Interspeech 2024
1-5 September 2024, Kos, Greece

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

Gábor Gosztolya[1,2], *Mercedes Vetráb*[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] HUN-REN Research Center for Linguistics, 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

In the past few years, self-supervised learning has revolutionalized automatic speech recognition. Self-supervised models such as wav2vec2, due to their generalization ability on huge unannotated audio corpora, were claimed to be state-of-the-art feature extractors in paralinguistic and pathological applications as well. In this study we test embeddings extracted from a wav2vec 2.0 model fine-tuned on the target language as features on a multiple sclerosis audio corpus, using three speech tasks. After comparing the resulting classification performances with traditional features such as ComParE functionals, ECAPA-TDNN and activations of a HMM/DNN hybrid acoustic model, we found that wav2vec2-based models, surprisingly, only produced a mediocre classification performance. In contrast, the decade-old ComParE functionals feature set consistently led to high scores. Our results also indicate that the number of features correlates surprisingly well with classification performance.
Index Terms: wav2vec 2.0, feature extraction, pathological speech processing, multiple sclerosis

1. Introduction

In the past few years, self-supervised learning has revolutionalized automatic speech recognition [1]. This learning approach allows deep models to be *pre-trained* on huge amounts of unannotated data by automatically generating training labels via masking, or by defining the training objective as predicting the next observation of the speech signal. Such pre-trained models can then be *fine-tuned* on the actual task at hand, even if the amount of training material is limited. This approach was shown to deliver state-of-the-art performance on a couple of ASR tasks [2, 3]as well as for computational paralinguistic tasks (e.g. emotion recognition) [4, 5].

In many cases, fine-tuning can be done by even keeping the lower blocks of the pre-trained network frozen, practically only attaching a few layers at the top of the network [6, 7]. This setup, although it still employs one deep (and perhaps end-to-end) network, practically corresponds to using the pre-trained network as a feature extractor, while the top (fine-tuned) layers can be considered as a classifier module in their own. (Of course, there are reasons for having the whole system as one network, such as simplicity and clarity.)

In some cases, however, the amount of data is so limited that it is not feasible to use a fine-tuned classifier. Such setups most notably include pathological speech processing tasks (e.g. screening Alzheimer's Disease or Parkinson's Disease), where one subject corresponds to one machine learning example. Due to this data scarcity, in these cases it is common to employ cross-validation (or nested cross-validation), where a classifier model is trained on the data of only a few dozen subjects, while at the same time up to hundreds of classifier models have to be trained. In such cases it might worth employing other classifier methods (such as Support Vector Machines, random forests or XGBoost), which are a lot quicker to train and deliver a more robust performance even for such a small amount of data [8, 9].

Nevertheless, even in such circumstances one could utilize deep models trained in a self-supervised manner for *feature extraction*. This setup, from a theoretical point of view, is not that different from fine-tuning a pre-trained network, the only difference being that the final classifier is replaced by some model which is quicker to train and which can be expected to deliver a more robust performance. Feature extraction is not even limited to the lower blocks of the pre-trained network, but embeddings can also be obtained from the higher layers of a network fine-tuned for a specific task (e.g. ASR for the given language). This is quite similar to previous research trends where networks trained for e.g. speaker identification (such as x-vectors [10] and ECAPA-TDNN [11]) were used as feature extractors to identify various diseases (e.g. Parkinson's Disease) [9, 12].

Perhaps the most widely-used self-supervised speech processing network type is wav2vec 2.0 [2], which was employed both by fine-tuning on the given task [3, 4] and as a feature extractor [8, 13, 14]. It is common to view it as a state-of-the-art feature extractor in paralinguistic and pathological applications as well. This study seeks to measure how much truth this view holds. We tested embeddings extracted from a wav2vec 2.0 model fine-tuned on the target language (Hungarian) as features on a multiple sclerosis audio corpus, using three speech tasks. Besides wav2vec 2.0 embeddings, we also tested various common feature extraction methods: ComParE functionals [15], eGeMAPS [16], x-vectors [10], ECAPA-TDNN [11], pause statistics [17] and embeddings obtained from DNN acoustic models of a HMM/DNN hybrid network [18].

