Fatigue-Indicator in Operational Settings: Vocal Changes

Heike Diepeveen[1,2], Maykel van Miltenburg[2(✉)], Alwin van Drongelen[2], Floris van den Oever[2], and Henk van Dijk[2]

[1] Faculty of Social and Behavioural Sciences, Utrecht University, 3584 CS Utrecht, The Netherlands
[2] Royal Netherlands Aerospace Centre - NLR, 1059 CM Amsterdam, The Netherlands
maykel.van.miltenburg@nlr.nl

Abstract. Fatigue is an important factor in aviation accidents and incidents. Since fatigue cannot always be prevented, it needs to be detected in real time so that countermeasures can be taken. This study researches whether vocal changes (in vocal intensity and fundamental frequency) can be used as a measure for fatigue in an operational aviation setting. Sixteen participants were measured two times. Before the first test moment, they were asked to sleep eight hours or more and before the second test moment six hours or less. During each test moment, they performed a PVT, filled in the KSS, and did two speech tasks. One task was aimed at free speech and one task was aimed at procedural speech. Pre-processing included segmentation of the speech into words and extracting fundamental frequency (f0) and intensity values. An overall mean of both variables was calculated for both free and procedural speech. Speech, PVT reaction time, PVT lapses and KSS scores were analyzed in SPSS using Paired Samples t-Tests and Wilcoxon Signed Ranks Tests. Participants slept significantly less during the night before the second test moment and scored significantly higher on the KSS. For the PVT, no differences in both reaction time and lapses were found. No significant differences in average f0 and intensity for both free and procedural speech were found either. The results did not show a significant relationship between fundamental frequency, intensity and fatigue. Further research is needed to examine if vocal changes can be used as a reliable fatigue measure.

Keywords: Fatigue · Fundamental frequency · F0 · Vocal intensity · Fatigue detection · Real-time · Non-invasive · Operational setting · Aviation

1 Introduction

Fatigue, defined as a state of performance impairment as a result of sleep loss, extended wakefulness, circadian phase and workload [1, 2], plays an important role in aviation accidents and incidents [3]. Attention, a fast reaction time, decision making and memory retention are critical for aviation personnel to perform their job in a safe and effective manner [4]. Fatigue significantly impairs these types of cognitive and executive functions. A fatigued individual can experience difficulty suppressing task-irrelevant stimuli,

N. L. Black et al. (Eds.): IEA 2021, LNNS 221, pp. 128–135, 2021.
https://doi.org/10.1007/978-3-030-74608-7_17

a reduced ability to correct behavior after a mistake, impaired vigilance, and a slower reaction time [4–7].

To prevent accidents it is important to be able to detect fatigue in real time, so that direct countermeasures can be taken. Examples of objective measures for fatigue include electroencephalogram (EEG) and the psychomotor vigilance task (PVT). These measures have proven themselves to be effective [8, 9], but are not practical to apply in an operational setting in their current form, due to their invasiveness and/or disruptiveness. On top of that, the PVT in particular does not provide a constant and real-time assessment of a person's fatigue level.

A subjective measure is the Karolinska Sleepiness Scale (KSS). This is a quick method to determine subjective sleepiness at a certain moment in time. Even though this measure is less invasive and disruptive, it does not provide a real-time representation of fatigue either.

A measure that has not been thoroughly researched yet, but that could circumvent the issues mentioned above, is the vocal change as a result of fatigue. In particular, changes in intensity and fundamental frequency (f0). In a number of papers, these concepts are referred to as volume and pitch, but since the current research focuses on the analysis of voice rather than the human perception of voice the terms intensity and f0 will be used. Intensity is defined as the average amplitude of the speech signals in Decibel. Fundamental frequency is defined as the rate of vocal fold vibration in Hertz [10, 11].

