Sleepiness detection on read speech using simple features

Vincent P. Martin, Jean-Luc Rouas, Pierre Thivel, Jarek Krajewski

▶ **To cite this version:**

Vincent P. Martin, Jean-Luc Rouas, Pierre Thivel, Jarek Krajewski. Sleepiness detection on read speech using simple features. 2019. hal-02132438v1

HAL Id: hal-02132438

https://hal.science/hal-02132438v1

Preprint submitted on 17 May 2019 (v1), last revised 10 Oct 2019 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Sleepiness detection on read speech using simple features

Vincent P. Martin[1], *Jean-Luc Rouas*[1], *Jarek Krajewski*[2]

[1]LaBRI CNRS UMR 5800 - Univ. Bordeaux - Talence, France
[2]Engineering Psychology, Rhenish University of Applied Science - Cologne, Germany
vincent.martin@labri.fr, jean-luc.rouas@labri.fr, j.krajewski@ixp-wuppertal.de

Abstract

This paper is about automatic sleepiness state detection using speech samples. Following previous research carried out for the Interspeech 2011 challenge, we use the Sleepy Language Corpus (SLC) for our experiments. However, as we are willing to record our own subjects with a collaboration project with the Bordeaux hospital, we focus only on the read speech samples of that database. Furthermore, we are looking for understandable cues that can guide clinicians to provide a diagnostic. Hence, we devised a set of meaningful features that are close to the signal and restrict the feature selection process to methods that do not use feature combinations. Thus, using simple correlations and a grid search procedure on the training and development parts of the database, we selected a final set of 23 features, reaching a performance on par with state-of-the-art systems. A discussion is proposed on the subjective ground truth used for the boundary between sleepy and non sleepy speech in this database. Finally, we discuss on the interpretation of the features and provide hints on the physiological causes.
Index Terms: Sleepiness detection, feature selection, prosody, read speech

1. Introduction

One of the major challenges for treating neuro-psychiatric pathologies is the follow-up of chronic patients in order to measure early relapses as well as observance and compliance to the treatment. Such a monitoring is possible thanks to connected medical devices (measuring for instance weight, blood pressure or physical activities) but crucial information about how the patients feels are difficult to measure. Regular in-person appointments between doctors and patients are thus required. The growing number of patients however increases the queuing time and often results in episodic followups with unevenly spaced interviews.

Apart from the clinical interviews, it is nonetheless possible to measure some symptoms (e.g.: sadness or sleepiness) with a range of techniques: looking at eye movements, measuring electroencephalographic responses, or examining verbal expressions or body movements. Thanks to recent advances in speech processing, it is now possible to detect precise cues in the voice allowing to characterise the state of a speaker (e.g. to measure the sleepiness level). This method has the following advantages: recording voice data is not invasive and does not require specific sensors nor complex calibration processes. It can thus be set up in various environments, outside laboratories, and allows regular and non-restrictive monitoring of patients.

Studies on sleepiness detection through voice has seen a peak of interest in the early 2010s, culminating with the 2011 Interspeech challenge [1]. Since then, there have been only few reported research on the subject [2]. The best performing automatic classification approach is still a system proposed in 2011 [3] and share the same feature set as most research [4, 5, 6] computed with openSmile [7]. While some have tried to suggest using different features [8, 9], the common drawback to all these systems is the lack of possible interpretation of the features due to feature combination techniques.

We intend to provide a decision on the sleepiness of the speaker, but that decision should be justified by the observation of specific traits in the voice that may have to be confirmed by clinicians. Thus, instead of researching new classifiers or new features selection processes, we wish to investigate in this paper if good performance can be obtained using a set of simple features that could be interpretable for clinicians.

2. Database presentation

2.1. Sleepy Language Corpus

The SLC (Sleepy Language Corpus) is the reference corpus for the Interspeech 2011 challenge [1] on detection of sleepiness through voice. This database is composed of multiple speech tasks conducted in parallel of other sleeping-deprivation studies. The speakers are German volunteers and all the speech samples are in German. More information about the experimental setup of the recordings can be found in [10, 11]. For extensive details about the dataset and the different experiments composing it, we invite the reader to see [12].

The additional information given in the database are the task, genre, attribution in train-development-test set and the mean of three Karolina Sleeping Scale (KSS) scores [13]. KSS is a subjective sleepiness scale ranging from 1 (very alert) to 9 (sleepy with difficulties to stay awake): the score used in the database is the mean of the perceptual value filled by the speaker and the score of two external observers trained to evaluate sleepiness.

