Investigating the Utility of wav2vec 2.0 Hidden Layers for Detecting Multiple Sclerosis

Gábor Gosztolya[1,2(✉)], László Tóth[2], Veronika Svindt[3], Judit Bóna[4], and Ildikó Hoffmann[3,5]

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

Abstract. Multiple sclerosis (MS) is a chronic autoimmune neurodegenerative disease, affecting the central nervous system. The disease can induce various symptoms, such as adversarily affecting the speech of the subject in various ways, therefore allowing the use of automatic speech analysis for the detection of MS and for monitoring the condition of the patient. Owing to data scarcity, however, deep neural networks are usually not employed for this task as classifiers, but are used as feature extractors. This is the case for self-supervised networks such as wav2vec 2.0 as well, where a straightforward source of embeddings (used as features) are the last layers of the convolutional (lower) and fine-tuned (higher) blocks. In this study we investigate whether extracting the embeddings from some other, inner layer of the fine-tuned (transformer) block can help improve MS detection performance. Tested on two speech tasks, we found that the lowest one-third of the 24 fine-tuned layers proved to be the most suitable for feature extraction, which led to statistically significant improvements in the AUC scores for both speech tasks.

Keywords: Multiple sclerosis · Pathological speech processing · Wav2vec 2.0 · Feature extraction

1 Introduction

Multiple sclerosis (MS) is a chronic autoimmune neurodegenerative disease, affecting the central nervous system, which can result in various cognitive and linguistic impairments of the subjects [29]. The progression of MS may vary considerably from subject to subject, and it can change over time. Several changes may occur as the disease progresses: an increase in disability (affecting walking,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
A. Karpov and V. Delić (Eds.): SPECOM 2024, LNAI 15299, pp. 297–308, 2024.
https://doi.org/10.1007/978-3-031-77961-9_22

balance, coordination, and other physical abilities of the patient); an increase in fatigue; sensory changes (affecting the ability to feel cold, heat and touch) and changes in cognitive and language functions. Noting these points, automatic speech analysis might contribute to detect the disease in an automatic, contact-free and (relatively) cheap way, or serve as a screening technique.

In the past decade, automatic speech analysis has developed into a broad area within speech technology. It includes *computational paralinguistics*, which seeks to automatically identify different speaker traits and states, such as emotion recognition [13,21], speaker age and gender determination [22], assessing the degree of sleepiness [15], whether the speaker has cold [33], or the presence of stuttering [12]. It also includes *pathological speech processing* tasks, where the aim is to automatically decide whether the speaker is suffering from a specific disease such as Parkinson's Disease [17,19], Alzheimer's Disease [16,27], mild cognitive impairment [25] or depression [9,18]. After the deep learning revolution, deep networks also found their way into the pathological speech processing area [9, 17,27].

Nowadays, with the emergence of self-supervised learning, perhaps the most widely-used speech processing network type is wav2vec 2.0 [2]. Besides direct speech processing applications [6,24], evaluating the network on a specific speech utterance and noting the activations of a specific hidden layer (i.e. the *embeddings*) and using these vectors as features (and thus, the whole network as a feature extractor) is a common approach as well [7,26]. Due to the scarcity of resources in the pathological speech processing area, deep networks are rarely used as classifiers there, but they primarily serve as feature extractors [7,27,30].

To employ neural networks (including those with a wav2vec 2.0 architecture) as feature extractors, one has to choose a specific layer to take the embeddings from. A wav2vec 2.0 network has two main blocks: the lower *convolutional* one and the higher *fine-tuned* one, and the straightforward sources of embeddings are the last layers of each block [20]. In this study, however, we investigate whether some inner layer of the fine-tuned block might supply better features. For this, we take a network fine-tuned on the target language (in our case, Hungarian), and test the embeddings taken from all of the inner layers of the fine-tuned block as machine learning features to distinguish multiple sclerosis patients from healthy control (HC) subjects.

2 The Hungarian Multiple Sclerosis Corpus

All the tests were carried out at the Neurology Department of Uzsoki Hospital, Budapest, Hungary, and at the Research Center for Linguistics of the Hungarian Research Network, Budapest, Hungary. The study was approved by the Ethics Committee of the Uzsoki Hospital, and it was conducted in accordance with the Declaration of Helsinki. In the current study we use the recordings of 23 MS subjects (5 males and 18 females) and 22 healthy controls (6 males and 16 females). All 23 MS subjects belonged to the relapsed-remitting MS subtype (RRMS). All the speakers involved in the study were native Hungarian speakers. The MS and

HC groups displayed no statistically significant difference in their demographic attributes (age in years, gender (male / female) and years of education).

