Received 18 January 2023, accepted 5 February 2023, date of publication 10 February 2023, date of current version 16 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3243986

RESEARCH ARTICLE

Automatic Voice Disorder Detection Using Self-Supervised Representations

DAYANA RIBAS[1], **MIGUEL A. PASTOR**[1], **ANTONIO MIGUEL**[1], **DAVID MARTÍNEZ**[2], **ALFONSO ORTEGA**[1], **AND EDUARDO LLEIDA**[1]

[1]ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, 50018 Zaragoza, Spain
[2]Lumenvox, 81379 Munich, Germany

Corresponding author: Dayana Ribas (dribas@unizar.es)

This work was supported in part by the European Union's Horizon 2020 Research and Innovation Programme under Marie Skłodowska-Curie under Grant 101007666, in part by Grant MCIN/AEI/10.13039/501100011033, in part by the European Union NextGenerationEU/PRTR under Grant PDC2021-120846-C41 and Grant PID2021-126061OB-C44, and in part by the Government of Aragon under Grant T36_20R.

ABSTRACT Many speech features and models, including Deep Neural Networks (DNN), are used for classification tasks between healthy and pathological speech with the Saarbruecken Voice Database (SVD). However, accuracy values of 80.71% for phrases or 82.8% for vowels /aiu/ are the highest reported for audio samples in SVD when the evaluation includes the wide amount of pathologies in the database, instead of a selection of some pathologies. This paper targets this top performance in the state-of-the-art Automatic Voice Disorder Detection (AVDD) systems. In the framework of a DNN-based AVDD system we study the capability of Self-Supervised (SS) representation learning for describing discriminative cues between healthy and pathological speech. The system processes the SS temporal sequence of features with a single feed-forward layer and Class-Token (CT) Transformer for obtaining the classification between healthy and pathological speech. Furthermore, there is evaluated a suitable data extension of the training set with out-of-domain data is also evaluated to deal with the low availability of data for using DNN-based models in voice pathology detection. Experimental results using audio samples corresponding to phrases in the SVD dataset, including all pathologies available, show classification accuracy values until 93.36%. This means that the proposed AVDD system achieved accuracy improvements of 4.1% without the training data extension, and 15.62% after the training data extension compared to the baseline system. Beyond the novelty of using SS representations for AVDD, the fact of obtaining accuracies over 90% in these conditions and using the whole set of pathologies in the SVD is a milestone for voice disorder-related research. Furthermore, the study on the amount of in-domain data in the training set related to the system performance show guidance for the data preparation stage. Lessons learned in this work suggest guidelines for taking advantage of DNN, to boost the performance in developing automatic systems for diagnosis, treatment, and monitoring of voice pathologies.

INDEX TERMS Voice disorder, pathological speech, Saarbruecken voice database, advanced voice function assessment database, self-supervised, class token, transformer, deep neural networks.

I. INTRODUCTION

With recent COVID-19 pandemic it has emerged a way of life where the remote access to health services has gained relevance. The availability of automatic systems for diagnosis, treatment and monitoring of voice pathologies has gained importance to help doctors provide timely assistance to patients and, at the same time, screen those who really need hospital visits. Currently, voice pathologies have an impressive prevalence among population. Previous studies [1] reported that almost 30% of general population have experienced a period of time with a problem of voice. While

The associate editor coordinating the review of this manuscript and approving it for publication was Junhua Li.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

in [2] authors reported that one in every 13 adults has voice problems annually, which means a prevalence of 7.5%. For professionals that use the voice as primary tool is even worst. Among teachers, the prevalence amounts to 58% [3], and there are many more professionals in similar circumstances, e.g. singers, actors, telemarketers, etc. There are many reasons why people with these problems do not go to the doctor in time for assessment or treatment. If voice affectation is not very annoying, it is usual to leave the problem aside while getting worse. Therefore, there is an opportunity to use smart solutions to assess voice pathologies as part of remote health services contributing to early diagnosis.

The task of automatic voice disorders detection opens the gate to health assistance systems, from the detection of a voice disorder to the specification of the disease and its severity. Then, tracing the pathology evolution and the treatment are also important tasks. Many research efforts have focused on this aim from the approach of binary classification between healthy and pathological voices [4], [5]. Saarbruecken Voice Database (SVD) [6] has been widely used in previous works considering that this one of the few options among freely available datasets with healthy/pathological voice recordings. Reported results in previous works are showing accuracy values of around 80% when the complete set of pathologies available in SVD is used. For instance, 80.71% for phrases in [7] and 82.8% for vowels /aiu/ in [8]. All reported accuracies higher than this figure, have been obtained with some selection of voice data using only certain pathologies [9], [10]. This closed-set scenario is not realistic for health applications. So, instead of the pathology selection, we think that the evaluation of systems with a high variability of pathologies reflects the real-life scenario of application, where the AVDD system would probably be exposed to many different voices. In this study we work with the wide amount of pathologies included in SVD's audio data for detecting which one correspond to healthy speech and which one to pathological speech.

The problem of data availability in voice disorder detection emerges from the difficulty of obtaining healthy and pathological voices manually labeled by experts. Usually, these data result are the outcome of research projects with medical institutions where frequently the condition is to keep data private. Therefore, the availability of datasets for developing AVDD systems is quite limited. SVD [6] is one of the scarce freely available databases for this task. More recently, AVFAD and VOICED appeared in similar conditions. In this work, we have used SVD motivated by its availability and the fact that it is almost a standard for assessing AVDD systems due to the number of related papers employing it. SVD is well endowed with recordings and speakers. However, it also presents some issues, such as fewer samples of healthy speech than pathological speech. There is great inequality in the distribution of individual pathologies - some pathologies have only one audio, and they can end up only in the testing set - which is a problem for training AVDD systems. In this work we propose a way to deal with these issues by using AVFAD database to implement a suitable data extension for expanding the training set of SVD.

A distinctive characteristic of Deep Learning approaches are the requirement of a large amount of data for training models. This issue has limited its use in AVDD, where voice data availability is scarce. This motivated us to explore the capability of SS representations, such as Wav2vec2.0 [11], HuBERT [12] and WavLM [13], for extracting features with information on healthy/pathological cues in the limited amount of data of SVD. For classification, we assess a basic feed-forward layer and a CT Transformer design based on a Multihead Self-Attention (MSA) mechanism [14], [15]. The classifiers refine the information in the temporal sequence of SS embeddings and provides a single vector for each utterance which capture the essence of pathological and healthy classes. We developed the AVDD framework inside the SUPERB toolkit [16] and contribute with a new downstream.[1]

Our contributions are:

1) Significant performance improvement in AVDD evaluated in the Saarbruecken Voice Database including all the set of pathologies without restriction. This system achieves an accuracy improvement of 4.1% over the baseline system and overcomes the unrealistic assumption of previous related work that used to select a set of pathologies for performing the AVDD system evaluation.
2) The first approach for evaluating SS representations and CT Transformer in voice disorder detection.
3) Evaluation of the strategy of training extension as a suitable solution for the low-resource characteristic of voice disorder-related tasks, that allows for effectively employing DNN-based solutions. This strategy achieves an accuracy improvement of 15.62% over the baseline system.
4) Open-source code for reproducing results and further investigations.

In the following, section II presents a review of related previous works that allows to establish the performance of state-of-the-art. Section III explains the experimental setup including databases and performance metrics. Then, section IV presents the AVDD system proposed and describes an strategy for making a suitable use of the data available for improving the performance of the small dataset used for evaluation. Section V presents the experimental configuration to evaluate the proposed system and discusses on obtained results. Finally section VI concludes the paper.

II. PREVIOUS WORKS

There are several approaches for developing AVDD systems using machine learning methods. Many of these works focused on studying suitable representations for pathological voice, such as spectral and cepstral features, voice quality and perturbation measures, and complexity measures [17].

