Check for updates

Automatic Assessment of Signs of Alcohol Dependency Syndrome from Spontaneous Speech

Gábor Gosztolya[1,2](✉), András Bence Lázár[3], Ildikó Hoffmann[3,4], Otília Bagi[3], Fruzsina Fanni Farkas[3], Janka Gajdics[3], László Tóth[2], and János Kálmán[3]

[1] HUN-REN–SZTE Research Group on Artificial Intelligence, Szeged, Hungary
ggabor@inf.u-szeged.hu
[2] University of Szeged, Institute of Informatics, Szeged, Hungary
[3] University of Szeged, Department of Psychiatry, Szeged, Hungary
[4] HUN-REN Research Center for Linguistics, Budapest, Hungary

Abstract. Alcohol is a progressive central nervous system depressant. Increased alcohol consumption leads to alterations in cognitive processes and also affects speech production. In this study we present a corpus of n=35 patients diagnosed with Alcohol Dependency Syndrome (ADS) and n=35 matched healthy controls, and attempt to automatically distinguish the two speaker groups based on their spontaneous speech. By using wav2vec 2.0 embeddings as features, we were able to identify the two speaker categories with quite high accuracy (EER scores between 9% and 20%, and AUC scores above 0.885). We also sought to find the difference between the two speech tasks (a general spontaneous task and an alcohol-related one) performed by the subjects. Lastly, we analyzed the amount of pauses present in the speech of the subjects. Based on our results, even three simple pause-related attributes are sufficient for the automatic identification of the ADS subjects with an acceptable performance for both speech tasks.

Keywords: Alcohol dependency syndrome · Pathological speech processing · Human-computer interaction · wav2vec 2.0

1 Introduction

Alcohol, as a progressive central nervous system depressant, can cause changes in cognitive functions. The prevalence of cognitive impairment in alcohol use disorder is 40%. Alcohol dependence can affect executive functions, may impair verbal fluency, working memory, recent memory, visuospatial abilities, visual recognition and processing speed [7,9,14]. Changes in cognitive functions can also affect language processes, speech perception and production [32]. There have been several studies on how alcohol dependence affects cognitive abilities, but only a few studies examine the effects of long-term alcohol use on language processes (e.g. [16,25,26]).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15300, pp. 18–29, 2025.
https://doi.org/10.1007/978-3-031-78014-1_2

From the aspect of automatic speech processing, investigating the effect of alcohol on speech production is not an uncommon topic. Several studies investigated the effects of alcohol intoxication, to distinguish speakers who are under the influence of alcohol from sober subjects [3,17,23,32]. What is common among these studies is that they focused on the short-term influence of alcohol. This is understandable as detecting whether the speaker is under the short-term effects of alcohol (i.e. is the speaker drunk?) has straightforward applications like denying the speaker to drive a motorized vehicle. However, to the best of our knowledge, no study has attempted to automatically detect the long-term effects of alcohol consumption on speech.

In this study we focus on the spontaneous speech production of patients suffering from Alcohol Dependency Syndrome (ADS) by performing machine learning experiments. In such experiments in the pathological speech processing area, due to the scarcity of data and the fact that each subject corresponds to one machine learning example, it is still common to employ traditional classification methods like a Support Vector Machine (SVM) instead of end-to-end deep neural network (DNN) systems [22]. This means that the feature extraction and classification steps are distinct, but the type of features used is not a trivial question. One still can find studies employing hand-crafted attributes (e.g. [8,13,22]), but utilizing general-purpose features like embeddings of deep learning models is becoming ever more common [15,27,31]. Due to this, we shall employ the embeddings of wav2vec 2.0 self-supervised models [2] as features.

The contribution of our study is four-fold: i) we present a corpus containing the speech of 35 patients with diagnosed Alcohol Dependency Syndrome and 35 matched healthy controls (HC); ii) we carry out machine learning experiments to identify two speaker groups based on the spontaneous speech production of the subjects; iii) we try to distinguish the two speech tasks by machine learning; iv) we investigate the extent of pauses present in the speech of the subjects, and compare the tendencies with the results of the classification experiments.

