Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Using reading mistakes as features for sleepiness detection in speech

Vincent P. Martin[1], *Gabrielle Chapouthier*[2], *Mathilde Rieant*[2], *Jean-Luc Rouas*[1], *Pierre Philip*[3]

[1]LaBRI - CNRS - UMR 5800 - Univ. Bordeaux - Bordeaux INP - F-33400 Talence (France)
[2]CFUOB - Univ. Bordeaux Sengalen - F-33076 Bordeaux (France)
[3]SANPSY - CNRS - USR 3413 - Univ. Bordeaux - CHU Pellegrin - F-33000 Bordeaux (France)

{vincent.martin, rouas}@labri.fr
{gabrielle.chapouthier, mathilde.rieant}@etu.u-bordeaux.fr
pierre.philip@u-bordeaux.fr

Abstract

Automatic detection of sleepiness can help to improve the follow-up of patients suffering from chronic diseases. Previous research on sleepiness detection has shown that this task is feasible using voice recordings. Most studies however rely on numerous features extracted from healthy subjects recordings and machine learning, the target being the output of subjective sleepiness questionnaires. In this paper, we propose to study the reading errors made by patients suffering from Excessive Daytime Sleepiness on the MSLT database collected at the Bordeaux hospital. This database differs from the others on two key points: patients are recorded instead of healthy subjects and their sleepiness level is assessed using multiple measurements, both subjective and objective. With the help of Speech Therapists, we defined and counted reading errors and confront these numbers with sleepiness measurements. We show that evaluating these reading errors can be useful to elaborate robust markers of objective sleepiness but also to elaborate exclusion criteria of the speakers not having a sufficient reading level.
Index Terms: reading mistakes, sleepiness detection, prosody

1. Introduction

One of the major challenges for treating neuro-psychiatric pathologies is the follow-up of patients suffering from chronic diseases in order to adapt treatment and measure early relapses. Regular in-person appointments between doctors and patients are useful but the growing number of patients increases the queuing time and often results in episodic follow-ups with unevenly spaced interviews. To tackle this issue, a virtual physician has already been developed [1]. Benefiting from the vocal interaction between the patient and the animated conversational agent, measuring sleepiness from voice provides useful complementary information.

Even if previous studies have already shown that it is possible to measure subjective sleepiness through voice using extracted vocal features [2, 3, 4], most of them were based on the Sleepy Large Corpus introduced during the Interspeech 2011 challenge [3] which included only healthy subjects. More recently, the SLEEP corpus has been elaborated for the Interspeech 2019 challenge, paving the way to the use of deep learning in sleepiness detection through voice.

Since our goal is to estimate objective sleepiness for patients suffering from Excessive Daytime Sleepiness, these datasets did not fit our expectations for three main reasons. First, the sleepiness level of this database is a combination of self-reported sleepiness and external behavioural sleepiness (measured by the Karolinska Sleepiness Scale [5] - KSS), while we attempt to estimate objective sleepiness (usually measured by EEG). Second, the vocal samples are collected through diverse tasks ranging from short sustained vowels to reading diverse texts, in English and German languages. This lead to heterogeneous samples that are barely comparable. Finally, but not least, we wish to estimate sleepiness from voice produced by patients suffering from Excessive Daytime Sleepiness. For this population, contrary to healthy subjects, objective and subjective sleepiness measurements may not correlate [6]. Moreover, these patients are more likely to suffer from depression [7], that can pollute the vocal production [8].

To achieve our goal, we have elaborated a new database recorded at the Bordeaux University Hospital Sleep Clinic, France. The MSLT database, extensively described in [9], contains speech samples collected on reading tasks and associated measurements for self-evaluated subjective sleepiness (KSS) and objective sleepiness measurements (Multiple Sleep Latency Test - MSLT - iteration value [10, 11]). The advantages on focusing on read speech are multiple: not only the vocal content is the same when comparing the patients, but it also ensures that the length of the samples is sufficient to ensure the detection of sleepiness.

