PHYSIOLOGICAL VS. SUBJECTIVE SLEEPINESS: WHAT CAN HUMAN HEARING ESTIMATE BETTER?

Vincent P. Martin[1,2], Aymeric Ferron[1], Jean-Luc Rouas[1],Takaaki Shochi[1,3], Lucile Dupuy[4], 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
[3]CLLE, Université de Toulouse, CNRS UMR 5263, F-31058, Toulouse, France
[4]Bordeaux Population Health, Univ. Bordeaux, INSERM U1219, F-33076, Bordeaux, France
Correspondence should adressed to: vincent.martin@labri.fr and rouas@labri.fr

ABSTRACT

This article investigates the perception of vocal manifestations of excessive sleepiness. Although previous efforts have demonstrated that naive listeners are able to estimate behavioral sleepiness, we aim to assess in this study this ability on both subjective (medical questionnaires) and physiological (polysomnography) measurements of sleepiness. We asked 71 naive French-speaking listeners to annotate a subset of the Multiple Sleep Latency Test corpus, each listener participating in two annotation sessions, using one among three annotation tools, with two different annotation paradigms (with or without reference). Based on these data, we then evaluated their ability to correctly annotate subjective or physiological sleepiness depending on the annotation tool or the paradigm of the test they undertook. We also measured the interaction between their performance and their characteristics, as well as the correlation between the listeners' performances and speakers' characteristics.

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

1. INTRODUCTION

Voice analysis is a promising tool for measuring sleepiness-related symptoms frequently and under normal patient living conditions [1]. Sleepiness is a peculiar symptom of interest in neuropsychiatric diseases that has a very high prevalence in the general population (up to 40% of the general population [2]) and induces negative consequences in both personal and public health, increasing the risk of disability and mortality [3].

Hence, automatic detection of sleepiness using speech samples has been the focus of two international challenges, proposed in parallel to the 2011 and 2019 Interspeech conferences. During the last challenge on the SLEEP corpus, the proposed systems did not lead to significant improvements in performance, the winner achieving a correlation coefficient between estimated and ground-truth values of sleepiness levels of $\rho = .387$ [4]. More recent propositions using the latest machine learning technologies have not reached much better performances [5, 6, 7].

To investigate the feasibility of the task, the team of Huckvale *et al.* conducted a perceptual study based on the SLEEP corpus [8]. Using 90 samples of the corpus, annotated by 26 listeners, they obtained a correlation of $r = .72$ between annotations processed using Wisdom of the Crowd and ground-truth values. Thus, they concluded that human hearing can estimate sleepiness through voice samples and that the bottleneck in the automatic estimation performances relies on the corpus content. However, this work suffers from three major limitations: (1) In a study using the exact same data, we failed to replicate these conclusions using the annotations made by 30 French-speaking listeners [9]; (2) In the SLEEP corpus, the samples are very short (between 3 and 5 seconds), whereas the minimum duration to estimate sleepiness with computational methods seems to be around 20 seconds [10]; (3) The ground-truth sleepiness label in the SLEEP corpus has never been validated in Sleep Medicine and is more a behavioral than a subjective sleepiness measurement [11], since it relies on the annotation by investigators of the behavioral manifestation of sleepiness, including voice (label contamination) [10].

Thus, to investigate the ability of human hearing to estimate sleepiness annotated with medicine-validated tools, we use in this study the Multiple Sleep Latency Test corpus (MSLTc) [10, 12], which

has been recorded during a MSLT, a gold standard polysomnographic test in sleep medicine [11], and contains physiological (polysomnography) and subjective (medical questionnaire) sleepiness labels. A specificity of MSLTc compared to other corpora is the recorded population: sleep clinic patients affected by various forms of hypersomnia, for which physiological and subjective sleepiness is uncorrelated [13].

This paper presents the results of the Endymion study, which takes advantage of this particularity to investigate which of subjective or physiological sleepiness is more easily detected in voice by naive listeners. The paper is organized as follows. We first introduce the methodology of our perceptual experiment in Section 2. Then, we present and discuss our results in Section 3. Finally, we draw conclusions in Section 4.