2 Data

Inpatients (n=35) admitted with a diagnosis of alcohol dependency syndrome (F.10.20) with Mini-Mental State Examination (MMSE) score of at least 28 were recruited at the Department of Psychiatry, University of Szeged, Hungary between July, 2022 and June, 2023. Patients suffering from alcohol withdrawal syndrome (CIWA $>$ 7) and with cognitive impairment (MMSE $<$ 28) were excluded. Recording the speech was performed on the eighth day of abstinence. In addition, the speech of demographically statistically matched healthy control subjects (n=35) was also recorded. The research was authorized by the Ethical Committe of the University of Szeged, Hungary. Collection of data was done in accordance with the Declaration of Helsinki.

For each subject (both for ADS and HC subjects), two speech recordings were collected: in the first one, as an emotionally neutral topic, they were asked to describe their previous day (task *Previous day*). Afterwards, as an alcohol-related topic, they had to describe their relationship to alcohol and situations

Table 1. Average durations and corresponding standard deviation values for the recordings (mean ± std), expressed in seconds.

	Speaker Groups	
Speech task	ADS	HC
Previous day	76.7 ± 45.0	84.1 ± 33.3
Alcohol-related	80.9 ± 45.8	86.4 ± 31.4

where they found it hard to resist to drink (task *Alcohol-related*). Recording was done in a clinical environment, by using a digital dictaphone; the recordings were later converted to 16 kHz mono format for digital processing.

The means and the standard deviations of the recording durations can be seen in Table 1. The responses of the subjects were around 1.5 min on average for both speech tasks and both speaker groups, although the HC subjects produced somewhat more speech on average (+8 and +6 s, *Previous day* and *Alcohol-related* speech tasks, respectively). Furthermore, there were significant individual variations, reflected by the high standard deviation values. The standard deviation of the durations of the speech produced by the HC speakers turned out to be notably lower than that of the responses of the ADS patients.

3 Methods

3.1 Wav2vec 2.0

wav2vec is a convolutional neural network (CNN) designed to process raw audio signals as input and generate representations suitable for automatic speech recognition (ASR) systems. The model is trained in a self-supervised manner, during which it learns to predict future observations for the given speech sample [24]. This self-supervised training allows the model to be pre-trained on large, unannotated corpora, enabling subsequent fine-tuning for specific audio processing tasks such as ASR for low-resource languages [19] or paralinguistic applications (e.g. emotion detection [20]). The **wav2vec 2.0** architecture further enhances this approach by incorporating masking during training. Specifically, raw audio is encoded using a block of convolutional neural networks, and small segments of the resulting latent speech representations are masked, akin to masked language modeling. These masked representations are then processed by a quantizer, which selects speech units from an inventory of learned units, and a transformer network, which incorporates information from the entire utterance [2]. Figure 1 shows the layout of the (fine-tuned) wav2vec 2.0 structure.

3.2 Wav2vec 2.0 for Feature Extraction

The outputs from the multi-layer convolutional block are the sequence of extracted feature vectors of the last convolutional layer, while the outputs from

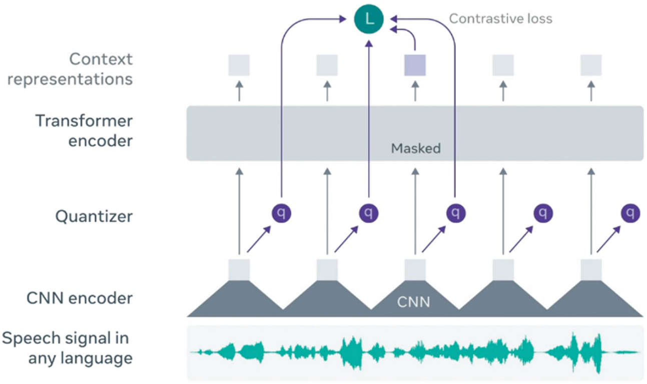

Fig. 1. The fine-tuned wav2vec 2.0 framework structure.

the second (fine-tuned) block comprise the sequence of the hidden states of the last layer of the block. These two types of feature vectors may carry relevant information for a large range of speech processing tasks: the former vector can be expected to capture lower-level information (e.g. pause-related information), while the fine-tuned layer can be expected to store phonetic-related information; so they could be used as features [10]. However, the number of these (frame-level) feature vectors is proportional to the length of the utterance. To employ them as *utterance-level* features, they have to be aggregated over the whole recording. To do this, taking the mean and/or the standard deviation of the values over the whole utterance is a generally accepted solution [11,21,30].