To our knowledge, all the previous works aiming at detecting the sleepiness level through voice have focused on extracting features concerning voice quality (frequency, energy, ...), usually extracted with the openSMILE toolbox [12]. We forecast that it is mainly due to the available datasets, which only propose audio samples and their labels. This article aims at benefiting from the fact that in our corpus, the voice is collected on read texts, allowing to label the reading errors for each sample and using this labels as new features for sleepiness detection. Indeed, if vocal features allow to study the influence of sleepiness on the neuro-muscular aspect of vocal production [13], we hypothesise that the reading mistakes are relevant markers to study the cognitive modifications due to sleepiness. We hence propose in this paper a new method to assess the sleepiness of a speaker based on its reading errors.

This paper is structured as follows. In Section 2, we provide a description of the five types of reading errors that we used to label our database. In Section 3 we briefly present our database and discuss on the criteria that we used to exclude some speakers. Results and discussion are presented in Section 4, while Section 5 presents a first attempt of classification. Finally, conclusions and future work are presented in Section 6.

2. Reading mistakes taken into account

In accordance with speech therapists, we have selected five types of errors that may be represented well enough in our data:

- Stumbling errors (”Achoppements” in French): ”hesita-

10.21437/SpeechProsody.2020-201

tions, breaks in the speech rhythm" [14].

These errors mainly measure the *assembling* capacities of the reader. *Assembling* is the fact to put together independent syllables so as to form a word: when a subject begins to read a word, stops, then continue, the process of assembling the word has been interrupted, leading to a stumbling. We have chosen to not take into account hesitations between words (breaks of the speech flow), but only breaks that occur inside words and unnatural vowel lengthening testifying hesitation.

- Paralexia ("Paralexies" in French) : "identification error of written words consisting in the production of a word instead of another"[14].

Contrary to stumbling errors, paralexia reflect the *addressing* capacities of the reader. In contrast to *assembling*, *addressing* can be defined as the fact to read a word wholly, without deciphering it or slicing it into syllables. Paralexia are symptomatic errors involving this type of reading. We have generalised this category to the pronunciation of any other word, existing or not, that is read instead of the correct one. For example, collapsing errors (the deletion of one syllable in a word) are counted as paralexia in this study. The pronounced word has however to be similar to the expected one, to differentiate this error from additions and deletions of words.

- Deletions of words: this error occurs when the speaker forgets to pronounce a word and goes directly to the next one. Even if self-correction occurs afterwards, the deletion error is counted.
- Additions of words: this error occurs when the speaker adds a word that is not present in the text. Even if self-correction occurs afterwards, the addition error is taken into account.
- Syntactic reversals: this error occurs when words in a sentence are inverted.

If a paralexia, deletion, addition or syntactic reversal error has already been counted, self-correction results in not taking into account an additional stumbling error, except if the patient mistakes during its resumption.

3. Description of the database

3.1. Database overview

The database used in this study is an extended version of the MSLT database, already presented in [9]. It consists of the voice recordings of 105 patients from the Sleep clinic of Bordeaux, France. Every patient has complaints about sleep disorders and undergoes a Multiple Sleep Latency Test [10] - MSLT - consisting on asking the patients to take 5 naps a day every two hours since 9am. The patients are highly phenotyped and physical characteristics, results of depression, fatigue and long-term sleepiness questionnaires are collected. The main advantage of this corpus lies in the fact that it gives both objective (sleep onset at each nap, called "MSTL iteration value") and subjective (KSS) values of sleepiness at each iteration. The voice samples are collected during the reading of a text, that is different at each session but the same for all the speakers at constant session.

3.2. Speaker selection (exclusion criteria)

Labelling the database with the errors described in Section 2, some reading profiles have been drawn to exclude subjects from this study. Admittedly, excluding speakers can reduce the size of the corpus but this however ensures that the computed vocal features and reading mistakes are mostly linked to sleepiness, excluding the influence of pathologies and reading disorders on these markers.

