IberSPEECH 2022
14-16 November 2022, Granada, Spain

S3prl-Disorder: Open-Source Voice Disorder Detection System based on the Framework of S3PRL-toolkit

Dayana Ribas[1], *Miguel A. Pastor*[1], *Antonio Miguel*[1], *David Martinez*[2], *Alfonso Ortega*[1], *and Eduardo Lleida*[1]

[1]ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain
[2]LumenVox, Germany

{dribas,mapastor,amiguel,ortega,lleida}@unizar.es, david.martinez@lumenvox.com

Abstract

This paper introduces S3prl-Disorder, an open-source toolkit for Automatic Voice Disorder Detection (AVDD) developed in the framework of the S3prl toolkit. It focuses on a binary classification task between healthy and pathological speech in the Saarbruecken Voice Database (SVD). However, the framework left room for following extensions to multi-class classification to differentiate among pathologies and to incorporate more datasets. This work aims to contribute on the development of automatic systems for diagnosis, treatment, and monitoring of voice pathologies in a common framework, that allows reproducibility and comparability among systems and results.
Index Terms: Voice disorder, SUPERB, pathological speech, Saarbruecken Voice Database, self-supervised, toolkit, deep neural networks.

1. Introduction

Voice pathologies have an impressive prevalence among population. Previous studies [1] reported that almost 30% of general population having experienced a period of time with a problem of voice. While in [2] authors reported that one in 13 adults has voice problems annually, which means a prevalence of 7.5% in adults. There is an opportunity to use smart solutions to assess voice pathologies as part of remote health services contributing with the increase of early diagnosis. 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. AVDD is the task that opens the gate of health assistance systems, from the detection of a voice disorder to the specification of the disease and its severity.

Many research efforts have focused on this aim, see many of them in following surveys [3, 4]. The extensive compilation of previous works in these papers show that there are many feature representations and classification models that have been explored for the task of AVDD. However, at the same time, there is a wide variety among experimental setups and results, as well as a lack of free available toolkits or implementations of the systems proposed. For instance, see in [3] the variance in reported results for binary classification using the same dataset, e.g. in SVD from 70% to 100% (this responds to different audio samples selection for train and test sets). This fact makes difficult to arrive into conclusions about the real performance of this kind of systems. Moreover, beyond some feature extraction toolkit, namely Multidimensional voice program parameters (MDVP) [5] -which is non-freely available-, we did not find other toolkit or author provided implementation for AVDD systems. Together with the variability of experimental setups described in papers, this hinders the reproduction of results. This is a problem that brings concerns to the research community related to voice disorder processing and there are some previous approaches to this issue [6, 7].

This work introduces an open-source AVDD system[1] that is mainly based on Deep Neural Network (DNN) and has been implemented in the framework of S3PRL-toolkit [8]. The system was developed with SVD [9], which is a database widely used among previous works considering that this is one of the few options of free available corpus with healthy and pathological voice recordings. The advantage of developing in the framework of S3prl is that all available feature extraction (FE) methods -called upstreams- can be used for the AVDD task. These include spectral and cepstral representations such as FFT, Filterbank and Mel Frequency Cesptral Coefficients (MFCC), as well as the Self-Supervised (SS) representations such as Wav2vec, Wav2vec2, and Hubert. The community around S3prl toolkit starts to grow up with contributions for enlarge the resources available [10, 11].

Our contribution is in the voice disorder detection system, integrated to S3prl as a new downstream. It includes two classification models based on DNN, a CNN architecture with Self-Attention and a Class-token (CT) Transformer previously presented in [12]. All of them can be combined with any FE method among available upstreams allowing to quickly design different AVDD systems. We believe this toolkit and the analysis around it, contribute to the reproducibility and comparability of results in the AVDD research community. Also it contributes to the SUPERB project.

In the following, section 2 comments on the problem of comparability and reproducibility of results in the field of voice disorder processing. Section 3 and 4 presents the AVDD system and describes the available resources for data pre-processing, representations, modeling, and evaluation metrics. Also we evaluate some combinations of FE and classification models for assessing the AVDD system and show classification results. In order to approach to a real-life scenario of disorder detection, we evaluate the system with the whole amount of data in SVD including the full variability of pathologies available. Finally, section 5 concludes the paper.