4 Experimental Setup

4.1 Feature Extraction

We used the wav2vec 2.0 model `wav2vec2-large-xlsr53-hungarian`. The base of this model is the `XLSR-53` model pre-trained by Facebook on the audio data of 53 languages simultaneously [1]. This base model was then fine-tuned by the user `jonatasgrosman` [12] on the Hungarian part of the Mozilla Common Voice 6.1 corpus (8 h). The last layer of the convolutional block of this model consists of 512 neurons, while the last layer of the fine-tuned block has 1024 neurons. By using mean and standard deviation of these frame-level embedding vectors, we obtained 1024 and 2048 utterance-level features, convolutional and fine-tuned embeddings, respectively.

4.2 Utterance-Level Classification

We employed the approach common in pathological speech processing studies (e.g. [4,13,28]): due to the low number of examples (subjects) from a machine learning perspective, we did not define separate training, development and test

Table 2. Equal Error Rate (EER) and Area Under the ROC Curve (AUC) values obtained when discriminating the ADS and HC subjects for the two speech tasks.

Speech task	Embedding	EER	AUC
Previous day	Convolutional	11.4%	0.947
	Fine-tuned	20.0%	0.885
Alcohol-related	Convolutional	16.6%	0.906
	Fine-tuned	9.1%	0.982

sets, but used cross-validation. Each fold consisted of the data of one ADS and one HC speaker, leading to 35 folds overall. Classification performance was measured via Equal Error Rate (EER) and the Area Under the ROC Curve (AUC) metrics, also commonly applied in pathological speech processing studies [4].

We applied Support Vector Machines for classification, using the LibSVM [6] library. We employed the nu-SVM method with a linear kernel; the value of C was tested in the range $10^{\{-5,\ldots,1\}}$. The optimal value for the C meta-parameter was determined by the technique called *nested cross-validation* [5]: for the 2×34=68 speakers being in the training fold in the actual CV step, we performed *another* cross-validation. We chose the C value which led to the highest AUC score in this “inner” cross-validation loop; “final” SVM model was trained on the data of the 68 speakers with this C value, and it was evaluated on the data of the last fold (i.e. two speakers). With this procedure we sought to avoid the bias in our scores which would have been present if we used standard cross-validation.

To reduce the random factor unavoidably present in our workflow, we repeated each classification experiment 5 times, using a different random seed value when constructing the folds for cross-validation. In the results, we always report the mean of the five EER and AUC scores. When comparing the significance of differences, we employed the Mann-Whitney U test (see [18], also known as the Wilcoxon rank-sum test).

5 Classification Results

Table 2 shows the EER and AUC values measured, averaged over the five runs, for the two speech tasks and the two embedding types. Overall, the results are quite good: the Equal Error Rate scores were between 9.1% and 20.0%, while the AUC values lay in the range 0.885 . . . 0.982. This, in our opinion, indicates that massive alcohol consumption, i.e. Alcohol Dependency Syndrome affects the speech production of the subjects significantly, and in a way which can be detected by automatic means. Regarding the usefulness of the speech tasks, it is hard to see any general trend, as we obtained better classification results for the *Previous day* speech task when using the convolutional embeddings ($p < 0.01$), and for the *Alcohol-related* speech task using the fine-tuned embeddings (again $p < 0.01$). This might indicate that changes in speech production are mostly not specific to these two speech tasks, but more general in nature such as tone, pronunciation, pausing patterns or speech rate. Still, the scores for the *Alcohol-related* task are somewhat better on the average, which might indicate that asking the subjects about their experiences with alcohol is somewhat better suited to reveal the differences in speech production of ADS patients compared to healthy controls.

Table 3. Equal Error Rate (EER) and Area Under the ROC Curve (AUC) values obtained when discriminating the two speech tasks.

