III. International Conference on Electrical, Computer and Energy Technologies (ICECET 2023)
16-17 November 2023, Cape Town-South Africa

Speech-based Screening of Multiple Sclerosis By Features Derived from Self-Supervised Models

Gábor Gosztolya
University of Szeged, Institute of Informatics
ELKH-SZTE Research Group on Artificial Intelligence
Szeged, Hungary
ggabor@inf.u-szeged.hu

José Vicente Egas-López
ELKH-SZTE Research Group on Artificial Intelligence
Szeged, Hungary
egasj@inf.u-szeged.hu

***Abstract*—Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous system. Since it, among other symptoms, adversely affects the speech of the subject, automatic speech analysis might offer a simple, inexpensive and remote tool for MS screening or monitoring the progression of the disease. We employ ten different wav2vec 2.0 models as the base of feature extraction and compare the performance with pre-trained and custom x-vector models. Based on our results, cross-lingual models perform better than the base wav2vec 2.0 networks, but the model size is crucial as the best results were obtained with a model having one billion trainable parameters. We found fine-tuning the application language to be beneficial to the classification performance, but for other languages, it did not improve the AUC scores. Surprisingly, though, we did not outperform standard x-vectors, which might be due to the standard, but perhaps too simple aggregation strategy of the frame-level embeddings.**

***Index Terms*—multiple sclerosis, pathological speech processing, wav2vec2**

I. Introduction

Multiple sclerosis is an inflammatory condition that affects the nervous system over a long course. Most of the cases in relation to MS are generally divided into three clinical steps: relapsing-remitting, primary progressive MS, and secondary progressive MS, which develops from relapsing-remitting, based on the ongoing existence of symptoms or their temporal absence. One of the main diagnostic characteristics of MS is a variety of motor skill deficits, and alterations in the subjects' motor abilities usually indicate a worsening of the condition. Due to the interconnected nature of language, cognition, and motor skills in the brain, modifications to one of these areas could have an impact on the others. As a result, automatic monitoring of alterations in speech production may be a useful technique for determining how MS is progressing.

Around 60-70% of the patients suffering from MS present distinct cognitive impairments such as decreased information processing speed, chronic fatigue, or various orientation disorders. Additionally, more than a third of MS patients report having transient or ongoing speech problems [1, 2], which also underlines the practical usability of automatic speech analysis. In individuals with multiple sclerosis, language and speech-related symptoms often manifest as motor speech disorders, including conditions like dysarthria and dysphonia, difficulties in recalling words, lack of verbal fluency [3], issues on sentence repetition and higher-level language processes [4,5], and limited inclination for communication [6].

Despite only one-third of the MS cases suffer from dysarthria, automatic speech analysis may be able to identify signs of a minor motor speech issue even before the disease [7]. These minor symptoms could, with a well-structured technique, indicate the beginning of cognitive deterioration. Some research has focused on the automatic processing of speech produced by individuals suffering from some sort of mental or physical disease like Alzheimer's disease [8], or depression [9].

A common approach in this area is to discriminate using Support Vector Machines (SVM) and use some deep neural network for feature extraction [8, 10]. Since the size of pathological datasets, generally, does not allow the training of this feature extractor DNN on the actual recordings of the subjects, typically standard ASR corpora are used for this step. These DNNs may be viewed as models for 'standard speech', while the feature extraction step expresses the difference between standard speech and the speech utterance produced by the actual subject.

In this study, we employ wav2vec 2.0 pre-trained neural network models as the base of feature extractors to identify MS subjects based on their speech. Since there are variations of these models, we experiment with 10 models overall: besides the generic "base" models, we test cross-lingual extractors (trained to recognize the phones of several different languages at the same time) as well as models fine-tuned for one specific language. Placing the results into a broader context, we will also give the performance of pre-trained and custom x-vector models [11], which were also widely employed in the pathological speech processing field as feature extractors [9,10].

979-8-3503-2781-6/23/$31.00 ©2023 IEEE

II. Data

Tests were conducted at the Neurology Department of Uzsoki Hospital, Hungary, and at the Research Institute for Linguistics of the Eötvös Loránd Research Network, Budapest, Hungary. Here, 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 participants included in the research were individuals who were native speakers of Hungarian. In line with the ethnic makeup of Hungary, they all belonged to the Caucasian ethnic group. None of the participants had any registered hearing impairments, a history of depression, or any recognized psychiatric disorders. The demographic characteristics (i.e., age, gender (male/female), and years of education) did not exhibit any statistically significant disparities between the MS and HC groups.

