Interspeech 2024
1-5 September 2024, Kos, Greece

Automatic Longitudinal Investigation of Multiple Sclerosis Subjects

Gábor Gosztolya[1,2], *Veronika Svindt*[3], *Judit Bóna*[4], *Ildikó Hoffmann*[3,5]

[1] HUN-REN–SZTE Research Group on Artificial Intelligence, Szeged, Hungary
[2] University of Szeged, Institute of Informatics, Szeged, Hungary
[3] Research Center for Linguistics, ELRN, Budapest, Hungary
[4] ELTE Eötvös Loránd University, Dept. of Applied Linguistics and Phonetics, Budapest, Hungary
[5] University of Szeged, Department of Psychiatry, Szeged, Hungary

ggabor @ inf.u-szeged.hu

Abstract

Multiple Sclerosis is a chronic inflammatory disease of the central nervous system. Over time, people with MS may experience significant changes in cognition, language and speech processes. In this study we investigate speech utterances recorded over the course of three years for 16 MS subjects and 12 healthy controls. Our examination is based on speaker category classification (healthy or MS) using wav2vec2 embeddings as features. We found that subject classification performance improved over time: the 0.745-0.844 AUC values from year one increased to 0.891-0.979 in the third year. By analyzing the posterior estimates, we measured a statistically significant improvement in the scores corresponding to the third year for the MS category, while for the control subjects there was no such tendency. This, in our view, indicates that the change is due to a subtle deterioration in the condition of MS patients, which was detected by our machine learning workflow.

Index Terms: multiple sclerosis, longitudinal investigation, wav2vec 2.0, posterior estimates

1. Introduction

Multiple sclerosis (MS) is a chronic autoimmune neurodegenerative disease that affects the central nervous system and it can result in various cognitive and linguistic impairments [1]. The progression of MS can vary considerably from person to person. in disability (difficulty with walking, balance, coordination, and other physical abilities); increase in fatique; changes in visual acuity or color perception; spasticity (increased stiffness, spasms, involuntary muscle contractions); sensory changes (ability to feel heat, cold and touch) and changes in cognitive and language functions. Therefore, investigating changes in the MS symptoms (i.e. longitudinal analysis) might provide information about the progression of the disease.

Longitudinal studies in MS have mainly focused on the effects of medication [2, 3, 4], neural changes [5], fatigue, depression and changes in cognitive ability [6, 7, 8, 9, 10, 11]. Although automatic speech analysis might offer a cheap and non-invasive tool to monitor the progression of the disease, there are only few studies to date that investigate the longitudinal effects of speech and language abilities in MS [12, 13, 14, 15].

Longitudinal analysis of speech, language and cognitive functions in MS may help to determine the natural course of the disease and monitor changes in the condition. Changes in cognitive function, including changes in attention, memory and executive functions, can be tracked. In this study we focus on the speech production of MS subjects. To automate the speech analysis as much as possible, we decided to employ machine learning. The speech of the subjects was fed into a Support Vector Machine (SVM), trained to distinguish the MS and healthy control (HC) subjects, and we focused on the output of the classifier. Still, in this (quite common) speech analysis setup, the choice of the right features is non-trivial. One might utilize hand-crafted attributes [16, 17, 18], which have the advantage of focusing on specific aspects of the speech production process. Another popular option is to employ feature extractor methods which are general in nature; such choices might be x-vectors [19, 20], ECAPA-TDNN [21, 22] or DNN acoustic models of a HMM/DNN hybrid [23].

In this study we opted for using a self-supervised model as feature extractor. Self-supervised learning allows models to learn from orders of magnitude more data, without training labels or any form of annotation. For example, in the wav2vec approach a neural network is trained to pick the correct next sample of the raw audio, which can be implemented even in the complete absence of transcriptions [24]. The weights obtained after this pre-training step can then be used to initialize a second neural network, trained for a given task where the number of samples is limited (fine-tuning) [24].

In this study we opted for the successor of wav2vec called wav2vec 2.0 [25], recently frequently employed in the pathological speech processing area [26, 27]. From wav2vec 2.0 we expect to represent the utterances of MS patients and healthy control subjects at the state-of-the-art level. We investigate the classification scores obtained on the utterances of 16 MS subjects and 12 healthy controls, recorded over the course of three years and involving three speech tasks. We examine the predictions both by yearly split and by overall, and also analyze the resulting posterior estimates statistically. We found significant differences for Year 3, which might indicate slight changes in the cognitive and/or speech abilities of the MS subjects, detected by our machine learning workflow.