Subjects	Embedding	EER	AUC
ADS	Convolutional	39.4%	0.605
	Fine-tuned	43.4%	0.576
HC	Convolutional	16.6%	0.892
	Fine-tuned	14.9%	0.893
ADS + HC	Convolutional	31.7%	0.699
	Fine-tuned	22.3%	0.829

5.1 Distinguishing the Two Speech Tasks

Although there was no great difference between the two speech tasks in classification performance (when we sought to distinguish the ADS subjects from the HC ones), we were also interested in how much the speech produced by the subjects differed in the two tasks. To this end, first we performed further binary classification experiments, where the classifiers were trained to distinguish the two speech tasks of the subjects. The experimental setup of these experiments mirrored those of our previous experiments: we utilized SVM with a linear kernel in a nested cross-validation setup, using the mean and standard deviation of (frame-level) wav2vec 2.0 embeddings as features. To avoid any further bias, each fold consisted of the two utterances of one speaker (so no SVM model was trained and evaluated on the speech of the same speaker).

The results obtained in this set of experiments can be seen in Table 3. For the ADS subjects, classification performance was above that of random guessing, but the results are quite low: we measured EER values of 39.4% and 43.4%, and AUC scores of 0.605 and 0.576, convolutional and fine-tuned embeddings, respectively (the AUC values displayed no statistically significant difference, although for the EER metric scores we measured $p = 0.0397$). This probably indicates that there was not much difference in the speech of the ADS subjects in the two tasks, at least for properties that the wav2vec 2.0 embeddings captured. In contrast, for the healthy control speakers, the two speech tasks proved to be markedly different, judging from the low EER (16.6% and 14.9%) and quite high AUC (0.892 and 0.893) values. (In these cases there was no significant difference between the two embedding types.) As might be expected, when we used all 70

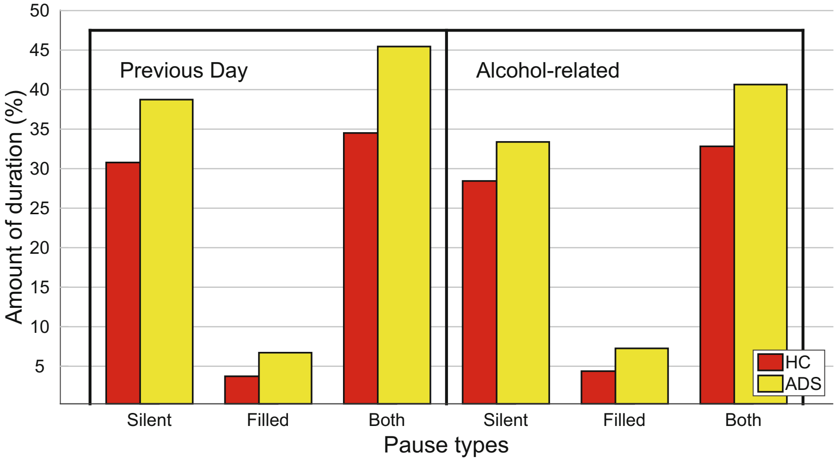

Fig. 2. The amount (percentage of duration) of silent and filled pauses for the two speaker categories and speech tasks (determined automatically by phone-level ASR).

subjects (i.e. all 140 utterances), the metric scores were in-between; in this case, the fine-tuned embeddings were significantly better ($p < 0.01$ for both metrics).

6 Pause Analysis

6.1 Classification Experiments Using the Amount of Pauses

As the last contribution of our study, we investigated the amount of pause present in the speech of the two speaker groups for the two speech tasks. For this, two pause types were distinguished: silent pause (the absence of speech for at least 30ms) and filled pause (vocalizations such as ‘er’, ‘um’ etc.). Pause identification was achieved using a standard HMM/DNN hybrid ASR system by performing phone-level recognition; the acoustic DNN model was trained on 240 h of noise-augmented Hungarian spontaneous speech. We treated a filled pause as a special phone, while several labels (breath intakes, sighs and gasps) were handled together with silent pauses. The amount of silent and filled pauses was calculated from the phonetic output of the ASR system as the total duration of the corresponding pause type divided by the duration of the utterance (and we calculated the third attribute corresponding to all pause occurrences as the sum of the amount of silent and filled pauses). For more details about this process, see [29].