The protocol for collecting the speech samples from the subjects was quite extensive, involving 17 different speech tasks. In the current study, due to space limitations, we use the recordings of two spontaneous speech tasks: in the **Opinion** task the subjects were asked to share their opinions about vegetarianism, while in the **Narrative Recall** speech task the subjects listened to a two-minute-long historical anecdote that was unknown to them beforehand, and they had to summarize the story heard as accurately as possible. Although participants have to produce coherent, complex narratives in both tasks, there are some significant differences in the cognitive requirements of these task. Namely, in the Narrative Recall task the speakers had to rely significantly on their working memory, and they had to inhibit irrelevant information, compared to the clearly simpler Opinion speech task.

The recording was performed with a Sony PCM-A10 digital dictaphone using a tie clip microphone with a sampling rate of 48 kHz; later the recordings were converted to 16 kHz mono with a 16 bit resolution.

3 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 [28]. 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 [24] or paralinguistic appications (e.g. emotion detection [26]). 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.

3.1 Wav2vec2 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 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, so they are quite popular as features [8,20]. Of course, these embeddings are at the frame level, so the number of these vectors is proportional to the length of the utterance. To

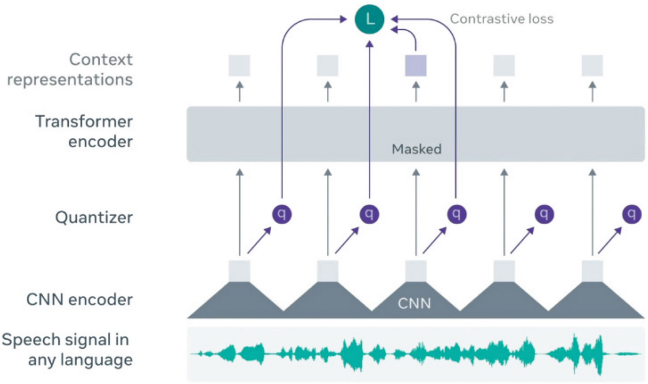

Fig. 1. The fine-tuned wav2vec 2.0 framework structure. Source: https://ai.meta.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio.

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 [10,27,32].

In this study we, however, focus on the *inner* layers of the fine-tuned block. Since in a standard XLSR-53 network (see e.g. [1]), there are 24 such transformer layers, we have 24 options of feature extraction. Besides the standard assumptions that lower-laying layers capture lower-level phenomena (e.g. silence, noise, acoustic conditions), while higher-level layers tend to capture high-level (e.g. phonetic) information, we do not have any further guidance. Due to this, in our experiments we tested the activations taken from all 24 layers for the two speech tasks for multiple sclerosis detection.

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, and it was ensured that the quantization module of the wav2vec 2.0 neural network also delivers multilingual quantized speech units [1]. This base model was then fine-tuned by the user `jonatasgrosman` [11] 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 all the layers of the fine-tuned block have 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. Since we focused on the fine-tuned embeddings, in most of our experiments we had feature vectors with a length of 2048.

4.2 Utterance-Level Classification

We employed the approach common in pathological speech processing studies (e.g. [3,14,31]): due to the low number of examples (subjects) from a machine learning perspective, we did not define separate training, development and test sets, but used cross-validation. Each fold consisted of the data of one MS and one HC speaker, leading to 23 folds overall. Classification performance was measured using the Area Under the ROC Curve (AUC) metric, also commonly applied in pathological speech processing studies [3,10].

We employed Support Vector Machines (SVM) for classification, using the LibSVM [5] 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* [4]: for the speakers of 22 folds in the training subset of the actual CV step, we performed *another* cross-validation. We chose the C value that gave the highest AUC score in this "inner" cross-validation loop; the "final" SVM model was then trained on the data of 22 folds 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 that would have been present if we had used standard cross-validation.

To measure the robustness of the AUC scores, we repeated each classification experiment five times, using a different random seed value when assigning the speakers to specific folds for cross-validation. In the results, we report the mean of the five AUC scores. When inspecting robustness, we calculate the standard deviation (*Std.*) of the five AUC values, and report the range (i.e. $[\min, \max]$) of the scores as well.

