informa healthcare

ORIGINAL ARTICLE

Impact of vocal load on breathiness: Perceptual evaluation

ANGÉLIQUE REMACLE[1], JEAN SCHOENTGEN[2], CAMILLE FINCK[3], AGNÈS BODSON[1] & DOMINIQUE MORSOMME[1]

[1]*Department of Psychology, Cognition and Behavior, University of Liège, Belgium,* [2]*L.I.S.T., Faculty of Applied Sciences, University of Bruxelles, Belgium,* [3]*ENT Department, CHU of Liège and Faculty of Psychology, University of Mons, Belgium*

PROOF ONLY

Abstract

Objectives. To evaluate the impact on voice of 2 hours of continuous oral reading. *Methods.* Fifty normophonic women underwent two sessions of voice loading in which the required intensity level varied: 60–65 dB(A) for the first session, and 70–75 dB(A) for the second session. Ten expert judges evaluated the breathiness of one sentence recorded before and after each loading session. Pairs of stimuli were presented randomly to the judges, who were asked to designate the breathiest sample. *Results.* A significant decrease in breathiness was observed following both sessions, suggesting an improvement of voice subsequent to loading. When comparing the two intensity levels, no difference was found for breathiness after vocal loading.

Key words: *Breathiness, pairwise comparison, perceptual analysis, vocal load*

Introduction

Vocal load corresponds to the amount of work accomplished by the laryngeal mechanism over time (1), mostly determined by the duration, intensity, and frequency (F0) of vocalization. Prolonged use of the voice has been identified as a risk factor for dysphonia, mainly when intensity and F0 are high (2,3). This study is part of a research project assessing the effects of vocal load by means of a reading task conducted in the laboratory. The project's objective is to improve our understanding of duration and intensity as loading factors. Fifty female speakers underwent two vocal load sessions. The first session consisted in 2 hours of reading at low intensity (LI) level. The second session comprised 2 hours of reading at high intensity (HI) level. We wanted to answer the following questions: 1) Does the voice vary during vocal loading? 2) Can differences be observed between the two vocal load sessions as a function of vocal intensity? In a previous study, we reported objective measurements and subjective self-ratings (4).

The present study is based on perceptual voice evaluation, which consists in assessing the signal perceived through an auditory input modality. Although it is sensitive to numerous sources of error and bias, perceptual analysis has the advantage of being convenient, inexpensive, and useful, in both clinical and research settings (5). In our clinical practice, testing 'by ear' is the first and most accessible modality for assessing the voice. However, it can be a difficult skill to master due to its subjective nature and its potential lack of sensitivity and reproducibility over time. Several studies have shown evidence of unreliability due to intra- or interrater variability (6–10). This unreliability may be explained by factors related to the judge, to the task, or to the interaction between the two (8). Judge-related unreliability is explained by the facts that judges use different strategies and that most perceptual analysis tasks require a comparison of stimuli with internal standards specific to each listener (8). Listeners develop individual, variable, and unstable internal standards, based on their own experience with voices (9,11). Task-related unreliability depends on the type of scale, its resolution, the voice samples, and the effects of context (8). The listening context can

Correspondence: Angélique Remacle, SLP, PhD, Unité Logopédie de la voix, Département de Psychologie, Cognition et comportement, Université de Liège, Rue de l'Aunaie 30 (B38), 4000 Liège, Belgium. E-mail: Angelique.Remacle@ulg.ac.be

(Received 13 September 2013; accepted 24 December 2013)

ISSN 1401-5439 print/ISSN 1651-2022 online © 2014 Informa UK, Ltd.
DOI: 10.3109/14015439.2014.884161

definitely influence listeners' internal standards. A given sample may sound breathier if it is presented after a very non-breathy sample than after a very breathy sample, just as an identical amount of light will be perceived as more intense if one is coming from a dark environment than if one is coming from a well-lit place.

To the best of our knowledge, there have been few studies of the perception of changes following vocal loading. In 1962, Sherman and Jensen assessed the effects of one-and-a-half hours of reading in a conversational voice, followed by 30 minutes of silence, in 15 men with normal voices and 15 men with harsh voice (12). The perceptual judgment concerned a spoken standardized text before reading (T1), after 45 minutes of reading (T2), after one-and-a-half hours of reading (T3), and after 30 minutes of silence (T4). Thirty-two seniors and graduate students, majors in speech pathology, evaluated the degree of harshness of each voice sample using an equal-appearing interval scale ranging from 1 (least) to 7 (most). Rather unexpectedly, the subjects with normal voice showed a decrease in harshness between T1 and T2, and between T1 and T3, followed by an increase between T3 and T4, returning to approximately the initial level of harshness observed at T1. No significant differences were found for subjects with harsh voice.