2. The multiple sclerosis corpus

All the tests were carried out at the Neurology Department of Uzsoki Hospital, Budapest, Hungary, and at the Research Institute 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. In the current

10.21437/Interspeech.2024-995

study we use the recordings of 23 MS subjects (5 males and 18 females, mean age 39 ± 8.11 years) and 22 healthy controls (6 males and 16 females, mean age 39.95 ± 7.22 years). 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 display 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, consisting of 17 different speech tasks (overall roughly 50 minutes). Due to space restrictions, now we narrowed it down to using three speech tasks, selected by their difference in the cognitive processes involved:

- **Opinion**: The subjects were asked to share their opinions about vegetarianism.
- **Narrative Recall**: The subjects listened to a two-minute-long anecdote that was unknown to them beforehand, and they had to summarize the story as accurately as possible.
- **Phonetics**: The subjects were asked to read aloud several specific non-words (consonant-vowel-consonant-vowel (CVCV) sequences, where the first CVs contained a voiceless plosive [p, t, k] and one of the vowels [i:, a:, u:]).

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. Methods

3.1. ComParE functionals & eGeMAPS

As the first feature extraction approach, we used the 'ComParE functionals' attributes developed by Schuller et al [19]. The feature set includes energy, spectral, cepstral (MFCC) and voicing related frame-level features, from which specific functionals (like the mean, standard deviation, 1st and 99th percentiles or peak statistics) are computed to provide 6373 utterance-level feature values. This feature set is still frequently applied both for pathological and paralinguistic tasks [20, 21].

Later, a minimalistic parameter set was also proposed, consisting only of 18 frame-level attributes (such as pitch, jitter, shimmer and formant frequencies), which was extended to 25 (by adding e.g. the first four MFCC components). By taking the mean of these frame-level attributes along with the coefficient of the variation for specific attributes, the so-called *extended Geneva Minimalistic Acoustic Parameter Set* (or *eGeMAPS*) was introduced (consisting of only 88 features) [16]. This feature set is also commonly used for paralinguistic and pathological speech processing applications [20, 22, 23]

Both feature sets were extracted by using the python port of the openSMILE tool [15].

3.2. x-vectors and ECAPA-TDNN

x-vectors are neural networks with a special structure, consisting of lower frame-level layers and higher segment-level (i.e. utterance-level) layers. The connection between the two parts is established by a special *statistics pooling* layer, which calculates the mean and the standard deviation of the activations of the last frame-level layer. This allows utterance-level training over frame-level features even for variable-length utterances [10]. This network is typically trained for speaker recognition; when used as a feature extractor, the activations of a segment-level layer are used as features [24, 25]. The ECAPA-TDNN model improves over the x-vector architecture on several points: it contains channel- and context-dependent attention mechanism, multi-layer feature aggregation and residual blocks [11]. It is in general considered to be superior to the x-vector architecture in speaker recognition and diarization [26], and it is commonly employed as a feature extractor as well [9].

We used the models `spkrec-xvect-voxceleb`[1] and `spkrec-ecapa-voxceleb`[2] (both released by the official `speechbrain` team [27]), both trained on the Voxceleb1 + Voxceleb2 training data. The networks calculate 512 and 192 embeddings (features), and have 8.2M and 22.2M million weights, x-vectors and ECAPA-TDNN, respectively.

3.3. Pause statistics

This method describes the amount of pause present in a given recording, which is known to be relevant in acoustic detection of multiple diseases [17, 28, 29]. In general, we distinguish two types of pauses: silent and filled pauses, where the latter correspond to vocalizations like 'um', 'uh', 'er' etc. The method consists of the following steps:

(i) A standard Deep Neural Network acoustic model (from a HMM/DNN hybrid model) is evaluated on the actual utterance, using frame-level features (e.g. MFCCs).

(ii) Based on the outputs provided by the DNN, we estimate the local posterior probability values of silent and filled pauses by adding up the posterior estimates of the corresponding phonetic classes (still at the frame level).

(iii) From the local posterior estimates calculated in step (ii), new representations are computed at the utterance level by calculating the ratio of frames where the posterior estimate of a particular pause type exceeds a given threshold.

Step (iii) was performed with thresholds between 0 and 1 with 0.02 increments, and was done for the silent pauses, for the filled pauses, and for both pause types (i.e. adding up the posterior estimates of the two types of pause), resulting in a total of $3 \times 50 = 150$ features for each utterance.

The acoustic DNN (employed in step (i)) was trained on a subset of the BEA Hungarian Spontaneous Speech Dataset [30]: 60 hours of data from 165 speakers was augmented by adding noise, background babble and reverberation, increasing the total amount of training data to 240 hours.Log-energies of Mel-scale filter banks along with $\Delta+\Delta\Delta$ were used as features over a 150 ms wide sliding window. The network had five hidden layers, each containing 1024 ReLU neurons, while it had 911 neurons corresponding to context-dependent phonetic states in the output layer (where we used the softmax activation function). This network consists of about 7M weights.

