between 10.0 dB/Hz and 17.7 dB/Hz), energy in FB4 of 12.5 dB/Hz (with ΔFB4 between 4.3 dB/Hz and 14.8 dB/Hz), energy in FB5 of 5.5 dB/Hz (with ΔFB5 between 0.2 dB/Hz and 9.5 dB/Hz), and energy in FB8 of 2.5 dB/Hz (with Δ between -5.1 dB/Hz and 12.1 dB/Hz). The differences between C1 and C4 are illustrated in the right line graphs of Figures 3, 4, and 5.

Additional information on the filtering by the RPMs is provided by the averaged spectra (showing energy per 100-Hz bin) in Figure 6. In general, all RPMs have attenuated the energy at the majority of frequency bins: most by C4 and least by C2. This is very similar for the recordings of both vowel and sentences. However, for the vowel recordings the main attenuation occurred from circa 1500 Hz to circa 5300 Hz, whereas for the sentences recordings the attenuation occurred from approximately 1400 Hz to at least 10000 Hz. Furthermore, surrounding 6 kHz all RPMs have boosted spectral amplitudes, and above that surrounding 1 kHz the transparent RPM raised the spectral amplitudes. Additional raise can be seen at some 100-Hz bins from circa 9000 Hz.

DISCUSSION

This VESPA study is similar to our previous study[11] in the context of reliability of acoustic voice measures. However, instead of assessing the influence of ambient noise and mobile communication devices on a set of selected measures, we now investigated the effect of mouth-and-nose-covering RPMs on several acoustic voice and speech measures using VESPA, a voice and/or speech sound-radiating