In 1973, Stone and Sharf studied the effect of the duration, intensity, and frequency of vocal loading in 10 men with normal voices (3). The task consisted in producing vowel lists for 20 minutes in nine different conditions (3 intensity levels × 3 frequencies). The three intensity levels were 75, 80, and 85 dB SPL measured 30 cm from the lips. The three frequency levels corresponded to 20%, 50%, and 80% of each speaker's frequency range. The nine conditions were administered on nine different days. Five graduate students in speech pathology and audiology conducted the perceptual analysis using an equal-appearing interval scale ranging from 0 (no change) to 6 (extreme change). For each speaker, the voice samples collected before and every 5 minutes during vocal loading were compared pairwise to determine the impact of intensity, frequency, and duration. The results showed a significant difference between the three frequency levels: the higher-pitched the voice, the greater the changes perceived during vocal loading. As for duration, a significant change was perceived after 20 minutes of loading at 80% of the frequency range, but not after low (20%) or medium (50%) frequency loading. In all conditions, the largest changes were observed in the first 5 minutes of vocal loading. On the other hand, no significant differences were observed between the three intensity levels of the task. It should be noted that when changes were perceived, the reported results do not enable deciding whether they represented an improvement or a deterioration, or which voice quality was affected by the change.

In 1987, Neils and Yairi studied the impact of the duration and intensity of vocal loading in six women with normal voices (13). Each participant read out loud for 45 minutes in three different background noise conditions: 50 dB, 70 dB, and 90 dB. The judges evaluated continuous speech samples before, after 15, 30, and 45 minutes of reading, and after 15, 30, and 45 minutes of silence, in each of the three background noise conditions. Nineteen graduate students of speech pathology assessed voice normalcy with an equal-appearing interval scale ranging from 1 (normal) to 7 (abnormal). The results did not show any significant effect of time or intensity.

In 2003, Yiu and Chan did a perceptual analysis of 20 karaoke singers at four points: 1) before singing, 2) after singing 10 songs, 3) after singing five additional songs, and 4) after the last song when the participant reported vocal fatigue and could not sing anymore (14). The vocal material used for the perceptual analysis comprised sustained /a/ sounds, plus the reading of a sentence. Three final-year speech pathology students with a year of clinical experience assessed roughness and breathiness on visual analog scales. Anchor points were used to illustrate different degrees of roughness and breathiness. No significant change in roughness or breathiness was perceived over the four recordings.

In 2009, McAllister et al. analyzed perceptually the voices of 10 children at three points: 1) in the morning, 2) at noon, and 3) in the afternoon during a normal day at a day care center (15). The vocal material involved the repetition of three sentences. Three speech and language pathologists assessed the voice samples on visual analog scales according to roughness, breathiness, hoarseness, and hyperfunction. Among girls, hyperfunction and breathiness tended to increase during the day, whereas hoarseness and hyperfunction tended to increase for the boys. These differences were, however, not statistically significant.

Only two of these studies addressed changes of breathiness consequently to prolonged voice use, and the results did not show significant differences following singing in adults (14) or speaking in children (15). However, breathiness has been identified as an effect of vocal fatigue (16). The present study aims to determine whether the breathiness of voice varied following vocal loading and as a function of the intensity of the load. The breathy characteristic of vocal quality is mainly evaluated in clinics with the B ('Breathiness') subscale of Hirano's GRBAS scale

PROOF ONLY

(17). Breathy voice is characterized by a lack of adduction of the vocal folds (hypoadduction) (18).

Few studies have examined the laryngeal effects of vocal loading. There is no consensus in the literature regarding the effect of vocal load on glottal adduction. Some studies suggest that adduction increases (19–21), whereas others tend to show the opposite (22–24). For example, Stemple and collaborators observed an anterior glottal chink (lack of adduction of the vocal folds) in 6 out of 10 women after 2 hours of reading at 75 to 80 dB (24). Solomon and DiMattia described spindle-shaped vibratory closure patterns (lack of adduction of the vocal folds) in 3 out of 4 women after 2 hours of reading at 75 to 80 dB SPL (23). Gelfer et al. noted a larger amplitude of glottal opening after 1 hour of reading in a group of untrained female singers, but not in trained singers (22). The lack of adduction of the vocal folds is characterized by a perception of breathiness (25). The glottal chink that certain studies have observed after vocal load led us to explore the breathy parameter of voice in order to determine whether it would increase following prolonged voice use. We also checked whether perceived breathiness would vary as a function of the intensity of vocal loading. In point of fact, glottal leakage and perceived breathiness have been found to decrease when vocal intensity is increased (25–27).