3.4. Embeddings from a DNN acoustic model

We also employed a standard feed-forward DNN acoustic model of a traditional HMM/ANN hybrid model, as this approach was shown to be an efficient feature extactor [18]. To do this, one first has to train such a model for a standard ASR corpus for frame-level phoneme identification, using traditional frame-level features like MFCCs or raw energies of Mel-frequency filter banks (FBANK). The embeddings of some hidden layer, obtained by evaluating the network on the target utterances, might serve as features for a further classification task. These frame-level activations should be aggregated over

[1] https://huggingface.co/speechbrain/spkrec-xvect-voxceleb

[2] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

Table 1: *Number of attributes (N), Equal Error rate (EER) and AUC values for the various feature sets sed for the three speech tasks*

Feature set		N	Opinion		Narrative Recall		Phonetics		Mean
			EER	**AUC**	**EER**	**AUC**	**EER**	**AUC**	**AUC**
ComParE functionals		6373	26.7%	0.810	17.8%	0.905	17.8%	0.879	0.865
eGeMAPS v02		88	22.2%	0.832	35.6%	0.759	22.2%	0.800	0.797
x-vectors		512	22.2%	0.824	26.7%	0.842	44.5%	0.555	0.740
ECAPA-TDNN		192	44.5%	0.480	51.1%	0.352	51.1%	0.449	0.427
Pause statistics		150	31.1%	0.702	51.1%	0.534	35.6%	0.745	0.660
DNN acoustic model (layer #2)	Mean	1024	17.7%	0.854	22.2%	0.828	31.1%	0.763	0.815
	Mean + std.	2048	13.3%	0.891	22.2%	0.854	31.1%	0.721	0.822
DNN acoustic model (layer #4)	Mean	1024	22.2%	0.759	17.8%	0.816	26.6%	0.769	0.781
	Mean + std.	2048	26.7%	0.798	26.7%	0.787	22.2%	0.783	0.789
wav2vec 2.0 (convolutional)	Mean	512	44.4%	0.656	31.1%	0.700	17.8%	0.872	0.743
	Mean + std.	1024	35.6%	0.713	31.1%	0.729	22.2%	0.874	0.772
wav2vec 2.0 (fine-tuned)	Mean	1024	31.1%	0.733	22.2%	0.816	17.8%	0.828	0.792
	Mean + std.	2048	26.6%	0.739	26.7%	0.820	26.7%	0.830	0.796
Average for all methods used			28.0%	0.753	29.4%	0.749	28.2%	0.759	

the whole utterance with functionals like mean and standard deviation, resulting in an utterance-level feature set the size of a few times the number of neurons in the given hidden layer.

Our actual model was the same as we used for calculating the pause statistics attributes (see Section 3.3). Out of the five hidden layers, we used the embeddings of layer #2 and #4; the former one can be expected to capture low-level information, while the second one can be expected to detect higher-order phenomena. We used the mean and the mean + standard deviation to aggregate the frame-level embeddings, resulting in 1024 and 2048 utterance-level features for both layers, respectively.

3.5. wav2vec 2.0

Next, we employed wav2vec 2.0 embeddings as features. The wav2vec network processes raw audio signals as input, and is trained in a self-supervised manner, where it learns to predict future observations for the given speech sample [1]. The wav2vec 2.0 architecture further enhances this by using masking during training, and employing a quantizer (which selects speech units from an inventory of learned units) and a transformer network (incorporating information from the whole utterance) [2].

The two straightforward layers to extract embeddings for feature extraction are the last layer of the convolutional block, and the last layer of the fine-tuned block. These two types of features might carry different kinds of relevant information: the convolutional layer (located at the lower region of the network) can be expected to capture lower-level information, while the fine-tuned layer can be expected to store more phonetic-related information [31].

Our wav2vec2 model[3] is based on the `XLSR-53` model pre-trained by Facebook on 53 languages simultaneously [32] (consisting of about 317M weights). This base model was fine-tuned on the Hungarian part of the Mozilla Common Voice 6.1 corpus (8 hours). The last layer of the convolutional block consists of 512 neurons, while the last layer of the fine-tuned block has 1024 neurons. Similarly to the case of the DNN acoustic model (see Section 3.4), we calculated the mean and the mean + standard deviation of the frame-level embedding vectors, obtaining 512, 1024 and 2048 utterance-level features overall.

