INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Automated Multiple Sclerosis Screening Based on Encoded Speech Representations

José Vicente Egas-López[1,2], *Veronika Svindt*[3], *Judit Bóna* [4], *Ildikó Hoffmann*[3,5], *Gábor Gosztolya*[1,2]

[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 Psychiatry, Szeged, Hungary

{ egasj, ggabor } @ inf.u-szeged.hu

Abstract

Multiple Sclerosis (MS) is a chronic disease affecting over 2.5 million people worldwide. Its early detection is crucial for the management and treatment of the disease. Here we present an approach for automatic MS screening based on encoded speech representations. Our methods rely on Wav2Vec2 models to extract relevant traits from speech recordings of patients, which are then fed into a Support Vector Machine. Besides employing Wav2Vec2 models pre-trained on large public corpora, we also fine-tune them on 85 hours of the target language (Hungarian) in two distinct ways: for ASR and for speaker identification. Both variations outperformed the original models and conventional methods (ComParE functionals, x-vectors, and ECAPA-TDNN). Our findings suggest that fine-tuning for the actual speaker provides more advantages than the typical approach of fine-tuning for ASR purposes. Still, we improved our best MS discrimination performance when we fused features from our two fine-tuned models.

Index Terms: speech processing, multiple sclerosis detection, wav2vec2, fine-tuning

1. Introduction

Multiple sclerosis is a chronic inflammatory disorder that affects the central nervous system. One of the main diagnostic characteristics of MS is the impairment of motor skills, which may indicate a decline in the patient's condition. As language, cognitive, and motor skills are interconnected in the brain, changes in one area may lead to changes in other areas as well. Therefore, monitoring changes in speech production could be an effective way to track the progression of the disease. Motor speech disorders such as dysarthria and dysphonia are commonly reported by patients with MS experiencing temporary or persistent speech difficulties [1]. Furthermore, word-finding difficulties, a limitation of verbal fluency [2], sentence repetition problems, and limitations of the higher-level language processes [3] are known to affect MS patients. Automatic speech analysis may be able to detect symptoms even before dysarthria develops [4]. It is known that dysarthria may cause changes in the rhythm and timing of speech, and also in the articulation's strength and clarity [5].

Screening multiple sclerosis using speech analysis has been addressed only by a handful of studies in the last few years. An et al. [6] used Convolutional Neural Networks (CNN) for lateral sclerosis (LS) early detection using speech. LS, akin to MS, may also affect the speech production of patients [7]. Gosztolya et al. [8] used acoustic deep neural network (DNN) embeddings for automatic MS assessment. In this study, we propose the use of contextual and convolutional embeddings derived from self-supervised architectures. In particular, we will rely on representations computed with Wav2Vec2 [9] for MS discrimination. In the context of automatic MS screening from the speech, Wav2Vec2 encodings can be employed to analyze changes in speech production over time, potentially providing a non-invasive and objective measure of the disease.

Wav2Vec2 has demonstrated state-of-the-art performances in Automatic Speech Recognition (ASR) [9, 10]. It builds on the principles of its predecessor, Wav2Vec, which seeks to create new forms of input vectors from raw, unlabeled audio data, which can then be utilized to construct an acoustic model [11]. Wav2Vec2 takes a step further by encoding speech representations from masked audio segments and passing them to a transformer network that builds contextualized representations. This method has been successfully applied in computational paralinguistics and pathological speech tasks, where pre-trained models were employed to estimate emotions [12], to assess Alzheimer's Disease [13], and even to detect COVID-19 [14].

Our proposed framework involves two strategies: fine-tuning on speech units, and fine-tuning on speakers. Additionally, we rely on pre-trained Wav2Vec2 models. In detail, we present the following contributions to the automation of multiple sclerosis assessment by means of encoded speech representations: 1) Examining the effectiveness of speech embeddings produced by distinct pre-trained Wav2Vec2 models after being fine-tuned; 2) Exploring how the feature extraction quality differs when the model is fine-tuned on speech units as in ASR, and fine-tuned on speakers as in speaker recognition; 3) Analyzing the sufficiency of different pre-trained Wav2Vec2 models as feature extractors for MS detection; 4) Investigating the robustness of both language-domain matching and cross-lingual models for the original language of the corpus used.