2. Previous work

There are several approaches to 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

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

10.21437/IberSPEECH.2022-28

[13]. There are some recent reviews about previous work that agglutinate brief system descriptions and results of many of the papers presented for voice disorder detection [3, 4].

The problem of results comparability. The review in [3] shows a large table of research works describing the database, methods, and system performance reported by authors. It is interesting that accuracies up to 100% are reported for SVD. However, note that each of these works use a selection of data with different set of pathologies, usually including those that have large amount of audio samples or have a high level of affectation. This fact makes hard the task of comparing with previous work. 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 [6]. Recently, Huckvale et al. [7] approached over this problem using SVD. They found that re-implementations of previous works reporting performances of 77% and 94% using DNN-based systems, actually achieved 70% and 71% (see table 3 in [7]).

The problem of reproducibility. From the discussion in the previous paragraph introduces the problem about reproducibility of results in the field of voice disorder detection. As is possible to find previous developments of computational tools in different areas related to voice disorders assessment, the availability of toolkits for AVDD systems is quite scarce. There are very useful toolkits for helping patients and clinicians to carry out speech and language therapy. For instance, there are VOCALIZA [14] and PEAKS [15]. Also there is this review of Chen at el. [16] about computer-aided systems for speech therapists of speech disorders. On the other hand, there are some interesting toolkits for speech analysis, such as NeuroSpeech[2] [17], which has available plenty of feature extraction implementations that are usually employed for representing healthy and pathological speech. Moreover, there are well-known toolkits to directly perform feature extraction, namely Opensmile [18] and MDVP [5].

Beyond the works mentioned, we did not find any other toolkit or author's provided implementation specifically for AVDD systems. At the end, most experiments reported end at creating the AVDD system by means of machine learning general toolkits, such as Sklearn or Weka, or even using implementations of models in general languages such as Python, Matlab or C++, as well as the performance metrics, namely Accuracy (ACC), Unweighted Average Recall (UAR), etc. The major issue in the case of own implementations are the availability of same list of data for training and evaluation than the state-of-the-art system. This is essential for comparison purposes. Also, in own re-implementations, the small details of the implementations are usually hard to follow from the paper description only.

The problem of meaningful results. Another issue in this field is the frequency of reports of non-meaningful results in terms of classification. For example, SVD has an approximately proportion of 30% healthy by 70% pathological samples. This means that in a binary classification approach, accuracy results less than 70% have the same meaning of a random classification. I.e. the classifier can output all data pathological and the result will be the same. You can see many examples of this issue in papers listed in the following survey [3].

[2]Recent version: https://github.com/jcvasquezc/DisVoice

3. AVDD system: Pre- and Post- Processing

Motivated to contribute to the effective comparability and reproducibility of developments towards AVDD systems this work presents S3prl-Disorder, a freely available toolkit for AVDD. In this section we describe the stages of pre- and post- processing of the AVDD system, including datasets available, configuration, performance metrics and visualization resources. This is implemented in the framework of S3PRL-toolkit [8], see the descriptive chart in Fig. 1), where methods in blue are those provided by S3prl and those in orange are our contributions, including functions for reporting results and some options for visualization of the system performance.