First, we kept away the patients that have medical history of stroke or transient ischaemic attacks: the errors made by these patients could possibly be sequelae of these events (alexia or visuo-attentional disorder are common sequelae of these pathologies). Based on this criteria, we have excluded three patients whom had a very slow reading flow and produced a lot of errors. In the same vein, patients with current neuromuscular pathologies (e.g. disphonia, myotonia, Huntington's chorea, epilepsy) are also excluded from the database: involuntary muscular contraction or lack of muscular control are likely to lead to the observation of numerous stumbling errors. Based on this criteria, three additional patients (respectively suffering from epilepsy, spasmodic disphonia and Steinert myotony) have been excluded.

Other criteria concerning more specifically the reading abilities include the Attention deficit disorder (associated or not to hyperactivity). As a matter of fact, as attention plays an important role when reading, these patient may skip words or lines. It has to be noted that sleepiness may also be the cause of such behaviour, but differentiating the origin of these mistakes may prove difficult. As a precaution, we therefore chose to exclude patients that produce such reading errors. Two patients having skipped a row and another being diagnosed with Attention deficit disorder, the three presenting incoherent readings, their recordings were removed from the study. Concerning fluency, we took into account disorders such as stuttering or cluttering as they are difficult to differentiate from the errors induced by sleepiness One patient presenting characteristics of cluttering has therefore been excluded from our corpus. Finally, we also excluded three patients suffering from dyslexia or disorder implying the deciphering of the text despite having read it silently a few minutes before.

One supplementary patient suffering from different serious inflammatory diseases (Crohn, Basedown and Ankylosing Spondylitise diseases) and not presenting a satisfactory reading level has been excluded.

Table 1: *Concise statistics about the clean database*

	Women	**Men**	**Total**
#subjects	53	37	91
#samples	115	340	455
mean Age (std)	34.4 (11.7)	37.9 (16.8)	35.8 (14.1)
mean Social Level (std)	5.8 (2.6)	4.5 (2.2)	5.3 (2.5)
mean MSLT (std)	11.82 (4.48)	10.07 (5.01)	11.11 (4.77)
mean KSS (std)	4.5 (1.1)	4.4 (1.4)	4.4 (1.3)
#SL	16	7	23
#NSL	38	30	68

Excluding these recordings, we have kept a total of 91 speakers out of the 105 original ones.

After having excluded the previously mentioned patients of the database, concise statistics concerning the studied subjects are presented in Table 1. We clustered patients between Sleepy (SL) and Non-Sleepy (NSL) using the 8 minutes medical limit on the mean MSLT value used to assess narcolepsy [10, 11]. Moreover, the social level is measured as the number of years of study after the French Certificate of General Education.

4. Statistical results

4.1. Speaker characteristics

To begin with, we plotted the distribution of errors averaged across all speakers in Figure 1 (plotted with Standard Error of the Mean - SEM). A first analysis shows that regarding the total count of all errors, the Sleepy speakers make more mistakes than their Non-Sleepy counterparts (Mann-Whitney, $U = 464.0, p = 1.9 \times 10^{-3}$). The same trend is observed when considering each type of error separately, except the Syntactic reversals (Mann-Whitney tests. Stumblings: $U = 588.5, p = 3.9 \times 10^{-2}$; Deletions: $U = 482.0, p = 2.7 \times 10^{-3}$; Additions: $U = 527.0, p = 7.7 \times 10^{-3}$; Paralexia: $U = 471.0, p = 2.2 \times 10^{-3}$). As there are few syntactic reversals, we decided not to consider them for the rest of the study.

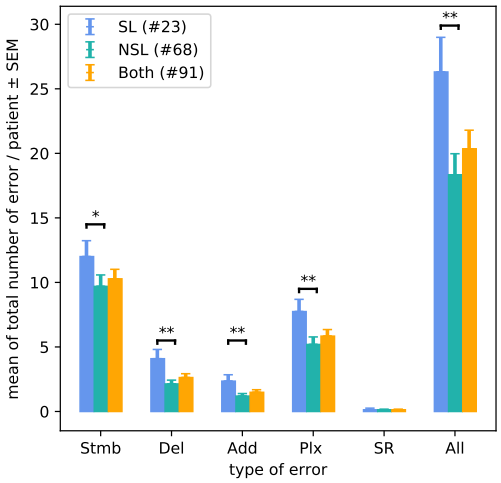

Figure 1: *Mean distribution of errors by speaker plotted with SEM. Stmb: Stumblings, Del: Deletions, Add: Additions, Plx: Paralexia, SR: Syntactic reversion.*
Mann-Whitney tests (: $p < 5 \times 10^{-2}$, **: $p < 10^{-2}$).*