2. The Multiple Sclerosis corpus used

All tests were carried out at the Neurology Department of Uzsoki Hospital, Budapest, Hungary, and at the Research Center for Linguistics of the Eötvös Loránd 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. Our corpus already contains the speech of over 80 subjects, but due to Covid-19 restrictions, we do not yet have three recordings from each subject. Therefore in the current study we use the recordings of

10.21437/Interspeech.2024-1009

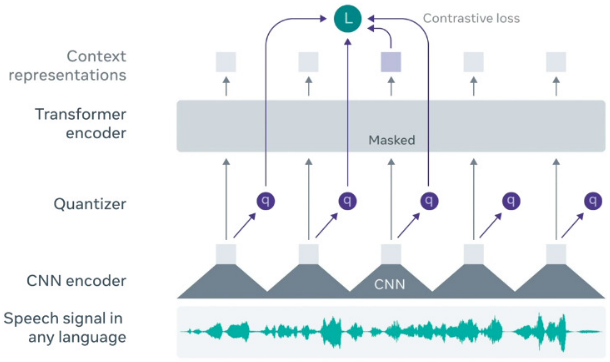

Figure 1: *Fine-tuned wav2vec 2.0 framework structure. Source:* `https://ai.facebook.com/blog`

16 MS subjects (6 males and 10 females, mean age in the first year: 43.6 years) and 12 healthy controls (2 males and 10 females, mean age in the first year: 40.7 years), using recordings made in three consecutive years (2020, 2021 and 2022). All the speakers involved in the study were native Hungarian speakers; and, mirroring the ethnic composition of Hungary, all of them were Caucasians. None of them had any hearing impairment, depression or any other known psychiatric condition.

The linguistic protocol for collecting the speech samples from the subjects was quite extensive; in this study we use three utterances from each subject. First, we asked them to describe the content of the widely-used **Boston Cookie Theft** picture; then the subjects were asked to share their **opinion** about vegetarianism (year 1), keeping pets in flats (year 2), and about advertisements (year 3). In the third task, the subjects were asked to read aloud several non-words, i.e. consonant-vowel-consonant-vowel (CVCV) sequences, in which the first CVs contained a voiceless plosive [p, t, k] and one of the vowels [i:, a:, u:] (task **Phonetics**). The tasks performed and investigated here differ in the activated cognitive processes and the rate of the cognitive load required for speech production.

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

To obtain wav2vec 2.0-based features we fine-tuned a self-supervised wav2vec 2.0 model [28]. This model comprises two main parts: (1) a feature encoder Convolutional Neural Network (CNN) block, (2) a connection-capturing transformer block. The CNN block converts the input raw audio into a series of multilingual quantized latent speech representations. This first part has a "dilated convolution" architecture. The transformer block converts the CNN output into a series of context representations to learn a representation that captures the relationships between features. The second part has a contextualized network architecture based on the BERT model. It includes a multi-head self- attention mechanism and a position-wise feed-forward network. The wav2vec 2.0 model can be trained on large unlabeled datasets using self-supervised learning, and it can be fine-tuned on a smaller annotated dataset by replacing the last layer with task-specific layers. The structure of a fine-tuned wav2vec 2.0 model can be seen in Fig. 1.

Due to the small size of our MS corpus, and the difficulty of using cross-validation for classification along with fine-tuning a wav2vec 2.0 model for our data, we used a model fine-tuned on another Hungarian corpus [29], using 17 hours of Hungarian speech. We used this fine-tuned network as an embedding extractor by freezing the weights and removing the last layers. We experiment with two setups, where we extract embeddings from: (1) the last layer of the CNN block, (2) the last layer of the Transformer block.

Table 1: *Utterance-level EER and AUC values for the two wav2vec 2.0 embedding types (convolutional and fine-tuned) measured when using all the utterances from all subjects*

Speech Task	**Embeddings**	**EER**	**AUC**
Boston Cookie Theft	Convolutional	16.7%	0.917
	Fine-tuned	33.3%	0.744
Opinion	Convolutional	28.6%	0.808
	Fine-tuned	30.9%	0.787
Phonetics	Convolutional	22.6%	0.879
	Fine-tuned	33.3%	0.792

To construct utterance-level features from the frame-level wav2vec 2.0 embeddings, we turned to the standard approach of taking their mean and standard deviations (see e.g. [30, 31, 32]). In the case of the convolutional embeddings, this led to 1024 utterance-level attributes, while for the contextualized activations we had obtained 2048 features. (In the following we will refer to the contextualized embeddings as "fine-tuned" ones.)