The linguistic protocol employed for recording the speech samples was quite comprehensive, encompassing a total of 17 distinct speech tasks. However, for the purposes of this study, we chose to utilize only the ”narrative recall” task. Here, participants were presented with a two-minute-long historical anecdote, and their objective was to provide an accurate summary of the story they had just heard. This particular task involves a range of cognitive processes, including focused attention, working memory, temporal orientation, organization, and sequencing, as highlighted in the work by Mar et al. [12]. The recordings were initially conducted at a sampling rate of 48 kHz. Subsequently, they were converted to a 16 kHz mono format with a 16-bit resolution.

III. Self-Supervised Learning

Self-supervised learning enables models to learn from significantly larger datasets, which is essential for capturing patterns in less common phenomena. Typically, ASR technology demands extensive quantities of transcribed data in order to achieve high performance, as noted by Amodei et al. [13]. An effective strategy to address this challenge is to employ neural network pre-training, especially in situations where labeled data is scarce. Pre-training involves training a neural network on a task that provides access to vast amounts of unlabeled data, often in an unsupervised or self-supervised manner. Subsequently, the learned weights from this pre-training phase are utilized to initialize a second neural network, fine-tuned for a specific task with limited available samples.

The **wav2vec** approach basically generates a representation suitable for an Automatic Speech Recognition system from raw audio. This architecture aims to forecast the upcoming observations from a specified utterance, as outlined by Schneider et al. [14]. The **wav2vec 2.0** architecture improves upon this by incorporating masking during the training process. In this approach, the raw audio is encoded using a series of convolutional neural networks. Following a methodology akin to masked language modeling, the wav2vec 2.0 approach involves the masking of small segments within the latent speech representations, which are shorter in duration than phonemes. These masked representations are subsequently

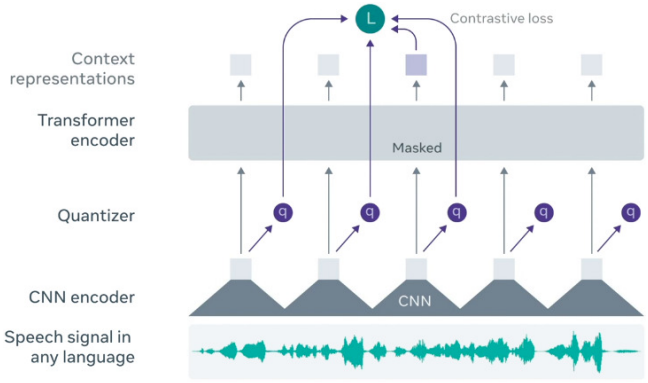

Fig. 1. wav2vec 2.0 architecture. Source: https://ai.facebook.com/blog

input into both a quantizer and a transformer network. The quantizer is responsible for selecting a speech unit from the latent audio representation, leveraging knowledge from a learned inventory of such units while the transformer network appends information from the entire utterance [15].

Once the pre-training phase is completed, the model undergoes fine-tuning for ASR using annotated corpora, employing Connectionist Temporal Classification (CTC) loss for sequence alignment. The architecture used in wav2vec 2.0 is depicted in Figure 1.

A. Cross-lingual Representation Learning

XLSR (Cross-lingual Speech Representations) is a multilingual representation method that relies on wav2vec 2.0, and addresses the challenge of working with languages that have limited or even no available unlabeled data. XLSR adopts a pre-training strategy where a model is simultaneously trained on multiple datasets from various languages.

One notable distinction in the XLSR architecture, as compared to wav2vec2, is the quantization module. In XLSR, this module is in charge of generating multilingual quantized speech units. Then, the transformer block as uses them as targets for learning via a contrastive task. This method allows for an effective handle of discrete tokens across different languages, making it a versatile solution for multilingual speech representation learning, as outlined by Conneau et al. [16].

B. Feature Extraction with wav2vec 2.0

The multi-layer convolutional block generates two distinct types of outputs. Firstly, it produces a sequence of features extracted from the last layer of the block of convolutions. Secondly, it generates a sequence of hidden states derived from the final part of the contextual block. These two kinds of embeddings can contain relevant information related to speakers, as indicated in the work by Lin et al. [17], as well as other information encoded within the speech signal, as explored by Fan et al. [18].

