Automatic Speech Recognition systems errors for accident-prone sleepiness detection through voice

Vincent P. Martin
LaBRI, CNRS UMR 5800
Univ. de Bordeaux, Bordeaux INP
Talence, France
vincent.martin@labri.fr

Jean-Luc Rouas
LaBRI, CNRS UMR 5800
Univ. de Bordeaux, Bordeaux INP
Talence, France
rouas@labri.fr

Florian Boyer
Airudit Speech Lab / LaBRI
Univ. de Bordeaux, Bordeaux INP
Talence, France
florian.boyer@airudit.com

Pierre Philip
SANPSY, CNRS USR 3413
Univ. de Bordeaux, CHU Bordeaux
Bordeaux, France
pierre.philip@u-bordeaux.fr

***Abstract*—Excessive Daytime Sleepiness (EDS), a symptom linked to chronic sleepiness, impacts everyday life and increases risks of work or road accidents of subjects affected by it. The detection of accident-prone EDS through voice benefits from its ease to be implemented in ecological conditions and to be sober in terms of data processing and costs.**

Contrary to previous works, this study focuses on long-term sleepiness detection through voice. Using the Multiple Sleep Latency Test corpus, we propose a feature selection pipeline inspired by clinical validation practices to classify accident-prone EDS – as measured by a threshold of 15 on the Epworth Sleepiness Scale – based on vocal clues. We propose three different approaches based on the acoustic quality of voice, reading mistakes, and a whole new approach, relying on Automatic Speech Recognition systems errors. The classification system achieves performances on the same scale as the state-of-the-art systems on short-term sleepiness detection through voice (74.2% of Unweighted Average Recall).

Moreover, we give insights into the decision process implied during classification and the system's specificity regarding the threshold delimiting the two classes Higher-risk driver and Lower-risk driver.

***Index Terms*—Sleepiness, Accidental risk, Excessive Daytime Sleepiness, Automatic Speech Recognition, Voice.**

I. Introduction

A. Excessive Daytime Sleepiness and risks of accident

Excessive Daytime Sleepiness – EDS – is one of the most frequent complaints reported to clinicians, with a prevalence estimated between 10% and 25% in the general population [1]. This chronic sleepiness impacts everyday life quality [2] and increases work and road accidental risks: if they have antecedents of sleepiness at the wheel, subjects affected by EDS are two to three times more likely to have a road accident [3]. EDS is usually measured either objectively, by electrophysiological recordings (EEG), or subjectively, using psychometric scales.

This project was supported by IS-OSA project funded by Région Nouvelle-Aquitaine and the national grants LABEX BRAIN (ANR-10-LABX-43)

The gold-standard measure to evaluate objective EDS is the Multiple Sleep Latency Test [4], evaluating the propensity to sleep of the subject. However, this medical procedure is expensive, both on human factors (it requires trained technicians to interpret the EEG signals) and monetary aspects (two nights and an entire day of full hospitalization).

Subjective EDS is measured by psychometric questionnaires such as the Epworth Sleepiness Scale – ESS [5]. It is an 8-items questionnaire on which the subject rates his or her chance to fall asleep in 8 situations encountered in daily life, the total score ranging from 0 to 24. A threshold of 10 is usually used to assess EDS [6] but other studies have determined that a threshold of 15 can predict accident-prone EDS. For example, in a study based on highway drivers [7], very-sleepy subjects (ESS $>$ 15) did significantly more inappropriate line crossings than the non-sleepy group (ESS $<$ 10) or the sleepy group ($11 \leq$ ESS ≤ 15), resulting in higher accident risks.

This study proposes a third method: detection of EDS through voice. This approach benefits from numerous advantages: it is implementable in various environments – including open environments outside laboratory conditions, it is not invasive, it requires neither specific sensors nor complex calibration processes and it is economical in data, requiring neither important calculus processors nor high-performance data networks. It is thus a choice technology for regular and non-restrictive monitoring of patients. Moreover, this sleepiness detection technique could be easily integrated into cutting-edge technologies such as virtual medical interviews [8].

