Automatic detection of schwa in French hypersomniac patients

C. Beaumard[1,2], V.P. Martin[1,2], Y. Wu[3,4,5], J-L. Rouas[1], P. Philip[2],

[1] CNRS, Université de Bordeaux, Bordeaux INP, LaBRI – UMR 5800
[2] CNRS, Université de Bordeaux, SANPSY – UMR 6033
[3] Université de Caen, CRISCO – EA4255
[4] CNRS, Sorbonne Nouvelle, LPP – UMR7018
[5] CNRS, Université Paris-Saclay, LISN – UMR9015

{colleen.beaumard, vincent.martin, jean-luc.rouas}@labri.fr, yaru.wu@unicaen.fr,
pierre.philip@u-bordeaux.fr

Résumé

La somnolence excessive est associée à diverses maladies et impacte la vie quotidienne et professionnelle. L'utilisation de la voix en conditions écologiques peut aider les médecins à la surveiller. La détection du schwa, voyelle optionnelle, par un système de Reconnaissance Automatique de la Parole peut être liée à la somnolence. Nous avons annoté un corpus de parole lue en français par des patients hypersomniaques et appliqué un système de Reconnaissance Automatique de la Parole sur les échantillons audio.

Mots-clés

Somnolence, schwa, Reconnaissance Automatique de la Parole

Abstract

Excessive sleepiness is a symptom associated with several diseases and impacts daily and professional life. Voice recordings collected in ecological conditions can help clinicians to monitor it. Schwa being optional, its detection by an Automatic Speech Recognition system can be related to sleepiness. We manually annotated a read-out-loud French corpus from hypersomniac patients and applied an Automatic Speech Recognition system to these audio samples.

Keywords

Sleepiness, schwa, Automatic Speech Recognition

1 Introduction

1.1 Context

Excessive daytime sleepiness is a symptom that negatively affects both daily and professional life [2] by being associated with several diseases of different origins (neurological, sleep, cardiovascular, etc.). Since it is an important factor of accidental risk [4], excessive daytime sleepiness increases the risk of disability and mortality [11] of hypersomniac patients. Thus, clinicians need a tool to collect symptoms regularly in ecological conditions, which is possible through Ecological Momentary Assessment (EMA). Besides the traditional questionnaires, voice is studied to monitor sleepiness thanks to its low-cost measurement (e.g. French application KANOPÉE [22]). Furthermore, voice data collection can be done passively via the smartphone and allows recording of read or spontaneous speech. According to Statista, 86% of the people possess a smartphone in 2023[1], which shows the potential to use the smartphone to monitor sleepiness through voice.

1.2 Previous work

Several corpora focus on automatic sleep detection through voice: the Sleepy Language Corpus (SLC), the SLEEP corpus (both described in [19, 15]), the Voiceome dataset [25] and the Multiple Sleep Latency corpus [18, 19].

The first two corpora use a measure of sleepiness not medically validated for their annotation while the third uses another measure that cannot distinguish between sleepiness and fatigue [12]. For these reasons, we have decided not to consider them for our analysis.

The Multiple Sleep Latency corpus (MSLTc) was introduced in 2021 and contains the recordings of read-out-loud texts of 125 French hypersomniac patients annotated using validated tools measuring both subjective and physiological sleepiness. Several features were used to evaluate sleepiness, such as acoustic features [20], reading mistakes [17, 16], reading pauses [13] and errors made by an Automatic Speech Recognition (ASR) system [17]. Martin and al. obtained 74.2% Unweighted Average Recall (UAR) [16] when discriminating between sleepy and non-sleepy patients with a threshold at $\leq$8 minutes and >8 minutes respectively for physiological sleepiness. They concatenated automatic speech recognition errors, acoustic features, reading errors, and reading pauses to achieve this performance.

1.3 Objective

Since clinicians need to monitor sleepiness under ecological conditions, we must focus on spontaneous speech rather

[1] https://www.bankmycell.com/blog/how-many-phones-are-in-the-world

than read speech. There are several differences between these two types of speech and, in particular, at the phonological level, since spontaneous speech is not prepared compared to read speech. Thus, words can be repeated several times, phonemes can be altered or not, hesitations can occur and be short or long, voiced or silent, etc. To our knowledge, no study focused on the phonological level of spontaneous speech in relation to sleepiness.