2.1. Classification and evaluation

We employed Support Vector Machines trained to predict whether the speakers belonged to the MS or HC group. We utilized the libSVM implementation [33] with a linear kernel (nu-SVR method); the C complexity parameter was set in the range 10^{-5}, ..., 10^{1}. Classification performance was measured in Equal Error Rate (EER) and area under the ROC curve (AUC).

We restricted our experiments to one speech task at a time. Since we had 16 MS and 12 HC subjects, and from each one we had one recording from each of the three years, we had 84 utterances in each experiment. Due to the small number of examples, we chose to perform leave-one-speaker-out cross-validation (CV); one fold always consisted of the three utterances of one subject (being either a healthy control or one having MS). To avoid any form of peeking, we employed *nested cross-validation* [34]. That is, each time we trained our model on the data of 27 speakers (81 utterances), *another* leave-one-speaker-out cross-validation step was performed on these recordings to find the C meta-parameter value with the highest AUC score. Afterwards, we trained an SVM model with this C value on all the data of all the 27 speakers, and this model was evaluated on the three recordings of the remaining speaker.

3. Classification results

First, we evaluated the trained models on all the speakers and utterances. In this experiment we treated the three recordings of a speaker independently, i.e. we measured the performance metrics on the utterance level. The achieved EER and AUC scores can be seen in Table 1. Overall, the measured scores are competitive to previously published results (e.g. [23]), although of course the values cannot be directly comparable, since they were not measured on the same data. In general, features derived from the convolutional embeddings led to better classification performances than those calculated from the fine-tuned

Table 2: *Utterance-level EER and AUC values for the convolutional wav2vec 2.0 embeddings, measured for all utterances from all subjects, and when using only the utterances recorded in a specific year*

Speech Task	Period	EER	AUC
Boston Cookie Theft	All years	16.7%	0.917
	Year 1	17.9%	0.839
	Year 2	14.3%	0.969
	Year 3	7.1%	0.979
Opinion	All years	28.6%	0.808
	Year 1	32.1%	0.745
	Year 2	32.1%	0.771
	Year 3	17.9%	0.891
Phonetics	All years	22.6%	0.879
	Year 1	25.0%	0.844
	Year 2	25.0%	0.854
	Year 3	17.9%	0.932

Table 3: *Utterance-level EER and AUC values for the contextualized ("fine-tuned") wav2vec 2.0 embeddings, measured for all utterances from all subjects, and when using only the utterances recorded in a specific year*

Speech Task	Period	EER	AUC
Boston Cookie Theft	All years	33.3%	0.744
	Year 1	25.0%	0.745
	Year 2	42.9%	0.656
	Year 3	25.0%	0.833
Opinion	All years	30.9%	0.787
	Year 1	17.9%	0.885
	Year 2	50.0%	0.641
	Year 3	25.0%	0.833
Phonetics	All years	33.3%	0.792
	Year 1	50.0%	0.693
	Year 2	32.1%	0.760
	Year 3	7.1%	0.938

(contextualized) ones. The reason for this could be that lower-level attributes (related to, for example, the amount of silent and filled pauses (vocalizations such as 'hmm' and 'er')), captured by the convolutional embeddings, are more useful for detecting Multiple Sclerosis than higher-level ones (e.g. phonetic-related ones, captured by the contextualized embeddings).

Among the three speech tasks, there were no notable differences when using the contextualized ("fine-tuned") embeddings: the EER values were 30.9% and 33.3%, while the AUC scores fell into the $0.744 \ldots 0.792$ range. Regarding the convolutional embeddings, clearly the "Opinion" speech task was the least suitable for MS assessment, although classification performance was not low either. Describing the Boston Cookie Theft picture and the "Phonetics" reading tasks, however, turned out to be significantly more effective, especially judging from the AUC scores (0.917 and 0.879, respectively). This is probably because in these two tasks the spoken content is less free-form than for the 'Opinion' task. Due to this, the difference between MS and control subjects is realized more at the phonetic level, suitable for the higher-level fine-tuned embeddings.