B. State of the art on EDS detection through voice

While the detection of short-term sleepiness through voice has already been the subject of two international challenges [9], [10], long-term sleepiness detection is a new task that has emerged during the last months. Using the Multiple Sleep Latency Test corpus (MSLTc) presented in [11], two approaches have been proposed to detect the objective EDS level of speakers. On one side, [12] proposed a system based

on simple acoustic features, that are explainable to physicians. On the other side, [13] introduced a new set of features based on reading mistakes the patients make during the reading of texts. This article differs from the previous works on this corpus by the studied EDS dimension: the specific case of accident-prone subjective EDS is exclusive of this article.

C. Aim of the study

The objective of the present study is threefold. First, we propose a new feature selection pipeline inspired by clinical validation practices, allowing to select features discriminating EDS independently from other speakers' traits. Second, we seek to validate the previously defined reading mistakes [13] as features for subjective EDS discrimination. Finally, we propose a new set of features based on the errors made by Automatic Speech Recognition systems, misled by the reduced articulation and prosody quality of the sleepy subjects.

This article is organized as follows. In Section II, we introduce the corpus and the ground truth label used in this study. In Section III, we present both the previous acoustic and reading mistakes features, and the new ASR features. The feature selection and classification pipelines are presented in Section IV and the results of this system are presented in Section V. Finally, we discuss these results in Section VI and we draw conclusions and propose future works in Section VII.

II. Corpus and EDS label

A. MSLT corpus

This study is based on the MSLT corpus (MSLTc), relying on the recordings of 106 patients of the Sleep Clinic of the Bordeaux University Hospital. They undertake an MSLT, consisting of taking a 35 minutes maximum nap every two hours, from 9 am to 5 pm. Before each of the 5 naps, the patients are recorded reading out loud texts that are approximately 200-words long extracted from *Le Petit Prince* by Saint-Exupéry. As a consequence, each speaker of the corpus is recorded 5 times during the same day, with different texts and different emotional, fatigue, and circadian states. To ensure consistency between the speakers, we only keep the 93 patients of the corpus affected by different forms of Hypersomnia.

B. Subjective EDS measure

In parallel with the MSLT test, the patients are asked to fill numerous fatigue and sleepiness-related questionnaires, including the Epworth Sleep Scale (ESS). This questionnaire evaluates the subjective propensity to sleep of the subjects. Usually used to discriminate subjective EDS using a threshold of 10, a threshold of 15 has been shown to predict accidental risk [7]. As almost all the patients of the corpus are affected by EDS, we will focus on the accident-prone dimension and classify speakers among Higher-Risk Drivers – HRD (ESS$>$15) and Lower-Risk Drivers – LRD (ESS$\leq$15). The sub-corpus of the MSLTc used in this study is described in Table I. For more information about the MSLTc, we redirect the reader to [11].

TABLE I
Distribution of the speakers across Sex and Sleepiness classes

Sex	HRD (ESS $>$ 15)	LRD (ESS $\leq$ 15)	TOTAL
Women	26	32	58
Men	13	22	35
TOTAL	**39**	**54**	**93**

III. Features

This Section introduces the three sets of features used in this study.

A. Acoustic features

The acoustic features used in this study are designed to be explainable to physicians. They are arranged into two categories. On one hand, statistics (total length and ratio) about voiced and vocalic segments are computed directly on each recording. On the other hand, features are computed on each voiced segment to characterize the regularity of the production of harmonic sounds. These features are averaged for each recording. They are divided into three classes: measurements (mean, var, max, min, extend) on the fundamental frequency and intensity; descriptive values (frequency, power, bandwidth) of harmonics and formants; cepstral peak prominence and HNR. This set of 47 acoustic markers has already been proven efficient for the detection of short-term sleepiness [14] and long-term objective EDS [12]. A complete description of these custom features can be found in [15].

B. Reading mistakes

Introduced in [13], these features rely on the manual annotations of the patients' errors during the reading of the texts. Elaborated with speech therapists, four errors have been designed for sleepiness detection: stumbling errors (defined as "hesitations and breaks in the speech rhythm" [16]), paralexia (i.e. "identification error of written words consisting in the production of a word instead of another" [16]), deletions of words and addition of words.