[1] https://github.com/dayanavivolab/s3prl/tree/voicedisorder

Table 6 in [17] shows accuracy figures for 12 feature sets, including MFCC, PLP, glottal source features, etc, that reach 76.19% and $EER = 26.2\%$ for SVD's phrases. Furthermore, Opensmile[2] and Multidimensional voice program parameters (MDVP) [18] (the latter is no freely available) are toolkits to extract feature sets that include measurements in the previously mentioned categories. In [7] authors report results with Opensmile for SVD's phrases with top accuracy of 80.71% (Table 2 in [7]) similar to accuracy of 82.8% reported in [8] for /aiu/ concatenated vowels in SVD. Those feature sets have been complemented with a variety of statistical classifiers such as GMM [19], SVM [9], among others that are further mentioned in [4]. More recently, some works have presented systems based on DNN including architectures such as CNN [20] reporting an accuracy of 71% for SVD's /a/ neutral vowel, LSTM [21] reporting accuracy of 68, 98% for SVD's /aiu/ sustained vowels, and ResNet [22] with accuracy of 69.37%[3] for /a/ neutral vowel in SVD, and also BLSTM [23]. There are other recent works employing DNN [24], [25], [26] with similar configuration and performances. Summarizing, table 1 shows reported accuracy results in the last three years as comparison starting point of the state-of-the-art performance. These previous works reported best accuracy values of around 80%, including the most recent approaches based on DNN.

TABLE 1. Summary of previous reported results of voice disorder detection using SVD dataset.

Work	Year	Result	Audio type
[7]	2021	UAR = 80.71%	Phrases
[17]	2020	UAR = 76.19%	Phrases
[8]	2020	UAR = 82.80%	Vowels aiu
[22]	2020	UAR = 69.57%	Vowel a

The review in [4] shows a large table of research works describing the database, methods used for AVDD systems, and the system performance reported by authors. It is interesting to see that accuracy values of 100% are also reported. However, note that each of these works uses a selection of data with a few pathologies, usually including those that have a large number of audio samples or are easily distinguishable from healthy speech. These experiments are designed in a close-set classification scenario, where the evaluation set contains only audio from a known set of pathologies. This is far from a real-life application scenario, so these reported results are hardly useful beyond the literature. Unfortunately, this kind of experimental design are very frequent among the voice disorder previous work [9], [24], [26], so this fact makes hard the task of comparing several approaches. We agree with other researchers that in this field the wide variety of AVDD reported performances and the reproducibility of results is quite an issue [22]. Recently, Huckvale and Buciuleac [7] approached over this problem using SVD. They found that re-implementations of previous works using DNN-based systems under-performed the reported results.

On the other side, in the framework of the recent flow of DNN-based solutions SS representations have raised as a general purpose frontend. The high capability of SS representation learning for finding underlying relations on data and providing substantial features has been assessed in several speech-related areas.[4] Related to the voice disorder topic there are a few recent works that use SS representations for automatic speech recognition with pathological speech [27], [28]. However, from the best of our knowledge there are not previous studies for SS representations in AVDD. From the classifier point of view, there are recent deep learning solutions based on Transformers [29] that are achieving very promising performances for classification tasks related to speech processing, for instance in biometric applications [30], [31], although not yet for pathological speech.

In voice disorder related task the use of DNN-based solutions is limited because the low availability of resources is a problem for using solutions based on deep learning. The augmentation of training data using different sets of data available has been previously explored with positive results in many speech related tasks. However, apparently in voice disorder related tasks, data augmentation strategies have not being widely used. Anyway there are a few related works that could be mentioned. In [32] authors evaluated their AVDD system with three databases exchanging train and test sets among them. In [23] authors evaluated an AVDD system with audio recorded with smartphone, so they employed domain adversarial training to increase the performance robustness against channel mismatch. In [33] and [34] authors explore the generation of synthetic samples for augmenting the minority class for handling imbalanced datasets.

Motivated from these issues, in this study we evaluate a training extension strategy as a suitable solution for the low-resource characteristic of voice disorder-related tasks that allows to effectively use DNN-based approaches. We design an AVDD system using SS representations and DNN-based classifiers as a framework for studying the performance of SS representations for healthy and pathological speech information. Then, we use this framework for assessing the performance of the training extension strategy. Worth highlighting that the experimental design consider the wide amount of pathologies and levels of severity intrinsic to SVD corpus for evaluation, because this data structure resemble a real life scenario of application where the AVDD system would face any kind of patient and pathology. Note that reported results using different features and classifiers for the AVDD system, including the most recent approaches based on DNN, have reached maximum accuracy values of around 80%. Then, beyond the issue of comparability with previous works, we can consider this performance figure as the state-of-the-art. Anyway, we choose the system presented

[2] https://github.com/audeering/opensmile

[3] This figure was taken from the reproduction of the experiment in [7].

[4] superbbenchmark.org

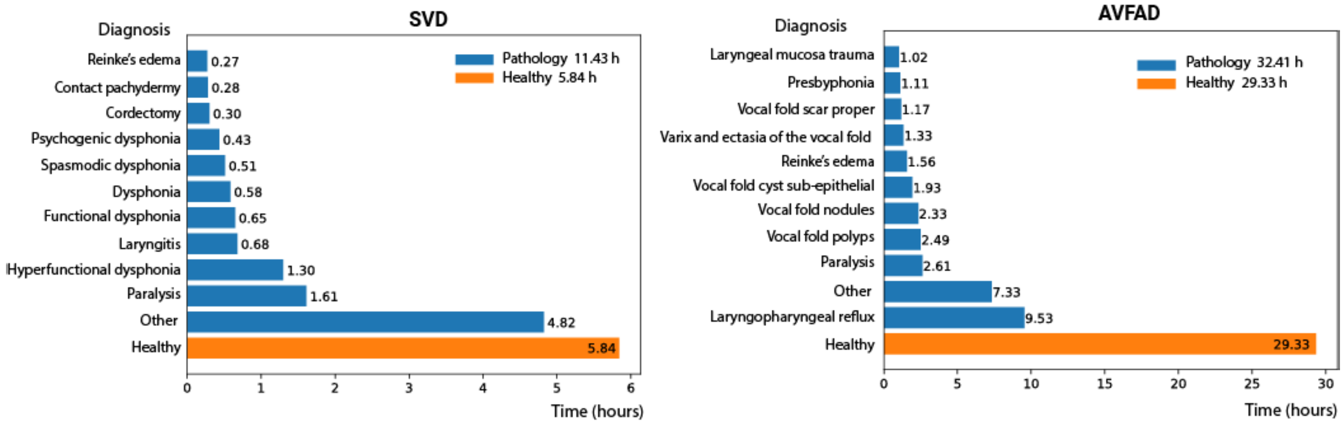


FIGURE 1. Distribution of more frequent pathologies in databases SVD and AVFAD.

in the previous paper [7], which has a performance among the best reported in literature, and recompute it with our data list for establishing the baseline.

III. EXPERIMENTAL SETUP

A. DATABASES

1) SAARBRUECKEN VOICE DATABASE (SVD)

This database in the German language [35] contains voice recordings of 687 healthy persons: 428 females and 259 males, and 1356 patients: 727 females and 629 males with one or more pathologies. The database includes 71 pathologies, although there is a great inequality in the distribution of individual pathologies even some pathologies have only one audio. Fig. 1 shows the pathologies with more audio representation.

For experiments we used the full set of pathologies without any selection, because this is the most related scenario to a real case in a hospital triage. We use the phrase included in each recording session: Sentence "Guten Morgen, wie geht es Ihnen?" ("Good morning, how are you?"). The original sample rate of the audio is 50 khz, but for these experiments we downsampled the audio to 16 Khz. The database was divided in 5-fold and a two-to-one ratio is maintained between pathological and healthy samples. Therefore, the minimum reasonable accuracy is around 70%, less than this, the system is worse than a random classification. The audio of speakers included in training is not in the test partition.

2) ADVANCED VOICE FUNCTION ASSESSMENT DATABASE (AVFAD)

This is an open-access dataset in the Portuguese language [36] with 363 subjects with no vocal alterations: 250 females and 113 males, and 346 clinically diagnosed subjects with vocal pathology: 249 females and 97 males. The patients are diagnosed with 26 different vocal pathologies, however most of them are not included in SVD. The most represented disorders in AVFAD are presented in Fig. 1. The original sample rate of the audio data is 48 khz, but for these experiments we downsampled the audio to 16 Khz.

