"PREDICTION OF SLEEPINESS RATINGS FROM VOICE BY MAN AND MACHINE": THE ENDYMION REPLICATION PERCEPTUAL STUDY

Vincent P. Martin[1,2], Jean-Luc Rouas[1], Aymeric Ferron[1], Pierre Philip[2]

[1]LaBRI, Univ. Bordeaux, CNRS, Bordeaux INP, UMR 5800, F-33400 Talence, France
[2] SANPSY, Univ. Bordeaux, CNRS UMR 6033, F-33076 Bordeaux, France
Correspondence should be addressed to: vincent.martin@labri.fr and rouas@labri.fr

ABSTRACT

Following the release of the SLEEP corpus during the Interspeech 2019 paralinguistic continuous sleepiness estimation challenge, a paper presented at Interspeech 2020 by Huckvale et al. examined the reasons for the poor performance of the models proposed for this task. They conducted a perceptual experiment on a subset of this corpus that seems to indicate that human hearing is, however, able to estimate sleepiness in this corpus.

In this study, we present the results of the Endymion replication study, in which the same samples were rated by thirty French-speaking naive listeners. We then discuss the causes of the differences between the two studies and examine the effect of listener and sample characteristics on annotation performances.

Keywords: Sleepiness, Voice, Perceptual study, Experimental study replication, Paralinguistics

1. INTRODUCTION

Sleepiness is a public health problem that increases the risk of disability and mortality [1]. The significant imbalance between the amount of sleep specialists and the prevalence of sleepiness (up to 40% of the general population [2]) and the need for physicians to better follow up their patients between consultations has led them to adopt Ecological Momentary Assessment with which they have access to patients' symptoms very regularly in their usual living conditions, paving the way to personalized treatments and real-time relapse prevention [3]. A promising tool to do so is speech processing. Indeed, voice is associated with the physiological state of the speaker [4] and it is possible to implement voice measurements in passive situations without requiring the patient to perform a specific task (e.g., interacting with a connected device).

In this way, sleepiness detection in speech has been at the heart of two international challenges proposed in parallel with the 2011 and 2019 Interspeech conferences. During the 2019 challenge, the SLEEP corpus was introduced [5] with the task of estimating sleepiness (correlation between predictions and ground truth values). It contains 16,492 random samples from 915 German-speaking subjects, whose sleepiness levels are annotated with the truncated average of three Karolinska Sleepiness Scale [6, KSS]: one is filled in by the subjects themselves, while the other two are annotated by assistants using video and audio [7]. Contrary to the expectations of the challenge organizers, the proposed systems did not show much improvement from the baseline (ρ = .387 for the best system [8] versus .343 for the baseline). Even more recent work on this corpus using the latest deep learning techniques did not perform better [9, 10].

To investigate the causes of this glass ceiling, Huckvale et al. [11] conducted a perceptual study to test the suitability of the corpus for the proposed regression task. Based on the annotations by 26 British English listeners of 90 samples extracted from the SLEEP corpus, and using *Wisdom of the Crowd*, they achieved performances far beyond those ever achieved by the systems proposed for sleepiness estimation tasks (r = .72). Thus, we claim that the study by Huckvale et al. supports that human hearing can estimate sleepiness from speech samples of the SLEEP corpus.

In this article, we propose to reproduce the perceptual study conducted in [11] (denoted as "original study") with naive French-speaking listeners to confirm or infirm this hypothesis. This paper is organized as follows. In Section 2, we introduce the methodology of our replication study on the SLEEP corpus. We present our results in Section 3 and discuss them in Section 4. Finally, we draw conclusions in Section 5.

2. METHOD

Thirty French-speaking listeners without hearing impairment were recruited by word-of-mouth.

Since we hypothesize that they can improve the perception of sleepiness in speech, we collected their understanding of German and their musical sensitivity [12]. All the characteristics of the listeners available in the original and replication studies are presented in Table 1. Using a KSS, they annotated the same 100 samples (10 for training) of the SLEEP corpus as in the original study. The order of the samples is the same in both studies, and the samples are shown and annotated one after the other (no browsing back).