5 Results with the Embeddings of the Last Layers

Table 1 shows the results obtained for both speech tasks when we used the embeddings from the last layers of the convolutional and the fine-tuned blocks. In general, the results are acceptable, with AUC values lying between 0.654 and 0.824, and mean AUC scores between 0.707 and 0.806. Focusing on the mean

Table 1. AUC values obtained for the two speech tasks, when using the embeddings from the last layers of the convolutional and the fine-tuned blocks. AUC is reported as the average (*Mean*) of the five values measured with the five random speaker fold assignments, along with the standard deviation (*Std.*) and the range ([min, max]).

Speech task	Embedding type	AUC		
		Mean	Std.	Range
Opinion	Convolutional	0.707	0.032	[0.654, 0.737]
	Fine-tuned	0.736	0.025	[0.698, 0.763]
Narrative Recall	Convolutional	0.724	0.008	[0.712, 0.733]
	Fine-tuned	0.806	0.014	[0.787, 0.824]

AUC values (averaged over the five classification runs, constructing the folds from different speaker pairs), we can see that the embeddings taken from the last layer of the fine-tuned block outperform those from the convolutional block. Furthermore, the *Opinion* speech task was less effective for detecting multiple sclerosis than the *Narrative Recall* task, since the mean AUC values were higher for both embedding types. When inspecting the standard deviations and the ranges of the AUC scores, we also see that the variance was definitely smaller for the *Narrative Recall* speech task than for the *Opinion* task, suggesting a more robust classification performance. Of course, as the main focus of our investigation is the embeddings taken from the inner fine-tuned layers, the values presented in Table 1 serve only as reference values.

6 Results with the Embeddings of the Inner Layers

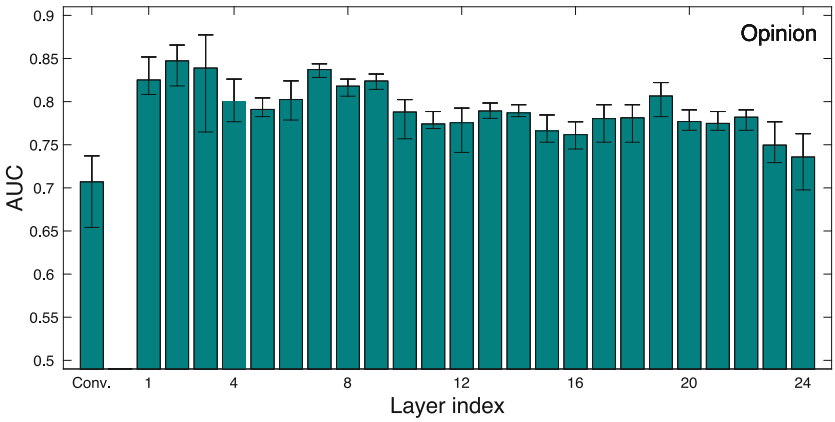

Fig. 2. Mean AUC values (bars) and the [min, max] range (error bars) obtained when using the embeddings from the last layer of the convolutional block (*Conv.*), and when using the hidden layers of the fine-tuned block, for the Opinion speech tasks.

Figure 2 shows the mean AUC values obtained for the *Opinion* speech task for the last layer of the convolutional block (*Conv*) and for all the layers of the fine-tuned block (1...24). (Of course, the 24^{th} layer is the last layer of the fine-tuned block, i.e. that shown in Table 1) Quite surprisingly, the embeddings taken from *any* inner layers (i.e. 1...23) outperformed both those taken from the convolutional layer and those taken from the last fine-tuned layer. The difference, measured by the Mann-Whitney U test (see [23], also known as the Wilcoxon rank-sum test), was statistically significant in almost all cases, with only three exceptions (the embeddings of the 15^{th}, 16^{th} and the 23^{th} layers relative to the last layer of the fine-tuned block). In general, lower layers tend to work better

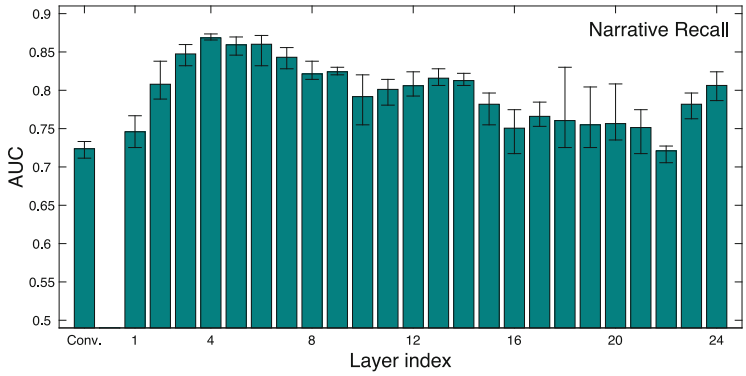

Fig. 3. Mean AUC values (bars) and the [min, max] range (error bars) obtained when using the embeddings from the last layer of the convolutional block (*Conv.*), and when using the hidden layers of the fine-tuned block, for the Narrative Recall speech task.

than those in the higher regions of the fine-tuned block, and we measured the highest mean AUC score with the 2^{nd} layer.