Previous studies found a reduction in both f0 and intensity when individuals were fatigued versus non-fatigued [1, 2, 12–14] This could indicate that voice is indeed a good measure for fatigue and could be used in an operational setting. However, because the studies were performed with small samples, the relationship between voice and fatigue needs to be studied further. In addition, it has been found that an increase in fundamental frequency is positively correlated with higher emotional stress [15]. This could mean that f0 may not reflect the level of fatigue when combined with a high stress work environment, and the results could give a distorted image.

The objective of this study was to find out if vocal intensity and f0 are good potential measures for fatigue in an operational setting. Previous studies focused on speech while driving a car [13, 14], speech of a pilot during a flight [2] and free speech measured by answering "would you rather" dilemmas [1]. The present study contributes to the current body of knowledge by integrating free and procedural speech, as seen in a normal work setting, and focusing specifically on procedural speech used in aviation. The effect of fatigue on f0 and vocal intensity in both procedural and free speech was studied. Based on the research described earlier [1, 2, 12–14], it was expected that fatigue decreases f0 and vocal intensity during both types of speech.

Due to the covid-19 pandemic, the experiment could not take place in a laboratory setting, as was originally planned. For this reason, the study was performed online, during which participants carried out tasks from home. This has not been done in previous studies, for which this study could give insight in the feasibility and effectivity of such a method as well.

2 Methods

A power analysis was performed to determine the minimum amount of participants necessary for a power of 0.8. For this, effect sizes found in the literature were used [13, 14]. An r-value of −.44 was found for intensity and a value of −.42 was found for fundamental frequency. These values were converted to d-values with the Eq. (1):

$$d = \frac{2\mathrm{r}}{\sqrt{1 - \mathrm{r}^2}} \tag{1}$$

Based on the analysis, a minimum of 12 participants was determined. A total of 16 participants were recruited through Sona Systems (an online researcher and participant platform used by universities), social media and snowball sampling. One participant was excluded from further analysis, because the audio recordings contained loud and consistent background noises, making a reliable analysis impossible. The resulting subject group consisted of 9 females and 6 males between the ages 19 and 55 (M = 28.73, SD = 9.94). The subjects were of Dutch, Filipino, German, Greek and Romanian descent.

The study had a total duration of six days with online measurements at two time points. The three nights before the first measurement, participants were instructed to sleep 8 h or more, while they had to track their sleep by means of a digital sleep diary. The two nights before the second measurement, participants were requested to sleep 6 h or less. This methodological choice was based on the assumption that 5 to 7 h of sleep per night, for a period of at least two days, can lead to cognitive impairments as a result of sleep debt [16].

In addition to the sleep protocol, participants were instructed to abstain from any alcohol and fatigue inducing medicine during the six days of the experiment and to avoid caffeine containing beverages two hours before sleep during the full six days of the experiment.

The measurement procedure was identical on both test days. The participants first performed a PVT on their mobile phone, followed by two speech tasks on their computer. Before, between and after the two speech tasks participants filled in the KSS questionnaire.

The PVT consisted of a 3-min simple reaction time task that was integrated in a mobile app called Psych Lab 101 [17]. The task comprised 75 trials in which a red square appeared, with an inter-stimulus interval range of 1000 to 2000 ms. The participants had to respond to this by tapping the screen as fast as possible. If the participant did not touch the screen before the next trial appeared, this was registered as a miss. Misses and reactions times above 500 ms were counted as lapses [18].

Two types of speech tasks were applied. The first speech task consisted of answering fifteen "would you rather" dilemmas (e.g. Would you rather lose the ability to read or the ability to speak?). The second task involved answering radio commands taken from the Multi-Attribute Task Battery II (MATB-II) [19] to which the participants had to respond verbally by saying: "*Roger, turning radio to frequency […]*". The task consisted of fifteen radio commands with the call sign NASA504 and ten radio commands with a different call sign. The participants were instructed to only answer commands with the call sign NASA504. The task-order was counterbalanced between participants, and the order did

not change within-subjects. The commands and dilemmas were different between test days to avoid that speech was influenced by the participants recognizing the dilemmas and commands, and therefore being able to answer them more fluently. Participants were instructed to make use of their own microphone and the standard Apple or Windows recorder on their computer.