We have decided to study specifically the French schwa /ə/, described as "a central vowel that can alternate with ∅ without changing the meaning of the word" [6, 7]. For example, *demain* (tomorrow) can be pronounced [dəmɛ̃] or [dmɛ̃] without changing the meaning. Furthermore, schwa is also used as French hesitation *euh* [ə:] in spontaneous speech because of its central articulatory position.

Since schwa is optional, we hypothesize the more patients are sleepy, the more they will pronounce /ə/ because of the additional mental effort made to counteract sleepiness. In a prior study [3], we found that the number of schwas pronounced may be linked to both physiological and subjective sleepiness with the Endymion subcorpus of the MSLTc. However, hand labeling schwa from audio recordings is costly in time and in expertise and needs to be automatized. In addition, due to its central articulatory position, schwa can be inteprated as other vowels, such as /ø/ and /œ/ [8] which make its automatic detection difficult.

This article aims to evaluate an Automatic Speech Recognition (ASR) system on /ə/ to later estimate physiological sleepiness in spontaneous speech by detecting /ə/ inside words or as hesitation. The ASR system must be robust to sleepiness, and detect the correct number of /ə/ no matter if the patient is sleepy or not.

The Multiple Sleep Latency Test corpus and the manual annotation method are described in Section 2. The ASR system is presented in Section 3 and the conclusion is given in Section 4 along with perspective for future work.

2 Method

2.1 Multiple Sleep Latency Test corpus

The Multiple Sleep Latency corpus (MSLTc) was introduced in 2020 by Martin et al. [18, 19] and contains 530 recordings of 106 French hypersomniac patients reading out loud texts extracted from *Le Petit Prince* (A. de Saint-Exupéry). Patients were recorded at the Bordeaux Sleep Clinic (France) and underwent a multiple sleep latency test (MSLT) [1]. MSLT consists of five nap opportunities every 2 hours from 9 am to 5 pm. Sleep latency is the time between the beginning of the test and the sleep onset, and is measured at each nap. Average sleep latency is a reference measure (validated by clinicians) of long-term physiological sleepiness [1, 15]. Before each nap opportunity, patients filled out the Karolinska Sleepiness Scale (KSS), which measures short-term sleepiness. They then read out loud a text extracted from *Le Petit Prince* (A. de Saint-Exupéry), each text from each nap being different. These texts are referenced as #1, #2, #3, #4 and #5. This corpus has been annotated with multiple sleepiness measurements (short-term/long-term, objective/subjective) and was validated by clinicians. MSLTc was specifically designed to differentiate symptomatic profiles in hypersomniac patients, explaining the absence of control speakers. Martin et al. have studied acoustic features [20] and reading performance by automatic analysis of reading errors [17, 16] and reading pauses [13]. We have decided to use the Endymion subcorpus [14], which contains 100 audio recordings from 20 hypersomniac patients because it contains the widest variations of short-term sleepiness.

2.2 Manual annotation

The texts were transcribed into phonemes thanks to our reference, the French *Lexique 3.83* lexicon [21] containing the standard French pronunciation of words. In the case a word present in the text was not transcribed in *Lexique 3.83*, we manually added this word's transcription. We then manually annotated the presence or absence of /ə/ on the audio recordings of the Endymion subcorpus based on the transcribed texts and reported the number of /ə/ pronounced for each speaker on each text. Table 1 refers to the number of phonemes and /ə/ (as well as its proportion) for each and all texts.

Phoneme	#1	#2	#3	#4	#5	All
All	735	726	634	689	734	3518
/ə/	30 4.1%	53 7.3%	42 6.6%	42 6.1%	35 4.8%	202 5.7%
/ø/	4 0.5%	7 1.0%	7 1.1%	7 1.0%	6 0.8%	31 0.9%
/œ/	4 0.5%	3 0.4%	3 0.5%	2 0.3%	0 0.0%	12 0.3%
/ə/+/ø/+/œ/	38 5.2%	63 8.7%	52 8.2%	51 7.4%	41 5.6%	245 7.0%

Table 1: Number and proportion of all phonemes, /ə/, /ø/ and /œ/ for each and all texts. #x: text