Our results suggest that fine-tuned Wav2Vec2 encoded speech representations can effectively identify relevant information for automatic multiple sclerosis discrimination. In this paper, we show that our approaches surpass the performances of well-known computational paralinguistic techniques like the ComParE functionals representations [15], as well as speaker verification methods like ECAPA-TDNN [16] and x-vector embeddings [17]. These approaches are used in our baseline systems.

10.21437/Interspeech.2023-234

2. The Corpus

The utterances were recorded 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 linguistic protocol for collecting the speech samples from the subjects were quite extensive, consisting of 17 different speech tasks. Here, we use the *narrative recall* task, which consisted of the subjects listening to a two-minute-long anecdote that was unknown to them beforehand. The task was to summarize the story heard as accurately as possible. *Narrative recall* requires a set of cognitive processes, such as focused attention, working memory, temporal orientation, organization, and sequencing [18]. All the subjects were native Hungarian speakers; we use the recordings of 23 MS subjects, and 22 healthy controls. The recordings were converted to 16 kHz mono with a 16-bit resolution.

3. Encoded Speech Representations

Models may learn from orders of magnitude more data thanks to self-supervised learning, which is essential for identifying and comprehending patterns in less prevalent representations. To operate successfully, voice recognition systems typically need enormous amounts of training data that has been transcribed [19]. We can handle this issue by pre-training neural networks, which is especially helpful when there is a scarcity of labeled data. This way, a model can learn general representations from massive volumes of information and then be applied to downstream tasks having fewer samples.

3.1. Wav2Vec2

Wav2Vec2, similar to its successor Wav2Vec [11], uses a self-supervised approach to learn representations from raw audio. It learns to predict the correct speech unit, but in this case, it does so for masked chunks of the audio. More specifically, Wav2Vec2 encodes raw audio using a block of convolutional neural networks, then akin to masked language modeling, it masks small segments (shorter than phonemes) of the latent speech representations (which are its output). These representations are inputted to a quantizer as well as to a transformer network. The former, based on an inventory of learned units, selects a speech unit for the latent audio representation, while the latter appends data from the whole utterance. In the end, the transformer network is exposed to a contrastive loss function [9]. During training, the model learns discrete speech units by means of a Gumbel softmax that chooses the quantized representations. After pre-training is done, the model is fine-tuned using labeled data relying on a Connectionist Temporal Classification (CTC) loss for sequence alignment.

3.2. Cross-lingual Representation Learning

A multi-lingual representation approach based on Wav2Vec2 named XLSR (Cross-lingual Speech Representations) addresses the issue for languages with a limited amount of *unlabeled* data. XLSR aims to pre-train a model on multiple corpora from different languages simultaneously. This approach has a similar structure to Wav2Vec2, meaning that it is trained to jointly learn context representations along with a discrete vocabulary of latent speech audio representations. The XLSR architecture differs from that of the Wav2Vec2 in the quantization module. This module in XLSR delivers multilingual quantized speech units; then these embeddings are fed to the transformer block which uses them as targets to learn via a contrastive task. This way, the model is capable of distributing discrete tokens across different languages [20].

4. The Experiments

In our experiments, we did not perform classification by extending the Wav2Vec2 architecture with one last linear layer via the sequence classification interface (e.g., as shown in [21]), since this would have been counterproductive given the limitation on the size of the MS corpus. Hence, we opted for extracting encoded representations generated by pre-trained models, as well as by models that we fine-tuned; and performed discrimination via SVM classifiers. We got the sequence of feature vectors from the last convolutional layer of the multi-layer convolutional block that constructs the low-level module of Wav2Vec2. Also, we fetched the outputs from the second block, that is, the sequence of hidden states (contextualized feature vectors). These two types of feature vectors, the *convolutional embeddings*, and the *contextualized representations* may capture relevant information related to speakers [22] and also information encoded in the speech signal [21].