The amount of pauses for the two speaker types and for the two speech tasks can be seen in Fig. 2. Perhaps the most obvious observation is that the ADS subjects produced more pauses on average than the healthy controls, which was the case for both pause types and both speech tasks (although the difference was quite small for filled pauses). Regarding the two speech tasks, the (mean) amount

Table 4. Equal Error Rate (EER) and Area Under the ROC Curve (AUC) values obtained when discriminating the ADS and HC subjects for the two speech tasks (up), and when discriminating the two speech tasks (down), using only the pause-related attributes.

Classification task	Data	EER	AUC
ADS vs. HC	Previous day	21.1%	0.826
	Alcohol-related	33.1%	0.730
Previous day vs. Alcohol-related	ADS	57.4%	0.409
	HC	53.4%	0.431
	ADS + HC	55.0%	0.497

of filled pause was practically the same both for the ADS patients and the HC subjects. In contrast, there were clearly more silent pauses present in the *Previous day* recordings of the ADS subjects than for their *Alcohol-related* responses. This is in sharp contrast with the contents of Table 3, where the responses of the HC subjects could be identified with a good classification performance, while the two speech recordings of the ADS speakers proved to be almost indistinguishable. This, in our opinion, indicates that the speech properties which are represented by the (means and standard deviations of) wav2vec 2.0 embeddings (be they convolutional or fine-tuned) are quite different from the amount of pauses.

Lastly, we performed classification experiments using only the three pause-related attributes. Since the experimental setup again was the same as that of our previous machine learning experiments, the EER and AUC scores presented next are directly comparable to those reported in Tables 2 and 3.

The two speaker groups could be identified with a surprisingly high efficiency (see the top half of Table 4), particularly for the *Previous day* speech task: the mean EER value of 21.1% and the mean AUC score of 0.826 fall quite close to the scores obtained with the fine-tuned embeddings (see Table 2). The statistical tests show no significant difference between the EER scores for the two feature sets ($p = 0.444$), although the AUC scores of the pause-related attributes are significantly ($p < 0.01$) worse. Compared to the convolutional embeddings, the performance gap is wider, just as for the *Alcohol-related* speech task with both embedding types. (The difference is statistically significant for all three cases and for both evaluation metrics.) According to the bottom half of Table 4, however, our classifier models were unable to distinguish the two speech tasks: the EER values exceeded 50% and the AUC scores were below 0.500, showing lower-than-chance level performance.

The tendencies of these scores is in accordance with Fig. 2. Clearly, the ADS subjects tend to produce more silent pauses for both speech tasks, which allowed an acceptable classification performance even when we used only the three pause statistic values. The difference between the amount of silent pauses produced by the two speaker groups is more apparent in the *Previous day* speech task (see Fig. 2 again), and indeed, the AUC value is higher and the EER value is lower for this speech task than those for the *Alcohol-related* one (see Table 3). However, although this difference allowed a better distinction of the two speaker groups for the *Previous day* speech tasks, it was not enough to adequately characterize the individual speech tasks. This might especially hold for the control subjects, where the difference between the amount of pauses between the two speech tasks was small. (Also note that Fig. 2 shows only the average values and does not reflect any individual variation.)

7 Conclusion and Discussion

In this study we investigated the spontaneous speech of subjects suffering from Alcohol Dependency Syndrome (ADS). We presented a corpus consisting of n=35 ADS patients and n=35 matched healthy controls, containing the recordings of two spontaneous speech tasks for each subject. In the first part of our study, we performed classification experiments in order to identify the speaker groups, using wav2vec 2.0 embeddings as features. We also investigated whether the two *speech tasks* can be distinguished, and found that they were quite different for the healthy control speakers, but differed only slightly for the ADS subjects (at least from the aspects which were represented by the wav2vec 2.0 embeddings). In the second part of our study we calculated the amount of (both silent and filled) pauses for each recording by a simple phone-level ASR system, and verified that the two speaker categories could be automatically identified by using these values alone as features. No difference was found, however, between the two speech tasks, in the sense that the metric scores reflected chance-level classification performance.