It’s worth noting that wav2vec was originally designed for speech recognition, resulting in the number of these feature

TABLE I.
AUC VALUES OBTAINED BY THE BASE WAV2VEC2 MODELS.

wav2vec 2.0 model	Embedding type	Aggregation	
		Mean	M. + Std.
wav2vec2-base	Convolutional	0.745	0.781
	Hidden	0.731	0.802
wav2vec2-base-960h	Convolutional	0.745	0.781
	Hidden	0.785	0.731
wav2vec2-large-960h	Convolutional	0.700	0.795
	Hidden	0.771	0.763

TABLE II.
AUC VALUES OBTAINED BY THE CROSS-LINGUAL MODELS.

wav2vec 2.0 model	Embedding type	Aggregation	
		Mean	M. + Std.
XLSR-53	Conv.	0.700	0.729
	Hidden	0.763	0.787
XLS-R-300M	Conv.	0.741	0.743
	Hidden	0.773	0.706
XLS-R-1B	Conv.	0.670	0.696
	Hidden	0.862	0.872

vectors being proportional to the dimension of the input recording. To utilize them as utterance-level features, they have to be aggregated across the entire recording. Here, we considered aggregation methods such as calculating the mean and standard deviation (Std.) of these values along the time axis. Note that in many cases, especially when dealing with pathological speech corpora, it is challenging to perform fine-tuning of the Deep Neural Network (DNN) models directly on the actual utterances due to the limited number of subjects and the potential risk of overfitting. Consequently, such studies typically leverage neural networks primarily as feature extractors in their analyses.

IV. EXPERIMENTAL SETUP

We employed Support Vector Machines as the classifier, we relied on the libSVM implementation [19] with a linear kernel (nu-SVR method); the C complexity parameter was set in the range 10^{-5}, ..., 10^{1}. Limited by the small size of the dataset, we opted for cross-validation (CV); one fold always consisted of the features of one control subject and one having MS. Seeking to avoid the presence of peeking, we carried out *nested cross-validation* [20]: each time, we trained the model on samples of 22 folds, *another* (22-fold) cross-validation session was performed, to find the C meta-parameter value that gave the highest AUC score within these subjects. Afterward, we trained the SVM model with the selected C value on all the data of these 22 folds, and then this model was evaluated on the speakers of the remaining fold. We measured the efficiency of MS classification by the area under the ROC curve (AUC) value; since there are only two speaker categories (i.e., classes), the AUC value of the two are identical.

V. RESULTS

A. Base wav2vec 2.0 Models

In our initial series of experiments, we tested base wav2vec 2.0 models. First, we employed `wav2vec2-base` [15], a model pre-trained on 53,000 non-annotated LibriSpeech samples (not fine-tuned). Next, we relied on `wav2vec2-base-960h` and `wav2vec2-large-960h`. These two were pre-trained and fine-tuned on 960 hours of annotated corpora, and their main difference lies in the number of parameters (i.e., 95 and 317 million parameters, respectively). The results obtained by these models can be seen in Table I.

The measured AUC values are competitive, but not outstanding, since for most cases they fall in the range $0.745 \ldots 0.802$. Regarding the utterance-level aggregation strategy, including the standard deviation of the activations seems to help classification, with the exception of the last hidden layers for the fine-tuned models (i.e. `base-960h` and `large-960h`), where taking only the mean of the values proved a better strategy.

As for the choice of embeddings, for the `base` model the last hidden layer produced slightly better performance than the convolutional one. For the two `960h` models, however, relying on the convolutional layers led to similar AUC scores as before, but the last hidden layer induced a drop in performance in the "Mean + Std." case. This is probably because the fine-tuning step made these layers focus on the spoken content of the speech signal more, so their activations were less suitable for MS detection.

B. Cross-Lingual Models

Next, we focused on the cross-lingual wav2vec 2.0 models: `XLSR-53` and `XLS-R`. The former was fitted relying on 53 distinct languages, and the latter (successor of XLSR) was pre-trained on 128 languages comprising around half a million hours of recordings [21]. Models of three different sizes are available; limited by computational resources, we made use of the two smaller versions. That is, `wav2vec2-XLS-R-300M` with 300 million parameters and `wav2vec2-XLS-R-1B`, with 1 billion parameters. Table II shows the AUC scores obtained by using these models.

Regarding the activations of the convolutional layers, surprisingly we got worse performance scores than with the base models: the AUC values fell in the range $0.670 \ldots 0.743$. Regarding the activations of the last hidden layer, with the `XLSR-53` model our results appeared to be around the level of the `wav2vec2-base` model and they were slightly better than with the two fine-tuned ("960h") base models (especially when we included the standard deviations as features). This suggests that this cross-lingual approach was in fact more effective than the mono-lingual fine-tuning on LibriSpeech, at least for distinguishing the speech of MS subjects from healthy controls. Regarding the `XLS-R` models, it proved to be ineffective than `XLSR-53` with "only" 300 million parameters. However, the larger model (having one billion parameters) was indeed more suitable: utilizing both the means

and the standard deviations of the activations of the hidden layer led to an AUC score of 0.872.