Early studies showed that pre-trained Wav2Vec2 models fine-tuned for English ASR may have an effect on the quality of encoded speech representations for tasks non-related to speech recognition, and also when there exists a language-domain mismatch, especially on those encodings taken from the contextual block of the model [12, 23]. This is different from cross-lingual approaches, which may also be capable of capturing traits relevant to computational paralinguistic tasks [13, 23]. Thus, we opted to dispense with pre-trained models that were fine-tuned for English ASR, and fine-tuned our own instances. We discuss our fine-tuning strategies in Section 4.2

4.1. Pre-trained Wav2Vec2 Models

Here, we employed four distinct cross-lingual pre-trained models to extract speech representations for MS discrimination. The first model was *wav2vec2-large-xlsr-53*, pre-trained on speech in 53 different languages [20]. The second and third cross-lingual models, *wav2vec2-XLS-R-300M* and *wav2vec2-XLS-R-1B*, were pre-trained on 128 languages [24]. To keep the size of the models comparable, we utilized wav2vec2-XLS-R's two smaller networks only (i.e., 300 million and 1 billion parameters), and avoided using the 2-billion-parameter version. Lastly, we relied on *wav2vec2-large-xlsr-53-hu* as the fourth model, which was pre-trained on 53 languages and fine-tuned for Hungarian ASR using around 19 hours of speech corpora (CommonVoice and CSS10).

The two blocks that make up the Wav2Vec2 architecture, output embeddings with variable-length sequences. In order to aggregate such encodings into fixed-size representations, we relied on statistical methods such as the mean, standard deviation, and maximum as pooling strategies. After conducting preliminary tests, we report the results for all our experiments using mean only.

4.2. Fine-tuned Wav2Vec2 Models

We employed two strategies for fine-tuning Wav2Vec2: first, we fine-tuned Wav2Vec2 for ASR using Hungarian corpora, which aligned with the language of the MS corpus; and second, we fitted Wav2Vec2 to perform classification at the utterance level during fine-tuning. Our motivation of employing the second strategy relied on the fact that Wav2Vec2 demonstrates state-of-the-art performance in extracting meaningful representations from speakers after being adapted to speaker recognition tasks [25]. Furthermore, it is worth noting that former (SOTA) deep learning-based speech recognition methods, such as x-vectors, have also been successfully adapted to paralinguistic and pathological speech tasks before [26, 27]. This fostered our confidence in the possibility of obtaining high-quality embeddings by fine-tuning Wav2Vec2 on speakers (i.e., an equivalent approach).

The limited size of the MS dataset prevented us from performing fine-tuning effectively. Hence, we made use of a portion of the BEA corpus [28], which contains Hungarian speech; this also allowed us to match the language-domain of the intended task. The subset included a total of 85 hours of spontaneous speech. Given that cross-lingual models tend to generate higher quality embeddings and they may be able to identify more paralinguistic information than their mono-lingual counterparts [23, 29], we chose the *wav2vec2-large-xlsr-53* pre-trained 300m parameter-model as the base for fine-tuning.

Our fine-tuning framework involved the following: *i)* we built our downstream model for Hungarian ASR optimized with CTC loss [30], and fine-tuned it with its low-level feature extractor part (i.e., the CNN blocks) frozen, as it was sufficiently fitted during pre-training [9]. And, *ii)* given that the BEA corpus consists of speaker-wise annotated data, we experimented by fine-tuning our *second* model on speakers. Despite our main task not being related to speaker verification, the proposed fine-tuning approach can be considered similar for feature extraction purposes. Here, this allowed us having a scenario where the model was fine-tuned on speakers rather than on speech units as in ASR. Overall, this would be the most convenient scenario if the MS corpus had an appropriate size for fine-tuning Wav2Vec2 architectures, however, this was not the case. For the fine-tuning process, we modified Wav2Vec2's sequence classification interface by adding a pooling layer for gathering information at utterance-level during fine-tuning. These 'pooled' encodings were sent to a fully connected layer as input for classifying (BEA) speakers with cross-entropy loss. We relied on the mean for the pooling method. And, similar to our first strategy, we froze the feature extractor for fine-tuning as well.

4.3. Baseline Systems

We relied on three different techniques as competitive baselines. *First*, we used a former state-of-the-art speaker verification (SV) method: the *x-vector* approach [17]. This technique has been adopted by a wide variety of speech analysis fields ranging from emotion recognition [26] to various pathological speech processing tasks [31, 32]. We trained our x-vector extractor on the same BEA Hungarian subset used for Wav2Vec2 fine-tuning. We used 40 MFCCs as frame-level features. For our *second* baseline, we relied on *ECAPA-TDNN* [16]; built upon x-vectors, it is the current state-of-the-art in SV. We employed a model that was pre-trained on Voxceleb2 and CN-Celeb [16]. As a *third* baseline, we computed *ComParE functionals* features [33], which is a popular choice in speech processing tasks [34, 35, 15]. These include energy, spectral, cepstral (MFCC) and voicing related frame-level attributes, which serve as the base of utterance-level aggregation by specific functionals (e.g., the mean, standard deviation, 1st and 99th percentiles, peak statistics etc.).