2. METHOD

2.1. Perceptual experiment

We recruited 71 native French-speaking annotators who had no experience in sleepiness assessment and no hearing problems. Since they can change the perception of speech, the sex [14], the age [15] and the musical sensibility [16] (any hobby or job related to music in this study) of the listeners were collected, as reported in Table 1.

N	Values	71
Age (years)	20-25	30
	25-30	26
	>30	15
Sex	F	22
	M	49
Musical Sensibility	Sensible	30
	Not. Sensible	41

Table 1: Listeners' characteristics.

The samples used in our experiment are extracted from the Multiple Sleep Latency Test corpus (MSLTc) [10, 12]. This corpus contains the recordings of 93 patients admitted to the sleep medicine department of the Bordeaux University Hospital for the diagnosis and/or treatment of rare hypersomnia diseases. They undertook a Multiple Sleep Latency Test (MSLT), consisting of asking them to take 20-minute naps every two hours, from 9a.m. to 5p.m. Before each nap, the patients were recorded reading texts of approximately 250 words extracted from *Le Petit Prince* (Saint-Exupéry). In order to reduce the variability of annotations due to the large number of speakers in the corpus, we selected 20 patients (10M/10F) from the MSLTc so that their variations of physiological (sleep latency) and subjective (Karolinska Sleepiness Scale and Cartoon Faces Scale) sleepiness across the naps were maximal. Furthermore, since the MSLTc recordings are long (mean duration: 77s), we kept only the first 30 seconds of each audio file to avoid fatiguing the listeners. The samples are then normalized in amplitude to -3dB. This constitutes the Endymion subcorpus.

The perceptual test consisted of asking the listeners to annotate one of the following three sleepiness measurements on these samples: (1) the duration after which patients fall asleep after the reading task, assessed by polysomnography. This value, named *sleep latency*, is a physiological measurement of short-term sleep propensity [11] and ranges from 0 to 20 minutes; (2) the score of the patients on the Karolinska Sleepiness Scale (KSS) [17], a nine-level scale that asks them to rate their level of sleepiness "during the last 10 minutes", filled just after each voice recording; (3) their score on the Cartoon Faces Scale (CFS) [18], also filled just after each voice recording, ranges from 0 to 4. The listeners annotated the KSS and CFS using the corresponding questionnaire, while sleep latency is annotated using a 100-level slider whose end anchors are 'Perfectly awake' on the left, and 'Extremely sleepy' on the right. The three annotation tools are represented in Figure 1.

During their contribution to the experiment, listeners were asked to participate in two annotation paradigms. On the one hand, a *Random* paradigm, during which they blindly annotated 10 samples drawn pseudorandomly among the Endymion subcorpus. On the other hand, a *Baseline* paradigm, for which the listeners had access to a reference sample of the patients when they are awake (recorded before an iteration of the MSLTc during which they stayed awake). The annotators then annotated the remaining four samples of the same patient, corresponding to the other four iterations of the MSLT. In this paradigm, each listener was asked to annotate the samples of two patients from the study subcorpus. In both paradigms, the samples were annotated one after the other (no browsing back). The order of the two paradigms and the used annotation tool was randomized.

2.2. Data Analysis

First, we scaled the annotations per annotator to remove their specific annotation behavior (z score). Then, to facilitate the comparison between the