3 Automatic Speech Recognition

3.1 Model

We used a chain-based automatic speech recognition system to maintain the time stamp provided by this approach [5]. It is a TDNN-HMM model trained with the LF-MMI objective function. The neural network is based on a subsampled time delay neural network (TDNN) with 7 TDNN layers and 1024 units in each layer. The time stride value is set to 1 in the first three layers, 0 in the fourth layer, and 3 in the final ones. The acoustic model is based on a 40-dimensional high-resolution MFCC vector concatenated with a 100-dimensional i-vector [10] and was trained using the Kaldi toolkit [23] on fine-selected sub-corpora of ESTER 1 and 2 [9] (231 hours augmented with 3-fold and volume perturbation). The whole ASR system achieves a Word Error Rate of 13.7% but has never been evaluated on the phoneme level [5].

The ASR system also contains a lexicon with pronunciation variants from the phoneme dictionary provided by *Laboratoire d'Informatique de l'Université du Mans* (LIUM). In

addition, a 3-gram word language model taking into account the context is implemented. It was trained using SRILM's n-gram counting method [24] with KN discounting, and it was limited to the most 50K most frequent words (in the training texts and the dictionary) when trained on ESTER corpora. We can choose to detect only phonemes (phoneme-based ASR) or to use in addition the lexicon and the language model implemented to directly output words (word-based ASR). Section 3.2 presents the results obtained with the phoneme-based ASR system and Section 3.3 the results obtained with the word-based ASR system. Figure 1 schematizes the ASR system.

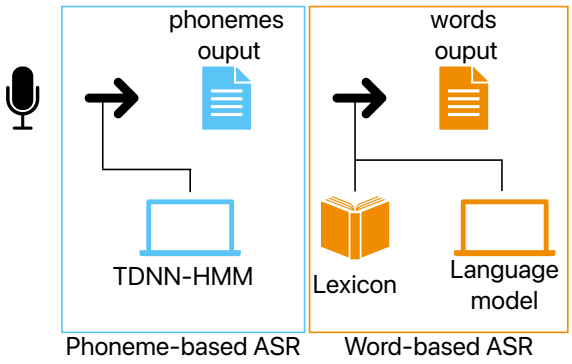

Figure 1: Schema of the ASR system.

3.2 Phoneme-based Automatic Speech Recognition system results

We calculated the mean absolute error (MAE) and the root mean squared error (RMSE) normalized by the number of /ə/ annotated for each and all texts to compare the number of /ə/ annotated on each recording and those detected by our phoneme-based ASR system. The lower these values, the better our phoneme-based ASR system because it is closer to the number of /ə/ annotated. These metrics are reported in Table 2.

Phoneme	Metric	#1	#2	#3	#4	#5	All
/ə/	%MAE	29.3	11.2	28.4	33.5	**50.2**	**28.3**
	%RMSE	35.7	13.5	30.5	36.2	**52.0**	**32.4**
/ø/	%MAE	50.0	14.5	44.8	42.6	**52.5**	**40.0**
	%RMSE	50.0	14.5	46.3	44.1	**42.5**	**44.6**
/œ/	%MAE	2.6	33.3	3.2	**20.0**	-	**12.5**
	%RMSE	5.1	36.7	9.7	**30.0**	-	**25**
/ə/+ /ø/+ /œ/	%MAE	17.5	9.6	16.5	21.5	**28.0**	**17.8**
	%RMSE	23.6	11.4	18.6	23.6	**30.7**	**20.9**

Table 2: %MAE and %RMSE for /ə/, /ø/, /œ/ and /ə/+/ø/+/œ/ on each and all texts with phoneme-based ASR system. *#x*: text.

When considering all texts, %MAE and %RMSE of /ə/ are high (28.3% and 32.4%, respectively), meaning the phoneme-based ASR system performed poorly regarding manual annotation. Furthermore, %MAE and %RMSE fluctuate according to the texts: the best performance is obtained on text #2 (11.2% and 13.5%, respectively), while the worst is obtained on text #5 (50.2% and 52.0%, respectively). The phoneme-based ASR system detects more /ə/ than the annotated number of occurrences.