B. PERFORMANCE METRICS

To evaluate the classification performance we use Accuracy (ACC) (eq.1) and Unweighted Average Recall (UAR) (eq.2). Among these indexes, we use values in the confusion matrix for computing the score, i.e., true and false positive and negative number (TP, FP, TN, FN). For balanced class distributions, ACC and UAR are quite similar. However, if this is not the case, UAR considers each class by itself, while ACC provides a more general metric.

$$ACC = \frac{TN + TP}{TN + TP + FN + FP} \tag{1}$$

$$UAR = 0.5 \cdot \frac{TP}{TP + FN} + 0.5 \cdot \frac{TN}{FP + TN} \tag{2}$$

Furthermore, as this is a detection task, we also used performance metrics such as the Area Under ROC Curve (AUC) [37] and Equal Error Rate (EER) [38]. AUC is a performance measurement for the classification problems at all threshold settings. It tells how much the model is capable of distinguishing between classes. So, the closest to 1 the AUC, the better the model is at distinguishing between healthy and pathological speech. On the other side, EER is the operation point where the false acceptance rate and false rejection rate are equal. It is a widely used statistic to show biometric performance, typically when operating in the verification task. In general, the lower the EER value, the higher the accuracy of the system.

For all experiments, we computed the classification scores in the 5-fold scheme. Then, we pull all the scores of the partitions together, and compute a single ACC, UAR, AUC and EER for the whole experiment. Note that all performance metrics are expressed in percent (%).

IV. METHODOLOGIES FOR AVDD SYSTEM

In this paper, we present a DNN-based AVDD system for the binary classification task -healthy vs. pathological speech- where the aim is to improve the performance by using the benefits of SS representation learning in an under-resourced task. In the last few years, a flood of DNN solutions has overspread many areas of speech research. However, the small amount

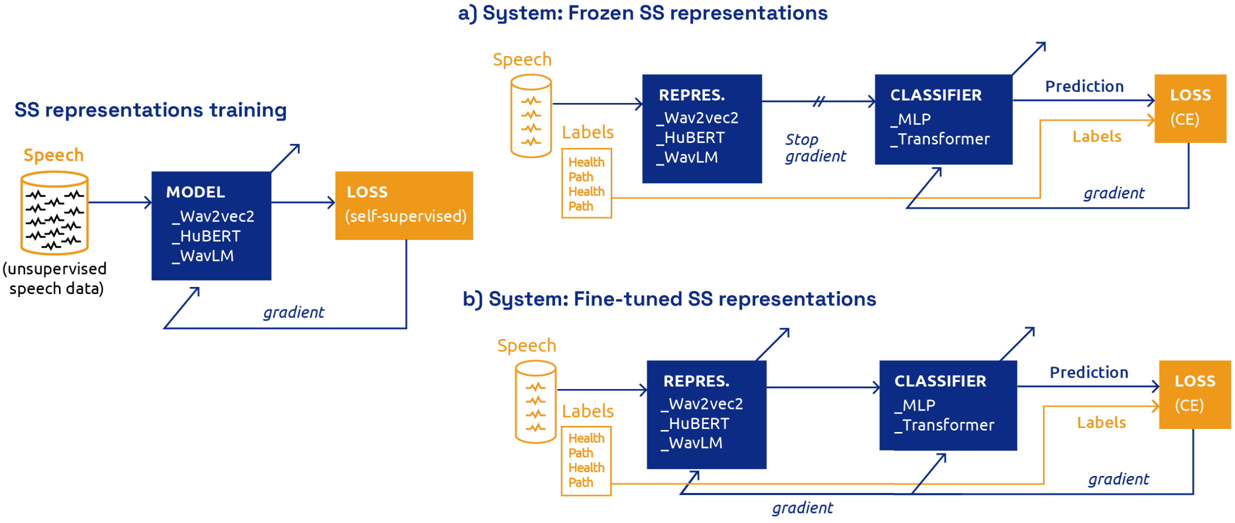


FIGURE 2. Flowchart of the AVDD system from the training of the SS representation models which are employed in two modalities, frozen and fine-tuned, for the binary classification of healthy and pathological speech samples.

of healthy and pathological voice data available in related datasets has limited the wide use of DNN in this field. Since the performance improvement is limited in reported results after studying handcrafted features and statistical classifiers suitable for the task, to find ways of assessing the data-driven paradigm is an interesting subject of study.

Fig. 2 depicts the AVDD system. It starts from the training of the SS models employed as speech representations. These models were previously trained with a huge amount of speech data without any labeling related to the medical condition. The other side of the figure illustrates the AVDD system for the task of healthy/pathological classification, so the speech data used in this case include labeling information indicating the class. There are two modalities for using the SS representations: frozen and fine-tuned. Fig. 2a) describes the frozen mode, where gradients are inhibited such that the back-propagation algorithm cannot modify the SS representation model with respect to the in-domain data. Fig. 2b) describes the fine-tuned mode, where the gradients are active such that the SS representation model can adapt to the in-domain data. The speech representation block is followed by a classification stage that trains an MLP or Transformer model using the Cross-Entropy (CE) loss function. The following subsections further explain the characteristics of these models, data, and loss function employed for training them.

We have developed the system in the framework of the SUPERB toolkit [16]. Then, we contribute to the toolkit development by releasing the code as a new downstream for AVDD system. Also, to ensure reproducibility, we share the audio lists employed in the experiments.

A. SELF-SUPERVISED REPRESENTATIONS

In this work we include three of the most remarkable methods in our study, namely: *Wav2vec2.0* [11], *HuBERT* [12], and *WavLM* [13]. All these, were previously trained with a significant amount of speech without any healthy/pathological awareness. Now we use them to create feature vectors for train/test partitions of SVD as part of the first processing stage of the system. They describe the sequential evolution of the utterance, so there is one feature vector by frame.

1) Wav2vec2.0

In this paper we used the Wav2vec2.0 base model, whose embeddings are 768-dimensional. The model architecture consists of a local encoder with several convolutional blocks. It encondes the raw audio into a sequence of embeddings with stride of 20 ms and receptive field of 25 ms.

The model is pre-trained in a self-supervised setting inspired by BERT [15]. In this setting, the model is trained to minimized a contrastive loss between the outputs of the contextualized encoders c_t and the quantized local encoder representations q_t of randomly masked contiguous time-steps t

$$L_m = -log \frac{exp(sim(c_t, q_t)/k)}{\sum_{\tilde{q} \in Q_t}(exp(sim(c_t, \tilde{q})/k)} \tag{3}$$

where $sim(c_t, q_t)$ is the cosine similarity between the contextualized encoder outputs c_t and the quantized convolutional encoder representations q_t, k is the temperature set to 0.1, Q_t represents the union of candidate representations $\tilde{q}$ including q_t, and $K = 100$ is the number of distractors, which are the outputs of the local encoder sampled from masked frames of the same utterance of q_t. Finally L_m is obtained by summing over all masked frames, a L2 regularization is added to L_m, and also a diversity loss to power the use of the quantized codebook representations. The quantization module is based on a Gumbel-Softmax layer [39]. It generates the targets of the model by the quantization of the local encoder representations, and then these representations input the contextualized encoder. The architecture employs 12 transformer encoder blocks [29] with 8 attention heads each. The pre-training process is optimized with ADAM [40] and the learning rate decays linearly after a warming up.

2) HuBERT

The HuBERT base model is the one employed in this study. In this model [12], the contiguous time steps from the local encoder representations are randomly masked similar to Wav2vec2.0. A k-means clustering mechanism is applied to 39-dimensional Mel Frequency Cepstral Features (MFCC) features to generate labels for the first pre-training iteration. In the following iterations, the k-means clustering uses the latent features from the previous iterations to generate better targets. In order to predict cluster labels, a projection layer is added over transformer blocks. The system is trained to minimize the cross-entropy loss (CE) defined as

$$L_m(f; X, \{Z^{(k)}\}_k, M) = \sum_{t\in M}\sum_{k} log(p_f^{(k)}(z_t^{(k)}|\tilde{X}, tz)) \quad (4)$$

where $Z^{(k)}$ is the target sequences generated by the k-means model, $M \subset [T]$ is the set of indices to be masked for a length$-T$ sequence X, and $\tilde{X} = r(X; M)$ is a noisy version of X where x_t is replaced by a mask embedding $\tilde{x}$ when $t \in M$. $\tilde{X}$ is the input of a masked prediction model f for predicting a distribution over the target indices at each timestep $p_f(|\tilde{X}; t)$. The pretraining is also based on the ADAM algorithm [40].