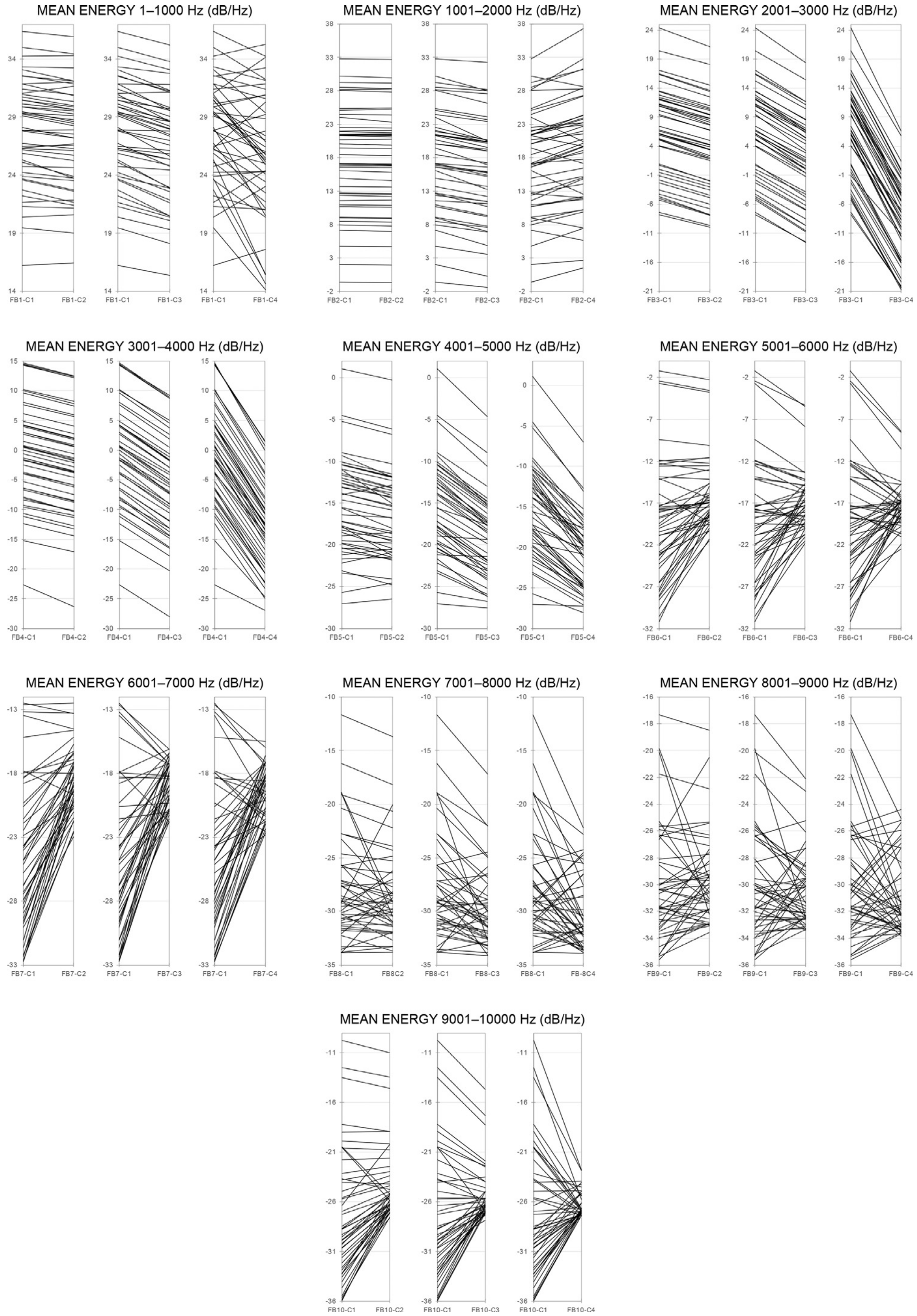

FIGURE 5. Multiple line plots illustrating differences per token in mean energy in 10 1-kHz frequency bands between without-mask condition (C1) and three with-mask conditions (C2, C3, and C4).

body-like model made out of spare parts. Because RPMs may add resistance to and filtering of airborne speech signals and thereby affect clinically relevant measures of speech sounds, a set of acoustic markers relevant to voice and speech clinics was selected: IL, f_O, CPPS, JL, SL, HNR, SS and AVQI as markers related to vocal physiology and voice quality;[15-17] F_1, B_{F1}, F_2 and B_{F2} as markers related to articulatory and resonatory phenomena such as vowel differentiation[18] and nasality;[19,20] SM1, SM2, SM3 and SM4 as markers related for example fricative differentiation,[21] and finally FB1-FB10 to address energy shifts in 1-kHz bands across the frequency-domain. An earlier small-sized and home-based exploration by Maryn[8] did not reveal consistent influence of disposable surgical RPM on f_O, IL, JL, SL, CPPS and AVQI. However, Corey et al[9] showed an important attenuation of frequencies above 1 kHz and even more above 4 kHz. From this, significant effect of this study's RPMs on spectrum-based markers (ie, CPPS, AVQI, SS, SM1-SM4, and FB1-FB10) was anticipated. Time-domain markers, on the other side, were

TABLE 2.
Differences in the Acoustic Markers on the Speech and/or Voice Signals Between the No-Mask and the Three With-Mask Recording Conditions

