Wav2vec 2.0 Embeddings Are No Swiss Army Knife – A Case Study for Multiple Sclerosis

Gábor Gosztolya[1,2], Mercedes Vetráb[1], 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

- Self-supervised learning revolutionalized ASR, using pre-trained models (via self-supervised learning) and fine-tuning them on the actual (even limited) training data
- This can also be employed for paralinguistic tasks (e.g. emotion recognition), starting from the same pre-trained layers (even frozen)
- It is equivalent of using the pre-trained model for feature extraction
- In pathological speech processing, training data is even more scarce, where DNNs are suboptimal. Here, one might use SVM or XGBoost, which are usually quicker to train and are more robust
- With these classifiers, one can still use embeddings from self-supervised models as features (similarly as using embeddings from x-vector or ECAPA-TDNN networks)

We now investigate how competitive are features obtained via self-supervised models (i.e. wav2vec 2.0 embeddings) for multiple sclerosis detection, compared to several straightforward feature types.

2. Experimental Setup

The Multiple Sclerosis Recordings Used

- 23 MS subjects (all RRMS), 22 Healthy Controls (HCs)

Three different speech tasks:

(1) Share their opinions about vegetarianism (**Opinion**)
(2) Summarize a heard short historical anecdote (**Narrative recall**)
(3) Read aloud specific non-words (CVCV sequences) (**Phonetics**)

Classification

- Support Vector Machines + linear kernel, **nested** cross-validation
- Area Under the ROC Curve (AUC) and Equal Error Rate (EER)

3. The Feature Sets Employed

ComParE functionals & eGeMAPS

- Standard 'general' feature sets, based on frame-level attributes and their statistics (e.g. mean, standard deviation, peak statistics)
- ComParE functionals: 6737 attributes, eGeMAPS: 88 features

x-vectors & ECAPA-TDNN

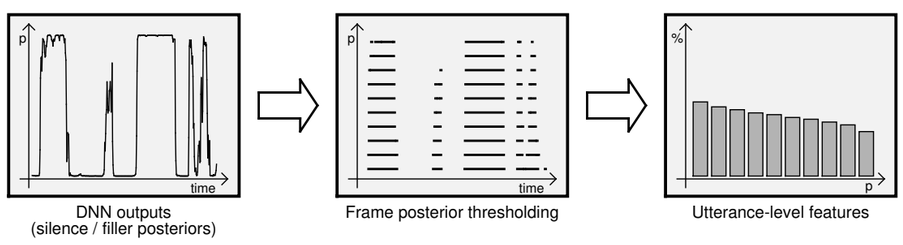
- Special-structure neural networks, trained for speaker identification
- Special *pooling layer*; embeddings taken from a layer above this point
- We used models trained on Voxceleb by the `speechbrain` team

Pause statistics

This method describes the amount of pause present in the recording

(1) A standard HMM/DNN acoustic model is evaluated
(2) The local probabilities of **silent** and **filled** pauses are calculated
(3) The ratio of frames where this probability exceeds a threshold is noted

It is repeated with thresholds between 0 and 1 with 0.02 increments.



Embeddings from a DNN Acoustic Model

- We also employed embeddings from a standard DNN acoustic model
- 5×1024 ReLU neurons trained on FBANK + Δ + $\Delta\Delta$ on 240 hours
- We used the embeddings from layer #2 (lower) and #4 (higher)
- Frame-level values were aggregated via mean and standard deviation

wav2vec 2.0 Embeddings

- A wav2vec 2.0 XLSR-53 model, fine-tuned on 8 hours
- Embeddings were taken from the last layers of the convolutional and of the fine-tuned block

Main references

[1] Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations", Advances in Neural Information Processing Systems, 2020.
[2] Eyben et al., "Opensmile: The Munich versatile and fast open-source audio feature extractor", Proc. ACM MM 2010.
[3] Pérez-Toro et al., "Alzheimer's detection from English to Spanish using acoustic and linguistic embeddings", Proc. Interspeech 2022.

4. Results

Feature set		N	Opin.	Narr.R.	Phon.	Mean
ComParE functionals		6373	0.810	0.905	0.879	0.865
eGeMAPS v02		88	0.832	0.759	0.800	0.797
x-vectors		512	0.824	0.842	0.555	0.740
ECAPA-TDNN		192	0.480	0.352	0.449	0.427
Pause statistics		150	0.702	0.534	0.745	0.660
DNN acoustic model (layer #2)	Mean	1024	0.854	0.828	0.763	0.815
	Mean + std.	2048	0.891	0.854	0.721	0.822
DNN acoustic model (layer #4)	Mean	1024	0.759	0.816	0.769	0.781
	Mean + std.	2048	0.798	0.787	0.783	0.789
wav2vec 2.0 (convolutional)	Mean	512	0.656	0.700	0.872	0.743
	Mean + std.	1024	0.713	0.729	0.874	0.772
wav2vec 2.0 (fine-tuned)	Mean	1024	0.733	0.816	0.828	0.792
	Mean + std.	2048	0.739	0.820	0.830	0.796
Average for all methods used			0.753	0.749	0.759	