We observe similar tendencies for the *Narrative Recall* task (see Fig. 3), although here the last layer of the fine-tuned block was more competitive. The inner layers outperformed the convolutional embeddings statistically significantly in 20 cases (with the exception of the 16^{th}, 18^{th} and 22^{nd} layers), but, compared to the last fine-tuned layer, the improvement (if any) was statistically significant only in 6 cases (lying in the $1 \ldots 9$ region). This suggests that it is worth exploring the inner hidden layers of the fine-tuned block for feature extraction, and that the lower hidden layers of this block might be more useful than those higher up in the wav2vec 2.0 structure, at least for detecting multiple sclerosis.

Table 2 shows the AUC values measured for some specific inner layers of the fine-tuned block; * and ** indicate a significant difference, $p < 0.05$ and $p < 0.01$, respectively, while "—" means there is no statistically significant improvement compared to the reference values. Symbols before and after the slash symbol (i.e. "/") show the difference compared to the last layer of the convolutional and the fine-tuned block, respectively. For the *Opinion* speech task, we obtained the best results with the 2^{nd} hidden layer (mean AUC value of 0.847), while this was the 4^{th} layer for the *Narrative Recall* task. Both variations brought significant improvements over both reference values (with $p < 0.01$), with absolute improvements of 0.111 and 0.060, Opinion and Narrative Recall speech tasks, respectively. Inspecting the standard deviation and range values, we can also see that classification models relying on the embeddings of these inner layers as features are somewhat more robust, having smaller standard deviation scores. In particular, for the Narrative Recall speech task, when we used the embeddings from the 4^{th} hidden layer, the five AUC values fell into a narrow range (0.866, 0.874).

Table 2. Area Under the ROC Curve (AUC) values obtained for the two speech tasks, when using the embeddings from specific fine-tuned layers. Here, * and ** indicate a statistically significant difference ($p < 0.05$ and $p < 0.01$, respectively), while "—" indicates there is no such difference.

Speech task	Embedding type	AUC		
		Mean	Std.	Range
Opinion	Fine-tuned (#2)$^{**/**}$	0.847	0.018	[0.818, 0.866]
	Fine-tuned (#4)$^{**/**}$	0.800	0.023	[0.777, 0.826]
	Fine-tuned (#6)$^{**/**}$	0.802	0.019	[0.779, 0.824]
	Fine-tuned (#8)$^{**/**}$	0.818	0.008	[0.806, 0.826]
	Last convolutional	0.707	0.032	[0.654, 0.737]
	Last fine-tuned	0.736	0.025	[0.698, 0.763]
Narrative Recall	Fine-tuned (#2)$^{**/—}$	0.808	0.022	[0.789, 0.838]
	Fine-tuned (#4)$^{**/**}$	0.868	0.004	[0.866, 0.874]
	Fine-tuned (#6)$^{**/**}$	0.860	0.016	[0.832, 0.872]
	Fine-tuned (#8)$^{**/—}$	0.821	0.010	[0.814, 0.838]
	Last convolutional	0.724	0.008	[0.712, 0.733]
	Last fine-tuned	0.806	0.014	[0.787, 0.824]

Lastly, Table 3 shows the mean, standard deviation and range of all the AUC values obtained for the lower, middle and top one-third of the fine-tuned layers. (In this table the range property is calculated by taking the 5^{th} and 95^{th} percentiles of the 40 AUC scores.) We can say that all three blocks significantly outperformed the convolutional layer, which is not surprising—as it appears that the convolutional embeddings are just too low-level to serve as a base for effective multiple sclerosis detection. Regarding the comparison with the last layer of the fine-tuned block, however, the lowest region of the fine-tuned block is the only one that is sufficiently robust. Although for the Opinion speech task, all three regions gave an improvement with a $p < 0.01$ significance level, for the Narrative Recall speech task only the layers #1...#8 brought a significant improvement ($p = 0.0377$). The middle region (layers #9...#16) were just on par with the last hidden layer, while the topmost one-third of the fine-tuned layers actually led to a significant decrease in the AUC values. This, in our opinion, suggests that the embeddings from the lower layers are, in general, more suited for automatic multiple sclerosis detection, but they still have to come from the fine-tuned block, as embeddings from the convolutional block performed the worst of all configurations tested for both speech tasks.