3.1. Classification results for the individual years

Next, we investigated the same classification metrics for the utterances corresponding to the individual years. Technically this means that we only kept the predictions that corresponded to one year (i.e. 16 examples belonging to the MS category and 12 examples corresponding to healthy controls), and did not train any further SVM models. The EER and AUC values measured this way when using the **convolutional** embeddings can be seen in Table 2. The values corresponding to both Year 1 and Year 2 are actually slightly worse than the metric values measured for all three years combined ("All"): the EER scores are higher by 1.2% to 3.5% absolute, while the AUC values are lower by $0.025 \ldots 0.078$. The exception to this is Year 2 for the Boston Cookie Theft speech task, where the EER value improved slightly (an absolute difference of 2.4%), while the AUC score was markedly higher than what was achieved in the first year (0.969 vs. 0.839). For Year 3, however, we can see notable increases in the classification metric values: the EER values fell to $7.1 \ldots 17.9$%, while AUC rose to $0.891 \ldots 0.979$.

We can observe similar tendencies for the **fine-tuned** (**contextualized**) embeddings (see Table 3). (Of course, just as we noted for Table 1, the values in this case are slightly worse than when we used the convolutional representations.) Although, the classification results for Year 1 turned out to be better than for Year 2 for two speech tasks out of the three (i.e. for Boston Cookie Theft and for Opinion), for Year 3 we can see notable improvements: the AUC scores (0.833-0.938) are higher than either in the 'All years' case (0.744-0.792) or in five cases out of the six in the first two years (0.641-0.760) (the exception is the first year for Opinion). The EER scores show a similar trend. Overall, we can see that classification performance was clearly superior for the Year 3 recordings than for the first two years.

4. Posterior Mean Statistics

In our classification experiments we found that it was easier to distinguish MS and HC subjects for Year 3 than for the first years, or for all years altogether. Next, we will investigate the posterior estimates provided by the SVM classifier. For this, we calculated the mean of the posteriors corresponding to the correct class (i.e. speaker category). For this experiment we did not train any new models, but used the posterior estimates provided by the same models as used in Section 3.

Fig. 2 shows the obtained mean values for all years combined, and for the individual years. For the convolutional embeddings (upper part) and for the fine-tuned representations (lower part), it is clear that the values for the MS subjects display a general trend: (posterior mean) values for Year 3 are consistently higher than those for the first two years. Although in some cases the year 2 scores are higher overall than those obtained for year 1 (i.e. fine-tuned embeddings for the Boston Cookie Theft and Phonetics tasks), these differences are much smaller. Perhaps more importantly, we cannot see such a difference in the posterior mean values for the HC category either, indicating that it is unlikely that the superior classification performance for year 3 reflects some difference in the speech task (e.g. different topic in 'Opinion', or that the subjects memorized some parts of the Boston Cookie Theft image).

To verify whether these notable differences are statistically significant, we used the Mann-Whitney U test. The posterior values of the actual speaker categories were selected, and compared with those of the subsequent year. Table 4 shows the p