Regarding the limitations of our study, perhaps the most apparent one is the number of subjects. Although 35 patients and the same number of (demographically matched) control subjects is not an uncommonly low number in pathological speech investigations, we plan to collect more data to increase the statistical significance of our findings. Another limitation might be the wav2vec 2.0 model used: although wav2vec 2.0 is considered to be a competitive feature extractor, and the actual model was fine-tuned for Hungarian, the amount of data used for fine-tuning was limited (only 8 h). Also, the XLSR-53 network architecture contains 24 layers in its contextualized (i.e. fine-tuned) block, from which we used only the embeddings taken from the last layer. Other studies already experimented with utilizing embeddings taken from an inner layer as features (see e.g. [22]). In the near future we plan to repeat our experiments both with self-supervised models fine-tuned on more data and with other types of acoustic features.

Furthermore, to describe the speech production characteristics of patients suffering from Alcohol Dependency Syndrome, it would make sense to employ interpretable attributes, which is not the case for embedding vectors. Although the amount of silent and filled pauses were indeed such interpretable features, and they also allowed satisfactory discrimination of the two subject groups, classification performance was significantly lower than what was obtained via the wav2vec 2.0 embeddings. Therefore, some speech properties of the ADS subjects was not captured by these attributes; finding further meaningful features which characterize the speech production of ADS patients is a straightforward extension of our current investigations.

Acknowledgements. This study was supported by the NRDI Office of the Hungarian Ministry of Innovation and Technology (grant no. TKP2021-NVA-09), and within the framework of the Artificial Intelligence National Laboratory Program (RRF-2.3.1-21-2022-00004). Bence András Lázár was supported by the Géza Hetényi Grant (SZTE-ÁOK-KKA-2024-HG).

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Babu, A., et al.: XLS-R: self-supervised cross-lingual speech representation learning at scale. In: Proceedings of Interspeech, pp. 2278–2282 (2022)
2. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: a framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. **33**, 12449–12460 (2020)
3. Bone, D., Li, M., Black, M.P., Narayanan, S.S.: Intoxicated speech detection: a fusion framework with speaker-normalized hierarchical functionals and GMM supervectors. Comput. Speech Lang. **28**(2), 375–391 (2014). https://doi.org/10.1016/j.csl.2012.09.004
4. Carvajal-Castaño, H.A., Pérez-Toro, P.A., Orozco-Arroyave, J.R.: Classification of Parkinson's disease patients - a deep learning strategy. Electronics **11**(17), 2684 (2022). https://doi.org/10.3390/electronics11172684
5. Cawley, G.C., Talbot, N.L.C.: On over-fitting in model selection and subsequent selection bias in performance evaluation. J. Mach. Learn. Res. **11**(Jul), 2079–2107 (2010)
6. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3), 1–27 (2011)
7. Courtney, K.E., Li, I., Tapert, S.F.: The effect of alcohol use on neuroimaging correlates of cognitive and emotional processing in human adolescence. Neuropsychology **33**(6), 781–794 (2020). https://doi.org/10.1037/neu0000555
8. Egas-López, J.V., et al.: Automatic screening of mild cognitive impairment and Alzheimer's disease by means of posterior-thresholding hesitation representation. Comput. Speech Lang. **75**(Sep), 101377 (2022)
9. Endreddy, A.R., Lakshmi, R.C., Seshamma, V.V.: A prospective study of amelioration of cognitive functions following alcohol abstinence in patients with alcohol dependence syndrome. Arch. Ment. Health **24**(2), 109–114 (2023). https://doi.org/10.4103/amh.amh_145_22
10. Fan, Z., Li, M., Zhou, S., Xu, B.: Exploring wav2vec 2.0 on speaker verification and language identification. In: Proceedings of Interspeech, pp. 1509–1513 (2021)
11. Gosztolya, G., Tóth, L., Svindt, V., Bóna, J., Hoffmann, I.: Using acoustic deep neural network embeddings to detect multiple sclerosis from speech. In: Proceedings of ICASSP, pp. 6927–6931. Singapore (2022)
12. Grosman, J.: Fine-tuned XLSR-53 large model for speech recognition in Hungarian. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-hungarian (2021)