Table 1: *The Area-Under-the-Curve (AUC) and Equal Error Rate (EER) scores obtained on the MS corpus. Both contextual and convolutional embeddings are reported.*

Model	**Embedding**	**AUC**	**EER**
x-vectors (baseline)	-	0.752	29.57%
ecapa-tdnn (baseline)	-	0.685	32.18%
ComParE fun. (baseline)	-	0.739	30.01%
Pre-trained			
wav2vec2-xls-r-300m	convolutional	0.770	29.83%
	contextualized	0.793	28.23%
wav2vec2-xls-r-1b	convolutional	0.756	30.38%
	contextualized	0.857	26.05%
wav2vec2-large-xlsr-53	convolutional	0.757	27.11%
	contextualized	0.813	22.50%
wav2vec2-large-xlsr-53-hu	convolutional	0.756	27.19%
	contextualized	0.831	21.23%
Fine-tuned			
wav2vec2-BEA-spk	convolutional	0.798	29.09%
	contextualized	**0.898**	**18.93%**
wav2vec2-BEA-asr	convolutional	0.798	29.09%
	contextualized	0.861	19.87%

4.4. Evaluation

We relied on linear Support Vector Machines (SVM) for classification; the C complexity parameter was set in the range 10^{-5}, . . ., 10^{1}. Seeking to avoid an optimistically-biased evaluation of the model, we opted for speaker-wise nested cross-validation. That is, each outer fold contained one speaker for test and the rest for training. During training, we carried out inner fold cross-validation to select the best hyper-parameters. This process was repeated for every single speaker, ensuring that each of them was used exactly one time during test across all folds. We employed evaluation metrics which are commonly used in biomedical studies (e.g. [36, 37]). Besides reporting area under the ROC curve (AUC) measures, we utilized Equal Error Rate (EER) as well. The EER is the point at which the false acceptance rate (FAR) is equal to the false rejection rate (FRR). This practice, in balanced binary-class distributions, leads to very similar accuracy, precision, recall and F-measure scores. Hence, we report only EER (i.e., $100\%-$ Accuracy). As we have only two speaker categories, the AUC value of the two appears to be the same.

5. Results and Discussion

Table 1 shows that although our baseline systems achieved competitive performances, they were surpassed by our pre-trained and fine-tuned Wav2Vec2 approaches. This may suggest the capability of contextual speech encodings over standard speaker embeddings for paralinguistic feature extraction. Overall, the contextualized embeddings got higher scores than their convolutional counterparts in all the experiments. This could be due to the fact that convolutional representations carry low-level information that may not be relevant for MS discrimination. On the other hand, contextual representations achieved higher scores in general. It appears that they were able to capture traits that are more relevant to paralinguistic tasks, as they are typically built from high-level semantic information at the utterance level.

Table 2: *Results of the experiments on feature combination in terms of AUC and EER. The best feature configurations are reported. N denotes the feature dimension, and only contextual embeddings were employed.*

Model	**N**	**AUC**	**EER**
wav2vec2-BEA-spk + asr	2048	**0.922**	19.61%
+ wav2vec2-large-xlsr-53-hu	3072	0.880	21.77%
+ wav2vec2-xls-r-1b	4352	0.845	23.94%

The best configuration corresponded to our model fine-tuned on speakers. That is, with *wav2vec2-BEA-spk*, we achieved an AUC score of 0.898, and equal error rate of 18.93% based on its contextualized representations; while its encoder embeddings achieved lower scores: AUC and EER of 0.798 and 29.04%, respectively. Compared to this configuration, similar scores were shown by our second system fine-tuned on speech units (i.e., *wav2vec2-BEA-asr*), where the convolutional embeddings yielded the same performance. This was an expected behavior since we froze the feature encoder during the fine-tuning process. Differently, the context features got lower scores: AUC and EER of 0.861 and 19.87%, correspondingly.