3) WavLM

In the same line of previously described representations Wav2vec2.0 and HuBERT, WavLM [13] is a more recent SS system built with transformer blocks and trained with several amount of speech data. It learns the speech representation by masking part of the speech signal to predict the hidden part. The model is based on a convolutional representation encoding over 25 ms of audio with stride 20 ms and a transformer encoder [29]. In the training stage, the masked acoustic features from the convolutional encoder get into the transformer and it outputs hidden states. The training objective forces the network to predict a discrete target sequence. Similar to HuBERT, to obtain these targets the system uses the k-means algorithm for clustering on the training data. This is an iterative processing, where MFCC features are used for the first step, and in the following iterations, the latent representations learned are used.

To optimize the network WavLM uses the mask prediction loss following HuBERT. The objective function is defined as

$$L_m = \sum_{l\in K}\sum_{t\in M} log(p(z_t|h_t^L)) \quad (5)$$

where M corresponds to the set of masked indices in time domain and h_t^L is the $L-$layer transformer output for step t.

In this paper we used the WavLM base model, which have 12 transformer encoder layers, 768-dimensional hidden states, and 8 attention heads. The model is pre-trained with 960 speech hours from Librispeech using the label generated by clustering the $6-th$ transformer layer output of the first iteration of HuBERT base model.

B. DNN-BASED CLASSIFIERS

For classification we included two models: a basic DNN-based architecture of Multi Layer Perceptron (MLP) and a more recent architecture based on Transformer.

1) MLP

This DNN-based architecture consists of an average pooling for obtaining a single representation vector from the sequence of SS vectors by frame. This is followed by a single feed-forward layer for classification between healthy and pathological. This is a basic model for classification that allows to evaluate the performance of the SS representation without much more processing.

2) CLASS-TOKEN TRANSFORMER

Furthermore we use a classifier based on Transformers that receive the full sequence of SS vectors and provides the classification between healthy and pathological speech. This is a more complex model than the previously described MLP, that allows to evaluate the performance of the AVDD system beyond the SS representation.

Recently, the Vision Transformer (ViT) [14] has employed the concept of Class-Token (CT) [15] to concentrate the class information in a single vector through several layers of self attention mechanisms in a classification task. This inspired us to use this learnable vector approach for refining the set of embeddings from SS representations in order to perform classification. The CT captures the relevant information for the final classification task from the whole sequence of embeddings through a configurable number of heads and layers in the MSA block (Fig. 3). Each attention head (HEAD j in Fig. 3) learns the weights to sum these embeddings for each layer and obtains a vector consisting on the concatenation of the attention heads (Hj in Fig. 3). This way the CT learns a global description of the utterance, where the multiple attention heads contribute to this final representation vector of the utterance. The multiple heads implied in the process are able to better capturing the underlying information in the sequence than the usual pooling alternative.

C. TRAINING EXTENSION STRATEGY

Considering the low availability of resources in this task, our purpose here is to explore a way of expanding the training set of SVD, that only includes one hour of speech data by fold, with the use of 32 hours of healthy and pathological speech from AVFAD dataset. Looking at the description of the databases in previous section III-A there is a clear difference between AVFAD and SVD in terms of the pathologies included, the language, the recording conditions, the speakers, and so on. However, they shared the common characteristic about the fact that they have labelled speech data from healthy and pathological speakers. Note that this

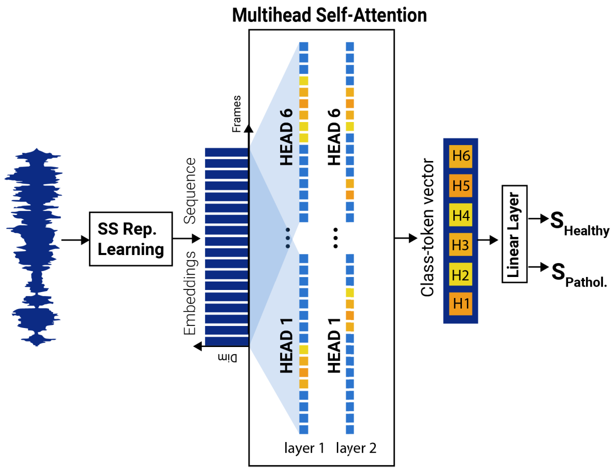


FIGURE 3. AVDD system with SS representation and CT Transformer for classification. MSA chart highlights only the attention relates to Class-Token.

could perfectly be the case in a real application, where we use to have access to some prerecorded dataset but with totally different conditions than the audio under evaluation.

Fig. 4 shows a preview of the behaviour of speech vectors obtained from the system using WavLM and CT-transformer projected with UMAP [41] to 2D. These corresponds to healthy and pathological utterances in the fold 2 using the model trained only with SVD training data (left) and with SVD+AVFAD (right). A priori, we can see that the training extension seems to increase the separability between classes. In the next section we present experiments to evaluate the training extension strategy.

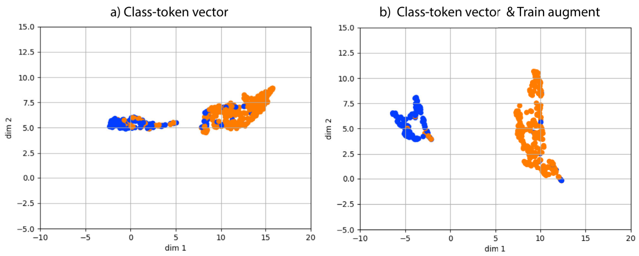


FIGURE 4. Representation vectors of healthy (blue) and pathological (orange) speech compressed with UMAP to 2D.

V. EXPERIMENTS AND RESULTS

In this section we assess the performance of the AVDD proposal by means of binary classification: healthy vs. pathological speech using all audios with phrases in SVD. Experiments are computed in a 5-fold cross-validation scheme.[5] The audio of speakers included in the training set are not used in test. Datasets included in the training set changes as follow:

1) *Train base*: Train SVD (Speech duration = 1 hour in each fold).
2) *Train extension*: Train SVD + AVFAD (Speech duration = 1 + 32 hour in each fold).

Segments of four seconds long of phrases, read, and spontaneous speech are employed for extending the training sets of SVD. For all partitions, the same audio set of AVFAD was appended to each SVD training fold.

A. SYSTEM CONFIGURATION

Experiments were carried out using the three sets of embeddings obtained with SS representations: *Wav2vec2.0*, *HuBERT* and *WavLM* available through SUPERB. For all representations we used the base model in two modalities: frozen and fine-tuned. Frozen means that the model was directly downloaded already pretrained and used for computing features. While in the fine-tuned modality the parameters of the SS models were adapted to the dataset characteristics. In this process, the hyper-parameters of *Wav2vec2.0*, *HuBERT*, and *WavLM* models were modified when processing the training set with healthy and pathological speech. In the first experiment, we used the training set of SVD, with five different folds, and in the second approach, we extended the audio list of the former five folds with phrases from AVFAD.

We use two classification approaches, MLP and CT-Transformer. MLP starts with a pooling layer for obtaining a single embedding from the sequence of feature vectors by frame followed by a feed-forward layer with two outputs for classification. The CT-Transformer has two-layer depth, each with an MSA module with six heads. For training the system, we reserved at each epoch 10% of the training data randomly as validation set. We run 30000 epochs. At the end, we checked the best result for the validation set and took this model for evaluating the test set and presenting it as the final result.