TABLE III.
AUC VALUES OBTAINED BY RELYING ON THE CONTEXTUALIZED ("HIDDEN") EMBEDDINGS OF THE FINE-TUNED CROSS-LINGUAL MODELS.

wav2vec 2.0 model	Aggregation	
	Mean	M. + S.
XLSR-53	0.763	0.787
XLSR-Hungarian-53	0.816	0.820
XLSR-Finnish-53	0.749	0.767
XLSR-German-53	0.791	0.783
XLSR-Spanish-53	0.761	0.737

C. Monolingual Fine-tuned XLSR Models

Next, we were interested in the performance of wav2vec 2.0 models fine-tuned for one specific language. For this, we focused on models which were fine-tuned using the XLSR-53 model. Since our recordings were in Hungarian, we chose a Hungarian model [1]. Next, we evaluated a Finnish model, as Finnish belongs to the same language family as Hungarian (both are Uralic languages). Furthermore, we were interested in the performance of models fine-tuned with a significant amount of training data, so we chose German and Spanish. (Note that German is phonetically similar to Hungarian, while it is not true for Spanish.) All four models were trained by the same team (`jonatasgrosman`), on the corresponding part of the Mozilla Common Voice 6.1 corpus, on 8, 1, 777 and 579 hours of data, Hungarian, Finnish, German and Spanish, respectively. As the convolutional part of the network was unaffected by the fine-tuning step, we only reported the values obtained by the contextualized (i.e. "Hidden") embeddings.

Table III shows the results obtained with the language-dependent fine-tuned wav2vec 2.0 models. Clearly, using the Hungarian model led to a nice improvement over the original AUC values. However, relying on the other three models produced quite similar AUC scores as the original cross-lingual XLSR-53 model did: the AUC values fell in the range $0.749 \ldots 0.791$ for mean (XLSR-53: 0.763) and $0.785 \ldots 0.812$ for standard deviation (XLSR-53: 0.791). This, in our opinion, shows that fine-tuning a model to a different language is not really beneficial; however, when the given wav2vec 2.0 model is fine-tuned to the actual target language, it can lead to an improved MS detection performance even when the training data is quite small (8 hours). This is most obvious in the "Mean + Std." case, where the AUC value improved from 0.787 to 0.820.

VI. Comparison With X-vectors

Lastly, we compare our results with those obtained with x-vectors, usually regarded as a competitive baseline. For this, we used the `SRE-16` pre-trained model by Snyder et al. [11], which was trained on a portion of Switchboard (28k recordings) and a subset of the NIST SRE corpus (63k utterances).

[1] jonatasgrosman/wav2vec2-large-xlsr-53-hungarian

TABLE IV.
AUC VALUES OBTAINED BY RELYING ON THE CONTEXTUALIZED ("HIDDEN") EMBEDDINGS OF THE FINE-TUNED CROSS-LINGUAL MODELS.

x-vector model	Features	AUC
SRE-16 (English)	MFCC	0.876
Custom (Hungarian)	MFCC	0.850
	FBANK	0.793
	Spectrogram	0.798

We also trained custom x-vector extractors on a 60 hours subset of the (Hungarian) BEA Hungarian Spoken Language Database [22], using MFCCs, FBANKs and spectrograms as inputs. The AUC values can be seen in Table IV.

Clearly, although the x-vector architecture is much simpler than the wav2vec 2.0 neural network structure and it has fewer parameters (around 10M), the scores achieved are quite high. Even the custom extractors (trained only on 60 hours of data, increased to 240 by adding noise and reverberation in the case of the FBANK and spectrogram models) matched the performance of most wav2vec 2.0 models, with the exception of `XLS-R-1B` and `XLSR-Hungarian-53`. The SRE-16 pre-trained extractor, however, led to an AUC score of 0.876, practically matching the wav2vec 2.0 model with one billion parameters. In our opinion, this (surprising) the result is due to the simple (though standard) aggregation approach used for the wav2vec 2.0 models. Although the statistics pooling layer employed in the x-vector networks also only takes the mean and the standard deviation of the frame-level activations, it is then followed by two further hidden layers (the x-vector embeddings are typically taken from the last hidden layer). These layers are also tuned during training, and this might lead to a more sophisticated and a more effective form of aggregation. This is why in the near future we plan to experiment with further aggregation strategies of the frame-level embedding vectors besides the widely-employed mean and standard deviation.