Acoustic marker	2-way ANOVA (C1-C2-C3-C4) *P*	Dunnett (C1−C2) *P*	Dunnett (C1−C3) *P*	Dunnett (C1−C4) *P*	Acoustic marker	2-way ANOVA (C1-C2-C3-C4) *P*	Dunnett (C1−C2) *P*	Dunnett (C1−C3) *P*	Dunnett (C1−C4) *P*
f_O	0.168	0.924	0.994	0.277	SM3	<0.001	<0.001	<0.001	0.010
IL	<0.001	0.349	<0.001	<0.001	SM4	<0.001	0.021	<0.001	0.012
JL	0.913	0.903	0.875	0.875	SS	<0.001	0.975	0.437	<0.001
SL	<0.001	0.595	0.102	0.000	FB1	<0.001	0.319	<0.001	<0.001
HNR	0.024	0.568	0.959	0.013	FB2	<0.001	0.906	<0.001	<0.001
CPPS	<0.001	<0.001	<0.001	<0.001	FB3	<0.001	<0.001	<0.001	<0.001
AVQI	<0.001	0.001	<0.001	<0.001	FB4	<0.001	<0.001	<0.001	<0.001
F_1	<0.001	0.818	0.819	<0.001	FB5	<0.001	<0.001	<0.001	<0.001
B_{F1}	0.145	0.684	0.891	0.394	FB6	<0.001	<0.001	<0.001	0.001
F_2	<0.001	0.908	0.011	<0.001	FB7	<0.001	<0.001	<0.001	<0.001
B_{F2}	<0.001	0.945	0.918	<0.001	FB8	<0.001	0.333	<0.001	<0.001
SM1	0.002	0.770	0.197	0.107	FB9	0.071	0.908	0.578	0.144
SM2	0.058	0.996	0.066	0.196	FB10	<0.001	<0.001	<0.001	0.114

C1, condition without mask; C2, condition with surgical mask; C3, condition with FFP2 mask; C4, condition with transparent mask; Z, Wilcoxon test value; f_O, median fundamental frequency; IL, median sound intensity level; JL, jitter local; SL. shimmer local dB; HNR, harmonics-to-noise ratio; CPPS, smoothed cepstral peak prominence; AVQI, Acoustic Voice Quality Index; F_1, first formant; B_{F1}, bandwidth of F_1; F_2, second formant; B_{F2}, bandwidth of F_2; SM, spectral moment; SS, spectral slope; FB, mean energy in 1-kHz frequency bands.

Darker grey boxes denote non-significant ($\alpha > .05$) findings.