B. BASELINE SYSTEM

To establish the baseline performance, we use the AVDD system in [7]. This system uses the feature set ComParE[6] designed for automatic recognition of paralinguistic issues. The ComParE acoustic feature set has 6373 parameters including spectral, cepstral, prosodic, and voice quality parameters of the speech signal obtained by applying a large set of statistical functionals to acoustic low-level descriptors [42]. For classification it employs a SVM model (Kernel Poly, 1-dim, C=1) implemented with sklearn.[7] The SVM configuration of the kernel function was selected after an auxiliary process of parameter sweeping where several combinations of parameters of the Polynomial and Radial Basis Function (RBF) kernels were evaluated for this database. Note that, despite this system is not based on DNN, we choose it for establishing the baseline because it presents the best results among all previously reported, including DNN-based systems.

[5] We use the same 5-fold audio partitions as in [7]: "all pathologies" thanks to the collaboration of authors.

[6] http://www.compare.openaudio.eu/

[7] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

TABLE 2. Average of metrics for the baseline AVDD system.

Baseline system: Opensmile+SVM				
Audio type	ACC	UAR	AUC	EER
phrase	83.30	80.11	89.86	18.17

C. TRAIN BASE EXPERIMENT

After presenting the performance of the baseline system, we evaluate the proposal by training the model with the SVD. From now on this experiment is refereed as *Train base*. Results presented in Table 3 show the performance of the system based on SS representations using both, the simpler classifier based on MLP, and the CT-transformer. In the first part of the table, SS models are frozen, while in the second part, they are fine-tuned using the training set.

TABLE 3. Performance metrics for the *Train base* experiment using MLP and CT-Transformer for classification (audio type: phrases).

Train base: SVD				
Classifier	**MLP**		**CT-Transformer**	
Features	ACC	UAR	ACC	UAR
frozen				
Wav2vec2.0	81.04	78.15	83.20	79.99
HuBERT	80.33	77.05	**83.45**	**81.27**
WavLM	**81.39**	**78.41**	82.39	79.65
fine-tuned				
Wav2vec2.0.0	85.76	84.01	**85.87**	84.09
HuBERT	**86.12**	**85.04**	85.61	84.07
WavLM	85.71	84.06	84.66	**84.21**

With frozen models the system was not able to outperform the accuracy of previously reported results in section II, namely the baseline system of Table 2. Then, by fine-tuning, the system improves the performance around 3 − 4% of accuracy. This is an encouraging result considering that the training data of SVD is very small (only 1 hour of speech). Note that despite the small amount of training data SS representation models are able to adapt to this specific domain. Fig. 5 shows the corresponding ROC curves for the experiments in Table 3 using WavLM and CT-Transformers. The behaviour of the curves confirmed the improvement of the fine-tuning over the frozen training for all operation points, as well as the corresponding AUC and EER results.

About SS representation performance, the results for frozen and fine-tuned systems show that the accuracy among SS representations is very similar. Furthermore, Fig. 6 shows ROC curves for the systems using SS representations evaluated - Wav2vec2.0, HuBERT, and WavLM - where we can see that the performance for all operation points is also comparable, as well as the AUC and EER values for the three systems evaluated.

About classifiers, accuracy results in Table 3 show a moderate improvement with CT-Transformer but only for the frozen models. When fine-tuning the SS representations, results are very similar between the system using MLP and CT-Transformer. Comparing the MLP and CT-Transformer curves of the fine-tuned models in Fig. 5 we can see that the performance for all operation points is also comparable, indicating that so far, SS representations are carrying the responsibility of the whole system performance.

D. TRAIN EXTENSION EXPERIMENT

In the previous section the AVDD system was able to obtain an improvement of 3 − 4% over the baseline. Anyway, we believe that this is still moderate considering the power of the DNN models employed. These previous results indicate that the capability of deep learning methods is not benefiting the system's performance. We think that the small amount of training data in SVD folds (1 hour of speech) is hindering the performance of the DNN classifier.

In this section we evaluate the performance of the AVDD system introducing an extension in the training set. In order to introduce out-of-domain information for providing more acoustic variability to the model, we propose the use of data from other corpus different from SVD. Then, we use audio from AVFAD corpora, which also has healthy and pathological audio phrases but recorded in Portuguese instead of in German, under different recordings conditions, and containing different pathologies than SVD. This experiment is called *Train extension* and obtained results are presented in Table 4.

TABLE 4. Performance metrics in the *Train extension* experiment (audio type: phrases).

Train extension: SVD+AVFAD				
Features	ACC	UAR	AUC	EER
Baseline	80.99	77.74	85.93	20.53

fine-tuned

Classifier	**MLP**		**CT-Transformer**	
Features	ACC	UAR	ACC	UAR
Wav2vec2.0	93.01	90.72	92.40	90.02
HuBERT	**94.22**	**93.07**	**94.27**	**93.36**
WavLM	93.96	92.76	93.96	93.22

Results in Table 4 show an impressive accuracy increase for all SS representations. Comparing to the previous experiment in Table 3 there is almost a 10% of absolute improvement in ACC and UAR, indicating that the use of AVFAD in the training set improves the performance. See that HuBERT and WavLM representations reached great improvements compared to *Train base*. For the baseline system, we can see that the training expansion does not look in favor of the system performance. Compared to the results in Table 2, the training extension makes the performance of the SVM-based system decreases.

Fig. 7 shows the ROC curves for the SS representations evaluated along with the corresponding curve for *Train base* experiment. First, the curves show a similar behaviour among all SS representation performances. Then, the comparison

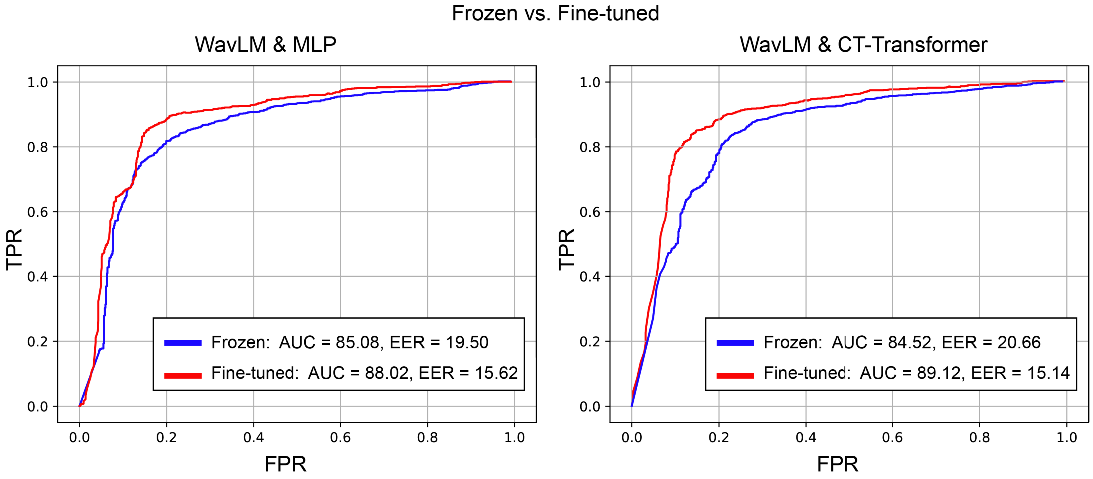


FIGURE 5. ROC curve (True Positive Ratio (TPR) vs. False Positive Ratio (FPR)) for the *Train base* experiment using WavLM and CT-Transformer in both training modes: frozen and fine-tuned (audio type: phrases).

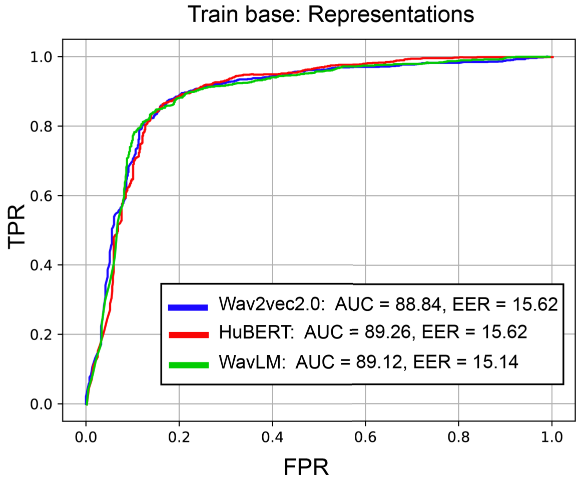


FIGURE 6. ROC curve for the *Train base* experiment using fine-tuned SS representations (Wav2vec2.0, HuBERT and WavLM) and CT-Transformer for classification (audio type: phrases).

between the systems of the *Train base* and the *Train extension* show a clear performance improvement for the last approach, which is consistent for all operation points in the ROC curve. So, we can confirm that the extension of the training set benefits the system performance with a significant increase in performance compared to the system trained only with SVD.