The annotation tools used in each study are shown in Figure 1. While the version used in the original study combines at the same time a Lickert-like scale (gradual textual description) and a Visual Analog Scale (continuous line with two anchors), the scale used in our replication study is the standard Lickert scale as it has been presented to speakers.

Characteristic	**Huckvale et al.** $n = 26$	**Endymion** $n = 30$
Age	18-60	20-60
Sex	-	M: 17 F: 13
Impairments in hearing	None	None
Native language	English	French
German language level	German $\neq$ first language	"Not at all" ($n = 19$) "At least a little" ($n = 11$)
Specific Musical Sensibility	-	No ($n = 16$) Yes ($n = 14$)
Compensation	£5 ($n = 20$) attendance credits ($n = 6$)	None

Table 1: Listeners' characteristics.

3. RESULTS

The results of the same analysis as in the original study are reported in Table 2.

Z-normalized raw scores. First, the annotations are z-scaled per listener to eliminate their individual characteristics. The resulting distributions in both studies are shown in Figure 2 (left). The raw z-scaled annotations in the Endymion replication study resulted in a slightly better Person and Kendall correlations than in the original study, but these achievements are still insufficient to accept the hypothesis that human hearing is able to estimate sleepiness from speech samples extracted from the SLEEP corpus.

Wisdom of the crowd. In a second step, a Wisdom-of-the-Crowd (WoC) procedure is applied: for each sample, all z-scaled annotations are averaged, resulting in an average predicted score. The resulting distributions are shown in Figure 2 (right). Compared to the original study, applying WoC to listener annotations in the Endymion replication study brings a smaller gain in the correlation between estimated values and ground truth.

Metric	**Huckvale et al.**	**Endymion**
	Z-scaled annotations	
Correlation Kendall's coefficient	$r = .249$ $\tau = .117$	$r = .318$ $\tau = .23$
	WoC z-scaled annotations	
Correlation Friedman test UAR F1 SL/NSL	$r = .72$ 1|2,3,4,5,6,7|8,9 93.6% .87/.96	$r = .41$ 1|2,3,4|5,6,7|8,9 69.6% .51/.81
	Complementary results (no normalization)	
ICC2-10	.668	**.975**
Std/listener *mean (std)*	1.83 (.38)	2.34 (.21)

Table 2: Comparison metrics between the original study and our replication study. *WoC*: Wisdom of the Crowd, *SL*: Sleepy, *NSL*: Not Sleepy.

In order to determine underlying groups in the annotations, a Friedman test and the corresponding post-hoc analysis are calculated with the Python package `Pingouin` v.0.4.0 [13]. In the original study, the annotations are grouped into three sleepiness levels based on a Friedman test: a 'sleepy' group (KSS > 7), a 'normal' group ($2 \leq$ KSS ≤ 8) and an 'aroused' group (KSS $= 1$). The same analysis applied to our replication study suggests a division into four groups ($W = .263, ddof = 8, p = .007$, pairwise Mann-Whitney): a 'sleepy' and an 'aroused' groups (resp. KSS $>$ 7 and KSS $= 1$), and two 'slightly aroused' and 'slightly sleepy' subgroups (KSS $\in \{2,3,4\}$ and KSS $\in \{5,6,7\}$). Sleepiness subgroups for each study can be observed in Figure 2 (right).

To calculate binary classification performance, the ground-truth KSS is binarized into two classes: Sleepy (KSS > 7) and Not Sleepy (KSS ≤ 7). A threshold of .26 on the z-scaled WoC annotation gives a binary classification UAR (Unweighted Average Recall) of 93.6% in Huckvale et al., while in our replication study the best cutoff value of .31 yielded a UAR of only 69.6%, which is lower than the classification performances typically achieved on the task. To complement these results for further discussion, we also calculated the F1 value for each class: in both studies, the F1 values are better in the NSL class than in the SL class.