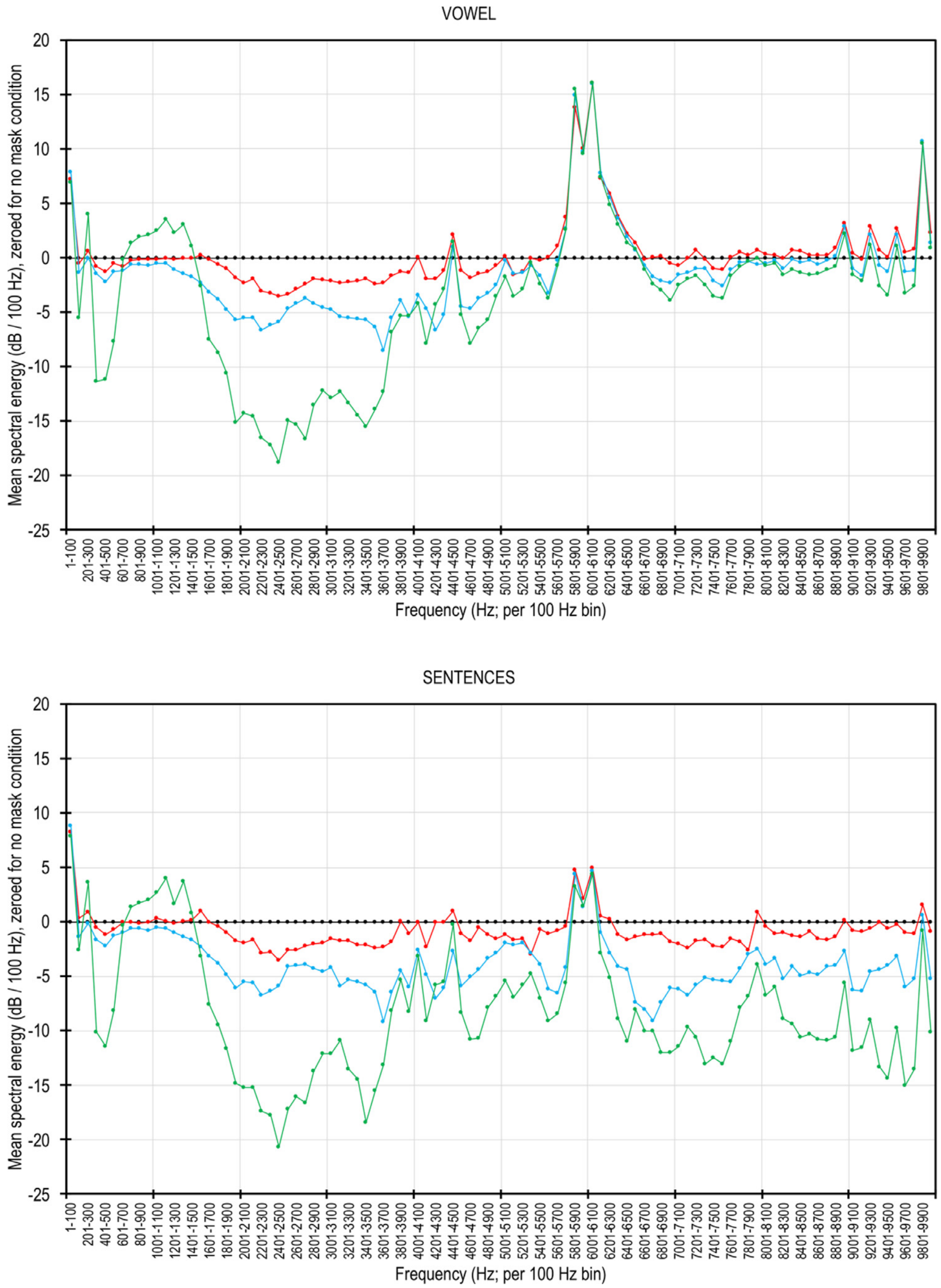

FIGURE 6. Averaged spectra (with frequency bins of 100 Hz) across the 47 vowel (top) and sentences tokens (bottom) for the four recording conditions, and after zeroing relative to the no mask spectra: no mask (black), surgical mask (red), FFP2 mask (blue), and transparent mask (green). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

anticipated not to be influenced by the RPMs. Also, after Corey et al,[9] surgical mask was expected to have least impact on the acoustic speech properties.

Although differences occurred for all the measures (only in FB1 and FB2 the differences were quasi null; see Table 1), only in nine measures this was statistically significant and even then the differences were relatively small from a clinical point of view. For example, |ΔAVQI| > 0.54 has been described to be a clinically relevant change in overall voice quality beyond test-retest variability.[22] In C2 this was found in 10 of the 47 (21.3 %) tokens and the largest change was |-0.85|. In C3 this occurred somewhat more: in 13 (27.7 %) and the largest change was |-1.66|. However, for C4 such clinically relevant influence emerged in 21 of the 47 (44.7 %) tokens and the largest change was |-2.40|. C2 thus clearly had the smaller effect on acoustically measured voice quality than C3 and C4. This mask was the lightest, thinnest and most foldable of the three nose-and-mouth-covering RPMs that were utilized in this study and we therefore hypothesize that it was least resistant to speech sound radiation, as evidenced by the average ΔIL of only 0.4 dB. The other RPMs, and especially the transparent mask from relatively rigid plastic material, were less transmissive for speech sounds, which resulted in significantly lower ILs and more decreased FB values even in the lowest kHz intervals. Probably as a consequence of RPMs' vibratory properties and sound filtering, also the formants' center frequencies and/or bandwidths altered. This became especially clear in C4 with significantly changed F_1, $F_{2,}$ and B_{F2}. In C3 only F_2 was affected by the RPM. Concerning the spectral moments: center of gravity nor standard deviation were changed by any of the RPMs in this study. Spectral skewness and peakedness, on the other side, were significantly influenced by all masks, but again be it most by C4 and least by C2.