About classifiers, we see that comparing MLP and Transformer for classification, the accuracy results in Table 4 are very similar. Fig. 8 shows that the similarity in performance is consistent for all operation points throughout the ROC curves, as well as for the AUC and EER results. This behaviour is consistent with the previous section result's.

When comparing *Train base* with *Train extension* experiments, the obtained results indicate that both classifiers needed more data to be able to perform. It is interesting to see that in the training set there is only a small amount of in-domain data. SVD data included in the training set amount only 1 hour, while AVFAD data are 32 hours of healthy and pathological speech. This fact indicates that even though the DNN-based system needs more audio, the possibility of using only a small set as in-domain data allows to keep it as a practical solution for low resource scenarios. We also tried to train the system using only AVFAD, but the accuracy went down to 30%, so this confirms the need for in-domain data even in a small proportion of the training set.

In conclusion, this experiment demonstrates the value of SS representation with a suitable use of the data on top of the model for classification. The extended training set SVD+AVFAD used for fine-tuning SS representations was able to accurately perform at separating healthy and pathological speech, without using further complex classifiers.

Note that this conclusion agrees with plenty of previous work on the usefulness of data augmentation for DNN approaches [43], [44]. However, this paper contributes with a first attempt to perform a suitable augmentation strategy for voice disorder classification task. This is a novel contribution in the field, that opens the gate for taking better advantage of deep learning solutions and also for further studies.

E. IN-DOMAIN DATA PROGRESSION

In the previous section V-D we saw the huge increase in the performance obtained with the training set expansion and using such a low amount of in-domain data. Namely 1 hour in-domain in a training set of 33 hours. In this section we study the evolution of performance with the progressive increase of in-domain data to the training set. We carry out a sequence of experiments starting from training with AVFAD alone and then progressively adding 5 minutes of speech data of SVD (in-domain). Experiments stop after having added the full training set of SVD (approximately 60 minutes). The AVDD system selected for these study is among one is

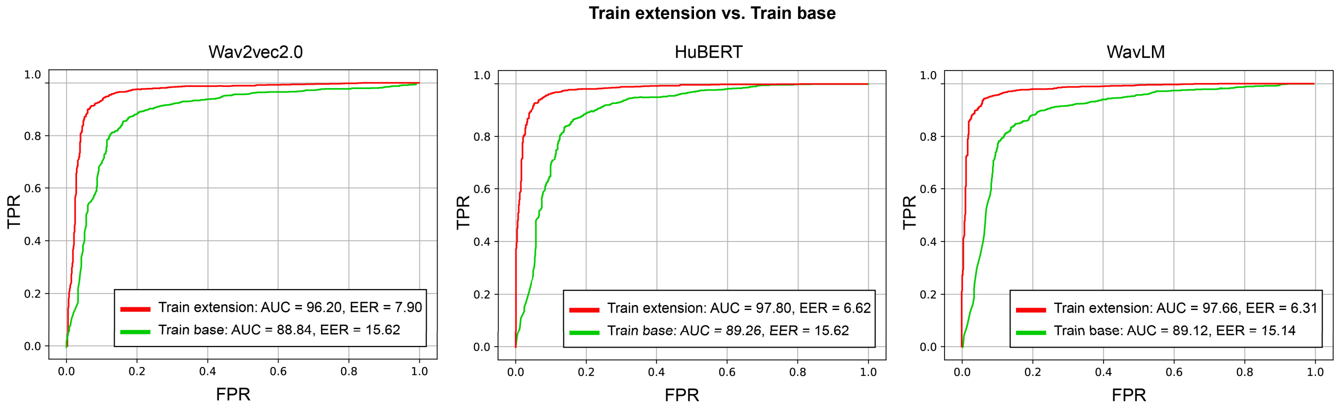


FIGURE 7. ROC curve throughout scores for all partitions in *Train extension* experiment vs. *Train base* experiment using SS representations (Wav2vec2.0, HuBERT and WavLM) and CT-Transformer for classification (audio type: phrases).

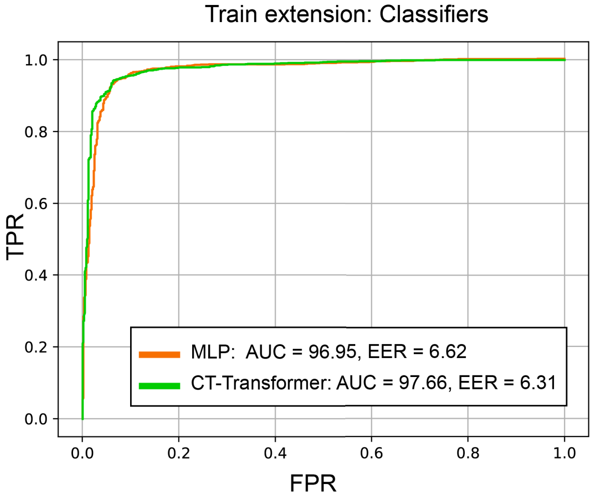


FIGURE 8. ROC curve throughout scores for all partitions in *Train extension* experiment using MLP and CT-Transformer for classification and WavLM for representation (audio type: phrases).

TABLE 5. ACC, UAR, AUC and EER for borderline experiments training with different sets and evaluating with SVD.

Borderline experiments				
Training set	ACC	UAR	AUC	EER
AVFAD	36.60	53.30	63.70	40.1
SVD	84.66	84.21	89.12	15.14
SVD + AVFAD	**93.96**	**93.22**	**97.66**	**6.31**

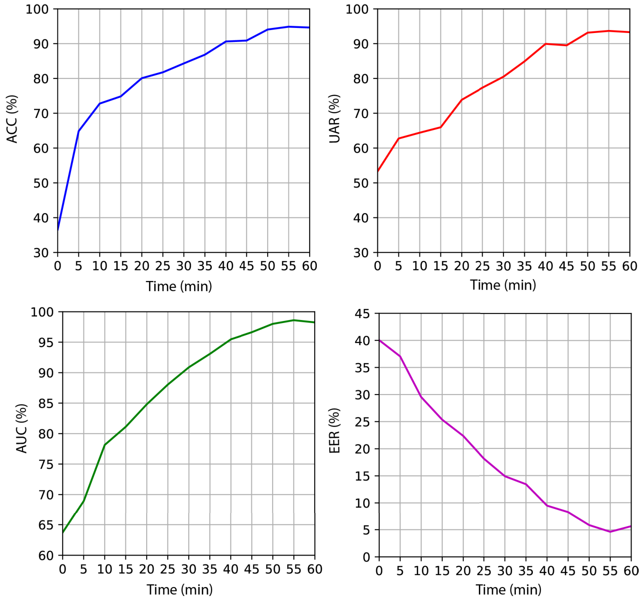


FIGURE 9. Performance metrics for the AVDD system with WavLM and CT-Transformer where x-axis is the amount of minutes of SVD's in-domain audio data added to the training set starting with only AVFAD's out-of-domain audio data.

the previously studied: the representation is WavLM and the classifier is CT-Transformer. Table 5 shows these results as borderlines of the progression in Fig. 9.

In the first experiment, data in training and evaluation sets are totally miss-matched according to language, recording conditions, speakers, text uttered, and pathologies. First row of Table 5 shows the results of this set-up. In this case the performance of the system decreases to a value of $ACC = 36.3\%$ and $EER = 40.1\%$. The evolution of the performance is shown in Fig. 9 showing the influence on the results of the inclusion of in-domain data in the training set. Note that with just five minutes of in-domain data there is a significant improvement in the system performance, going from the system being useless to having results that are better than a random classification. Then, gradually with the addition of in-domain data all metrics improve.