Extension to other phonemes. We performed quantitative analysis on the main errors between the phoneme-based ASR system and *Lexique 3.83* for schwa and found /ə/ was replaced either by /ø/ (***deux***, two) or by /œ/ (***oeuf***, egg), /ə/ being often realized as [ø] or [œ] phonetically [8]. Indeed, some transcriptions from the phoneme-based ASR system could not correspond to the transcription of standard French included in our reference. For example, the phoneme-based ASR system produced the word *premier* (first) as [pʁəmje], while the standard pronunciation included in *Lexique 3.83* is /pʁømje/. The phoneme-based ASR system does not have any lexicon at the phoneme level, only an acoustic model. To remove errors due to the mismatch between these three phonemes, we therefore extended the considered phonemes to /ə/, /ø/, /œ/ and their combination (/ə/+/ø/+/œ/). Table 1 contains the number of their individual and combined occurrences in each and all texts.

The %MAE and %RMSE of /ø/ are high in every and all texts (all texts: 40.0% and 44.6%, respectively) while those of /œ/ are lower (all texts: 12.5% and 25%, respectively). Text #5 obtained the worst performance from the phoneme-based ASR system for /ø/ (%MAE=52.5%; %RMSE=42.5%), like /ə/, while it is text #4 for /œ/ (%MAE=20.0%; %RMSE=30.0%) since there are no /œ/ in text #5. However, these results should be taken cautiously concerning /ø/ and /œ/ regarding their low number of occurrences in each text.

When combining all three phonemes, the %MAE and %RMSE reached represent respectively 17.8% and 20.9% of the avg. ground truth for all texts, which is slightly better than those obtained individually on /ə/, /ø/, and /œ/ (except for the %MAE of /œ/). In addition, as observed earlier, the performances of the phoneme-based ASR system on these phonemes are inequal across texts, following the %MAE and %RMSE of /ə/ due to its high number of occurrences in each text.

Statistical analysis. To identify if there are factors affecting the detection errors of /ə/, /ø/, /œ/, by this system, we processed a multivariate ANOVA with repeated measurements on absolute errors considering each recording from each speaker. The factors considered are the texts, MSLT (long-term physiological sleepiness), and KSS (short-term sleepiness). The results are presented in Table 3.

No significant effect of either texts, long-term sleepiness or short-term sleepiness on inter-speaker variations of absolute errors in /ə/, /ø/, /œ/ detection was found. This is however the case for intra-speaker variations: texts have a great effect on phonemes detection (individually and combined) by the phoneme-based ASR system. It can be explained by the fact that each text is different, and so on by the difference of /ə/, /ø/, and /œ/ number of occurrences for each text

Factor	/ə/	/ø/	/œ/	e
Texts	*** 1.84e-11	*** <2e-16	*** 1.43e-12	*** 6.77e-6
Texts:MSLT	-	* 3.58e-2	* 4.22e-2	-
Texts:KSS	-	*** 2.06e-4	-	-

Table 3: Results of the multivariate ANOVA with repeated measurements for intra-speaker variations (no significant effect on inter-speaker variations) with phoneme-based ASR system. td.: -: no significant effect; .: $p < 0.1$; ***: $p < .001$

(Table 1). In addition to texts, both /ø/ and /œ/ are lightly affected by the combination of texts and MSLT, which corresponds to long-term physiological sleepiness. /ø/ is also highly affected by the combination of texts and KSS, which corresponds to short-term sleepiness.
Since there is an effect of long-term physiological (MSLT) and short-term sleepiness (KSS) on errors, the performances of the phoneme-based ASR are not robust enough to sleepiness for the detection of /ə/, /ø/, and /œ/. To try to improve their detection and remove the effect of short and long-term physiological sleepiness, we used the word-based ASR system, which benefits from a language model and a lexicon compared to the previous system. The words were transcribed into phonemes using our reference (*Lexique 3.83* which contains only the French standard pronunciation). If a word was not present in *Lexique 3.83*, we manually transcribed it.

3.3 Word-based Automatic Speech Recognition system results

The same metrics as in Section 3.2 (%MAE and %RMSE) were calculated for the word-based ASR system and are reported in Table 4.