The Karolinska Sleepiness Scale (KSS) [20] was used to measure the subjective level of sleepiness during the measurements. The KSS version used had a ten-point Likert scale [20], where 1 is defined as being "extremely alert" and 10 represents being "extremely sleepy, can't stay awake".

The pre-processing and analysis of vocal intensity and fundamental frequency were done using PRAAT [21], which is an open source software tool for speech analysis. Firstly, all speech was segmented into phrases and words. Outliers caused by background noise were excluded from segmentation. After this, f0 and intensity values were extracted for each word using a pitch range of 75–300 Hz for males and 75–500 Hz for females. These ranges were determined based on the speech range of the participant sample [11]. The intensity values were extracted manually by selecting each word segment, letting PRAAT calculate the value, and logging the value. The f0 values were extracted using a TextGrid script [22]. Finally, the average pitch and intensity of all the extracted values per speech sample was calculated [23], resulting in four values per participant for both the well-rested and sleep deprived condition: average f0 for commands, average f0 for dilemmas, average intensity for commands, and average intensity for dilemmas.

3 Results

The Paired Samples t-Test showed a significant difference between the average sleep duration before the first measurement and the second measurement ($t = 15.17, p < 0.01$). The subjects slept 8 h and 39 min (SD = 0.72) the days before the first measurement, and on average 5 h and 55 min (SD = 0.23) before the second measurement. Table 1 shows the mean (M) and standard deviations (SD) of the following measured variables.

A significant difference was found for the average KSS score ($t = -2.69, p < 0.05$) (measured before, between and after the two speech tasks) of the two measurement days. The Wilcoxon Signed Rank Test showed no significant difference in reaction time ($Z = -1.08, p = 0.28$) and the number of lapses ($Z = -0.89, p = 0.37$) between the test days.

No significant differences were found for the voice measures either. The Wilcoxon Signed Rank Test showed no significant difference in average f0 between the two conditions for both the radio commands ($Z = -0.23, p = 0.82$) and the dilemmas ($Z = -1.08, p = 0.28$). This is shown in Fig. 1a. No significant differences were found with the Paired Samples T-tests between the two conditions for the intensity of the commands ($t = -0.99, p = 0.34$) and dilemmas ($t = -1.43, p = 0.17$). This is shown in Fig. 1b. Based on the observed trend between the commands and dilemmas, post-hoc analyses were performed to study the differences between the two speech tasks during the test days. A Wilcoxon Signed Ranks Test showed a significant difference between the average f0 of the dilemmas and the commands task during the first test moment ($Z = -3.07, p < 0.01$), and during the second test moment ($Z = -3.41, p < 0.01$). In addition, a Paired Samples T-test showed a significant difference between the intensity of the dilemmas

and commands tasks during the first test moment ($t = 4.88$, $p < 0.01$), and the second test moment ($t = 4.70$, $p < 0.01$).

Table 1. Mean (M) and standard deviations (SD) of the measured variables.

Variable	$M_{test\ 1}$	$SD_{test\ 1}$	$M_{test\ 2}$	$SD_{test\ 2}$
KSS score	3.4	1.5	4.5	2.3
Reaction time (in ms)	259.6	32.9	253.7	25.5
Lapses	0.9	1.3	1.9	3.3
f0 commands (in Hz)	178.9	48.5	179.8	48.8
f0 dilemmas (in Hz)	161.1	37.0	159.3	38.8
Intensity commands (in dB)	59.8	9.7	60.7	8.9
Intensity dilemmas (in dB)	55.3	8.8	57.1	9.3