2.2. Selection of a read subset of the database

This research is a preliminary study for a project aiming at providing followup for patients treated for sleepiness problems at the Sleep Clinic of Bordeaux [1]. Since our target population may already be under treatment and have interaction difficulties, we decided to consider only a reading task which have a light cognitive load [14]. Furthermore, we selected only the reading tasks with an average duration over 8 seconds: the reading of the fable "Nordwind und Sonne" ("The Northwind and the Sun") (mean duration: 36.5 seconds for *northw*), the reading of two simulated air traffic control communications (mean duration: 9.7 s for *flight1* and 13.8 s for *flight2*) and the reading of a simulated air traffic controller sentence (mean duration: 8.5 s for *roger1*).

[1]Centre Hospitalier Universitaire de Bordeaux

2.3. Ground truth: The Karolinska Sleeping Scale

The KSS being a semi-continuous measure, the choice is made to split the dataset in two classes: following [15, 4, 9, 6], the samples with a mean KSS$>$7.5 will be considered as Sleepy Language (SL). On the contrary, the samples with a KSS$\leq$7.5 are labelled as Non Sleepy Language (NSL). Some statistics of the database with the KSS limit fixed at 7.5 are presented in Table 1. It should be mentioned that others choices for that boundary may be made, such as a limit of 7 as in [16] or 8 in [10]. However, setting the limit between sleepy and non-sleepy is an arbitrary choice that may be different according to sleepiness specialists. We will further comment on this limit in the discussions on the results in Section 4.3.

Table 1: *Statistics (number of samples and total duration of the selected read parts of the SLC database with KSS limit fixed at 7.5.*

Samples		Train	Dev	Test	All
Male	NSL	58 1089s	36 680.5s	60 1035.3s	154 1872.8
	SL	29 364.9s	47 680.5s	22 297.0s	98 2804.8s
Female	NSL	119 1624.6s	101 1286.6s	93 1340.7s	313 1342.4s
	SL	96 941.7s	63 753.7s	66 806.9s	225 4251.9s
Total		303 4019.6s	247 2720.8s	241 3478.9s	791 10219.3s

3. Features extraction

The goal of the project is to focus on features that can be understood and used by physicians. As a consequence, several features are extracted on two time scales. On one hand, excerpt level features are computed directly on each recording, using either an automatic vocalic segments detection algorithm [17] or voiced segments detected using a fundamental frequency extraction algorithm [18]. On the other hand, other features are computed on each voiced segment to characterise the regularity of production of harmonic sounds. These features are averaged for each recording.

3.1. Excerpt level features

The statistics on the duration/proportion of voiced segments or automatically detected vowels should reflect the global behaviour of the speaker. An example of excerpt level feature extraction is given on Figure 1.

The features extracted using this time-frame paradigm are: durvoiced: the total duration of voiced parts (in s.), pervoiced: the percentage in duration of voiced parts, durvowel: the total duration of vocalic segments (in s.), pervowel: the percentage in duration of vocalic segments. This feature set provides 4 features per recording.

3.2. Voiced segments features

The voiced segment feature extraction is illustrated on Figure 2. These features include measurements on the fundamental frequency and intensity curves: F0MEAN: mean of fundamental frequency over a voiced segment, F0VAR: variance of fundamental frequency over a voiced segment, F0SLOPE: slope of

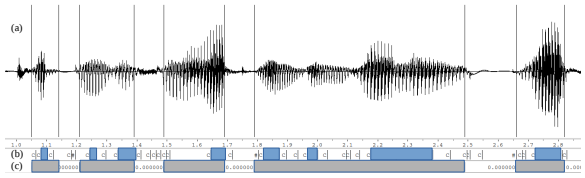

Figure 1: *Illustration of the result of the pre-processing steps for excerpt level features on the sentence "... sich Nordwind und Sonne, wer...": (a) signal, (b) vocalic segments (c) voiced segments.*

the linear approximation of the fundamental frequency over a voiced segment, F0MAX: maximum of fundamental frequency over a voiced segment, F0MIN: minimum of fundamental frequency over a voiced segment, F0EXTEND: extend of fundamental frequency values over a voiced segment. The same features are computed on the intensity curve (NRJMEAN, NRJVAR, NRJMAX, NRJMIN, NRJEXTEND). This results in 12 more features (6 on F0, 6 on intensity). We also computed the F0MEAN, F0VAR, NRJMEAN and NRJVAR features on vocalic segments, adding 4 features to the set.