Then, we investigated if the total number of errors in each category correlates (Spearman's ρ correlation) with the different medical and social data available in our database. The social level or age of the subject does not correlate with the error production on our 91 patients, contrary to what we expected.

Interesting results include a correlation between the total number of additions per patient and their mean MSLT values over the five iterations of the test ($\rho = 0.30, p = 3.0 \times 10^{-3}$). This correlation means that the more the patients are affected by long-term sleepiness diseases, the more they make additions errors. The same observation is made for the total number of paralexia, that correlates not only with the mean objective sleepiness value ($\rho = 0.25, p = 1.5 \times 10^{-2}$) but also with the mean subjective sleepiness KSS measure ($\rho = 0.27, p = 9.5 \times 10^{-3}$). The fact that paralexia positively correlates with both objective and subjective mean sleepiness measures indicates that the production of this kind of errors increases with the severity of their long-term sleepiness disease but also with the feeling patients have of their sleepiness. This has the advantage to help detecting both types of sleepiness but has the drawback to prevent from differentiating them: a high production of paralexia errors can be due to either a high objective sleepiness level or the feeling that it is high. Comparatively, the number of additions errors seems a more robust bio-marker of the sleepiness level of patients as it only correlates with the mean objective sleepiness.

4.2. Iteration level

Although a correlation between the total number of errors per patient and their long-term sleepiness, the peculiar influence of the texts on the error production has to be unraveled. Indeed, the texts have neither exactly the same size nor the same difficulty level or the same number of dialogue. Moreover, some time-dependant variables can affect both the sleepiness level and the error production, such as the fact that the patients have breakfast before the first iteration (9am), that they have lunch shortly before the third iteration (1pm) or that they usually express boredom or fatigue during the last iteration (5pm). In the following, "iteration influence" means either influence of the text or time-dependant variables, the two influences not being separable.

Addition, paralexia and MSLT through the iterations of the test are represented in Figure 2. Paralexia (resp. Additions) and KSS variations are represented in Figure 3 (resp. Figure 4). To sort out the contribution of time, MSLT and KSS over the variations of addition and paralexia errors, we conducted a multivariate ANOVA with repeated measures using R [15]. Concerning the paralexia and addition errors, we studied separately the Sleepy (mean MSLT value $\leq$ 8, abbreviated "SL") and Non-Sleepy (mean MSLT value >8, abbreviated "NSL") subjects.

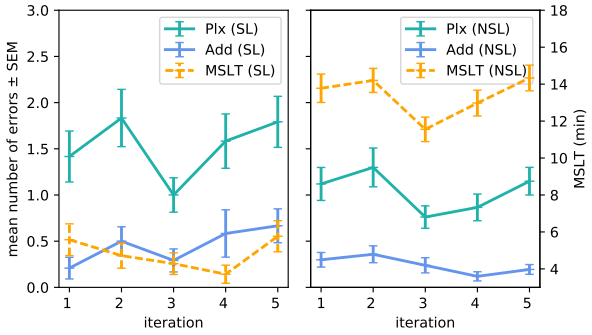

Figure 2: *Additions, Paralexia and MSLT depending on the iteration (plotted with SEM). Plx: Paralexia, Add: Additions.*

Observing the variations of MSLT values and additions, there seems to be an influence of objective sleepiness on additions. This is confirmed by the ANOVA tests, that give almost significant results on both SL and NSL subjects (NSL: $F = 2.89, p = 9 \times 10^{-2}$, SL: $F = 3.40, p = 8.0 \times 10^{-2}$). It is also to be noted that the variations of the number of paralexia are linked to the variations of the KSS (NSL: $F = 3.49, p = 6.3 \times 10^{-3}$, SL: $F = 7.58, p = 1.2 \times 10^{-2}$). This did not appear when computing correlation between the total number of errors and the mean KSS of the patients (cf Section 3.2) but appears on Figure 4. The session not having a significant effect on the production of these errors, we hypothesise that the variations of the addition errors are mainly due to the variations of objective and subjective sleepiness, and that they are independent from text and other time-dependant variables.