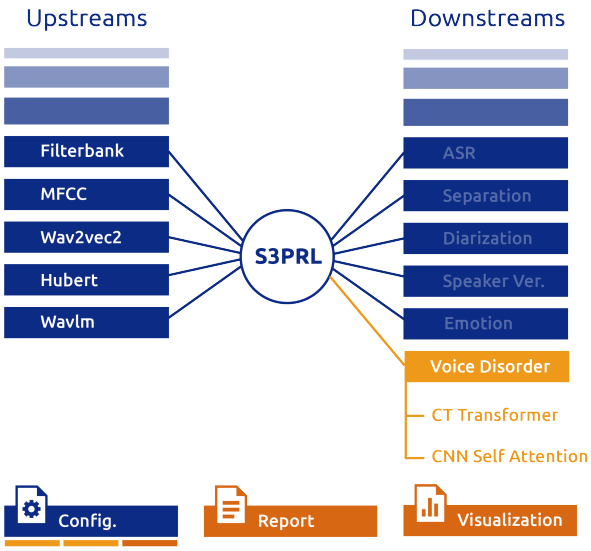

Figure 1: *S3prl-Disorder: AVDD system in the framework of S3prl toolkit.*

3.1. Data preprocessing

3.1.1. Databases

Saarbruecken Voice Database (SVD)[3] is a dataset of healthy and pathological speech in the German language. It contains voice recordings of 687 healthy persons: 428 females and 259 males, and 1356 non-healthy persons: 727 females and 629 males with one or more of the included 71 pathologies (see more details in [19]). Each recording session includes the vowels /a,i,u/ with low, high and neutral pitch, and the phrase "Guten Morgen, wie geht es Ihnen?" ("Good morning, how are you?"). From the binary point of view, the dataset distribution is one healthy by two pathological, therefore the minimum reasonable accuracy is around 70%, less than this, the system is not contributing. For evaluating the AVDD system proposed we used audio from phrases.

3.1.2. Train and Evaluation Sets

The system works with 5-fold train and test lists of audio and labels created with all data available in SVD. We used the same audio subset as in [7]: "all pathologies" thanks to the collaboration of authors. As labels we use HEALTH for healthy and

[3]http://www.stimmdatenbank.coli.uni-saarland.de/help_en.php4

PATH for pathological. Folds in SVD have one healthy by two pathological samples, and the audio of speakers included in the training is not in the test partition.

To illustrate the performance of the toolkit, the experiments in this paper are computed with phrases. However, list for sustained vowels are also included in the repository. In this first approach, the classification models included are configured for the binary classification task between healthy vs. pathological speech. This task has applicability in the use case of diagnosis support where the doctor performs a quick test to decide the need of a hospital visit [4]. Next steps will include the multiclass classification for considering the bulk of pathologies labels in the dataset.

3.2. Configuration

To define configuration details for setting parameters in each stages of the system there is available a document in 'yaml' format. See a guide at "$config\_README.yaml''$ that includes a description for each parameter and the value options according to the task. This configuration file is based on the one in S3prl toolkit, but additionally, we have introduced new parameters related to the AVDD task.

3.3. Execution guide

To run a test you can use the script "$run\_disorder.sh''$. This script executes a loop over the five folds of SVD to perform train and evaluation. The following options are configurable in the script, you just need to look in in the script and change the options accordingly:

- For training there is option to run use representation frozen "basic" or to perform finetunning "finetune".
- Choose the audiotype for selecting SVD lists: "phrase", "aiu" (concatenated neutral vowels a, i, u), "a_n" (the neutral vowel a).
- Select the upstream and models for classification.

3.4. Performance metrics

Reports in "$log\_acc\_auc.log''$ with the following metrics are used to evaluate the classification performance: ACC (eq:1) and Balanced Classification Accuracy, also known as 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 rates (TP, FP, TN, FN). If data are balanced, ACC and UAR should be 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 Area Under ROC Curve (AUC) [20] and Equal Error Rate (EER) [21] because they can provide a measure independent of the threshold and operation point.

Besides these metrics, in the result report of the toolkit there are available the precision consisting of the ratio of all correctly positively classified samples (TP) to all positive classified samples (TP and false positive—FP). Recall, consisting of the ratio of all correctly positively classified samples (TP) to the number of all samples in a tested subgroup (TP and false negative FN), indicating a class-specific recognition accuracy. Note that, the recall and precision considering both classes are computed as the average of recall and precision for healthy and pathology individual classes. F1-score, defined as the harmonic mean of the precision and recall.

In "$truth\_predict\_score.txt''$ there is the classification result for each audio in the test set in the following format: audio_name truth_label predicted_label score.