Acoustic markers f_O, JL, B_{F1}, SM1, SM2, and FB9 were not affected by the RPM's in this study. For f_O this came as no surprise as earlier research already indicated its robustness against factors like recording system and environmental noise (see Maryn et al.[11] for an overview). JL, on the other side, is known to lose accuracy and reliability when recording-related noise exceeds certain levels.[23] However, in the present study data acquisition system and environment stayed the same across recordings. As far as we know, such influences have not yet been investigated in B_{F1}, SM1, SM2, and FB9. Filtering by RPM, however, had no influence on the magnitude-weighted mean of the frequencies in the spectrum (spectral center of gravity or SM1), on how much the frequencies in a spectrum deviate from the center of gravity (spectral standard deviation or SM2), nor on the bandwidth of the first formant (B_{F1}). Finally, only the ninth 1-kHz frequency band remained unaffected by the RPMs in this study.

Energy in 1-kHz frequency bands from 2001 to 7000 Hz were affected by all RPMs in this study. Energy of lower frequencies were only influenced by the FFP2 and transparent masks. This is consistent with Corey et al,[9] who also found effects from 1 kHz (and especially from 4 kHz), and least effect by the surgical mask. This may be relevant for several phoneme groups, but especially for voiceless ([s], [S], [T], [f]) as well as voiced ([z], [Z], [D], [v]) fricatives that have their main energy content in higher frequency intervals.[21,24]

Healthcare workers as well as other groups can choose a mask depending on what requires priority. For example, FFP masks reduce risk of infection more than surgical masks,[25,26] but they also attenuate speech sounds more than surgical masks [[9], present study]. So, when in unsafer situations with physical distances of < 1 m,[25] one will have to prioritize respiratory protection and choose FFP mask. However, in situations where sufficient physical distance can be maintained and adequate spoken communication or least-filtered speech recording is required, as required for clinical assessment of acoustic voice/speech signals,[11,23] one can opt for surgical/medical mask.

Based on the results of this study, interpretation of acoustic voice/speech measures may depend on which RPM is used. For example, when patients wear surgical/medical mask, the clinician won't need to deviate from the normative/reference information or diagnostic thresholds for f_O, IL, JL, SL, HNR, F_1, B_{F1}, F_2, SM1, SM2 and SS. For AVQI and CPPS there was a significant influence of this RPM, and therefore their norm references (as for example in Batthyany et al[27]) should be 0.25 higher and 0.57 lower, respectively, based on the mean differences between C1 and C2. For comparability of data across time or treatment, it is essential to establish as much as possible the same recording conditions.[11,23] This also includes masks/shields between speaker and microphone, and therefore clinicians should guard that same type of RPM is worn during pre-, within- and post-treatment recordings.

LIMITATIONS AND FUTURE DIRECTIONS

Although a representative set of voice/speech signals was used and VESPA was considered an acceptable approximation of a human speech producer, there are several limitations regarding the present study that may restrict generalizability of the findings and provide direction for future research. First, in VESPA for example the round loudspeaker of 6.5 cm diameter is fixed immediately after a circle opening of 3.5 cm. This design leaves space between loudspeaker and fiber glass by which sound may have been restricted to radiate. In real speakers, however, similar partial occlusion of the mouth occurs by the lips and/or teeth. Furthermore, all signals in all recording conditions were presented in the same way. Therefore, VESPA's design was not expected to influence comparison between recording conditions. Nevertheless, more representative models are commercially available and have already been used in speech research (eg, Bottalico et al[28]). They could also be applied for well-controlled investigation of acoustic effects of RPMs on speech signals. Second, because we worked with a model instead of human speakers, no speech movements were involved and effects of RPMs on speech behaviors and intentions to move articulators while wearing a nose-and-mouth-covering mask was not investigated. Additionally, added noise (if any) due to friction of facials hairs against RPM while speaking was not taken into account either. However, Corey et al[9] used both a simulator/loudspeaker and a bearded human. Although differences between these two situations were not statistically analyzed, similar sound attenuation especially above 1 kHz emerged (be it less for human in most RPMs). Nonetheless, future research should include humans, as in eg, Corey et al,[9] to consider these influences on clinically relevant speech and voice measures such as formant properties, spectral moments, HNR, CPPS and AVQI. Third, we used three RPM types with different materials, weight, thickness and pliability. However, there are many other RPMs available (eg, cloth masks with/without air filtering pads), and it would be interesting to expand this study with other masks. Fourth, we analyzed RPM influence on 26 acoustic markers that are considered relevant to voice (eg, f_O and voice quality), speech (eg, vowel differentiation and fricative distinction), and consequently also intelligibility.[29] It would also be interesting to evaluate how listeners react to speakers wearing various nose-and-mouth-covering RPMs, and how aspects like voice quality, articulatory precision, nasality, speech intelligibility, speech acceptability, etc. are perceived in these sound-filtered speaker conditions.

