Voice Disorder Analysis: a Transformer-based Approach

Interspeech 2024
1-5 September 2024, Kos, Greece

Voice Disorder Analysis: a Transformer-based Approach

Alkis Koudounas[♣], *Gabriele Ciravegna*[♣], *Marco Fantini*[♢♡],
Giovanni Succo[♡♠], *Erika Crosetti*[♡], *Tania Cerquitelli*[♣], *Elena Baralis*[♣]

[♣]Politecnico di Torino, Turin, Italy [♢]San Feliciano Hospital, Rome, Italy
[♡]SCDU Otorinolaringoiatria, Head Neck Cancer Unit, Ospedale San Giovanni Bosco, Turin, Italy
[♠]Dipartimento di Oncologia, Università degli Studi di Torino, Turin, Italy
alkis.koudounas@polito.it, gabriele.ciravegna@polito.it

Abstract

Voice disorders are pathologies significantly affecting patient quality of life. However, non-invasive automated diagnosis of these pathologies is still under-explored, due to both a shortage of pathological voice data, and diversity of the recording types used for the diagnosis. This paper proposes a novel solution that adopts transformers directly working on raw voice signals and addresses data shortage through synthetic data generation and data augmentation. Further, we consider many recording types at the same time, such as sentence reading and sustained vowel emission, by employing a Mixture of Expert ensemble to align the predictions on different data types. The experimental results, obtained on both public and private datasets, show the effectiveness of our solution in the disorder detection and classification tasks and largely improve over existing approaches.
Index Terms: pathological voice disorder, transformer model, synthetic data, mixture of experts, data augmentation

1. Introduction

Vocal disorders are important pathologies affecting a significant portion of the population and exerting a substantial impact on patient quality of life [1–4]. These disorders may originate from various causes, including both benign and malignant conditions, and neurodegenerative disorders [5–7]. Diagnosis often relies on clinician auditory assessments of patient voices. An early diagnosis and treatment are crucial, as they greatly improve patient prognosis. Thus, an automated tool that can detect and classify these disorders would be of crucial importance.

Deep learning is transforming the field of medicine by enabling new ways of diagnosing and treating diseases [8]. One of the promising applications of deep learning is non-invasive diagnostics, which can reduce the need for invasive procedures and improve the quality of life for patients [9]. However, while deep learning has made notable strides in non-invasive diagnostics for skin cancer [10], diabetic retinopathy [11], and atrial fibrillation [12], voice analysis remains an under-explored domain. The main reasons are related to (i) the scarcity of voice data related to pathological issues that prevents effectively employing powerful models, and (ii) the intrinsic complexity of pathological voice data and the diversity of the data used for the diagnosis.

In this paper, we show that it is possible to overcome these challenges respectively by (i) generating synthetic data based on Text-to-Speech (TTS) technology and designing a strong data augmentation pipeline to enrich and balance the training data, and (ii) employing a Mixture of Experts (MoE) ensemble, combining Transformer models [13] trained on different recording types (e.g., vowel emission and sentence readings) to capture the voice nuances available across data types.

We tested our approach on two public datasets, namely SVD [14] and AVFAD [15], and on an internal Italian Pathological Voice (IPV) dataset. The experimental validation shows that our solution significantly improves both the AUC in voice disorder detection and the F1 score in pathology classification with respect to existing models.

Related Works. Automatic voice disorder analysis has been the subject of several studies in the literature. On the one side, a few works employed multi-layer perceptron over extracted features, including acoustic features and Mel-frequency cepstral coefficients (MFCC) [16, 17]. On the other side, Convolutional Neural Networks (CNN) have been applied to 2D representations of the audio signal, such as the Mel spectrograms and MFCCs cepstrograms [18, 19]. Hybrid architectures combining CNNs and Recurrent Neural Networks (RNNs) networks have also been proposed to improve the performance over long voice signal [20]. Other studies have utilized models working directly on the audio signal, such as 1D-CNN [21] and Transformers [22]. The latter is reported to boost the performance also in the related context of dysarthric speech [23–25]. In this paper, we use transformer models working on raw speech data as well. However, unlike [22] that pre-trains the model on a separate dataset before fine-tuning it on the voice pathology detection task, we show that training on augmented and synthetic data and using an ensemble model allows us to achieve high performance.