Phoneme	Metric	#1	#2	#3	#4	#5	All
/ə/	%MAE	8.9	6.4	11.5	10.2	**25.3**	**11.4**
	%RMSE	12.9	9.5	13.4	13.7	**30.2**	**15.8**
/ø/	%MAE	0.4	**17.1**	16.7	7.4	11.3	**10.8**
	%RMSE	0.1	17.1	**21.2**	10.3	15	**16.9**
/œ/	%MAE	2.6	6.7	6.5	**25.0**	-	**8.3**
	%RMSE	7.7	16.7	12.9	**35.0**	-	**20.8**
/ə/+ /ø/+ /œ/	%MAE	7.5	5.9	7.4	7.8	**18.8**	**8.9**
	%RMSE	10.6	8.8	9.3	11.7	**22.4**	**12.4**

Table 4: %MAE and %RMSE for /ə/, /ø/, and /œ/ for each and all text with word-based ASR system. #x : text.

The %MAE and %RMSE of all phonemes and their combination have globally greatly decreased, indicating the word-based ASR system performed better than the phoneme-based ASR system for the detection of /ə/, /ø/ and /œ/ and their combination.
/ə/ detection on all texts improved from 28.3% for %MAE and 32.4% for %RMSE, to 11.4% and 15.8%, respectively. These metrics improved for each text as well: text #5 is still the text with the worst performance (%MAE=25.3%, %RMSE=30.2%) while text #2 has the best performance (%MAE=6.4%, %RMSE=9.5%). As a recall, these texts had an %MAE of 50.2% and 11.2%, and an %RMSE of 52.0% and 13.5%, respectively, with the phoneme-based ASR.
/ø/, /œ/, and the combination of the three phonemes detection improved as well for all texts. Text #5 is no longer the text with the worst performance for /ø/, it is text #2 for %MAE (17.1%) and text #3 for %RMSE (21.2%) while it is still text #4 for /œ/ (%MAE=25.0%, %RMSE=35.0%) and text #5 for their combination (%MAE=18.8%, %RMSE=22.4%).
However, some %MAE and %RMSE were degraded for some texts. It is the case for example for the %MAE and %RMSE of /œ/ for text #4: %MAE is equal to 25.0% and %RMSE to 35.0% while they were equal to 20.0% and 30.0%, respectively with the phoneme-based ASR system.
To resume, even if there are some degraded %MAE and %RMSE for specific phonemes and texts, the performance of the word-based ASR system has improved compared to the phoneme-based ASR system: the detected number of /ə/, /ø/ and /œ/ is closer to the manual annotation than before.

Statistical analysis. As in Section 3.2, we performed a multivariate ANOVA with repeated measurements for all phonemes considered and their combination with the same factors. The results are presented in Table 5.

Factor	/ə/	/ø/	/œ/	e
Texts	*** 6.86e-6	*** 8.54e-12	-	*** 1.35e-3
Texts:MSLT	-	-	-	-
Texts:KSS	-	-	-	-

Table 5: Results of the multivariate ANOVA with repeated measurements for intra-speaker variations (no significant effect on inter-speaker variations) with word-based ASR. td.: -: no significant effect; .: $p < 0.1$; ***: $p < .001$

Like before, no significant effect of texts, short-term sleepiness (KSS), or long-term physiological sleepiness (MSLTc) was found on inter-speaker variations for the detection of /ə/, /ø/ and /œ/, as well for their combination. In terms of intra-speaker variations, texts highly correlate with /ə/, /œ/, and their combination, but not anymore with /ø/. However, there is no effect of short-term sleepiness and long-term physiological sleepiness on the detection of these three phonemes and their combination, which means that the word-based ASR system is more robust to sleepiness than the phoneme-based ASR system. The number of /ə/, /ø/, and /œ/ detected is no more affected by sleepiness and only depends on the texts (except for /œ/).

4 Conclusion

Our goal was to evaluate an ASR system for the detection of /ə/, /ø/, /œ/ and their combination to later evaluate physi-

ological sleepiness in spontaneous speech. The word-based ASR system performed greatly for each phoneme and their combination, with no effect of either short-term sleepiness or long-term physiological sleepiness for their detection. The next steps are first to apply the word-based ASR system to a spontaneous speech corpus annotated with sleepiness measures validated by clinicians, and second to evaluate if the number of /ə/, /ø/, and /œ/ is sufficient to estimate sleepiness. If not, analyzing the acoustic and temporal properties of these phonemes in addition to their number is considered.
In the future, we plan to analyze the duration of these three phonemes and add two other phonemes (/e/ and /ɛ/) to search for a potential correlation with sleepiness. Moreover, studying more deeply the link between each category of phoneme (stop consonants, fricatives, etc.) and sleepiness is envisaged.