CONCLUSION

Effects of RPMs on various speech sound properties were least in case of disposable surgical mask and strongest with the plastic transparent mask. RPM should be added to the list of "noise" factors in the context of spoken communication and its acoustic measures. Future research is warranted to better understand how RPMs affect speech sound production and propagation. Finally, healthcare personnel as well as other people with professional and/or recreational speech activities are warranted to consider the present results when choosing between RPMs, next to other arguments related to respiratory protection, face dermatology, ecological ballast, nonverbal communication, etc.

AUTHOR CONTRIBUTIONS

Youri Maryn: conceptualization, methodology, software, formal analysis, writing - original draft, visualization. Floris L. Wuyts: formal analysis, writing - review & editing. Andrzej Zarowski: writing - review & editing, supervision.

CONFLICT OF INTEREST

No conflict of interest exists.

ACKNOWLEDGMENTS

We thank Carine Franck (professional window dresser) for providing the fashion mannequin doll, and Wilfried De Coninck (retired electromechanical engineer) for recuperating the loudspeaker that was used in VESPA.

FUNDING

No funding was received for this work.

REFERENCES

1. Johnson AT. Respirator masks protect health but impact performance: a review. *J Biol Eng*. 2016;10:4.
2. Chodosh J, Weinstein BE, Blustein J. Face masks can be devastating for people with hearing loss. *BMJ*. 2020;370:m2683.
3. Caretti DM, Strickler LC. Speech intelligibility during respirator wear: influences of respirator speech diaphragm size and background noise. *AIHA J*. 2003;64:846–850.
4. Mendel LL, Gardino JA, Atcherson SR. Speech understanding using surgical masks: a problem in health care? *J Am Acad Audiol*. 2008;19:686–695.
5. Radonovich LJ, Yanke R, Cheng J, et al. Diminished speech intelligibility associated with certain types of respirators worn by healthcare workers. *J Occup Environ Hyg*. 2010;7:63–70.
6. Willum KJ, Feth L, Hoglund E. The effects of surgical masks on speech perception in noise. *Proc Meet Acoust*. 2013;19: 060125.
7. Atcherson SR, Mendel LL, Baltimore WJ, et al. The effect of conventional and transparent surgical masks on speech understanding in individuals with and without hearing loss. *J Am Acad Audiol*. 2017;28:58–67.
8. Maryn Y. *Initial Exploration at Home: Acoustic Voice Markers With and Without Disposable Surgical Face Mask*. 2020.. Lokeren, Belgium; https://www.phonanium.com/videos-and-tutorials/#toggle-id-14.
9. Corey RM, Jones U, Singer AC. Acoustic effects of medical, cloth, and transparent face masks on speech signals. *J Acoust Soc Am*. 2020;148: 2371.
10. Maryn Y, De Bodt M, Barsties B, et al. The value of the Acoustic Voice Quality Index as a measure of dysphonia severity in subjects speaking different languages. *Eur Arch Otorhinolaryngolog*. 2014;271: 1609–1619.
11. Maryn Y, Ysenbaert F, Zarowski A, et al. Mobile communication devices, ambient noise, and acoustic voice measures. *J Voice*. 2017; 31:248.e11–248.e23.