2. Methodology

We propose a pipeline to fully exploit the generalization capability of transformer models in the context of non-invasive voice disorder analysis. Similarly to other deep learning models, transformers require a huge quantity of data to be trained, or even fine-tuned. Moreover, each class must be sufficiently represented to avoid implicit biases. On the contrary, publicly available datasets tend to be small and often unbalanced, exhibiting an under-representation of either pathological or healthy individuals. In this work, we propose to overcome this issue by generating synthetic data conditioned by the voices of actual patients, both healthy and pathological. This approach, together with a strong standard data augmentation pipeline, aims to enhance and balance the distribution of the training datasets.

To get the most information from the available data, we also propose to consider all data types at the same time. Indeed, the detection of a voice disorder, as well as the classification of the specific pathology causing the disorder, may benefit differently from the type of recorded sample. Unlike current literature focusing on a single recording type (i.e., either on sustained vowels [17–19, 21] or on read sentences [16, 22]), in this paper we

10.21437/Interspeech.2024-1122

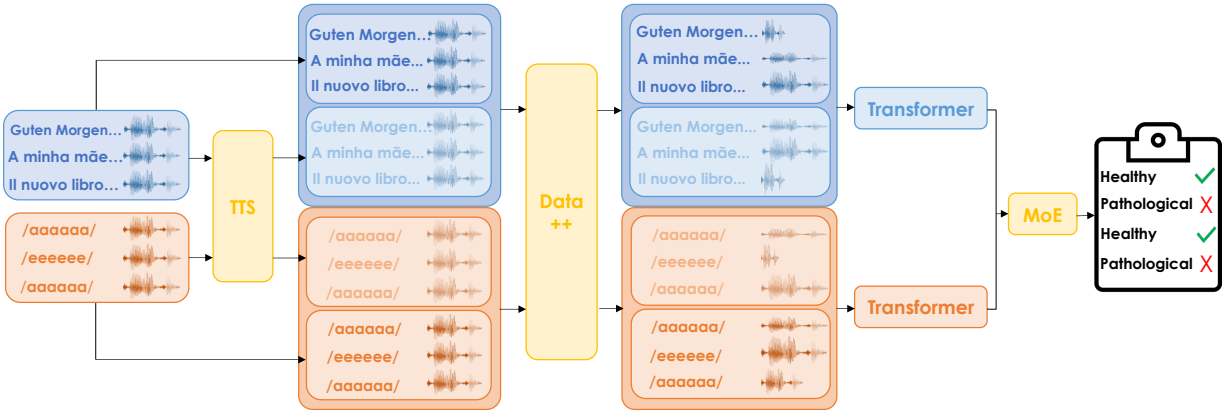

Figure 1: ***Schematic diagram of our pipeline.*** *We synthesize (`TTS`) and augment (`data++`) the training data, separately for sentences and sustained vowels, to train two dedicated `transformer` models. We then align their predictions with a Mixture of Experts (`MoE`).*

employ a MoE ensemble, combining the predictions of transformer models analyzing different data types. In the following, we introduce the main components of the proposed pipeline, which is represented in Figure 1.

Synthetic Data Creation. To generate both healthy and pathological voices, we propose employing Text-to-Speech (`TTS`) and conditioning the generation process to the specific class. The TTS generation process is conditioned by incorporating the learned embeddings derived from the vocal characteristics of the real data within each class. We employ a state-of-the-art multilingual TTS model [26] that allows the automatic generation of voices for diverse datasets in different languages. Synthesizing voices for pathological conditions through TTS may introduce potential pitfalls, as the synthesized voice may not accurately capture the nuanced characteristics of the specific pathologies experienced by the patient. To avoid this issue, we checked the generalization capability of the synthesized voices by training a model entirely on synthetic data and testing it on the original data (see Section 4). To further enrich our data, we also apply a strong data augmentation pipeline (`data++`) with pitch shifting, time stretching, and noise addition. We better describe it in Section 3.