The questions we sought to answer by means of perceptual analysis are the following: 1) Is the voice breathier after vocal loading? 2) Is the voice breathier when vocal load intensity is higher?

Materials and methods

Vocal load

Subjects. Fifty women (mean age = 25.4 years, SD = 4.9, range = 21–47 years) underwent two sessions of vocal loading. A videolaryngostroboscopic examination (EndoSTROB Stroboscop; Xion GmbH, Berlin, Germany) and an in-depth anamnesis ruled out all vocal pathologies. The exclusion criteria in choosing participants were as follows: smoking, history of voice problems, voice rehabilitation in the past or present, hearing disorders, upper respiratory infection at the time of the study, and professional or recreational activity involving frequent use of the voice (e.g. singing, theater).

Loading task. The vocal loading task consisted in reading a novel aloud for 2 hours. Each speaker underwent two vocal loading sessions separated by a minimum of 5 days, to ensure voice recovery between the two sessions. Using a decibel meter (DVM805; Velleman, China) placed at a distance of 40 cm from the lips, vocal intensity was constantly monitored to ensure that it was always between 60 and 65 dB(A) in the first session and 70 and 75 dB(A) in the second session. The examiner verbally encouraged the participants to correct the intensity level when it differed from the target level. While reading, the participants were seated in a quiet room (ambient noise < 30 dB(A)). Relative humidity was monitored with a hygrometer (P600; Dostmann Electronic, Wertheim-Reicholzheim, Germany) and maintained at 30% ± 10%. Every 30 minutes, participants took a break and were encouraged to drink a glass of water.

Recording equipment and procedures. The voice samples were acquired in a sound-proof booth (213 × 194 × 219 cm). Recordings were made with Computer Speech Lab software (Kay Elemetrics, Lincoln Park, NJ, USA) and a head-worn microphone with a frequency range of 20 to 20,000 Hz (AKG C420; Harman, Stamford, CT, USA), placed at a distance of 7 cm from the lips. The perceptual analysis was done on voice samples collected PRE and POST 2 hours of vocal loading at LI and HI levels.

Perceptual evaluation

Voice samples. The voice material used for the perceptual judgment task was the French sentence ‘*A cet instant, Vick sortit contempler le jour naissant*’. This was the second sentence from the reading of a phonetically balanced text, at a comfortable frequency and intensity. The sentence was selected and segmented with PRAAT freeware, designed by Paul Boersma and David Weenink (Phonetic Sciences, University of Amsterdam, The Netherlands). We have used a read sentence because we consider that it is more similar to connected speech than a sustained vowel.

The voice samples of the 50 speakers were then classified into four different files: 1) PRE LI session samples ($n = 50$); 2) POST LI session samples ($n = 50$); 3) PRE HI session samples ($n = 50$); and 4) POST HI session samples ($n = 50$).

Perceptual judgment task. A variety of perceptual analysis methods exist. The most widely used are equal interval scales (8), of which Hirano’s GRBAS (17) scale is the best known. Despite the widespread use of these perceptual analysis scales in the clinical setting, their main weakness is the lack of intra- and interrater reliability. According to Teston, reliability is enhanced if the perceptual evaluation is done in comparative mode, with an instantaneous transition between the samples to be judged (28). The stimulus to be judged can be compared with an

PROOF ONLY

explicit external standard (anchored scale), or two voice samples can be compared (pairwise comparison). Kacha and colleagues showed that pairwise comparison increased intra- and interrater reliability compared to the GRBAS scale, for both novice and expert judges (29). When they rate pairwise, the judges do not need to refer to their internal standards because they are comparing two voice samples with each other (9). We therefore opted for this perceptual analysis method to avoid reference to the judges' internal standards, and we expect judgment reliability to increase. Another reason for choosing pairwise comparison is that the perceptual differences between samples were minimal. Therefore, comparing samples pairwise is expected to be easier than scoring on interval scales. Also, comparative ratings are particularly appropriate for confronting a subject with herself PRE and POST vocal loading.