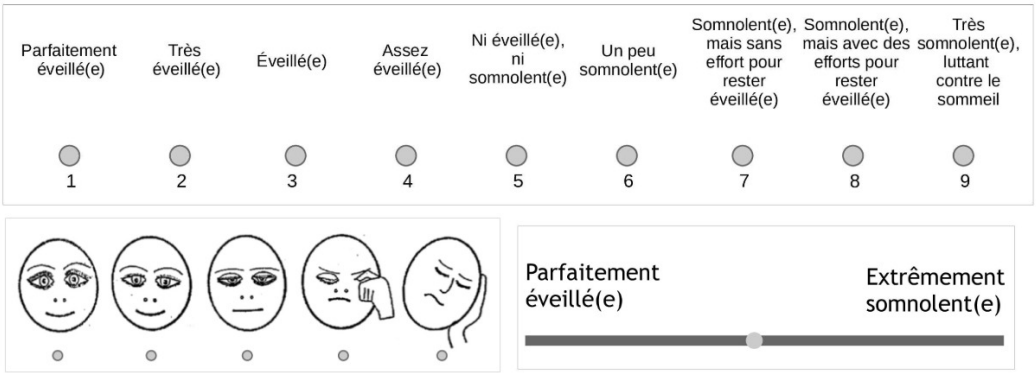

Figure 1: Annotation tool used in this perceptual experiment. (Top) Karolinska Sleepiness Scale. (Bottom left) Cartoon Faces Scale. (Bottom Right) Slider.

scales, we normalized the scores of each scale between 0 and 1. The same transformation is applied to the ground-truth values. To be compared with other scales, the annotated score on the slider and the corresponding ground-truth sleep latency are inverted, so that 0 corresponds to sleep latencies of 20 min (the patient did not fall asleep) and 1 corresponds to sleep latencies of 0 min (the patients felt asleep immediately). We thus obtain for each task a "sleepiness score" between 0 and 1 that can be compared between paradigms and annotation tools.

Since the number of points per annotator is too small to use correlation metrics [19], we measured annotation performance using the Mean Absolute Error (MAE), which is the arithmetic mean of the absolute error between annotation and ground truth across the samples. As a consequence, a lower MAE means fewer errors and thus better annotation performances.

3. RESULTS AND DISCUSSION

3.1. Paradigm and task

To identify the combination of paradigm and task favoring annotation performances, the MAE between the annotation of the listeners and the expected ground-truth values for each combination of task and paradigms are reported in Table 2.

The predominant effect affecting performances is the task: the best MAEs (.31-.28) are obtained annotating subjective sleepiness (KSS or CFS – with few differences between them) while the performances of annotating physiological sleepiness using the slider are distinctly worse (.47-.39). One hypothesis to explain this observation is that subjective sleepiness affects predominantly the acoustic quality of voice [20], while physiological sleepiness interfer with reading abilities [21, 22].

Paradigm	KSS	CFS	Sleep Latency
Random	**.30** (n=220)	.31 (n=250)	.40 (n=240)
Baseline	.29 (n=192)	**.28** (n=192)	.45 (n=184)
Both	**.29** (n=412)	.30 (n=442)	.46 (n=424)

Table 2: MAE between normalized annotation and ground truth for each paradigm and each annotation task.

Since the listeners only had access to short extracts (30s), we hypothesize that their annotation relies mainly on acoustic and prosodic information, hence detecting subjective sleepiness.

The second predominant effect is a small but notable difference between the *Random* and the *Baseline* paradigms, the latter leading to a better MAE (-1% of absolute MAE for the KSS, -3% for the CFS, -2% for the slider). We hypothesize that these differences are due to the reference audio file. Indeed, when picking a random audio sample, it is almost impossible to distinguish the sleepiness state of the speaker from all the other traits expressing through voice. By proposing such a reference, the listener may estimate accurately the sleepiness state of the patients independently from their traits, that are also expressing in the reference audio file in which they are not sleepy. Moreover, annotating the same speaker four times in a row in four different states could reinforce this distinction between the expression of the speaker's state and traits through voice.

3.2. Influence of listeners characteristics

In order to identify features of the listeners that could improve or interfere with their annotation capabilities, we tested (Mann-Whitney's U) on all tasks and paradigms whether the MAE per

listener before normalization was different between the following different categories: "first annotation session" vs. "second annotation session"; "M" vs. "F"; "Musical Sensitivity" vs. "No musical sensitivity". Moreover, regarding differences through age, we computed an univariate ANOVA across the three categories.