Table 3. Area Under the ROC Curve (AUC) values obtained for the two speech tasks, when using the embeddings from specific regions of fine-tuned layers. Here, * and ** indicate a statistically significant difference ($p < 0.05$ and $p < 0.01$, respectively), while "—" indicates there is no such difference.

Speech task	Embedding type	AUC		
		Mean	Std.	Range
Opinion	Fine-tuned (#1...#8)$^{**/**}$	0.820	0.028	[0.778, 0.867]
	Fine-tuned (#9...#16)$^{**/**}$	0.783	0.022	[0.749, 0.827]
	Fine-tuned (#17...#24)$^{**/**}$	0.773	0.025	[0.729, 0.810]
Narrative Recall	Fine-tuned (#1...#8)$^{**/*}$	0.832	0.040	[0.740, 0.872]
	Fine-tuned (#9...#16)$^{**/—}$	0.798	0.026	[0.741, 0.828]
	Fine-tuned (#17...#24)$^{**/—}$	0.762	0.033	[0.719, 0.817]

7 Conclusion and Discussion

In this study we investigated whether multiple sclerosis could be automatically detected from the speech of the subjects. For this, we built a workflow consisting of a wav2vec 2.0 model for feature extraction and an SVM model for classification. We used the speech recordings of 45 native Hungarian speakers (23 MS patients of the relapsing-remitting subtype, and 22 healthy controls), performing two spontaneous speech tasks. Besides using the last layers of the convolutional and the fine-tuned blocks of the wav2vec 2.0 model, we experimented with the other hidden layers of the fine-tuned block as potential sources of the embedding vectors. We found that most inner layers were indeed more effective than the final layers of the two blocks: we achieved statistically significant improvements over the convolutional embeddings in 43 cases out of 46, while the last fine-tuned layer was significantly outperformed in roughly half the cases (i.e. 26 times out of 46). Regarding tendencies, we found that the lower-lying hidden layers were more effective for both speech tasks, indicating that lower-level information might be more suitable for multiple sclerosis detection than high-level one, but the convolutional layers alone cannot capture this information. The reason for this might lie in the efficiency of transformers, which are present only in the fine-tuned block.

Of course, the information stored by the independent layers is not mutually exclusive. Therefore, it might be worth using the embedding vectors obtained from different hidden layers in some way, as this combination might improve the classification performance further. However, a fair validation of such combination algorithms might require more subjects than our 45 (which, in other respects, is a fair number of speakers in the pathological speech processing area). Still, we aim to perform such combination experiments in the near future.

Acknowledgments. This study was supported by the NRDI Office of the Hungarian Ministry of Innovation and Technology (grants K-132460 and TKP2021-NVA-09), and within the framework of the Artificial Intelligence National Laboratory Program (RRF-2.3.1-21-2022-00004).

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. 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
4. 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**, 2079–2107 (2010)
5. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3), 1–27 (2011)
6. Chen, L.W., Rudnicky, A.: Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition. In: Proceedings of ICASSP, Rhodes Island, Greece (2023). https://doi.org/10.1109/ICASSP49357.2023.10095036
7. Egas-López, J.V., Svindt, V., Bóna, J., Hoffmann, I., Gosztolya, G.: Automated multiple sclerosis screening based on encoded speech representations. In: Proceedings of Interspeech, Dublin, Ireland, pp. 3003–3007 (2023)
8. 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)
9. Fara, S., Hickey, O., Georgescu, A., Goria, S., Molimpakis, E., Cummins, N.: Bayesian Networks for the robust and unbiased prediction of depression and its symptoms utilizing speech and multimodal data. In: Proceedings of Interspeech, Dublin, Ireland, pp. 1728–1732 (2023). https://doi.org/10.21437/Interspeech.2023-1709
10. 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, Singapore, pp. 6927–6931 (2022)
11. Grosman, J.: Fine-tuned XLSR-53 large model for speech recognition in Hungarian (2021). https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-hungarian
12. Grósz, T., Porjazovski, D., Getman, Y., Kadiri, S., , Kurimo, M.: Wav2vec2-based paralinguistic systems to recognise vocalised emotions and stuttering. In: Proceedings of ACM Multimedia, Lisboa, Portugal, pp. 7026–7029 (2022)
13. Grósz, T., Virkkunen, A., Porjazovski, D., Kurimo, M.: Discovering relevant subspaces of BERT, Wav2Vec 2.0, ELECTRA and ViT embeddings for humor and mimicked emotion recognition with integrated gradients. In: Proceedings of MuSe, Ottawa, Canada, pp. 27–34 (2023). https://doi.org/10.1145/3606039.3613102
14. 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. Psychiat. **13**, 879896 (2022). https://doi.org/10.3389/fpsyt.2022.879896