13. Hajduska-Dér, B., Kiss, G., Sztahó, D., Vicsi, K., Simon, L.: The applicability of the beck depression inventory and hamilton depression scale in the automatic recognition of depression based on speech signal processing. Front. Psychiatry **13**, 879896 (2022). https://doi.org/10.3389/fpsyt.2022.879896
14. Kandasamy, V., Abdul Rahuman, M., Ramanujam, G.: A study of cognitive impairment and its neuroimaging correlates in patients with alcohol dependence a cross-sectional study. Asian J. Med. Sci. **14**(10), 781–794 (2023). https://doi.org/10.3126/ajms.v14i10.53860
15. Klumpp, P., et al.: The phonetic footprint of Parkinson's disease. Comput. Speech Lang. **72**(Mar), 101321 (2022)
16. Kung, J., et al.: Semantic and phonemic fluency in alcohol dependent individuals. In: Proceedings of 51st Annual Meeting of INS, pp. 799. San Diego, CA, USA (2023)
17. Laptev, P., Litovkin, S., Kostyuchenko, E.: Determining alcohol intoxication based on speech and neural networks. In: Proceedings of SPECOM, pp. 107–115. Gurugram, India (2023). https://doi.org/10.1007/978-3-031-48309-7_9
18. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. **18**(1), 50–60 (1947)
19. Mihajlik, P., Balog, A., Gráczi, T.E., Kohári, A., Tarján, B., Mády, K.: BEA-Base: a benchmark for ASR of spontaneous Hungarian. In: Proceedings of LREC, pp. 1970–1977. Marseille, France (2022)
20. Pepino, L., Riera, P., Ferrer, L.: Emotion recognition from speech using wav2vec 2.0 embeddings. In: Proceedings of Interspeech, pp. 3400–3404. Brno, Czechia (2021). https://doi.org/10.21437/Interspeech.2021-703
21. Pérez-Toro, P., et al.: Alzheimer's detection from English to Spanish using acoustic and linguistic embeddings. In: Proceedings of Interspeech, pp. 2483–2487 (2022)
22. Pérez-Toro, P.A., et al.: Automatic assessment of Alzheimer's across three languages using speech and language features. In: Proceedings of Interspeech, pp. 1748–1752. Dublin, Ireland (2023). https://doi.org/10.21437/Interspeech.2023-2079
23. Schiel, F., Heinrich, C.: Laying the foundation for in-car alcohol detection by speech. In: Proceedings of Interspeech, pp. 983–986. Brighton, United Kingdom (2009). https://doi.org/10.21437/Interspeech.2009-292
24. Schneider, S., Baevski, A., Collobert, R., Auli, M.: wav2vec: unsupervised pre-training for speech recognition. In: Proceedings of Interspeech, pp. 3465–3469 (2019)
25. Sebold, M., et al.: When habits are dangerous: alcohol expectancies and habitual decision making predict relapse in alcohol dependence. Biol. Psychiat. **82**(11), 847–856 (2017)
26. Stavro1, K., Pelletier, J., Potvin, S.: Widespread and sustained cognitive deficits in alcoholism: a meta-analysis. Addict. Biol. **18**, 203–213 (2012)
27. Sztahó, D., Fejes, A.: Effects of language mismatch in automatic forensic voice comparison using deep learning embeddings. J. Forensic Sci. **88**(3), 871–883 (2023). https://doi.org/10.1111/1556-4029.15250
28. Tóth, L., et al.: Automatic detection of mild cognitive impairment from spontaneous speech using ASR. In: Proceedings of Interspeech, pp. 2694–2698. Dresden, Germany (2015)
29. Tóth, L., et al.: A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. Curr. Alzheimer Res. **15**(2), 130–138 (2018)

30. Vaessen, N., Van Leeuwen, D.A.: Fine-tuning wav2vec2 for speaker recognition. In: Proceedings of ICASSP, pp. 7967–7971 (2021)
31. Wagner, D., et al.: Multi-class detection of pathological speech with latent features: How does it perform on unseen data? In: Proceedings of Interspeech, pp. 2318–2322. Dublin, Ireland (2023). https://doi.org/10.21437/Interspeech.2023-464
32. Wayland, R., Tang, K., Wang, F., Vellozzi, S., Sengupta, R.: Neural networks' posterior probability as measure of effects of alcohol on speech. J. Acoust. Soc. Am. **153**(3), A293 (2023). https://doi.org/10.1121/10.0018898