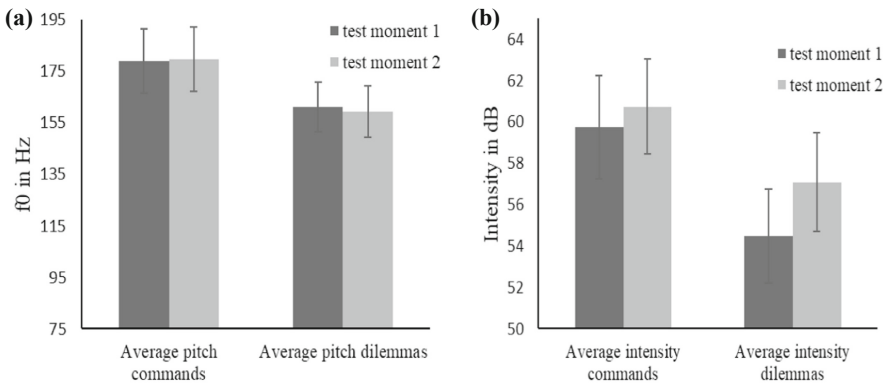

Fig. 1. (a) shows the mean pitch on both test moments for both the radio comments and dilemmas with error bars (SD). (b) shows the mean intensity on both test moments for both the radio commands and dilemmas, with error bars (SD).

4 Discussion

The purpose of this study was to explore whether changes in vocal intensity and fundamental frequency (f0) can be used as measures for fatigue detection in an operational setting. The outcomes of previous studies suggest that fatigue can lead to a reduction in both vocal intensity and f0 [1, 2, 12–14]. This relationship needed to be studied further, since only few studies have addressed this topic. The objective of the current study was to contribute to the available body of knowledge, while also including an operational element, namely different types of speech. Due to time constraints, the sample size was rather small, although the power analysis showed that a minimum of twelve participants

should be sufficient to find a significant difference on vocal changes. A distinction was made between free and procedural speech by means of two different speech tasks, and the vocal measures were compared using two fatigue conditions. In the first, well-rested condition, participants were asked to sleep 8 h or more. In the second condition, participants were asked to sleep 6 h or less during the two days before the test day. It was assumed that the latter led to sleep-deprivation and a subsequent impaired alertness during the second test day. This was confirmed by the findings, where the KSS ratings during the sleep-deprived condition were significantly higher. The analyses on the vocal measures however showed no significant difference in vocal intensity and f0 between the two test days for both speech tasks. We could therefore not conclude that fatigue can be detected by measuring vocal changes. These results might be explained by a number of factors.

Firstly, it could be that the participants were not fatigued to the extent that it affected vocal intensity and f0. Even though the participants slept significantly less and scored significantly higher on the KSS, the average scores represented a limited degree of fatigue, ranging between "rather alert" and "neither alert, nor sleepy". In comparison, another study looking at fatigue and vocal changes, found a mean KSS score of 7.47 (on a scale from 1 to 10) as a result of sleep deprivation, which is a degree of fatigue in between "sleepy, but no effort to stay awake" and "sleepy, but some effort to stay awake" [14]. It could therefore be that changes in voice measures only occur when sleepiness reaches these type of levels. On top of this, no significant differences were found for the PVT measures reaction time and lapses in the current study, which could indicate that the participants experienced little cognitive impairment as a result of sleep deprivation.

The task difficulty may have played a role as well. Two studies that did find vocal changes as a result of fatigue, used a more cognitively demanding task, namely driving a car [13, 14]. In the current study the participants only had to focus on one task at a time (listening to and answering radio commands and answering dilemmas), presumably resulting in lower workload. Higher perceived workload might lead to higher subjective ratings of fatigue. On the other hand, a higher workload might also lead to a higher level of perceived stress, which might have resulted in the vocal changes found in the previous studies.