VII. Conclusions

In this study, we focused on the automatic detection of Multiple Sclerosis from the speech of the subjects. For this, we employed ten different transformer-based wav2vec 2.0 models as feature extractors; the frame-level embeddings were aggregated to form utterance-level features by taking their mean and standard deviation over the whole recording. We found that cross-lingual fine-tuning of the unsupervised pre-trained models was not really beneficial for MS detection, with the exception of the quite large `XLS-R` model (`XLS-R-1B`). Still, even the fine-tuning of a smaller model turned out to be useful when the language of the training material matched that of the MS and healthy control subjects (in our case, Hungarian). However, surprisingly, the results of the wav2vec 2.0 models did not exceed those of the standard `SRE-16` x-vector extractor (AUC values of 0.872 and 0.876, respectively), which contains only a fraction of the parameters the wav2vec 2.0 model has. Because of this, in the near future, we will focus

on more sophisticated aggregation strategies of wav2vec 2.0 embeddings.

VIII. 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).

References

[1] K. Laakso, K. Brunnegård, L. Hartelius, and E. Ahlsén, "Assessing high-level language in individuals with multiple sclerosis: A pilot study," *Clinical Linguistics & Phonetics*, vol. 14, no. 5, pp. 329–349, 2000.

[2] S. Renauld, L. Mohamed-Saïd, and J. Macoir, "Language disorders in multiple sclerosis: A systematic review," *Multiple Sclerosis and Related Disorders*, vol. 10, no. Nov, pp. 103–111, 2016.

[3] A. Delgado-Álvarez, J.A. Matias-Guiu, C. Delgado-Alonso, L. Hernández-Lorenzo, A. Cortés-Martínez, L. Vidorreta, P. Montero-Escribano, V. Pytel, and J. Matias-Guiu, "Cognitive processes underlying verbal fluency in multiple sclerosis," *Frontiers in Neurology*, vol. 11, 2021.

[4] F.L. Darley, J.R. Brown, and N.P. Goldstein, "Dysarthria in multiple sclerosis," *Journal of Speech and Hearing Research*, vol. 15, no. 2, pp. 229–245, 1972.

[5] L. Hartelius, B. Runmarker, and O. Andersen, "Prevalence and characteristics of dysarthria in a multiple-sclerosis incidence cohort: Relation to neurological data," *Folia Phoniatrica et Logopaedica*, vol. 52, no. 4, pp. 160–177, 2000.

[6] F.J.F. Gerald, B.E. Murdoch, and H.J. Chenery, "Multiple sclerosis: Associated speech and language disorders," *Australian Journal of Human Communication Disorders*, vol. 15, no. 2, pp. 15–35, 1987.

[7] D. Mulfari, G. Meoni, M. Marini, and L. Fanucci, "Machine learning assistive application for users with speech disorders," *Applied Soft Computing*, vol. 103, no. May, 2021.

[8] 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 *Proceedings of Interspeech*, 2022, pp. 2483–2487.

[9] 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 *Proceedings of ICASSP*, 2022, pp. 8502–8506.

[10] L. Jeancolas, D. Petrovska-Delacrétaz, G. Mangone, B.-E. Benkelfat, J.-C. Corvol, M. Vidailhet, S. Lehéricy, and H. Benali, "X-vectors: New quantitative biomarkers for early Parkinson's Disease detection from speech," *Frontiers in Neuroinformatics*, vol. 15, 2021.

[11] D. Snyder, Daniel Garcia-Romero, G. Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust DNN embeddings for speaker verification," in *Proceedings of ICASSP*, 2018, pp. 5329–5333.

[12] R.A. Mar, "The neuropsychology of narrative: Story comprehension, story production and their interrelation," *Neuropsychologia*, vol. 42, no. 10, pp. 1414–1434, 2004.

[13] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, et al., "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proceedings of ICML*, 2016, pp. 173–182.

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

[15] A. Baevski, Y. 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. 12449–12460, 2020.

[16] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proceedings of Interspeech*, 2021, pp. 2426–2430.

[17] W.-W. Lin and M.-W. Mak, "Wav2spk: A simple DNN architecture for learning speaker embeddings from waveforms," in *Proceedings of Interspeech*, 2020, pp. 3211–3215.

[18] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," in *Proceedings of Interspeech*, 2021, pp. 1509–1513.

[19] 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.

[20] 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.

[21] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, et al., "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proceedings of Interspeech*, 2022, pp. 2278–2282.

[22] T. Neuberger, D. Gyarmathy, T.E. Gráczi, V. Horváth, M. Gósy, and A. Beke, "Development of a large spontaneous speech database of agglutinative Hungarian language," in *Proceedings of TSD*, 2014, pp. 424–431.