[3]https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-hungarian

4. Experimental Setup

We utilized Support Vector Machines (SVM) to predict whether the speakers belonged to the MS or the HC group. We used the libSVM implementation [33] with a linear kernel (nu-SVR method); the C complexity parameter was set in the range 10^{-5}, ..., 10^{1}. Due to the small number of examples, we performed cross-validation (CV), being common in pathological speech processing (see e.g. [8, 14]), where one fold always consisted of the features of one control subject and one having MS. To avoid any form of peeking, we employed *nested cross-validation* [34]: each time we trained our model on the data of 22 folds, *another* (22-fold) cross-validation session was performed, to find the C meta-parameter value that gave the highest AUC score. Afterwards, we trained an 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.

Before classification, all the feature sets were standardized to zero mean and unit variance. Performance was measured via Equal Error Rate (EER) and Area Under the ROC Curve (AUC).

5. Results

Table 1 shows the EER and AUC scores obtained for the different feature sets and for the three speech tasks, the mean of the three AUC values and the size of the feature vectors (N). First, notice that there is no significant difference between the speech tasks, at least from the perspective of automatically identifying the MS subjects: the EER values, averaged over all approaches tested (see the last line of Table 1) lie in the range $28.0\% \ldots 29.4\%$, while the AUC scores are even closer to each other ($0.749 \ldots 0.759$). We verified the lack of any significant difference by the Mann-Whitney U test (or Wilcoxon rank-sum test, see [35]), and we found the p values fall between 0.573 and 0.939 (so $p > 0.05$) for all pairwise comparisons.

However, we found notable differences among the values obtained by the various feature sets. (Unfortunately, as both SVM and most of the used feature extraction approaches are deterministic, we were unable to verify the significance of the performance differences by statistical tests.) Regarding the focus of our study, the wav2vec 2.0 models gave mediocre performance compared to the other tested methods, with mean AUC values between 0.743 and 0.796. The convolutional embed-

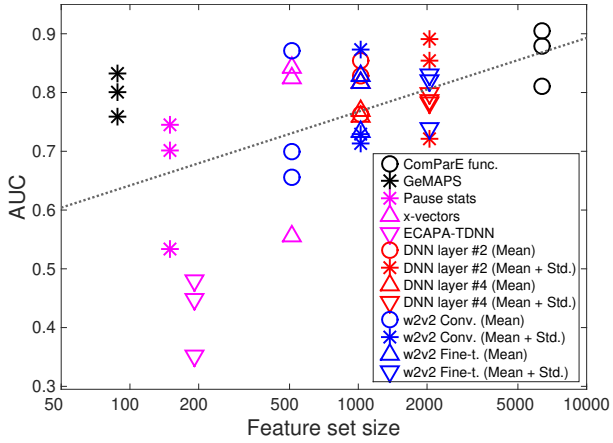

Figure 1: *Feature set sizes (on logarithmic scale) versus the AUC score measured on the three tasks.*

dings were somewhat less effective than the fine-tuned ones, although we obtained the highest AUC scores (0.872 and 0.874 for *Phonetics*) with these embeddings. Regarding the aggregation functions, there were no huge differences.

The embeddings of the DNN acoustic model were perhaps the most similar to the wav2vec 2.0 embeddings in nature, but the trends found were actually quite different. Taking embeddings from the lower region of the network (i.e. layer #2) was a better strategy in this case than taking higher-order embeddings (i.e. from layer #4). The most difficult speech task for these features was actually the reading task *Phonetics*, with EER values between 22.2% and 31.1%, outperformed by both spontaneous speech tasks (i.e. *Opinion* and *Narrative Recall*). In contrast, we obtained the best scores for all types of wav2vec 2.0 embeddings on the *Phonetics* task.

The pause statistics method produced quite low scores as well, especially when we compare the EER and AUC values to those obtained via the DNN acoustic model embeddings. (Recall that the very same network was used for these feature extraction strategies.) This suggests that it is not worth incorporating expert knowledge (i.e. pause-related information) into the feature extraction step, since a better classification performance can be achieved by using raw statistical data (i.e. mean and standard deviation of the activations).