Stress could also have influenced the results of the current study if participants felt more stressed during the second measurement, resulting in an increase in voice pitch [15], and a mitigating effect on the outcome measures. According to the literature, sleep deprivation leads to a higher sensitivity to emotional and stressful stimuli [24]. The speech tasks and other external factors (i.e. covid-19) might have been possible stressors for participants after a period of reduced sleep. Stress could also explain the differences found in f0 and intensity between the speech tasks. Moreover, many subjects mentioned that they experienced stress during the radio commands task when asked to answer as rapidly and accurately possible.

Another, more methodological factor that might have played a role is the participants distance to the microphone during the experiment. According to the guidelines for voice production research [25] the intensity level decreases by approximately 6 dB when the mouth-to-microphone distance is doubled. This may have slightly affected the average intensity values of the participants. In future studies, participants can be asked to play a

certain sound on both test days so a possible difference in loudness can be determined beforehand, and taken into account during the analysis.

Since the results of the current study do not confirm the findings of earlier studies, further research is needed. A better controlled, lab based study with more participants, would be preferable, since more objective control measures can be used in a lab setting (e.g. EEG) and there is more surveillance on how participants are performing the tasks. However, an online method as in the current study could also be used again, although a few factors should be considered. Firstly, a more severe sleep restriction should be applied in order to induce a higher level of sleepiness. A more effective protocol could be to ask participants to sleep 4 h for a consecutive period of 5 nights [16]. Sleep wearables or actigraphs could be used to objectively assess sleep duration. Secondly, possible confounding factors such as stress, should be measured (either objectively or subjectively) and controlled for. Furthermore, a task with a higher workload might be added to the measurement procedure (e.g. a difficult version of the MATB-II). Lastly, to control for microphone distance on both test days participants should be asked to hold their phone close to their mouth while playing a sound at maximum volume, while the laptop records the sound. This way a potential difference in sound intensity could be determined and controlled for.

Another possible study idea is to analyse voice recordings of pilots and air traffic controllers. A group of pilots and air traffic controllers can be asked to fill in the KSS questionnaire every hour during their normal duties. Afterwards, voice recordings during duty hours with low levels of fatigue could be compared to voice recordings during duty hours with high levels of fatigue. On top of that, other voice measures, such as disfluency, can be explored to determine if a combination of voice measures leads to a more accurate detection of fatigue.

5 Conclusion

In the current study no difference was found for vocal intensity and fundamental frequency (f0) between a well-rested and a sleep-restricted condition. The results of this study suggest that sleep restriction does not have an effect on f0 and vocal intensity. However, these findings do not have to imply that speech is an ineffective measure for fatigue detection altogether. Based on previous studies, it might still be possible that higher levels of fatigue do provoke an effect on voice measures. Since people working in the aviation sector often deal with a high workload, time zone shifts and night shifts, higher levels of fatigue are likely to occur. Additional research is needed to better understand the relationship between vocal changes and fatigue, taking into account the suggestions described earlier.

References

1. Roelen, A.L.C., Stuut, R.: Association of sleep deprivation with speech volume and pitch. In: Ergonomics & Human Factors 2016, Daventry, United Kingdom, pp. 19–21 (2016)
2. de Vasconcelos, C.A., Vieira, M.N., Kecklund, G., Yehia, H.C.: Speech analysis for fatigue and sleepiness detection of a pilot. Aerosp. Med. Hum. Perform. **90**(4), 415–418 (2019)