Mixture of Experts. To exploit all available data and acknowledging the distinct characteristics of each data type in representing one's pathologies, we propose to train separate models for each input type. We then introduce a "shallow" Mixture of Expert (`MoE`) framework to align the predictions of all models. Specifically, we select the predictions from the model providing higher confidence. Model confidence is estimated by the entropy of predicted probabilities: lower entropy implies higher certainty, indicating a confident prediction for the given sample. By leveraging this integrated approach, we extract a comprehensive understanding of the pathology representation within the voice data. The experimental results show that this approach works better with respect to considering all data types within a single model.

We further consider a different pre-training for each model. Specifically, we propose to fine-tune on sustained vowels a model pre-trained on the Audioset dataset [27], and on sentence reading a model pre-trained on the classic LibriSpeech dataset [28]. The rationale behind this choice is based on the distinct strengths of each dataset. LibriSpeech is well suited for capturing the nuances of sentence readings while, following the intuition of [18], a model pre-trained on Audioset is expected to better represent vowels, as it comprises speech and vocalization samples, such as laughter, screaming, humming, etc. In the following, we will refer to this ensemble as `MoE`*.

3. Experimental Setup

This section details the datasets, models, and training procedures used for the experiments.

3.1. Datasets

We consider two publicly-available datasets, the German SVD [14] and the Portuguese AVFAD [15]. We also assess the performance of our approach on an internal Italian dataset (IPV). We refine public datasets to focus on similar recordings, i.e., only readings of given sentences and emissions of sustained vowels, both in a typical setting (i.e., with normal pitch). Table 1 summarizes the characteristics of each dataset. Notice that the same patients may be involved in several recordings.

SVD. The Saarbruecken Voice Database (SVD)[1] [14] includes voice recordings and electroglottography (EGG) data, with 13 files per recording session, incorporating vowels /a, i, u/ with pitch variations (normal, high, low, rising-falling) and a sentence reading task. For consistency, we only considered the sentence reading and the normal pitch vowels. The dataset does not aggregate pathologies into macro-classes. Thus, for the disorder classification task, we considered the 6-most frequent classes.

AVFAD. The Advanced Voice Function Assessment Databases (AVFAD)[2] [15] collect audio recordings capturing participants performing various vocal tasks, including sustaining vowels /a, e, o/, reading six sentences, reading a phonetically balanced text, and engaging in spontaneous speech, everything repeated three times. For consistency, we concatenate the six sentences together, having three repetitions for each obtained audio. We also concatenate the sustained vowels by repetition, thus obtaining 3 audios, one for each vowel.

IPV. The Italian Pathological Voice (IPV) is a novel dataset we use for further testing the proposed method. The study recruited participants from several private phoniatric and speech-therapy practices and hospitals in Italy. Participants included both eu-

[1] https://stimmdb.coli.uni-saarland.de/
[2] https://acsa.web.ua.pt/AVFAD.html

Table 1: *Characteristics of employed datasets (Ds), including language (L), number of healthy (#H) and pathological (#P) individuals, number of sentence readings (#S) and sustained vowel recordings (#V), number of pathological classes (#C), macro classes (#MC) and average audio duration (T (s)).*