None of the factors taken into account led to significant differences in annotation performance in terms of MAE ($p > .05$): if some listeners' characteristics influence the way they annotate sleepiness from voice samples, they have not been captured by our study and would eventually require further investigation.

3.3. Influence of patients' characteristics

Factor	Paradigm	Task	MAE	
			ρ	p
Fatigue	Random	CFS	.51	.02
		All	.48	.04
Anxiety	Random	Slider	.56	.01
		All	.49	.03
Education level	Random	Slider	-.47	.04
	Baseline		-.57	.009
	All		-.57	.009

Table 3: Significant correlations between MAE per patient and patients' characteristics.

To measure the characteristics of the speakers that could influence the annotation of their level of sleepiness, we calculated the correlation (Spearman ρ) between MAE per speaker (before normalization) and the following factors: age, BMI, neck circumference, educational level (highest level of study after the French Certificate of general education), fatigue [23], Alertness [24], and Anxiety and Depression [25]. We also tested sex differences (Mann-Whitney's U). The significant correlations ($p < .05$) are reported in Table 3.

The speakers' fatigue level interferes with the annotation performance on the *Random* paradigm: when speakers report high levels of fatigue, listeners annotating their samples make more errors (i.e. MAEs are higher). This difference was not observed in the baseline paradigm. Similarly, the speaker's anxiety level interferes with the performance of listeners on the *Random* paradigm. Finally, the speakers' education level anticorrelates with the MAE using the slider: when the education level of the speaker increases, the errors made by the listeners when estimating their sleep latency using the slider decrease.

Fatigue and anxiety interfere with the correct annotation of sleepiness based on voice samples on the *Random* paradigm but not on the *Baseline* paradigm: we assume that the *Baseline* paradigm effectively allows listeners to annotate the speaker's sleepiness states independently of their trait features – such as fatigue or anxiety. An exception is the educational level of the speakers, which interferes with the annotation of physiological sleepiness on all the paradigms (*Random*, *Baseline*, and All). We hypothesize that in that task, physiological sleepiness has the same impact over voice as a lower educational level: the sleepy patients make more reading errors [21] and have different reading behaviors regarding the location and duration of the reading pauses [22].

4. CONCLUSION

This perceptual experiment is the first to investigate the ability of human hearing to differentiate between subjective and physiological sleepiness. The collected annotations show a better estimation of subjective sleepiness than physiological sleepiness by naive listeners. This observation leads us to hypothesize that the acoustic correlates of a subjective level of sleepiness are more numerous than those of its physiological counterpart, which is corroborated by the interference of education level with the estimation of physiological sleepiness only. Assessing physiological sleepiness using voice and speech may thus require additional information, i.e. features related to cognitive planning such as reading proficiency, length of pauses or sustained vowels. Moreover, the better performances obtained using the *Baseline* paradigm than using the *Random* one suggest that human earing estimate better variations of sleepiness states than its absolute level, opening new research paradigms in the automatic estimation of sleepiness.

Finally, no measured listeners' characteristics interacted with performances, while errors were correlated with speakers' fatigue, anxiety, and educational level: these results should incite to consider these parameters when automatically estimating sleepiness using voice features. However, such results may also be linked to the reading task on which the patients have been recorded: they still have to be confirmed using spontaneous speech recordings.

5. ACKNOWLEDGEMENTS

We are deeply indebted to all the participants who spent time participating in the Endymion study.

6. REFERENCES

[1] 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.

[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] 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.

[4] 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.

[5] S. Amiriparian, P. Winokurow, V. Karas, S. Ottl, M. Gerczuk, and B. W. Schuller, "A Novel Fusion of Attention and Sequence to Sequence Autoencoders to Predict Sleepiness From Speech," arXiv 2005.08722, 2020, _eprint: 2005.08722.

[6] J. V. Egas-Lopez and G. Gosztolya, "Deep Neural Network Embeddings for the Estimation of the Degree of Sleepiness," in *ICASSP 2021*, Toronto, ON, Canada, 2021, pp. 7288–7292.