C. ASR features

Attempting at automatizing the labeling of reading errors, we measured the errors made by ASR systems. Indeed, when a subject feels sleepy, his or her articulation and prosody are impaired [17] while the number of hesitations and repeats increases. This alteration of speech due to sleepiness may induce errors in ASR systems that could be used as biomarkers of sleepiness. Thanks to recent advances in end-to-end ASR systems allowing intermediate transcription units such as characters or tokens, it is possible to transcribe not only words but also portions of words (Byte Pair Encoding – BPE).

In this study, we use an end-to-end system based on RNN transducers with attention, using either words, BPE, or characters, to transcribe words or BPE. The language model is trained on a word, BPE, or character version of the ESTER corpus [18]. A complete review of such systems and their performances is proposed in [19]. The end-to-end system achieving the best performances is the character-based one

with a word-based RNN language model achieving 17.6% of Word Error Rate on the ESTER corpus.

Three types of errors are considered in this study: deletions, insertions, and substitutions, to which we add the number of correctly recognized units. Each type of error is computed on tokens and words, and we consider both the raw number of errors and their ratio over the total number of transcription units, leading to 16 features for each of the 5 ASR systems considered in this study.

IV. Classification pipeline

A. Feature selection

Each previously presented set of features is computed on each of the 5 naps of the MSLT, aggregated with the mean and the standard deviation across the naps, resulting in 7 measurements for each feature.

The feature selection pipeline is described in Figure 1. It is based on statistical methods and divided into two steps:

1) A discriminating test (Mann-Whitney's U) to eliminate features not having a sufficient discrimination power to be statistically different across the two classes ($p > 0.05$);
2) Statistical tests to certify that the selected features discriminate only sleepiness and that they are not correlated to speaker traits that could interfere with voice production. As a consequence, features correlating (Spearman's ρ, $p < 0.05$) with Age, Body Mass Index (BMI), Neck size, Anxiety or Depression score (measured by the Hospital Anxiety and Depression scale [20]), or discriminating Sexes (Mann-Whitney's U, $p < 0.05$) or subtypes of hypersomnia (univariate ANOVA, $p < 0.05$) are eliminated.

The selected features are then aggregated before the classification pipeline (early fusion).

This procedure, usually employed when validating psychometric questionnaires, ensures that the selected features are specific to EDS independently from all the other factors. This pipeline has two major benefits compared with other traditional features selection pipelines: 1) it is compatible with small datasets, as statistical tests do not require a large amount of data; and 2) it is independent of the performances metrics: the selected features do not vary depending on the chosen metric compared with performance-driven pipelines.

B. Classification

To avoid overlearning and to allow generalization, the classification process is carried out under Leave One Speaker Out Cross Validation (LOSOCV): each speaker is turn-by-turn isolated as a test sample, while the classification system is trained on the others. The predicted and ground-truth classes of the test speakers are aggregated and the performance metrics are computed on this aggregation, allowing to validate the reliability of the whole pipeline.

Regarding the classification process, this study aims not to optimize classification performances, but to validate new features for sleepiness detection through voice. As a consequence, a Principal Components Analysis (PCA) followed by logistic regression was sufficient to achieve convincing performances. Features are scaled independently at each iteration of the LOSOCV and the PCA is re-computed each turn, with a number of components that are chosen to reproduce at least 80% of the initial variance. The logistic regression was processed using the Python module sci-kit learn [21] with a *newton-cg* solver and a balanced class-weighting.

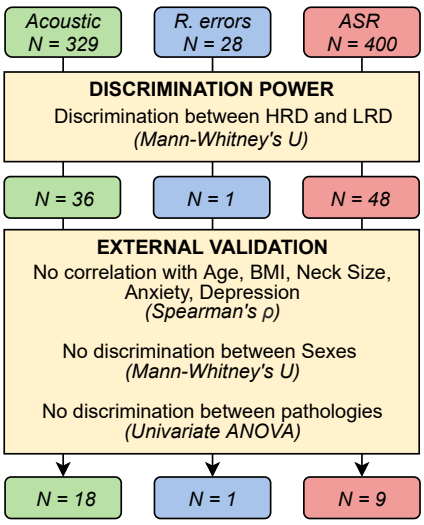


Fig. 1. Features selection pipeline and the number of selected features after each step. *Acoustic*: Acoustic features. *R. errors*: Reading errors. *ASR*: Automatic Speech Recognition features. HRD: Higher-Risk Driver, LRD: Lower-Risk Driver