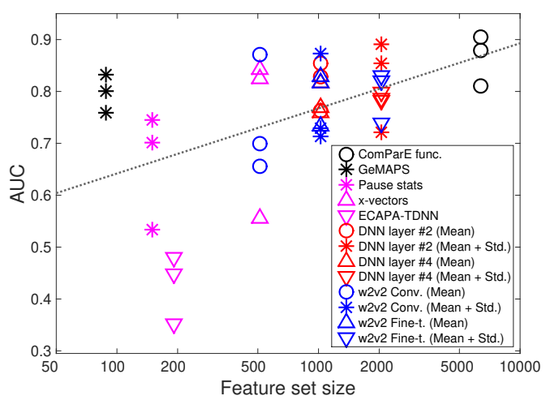
Results with Area Under the ROC Curve (AUC)

Feature set		N	Opin.	Narr.R.	Phon.	Mean
ComParE functionals		6373	26.7%	17.8%	17.8%	20.8%
eGeMAPS v02		88	22.2%	35.6%	22.2%	26.7%
x-vectors		512	22.2%	26.7%	44.5%	31.1%
ECAPA-TDNN		192	44.5%	51.1%	51.1%	48.9%
Pause statistics		150	31.1%	51.1%	35.6%	39.3%
DNN acoustic model (layer #2)	Mean	1024	17.7%	22.2%	31.1%	23.7%
	Mean + std.	2048	13.3%	22.2%	31.1%	22.2%
DNN acoustic model (layer #4)	Mean	1024	22.2%	17.8%	26.6%	22.2%
	Mean + std.	2048	26.7%	26.7%	22.2%	25.2%
wav2vec 2.0 (convolutional)	Mean	512	44.4%	31.1%	17.8%	31.1%
	Mean + std.	1024	35.6%	31.1%	22.2%	29.6%
wav2vec 2.0 (fine-tuned)	Mean	1024	31.1%	22.2%	17.8%	23.7%
	Mean + std.	2048	26.6%	26.7%	26.7%	26.7%
Average for all methods used			28.0%	29.4%	28.2%	

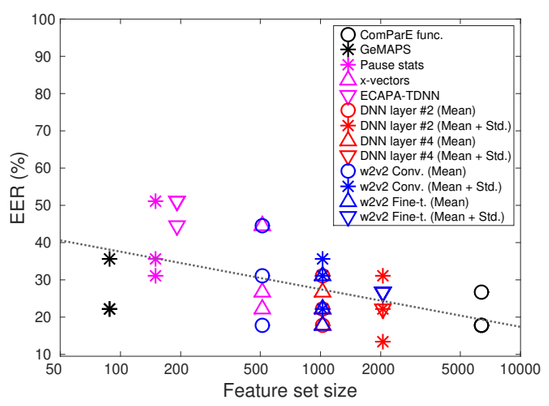
Results with Equal Error Rate (EER)

- No significant difference between the speech tasks ($p \geq 0.573$)
- wav2vec 2.0 embeddings gave mediocre performance (mean AUC between 0.743. . .0.796, mean EER between 23.7%. . .31.1%)
- Fine-tuned embeddings are somewhat better than convolutional ones
- x-vectors and (especially) ECAPA-TDNN led to low performance
- Pause statistics was not efficient either
- Embeddings from the very same DNN model were quite good (lower-level embeddings were perhaps somewhat better)
- eGeMAPS was competitive, but ComParE functionals was really good (also quite consistent over the speech tasks)

Classification Performance and the Number of Features



Feature set size vs. AUC
(CC = 0.507)



Feature set size vs. EER
(CC = -0.513)

- There is only a moderate correlation between the two values, but a trend is clearly visible
- More features is usually beneficial for classification performance
- The best model (ComParE functionals) also had the most features (although the tiny eGeMAPS also turned out to be fine)
- Including standard deviation in aggregation doubled the number of features, but helped the classification procedure only slightly
- wav2vec 2.0 was by far the largest network (317M weights), but its performance was far from outstanding

DNN Model	#Weights	#Embeddings
x-vectors	8M	512
ECAPA-TDNN	22M	192
Pause statistics	7M	150
HMM/DNN acoustic model	7M	1024
wav2vec 2.0 XLSR-53	317M	512 (conv.) 1024 (fine-tuned)

5. Conclusions

- We compared the performance of wav2vec 2.0-based features to several other methods in a pathological speech processing task
- It turned out that it is not obvious that self-supervised models automatically lead to high classification performance
- Even classic hand-crafted features can outperform them
- The no. of features correlated well with classification performance

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).