Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

MVP: Multi-source Voice Pathology detection

Alkis Koudounas[1], Moreno La Quatra*[2], Gabriele Ciravegna[1], Marco Fantini[3,4], Erika Crosetti[4], Giovanni Succo[4,5], Tania Cerquitelli[1], Sabato Marco Siniscalchi[6], Elena Baralis[1]*

[1]Politecnico di Torino, Italy [2]Kore University of Enna, Italy [3]San Feliciano Hospital, Italy
[4]Ospedale San Giovanni Bosco, Italy [5]Università degli Studi di Torino, Italy
[6]Università degli Studi di Palermo, Italy

alkis.koudounas@polito.it moreno.laquatra@unikore.it

Abstract

Voice disorders significantly impact patient quality of life, yet non-invasive automated diagnosis remains under-explored due to both the scarcity of pathological voice data, and the variability in recording sources. This work introduces MVP (Multi-source Voice Pathology detection), a novel approach that leverages transformers operating directly on raw voice signals. We explore three fusion strategies to combine sentence reading and sustained vowel recordings: waveform concatenation, intermediate feature fusion, and decision-level combination. Empirical validation across the German, Portuguese, and Italian languages shows that intermediate feature fusion using transformers best captures the complementary characteristics of both recording types. Our approach achieves up to +13% AUC improvement over single-source methods.

Index Terms: voice pathology detection, multi-source analysis, fusion strategies, transformers

1. Introduction

Voice disorders affect approximately 30% of the general population during their lifetime [1–3]. These conditions impact daily communication, work performance, and overall quality of life. Disorders range from functional issues such as muscle tension dysphonia to organic pathologies such as vocal cord nodules [4]. Early detection is crucial, as untreated disorders often progress to chronic conditions and cause psychological distress [5]. Current clinical diagnosis relies on specialized equipment and clinicians. This approach is costly, invasive, and limited by specialist availability. There is thus a clear need for accessible, non-invasive screening methods for early detection.

Automated voice pathology detection offers a promising solution through machine learning analysis of voice recordings [6]. Previous work has explored various approaches, including multilayer perceptrons trained on hand-crafted features such as Mel-frequency cepstral coefficients (MFCCs) [7, 8], as well as convolutional neural networks (CNNs) applied to spectrograms [9, 10]. Hybrid models combining CNNs and recurrent neural networks (RNNs) have also been investigated to better capture temporal dependencies in voice signals [11]. More recently, researchers have explored architectures that operate directly on raw audio, such as 1D-CNNs [12] and transformers [13], with the latter showing promising results in both voice pathology and related speech disorders like dysarthria [14–18].

Current methods analyze either sustained vowels or continuous speech in isolation [13,15,17]. Sustained vowels offer stable conditions for voice analysis but miss the dynamics of natural speech. Conversely, continuous speech captures everyday voice use and adds complexity through linguistic content and prosody. Voice pathologies show different patterns across these speaking tasks, limiting the effectiveness of single-source analysis. For instance, vocal nodules may affect sustained phonation more noticeably, while muscle tension disorders might be more evident during continuous speech [19, 20].

To address this limitation, we propose MVP (**M**ulti-source **V**oice **P**athology detection), a framework that leverages both sustained vowels and sentence readings through specialized transformer models [21–23], and allows a comprehensive assessment of vocal health. We specifically employ models pre-trained on LibriSpeech [24] to analyze sentences and models pre-trained on AudioSet [25] to process sustained vowels. Our method builds on recent advances in transformer-based voice pathology detection [6], but while previous work exploited ensemble model separately trained on different sources, our approach explicitly explores multi-source integration at multiple levels. We investigate three types of fusion strategy: waveform concatenation, intermediate representation fusion, and decision-level combination. Each strategy addresses different aspects of multi-source integration, from raw signal combination to high-level feature fusion. By employing specialized transformer architectures for each recording type, MVP captures their unique acoustic characteristics while maintaining crucial temporal relationships for pathology detection. The contributions of this work are as follows.

- *Multi-source framework.* MVP combines sustained vowels and sentence readings through specialized architectures.
- *Different fusion strategies.* We compare different strategies and show the advantages of intermediate representation fusion in this context.
- *Extensive evaluation.* Experiments on three datasets demonstrate robust performance across different languages and recording conditions, with up to 13% AUC improvement over state-of-the-art single-source methods.

2. MVP Framework

Our MVP framework addresses voice pathology detection by combining multiple recording sources. It adopts a three-stage architecture: (i) source-specific backbone models for feature extraction, (ii) fusion strategies to combine information from different sources, and (iii) a decision-making module for pathology detection. Fig. 1 provides an overview of the method.

2.1. Backbone Models

We denote a sustained vowel recording as X_{SV} and a sentence recording as X_{S}. For waveform concatenation, a single backbone processes the combined input. For intermediate fusion and

* Both authors contributed equally to this work.