[7] 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.

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

[9] Anonymous, ""Prediction of Sleepiness Ratings from Voice by Man and Machine": the Endymion replication study," in *sumbitted to ICPhS 2023*, 2023.

[10] V. P. Martin, J.-L. Rouas, J.-A. Micoulaud-Franchi, P. Philip, and J. Krajewski, "How to Design a Relevant Corpus for Sleepiness Detection Through Voice?" *Frontiers in Digital Health*, vol. 3, p. 686068, Sep. 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdgth.2021.686068/full

[11] V. P. Martin, R. Lopez, Y. Dauvilliers, J.-L. Rouas, P. Philip, and J.-A. Micoulaud-Franchi, "Sleepiness in adults: An umbrella review of a complex construct," *Sleep Medicine Reviews*, p. 101718, Nov. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1087079222001319

[12] V. P. Martin, J.-L. Rouas, J.-A. Micoulaud-Franchi, and P. Philip, "The Objective and Subjective Sleepiness Voice Corpora," in *LREC 2020*, Marseille, France, 2020, p. 6525-6533. [Online]. Available: https://aclanthology.org/2020.lrec-1.803

[13] R. B. Sangal, "Subjective sleepiness ratings (Epworth sleepiness scale) do not reflect the same parameter of sleepiness as objective sleepiness (maintenance of wakefulness test) in patients with narcolepsy," *Clinical Neurophysiology*, vol. 110, no. 12, pp. 2131–2135, 1999.

[14] M. Sato, "The neurobiology of sex differences during language processing in healthy adults: A systematic review and a meta-analysis," *Neuropsychologia*, vol. 140, p. 107404, Mar. 2020.

[15] H. Goy, M. Kathleen Pichora-Fuller, and P. van Lieshout, "Effects of age on speech and voice quality ratings," *The Journal of the Acoustical Society of America*, vol. 139, no. 4, pp. 1648–1659, Apr. 2016.

[16] 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.

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

[18] C. C. Maldonado, A. J. Bentley, and D. Mitchell, "A Pictorial Sleepiness Scale Based on Cartoon Faces," *Sleep*, vol. 27, no. 3, pp. 541–548, 2004.

[19] M. Rhemtulla, P. Brosseau-Liard, and V. Savalei, "When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions." *Psychological Methods*, vol. 17, no. 3, pp. 354–373, Sep. 2012.

[20] V. P. Martin, J.-L. Rouas, P. Thivel, and J. Krajewski, "Sleepiness detection on read speech using simple features," in *10th Conference on Speech Technology and Human-Computer Dialogue*, Timisoara, Romania, 2019.

[21] V. P. Martin, G. Chapouthier, M. Rieant, J.-L. Rouas, and P. Philip, "Using reading mistakes as features for sleepiness detection in speech," in *Speech Prosody 2020*, Tokyo, Japan, 2020, pp. 985–989.

[22] V. P. Martin, B. Arnaud, J.-L. Rouas, and P. Philip, "Does sleepiness influence reading pauses in hypersomniac patients?" in *Speech Prosody 2022*. ISCA, 2022, pp. 62–66.

[23] L. B. Krupp, N. G. LaRocca, J. Muir-Nash, and A. D. Steinberg, "The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus," *Archives of Neurology*, vol. 46, no. 10, pp. 1121–1123, 1989.

[24] A. Shahid, S. Chung, L. Maresky, A. Danish, A. Bingeliene, J. Shen, and C. Shapiro, "The Toronto Hospital Alertness Test scale: relationship to daytime sleepiness, fatigue, and symptoms of depression and anxiety," *Nature and Science of Sleep*, p. 41, 2016.

[25] A. S. Zigmond and R. P. Snaith, "The hospital anxiety and depression scale," *Acta Psychiatrica Scandinavica*, vol. 67, no. 6, pp. 361–370, 1983.