Ds	L	#H	#P	#S	# V	#C	#MC	T (s)
SVD [14]	DE	687	1356	2043	6129	71	(6)[4]	1.73
AVFAD [15]	PT	346	363	1989	1989	25	8	15.86
IPV	IT	173	340	513	513	15	6	12.89

phonic individuals seeking otolaryngological evaluations and participants with various degrees of dysphonia. Dysphonic participants exhibited organic and functional voice disorders of varying severity. Each participant underwent videolaryngostroboscopic examinations, perceptual voice evaluations, and acoustic voice analysis conducted by experienced physicians. Data collection involved two tasks: sustained production of the vowel /a/, and a reading of five phonetically balanced sentences derived from the Italian adaptation of the CAPE-V [29]. Voice samples were recorded under standardized conditions, keeping a consistent mouth-to-microphone distance of 30 cm and ensuring a quiet environment with a signal-to-noise ratio exceeding 30 dB. The study adhered to the principles of the Declaration of Helsinki, with all participants providing informed consent. Data analysis was conducted retrospectively and anonymously on the recorded voice samples. Further information regarding the dataset is available in the project repository[3].

3.2. Models and training procedure

Compared Models. We reproduced the 1D and 2D CNN architectures presented in [18, 21], respectively.[5] For the CNN-2D, we experimented with various MFCCs. We show the results for MFCC=40, which yielded the best outcomes overall. We also report the performance of vanilla transformers. Following [22], the suite of transformers evaluated in this work includes wav2vec 2.0 [30], WavLM [31], and HuBERT [32], in their base sizes. For wav2vec 2.0 and HuBERT, we further employ the version pre-trained on Audioset introduced in [33].

Training procedure. We performed a 10-fold cross-validation (CV) for all the considered datasets. A robust data augmentation pipeline was employed to randomly replicate various noisy environments and introduce alterations in pitch (higher and lower) and time (stretched and compressed), either individually or in combination. We applied a more frequent and intense data augmentation for sentences and a slightly weaker approach for vowels, as overly aggressive augmentation for this audio type reduces the quality of the recording. Detailed information on the hyperparameter setup can be found in the project repository.

4. Results and Discussion

We tested the effectiveness of the proposed pipeline for the voice disorder detection and classification tasks against the compared methods. To evaluate the contribution of each part of the proposed pipeline, we also report an ablation study over each considered dataset. Finally, to assess the quality of the synthetic data, we show the results of the models when trained on synthetic data only and evaluated on real data. All results are reported in terms of average Accuracy, F1 Macro, and AUC (for disorder detection only).

Voice disorder detection. It is a binary task, that aims at separating pathological and healthy patients. In Table 2 we report the performance on the voice disorder detection task of the proposed pipeline (Ours) against the compared models. The advantage of employing the proposed pipeline is significant, with +.20-.36 points in terms of AUC with respect to a 1D-CNN, +.10-.26 with respect to a 2D-CNN and +.05-.13 with respect to a plain transformer model (HuBERT).

Table 2: ***Voice Disorder Detection***. *Mean ±std of 10-fold CV for all datasets. Best results are in bold.*

Ds	Approach	Accuracy	AUC	F1 Macro
SVD	CNN-1D	.746±.041	.705±.041	.722±.041
	CNN-2D	.799±.025	.734±.025	.747±.024
	HuBERT	.862±.040	.844±.041	.842±.038
	Ours	**.909±.006**	**.911±.005**	**.907±.007**
AVFAD	CNN-1D	.712±.028	.719±.029	.711±.029
	CNN-2D	.835±.019	.834±.021	.834±.021
	HuBERT	.872±.015	.877±.015	.871±.014
	Ours	**.927±.004**	**.931±.004**	**.926±.004**
IPV	CNN-1D	.673±.025	.616±.024	.637±.025
	CNN-2D	.788±.021	.721±.021	.737±.021
	HuBERT	.875±.024	.847±.026	.870±.026
	Ours	**.981±.005**	**.983±.006**	**.978±.005**

Voice disorder classification. This multilabel classification task aims at distinguishing among different pathologies causing the disorder (e.g., polyps, nodules, cysts). In Table 3 we report the results for voice disorder classification. The performance gap becomes very significant, particularly when considering the F1 macro, which takes into account class imbalance. Our method reports an increase for the F1 score of +.48-.57 with respect to the CNN-1D, +.38-.52 with respect to the CNN-2D, +.11-.15 with respect to a plain transformer.