15. Huckvale, M., Beke, A., Ikushima, M.: Prediction of sleepiness ratings from voice by man and machine. In: Proceedings of Interspeech, Shanghai, China, pp. 4571–4575 (2020)
16. Ivanova, O., Martínez-Nicolás, I., Meilán, J.J.G.: Speech changes in old age: methodological considerations for speech-based discrimination of healthy ageing and alzheimer's disease. Int. J. Lang. Commun. Disord. **59**(1), 13–37 (2023)
17. Jenei, A.Z., Kiss, G., Sztahó, D.: Detection of speech related disorders by pretrained embedding models extracted biomarkers. In: Proceedings of SPECOM, Gurugram, India, pp. 279–289 (2022)
18. Kiss, G., Tulics, M.G., Sztahó, D., Vicsi, K.: Language independent detection possibilities of depression by speech. In: Proceedings of NoLISP, pp. 103–114 (2016)
19. Klumpp, P., et al.: The phonetic footprint of Parkinson's disease. Comput. Speech Lang. **72**, 101321 (2022)
20. Kodali, M., Kadiri, S.R., Alku, P.: Classification of vocal intensity category from speech using the wav2vec2 and whisper embeddings. In: Proceedings of Interspeech, pp. 4134–4138 (2023). https://doi.org/10.21437/Interspeech.2023-2038
21. Kondratenko, V., Karpov, N., Sokolov, A., Savushkin, N., Kutuzov, O., Minkin, F.: Hybrid dataset for speech emotion recognition in Russian language. In: Proceedings of Interspeech, pp. 4548–4552 (2023). https://doi.org/10.21437/Interspeech.2023-311
22. Kumar, N., Nasir, M., Georgiou, P., Narayanan, S.S.: Robust multichannel gender classification from speech in movie audio. In: Proceedings of Interspeech, San Francisco, CA, USA, pp. 2233–2237 (2016)
23. 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)
24. 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 (2022)
25. Mirheidari, B., O'Malley, R., Blackburn, D., Christensen, H.: Identifying people with mild cognitive impairment at risk of developing dementia using speech analysis. In: Proceedings of ASRU (2023). https://doi.org/10.1109/ASRU57964.2023.10389623
26. Pepino, L., Riera, P., Ferrer, L.: Emotion recognition from speech using wav2vec 2.0 embeddings. In: Proceedings of Interspeech, Brno, Czechia, pp. 3400–3404 (2021). https://doi.org/10.21437/Interspeech.2021-703
27. 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)
28. Schneider, S., Baevski, A., Collobert, R., Auli, M.: wav2vec: unsupervised pretraining for speech recognition. In: Proceedings of Interspeech, pp. 3465–3469 (2019)
29. Szirmai, I.: Neurológia. Medicina, Budapest (2006)
30. Thienpondt, J., Speksnijder, C.M., Demuynck, K.: Behavioral analysis of pathological speaker embeddings of patients during oncological treatment of oral cancer. In: Proceedings of Interspeech, pp. 3018–3022 (2023). https://doi.org/10.21437/Interspeech.2023-1868
31. Tóth, L., et al.: Automatic detection of mild cognitive impairment from spontaneous speech using ASR. In: Proceedings of Interspeech, Dresden, Germany, pp. 2694–2698 (2015)

32. Vaessen, N., Van Leeuwen, D.A.: Fine-tuning wav2vec2 for speaker recognition. In: Proceedings of ICASSP, pp. 7967–7971 (2021)
33. Warule, P., Mishra, S.P., Deb, S.: Significance of voiced and unvoiced speech segments for the detection of common cold. Signal Image Video Process. **17**, 1785–1792 (2023)