Acknowledgments

This project has received financial support from the CNRS through the MITI interdisciplinary programs.

References

[1] Donna Arand, et al. The Clinical Use of the MSLT and MWT. *Sleep*, 28(1):123–144, 2005.

[2] Christopher M. Barnes et al. Why healthy sleep is good for business. *Sleep Med. Rev.*, 47:112–118, 2019.

[3] Colleen Beaumard, et al. Somnolence et schwas : La somnolence influence-t-elle la production des schwas chez les patients hypersomniaques ? In *9èmes Journées de Phonétique Clinique (poster)*, 2023.

[4] Stéphanie Bioulac, et al. Risk of Motor Vehicle Accidents Related to Sleepiness at the Wheel: A Systematic Review and Meta-Analysis. *Sleep*, 40(10), 2017.

[5] Florian Boyer et al. End-to-End Speech Recognition: A review for the French Language. 2019.

[6] Audrey Bürki, et al. What affects the presence versus absence of schwa and its duration: A corpus analysis of French connected speech. *J. Acoust. Soc. Am.*, 130(6):3980–3991, 2011.

[7] Jacques Durand. À la recherche du schwa : données, méthodes et théories. *SHS Web of Conferences*, 8:23–43, 2014.

[8] Cécile Fougeron, et al. On the acoustic characteristics of French schwa. Saarbrücken, 2007.

[9] Sylvain Galliano, et al. The ester 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Interspeech 2009*, pages 2583–2586, 2009.

[10] Vishwa Gupta, et al. I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription. In *ICASSP*, pages 6334–6338, 2014.

[11] Maki Jike, et al. Long sleep duration and health outcomes: A systematic review, meta-analysis and meta-regression. *Sleep Med. Rev.*, 39:25–36, 2018.

[12] Alistair W. Maclean, et al. Psychometric evaluation of the Stanford Sleepiness Scale. *J. Sleep Res.*, 1(1):35–39, 1992.

[13] Vincent P. Martin, et al. Does sleepiness influence reading pauses in hypersomniac patients? pages 62–66, 2022.

[14] Vincent P. Martin, et al. Physiological vs. Subjective sleepiness: what can human hearing estimate better? Insights from the French Endymion study. In *ICPhS 2023*, 2023.

[15] Vincent P. Martin, et al. Sleepiness in adults: An umbrella review of a complex construct. *Sleep Med. Rev.*, 67:101718, 2023.

[16] Vincent P. Martin, et al. Automatic Speech Recognition systems errors for accident-prone sleepiness detection through voice. In *EUSIPCO*, pages 541–545, 2021.

[17] Vincent P. Martin, et al. Automatic Speech Recognition Systems Errors for Objective Sleepiness Detection Through Voice. In *Interspeech 2021*, pages 2476–2480, 2021.

[18] Vincent P Martin, et al. The Objective and Subjective Sleepiness Voice Corpora. *LREC*, 2020.

[19] Vincent P. Martin, et al. How to Design a Relevant Corpus for Sleepiness Detection Through Voice? *Front. digit. health.*, 3:686068, 2021.

[20] Vincent P Martin, et al. Détection de la somnolence dans la voix: nouveaux marqueurs et nouvelles stratégies. 2020.

[21] Boris New, et al. Lexique 2 : A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3):516–524, 2004.

[22] Pierre Philip, et al. Smartphone-Based Virtual Agents to Help Individuals With Sleep Concerns During COVID-19 Confinement: Feasibility Study. *JMIR*, 22(12):e24268, 2020.

[23] Daniel Povey, et al. The Kaldi Speech Recognition Toolkit.

[24] Andreas Stolcke. SRILM - an extensible language modeling toolkit. In *7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 901–904, 2002.

[25] Bang Tran, et al. Speech Tasks Relevant to Sleepiness Determined With Deep Transfer Learning. In *ICASSP*, pages 6937–6941, 2022.