We also propose leveraging an initial pre-training on the voice disorder detection task to further enhance the performance on the classification task. We ensured that, in the first training, the model did not encounter the speakers included in the test set for the classification task. We call this approach ours*. As expected, the pre-training boosts the final performance, with a further increase of +.01-.05 over the version without pre-training.

These results show that the employment of synthetic data and data augmentation, together with an ensemble model to address different data types, fully exploits the generalization capability of transformer models. In the following, we will analyze the contribution of the different pipeline components through an ablation study.

Ablation study. We report in Table 4 the AUC performance of our model when adding the different elements of the proposed pipeline. More precisely, we tested the improvement given by adding to a transformer base model trained on the sentence reading data type[6] (base), the data augmentation (data++), and

[3] https://github.com/koudounasalkis/AI4Voice

[4] As macro-classes are not given in SVD, we considered the six most frequent classes.

[5] For a fair comparison, we implemented and trained these models.

[6] Please note that the base model trained on sustained vowels achieves worse performance w.r.t. its counterpart on sentences, thus we only report it in the project repository.

Table 3: ***Voice Disorder Classification.*** *Mean ±std of 10-fold CV for* AVFAD *and* IPV *datasets. Best results are in bold.*

Ds	Model	Accuracy	F1 Macro
SVD	CNN-1D	.437±.025	.280±.024
	CNN-2D	.539±.021	.348±.023
	HuBERT	.771±.022	.712±.020
	Ours	.874±.017	.859±.014
	Ours*	**.888±.015**	**.868±.016**
AVFAD	CNN-1D	.401±.027	.167±.028
	CNN-2D	.509±.025	.266±.024
	HuBERT	.693±.028	.538±.026
	Ours	.782±.024	.648±.023
	Ours*	**.808±.021**	**.703±.020**
IPV	CNN-1D	.419±.022	.278±.024
	CNN-2D	.521±.019	.335±.021
	HuBERT	.764±.025	.710±.023
	Ours	.867±.010	.854±.007
	Ours*	**.883±.007**	**.871±.006**

the synthetic data (TTS). We then tested the employment of both sentence reading and sustained vowels with an ensemble model (MoE). By further differentiating the pre-training of each model on ad-hoc datasets (LibriSpeech for sentence reading, Audioset for sustained vowels), we obtain our model (MoE*).[7]

We can observe from Table 4 that the data augmentation and the synthetic data creation yield significant improvements, albeit with varying degrees across datasets and models, ranging from +0.01 on SVD to +0.10 on IPV. Conversely, employing an ensemble model (with ad-hoc pre-training) consistently enhances the performance by approximately +0.04 across all datasets and models. Employing all data types at the same time for a single model (ALL) results instead in a bad strategy, decreasing the performance of the model below the baseline. These results confirm our intuition that different data types provide complementary information that can improve prediction accuracy. At the same time, their diversity prevents considering them together, and requires separate model training for each data type.

Generalization performance of synthetic data. As previously observed, while altering and creating more variety in the training distribution by means of synthetic data, we still need to preserve the characteristics of the original data. We assessed the quality of the synthetic data by training the models only on the synthetic data and checking their performance on the real data. In this case, our model, while not incorporating additional synthetic data, integrates the HuBERT model with data augmentation and the MoE ensemble technique. In Table 5 we present the results on the IPV dataset. The drop in accuracy, compared to using solely real data, ranges from -.07 to -.14 in detection and -.05 to -.12 in classification for both the CNN and transformer approaches. This small decline underscores the quality of synthetic data in emulating real data characteristics, as model performance remain significantly above random chance and close to the original performance.

[7] A pre-trained version of WavLM on the AudioSet dataset is not publicly available. Thus the MoE* result is not reported in this case.