12. Svec JG, Granqvist S. Guidelines for selecting microphones for human voice production research. *Am J Speech Lang Pathol*. 2010;19:356–368.
13. Maryn Y, Zarowski A. Calibration of clinical audio recording and analysis systems for sound intensity measurement. *Am J Speech Lang Pathol*. 2015;24:608–618.
14. Maryn Y, Weenink D. Objective dysphonia measures in the program Praat: smoothed cepstral peak prominence and Acoustic Voice Quality Index. *J Voice*. 2015;29:35–43.
15. Jiang J, Lin E, Hanson DG. Vocal fold physiology. *Otolaryngol Clin North Am*. 2000;33:699–718.
16. Maryn Y, Roy N, De Bodt M, et al. Acoustic measurement of overall voice quality: a meta-analysis. *J Acoust Soc Am*. 2009;126:2619–2634.
17. Maryn Y, Corthals P, Van Cauwenberge P, et al. Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels. *J Voice*. 2010;24:540–555.
18. Hillenbrand J, Getty LA, Clark MJ, et al. Acoustic characteristics of American English vowels. *J Acoust Soc Am*. 1995;97:3099–3111.
19. Hawkins S, Stevens KN. Acoustic and perceptual correlates of the non-nasal−nasal distinction for vowels. *J Acoust Soc Am*. 1985;77: 1560–1575.
20. Styler W. *On the Acoustical and Perceptual Features of Vowel Nasality*. Boulder, CO: University of Colorado; 2015. [unpublished PhD thesis].
21. Maniwa K, Jongman A, Wade T. Acoustic characteristics of clearly spoken English fricatives. *J Acoust Soc Am*. 2009;125:3962–3973.
22. Barsties B, Maryn Y. Test-Retest-Variabilität und interne Konsistenz des Acoustic Voice Quality Index [Test-retest variability and internal consistency of the Acoustic Voice Quality Index]. *HNO*. 2013;61:399–403.
23. Deliyski DD, Shaw HS, Evans MK. Adverse effects of environmental noise on acoustic voice quality measurements. *J Voice*. 2005;19:15–28.
24. Kent RD, Read WC. *The Acoustic Analysis of Speech*. 2nd ed. San Diego, CA: Singular Publishing Group; 2002.
25. Chu DK, Akl EA, Duda S, et al. COVID-19 Systematic Urgent Review Group Effort (SURGE) study authors. Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARSCoV-2 and COVID-19: a systematic review and meta-analysis. *Lancet*. 2020;395:1973–1987.
26. Wilson AM, Abney SE, King M-F, et al. COVID-19 and use of non-traditional masks: how do various materials compare in reducing the risk of infection for mask wearers? *J Hosp Infect*. 2020;105:640–642.
27. Batthyany C, Maryn Y, Trauwaen I, et al. A case of specificity: how does the Acoustic Voice Quality Index perform in normophonic subjects? *Appl Sci*. 2019;12:2527.
28. Bottalico P, Codino J, Cantor-Cutiva LC, et al. Reproducibility of voice parameters: the effect of room acoustics and microphones. *J Voice*. 2020;34:320–334.
29. De Bodt MS, Hernández-Díaz HM, Van De Heyning PH. Intelligibility as a linear combination of dimensions in dysarthric speech. *J Commun Disord*. 2002;35:283–292.
30. Titze I, Baken R, Bozeman K, et al. Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. *J Acoust Soc Am*. 2015;137:3005–3007.
31. Titze IR. Some consensus has been reached on the labeling of harmonics, formants, and resonances. *J Voice*. 2016;30:129.