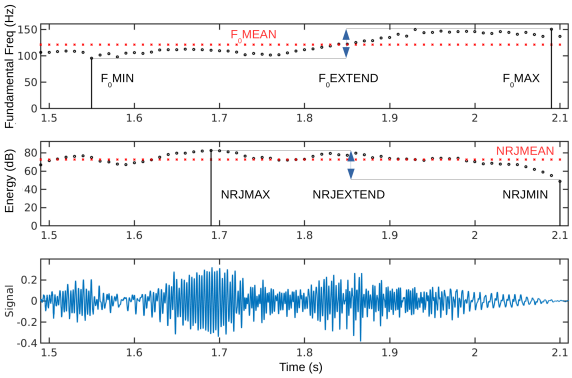

Figure 2: *Illustration of the extraction of features on a voiced segment. Upper pane: fundamental frequency in Hz. Middle pane: intensity in dB. Bottom pane: signal.*

Furthermore, additional features are computed using the COVAREP matlab toolkit [19] which we modified to add some features and compute them only on voiced segments. These features have already been used for characterising singing styles [20] and for spoken social attitudes classification [21] and are the following ones: harmonics amplitude (H1, H2, H4), formants amplitude (A1, A2, A3), frequencies (F1, F2, F3, F4) and bandwidth (B1, B2, B3, B4), differences between harmonics amplitude (H1-H2, H2-H4), differences between amplitude of harmonics and formants (H1-A1, H1-A2, H1-A3), cepstral peak prominence (CPP), harmonics to noise ratios on different frequency bands (HNR05, HNR15, HNR25, HNR35). All these features are averaged over each recording, yielding an additional set of 24 features per recording. We thus extract a total of 44 features.

3.3. Open Smile features

For comparison purposes, we also extract the most widely adopted feature set consisting of 59 low-level descriptors (4 energy related descriptors, 50 spectral descriptors and 5 voice related descriptors), combined with 33 base functionals and 5

F0 functionnals, leading into 4368 features. A more complete description of these features is presented in [1].

4. Proposed system

All the features are centered with the mean values of the features for all the samples of a given speaker (number of speech samples for each speaker: mean=13.4, std=19.4).

4.1. Features selection by statistical methods

In our framework, the reduction of the number of features is constrained by two limits. First, we do not wish to carry out dimensional reduction methods that lead to incomprehensible features such as Principal Components Analysis or Linear Discriminant Analysis, as our goal is to link sleepiness to vocal physiological events. Second, not only the reliability of the KSS as a measure of sleepiness is not certain but it is a semi-continuous measure: a threshold has to be set to label the samples into the sleepy class (SL) or non-sleepy class (NSL). This uncertainty encourage us to prefer methods that do not use a strict limit for the KSS.

One way to choose the features is to select those who have the highest correlation to the KSS measures and give good classification results. After a Shapiro test that ensure that the data is not normally distributed, we conduct a Spearman ρ test to measure the correlation between each of the features and the KSS values. This computation is done only on the aggregated *train+dev*.

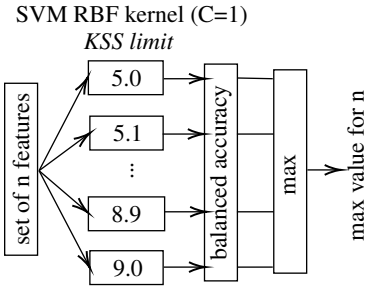

Figure 3: *Flowchart of the gridsearch to compute the best performance with n features. The SVM are trained with the train dataset and tested with the development dataset.*

Then, we perform a grid search on the two parameters of the system: the KSS limit and the number of features. For each number n of features, we keep the n features that correlate the most with the KSS. The train set is then splitted according to the various possible KSS limits (between 5: "neither sleepy neither awaken") and 9: "very sleepy with great efforts to stay awake"). A Support Vector Machine (SVM) using the Radial Basis Function (RBF) kernel is implemented with the Python library *sklearn.SVM*[22] and trained using the training set and tested against the development set. Finally, for a given set of n features, the reported result will be the best result over all the KSS limit values. The flowchart is presented in Figure 3. This procedure is carried out using our complete set of 44 features and the 75 OpenSmile features that correlate the most with the KSS score.

