Automatic Longitudinal Investigation of Multiple Sclerosis Subjects

GÁBOR GOSZTOLYA[1,2], VERONIKA SVINDT[3], JUDIT BÓNA[4], ILDIKÓ HOFFMANN[3,5]

[1] University of Szeged, Institute of Informatics, Szeged, Hungary
[2] ELRN-SZTE Research Group on Artificial Intelligence, 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 Linguistics, Szeged, Hungary

1. The Focus of This Study

- Multiple sclerosis, among other symptoms, might deteriorate the patient's speech
- Due to this, automatic speech analysis can serve as a tool to detect the disease, or **monitor its progression**
- We conducted a **longitudinal** study of MS
- We employed a standard pathological speech processing workflow (wav2vec 2.0 embeddings as features, SVM as classifier in leave-one-subject-out nested cross-validation)
- We analyzed the results of the individual years, and found that the best classification performance was achieved on the recordings of the last year

2. Multiple Sclerosis

Multiple Sclerosis and Language

- A chronic inflammatory disease of the central nervous system
- Impairments in the patient's gross and fine motor skills
- cca. 60% of MS subjects have some cognitive impairments (cognitive flexibility, disorders of orientation, working-memory limitation, information processing speed) [1]
- cca. one-third of MS patients report (temporary or persistent) speech disorders
- Motor speech disorders (dysarthria, dysphonia); word finding difficulties; limitations of the higher level cognitive processes

The Multiple Sclerosis Recordings Used

- 16 MS subjects (6/10 m/fm), 12 Healthy Controls (HC) (2/10 m/fm)
- Recordings collected in **three consecutive years** (2020-2022)

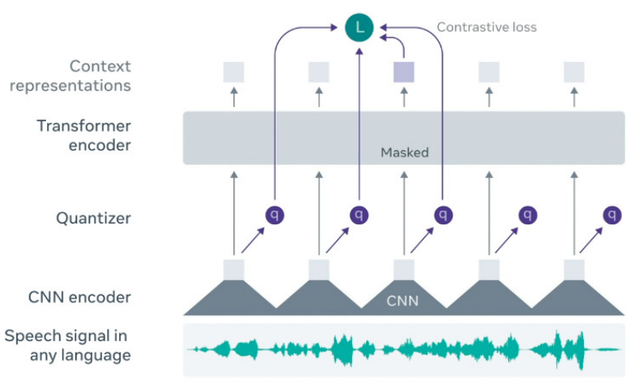
Three different speech tasks:

(1) Picture description task (**Boston Cookie Theft**)
(2) Share their opinions about vegetarianism (year 1), keeping pets in flats (year 2), advertisements (year 3) (**Opinion**)
(3) Read aloud specific non-words (CVCV sequences) (**Phonetics**)

3. Experimental Setup

Classification and Evaluation

- Features: embeddings from a wav2vec 2.0 model (XLS-53) fine-tuned on 17 hours of the target language (Hungarian)
- Embeddings from the last layers of the **convolutional** and contextualized (**fine-tuned**) blocks, aggregated over time via mean and standard deviation



- Support Vector Machines + linear kernel, **nested** cross-validation
- All 84 utterances ((16 + 12) $\times$ 3), leave-one-subject-out
- Area Under the ROC Curve (AUC) and Equal Error Rate (EER)

4. Results for All the Years

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

- The results are competitive, and similar to those of previous studies
- Contextualized embeddings are somewhat more suitable for MS detection than fine-tuned ones
- All three tasks are similar with the fine-tuned embeddings
- For the convolutional ones, the 'Opinion' task seems less suited for MS detection than 'Boston Cookie Theft' or 'Phonetics'

Main references

[1] Dobson et al., "Multiple sclerosis – a review", European Journal of Neurology 2019.
[2] Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations", Advances in Neural Information Processing Systems, 2020.

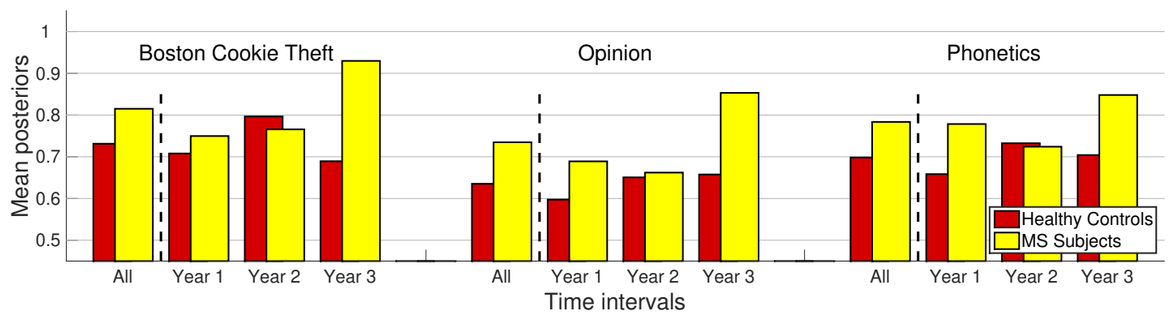
5. Results for the Individual Years

Due to data scarcity, we trained no further classifier models, only filtered the recordings (posteriors) and calculated EER / AUC for the values for the specific year.