The voice samples were presented and the scores were calculated with Pairwise software, developed by Ali Alpan (L.I.S.T., Faculty of Applied Sciences, University of Brussels). This software creates one-to-one comparisons between the samples so that each speaker is compared with herself for task 1 (PRE LI session versus POST LI session), task 2 (PRE HI session versus POST HI session), and task 3 (POST LI session versus POST HI session). For each pair of stimuli, the judges were asked to answer the question 'In your opinion, which voice is breathier?' The objective was to determine in which sample the breathy parameter was more evident and not to score this parameter in the samples to be judged. The judges could listen to the voice samples as many times as they wished before clicking on the button corresponding to their answer. The judges were required to choose between the two sounds played; they were not given the possibility of answering that the voices were similar in the aspect to be evaluated or that they did not perceive that aspect in either sample. Thus, they had to make a forced choice.

All the tasks were performed on a 13-inch MacBook Pro portable computer. Samples were presented over professional headphones with a frequency range of 18 to 18,000 Hz (Sennheiser HD 202; Sennheiser Electronic GmbH & Co. KG, Wedemark, Germany). The intensity was set at a comfortable level for each judge. The listening sessions were administered individually, in a quiet room. Before each session, the judge was given a written explanation of the task and a definition of the breathiness. Breathy voice was defined as follows: 'Breathy voice is a characteristic of voice quality that is usually clinically evaluated using the GRBAS scale. The perceived breathiness of the voice corresponds to an escape of air from the larynx, caused by the incomplete closure of the vocal folds. The glottis is then expanded, which results in excessive air flow during phonation, and occasionally a dull voice due to reduced timbre.'

To assess the agreement among the different judges (interrater reliability), the same tasks composed of the same voice samples were administered to all of them. However, the randomized order of presentation was different for all judges. To evaluate intrarater reliability, a retest was realized after 7 to 14 days. Each judge therefore completed two listening sessions (test and retest), composed of all the tasks.

Judges. Our jury was made up of 10 expert judges aged 25 to 60 years (mean = 37.4 years). The expert judges, who were recruited among our professional contacts, had theoretical knowledge and regular practical experience in perceptual voice analysis. Of these 10 judges, eight were speech therapists and two were otorhinolaryngologists specializing in voice disorders. All were native French speakers, none had hearing problems, and all were naive regarding the study hypotheses. Table I describes the judges.

Statistical analyses

All data were processed with Statistica software (version 10, StatSoft Inc., Tulsa, OK, USA). Cohen's kappa coefficient was used to measure intrarater reliability. It allows one to measure the agreement between two qualitative variables (test and retest) with the same modalities. Fleiss's kappa coefficient was used to test interrater reliability. It allows one to measure the agreement among several judges who are making a qualitative evaluation with the same modalities. The value of kappa always falls between –1 and 1. To interpret it, we used the classification established by Landis and Koch (Table II) (30).

Finally, to determine whether the duration and intensity of vocal loading had an impact, the judges' responses to each of the test tasks were analyzed. For

Table I. Description of the judges.

Judge	Sex	Age, years	Profession (years of practice)
J1	F	25	Speech therapist (1)
J2	F	26	Speech therapist (2)
J3	F	26	Speech therapist (2)
J4	F	31	Speech therapist (3)
J5	F	37	Speech therapist (12)
J6	F	44	Speech therapist (4)
J7	F	45	Speech therapist (20)
J8	F	60	Speech therapist (37)
J9	M	29	Otorhinolaryngologist specializing in voice disorders (5)
J10	F	51	Otorhinolaryngologist specializing in voice disorders (27)

PROOF ONLY

Table II. Strength of the agreement according to the Kappa statistic (24).

Kappa statistic	Strength of agreement
<0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

each task, the averages of the responses by the 10 judges were calculated for the two possible choices, then transformed into percentages. This mean (as a percentage) was then compared to a standard of 50%. Our aim was to determine whether there was a difference between the judges' mean responses and an identical distribution between the two possible choices (i.e. the chance level of 50% for each choice). The comparison of the mean to a standard tests the null hypothesis that the mean equals 50%. The significance level was set at $P<0.05$.

Results

Reliability

As can be seen in Table III, results of Cohen's kappa indicated that the agreement between the responses given by the judges at test and retest was poor to fair. Concerning interrater reliability, results indicated poor agreement for task 1 (Fleiss's kappa $=-0.028$) and fair agreement for task 2 (Fleiss's kappa $=0.040$) and task 3 (Fleiss's kappa $=0.043$).

Effect of vocal load duration

For task 1, 55.6% of voice samples were judged to be breathier PRE LI session than POST LI session (Figure 1). The null hypothesis tested is that PRE LI session breathiness equals 50%. The rejection of this hypothesis ($P=0.006$) means that voices were significantly breathier PRE than POST LI session.