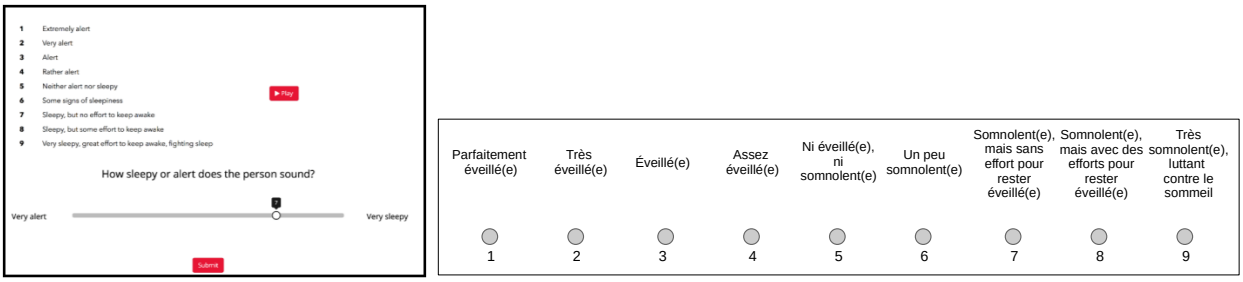

Figure 1: KSS annotation tool proposed in the original study (left), and our replication study (right).

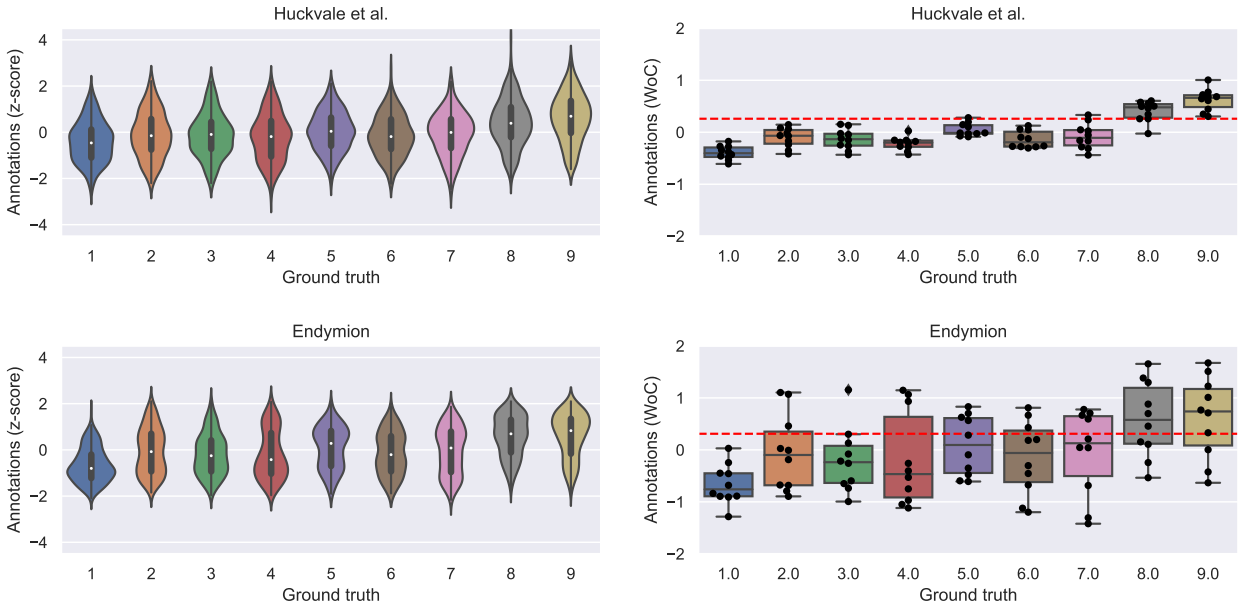

Figure 2: (Left) Violin plots of z-scaled annotations by ground-truth value, in the study of Huckvale et al. (top) and in the Endymion study (bottom). (Right) Box plot of the WoC z-scaled annotations depending on the ground-truth KSS. Each dot represents a sample, and the red dashed line represents the cut-off value giving the best UAR.

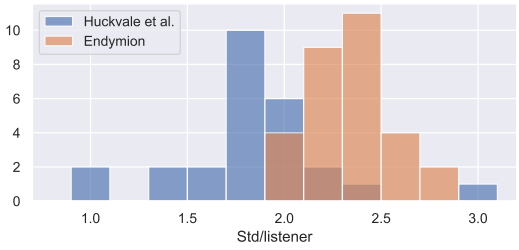

Figure 3: Distribution of standard deviation of annotations per listener (before z-normalization). The observed difference is statistically significant (t-test, $p < .0001$).