The result of this experiment is shown on Figure 4. On the development set, the best performances are obtained using 23 features from our set (68.1% of UAR) and 59 for the openSMILE features (71.3%). Even though we do not report the results here, the performances obtained using the openSMILE features do not evolve significantly from 70 features on (we tested using up to 200 features).

The selected 23 features from our set are presented in Table 2 with their Spearman correlation value, their significance and the order in which they are added to the system.

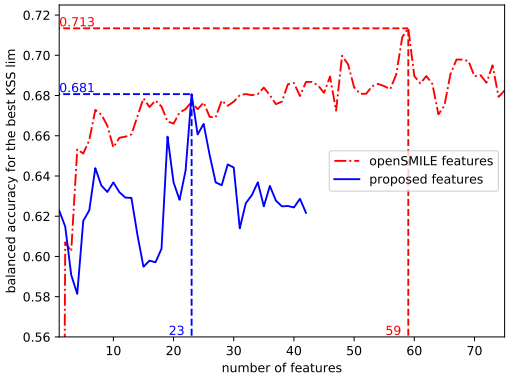

Figure 4: *Unweighted Average Recall (UAR) for the n features that correlate the most with the KSS. The KSS limit is varying from 5 to 9: for each set of features, only the performance of the best system is kept.*

4.2. Results using the conventional KSS limit

Using the 23 selected features, we carry out a classification task using the most conventional limit for the KSS: 7.5. For this experiment, we merge and scale both the training and development set. They are then used to train a SVM classifier using the same parameters as in Section 4.1.The test set is scaled using the mean and variance of the *train+dev* set. The obtained results with our 23 features and the openSMILE features are given in the Table 3. For comparison purposes, we decided to also report the results obtained using the same number of features (23) for the openSMILE features. In this experiment, our proposed set of 23 features outperforms the openSMILE features: the UAR is 76.4% versus 73% for the 59 openSMILE features and 71% if using only 23 openSMILE features.

To compare to the best reported results on the SLC database, we also performed an experiment using the entire database (i.e. not only the read samples). These results are also given in Table 3. While the state-of-the-art performances (using openSMILE features) are around 72% of UAR, the proposed system achieves only 58.3%. This can be due to the fact our features may be more suited to characterise read speech and that many short samples are present in the database: an additional experiment on the duration needed for our features to reach stability show that at least 8 seconds of speech are required.

4.3. Sensibility of the system to the KSS limit

Although most research uses a KSS limit fixed at 7.5, a clear consensus on this value does not seems to exist. Furthermore, fixing a threshold between Sleepy and Non Sleepy patients may be a question left to the appreciation of a specialist. To answer this problem, we compute for different KSS limits the UAR of the output of the system, keeping identical all other parameters,

Table 2: *The selected 23 features achieving the best performances, their Spearman ρ, p-value and rank. $p<0.05$:*, $p<0.01$:**, $p< 0.001$:****

Features	Spearman ρ	p-value	rank
durvoiced	0.057	0.17	23
durvowel	0.045	0.29	19
F0 Mean (vowels)	-0.32	***	1
F0 Mean	-0.27	***	2
F0 Slope	-0.085	*	16
F0 Min	-0.20	***	4
F0 Max	-0.24	***	3
F0 Extend	-0.08	0.053	17
Energy Var (vowels)	-0.07	0.1	21
Energy Var	-0.07	0.1	22
Energy Slope	0.13	**	10
Energy Min	0.14	***	8
Energy Extend	-0.16	***	6
H1	0.10	*	14
H2	0.13	**	9
A2	-0.076	0.073	18
A3	-0.10	*	13
F1	-0.19	***	5
B1	-0.10	*	15
H1A1	0.073	*	20
H1A2	0.12	**	11
H1A3	0.14	***	7
HNR05	-0.12	**	12

Table 3: *Results of the system.*

Features (#)	Sensibility	Specificity	UAR
proposed (23)	75%	77.78%	**76.39%**
openSMILE (23)	69.3%	74.5%	71.9%
openSMILE (59)	63.6%	82.4%	73.0%
Entire SLC database (2808 samples)			
proposed (23)	23.4%	93.1 %	58.3%
State of the art [4]	64.3%	79.1%	71.7%

using the same method as in section 4.2. The results are presented on Figure 5.