For task 2, 58.4% of the voice samples were judged to be breathier PRE than POST HI session (Figure 1). The null hypothesis tested is that PRE HI session breathiness equals 50%. The rejection of this hypothesis ($P=0.002$) means that voices were significantly breathier PRE than POST HI session.

Effect of vocal load intensity

For task 3, 52.2% of voice samples were judged to be breathier POST LI session than POST HI session (Figure 2). The null hypothesis tested is that breathiness POST LI session equals 50%. The acceptance of this hypothesis ($P=0.411$) means that there was no significant difference in breathiness between voices POST LI session and POST HI session.

Discussion

Methodological aspects

In this study, we have evaluated perceptually the effects of vocal loading on breathiness, as a complement to the objective measurements and self-ratings reported in a previous study (4). Our study as well as previous ones of the perception of changes in voice quality following vocal loading relied on expert judges, either students at the end of their training or voice professionals (3,12–15). We turned to expert judges because of the difficulty of the task. Indeed, the differences between the test samples have often been small. Consequently, most judges reported that the task was difficult and tiring.

As in our study, Stone and Sharf used pairwise comparisons PRE and POST vocal loading (3). They asked the judges to calculate the amplitude of the change observed between two samples compared with a 7-interval scale. In our study, we asked judges to choose the breathiest of two samples played, without any possibility of grading the perceived difference. We did not give judges the possibility of saying that the aspect in question was identical in the two samples played. The disadvantage of this method is that the judges were forced to choose an answer, even if they did not hear breathiness in the stimuli presented. Nevertheless, the forced choice had the aim of pushing judges to examine samples as carefully as possible before answering.

Reliability

For all three tasks, Cohen's kappa indicated poor to fair agreement between the test and retest. There was little difference between judges. As in previous studies, there did not appear to be any correlation between a judge's reliability and his or

Table III. Values of Cohen's kappa.

	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10
Task 1	0.000	0.040	−0.040	0.000	0.240	0.400	0.000	−0.240	0.160	0.040
Task 2	0.000	−0.120	0.040	0.000	0.080	0.240	−0.120	0.080	0.040	0.080
Task 3	0.040	−0.160	0.040	0.080	0.240	0.320	−0.240	−0.200	−0.160	−0.080

PROOF ONLY

mono for print colour online

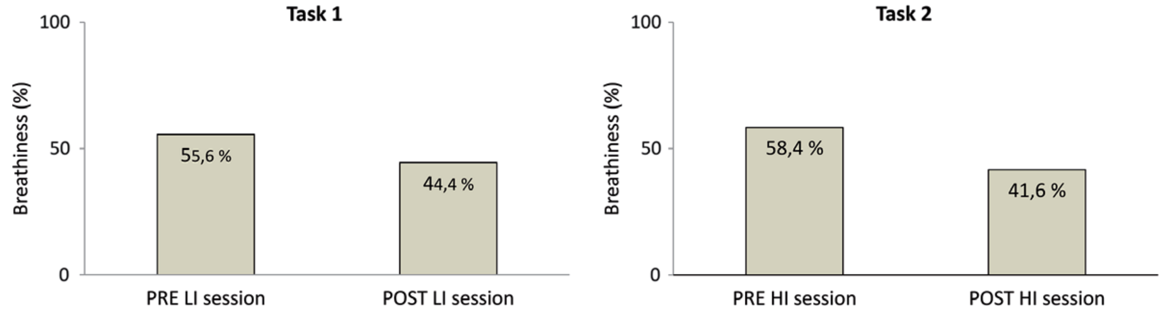

Figure 1. Effect of vocal load duration on breathiness.

her professional experience (8,31,32). Although the intrarater agreement was low, we retained the responses of all judges in our analyses of the results. In fact, the lack of difference in intrarater reliability levels meant we could not identify any particular judges as being clearly less reliable. Moreover, the judges appeared to show a degree of reliability that varied according to task. Regarding interrater reliability, tasks 1 revealed poor agreement, whereas agreement in tasks 2 and 3 was fair.