3.5. Analysis and visualization

At each epoch, there is a drawing to see the evolution of accuracy with epochs ("$results\_by\_epoch.png''$). Optionally you can visualize the ROC curve at each epoch by setting 'roc: 1' in the configuration file. This remain saved at a folder called 'figures'. Also you can save the embeddings of the test set in pickle format. For this you need to set 'embeddings: 1' in the configuration file and they will be stored in a folder called 'embeddings'.

There is also the script "$compute\_umap\_tsne.py''$ available for computing visualizations of embeddings using dimensionality reduction with TSNE [22] and UMAP [23]. This can be optionally executed using the embeddings saved in each evaluation round if the config parameter 'embeddings: 1' was set. Results are png files saved in a folder called '$umap\_tsne'$, created in the same folder where embeddings are located.

4. AVDD system: Representations and Models

In this section we describe the main stages of the AVDD system and the methods available for representation and classification. As it is implemented in the framework of S3PRL-toolkit [8] some methods are re-used from S3prl and others are introduced by us.

4.1. Representations

There are several approaches to AVDD systems studying suitable representations for pathological voice (see the review in [13]). Among them, spectral and cepstral features, such as Filterbank or MFCC, are frequently employed for representing healthy and pathological cues [24, 25, 26, 27].

More recently, SS representations are widely used in several speech-related areas (see results in the following benchmark: `superbbenchmark.org`). Models such as Wav2vec2 [28], and Hubert [29] are known to be able of finding underlying relations on data and providing substantial representations. These models are trained with a significant amount of general speech data, without any healthy/pathological awareness. Then, we use them to create feature vectors for train/test sets as part of the first processing stage of the system. They describe the sequential evolution of the utterance, i.e. there is one feature by frame.

In order to illustrate the performance of these features with healthy and pathological speech, we carried out an experiment of binary classification using SVD. We employed some spectral, cepstral and SS representations followed by and attentive pooling (eq. 3) to unify the sequence of feature vectors for each frame, and then a feed-forward linear layer to classify in healthy and pathological.

$$P = Softmax(W_p F^T) F \tag{3}$$

where F is the sequence of feature vectors or SS embeddings by frame. W_p is the trainable matrix.

Table 1 shows the performance metrics for binary classification between healthy and pathological speech using the AVDD system with mentioned representations. For spectral and cepstral features we have used the default configuration in S3prl toolkit implemented with torchaudio toolkit. FFT consists in a 256-FFT with windowing of 25 ms and overlap of 10 ms. Fbank is a Mel Filterbank with 80 dimensions plus delta and double delta, and finally MFCC is 13 dimensional plus delta and double delta.

For SS representations we have used the base version of the pretrained models of Wav2vec, Wav2vec2, and Hubert to create embeddings for the audio phrases files in SVD. SS representations outperformed spectral and cepstral feature extraction sets. However, the performance among SS representations is very similar.

Represent.	ACC(%)	UAR(%)	AUC(%)	EER(%)
Spectral/Cepstral Representations				
FFT	73.04	69.01	71.42	31.86
Fbank	76.31	72.33	75.51	27.44
MFCC	72.74	67.82	69.93	32.42
Self-Supervised Representations				
Wav2vec	80.18	76.77	85.29	23.03
Wav2vec2	**81.04**	**78.15**	**82.85**	**21.05**
Hubert	80.33	77.05	86.84	22.40

Table 1: *ACC, UAR, AUC and EER metrics for all test samples in 5-folds for different representations in a binary classification between healthy and pathological speech using a linear layer neural network as classifier.*

4.2. Models

For classification we included two models based on DNN, a CNN-Self Attention architecture already available in the *Emotion* downstream of S3prl and a Class-Token Transformer inspired in the Vision Transformer (ViT) [30] and the concept of Class-Token (CT) [31]. This consists of concentrating the class information in a single vector through several layers of Self Attention Mechanisms (MSA) for the classification task. The CT vector learns a global description of the utterance by processing the relevant information from the whole sequence of SS embeddings through a configurable number of heads and layers in the MSA block.