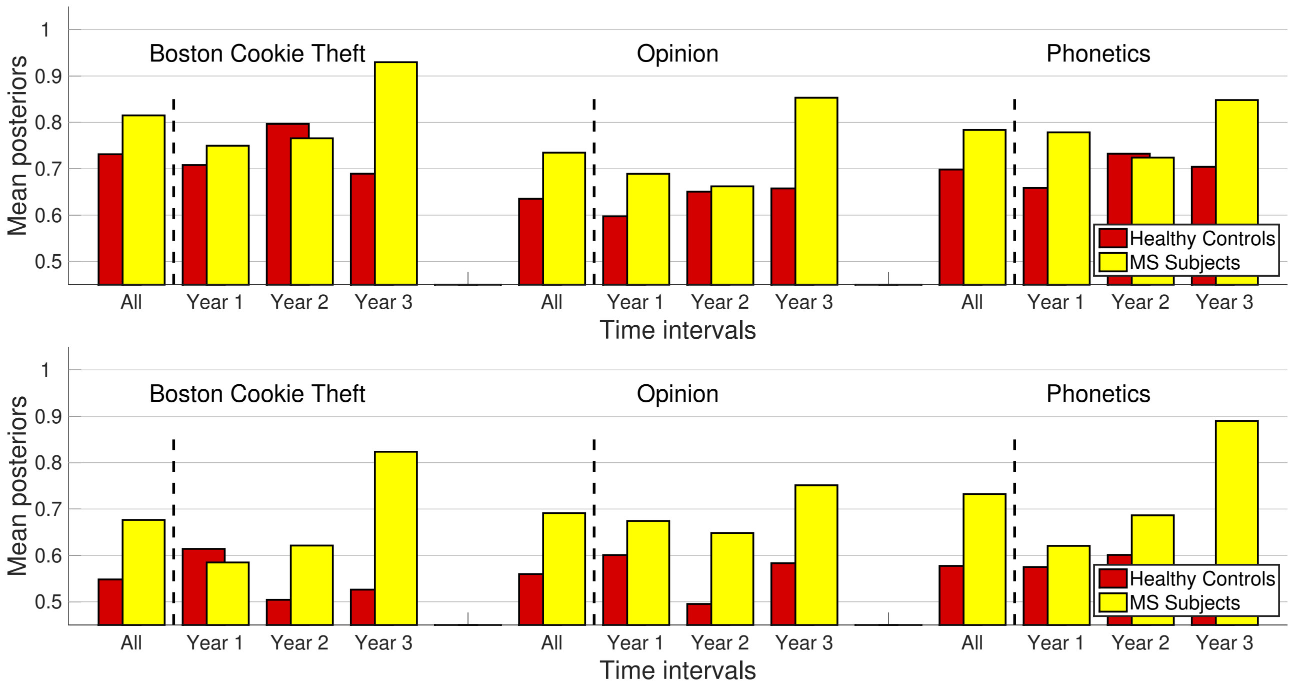

Figure 2: *Mean posteriors for the different speech tasks and for the two speaker categories, obtained for all utterances and for each year separately; convolutional (up) and fine-tuned (down) representations.*

Table 4: *Significances (p) of the posterior estimates between consecutive years; $p < 0.05$ cases are shown in* **bold**

Embeddings	**Speech Task**	**Periods**	**HC**	**MS**
Convolutional	BCT	Year 1 vs. 2	0.471	0.955
		Year 2 vs. 3	0.544	**0.009**
	Opinion	Year 1 vs. 2	0.977	1.000
		Year 2 vs. 3	0.885	**0.037**
	Phonetics	Year 1 vs. 2	0.260	0.337
		Year 2 vs. 3	0.471	**0.040**
Fine-tuned	BCT	Year 1 vs. 2	0.194	0.692
		Year 2 vs. 3	0.624	**0.010**
	Opinion	Year 1 vs. 2	0.260	0.836
		Year 2 vs. 3	0.371	**0.044**
	Phonetics	Year 1 vs. 2	0.751	0.720
		Year 2 vs. 3	0.403	**0.002**

values measured; statistically significant values ($p < 0.05$) are shown in **bold**. The values follow a quite clear pattern; namely, there is a clear difference in the posterior estimates between the second and the third years, but only for the MS patients. Between the posterior estimates for the first two years for the MS subjects, and for either year pairs for healthy controls, the differences between the posterior estimates for the correct speaker category were statistically not significant. This also supports our observation drawn from Fig. 2, and also that automatic speech analysis might be able to capture the subtle deterioration of MS subjects in their speech and in their cognitive abilities. Of course, we cannot rule out other factors, such as a sudden change in the acoustic conditions of the MS patients in Year 3, which the wav2vec 2.0 embeddings (and the subsequent classification process) might have been able to capture.

5. Conclusions

In this study we performed a longitudinal investigation of Multiple Sclerosis patients and healthy control subjects. We investigated the speech recordings of 16 MS subjects and 12 healthy controls, using three speech tasks (Boston Cookie Theft picture description, asking their opinion, and reading aloud specific non-existent words) over the course of three years. We utilized machine learning to distinguish the speaker groups (i.e. MS and controls): we made the straightforward choice of using the mean and standard deviation of wav2vec 2.0 embeddings as features, while as classifiers, due to data scarcity, we applied Support Vector Machines in a nested cross-validation setup.