4. DISCUSSION

Differences between the studies To find the underlying cause of the differences between these analyzes, we calculate the intraclass correlation (ICC2-10) on the raw annotations before z-normalization for each study, which is an indicator of overall agreement between the annotators [14]. It shows that there is a lower agreement between annotators in the original study (ICC = .668) than in our replication study (ICC = .975). This could be the source of the small performance gain bought by WoC in the Endymion study: averaging already converging opinions over a sample yields much less information than averaging dissenting opinions. To account for the variety of levels used by listeners in each study, we also calculate the standard deviation per listener of their annotations before z-normalization (see Figure 3).

The listeners of the original study use significantly fewer different levels than those of the Endymion study. However, in the present study, listeners use a greater variety of levels, creating greater finesse in the annotations, albeit with less contrast (4 subgroups in the Endymion study vs. 3 subgroups in Huckvale et al.). These

different behaviors may find their source in the presentation of the annotation scales between the two studies: while in the present study the textual description is directly above the selected value, some listeners may have inadvertently used the annotation scale in Huckvale et al. as a simple visual analog scale, without referring to the text description at the top of the screen, creating their own rating scale [15].

What is the influence of the listeners' characteristics on their annotation performance? In the Endymion study, two listener characteristics of particular interest for the task were collected: (1) their musical sensitivity, since music practice or a hobby related to music might improve the perception of speech [12, 16]; (2) their understanding of the German language, as annotators who understand the language might have access to additional linguistic information. Therefore, for each annotator, we calculate the Mean Absolute Error (MAE) between the annotations (after z-scaling) and the corresponding ground truth values. Then, we min-max normalize them and compute Mann-Whitney tests with the aim of distinguishing the MAE between each group.

We find no significant difference for either of the two previous factors (resp. $p = .190$ and $p = .228$ for musical sensitivity and German language comprehension). Thus, if some listener characteristics influence the way they estimate sleepiness from speech samples, they are not captured by our study.

Which samples are the hardest to annotate? For a given sample, we hypothesize that two effects could explain the differences between annotations and ground truth. First, listener fatigue, who may not annotate the last samples as carefully as the first ones, for whom concentration may be easier (influence of the order of the samples). Second, the speaker's level of sleepiness, since human hearing might be able to detect some levels of sleepiness more easily than others (influence of the KSS). For each study and each sample, we computed the MAE between the ground truth values and the listeners' z-score annotation (MAE per sample). To make the ground-truth and z-scale annotation values comparable, we min-max normalize them so that their minimum value is 0 and their maximum value is 1. Then we calculated the correlation (Spearman's ρ) between the MAE per sample and their order and between the MAE per sample and the ground-truth KSS. The results are presented in Table 3.

The MAE does not correlate with the sample index in either study, which excludes the hypothesis of listener fatigue. However, the per-sample MAE correlates weakly with the KSS ground truth of the original study and more strongly with the Endymion study: the higher the KSS, the larger the errors between annotations and ground truth. We have put forward two hypotheses about this result. First, the human auditory system may be more sensitive to vocal expressions of alertness than to sleepiness, which explains the increase in MAE with KSS. Coming back to the classification results from Section 3, the F1 values are better in the NSL class than in the SL class, which supports this hypothesis. Second, we cannot exclude the hypothesis that some speakers could have completed a KSS indicating a high level of sleepiness at the time of their evaluation, but were then stimulated by the various recording tasks. Therefore, sleepy subjects may make (involuntary) efforts to compensate for their sleepiness in order to complete the task, creating a difference between their self-reported level of sleepiness, assessed before the task, and the expression of their sleepiness in their voice.

Factor	Huckvale et al.	Endymion
Order	$\rho = .09, p = .39$	$\rho = -.13, p = .24$
KSS	$\rho = .20, p = .05$	$\rho = .40, p < 10^{-3}$

Table 3: Correlation between the MAE per sample and their order and ground truth KSS.