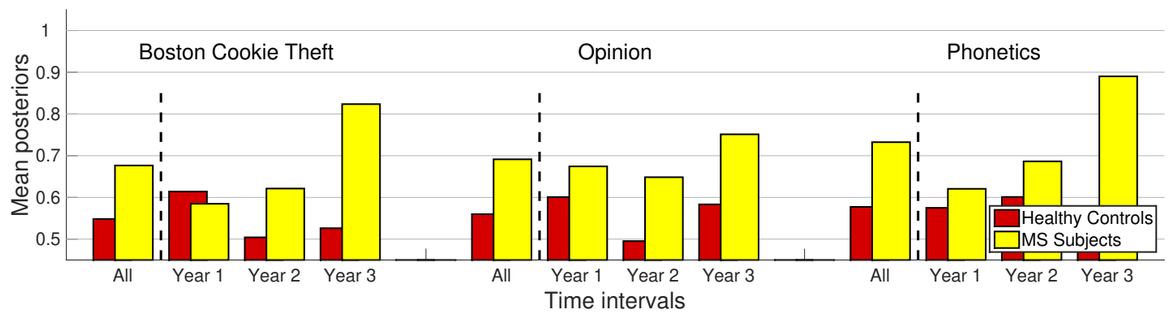
Speech Task	Period	Convolutional EER	Convolutional AUC	Fine-tuned EER	Fine-tuned AUC
Boston Cookie Theft	All years	16.7%	0.917	33.3%	0.744
	Year 1	17.9%	0.839	25.0%	0.745
	Year 2	14.3%	0.969	42.9%	0.656
	Year 3	7.1%	0.979	25.0%	0.833
Opinion	All years	28.6%	0.808	30.9%	0.787
	Year 1	32.1%	0.745	17.9%	0.885
	Year 2	32.1%	0.771	50.0%	0.641
	Year 3	17.9%	0.891	25.0%	0.833
Phonetics	All years	22.6%	0.879	33.3%	0.792
	Year 1	25.0%	0.844	50.0%	0.693
	Year 2	25.0%	0.854	32.1%	0.760
	Year 3	17.9%	0.932	7.1%	0.938

- For the convolutional embeddings, Year 1 and Year 2 are typically worse than the "All years" case
- Values for Year 3 are always better (AUC in the range 0.891...0.979, EER in the range 7.1%...17.9%)
- For the fine-tuned embeddings, the trend is similar: some variation between Year 1 and Year 2, but Year 3 always outperforms "All years"

6. Posterior Mean Statistics



Mean Posterior Values for the Convolutional Embeddings



Mean Posterior Values for the Fine-tuned Embeddings

- The mean posteriors for the convolutional embeddings are higher than those for the fine-tuned embeddings
- The values for MS subjects (yellow bars) in Year 3 are much higher than for Year 1 & Year 2
- No similar phenomenon for the HC subjects (red bars)

Speech Task	Periods	Convolutional HC	Convolutional MS	Fine-tuned HC	Fine-tuned MS
Boston Cookie Theft	Year 1 vs. 2	0.741	0.955	0.194	0.692
	Year 2 vs. 3	0.544	**0.009**	0.624	**0.010**
Opinion	Year 1 vs. 2	0.977	1.000	0.260	0.836
	Year 2 vs. 3	0.885	**0.037**	0.371	**0.044**
Fine-tuned	Year 1 vs. 2	0.260	0.337	0.751	0.720
	Year 2 vs. 3	0.471	**0.040**	0.403	**0.002**

Results with Equal Error Rate (EER)

- We used the Mann-Whitney U test to verify the significance of these differences (significant p values are shown as **bold**)
- Only the Year 2 vs. Year 3. cases are significant, and only for the MS subjects
- It might be caused from a slight deterioration of MS subjects in Year 3
- It can also be due to some speech property or acoustic artifact

7. Conclusions

- We performed a longitudinal investigation of multiple sclerosis patients and healthy control subjects
- We used 3 speech recordings from 16 MS and 12 HC subjects recorded over 3 consecutive years
- MS identification was much better for Year 3 than for the first two years
- This was also verified by posterior statistics and significance tests

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).