Even if the KSS limit of 7.5 leads to satisfying results, the best results are obtained for both systems with a KSS limit of 7, achieving respectively 77.6% and 73.7% of UAR using our set of features and the openSMILE features. Not only the proposed features show satisfying performances for a conventional KSS limit (7 or 7.5), but they are significant of a more global state of sleepiness, whatever the KSS limit being between 6 and 8. For a KSS limit near 6 (between 'Neither Awaken nor Sleepy' and 'Sleepy without efforts to maintain wakefulness'), the system identify the beginning of the sleepiness state. On the contrary, for a KSS limit near 8 (between 'Sleepy without efforts to maintain wakefulness' and 'Sleepy fighting to stay awake'), the system identify an advanced sleepiness. This study allows to let the specialised physicians to determine what state of sleepiness they need to evaluate, the proposed system achieving in all the cases relevant performances.

5. Physiological interpretation

One of the biggest constraint of this work is to select features that can relate the physiological modifications of the voice of the patient caused by sleepiness.

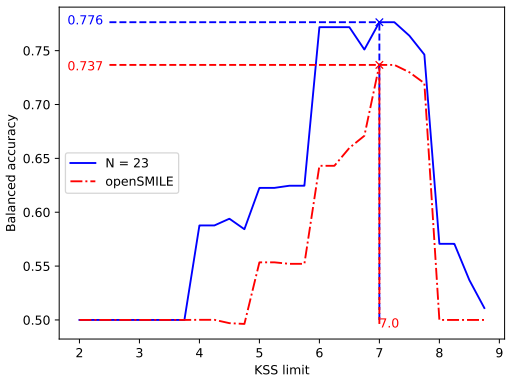

Figure 5: *Sensibility of the system to the limit of the KSS.*

First, as in [23], an augmentation of the voiced and vowels parts is observed. This observation can be clue to the augmentation of hesitations of the sleepy speakers. The diminution of the values of F0 Mean, F0 Min, F0 Max, F0 Extend, F0 Slope, frequency F1 (also observed in [10, 24]), bandwidth of F1 and amplitude of second and third formants witness a shift of the frequencies contained in the voice towards lower values as already observed in [25, 10, 26]. Moreover, the diminution of the values of F0 Extend and F0 Slope are clues of a reduction of the bandwidth used during the vocal process.

Contrary to the observations made in [10], the energy extend, the absolute value of the energy slope and the variance of the energy decrease with sleepiness. Added to the rise of the lower frequencies, the higher energies staying constant, these observations express a diminution of nuances in the Sleepy Language. We hypothesise that the slight augmentation of the first harmonic frequency, that seems contrary to previous observations, is due to the modification of the exhaled air flux that modifies the distribution of harmonics but not formants [27]. This is consistent with the diminution of the HNR (HNR05 in our case) also observed in [28].

All these observations lead to the hypothesis that the sleepy speakers struggle to produce the same variety of nuances of frequencies, energy and quality of voiced parts.

6. Conclusion & Perspectives

In this paper, we have proposed a novel strategy for sleepiness detection in voice, with possible applications in the medical field. We have shown that the subset of reading tasks leads to better results for sleepiness detection and we have developed a set of adapted features. Our system performances are comparable to state of the art methods. The careful selection of features as well as the choice of the subset of the SLC enhance the detection of sleepiness through voice. Moreover, we have proposed a physiological analysis of the vocal parameters for various levels of sleepiness. In the future, we would like to apply these results on the elaboration of a new database in collaboration with the Bordeaux hospital.

7. Acknowledgements

This work is carried out in the framework of the IS-OSA project funded by the French Region Nouvelle Aquitaine.

8. References

[1] B. Schuller, S. Steidl, A. Batlinder, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Interspeech*, 2011.

[2] N. Cummins, A. Baird, and B. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Health Informatics and Translational Data Analytics*, vol. 151, pp. 1–54, 2018.

[3] D.-Y. Huang, Y. Tsao, H. Chiori, and H. Kashioka, "Feature Normalization and Selection for Robust Speaker State Recognition," *IEEE - International Conference on Speech Database and Assessments*, 2011.

[4] D.-Y. Huang, Z. Zhang, and S. S. Ge, "Speaker State Classification Based on Fusion of Asymmetric Simple Partial Least Squares (SIMPLS) and Support Vector Machines," *Comput. Speech Lang.*, vol. 28, no. 2, pp. 392–419, 2014.

[5] J. Krajewski, S. Schnieder, C. Monschau, R. Titt, and D. Sommer, "Large Sleepy Reading Corpus (LSRC): Applying Read and Speech for Detecting Sleepiness," *Speech Communication; 12. ITG Symposium*, pp. 1–4, 2016.