Table 4: ***Ablation study*** *on voice disorder detection to quantify the contribution of each term in terms of AUC. Best results for each model in bold, best results overall in light-green.*

Ds	Approach	HuBERT	wav2vec 2.0	WavLM
SVD	base	.844±.041	.842±.038	.842±.035
	+ data++	.851±.032	.849±.036	.849±.032
	+ TTS	.871±.051	.855±.032	.859±.039
	+ MoE	.903±.012	.888±.013	**.884±.019**
	+ MoE*	**.911±.005**	**.894±.011**	-
	ALL	.791±.031	.787±.029	.784±.038
AVFAD	base	.877±.015	.875±.018	.872±.016
	+ data++	.889±.019	.879±.017	.881±.014
	+ TTS	.894±.015	.882±.015	.885±.019
	+ MoE	.917±.007	.902±.014	**.908±.007**
	+ MoE*	**.931±.004**	**.921±.009**	-
	ALL	.865±.012	.861±.017	.863±.013
IPV	base	.847±.026	.874±.021	.832±.029
	+ data++	.903±.016	.894±.019	.845±.022
	+ TTS	.948±.021	.933±.021	.929±.020
	+ MoE	.977±.013	.943±.015	**.930±.012**
	+ MoE*	**.983±.006**	**.970±.005**	-
	ALL	.831±.015	.818±.018	.787±.016

Table 5: ***Synthetic Data Only, IPV dataset.*** *Mean ±std of 10-fold CV. Best results in bold. Difference with real data in brackets.*

Task	Model	Accuracy		F1 Macro	
Detection	CNN-1D	.573±.029	(-.100)	.564±.030	(-.073)
	CNN-2D	.659±.031	(-.129)	.592±.033	(-.145)
	HuBERT	.794±.028	(-.081)	.782±.025	(-.088)
	Ours	**.868±.012**	(-.113)	**.854±.011**	(-.124)
Classification	CNN-1D	.308±.034	(-.111)	.221±.036	(-.057)
	CNN-2D	.387±.033	(-.234)	.284±.032	(-.051)
	HuBERT	.651±.027	(-.113)	.622±.026	(-.088)
	Ours	.749±.025	(-.118)	.743±.026	(-.111)
	Ours*	**.755±.021**	(-.128)	**.751±.022**	(-.120)

5. Conclusion

We proposed a novel approach to accurately detect and classify voice disorders. Its experimental evaluation shows that using a strong data synthesis and augmentation strategy, and employing an ensemble of transformer models allows achieving significant performance improvements on both tasks.

Limitations and future work. The main limitations of our approach are related to (i) the model size and (ii) the real-world generalization of the model. For (i), while the MoE boosts the performance over a model trained on a single data type, it doubles the model size. A possible approach to weight sharing could entail employing a single transformer working on both data types at the same time. For (ii), the models have been trained on data recorded under expert guidance. As future work, we would like to test and extend our model by considering real-world scenarios. To this aim, we are embedding our model in a web-based application that will be deployed in a number of selected private and public otolaryngology clinics.

6. Acknowledgements

This work is partially supported by the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

7. 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," *The Laryngoscope*, 2005.

[2] S. M. Cohen, "Self-reported impact of dysphonia in a primary care population: An epidemiological study," *The Laryngoscope*, 2010.

[3] N. Bhattacharyya, "The prevalence of voice problems among adults in the united states," *The Laryngoscope*, 2014.

[4] N. Spantideas, E. Drosou, A. Karatsis, and D. Assimakopoulos, "Voice disorders in the general greek population and in patients with laryngopharyngeal reflux. prevalence and risk factors," *Journal of Voice*, 2015.

[5] E. Brunner, K. Eberhard, and M. Gugatschka, "Prevalence of benign vocal fold lesions: Long-term results from a single european institution," *Journal of Voice*, 2023.

[6] I. Karabayir, S. M. Goldman, S. Pappu, and O. Akbilgic, "Gradient boosting for parkinson's disease diagnosis from voice recordings," *BMC Medical Informatics and Decision Making*, 2020.