After adding 30-35 minutes, the system achieves a performance similar to the system trained using only SVD without domain missmatch, when the train and test are fully matched in domain (second row of results in Table 5). This result shows the value of the augmentation considering the difference of 25-30 minutes of in-domain data in this experiment with respect to SVD's full training set which amounts 1 hour of speech. Note that in a realistic scenario of application, the availability of in-domain data is usually the weak point (this is the role of SVD in this experiment). While the availability of a data with similar characteristic of the problem is feasible to acquired (this is the role of AVFAD in this experiment). Hence, a difference of 25-30 minutes is significant in this case. In the third row of Table 5 we can see that the system highly improves the performance obtaining $ACC = 93.96\%$ and $EER = 6.31\%$ and using only 1 hour of in-domain data. Looking at the previous experiment training only with SVD, the training expansion (SVD+AVFAD) is able to increase the performance for ACC in 9.3% and EER in 8.8%. Note that looking at the ACC curve in Fig. 9, from 50 minutes there is a flat behaviour of the curve. However, we can not be sure if this is a maximum of performance because as usually happen in realistic applications, we don't have more data available for progressing the curve. Anyway, undoubtedly, this result is already remarkable for the development of voice disorder related technologies, considering the previous works using or not DNN were not able to go beyond 90% of accuracy.

VI. CONCLUSION AND FUTURE DIRECTIONS

Voice disorder related tasks are usually in low resources scenarios, i.e. very often the datasets are small. This is a clear problem for applying DNN-based solution. However, in this paper we have found the way to face these challenges by proposing an AVDD system based on DNN by using SS representations. The system achieves an impressive capability for discriminating between healthy and pathological speech either using a simple classifier such as the MLP or a more complex one such as the CT-Transformer. Jointly to a suitable strategy for expanding the training set considering the low availability of resources, this system reached a significant performance increase with more than 10% of absolute improvement in accuracy compared to the previous work. Results indicate the feasibility of using other healthy/pathological data even though this is out of the evaluation domain.

It is also remarkable that the performance achieved in these experiments was using phrases. We believe this is due to the increase in information provided by phonetic richness. The clinical practice usually employs sustained vowels, so most works in literature also utilize vowels. However, obtained results indicate that automatic systems make an appropriate use of phrases. So these paper results complement previous studies by including phonetic variability, which also increases the flexibility of applications where specific restrictions of the phonetical content are not feasible.

Conclusions about the usefulness of training extension line up with previous work for DNN approaches. However, our contribution is a suitable strategy for implementing data augmentation considering the resources available for voice disorder-related tasks. Namely, some set of prerecorded data but out of the domain of the application. The fact that obtaining performances over 90% of accuracy in these conditions and using the whole set of pathologies of the database is a milestone for voice disorder-related research. To contribute to reproducibility in the research community, we released the code and shared the lists for experiments https://github.com/dayanavivolab/s3prl/tree/voicedisorder.

In the future, we plan to approach the multiclass classification for detecting the specific pathology beyond the binary classification of the speech sample in healthy and pathological. For this task, we will study the clustering of the voice pathologies considering their expression in the speech signal. The main objective of this task is to make it feasible the classification of the AVDD system, which only evaluates the information provided by the speech signal. On the other side, we will explore the multi-head self-attention configuration architecture for adjusting better the DNN architecture.

REFERENCES

[1] N. Roy, R. M. Merrill, S. D. Gray, and E. M. Smith, "Voice disorders in the general population: Prevalence, risk factors, and occupational impact," *Laryngoscope*, vol. 115, no. 11, pp. 1988–1995, Nov. 2005.

[2] N. Bhattacharyya, "The prevalence of voice problems among adults in the United States," *Laryngoscope*, vol. 124, no. 10, pp. 2359–2362, Oct. 2014.

[3] N. Roy, R. M. Merrill, S. Thibeault, R. A. Parsa, S. D. Gray, and E. M. Smith, "Prevalence of voice disorders in teachers and the general population," *J. Speech, Lang., Hearing Res.*, vol. 47, no. 2, pp. 281–293, Apr. 2004.

[4] S. Hegde, S. Shetty, S. Rai, and T. Dodderi, "A survey on machine learning approaches for automatic detection of voice disorders," *J. Voice*, vol. 33, no. 6, p. 947, 2019.

[5] F. T. Al-Dhief, N. M. A. Latiff, N. N. N. A. Malik, N. S. Salim, M. M. Baki, M. A. A. Albadr, and M. A. Mohammed, "A survey of voice pathology surveillance systems based on Internet of Things and machine learning algorithms," *IEEE Access*, vol. 8, pp. 64514–64533, 2020.

[6] M. Pützer and W. Barry, "Saarbrücken voice database," Institute of Phonetics, Univ. Saarland, Saarbrücken, Germany, Tech. Rep., [Online]. Available: http://www.stimmdatenbank.coli.uni-saarland.de/

[7] M. Huckvale and C. Buciuleac, "Automated detection of voice disorder in the Saarbrücken voice database: Effects of pathology subset and audio materials," in *Proc. Interspeech*, Aug. 2021, pp. 1399–1403.

[8] P. Barche, K. Gurugubelli, and A. K. Vuppala, "Towards automatic assessment of voice disorders: A clinical approach," in *Proc. Interspeech*, Oct. 2020, pp. 2537–2541.

[9] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat, K. H. Malki, and M. A. Bencherif, "An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification," *J. Voice*, vol. 31, no. 1, p. 113, Jan. 2017.

[10] M. Alhussein and G. Muhammad, "Automatic voice pathology monitoring using parallel deep models for smart healthcare," *IEEE Access*, vol. 7, pp. 46474–46479, 2019.

[11] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "WAV2VEC 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds. Red Hook, NY, USA: Curran Associates, 2020, pp. 12449–12460. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[12] W.-N. Hsu, B. Bolte, Y.-H. Hubert Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," 2021, *arXiv:2106.07447*.

[13] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-Scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2021.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–21.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[16] S.-W. Yang, P.-H. Chi, Y.-S. Chuang, C.-I.-J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K.-T. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-Y. Lee, "SUPERB: Speech processing universal PERformance benchmark," in *Proc. Interspeech*, Aug. 2021, pp. 1194–1198.

[17] S. R. Kadiri and P. Alku, "Analysis and detection of pathological voice using glottal source features," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 367–379, Feb. 2020.

[18] *Kay Elemetrics, Multi-Dimensional Voice Program (MDVP) (Computer Program)*, Kay Elemetric Corp., Lincoln Park, NJ, USA, 2012.

[19] D. M. González, E. Lleida, A. Ortega, A. Miguel, and J. A. V. López, "Voice pathology detection on the saarbrücken voice database with calibration and fusion of scores using multifocal toolkit," in *Advances in Speech and Language Technologies for Iberian Languages* (Communications in Computer and Information Science), vol. 328. Madrid, Spain: Springer, Nov. 2012, pp. 99–109.

[20] H. Wu, J. Soraghan, A. Lowit, and G. Di-Caterina, "A deep learning method for pathological voice detection using convolutional deep belief networks," in *Proc. Interspeech*, Sep. 2018, pp. 446–450.

[21] P. Harár, J. Alonso, J. Mekyska, Z. Galáz, R. Burget, and Z. Smekal, "Voice pathology detection using deep learning: A preliminary study," in *Proc. Int. Conf. Workshop Bioinspired Intell. (IWOBI)*, Jul. 2017, pp. 1–4.

[22] M. A. Mohammed, K. H. Abdulkareem, S. A. Mostafa, M. K. A. Ghani, M. S. Maashi, B. Garcia-Zapirain, I. Oleagordia, H. Alhakami, and F. T. AL-Dhief, "Voice pathology detection and classification using convolutional neural network model," *Appl. Sci.*, vol. 10, no. 11, p. 3723, May 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/11/3723

[23] Y.-T. Hsu, Z. Zhu, C.-T. Wang, S.-H. Fang, F. Rudzicz, and Y. Tsao, "Robustness against the channel effect in pathological voice detection," 2018, *arXiv:1811.10376*.

[24] M. Alhussein and G. Muhammad, "Voice pathology detection using deep learning on mobile healthcare framework," *IEEE Access*, vol. 6, pp. 41034–41041, 2018.