We decided to include these architectures to be able of testing different alternatives for classification in the AVDD system. However, as these are not the only models useful for this task, there is room to include any other model based on DNN we would like to evaluate.

To illustrate the performance of the system using a DNN-based model for classification, the table 2 shows performance results for the CNN-Self Attention Classifier with the representations evaluated in the previous section. Comparing with results obtained in the previous table 1, where we used a simple linear layer to see the representations performance, we can see that the introduction of the CNN-based classifier achieves improvement for all representations. These improvements are around $2-3\%$ for most representations, with a maximum of 4.42% in terms of ACC or 9.91% in terms of EER for MFCC features. We see that the performance among SS representations is again very similar. Anyway, the best performance among all representations agree with results in previous table 1, this is achieved by Wav2vec2 with ACC=83.80% and EER=18.09%.

Represent.	ACC(%)	UAR(%)	AUC(%)	EER(%)
CNN-Self Attention based Classifier				
FFT	75.96	74.09	81.14	25.55
Fbank	78.02	75.81	83.31	23.41
MFCC	77.16	76.69	82.91	23.41
Wav2vec	83.00	79.89	87.02	18.61
Wav2vec2	**83.80**	**81.86**	**86.74**	**18.09**
Hubert	82.95	80.18	85.52	18.30

Table 2: *ACC, UAR and EER metrics for all test samples in 5-folds with SVD for binary classification (with and without pathology) using the CNN-based classifier.*

5. Conclusions and future work

In this paper we have presented an open source system for Automatic Voice Disorder Detection in the framework of S3prl toolkit. Our main objective for releasing this toolkit is contributing to the effective comparability and reproducibility of developments in the field of pathological speech automatic processing. Such that the idea of having systems for helping medical doctors for the diagnosis, treatment, and monitoring of voice pathologies become a fact as soon as possible.

To evaluate the toolkit performance, we presented results for binary classification between healthy and pathological speech using some Self-Supervised representations available in S3prl toolkit and the DNN models included in the Voicedisorder downstream, namely a CT-Transformer and a CNN Self Attention architecture. Anyway, the framework also left room for next extensions to further DNN models for classification. Experimental results showed the benefits of SS representations on top of spectral and cepstral features for classifying between healthy and pathological speech using Saarbruecken Voice Database. In next steps we plan to introduce other free available databases for evaluation, for instance AVFAD in Portuguese and VOICED in Italian. We also plan to extend the classifier design to evaluate multiple classes in order to differentiate among pathologies.

6. Acknowledgement

This work was supported in part by the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie Grant 101007666; in part by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU" / PRTR under Grants PDC2021-120846-C41 PID2021-126061OB-C44, and in part by the Government of Aragon(Grant Group T36_20R).

7. References

[1] N. Roy, R. M. Merrill, and S. D. Gray, "Voice disorders in the general population: prevalence, risk factors, and occupational impact." *The Laryngoscope*, vol. 115, pp. 1988–1995, 2005.

[2] N. Bhattacharyya, "The prevalence of voice problems among adults in the United States." *The Laryngoscope*, vol. 124, p. 2359–2362, 2014.

[3] S. Hegde, S. Shetty, S. Rai, and T. Dodderi, "A Survey on Machine Learning Approaches for Automatic Detection of Voice Disorders," *Journal of Voice*, vol. 33, no. 6, pp. 947.e11–947.e33, 2019.

[4] 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. 64 514–64 533, 2020.

[5] "Kay Elemetrics, Multi-dimensional voice program (MDVP) (computer program)," 2012.

[6] M. A. Mohammed, K. H. Abdulkareem, S. A. Mostafa, M. Khanapi Abd 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," *Applied Sciences*, vol. 10, no. 11, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/11/3723

[7] M. Huckvale and C. Buciuleac, "Automated Detection of Voice Disorder in the Saarbrücken Voice Database: Effects of Pathology Subset and Audio Materials," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2021, pp. 1399–1403.

[8] S. wen 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. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2021, pp. 1194–1198.