[7] H. Vieira, N. Costa, T. Sousa, S. Reis, and L. Coelho, "Voice-based classification of amyotrophic lateral sclerosis: where are we and where are we going? a systematic review," *Neurodegenerative Diseases*, 2020.

[8] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "Ai in health and medicine," *Nature medicine*, vol. 28, no. 1, pp. 31–38, 2022.

[9] J.-R. Rueda, I. Sola, A. Pascual, and M. S. Casacuberta, "Non-invasive interventions for improving well-being and quality of life in patients with lung cancer," *Cochrane Database of Systematic Reviews*, no. 9, 2011.

[10] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, 2017.

[11] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[12] Z. I. Attia, S. Kapa, F. Lopez-Jimenez, P. M. McKie, D. J. Ladewig, G. Satam, P. A. Pellikka, M. Enriquez-Sarano, P. A. Noseworthy, T. M. Munger *et al.*, "Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram," *Nature medicine*, vol. 25, no. 1, pp. 70–74, 2019.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[14] B. Woldert-Jokisz, "Saarbruecken voice database," 2007.

[15] 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*. IntechOpen, 2017.

[16] L. Salhi, M. Talbi, and A. Cherif, "Voice disorders identification using hybrid approach: Wavelet analysis and multilayer neural networks," *International Journal of Electrical and Computer Engineering*, vol. 2, no. 9, pp. 3003–3012, 2008.

[17] J. D. Arias-Londoño, J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "Byovoz automatic voice condition analysis system for the 2018 femh challenge," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018.

[18] X. Peng, H. Xu, J. Liu, J. Wang, and C. He, "Voice disorder classification using convolutional neural network based on deep transfer learning," *Scientific Reports*, vol. 13, no. 1, p. 7264, 2023.

[19] X. Xie, H. Cai, C. Li, Y. Wu, and F. Ding, "A voice disease detection method based on mfccs and shallow cnn," *Journal of Voice*, 2023.

[20] U. K. Lilhore, S. Dalal, N. Faujdar, M. Margala, P. Chakrabarti, T. Chakrabarti, S. Simaiya, P. Kumar, P. Thangaraju, and H. Velmurugan, "Hybrid cnn-lstm model with efficient hyperparameter tuning for prediction of parkinson's disease," *Scientific Reports*, vol. 13, no. 1, p. 14605, 2023.

[21] R. Islam, E. Abdel-Raheem, and M. Tarique, "Voice pathology detection using convolutional neural networks with electroglottographic (egg) and speech signals," *Computer Methods and Programs in Biomedicine Update*, vol. 2, p. 100074, 2022.

[22] D. Ribas, M. A. Pastor, A. Miguel, D. Martínez, A. Ortega, and E. Lleida, "Automatic voice disorder detection using self-supervised representations," *IEEE Access*, 2023.

[23] S. R. Shahamiri, V. Lal, and D. Shah, "Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

[24] A. Almadhor, R. Irfan, J. Gao, N. Saleem, H. T. Rauf, and S. Kadry, "E2e-dasr: End-to-end deep learning-based dysarthric automatic speech recognition," *Expert Systems with Applications*, vol. 222, p. 119797, 2023.

[25] L. Ilias, D. Askounis, and J. Psarras, "Detecting dementia from speech and transcripts using transformers," *Computer Speech & Language*, vol. 79, p. 101485, 2023.

[26] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, and J. Weber, "Xtts: a massively multilingual zero-shot text-to-speech model," 2024.

[27] J. F. Gemmeke and et al., "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*, 2015.

[29] G. B. Kempster, B. R. Gerratt, K. V. Abbott, J. Barkmeier-Kraemer, and R. E. Hillman, "Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol," 2009.

[30] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

[31] C. Sanyuan and et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Top. Signal Process.*, 2022.

[32] W.-N. Hsu and et al., "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[33] M. La Quatra, A. Koudounas, L. Vaiani, E. Baralis, P. Garza, S. Cagliero, Luca, and S. Marco, "Benchmarking representations for speech, music, and acoustic events," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024.