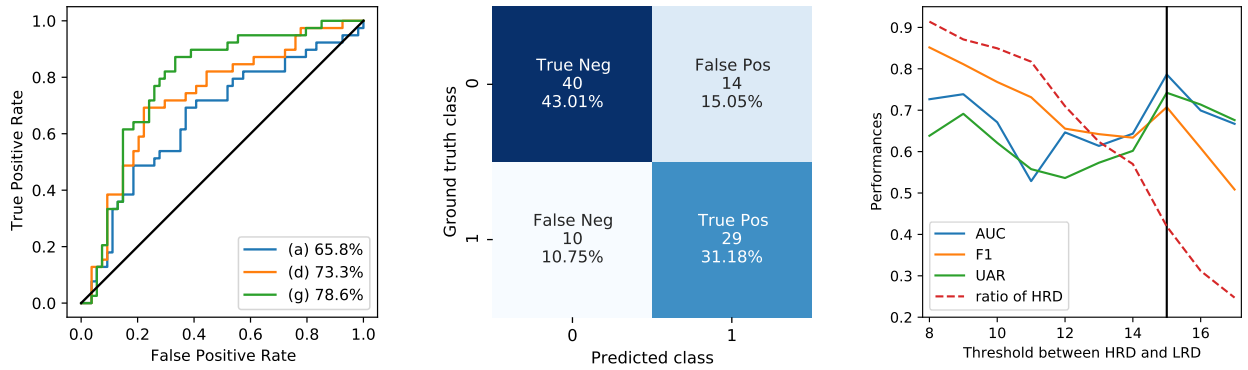
V. Results

TABLE II
Classification performances of the proposed pipeline

	Features	UAR	F1	AUC
(a)	ASR	63.5%	59.5%	65.8%
(b)	Acoustic	64.2%	61.3%	69.2%
(c)	R. errors	61.2%	49.2%	35.4%
(d)	ASR + Acoustic	69.2%	65.9%	73.3%
(e)	ASR+ R. errors	64.5%	60.2%	66.2%
(f)	Acoustic + R. errors	66.1%	62.8%	70.8%
(g)	ASR + Acoustic + R. errors	**74.2%**	**70.7%**	**78.6%**

The obtained Unweighted Average Recall (UAR), F1-score, and Area Under the ROC Curve (AUC) for different sets of features are presented in Table II. The corresponding ROC curves, measuring the discrimination power of the classifier, are presented in Figure 2a.

When taken separately, acoustic and ASR features perform identically (systems (a) and (b), $\approx$ 64% of UAR). However, their combination outperforms all the combinations of two sets of features (69.2% of UAR for system (d)). On the contrary, reading errors perform poorly when they are taken alone (61.2% of UAR for system (c)) or combined with any other set of features (64.5% of UAR for system (e), 66.1% of UAR for system (f)). However, the reading mistakes are not sufficient for correct classification but they seem to be necessary: the



(a) (b) (c)

Fig. 2. **(a)** ROC of the systems (a), (d), and (g) and their respective AUC. **(b)** Confusion matrix of the system (g). **(c)** Performances of the system (g) depending on the threshold between Higher Risk Drivers and Lower Risk Drivers classes

best results are obtained by combining the three set of features and achieves 74.2% of UAR (system (g), 70.7% of F1-score, 78.6% of AUC). The corresponding classification matrix is presented in Figure 2b.

Even if it is not rigorously comparable with other existing systems, the system (g) achieves performances that are in the same vein as the state-of-the-art performances on short-term sleepiness detection (71.7% in [22], 76.4% in [14]). As a consequence, the proposed pipeline selects relevant features that allow the detection of accident-prone EDS independently from other speaker traits (age, sex, BMI, ...).