Low intra- and interrater reliability is a well-known problem, inherent in perceptual judgment. Many studies have attempted to overcome these difficulties by making use of different kinds of judges (31), different scales (33), or different phonetic materials (34). Anchor points or learning protocols have also been used to try to improve reliability (35,36). However, there is still no consensus regarding the ideal perceptual analysis method. Like other methods, pairwise comparisons have limitations related to reliability. In our study, the lack of reliability may be related to the task design, which did not allow judges to say that the two samples to be compared were similar or that the aspect being assessed did not exist in either sample. For each pair, the judges had to answer the question 'which voice is breathier?' If the two samples were identically breathy, or if neither sample was breathy, then the judges may have chosen their answers by chance, resulting in randomness among the judges as well as between test and retest. Thus, the restrictive response possibilities may explain the low reliability levels.

mono for print colour online

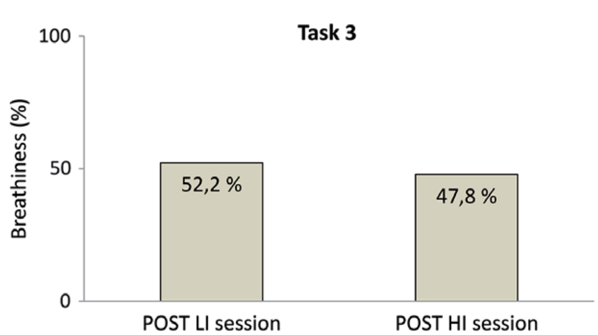

Figure 2. Effect of vocal load intensity on breathiness.

Another possible explanation is that the judges were basing their judgments of a particular aspect on different dimensions. After the perceptual judgment tasks, we asked the judges about the strategies underlying their judgments. In addition to breathiness due to glottal leakage, the judges reported basing their assessments on asthenia, and harmonic richness. Some of the judges said they had based their responses on the general quality of the voice in some cases, since breathiness was not very obvious. Although the term *breathy* was defined before the tasks, it appears that the judges used a variety of criteria in making their judgments. Moreover, as Kent noted, the discriminable differences for a stimulus are not necessarily *isomorphic* (5). A given judge probably did not apply one single criterion to all the sample pairs, and if several criteria were used, it is impossible to know how much weight the judge attributed to each one. These factors reflect the subjective nature of perceptual analysis. Kreiman and Gerratt suggest that judges are incapable of being consistent in their judgments of specific voice characteristics because it is difficult to isolate the individual dimensions of complex signals (7). These authors also question the value of approaches based on a one-dimensional scale for perceptual evaluation.

Finally, one last explanation of the lack of reliability is that the differences between the comparison samples were really minimal. The smaller the differences between samples to be compared, the more difficult the task is for the judges; the consequence is reduced reliability.

Impact of vocal load duration

The results showed that breathiness was significantly lower POST vocal loading, in both task 1 (LI session) and task 2 (HI session). This suggests that an improvement was perceived following 2 hours of vocal loading. This decline in breathiness may reflect 1) a decrease in laryngeal air leakage, contrary to studies describing the increase of glottal leakage after vocal loading (22–24); 2) a decrease in asthenia, which is associated with breathiness by some judges;

PROOF ONLY

or 3) a change in harmonic richness, characterized by a more brilliant timbre and a less dull voice. At the end of the task, some judges stated that they had based themselves on timbre or on a 'lack of brilliance' because they did not perceive any breathiness due to glottal air leakage.

This decrease in breathiness POST vocal loading suggests that speakers' voices improved and is comparable to the results of Sherman and Jensen (12), who observed a reduction in harshness during one-and-a-half hours of reading in men with normal voices. Several speakers in Sherman and Jensen's study reported that they had thought they might be unable to complete the task, or at the least found it difficult, due to the increase in vocal effort experienced during the first 30 minutes of reading. After that, though, they felt an improvement in their vocal performance, as if they could continue to read indefinitely. Adaptation of voice to loading is one possible interpretation of the improved harshness reported by Sherman and Jensen (12), and of the improvement in breathiness in our study. In fact, participants knew that they would have to read for a long time (one-and-a-half hours in Sherman and Jensen's study; 2 hours in the present study). It is possible that muscular, respiratory, and resonance adjustments were made to deal with vocal demand and ensure vocal effectiveness throughout the task. The hypothesis that subjects adapt to vocal loading is supported by the improvement in certain objective cues such as an increase of the maximum phonation time, a decrease of shimmer and a tendency of jitter to decrease during the 2 hours of reading, as reported in our earlier study (4). Previous studies have reported a correlation between breathiness and shimmer, as well as between breathiness and jitter (37–39). The decrease in breathiness, shimmer, and jitter may reflect an improvement in voice quality. Finally, interpreting our results as showing adaptation seems plausible given that we observed the effects of vocal loading in women with normal voices, who had never reported any voice problems.