Quite surprisingly, the most effective feature set was the 'ComParE functionals' with a mean AUC score of 0.865; also notice that the performance was quite consistent over the three speech tasks, the lowest AUC score being 0.810 and the highest EER value being 26.7% (for the *Opinion* task). eGeMAPS was less effective, but the mean AUC value 0.797 is still competitive.

x-vectors turned out to be acceptable for two speech tasks (AUC values over 0.800), but performed quite badly for *Phonetics*. ECAPA-TDNN, however, gave really bad results, with AUC values below 0.500 and EER scores around 50% in all cases. This, in our opinion, indicates that ECAPA-TDNN, as opposed to its precedessor x-vectors, is just not robust enough to be employed as a feature extractor for a pathological speech processing task (or, at least, for multiple sclerosis subjects). (As the x-vector and the ECAPA-TDNN models were trained by the same group on the same data, it is straightforward to attribute their performance differences only to the network structure.)

Figure 1 shows the AUC values as a function of the (logarithmic) feature set sizes. Although there is only a moderate correlation between the two values, a trend is clearly visible: having more features is usually beneficial for classification performance. This is reinforced by the high AUC values of our largest attribute set tested (ComParE functionals with 6373 attributes). However, there are some exceptions to this trend, like the most compact attribute set (eGeMAPS) also leading to competitive scores. Also, using both mean and standard deviation for aggregation (which doubled the number of utterance-level attributes) usually brought only slight improvements, or in some cases it outright harmed classification performance (especially for the EER metric). Feature size and AUC had a correlation coefficient of 0.507, which indicates a moderate-strength relation. (We found a similar connection between feature set size and Equal Error Rate, with a correlation coefficient of -0.513.)

6. Conclusions

In this study we used features calculated from wav2vec 2.0 embeddings to detect multiple sclerosis from speech. For a large-scale comparison, we also used a number of other feature sets, employing both hand-crafted ones and embeddings from different neural networks. We found that wav2vec 2.0 embeddings were by no means the most competitive features. Although they outperformed x-vectors, ECAPA-TDNNs and statistical features describing the amount of pause in the utterance, they lagged behind embeddings taken from a standard DNN acoustic model, the ComParE functionals feature set, and they could not outperform the tiny eGeMAPS attribute set either (which consists of only 88 attributes). This indicates that hand-crafted features are not necessarily outdated, as they might be a better solution to obtain reliable and effective features.

We also investigated the relation of classification performance and the number of features. We found moderate correspondence (with coefficients of -0.513 and 0.507, EER and AUC, respectively), which was confirmed by the observation that taking means and standard deviations of frame-level embeddings was more effective than using the means alone. Another factor worth considering, at least for DNN-based feature extraction approaches, is network size. The employed wav2vec 2.0 model consists of over 300 million weights, and although it was able to outperform the x-vector (8M weights) and the ECAPA-TDNN (22M weights) networks, it lagged behind the DNN acoustic model, having only 7M (or, by the second hidden layer, only 2.9M) weights. The fact that wav2vec2 embeddings were outperformed both by a decade-old hand-crafted attribute set and by a standard (and, by today's standards, tiny) DNN acoustic model, in our opinion, indicates that it is not obvious that self-supervised models automatically lead to high classification performance, but (depending on the actual task) other techniques might turn out to be more competitive alternatives.

7. Acknowledgements

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).

8. References

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

[2] 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. 12 449–12 460, 2020.

[3] P. Mihajlik, A. Balog, T. E. Gráczi, A. Kohári, B. Tarján, and K. Mády, "BEA-Base: A benchmark for ASR of spontaneous Hungarian," in *Proceedings of LREC*, 2022, pp. 1970–1977.

[4] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *Proceedings of ICASSP*, Rhodes Island, Greece, 2023.

[5] T. Grósz, A. Virkkunen, D. Porjazovski, and M. Kurimo, "Discovering relevant sub-spaces of BERT, Wav2Vec 2.0, ELECTRA and ViT embeddings for humor and mimicked emotion recognition with integrated gradients," in *Proceedings of MuSe*, Ottawa, Canada, 2023, pp. 27–34.

[6] A. Romana and K. Koishida, "Toward a multimodal approach for disfluency detection and categorization," in *Proceedings of ICASSP*, Rhodes Island, Greece, 2023.

[7] B. W. Schuller, A. Batliner, S. Amiriparian, A. Barnhill, M. Gerczuk, A. Triantafyllopoulos, A. E. Baird, P. Tzirakis, C. Gagne, A. S. Cowen, N. Lackovic, M. J. Caraty, and C. Montacié, "The ACM Multimedia 2023 Computational Paralinguistics Challenge: Emotion share & Requests," in *Proceedings of ACM Multimedia*, Oct 2023, pp. 9635–9639.