VI. Discussion

A. Specificity of the measure

In the same vein as presented in [14], we represented in Figure 2c the variations of the performances depending on the limit to distinguish the two classes during the classification. This measure reflects the specificity of the selected features and the pipeline: the best performances (UAR and AUC) are observed for the classification limit of 15 that we have selected for our classification. Another peak is observed around a limit of 9, but the imbalance between classes (less than 15% of LRD) makes any conclusion perilous. As a consequence, our system seems specific to the threshold of 15 on the ESS, corresponding to an accident-prone dimension of EDS.

B. Need for PCA

At a first glance, the low number of selected features (28) should allow classification without the need for dimension reduction techniques such as PCA. Thus, we applied the same pipeline as in system (g) without the PCA. This led to classification performances noticeably lower than the same pipeline with PCA (UAR: 68.2%, F1: 63.3%, AUC: 69.7%). Indeed, not only PCA can be used as a dimension reduction technique, but it is also an orthogonalizing process, optimizing the following logistic regression.

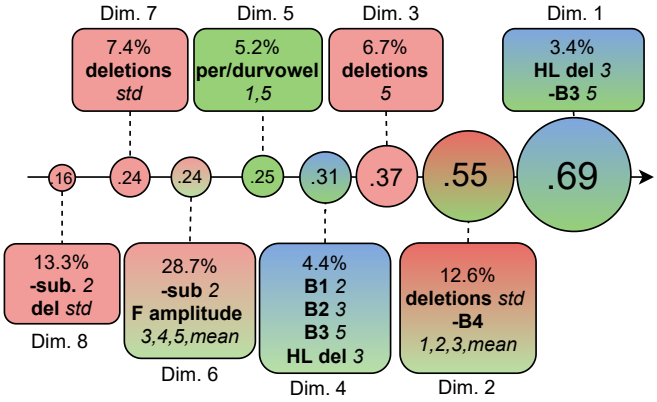


Fig. 3. PCA components and the associated weight in the logistic regression. From top to bottom: mean ratio of explained variance in the PCA, feature (in bold), condition (in italic). Green background: Acoustic features; blue background: Reading errors; red background: ASR features.
- : negative PCA weight; F: Formant; Sub: ASR substitutions; Del: ASR deletions; HL del: Hand-Labeled deletions; Dim: Dimension

C. PCA dimensions analysis

Along the cross-validation process, the parameters of the PCA and the weights of the logistic regression are averaged. Figure 3 represents the eight different PCA dimensions and their corresponding weights in the logistic regression.

1) Reading errors: The most important feature is the PCA dimension partly directed by the hand-labeled deletions of the third nap (mean coefficient across the LOSOCV, $\alpha_1 = 0.69$). In the first or fourth component, hand-labeled reading errors share PCA dimensions only with acoustic features: even if the selected ASR errors comprise deletions, they are not to replace hand-labeled errors but a whole complementary measure of sleepiness expression through voice.

2) Acoustic features: The relevant acoustic features are linked to the bandwidth of the formants: the third formant (B3) during the fifth naps (Dim. 1); the fourth formant (B4) during the first, second, and third naps and its average value across

the nap ($\alpha_2 = 0.55$); the first, second and third formants resp. during the second, third, and fifth naps ($\alpha_4 = 0.31$). Their amplitude during third, fourth, and fifth naps and their mean value are also relevant ($\alpha_6 = 0.24$). Finally, the percentage and duration of vowel parts extracted from audio during the first and fifth naps take part in the decision of the classifier ($\alpha_5 = 0.24$).

3) ASR features: Regarding ASR, the most relevant features are the standard deviation of the deletions (Dim. 2, 7, and 8) and its value during the fifth nap (Dim. 3). Substitutions during the second nap are also important when contributing to the sixth and eighth dimensions. These errors both come from a character-based ASR system with a language model trained on characters or words, that are among the systems achieving the best recognition performances [19].

D. Condition of the features

During feature selection, the selected acoustic features were all statistics on the formants and were computed on different naps (or averaged across the naps), whereas only one of the reading error features on only one nap has been selected. Two phenomena could explain these observations.

First, the text influences the features. Indeed, we hypothesize that if deletions are computed only on the third nap (Dim. 1 and 4.), it is because the corresponding text favors the differences between HRD and LRD. This also could be the case for ASR-based features, for which the ASR quality could depend on the content of the texts, or acoustic features for which the phoneme distributions across the texts are unbalanced.