Impact of vocal load intensity

According to the literature, vocal behavior becomes hyperfunctional and the voice is perceived less breathy when voice intensity increases (27,40). Contrary to our expectations, the results did not show any significant difference in breathiness between voices POST LI session and POST HI session (task 3). The lack of an intensity effect on perceptual analysis suggests that high-intensity vocal load does not entail a less breathy voice. The management of vocal load intensity therefore depends more on control of the respiratory muscles and effective use of resonators than on a strategy involving a modification of the glottal resistance. Thus, the vocal behavior our speakers engaged in seems to be appropriate and effective. Similarly, Stone and Sharf (3) did not find significant changes as a function of intensity in men with normal voices after vowel repetition for 20 minutes, at three different intensity levels. Neils and Yairi (13) showed that intensity had no effect on women with normal voices who read for 45 minutes in three different background noise levels, involving different voice intensity levels.

Conclusion

In this study, perceptual analysis was used to determine the impact of duration and intensity of vocal loading on breathiness. The voices of 50 female speakers were perceptually analyzed by 10 expert judges before and after 2 hours of reading at LI level, and before and after 2 hours of reading at HI level. A pairwise comparison method was used to reduce the subjectivity inherent in perceptual judgments, with the aim of avoiding the variability related to comparison of stimuli with judges' internal standards and increasing the reliability of judgments. Despite these efforts, intra- and interrater reliability ranged from poor to fair. This low reliability may have been caused by the restrictive response possibilities the judges were given, the fact that the judges may rely on different perceptual indices to judge a single voice quality, and the fact that the perceptual differences between samples were minimal.

Regarding the effect of duration of vocal loading, voices were significantly less breathy after 2 hours of reading, in both the LI and HI sessions. The perceived improvement in breathiness can be interpreted as an adaptation of voice to loading. It is possible that muscular, respiratory, and resonance adjustments were instituted to cope with vocal demand and ensure effectiveness throughout the task. Finally, no effect of vocal load intensity was observed on breathiness when comparing prolonged reading tasks at 60–65 dB versus 70–75 dB. This study confirmed that perceptual evaluation of the vocal load effects remains challenging, due to the subjectivity of the method itself and the relatively small differences between the samples to be judged.

Acknowledgements

We are extremely grateful to the speakers who took part in the vocal load task and the judges who conducted the perceptual analysis. We also thank Professor Etienne Quertemont for his advice regarding the statistical analyses.

PROOF ONLY

Declaration of interest: The authors report no conflicts of interest.