The influence of sleepiness on the production of paralexia errors is more complex to analyse. Indeed, if the MSLT seems to have an effect on the variations of the production of paralexia errors when studying the SL subjects (SL: $F = 3.087, p = 8.2 \times 10^{-2}$), there does not seem to have an effect of the sleepiness over this type of errors when observing the mistakes production of NSL patients (NSL: $F = 0.76, p = 0.39$). How-

ever, there is a significant influence of the iteration on both groups (NSL: $F = 2.92, p = 2.1 \times 10^{-2}$, SL: $F = 2.53, p = 4.6 \times 10^{-2}$). When labelling the database, it has been observed that some words are almost systematically mistaken with paralexia (for example the word "méditatif" is often pronounced as "médiatif"): the influence of the iteration certainly comes from this bias. As the total number of paralexia per patients correlates with the mean KSS, we also plotted Paralexia mistakes and KSS depending on the iterations in Figure 3. Even if the influence of the KSS is not shown using ANOVA, the paralexia varies in the same way than the KSS, ascertaining the hypothesis that they are linked to both subjective and objective sleepiness.

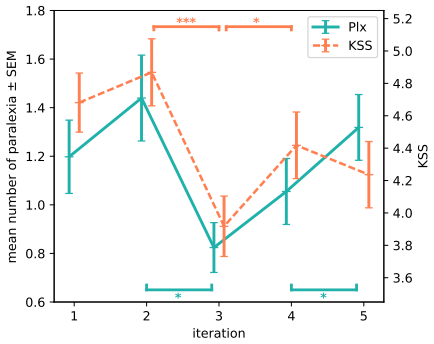

Figure 3: *Paralexia and KSS across iterations, plotted with SEM. Mann-Whitney tests (*: $p<5 \times 10^{-2}$, ***: $p<10^{-3}$).*

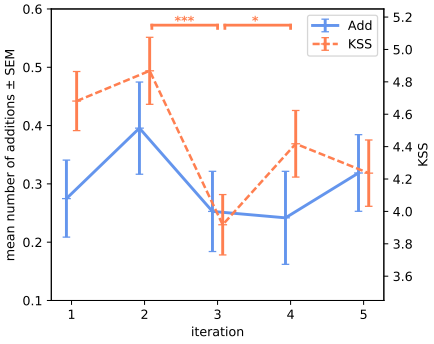

Figure 4: *Additions and KSS across iterations, plotted with SEM. Mann-Whitney tests (*: $p<5 \times 10^{-2}$, ***: $p<10^{-3}$).*

Although we observed a difference between Sleepy and Non-Sleepy subjects for almost all the categories of errors (Figure 1), only the total number of paralexia and additions have shown correlation with mean objective or subjective sleepiness. They are also the only errors whose variations are explained by the variations of the sleepiness of the subject.

Concerning stumbling errors, it has been chosen to ignore the stumbling between words, as they are hard to differentiate from specific accents or unusual breathing. This could be the reason for not observing differences between SL patients and their NSL counterparts, preventing from studying inter-words hesitations observed during the labelling of the database.

Concerning the deletions of words, we remarked that the forgotten words are almost always the same. They concern small transition words that are usually skipped in oral language (for example "Il me répéta" instead of "Et il me répéta" i.e. "He repeated" instead of "And he repeated"). We conducted an ANOVA test that confirms that the variations of the deletions errors are strongly dependent of the text (influence of the iteration: $F = 5.93, p = 1.6 \times 10^{-4}$), preventing them from being robust biomarkers of the sleepiness state of the speaker. This raises the necessity to make an in-depth study of the content of the text, to prevent this phenomena but also to ensure that they are visually equivalent (dialogues have been observed to imply some visuo-attentionnal errors) and that the difficulty of the texts is not the source of the errors made by the subjects.