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

[9] D. Sztahó and A. Fejes, "Effects of language mismatch in automatic forensic voice comparison using deep learning embeddings," *Journal of Forensic Sciences*, vol. 88, no. 3, pp. 871–883, 2023.

[10] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep Neural Network embeddings for text-independent speaker verification," in *Proceedings of Interspeech*, Stockholm, Sweden, Aug 2017, pp. 999–1003.

[11] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN based speaker verification," in *Proceedings of Interspeech*, 2020, pp. 3830–3834.

[12] L. Moro-Velazquez, J. Villalba, and N. Dehak, "Using x-vectors to automatically detect Parkinson's Disease from speech," in *Proceedings of ICASSP*, Barcelona, Catalonia, Spain, 2020, pp. 1970–1977.

[13] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Proceedings of Interspeech*, Brno, Czechia, 2021, pp. 3400–3404.

[14] J. V. Egas-López, V. Svindt, J. Bóna, I. Hoffmann, and G. Gosztolya, "Automated multiple sclerosis screening based on encoded speech representations," in *Proceedings of Interspeech*, Dublin, Ireland, Aug 2023, pp. 3003–3007.

[15] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of ACM Multimedia*, 2010, pp. 1459–1462.

[16] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing," *IEEE Transactions for Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[17] J. V. Egas-López, R. Balogh, N. Imre, I. Hoffmann, M. K. Szabó, L. Tóth, M. Pákáski, J. Kálmán, and G. Gosztolya, "Automatic screening of mild cognitive impairment and Alzheimer's disease by means of posterior-thresholding hesitation representation," *Computer, Speech & Language*, vol. 75, no. Sep, 2022.

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

[19] B. W. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The Interspeech 2013 Computational Paralinguistics Challenge: Social signals, Conflict, Emotion, Autism," in *Proceedings of Interspeech*, Lyon, France, Sep 2013, pp. 148–152.

[20] B. Mirheidari, A. Bittar, N. Cummins, J. Downs, H. L. Fisher, and H. Christensen, "Automatic detection of expressed emotion from five-minute speech samples: Challenges and opportunities," in *Proceedings of Interspeech*, 2022, pp. 2458–2462.

[21] Z. Liu, M. Huckvale, and J. McGlashan, "Automated voice pathology discrimination from continuous speech benefits from analysis by phonetic context," in *Proceedings of Interspeech*, 2022, pp. 2158–2162.

[22] D. Woszczyk, A. Hedlikova, A. Akman, S. Demetriou, and B. Schuller, "Data augmentation for dementia detection in spoken language," in *Proceedings of Interspeech*, 2022, pp. 2858–2862.

[23] S. Fara, O. Hickey, A. Georgescu, S. Goria, E. Molimpakis, and N. Cummins, "Bayesian Networks for the robust and unbiased prediction of depression and its symptoms utilizing speech and multimodal data," in *Proceedings of Interspeech*, Dublin, Ireland, 2023, pp. 1728–1732.

[24] M. Huckvale, A. Beke, and M. Ikushima, "Prediction of sleepiness ratings from voice by man and machine," in *Proceedings of Interspeech*, Shanghai, China, Oct 2020, pp. 4571–4575.

[25] J. V. Egas-López and G. Gosztolya, "Deep Neural Network embeddings for the estimation of the degree of sleepiness," in *Proceedings of ICASSP*, Toronto, Canada, Jun 2021.

[26] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, and H. Na, "ECAPA-TDNN embeddings for speaker diarization," in *Proceedings of Interspeech*, Brno, Czechia, 2021, pp. 3560–3564.

[27] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[28] R. Balogh, N. Imre, G. Gosztolya, I. Hoffmann, M. Pákáski, and J. Kálmán, "The role of silence in verbal fluency tasks – a new approach for the detection of mild cognitive impairment," *Journal of the International Neuropsychological Society*, vol. 29, no. 1, pp. 46–58, 2023.

[29] V. Svindt, G. Gosztolya, and T. E. Gráczi, "Narrative recall in relapsing-remitting multiple sclerosis: A potentially useful speech task for detecting subtle cognitive changes," *Clinical Linguistics & Phonetics*, vol. 37, no. 4-6, pp. 549–566, 2023.

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

[31] 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.

[32] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proceedings of Interspeech*, 2022, pp. 2278–2282.

[33] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, 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.

[35] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.