3. Caldwell, J.A.: Fatigue in aviation. Travel Med. Infect. Dis. **3**(2), 85–96 (2005)
4. Ruudin-Brown, C., Rosberg, A., Krukowski, D.: If we'd only listen! What can tell us about aircrew fatigue. In: 20th International Symposium on Aviation Psychology, pp. 319–324 (2019)
5. Boksem, M.A.S., Meijman, T.F., Lorist, M.M.: Effects of mental fatigue on attention: An ERP study. Cogn. Brain Res. **25**(1), 107–116 (2005)
6. van der Linden, D., Frese, M., Meijman, T.F.: Mental fatigue and the control of cognitive processes: effects on perseveration and planning. Acta Psychol. **113**(1), 45–65 (2003)
7. Lorist, M.M., Boksem, M.A.S., Ridderinkhof, K.R.: Impaired cognitive control and reduced cingulate activity during mental fatigue. Cogn. Brain Res. **24**(2), 199–205 (2005)
8. Lee, I., Bardwell, W.A., Ancoli-Israel, S., Dimsdale, J.E.: Number of lapses during the psychomotor vigilance task as an objective measure of fatigue. J. Clin. Sleep Med. **6**(2), 163–168 (2010)
9. Zhang, C., Yu, X.: Estimating mental fatigue based on electroencephalogram and heart rate variability. Pol. J. Med. Phys. Eng. **16**(2), 67–84 (2010)
10. Aalto University Wiki: Fundamental frequency (F0) - Introduction to Speech Processing, https://wiki.aalto.fi/display/ITSP/Introduction+to+Speech+Processing. Accessed 2019
11. Intro 4.2. Configuring the pitch contour. https://www.fon.hum.uva.nl/praat/manual/Intro_4_2__Configuring_the_pitch_contour.html. Accessed 2019
12. Baykaner, K., Huckvale, M., Whiteley, I., Ryumin, O., Andreeva, S.: The prediction of fatigue using speech as a biosignal. In: Dediu, A.H., Martín-Vide, C., Vicsi, K. (eds.) Statistical Language and Speech Processing. SLSP. Lecture Notes in Computer Science, vol. 9449, pp. 8–17. Springer, Cham (2015)
13. Krajewski, J., Batliner, A., Golz, M.: Acoustic sleepiness detection: framework and validation of a speech-adapted pattern recognition approach. Behav. Res. Methods **41**(3), 795–804 (2009)
14. Krajewski, J., Trutschel, U., Golz, M., Sommer, D., Edwards, D.: Estimating fatigue from predetermined speech samples transmitted by operator communication systems. In: Proceedings of the 5th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design : Driving Assessment 2009, pp. 468–474 (2009)
15. Ruiz, R., Legros, C., Guell, A.: Voice analysis to predict the psychological or physical state of a speaker. Aviat. Space Environ. Med. **61**, 266–271 (1990)
16. Alhola, P., Polo-Kantola, P.: Sleep deprivation: Impact on cognitive performance. Neuropsychiatric Dis. Treat. **3**(5), 553–567 (2007)
17. Psychlab 101 (Version 2.1.0) [Software]: Neurobehavioral Systems. https://www.neurobs.com. Accessed 2020
18. Anderson, C., Wales, A.W.J., Home, J.A.: PVT lapses differ according to eyes open, closed, or looking away. Sleep **33**(2), 197–204 (2010)
19. NASA: The Multi-Attribute Task Battery II (MATB-II) software for human performance and workload research: a user's guide (NASA/TM–2011–217164) (2011)
20. Shahid, A., Wilkinson, K., Marcu, S., Shapiro, C.M.: Karolinska sleepiness scale (KSS). In: STOP, THAT and One Hundred Other Sleep Scales, pp. 209–210. Springer (2012)
21. Boersma, P., Weenink, D.: Praat (Version 6.1.08): doing phonetics by computer [Computer program]. https://www.praat.org/. Accessed 2019
22. Script for analysing pitch with a TextGrid. (2014). https://www.fon.hum.uva.nl/praat/manual/Script_for_analysing_pitch_with_a_TextGrid.html
23. Microsoft Corporation: Microsoft Excel (2018)
24. Vandekerckhove, M., Cluydts, R.: The emotional brain and sleep: an intimate relationship. Sleep Med. Rev. **14**(4), 219–226 (2010)
25. Švec, J.G., Granqvist, S.: Guidelines for selecting microphones for human voice production research. Am. J. Speech-Lang. Pathol. **19**(4), 356–368 (2010)