Although both strategies showed comparable performance scores on the given task, it appears that fine-tuning on speakers led to a better quality of contextual speech representations compared to its ASR-based counterparts (i.e., our *wav2vec2-BEA-asr*, and the *wav2vec2-large-xlsr-53-hu*). The difference between the representations generated by each fine-tuning approach may rely on the fact that the sequence of embeddings in the 'spk' method was pooled into a 'single-utterance' encoding (similar to [38]) within the fine-tuning process, which can be viewed as a summary of the entire input recording, and later were optimized via cross-entropy. Conversely, the 'asr' method takes each output from the sequence of contextual embeddings and labels it based on the vocabulary of the task with a fully connected layer, and optimizes via CTC loss, this may carry less relevant (paralinguistic) information utterance-wise.

As shown in Table 1, *wav2vec2-large-xlsr-53-hu* and *wav2vec2-xls-r-1b* attained comparable performances to our fine-tuned solutions based on their contextual feature representations. The former (wav2vec2-large-xlsr-53-hu) produced higher quality embeddings than the latter, most probably due to its fine-tuning on Hungarian language; this indicated the effectiveness of language matching for fine-tuning. Nevertheless, this was not sufficient to equalize the performance of either of our proposed configurations, especially that of the 'wav2vec2-BEA-asr' counterpart. Although both models were fine-tuned using (distinct) Hungarian language corpora, the difference in performance may be due to the following, i) the acoustic and recording conditions, ii) their size, the number of hours of the BEA corpus exceeded CommonVoice and CSS combined significantly (85 vs. 19 hours). This may suggest the relative importance of the corpus size for downstream tasks. Conversely, the 1 billion parameters of the wav2vec2-xls-r-1b model were found to be competitive as their contextual embeddings outperformed the smaller version of the same model (i.e., 300m).

5.1. Feature Combination

This involves combining multiple feature vectors to create a new feature set for training. In our last set of experiments, attempting to capture complex, non-linear relationships between the different sets of latent speech embeddings, we combined the feature encodings from our best-performing configurations. Although a combination can be executed in different ways, here we relied on a simple yet effective concatenation method. Table 2 shows the results of these experiments. Combining our best models (i.e., wav2vec2-BEA-spk + wav2vec2-BEA-asr) led to even higher performances: AUC of 0.922. This process contributed to make the SVM model more robust by reducing the impact of outliers that may be present in the feature set, while keeping the number of dimensions at an adequate size for the SVM classifier. Consequently, although appending more features from the subsequent best models just led to a worsening of the performances, our new results still surpassed all the previous configurations, except for 'wav2vec2-BEA-spk'. The decrease in scores when more 'external' features were involved may be due to the SVM classifier being over-fitted because of the increased dimensionality while the number of samples remained the same; and, the (lower) quality of the successive appended embeddings may have led to more 'noisy' information being added during training.

6. Conclusions

This paper presented the use of speech-encoded representations for automatic multiple sclerosis discrimination using audio recordings. More precisely, we showed that embeddings derived from distinct methods employed for fine-tuning Wav2Vec2 pre-trained instances may contain relevant paralinguistic information. In the experiments, we employed both automatic speech recognition and speaker verification techniques in our fine-tuning frameworks, the former following standard ASR fine-tuning approaches for Wav2Vec2 (i.e., learning speech units); and the latter was inspired by speaker verification where we rather fine-tune on speakers (e.g., learning information from the whole utterance). Our methods exceeded performance scores obtained from different cross-lingual Wav2Vec2 pre-trained models, which were used for feature extraction as well. Overall, our results confirm that contextual features consistently surpass the quality of convolutional representations. In addition, we noted an improvement in our classification scores by adding robustness to our classifier through feature combination from our top models. However, further combinations proved to be counterproductive, most likely due to the ratio between the number of samples and feature dimension (curse of dimensionality). Lastly, our investigation demonstrated the superiority of fine-tuning Wav2Vec2 on speakers rather than on speech units for extracting paralinguistic representations for MS screening.

7. 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).

8. References

[1] 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.

[2] A. Delgado-Álvarez, J. 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.

[3] 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.

[4] 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.

[5] M. Walshe and N. Miller, "Living with acquired dysarthria: the speaker's perspective," *Disability and rehabilitation*, vol. 33, no. 3, pp. 195–203, 2011.