We noted that MS is known to cause a deteriorating speech performance over time, as well as to adversely affect various cognitive functions. Due to this, besides analyzing the effect of the speech task and the type of embeddings (convolutional and fine-tuned / contextualized), we were also interested in the effect of the year of recording. In line with our hypothesis, when we limited the classification evaluation to a specific year, we saw an increased MS detection performance for the third year compared to the first two years, and this trend was quite noticeable for all three speech tasks. As this improvement could possibly have been caused by a change in the speech of the control subjects, we investigated the posterior estimates provided by the SVM classifier. We found a significant difference in the posterior values of the MS subjects in the third year for all three tasks and for both embedding types, but not in any other cases. This might indicate that there is a slight deterioration of the cognitive and/or speech abilities of the MS subjects over time, which was detected by the wav2vec 2.0-derived features and the SVM classifier. Of course, the exact phenomenon causing this sudden improvement in our scores (be it either some speech property or, perhaps, some acoustic artifact) needs to be looked into, which we plan to do in the near future.

6. Acknowledgements

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

7. References

[1] I. Szirmai, *Neurológia*. Medicina, Budapest, 2006.

[2] B. Nourbakhsh, L. Julian, and E. Waubant, "Fatigue and depression predict quality of life in patients with early multiple sclerosis: a longitudinal study," *European Journal of Neurology*, vol. 23, p. 1482–1486, 2016.

[3] C. Hemond, B. Healy, S. Tauhid, M. Mazzola, F. Quintana, R. Gandhi, H. Weiner, and R. Bakshi, "MRI phenotypes in MS: Longitudinal changes and miRNA signatures," *Neurology Neuroimmunology & Neuroinflammation*, vol. 6, 2019.

[4] C. C. Hemond, J. Baek, C. Ionete, and D. S. Reich, "Paramagnetic rim lesions are associated with pathogenic CSF profiles and worse clinical status in multiple sclerosis: A retrospective cross-sectional study," *Multiple Sclerosis Journal*, vol. 28, no. 13, pp. 2046–2056, 2022.

[5] M. Fartaria, T. Kober, C. Granziera, and M. Bach Cuadra, "Longitudinal analysis of white matter and cortical lesions in multiple sclerosis," *NeuroImage: Clinical*, vol. 23, p. 101938, 2019.

[6] S. Morrow, B. Weinstock-Guttman, F. Munschauer, D. Hojnacki, and R. Benedict, "Subjective fatigue is not associated with cognitive impairment in multiple sclerosis: Cross-sectional and longitudinal analysis," *Multiple Sclerosis*, vol. 15, pp. 998–1005, 2009.

[7] A. Rabinowitz and P. Arnett, "A Longitudinal Analysis of Cognitive Dysfunction, Coping, and Depression in Multiple Sclerosis," *Neuropsychology*, vol. 23, pp. 581–91, 2009.

[8] J. Sumowski, M. Rocca, V. Leavitt, J. Dackovic, S. Mesaros, J. Drulovic, J. Deluca, and M. Filippi, "Brain reserve and cognitive reserve protect against cognitive decline over 4.5 years in MS," *Neurology*, vol. 82, pp. 1776–83, 2014.

[9] R. Barbu, J. Berard, L. Gresham, and L. Walker, "Longitudinal Stability of Cognition in Early-Phase Relapsing-Remitting Multiple Sclerosis. Does Cognitive Reserve Play a Role?" *International Journal of MS Care*, vol. 20, p. 173–179, 2018.

[10] A. Andravizou, V. Siokas, A. Artemiadis, C. Bakirtzis, A. Aloizou, N. Grigoriadis, M. Kosmidis, G. Nasios, L. Messinis, G. Hadjigeorgiou, E. Dardiotis, and E. Peristeri, "Clinically reliable cognitive decline in relapsing remitting multiple sclerosis: Is it the tip of the iceberg?" *Neurological and Neurosciences*, vol. 42, pp. 575–586, 2020.

[11] A. Pike, G. James, P. Drew, and R. Archer, "Neuroimaging Predictors of Longitudinal Disability and Cognition Outcomes in Multiple Sclerosis Patients: A Systematic Review and Meta-Analysis," *Multiple Sclerosis and Related Disorders*, vol. 57, p. 103452, 2021.

[12] B. Duque, J. Sepulcre, B. Bejarano Herruzo, L. Samaranch, P. Pastor, and P. Villoslada, "Memory decline evolves independently of disease activity in MS," *Multiple Sclerosis*, vol. 14, pp. 947–53, 2008.