Second, the state of the speakers impacts their voice when they are recorded. Indeed, numerous studies have shown that short-term sleepiness impacts voice [9], [10], but it also the case of emotion or the circadian state of the speaker. For example, when recording the MSLTc, numerous subjects complained about fatigue or boredom: the state of the speaker at this moment could favor vocal traits linked to sleepiness. This could explain why ASR deletions are computed on the standard deviation of naps (Dim. 7) but also on the fifth nap (Dim. 5).

VII. Conclusion and future work

This article has proposed a novel feature selection pipeline based on clinical validation to achieve accident-prone EDS classification through voice. It has been tested with acoustic, reading errors, and new ASR systems errors, and achieves 74.2% of performances.

In the future, we plan to extend the scope of the proposed pipeline to further sleepiness-related phenomena, such as subjective sleepiness or objective EDS.

References

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

[2] M. M. Ohayon, C. F. Reynolds, and Y. Dauvilliers, "Excessive sleep duration and quality of life: Excessive Sleep in USA," *Annals of Neurology*, vol. 73, no. 6, pp. 785–794, Jun. 2013.

[3] P. Philip and P. Sagaspe, "Sleep and accidents," *Bull. Acad. Natl. Med.*, vol. 195, no. 7, pp. 1635–1643, 2011.

[4] D. Arand, M. Bonnet, T. Hurwitz, M. Mitler, R. Rosa, and R. B. Sangal, "The Clinical Use of the MSLT and MWT," *SLEEP*, vol. 28, no. 1, pp. 123–144, 2005.

[5] M. W. Johns, "A New Method for Measuring Daytime Sleepiness: The Epworth Sleepiness Scale," *Sleep*, vol. 14, no. 6, pp. 540–545, 1991.

[6] ——, "Sleepiness in Different Situations Measured by the Epworth Sleepiness Scale," *Sleep*, vol. 17, no. 8, pp. 703–710, Dec. 1994.

[7] P. Philip, P. Sagaspe, J. Taillard, G. Chaumet, V. Bayon, O. Coste, B. Bioulac, and C. Guilleminault, "Maintenance of Wakefulness Test, obstructive sleep apnea syndrome, and driving risk," *Annals of Neurology*, vol. 64, no. 4, pp. 410–416, Oct. 2008.

[8] P. Philip, L. Dupuy, M. Auriacombe, F. Serre, E. de Sevin, A. Sauteraud, and J.-A. Micoulaud-Franchi, "Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients," *npj Digital Medicine*, vol. 3, no. 1, p. 2, 2020.

[9] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Interspeech 2011*, 2011, pp. 3201–3204.

[10] 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.

[11] V. P. Martin, J.-L. Rouas, J.-A. Micoulaud-Franchi, and P. Philip, "The Objective and Subjective Sleepiness Voice Corpora," in *12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020, p. 6525-6533.

[12] V. P. Martin, J.-L. Rouas, and P. Philip, "Détection de la somnolence dans la voix : nouveaux marqueurs et nouvelles stratégies," *TAL*, vol. 61, no. 2, pp. 67–90, 2021.

[13] V. P. Martin, G. Chapouthier, M. Rieant, J.-L. Rouas, and P. Philip, "Using reading mistakes as features for sleepiness detection in speech," in *10th International Conference on Speech Prosody 2020*, Tokyo, Japan, 2020, pp. 985–989.

[14] 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.

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

[16] F. Brin, C. Courrier, E. Lederle, and V. Masy, *Dictionnaire d'orthophonie - 4ème édition*, orthoedition ed., Sep. 2018.

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

[18] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts," in *Interspeech 2009*, 2009, pp. 2583–2586.

[19] F. Boyer and J.-L. Rouas, "End-to-End Speech Recognition: A review for the French Language," *[Unpublished]*, 2019, arXiv: 1910.08502. [Online]. Available: http://arxiv.org/abs/1910.08502

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

[21] 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.

[22] B. Schuller, S. Steidl, A. Batliner, 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.*, vol. 28, no. 2, pp. 346–374, 2013.