[6] K. An, M. J. Kim, K. Teplansky, J. R. Green, T. F. Campbell, and Y. e. a. Yunusova, "Automatic early detection of amyotrophic lateral sclerosis from intelligible speech using convolutional neural networks." in *Interspeech*, 2018, pp. 1913–1917.

[7] T. Makkonen, H. Ruottinen, R. Puhto, M. Helminen, and J. Palmio, "Speech deterioration in amyotrophic lateral sclerosis (als) after manifestation of bulbar symptoms," *International journal of language & communication disorders*, vol. 53, no. 2, pp. 385–392, 2018.

[8] G. Gosztolya, L. Tóth, V. Svindt, J. Bóna, and I. Hoffmann, "Using acoustic deep neural network embeddings to detect multiple sclerosis from speech," in *ICASSP 2022*, 2022, pp. 6927–6931.

[9] 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.

[10] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," *arXiv:2010.10504*, 2020.

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

[12] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[13] Y. Qin, W. Liu, Z. Peng, S.-I. Ng, J. Li, H. Hu, and T. Lee, "Exploiting pre-trained asr models for alzheimer's disease recognition through spontaneous speech," *arXiv:2110.01493*, 2021.

[14] X.-Y. Chen, Q.-S. Zhu, J. Zhang, and L.-R. Dai, "Supervised and self-supervised pretraining based covid-19 detection using acoustic breathing/cough/speech signals," *arXiv preprint arXiv:2201.08934*, 2022.

[15] B. Schuller, A. Batliner, S. Amiriparian, C. Bergler, M. Gerczuk, N. Holz, P. Larrouy-Maestri, S. Bayerl, K. Riedhammer, A. Mallol-Ragolta *et al.*, "The acm multimedia 2022 computational paralinguistics challenge: Vocalisations, stuttering, activity, & mosquitoes," in *ACM MM*, 2022, pp. 7120–7124.

[16] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

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

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

[19] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *ICML*. PMLR, 2016, pp. 173–182.

[20] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[21] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," *arXiv preprint arXiv:2012.06185*, 2020.

[22] W.-W. Lin and M.-W. Mak, "Wav2spk: A simple dnn architecture for learning speaker embeddings from waveforms." in *INTERSPEECH*, 2020, pp. 3211–3215.

[23] T. Grósz, D. Porjazovski, Y. Getman, S. Kadiri, and M. Kurimo, "Wav2vec2-based paralinguistic systems to recognise vocalised emotions and stuttering," in *ACM*, 2022, pp. 7026–7029.

[24] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[25] N. Vaessen and D. A. Van Leeuwen, "Fine-tuning wav2vec2 for speaker recognition," in *ICASSP*. IEEE, 2022, pp. 7967–7971.

[26] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker verification," in *ICASSP*, 2020, pp. 7169–7173.

[27] S. Zargarbashi and B. Babaali, "A multi-modal feature embedding approach to diagnose Alzheimer's disease from spoken language," *arXiv preprint arXiv:1910.00330*, 2019.

[28] 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*, Brno, Czech Republic, Sep 2014, pp. 424–431.

[29] Z. Zhang, X. Zhang, M. Guo, W.-Q. Zhang, K. Li, and Y. Huang, "A multilingual framework based on pre-training model for speech emotion recognition," in *APSIPA ASC*. IEEE, 2021, pp. 750–755.

[30] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006, pp. 369–376.

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

[32] 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*, Singapore, May 2022, pp. 8502–8506.

[33] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The Interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Interspeech*, San Francisco, CA, USA, 2016, pp. 2001–2005.

[34] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, Parkinson's & eating condition," in *Interspeech*, 2015, pp. 478–482.

[35] B. W. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S. Roelen, S. Schnieder, E. Bergelson, A. Cristia, and A. S. et. al, "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity," in *Interspeech*, Graz, Austria, Sep 2019, pp. 2378–2382.

[36] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, J. Villalba, J. R. Orozco-Arroyave, and N. Dehak, "Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's Disease," *Applied Soft Computing*, vol. 62, no. 10, pp. 649–666, 2018.

[37] J. Fritsch, S. Wankerl, and E. Nöth, "Automatic diagnosis of Alzheimer's Disease using Neural Network language models," in *ICASSP*, 2019, pp. 5841–5845.

[38] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American*. Association for Computational Linguistics, 2019, pp. 4171–4186.