[13] R. Ehling, M. Amprosi, B. Kremmel, G. Bsteh, K. Eberharter, M. Zehentner, R. Steiger, N. Tuovinen, E. Gizewski, T. Benke, T. Berger, C. Spöttl, C. Brenneis, and C. Scherfler, "Second language learning induces grey matter volume increase in people with multiple sclerosis," *PLOS ONE*, vol. 14, p. e0226525, 2019.

[14] A. Kever, K. Buyukturkoglu, S. Levin, C. Riley, P. De Jager, and V. Leavitt, "Associations of social network structure with cognition and amygdala volume in multiple sclerosis: An exploratory investigation," *Multiple Sclerosis Journal*, vol. 28, pp. 1–9, 2021.

[15] J. Sepulcre, H. Peraita, J. Goni, G. Arrondo, I. Martincorena, B. Duque, N. Velez de Mendizabal, J. Masdeu, and P. Villoslada, "Lexical access changes in patients with multiple sclerosis: A two-year follow-up study," *Journal of Clinical and Experimental Neuropsychology*, vol. 33, pp. 169–75, 2011.

[16] J. Wagner, D. Schiller, A. Seiderer, and E. Andre, "Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant?" in *Proc. Interspeech*, 2018, pp. 147–151.

[17] M. Huckvale, "Neural network architecture that combines temporal and summative features for infant cry classification in the interspeech 2018 computational paralinguistics challenge," in *Proc. Interspeech*, 2018, pp. 137–141.

[18] P. Barche, K. Gurugubelli, and A. K. Vuppala, "Towards automatic assessment of voice disorders: A clinical approach." in *Proc. Interspeech*, 2020, pp. 2537–2541.

[19] L. Moro-Velazquez, J. Villalba, and N. Dehak, "Using x-vectors to automatically detect Parkinson's disease from speech," in *Proc. ICASSP*, 2020, pp. 1155–1159.

[20] J. V. Egas-López, G. Kiss, D. Sztahó, and G. Gosztolya, "Automatic assessment of the degree of clinical depression from speech using x-vectors," in *Proc. ICASSP*, Singapore, 2022, pp. 8502–8506.

[21] D. Wang, Y. Ding, Q. Zhao, P. Yang, S. Tan, and Y. Li, "ECAPA-TDNN based depression detection from clinical speech," in *Proc. Interspeech*, 2022, pp. 3333–3337.

[22] A. Z. Jenei, G. Kiss, and D. Sztahó, "Detection of speech related disorders by pre-trained embedding models extracted biomarkers," in *SPECOM*, Gurugram, India, 2022, pp. 279–289.

[23] G. Gosztolya, L. Tóth, V. Svindt, J. Bóna, and I. Hoffmann, "Using acoustic Deep Neural Network embeddings to detect Multiple Sclerosis from speech," in *Proc. ICASSP*, Singapore, May 2022, pp. 6927–6931.

[24] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Interspeech*, 2019, pp. 3465–3469.

[25] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[26] Y. Zhu, A. Obyat, X. Liang, J. A. Batsis, and R. M. Roth, "WavBERT: Exploiting semantic and non-semantic speech using wav2vec and BERT for dementia detection," in *Proc. Interspeech*, 2021, pp. 3790–3794.

[27] P. A. Pérez-Toro, P. Klumpp, A. Hernandez, T. Arias, P. Lillo, A. Slachevsky, A. M. García, M. Schuster, A. K. Maier, E. Nöth, and J. R. Orozco-Arroyave, "Alzheimer's detection from English to Spanish using acoustic and linguistic embeddings," in *Proc. Interspeech*, Incheon, South Korea, 2022, pp. 2483–2487.

[28] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-lingual representation learning for Speech Recognition," 2020.

[29] J. Grosman, "Fine-tuned XLSR-53 large model for speech recognition in Hungarian," https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-hungarian, 2021.

[30] P. Ghahremani, P. S. Nidadavolu, N. Chen, J. Villalba, D. Povey, S. Khudanpur, and N. Dehak, "End-to-end deep neural network age estimation." in *Proc. Interspeech*, 2018, pp. 277–281.

[31] H. Kaya, D. Fedotov, A. Yesilkanat, O. Verkholyak, Y. Zhang, and A. Karpov, "LSTM based cross-corpus and cross-task acoustic emotion recognition," in *Proc. Interspeech*, 2018, pp. 521–525.

[32] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv preprint arXiv:1803.10963*, 2018.

[33] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[34] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, no. Jul, pp. 2079–2107, 2010.