5. Classification experiment

As a first experiment to classify between SL and NSL speakers, we concatenate the Paralexia and Additions errors from the five iterations of the test and use this vector as input of a Support Vector Machine Classifier (linear kernel, $C = 1 \times 10^{-2}$). As our database has few samples (91 patients), we use Leave One Speaker Out Cross Validation (LOSOCV): each speaker is turn by turn isolated to be considered as test, while the others form the train set. The result of the estimated class for the test sample is then added to a global confusion matrix. Scaling the train set, training the SVM parameters and evaluating the system for each iteration of the LOSOCV, we obtain an unweighted accuracy recall (UAR) computed from the global confusion matrix reaching 61.5% (Sensibility: 36.0%, Specificity: 83.8%). For comparison purposes, we evaluate the same system with the concatenation of all type of errors, leading to an UAR of 61.4% (Sensibility: 52.2%, Specificity: 70.6%). While the performances obtained by this system are not very good, it will be interesting to see how it may combine with classical systems using very different features.

6. Conclusions

In this paper, we have elaborated new features for sleepiness detection based on measurements of reading errors. Measuring these errors provides two advantages: they may be used to exclude speakers from databases (either related to their reading level or to pathologies) but also to assess their sleepiness level. Correlations are indeed shown between some of our measurements and objective and subjective sleepiness measurements. We have also shown that these new features can be used for classification. This could be used to improve classical systems relying on different kinds of features.

The next step of our research will be to find a way to automatically detect the reading errors we presented here, using automatic speech transcription systems. Future works will also include the elaboration of measurements adapted to spontaneous speech and fusion of classical systems features and our new feature set.

7. Acknowledgements

This work is carried out in the framework of the IS-OSA project funded by the French Region Nouvelle Aquitaine and by the SOMVOICE project sponsored by the Labex BRAIN (University of Bordeaux, France).

8. References

[1] P. Philip, J.-A. Micoulaud-Franchi, P. Sagaspe, E. De Sevin, J. Olive, S. Bioulac, and A. Sauteraud, "Virtual human as a new diagnostic tool, a proof of concept study in the field of major de-

pressive disorders," *Scientific Reports*, vol. 7, no. 1, pp. 426–456, 2017.

[2] 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*, 2019.

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

[4] 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.

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

[6] R. 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.

[7] E. O. Bixler, A. N. Vgontzas, H.-M. Lin, S. L. Calhoun, A. Vela-Bueno, and A. Kales, "Excessive Daytime Sleepiness in a General Population Sample: The Role of Sleep Apnea, Age, Obesity, Diabetes, and Depression," *The Journal of Clinical Endocrinology & Metabolism*, vol. 90, no. 8, pp. 4510–4515, Aug. 2005.

[8] C. Espy-Wilson, A. Lammert, N. Seneviratne, and T. Quatieri, "Assessing Neuromotor Coordination in Depression Using Inverted Vocal Tract Variables," in *Interspeech 2019*, 2019, pp. 1448–1452.

[9] 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*, 2020.

[10] M. R. Littner, C. Kushida, M. Wise, D. G. Davila, T. Morgenthaler, T. Lee-Chiong, M. Hirshkowitz, D. L. Loube, D. Bailey, R. B. Berry, S. Kapen, and M. Kramer, "Practice Parameters for Clinical Use of the Multiple Sleep Latency Test and the Maintenance of Wakefulness Test," *Sleep*, vol. 28, no. 1, pp. 113–121, 2005.

[11] M. S. Aldrich, R. D. Chervin, and B. A. Malow, "Value of the multiple sleep latency test (MSLT) for the diagnosis of narcolepsy," *Sleep*, vol. 20, no. 8, pp. 620–629, 1997.

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

[13] J. Krajewski, A. Batliner, 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.

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

[15] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2017. [Online]. Available: https://www.R-project.org/