References

1. Morrow SL, Connor NP. Voice amplification as a means of reducing vocal load for elementary music teachers. J Voice. 2011;25:441–6.
2. Chen SH, Chiang S-C, Chung Y-M, Hsiao L-C, Hsiao T-Y. Risk factors and effects of voice problems for teachers. J Voice. 2010;24:183–90.
3. Stone R, Scharf D. Vocal change associated with the use of atypical pitch and intensity levels. Folia Phoniatr Logop. 1973;25:91–103.
4. Remacle A, Finck C, Roche A, Morsomme D. Vocal impact of a prolonged reading task at two intensity levels: objective measurements and subjective self-ratings. J Voice. 2012;26: 177–86.
5. Kent RD. Hearing and believing: some limits to the auditory-perceptual assessment of speech and voice disorders. Am J Speech Lang Pathol. 1996;5:7–23.
6. Bele IV. Reliability in perceptual analysis of voice quality. J Voice. 2005;19:555–73.
7. Kreiman J, Gerratt BR. Sources of listener disagreement in voice quality assessment. J Acoust Soc Am. 2000;108: 1867–76.
8. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. J Speech Hear Res. 1993;36:21–40.
9. Kreiman J, Gerratt BR, Precoda K, Berke GS. Individual differences in voice quality perception. J Speech Hear Res. 1992;35:512–20.
10. Schoentgen J, Fraj S, Lucero J. Testing the reliability of grade, roughness and breathiness scores by means of synthetic speech stimuli. Logoped Phoniatr Vocol. 2013 Oct 11. [Epub ahead of print].
11. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. J Voice. 2006;20:527–44.
12. Sherman D, Jensen PJ. Harshness and oral-reading time. J Speech Hear Disord. 1962;27:172–7.
13. Neils LR, Yairi E. Effects of speaking in noise on vocal fatigue and vocal recovery. Folia Phoniatr (Basel). 1987;39:104–12.
14. Yiu EM, Chan RM. Effect of hydration and vocal rest on the vocal fatigue in amateur karaoke singers. J Voice. 2003; 17:216–27.
15. McAllister AM, Granqvist S, Sjolander P, Sundberg J. Child voice and noise: a pilot study of noise in day cares and the effects on 10 children's voice quality according to perceptual evaluation. J Voice. 2009;23:587–93.
16. Welham NV, Maclagan MA. Vocal fatigue: current knowledge and future directions. J Voice. 2003;17:21–30.
17. Hirano M. Clinical examination of voice. New York: Springer; 1981.
18. Titze IR. Principles of voice production. 2nd ed. Iowa City: National Center for Voice and Speech; 2000.
19. Lauri ER, Alku P, Vilkman E, Sala E, Sihvo M. Effects of prolonged oral reading on time-based glottal flow waveform parameters with special reference to gender differences. Folia Phoniatr Logop. 1997;49:234–46.
20. Vilkman E, Lauri ER, Alku P, Sala E, Sihvo M. Effects of prolonged oral reading on F0, SPL, subglottal pressure and amplitude characteristics of glottal flow waveforms. J Voice. 1999;13:303–12.
21. Vintturi J, Alku P, Lauri ER, Sala E, Sihvo M, Vilkman I. Objective analysis of vocal warm-up with special reference to ergonomic factors. J Voice. 2001;15:36–53.
22. Gelfer MP, Andrews ML, Schmidt CP. Documenting laryngeal change following prolonged loud reading: a videostroboscopic study. J Voice. 1996;10:368–77.
23. Solomon NP, DiMattia MS. Effects of a vocally fatiguing task and systemic hydration on phonation threshold pressure. J Voice. 2000;14:341–62.
24. Stemple JC, Stanley J, Lee L. Objective measures of voice production in normal subjects following prolonged voice use. J Voice. 1995;9:127–33.
25. Sodersten M, Lindestad PA. Glottal closure and perceived breathiness during phonation in normally speaking subjects. J Speech Hear Res. 1990;33:601–11.
26. Sodersten M, Hammarberg B. Effects of voice training in normal-speaking women: videostroboscopic, perceptual, and acoustic characteristics. Scand J Logop Phoniatr. 1993;18:33–42.
27. Sodersten M, Hertegard S, Hammarberg B. Glottal closure, transglottal airflow, and voice quality in healthy middle-aged women. J Voice. 1995;9:182–97.
28. Teston B. L'évaluation instrumentale des dysphonies: État actuel et perspectives d'évolution. In: Giovanni A, editor. Le bilan d'une dysphonie: État actuel et perspectives. Marseille, France: Solal; 2004. p. 105–69.
29. Kacha A, Grenez F, Schoentgen J. Voice quality assessment by means of comparative judgments of speech tokens. Proceedings of the European conference on speech communication and technology, Interspeech, Lisbon (Portugal), 2005. pp. 1733–1736.
30. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.
31. De Bodt MS, Wuyts FL, Van de Heyning PH, Croux C. Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. J Voice. 1997;11:74–80.
32. Fraj S, Schoentgen J, Grenez F. The reliability of perceptual scores of grade, roughness and breathiness assigned to disordered voices does not depend on the number of years of professional experience of the raters. Vocologie: Stem en Stemstoornissen. 2011;4:81–6.
33. Wuyts FL, De Bodt MS, Van de Heyning PH. Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia. J Voice. 1999;13:508–17.
34. Revis J, Giovanni A, Wuyts F, Triglia J. Comparison of different voice samples for perceptual analysis. Folia Phoniatr Logop. 1999;51:108–16.
35. Chan KM, Yiu EM. The effect of anchors and training on the reliability of perceptual voice evaluation. J Speech Lang Hear Res. 2002;45:111–26.
36. Ghio A, Dufour S, Rouaze M, Bokanowski V, Pouchoulin G, Révis J, et al. [Perceptual assessment of dysphonia: a training protocol with natural speech]. Rev Laryngol Otol Rhinol (Bord). 2011;132:1–9.
37. Eskenazi L, Childers DG, Hicks DM. Acoustic correlates of vocal quality. J Speech Lang Hear Res. 1990;33:298–306.
38. Wolfe V, Martin D. Acoustic correlates of dysphonia: type and severity. J Commun Disord. 1997;30:403–15.
39. Wolfe V, Fitch J, Martin D. Acoustic measures of dysphonic severity across and within voice types. Folia Phoniatr Logop. 1997;49:292–9.
40. Sodersten M, Ternstrom S, Bohman M. Loud speech in realistic environmental noise: phonetogram data, perceptual voice quality, subjective ratings, and gender differences in healthy speakers. J Voice. 2005;19:29–46.

PROOF ONLY