[25] L. Verde, G. De Pietro, M. Alrashoud, A. Ghoneim, K. N. Al-Mutib, and G. Sannino, "Leveraging artificial intelligence to improve voice disorder identification through the use of a reliable mobile app," *IEEE Access*, vol. 7, pp. 124048–124054, 2019.

[26] L. Verde, N. Brancati, G. De Pietro, M. Frucci, and G. Sannino, "A deep learning approach for voice disorder detection for smart connected living environments," *ACM Trans. Internet Technol.*, vol. 22, no. 1, pp. 1–16, Oct. 2021, doi: 10.1145/3433993.

[27] L. P. Violeta, W. C. Huang, and T. Toda, "Investigating self-supervised pretraining frameworks for pathological speech recognition," in *Proc. Interspeech*, Sep. 2022, pp. 41–45.

[28] A. Hernandez, P. A. Pérez-Toro, E. Noeth, J. R. Orozco-Arroyave, A. Maier, and S. H. Yang, "Cross-lingual self-supervised speech representations for improved dysarthric speech recognition," in *Proc. Interspeech*, Sep. 2022, pp. 51–55.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[30] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, "Object-centric learning with slot attention," in *Proc. NeurIPS*, 2020, pp. 1–14.

[31] V. Mingote, A. Miguel, A. Ortega, and E. Lleida, "Class token and knowledge distillation for multi-head self-attention speaker verification systems," 2021, *arXiv:2111.03842*.

[32] Z. Ali, M. Alsulaiman, G. Muhammad, I. Elamvazuthi, A. Al-Nasheri, T. A. Mesallam, M. Farahat, and K. H. Malki, "Intra- and inter-database study for arabic, english, and German databases: Do conventional speech features detect voice pathology?" *J. Voice*, vol. 31, no. 3, p. 386, May 2017. [Online]. Available: http://repository.essex.ac.uk/27215/1/Intra_and_Inter_Database_Study_for_Arabic_English_and_German_Databases.pdf

[33] Z. Fan, Y. Wu, C. Zhou, X. Zhang, and Z. Tao, "Class-imbalanced voice pathology detection and classification using fuzzy cluster oversampling method," *Appl. Sci.*, vol. 11, no. 8, p. 3450, Apr. 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/8/3450

[34] K. T. Chui, M. D. Lytras, and P. Vasant, "Combined generative adversarial network and fuzzy C-means clustering for multi-class voice disorder detection with an imbalanced dataset," *Appl. Sci.*, vol. 10, no. 13, p. 4571, Jul. 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/13/4571

[35] M. Pützer and J. Koreman, "A German database of pathological vocal fold vibration," Inst. Phonetics, Univ. Saarland, Saarbruecken, Germany, Tech. Rep., 1997, pp. 143–153. [Online]. Available: http://www.stimmdatenbank.coli.uni-saarland.de/help_en.php4

[36] L. M. Jesus, I. Belo, J. Machado, and A. Hall, "The advanced voice function assessment databases (AVFAD): Tools for voice clinicians and speech research," in *Advances in Speech-language Pathology*, F. D. M. Fernandes, Ed. Rijeka: IntechOpen, 2017, ch. 14, doi: 10.5772/intechopen.69643.

[37] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[38] N. Brümmer and J. D. Preez, "Application-independent evaluation of speaker detection," *Comput. Speech Lang.*, vol. 20, nos. 2–3, pp. 230–275, 2006.

[39] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, Apr. 2017, pp. 1–13. [Online]. Available: https://openreview.net/forum?id=rkE3y85ee

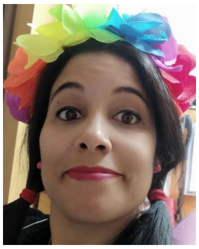
[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[41] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.

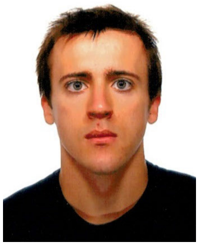
[42] F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: What speech, music, and sound have in common," *Frontiers Psychol.*, vol. 4, p. 292, May 2013, [Online]. Available: https://archive-ouverte.unige.ch/unige:97889

[43] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, p. 60, 2019, doi: 10.1186/s40537-019-0197-0.

[44] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, Sep. 2019, doi: 10.21437/Interspeech.2019-2680.



DAYANA RIBAS received the joint Ph.D. degree from the Technological University of Havana, in 2016, in the framework of research collaboration with INRIA, France, and Zaragoza University, Spain. Currently, she is a Scientific Researcher in the field of machine learning and statistical data analysis applied to speech and audio signals. Since 2018, she has been with the ViVoLab, Aragón Institute for Engineering Research (I3A). Her research interests include robust speech processing field, focusing mainly on robust speech processing, realism in speech and language processing, pathological speech processing, speech quality measures, and speaker recognition in noisy environments.



MIGUEL A. PASTOR received the master's degree in telecommunication engineering from the University of Zaragoza, in 2022. He is currently working as a Novel Researcher with the ViVoLab in topics related to the processing of para-linguistic information in speech signal with deep learning methods, including voice disorders and emotions. He is supervised by Dr. Dayana Ribas and Dr. Alfonso Ortega. His research interests include speech emotion recognition, voice disorder detection and assessment, health applications, and digital processing of audio from call center.

ANTONIO MIGUEL received the M.Sc. degree in telecommunication engineering and the Ph.D. degree from the University of Zaragoza, Spain, in 2001 and 2008, respectively. From 2000 to 2006, he was with the Communication Technologies Group, Department of Electronic Engineering and Communications, under a research grant. Since 2006, he has been an Associate Professor with the Department of Electronic Engineering and Communications, University of Zaragoza. His current research interests include acoustic modeling for speech and speaker recognition.

DAVID MARTÍNEZ received the M.Sc. and Ph.D. degrees from the University of Zaragoza, in 2009 and 2015, respectively. During this time, he visited the Brno University of Technology, SRI International, and University of Sheffield, working on language and speaker identification and voice pathology processing. Since 2015, he has been working with Agnitio, Cirrus Logic, and LumenVox as a Principal Technologist on different speech technologies.

ALFONSO ORTEGA received the degree in telecommunication engineering and the Ph.D. degree from the University of Zaragoza, Spain, in 2000 and 2005, respectively. He is currently an Associate Director of the Aragon Institute for Engineering Research (I3A), University of Zaragoza, where he is also an Associate Professor with the Department of Electronic Engineering and Communications. In 2006, he was a Visiting Scholar with the Center for Robust Speech Systems, University of Texas at Dallas, Dallas, TX, USA. He has participated in more than 50 research projects funded by national or international public institutions and more than 30 research projects for several companies. He is the author of more than 100 papers published in international journals or conference proceedings and several international patents. His research interests include speech processing, analysis and modeling, automatic speaker verification, and automatic speech recognition. His Ph.D. thesis, advised by Dr. Eduardo Lleida. He was a recipient of the Ph.D. Extraordinary Award and the Telefónica Chair Award to the best technological Ph.D.

EDUARDO LLEIDA received the M.Sc. degree in telecommunication engineering and the Ph.D. degree in signal processing from the Universitat Politecnica de Catalunya (UPC), Barcelona, Spain, in 1985 and 1990, respectively. From 1986 to 1988, he was involved in his doctoral work with the Department of Signal Theory and Communications, UPC. From 1989 to 1990, he was an Assistant Professor and from 1991 to 1993, he was an Associate Professor with the Department of Signal Theory and Communications, UPC. From February 1995 to January 1996, he was with the AT&T Bell Laboratories, Murray Hill, NJ, USA, as a Consultant in speech recognition. He is currently a Full Professor in signal theory and communications with the Department of Electronic Engineering and Communications, University of Zaragoza, Spain. He is a member of the Aragón Institute for Engineering Research (I3A), where he is heading the ViVoLab Research Group in speech technologies. He is the doctoral advisor of 12 doctoral students. He has managed more than 50 speech-related projects, being an inventor in several worldwide patents, and coauthored more than 200 technical papers in the field of speech, speaker and language recognition, speech enhancement and recognition in adverse acoustic environments, acoustic modeling, confidence measures, and spoken dialogue systems.