[9] M. Pützer and W. Barry, "Saarbrücken Voice Database," institute of Phonetics, Univ. of Saarland, http://www.stimmdatenbank.coli.uni-saarland.de/.

[10] W.-C. Huang, S.-W. Yang, T. Hayashi, H.-Y. Lee, S. Watanabe, and T. Toda, "S3prl-vc: Open-source voice conversion framework with self-supervised speech representations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6552–6556.

[11] H.-S. Tsai, H.-J. Chang, W.-C. Huang, Z. Huang, K. Lakhotia, S.-w. Yang, S. Dong, A. Liu, C.-I. Lai, J. Shi, X. Chang, P. Hall, H.-J. Chen, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee, "SUPERB-SG: Enhanced speech processing universal PERformance benchmark for semantic and generative capabilities," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8479–8492. [Online]. Available: https://aclanthology.org/2022.acl-long.580

[12] D. Ribas, M. A. Pastor, A. Miguel, D. Martinez, E. Lleida, and A. Ortega, "Automatic voice disorder detection using self-supervised representations and token-class transformer." 2022.

[13] S. R. Kadiri and P. Alku, "Analysis and Detection of Pathological Voice using Glottal Source Features," *Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 367–379, 2020.

[14] C. Vaquero, O. Saz, E. Lleida, J. Marcos, and C. Canalís, "Vocaliza: an application for computer-aided speech therapy in spanish language," in *Proceedings of IV Tecnologías del habla*, 2006, p. 321–326.

[15] P. Kitzing, A. Maier, and V. Ahlander, "Automatic speech recognition (asr) and its use as a tool for assessment or therapy of voice, speech, and language disorders," *Logop. Phoniatr. Vocology*, vol. 34, no. 6, p. 91–96, 2009.

[16] Y.-P. Chen, C. Johnson, P. Lalbakhsh, T. Caelli, G. Deng, D. Tay, S. Erickson, P. Broadbridge, A. Refaie, W. Doube, and M. Morris, "Systematic review of virtual speech therapists of speech disorders," *Comput. Speech Lang.*, vol. 37, p. 98–128, 2016.

[17] J. Orozco-Arroyave, J. Vásquez-Correa, J. Vargas-Bonilla, R. Arorac, N. Dehakc, P. Nidadavoluc, H. Christensen, F. Rudzicze, M. Yancheva, H. Chinaei, A. Vann, N. Vogler, T. Bockleth, M. Cernaki, J. Hannink, and E. Noth, "NeuroSpeech: An open-source software for Parkinson's speech analysis," *Digital Signal Processing*, vol. 77, p. 207–221, 2018.

[18] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proceedings of ACM Multimedia (MM)*, 2010, pp. 1459–1462.

[19] M. Pützer and J. Koreman, "A German database of pathological vocal fold vibration," pp. 143–153, 1997. [Online]. Available: https://www.coli.uni-saarland.de/publikationen/softcopies/Putzer:1997:GDP.pdf

[20] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[21] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech Language*, vol. 20, no. 2, pp. 230–275, 2006.

[22] L. van der Maaten and G. E. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[23] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2018. [Online]. Available: http://arxiv.org/abs/1802.03426

[24] J. I. Godino-Llorente and P. G. Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 2, p. 380–384, 2004.

[25] C. R. Watts and S. N. Awan, "Use of spectral/cepstral analyses for differentiating normal from hypofunctional voices in sustained vowel and continuous speech contexts," *J. Speech, Lang., Hearing Res.*, vol. 54, no. 6, p. 1525–1537, 2011.

[26] 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 - IberSPEECH 2012 Conference, Madrid, Spain, November 21-23, 2012. Proceedings*, ser. Communications in Computer and Information Science, vol. 328. Springer, 2012, pp. 99–109.

[27] J. A. G. García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art," *Biomed. Signal Process. Control*, vol. 51, p. 181–199, 2019.

[28] 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*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[29] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *CoRR*, vol. abs/2106.07447, 2021. [Online]. Available: https://arxiv.org/abs/2106.07447

[30] 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 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

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