5. CONCLUSION

To conclude, our replication study did not provide results as conclusive as the previous study conducted by Huckvale et al. [11]. Therefore, we cast doubt on the assumption that the human ear is capable of correctly assessing sleepiness from speech samples extracted from the SLEEP corpus. Regarding the factors influencing the annotation of the samples, we did not identify any influence of the annotators' characteristics. On the other hand, we found a link between the level of sleepiness of the speakers and the quality of sleepiness annotation in these two studies, with listeners having more difficulty in estimating very sleepy speakers.

6. ACKNOWLEDGEMENTS

We are deeply indebted to all the participants who spent time participating in the Endymion study. We gratefully thank Pr. Jarek Krajewski for having given us access to the SLEEP corpus, and Pr. Mark Huckvale for having given us the necessary material for the replication of his study.

7. REFERENCES

[1] A. J. Scott, T. L. Webb, M. Martyn-St James, G. Rowse, and S. Weich, "Improving sleep quality leads to better mental health: A meta-analysis of randomised controlled trials," *Sleep Medicine Reviews*, vol. 60, p. 101556, 2021.

[2] T. B. Young, "Epidemiology of daytime sleepiness: definitions, symptomatology, and prevalence," *The Journal of Clinical Psychiatry*, vol. 65 Suppl 16, pp. 12–16, 2004.

[3] S. Shiffman, A. A. Stone, and M. R. Hufford, "Ecological Momentary Assessment," *Annual Review of Clinical Psychology*, vol. 4, no. 1, pp. 1–32, 2008. [Online]. Available: https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

[4] G. Fagherazzi, A. Fischer, M. Ismael, and V. Despotovic, "Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice," *Digital Biomarkers*, pp. 78–88, Apr. 2021.

[5] B. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychocz, R. Vollman, S.-D. Roelen, S. Schnieder, E. Bergelson, A. Cristia, A. Seidl, A. Warlaumont, L. Yankowitz, E. Nöth, S. Amiriparian, S. Hantke, and M. Schmitt, "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity," in *Interspeech 2019*, 2019.

[6] T. Åkerstedt and M. Gillberg, "Subjective and objective sleepiness in the active individual." *Int J Neurosci*, vol. 52, pp. 29–37, 1990.

[7] F. Hönig, A. Batliner, E. Nöth, S. Schnieder, and J. Krajewski, "Acoustic-Prosodic Characteristics of Sleepy Speech - Between Performance and Interpretation," in *Speech Prosody 2014*, 2014, pp. 864–868.

[8] G. Gosztolya, "Using Fisher Vector and Bag-of-Audio-Words Representations to Identify Styrian Dialects, Sleepiness, Baby & Orca Sounds," in *Interspeech 2019*, 2019, pp. 2413–2417.

[9] J. V. Egas-López, R. Busa-Fekete, and G. Gosztolya, "On the Use of Ensemble X-Vector Embeddings for Improved Sleepiness Detection," in *Speech and Computer*, ser. Lecture Notes in Computer Science, S. R. M. Prasanna, A. Karpov, K. Samudravijaya, and S. S. Agrawal, Eds. Cham: Springer International Publishing, 2022, pp. 178–187.

[10] E. L. Campbell, L. Docio-Fernandez, C. Garcia-mateo, A. Wittenborn, J. Krajewski, and N. Cummins, "Automatic detection of short-term sleepiness state. Sequence-to-Sequence modelling with global attention mechanism." in *Workshop on Speech, Music and Mind*, 2022.

[11] M. Huckvale, A. Beke, and M. Ikushima, "Prediction of Sleepiness Ratings from Voice by Man and Machine," in *Interspeech 2020*, 2020.

[12] S. S. Asaridou and J. M. McQueen, "Speech and music shape the listening brain: evidence for shared domain-general mechanisms," *Frontiers in Psychology*, vol. 4, 2013.

[13] R. Vallat, "Pingouin: statistics in Python," *Journal of Open Source Software*, vol. 3, no. 31, p. 1026, 2018.

[14] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, Jun. 2016.

[15] C. C. Preston and A. M. Colman, "Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences," *Acta Psychologica*, vol. 104, no. 1, pp. 1–15, Mar. 2000.

[16] W. F. Thompson, E. G. Schellenberg, and G. Husain, "Decoding speech prosody: Do music lessons help?" *Emotion*, vol. 4, no. 1, pp. 46–64, 2004.