[6] Y. Zhang, F. Weninger, and B. Schuller, "Cross-Domain Classification of Drowsiness in Speech: The Case of Alcohol Intoxication and Sleep Deprivation," in *Insterspeech*, 2017.

[7] F. Eyben and B. Schuller, "Opensmile," *ACM SIGMultimedia Records*, vol. 6, pp. 4–13, 2015.

[8] J. Krajewski, S. Schnieder, D. Sommer, A. Batlinder, and B. Schuller, "Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech," *Neurocomputing*, vol. 84, pp. 65–75, 2011.

[9] C. Sezgin, B. Günsel, and J. Krajewski, "Medium term speaker state detection by perceptually masked spectral features," *Speech Communication*, vol. 67, pp. 26–41, 2015.

[10] J. Krajewski, A. Batlinder, and M. Golz, "Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach," *Behavior Research Methods*, vol. 41, no. 3, pp. 795–804, 2009.

[11] M. Golz, D. Sommer, M. Chen, D. Mandic, and U. Trutschel, "Feature Fusion for the Detection of Microseelp Events," *Journal of VLSI Signal Processing*, vol. 49, pp. 329–342, 2007.

[12] B. Schuller, S. Steidl, A. Batlinder, F. Schiel, J. Krajewski, F. Weninger, and F. Eyben, "Medium-term speaker states-A review on intoxication, sleepiness and the first challenge," *Comput. Speech Lang.*, 2013.

[13] A. Shahid, K. Wilkinson, S. Marcu, and C. M. Shapiro, "Karolinska Sleepiness Scale (KSS)," *STOP, THAT and One Hundred Other Sleep Scales*, pp. 209–210, 2011.

[14] G. Christodoulides, "Effects of Cognitive Load on Speech Production and Perception," Ph.D. dissertation, 2016.

[15] B. Günsel, C. Sezgin, and J. Krajewski, "Sleepiness detection from speech by perceptual features," in *IEEE - ICASSP*, 2013, pp. 788–792.

[16] H. Martensson and O. Keelan, "Feature Engineering and Machine Learning for Driver Sleepiness Detection," Ph.D. dissertation, 2017.

[17] F. Pellegrino and R. Andre-Obrecht, "Automatic language identification: an alternative approach to phonetic modelling," *Signal Processing*, vol. 80, no. 7, pp. 1231–1244, 2000.

[18] K. Sjölander, "The Snack Sound Toolkit," 2004. [Online]. Available: http://www.speech.kth.se/snack/

[19] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP — A collaborative voice analysis repository for speech technologies," in *IEEE - ICASSP*, 2014, pp. 960–964.

[20] J.-L. Rouas and L. Ioannidis, "Automatic Classification of Phonation Modes in Singing Voice: Towards Singing Style Characterisation and Application to Ethnomusicological Recordings," in *Interspeech*, 2016, pp. 150–154.

[21] J.-L. Rouas, T. Shochi, M. Guerry, and A. Rilliard, "Categorisation of spoken social affects in Japanese: human vs. machine," in *ICPhS*, 2019.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[23] L. S. Dhupati, S. Kar, A. Rajaguru, and A. Routray, "A novel drowiness detection scheme based on speech analysis with validation using simultaneous EEG recordings," in *IEEE - Int. CASE*, 2010, pp. 917–921.

[24] H. P. Greeley, E. Friets, J. P. Wilson, S. Raghavan, J. Picone, and J. Berg, "Detecting Fatigue From Voice Using Speech Recognition," in *IEEE International Symposium on Signal Processing and Information Technology*, 2006, pp. 567–571.

[25] T. L. Nwe, H. Li, and D. Minghui, "Analysis and Detection of Speech under Sleep Deprivation," in *Interspeech*, 2006.

[26] E. L. McGlinchey, L. S. Talbot, K.-h. Chang, K. A. Kaplan, R. E. Dahl, and A. G. Harvey, "The Effect of Sleep Deprivation on Vocal Expression of Emotion in Adolescents and Adults," *SLEEP*, vol. 34, pp. 1233–1241, 2011.

[27] J. Hillenbrand, R. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 4, pp. 769–778, 1994.

[28] S. Boyer, R. El-Yagoubi, M. Tiberge, R. Ruiz, and A. Daurat, "Paramètres Acoustiques de la Voix et Privation de Sommeil," in *CFA/VISHNO*, 2016.