 10.21437/Interspeech.2025-1868

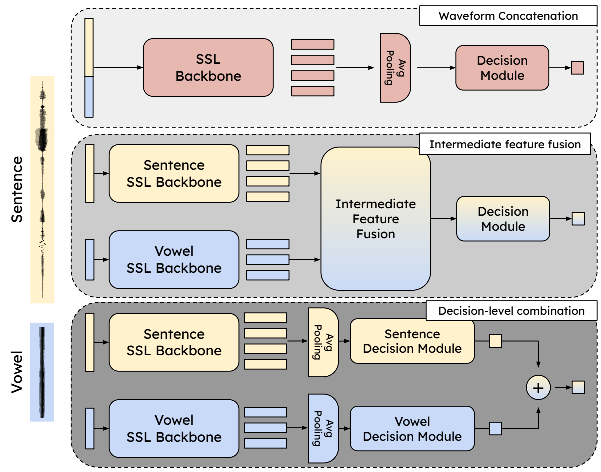

Figure 1: ***Proposed MVP framework**: waveform concatenation (top panel), intermediate feature fusion (mid panel), and decision-level combination (bottom panel).*

decision-level combination, each input is processed by specialized backbones. For sentence recordings, we use HuBERT pre-trained on LibriSpeech [24], which captures linguistic content and prosodic variations. For sustained vowel recordings, we use HuBERT pre-trained on AudioSet [25], focusing on vocal quality and non-semantic features. The frame-level representations extracted by each backbone are:

$$H_{\text{SV}} = \text{HuBERT}_{\text{AS}}(X_{\text{SV}}), \quad H_{\text{S}} = \text{HuBERT}_{\text{LS}}(X_{\text{S}}) \tag{1}$$

where $H_{\text{SV}} \in \mathbb{R}^{T_{\text{SV}} \times d}$ and $H_{\text{S}} \in \mathbb{R}^{T_{\text{S}} \times d}$ contain temporal information from each recording type, T_{SV}, T_{S} represents the source lengths and d is the latent dimensionality of the representations.

2.2. Fusion Strategies

We explore three strategies for combining information from sustained vowel and sentence recordings, each operating at a different level of abstraction.

- *Waveform Concatenation (WC).* Raw audio signals are concatenated and processed by a single backbone, preserving all information but potentially introducing artifacts due to the different characteristics of each source (Fig. 1 (top)).
- *Intermediate Feature Fusion (IFF).* Features from specialized backbones are combined through fusion methods (Section 2.2.1). This preserves source-specific characteristics while enabling cross-source learning (Fig. 1 (middle)).
- *Decision-Level Combination (DLC).* This approach mimics the pipeline proposed in [6], where an ensemble dynamically selects between source-specific predictions by averaging individual backbones probabilities (Fig. 1 (bottom)).

2.2.1. Intermediate Feature Fusion

Given two feature sequences $H_{\text{SV}} \in \mathbb{R}^{T_{\text{SV}} \times d}$ and $H_{\text{S}} \in \mathbb{R}^{T_{\text{S}} \times d}$ from sustained vowels and sentence recordings, respectively, we explore five methods to fuse them into a single vector $Z_{\text{fused}} \in \mathbb{R}^d$. Each method represents a different approach to capturing cross-source interactions [26].

Simple Concatenation (Baseline). We first apply average pooling to each sequence independently to obtain global representations $Z_{\text{SV}}^{\text{Avg}}$ and $Z_{\text{S}}^{\text{Avg}}$ for each source. The fused representation is their concatenation:

$$Z_{\text{Concat}} = [Z_{\text{SV}}^{\text{Avg}}; Z_{\text{S}}^{\text{Avg}}] \in \mathbb{R}^{2d} \tag{2}$$

This approach preserves the global characteristics of each source while providing a strong baseline for comparison.

Attention Pooling (AP). We concatenate sequences on the time dimension and apply a learned attention mechanism to capture the relative importance of different time steps. The attention scores α_t are computed using a learnable vector $w \in \mathbb{R}^d$:

$$\alpha_t = \frac{\exp(w^\top H_t)}{\sum_{t'} \exp(w^\top H_{t'})}, \quad Z_{\text{AP}} = \sum_t \alpha_t H_t \in \mathbb{R}^d \tag{3}$$

where H_t represents features at time step t from the concatenated sequence $[H_{\text{SV}}; H_{\text{S}}]$. This method emphasizes the most informative temporal features across both sources.

Gating Mechanism. This method introduces adaptive weighting between sources through a learned gating mechanism [27]. We first obtain source-specific vectors using AP, then compute an adaptive gating vector:

$$G = \sigma(W[Z_{\text{SV}}^{\text{AP}}; Z_{\text{S}}^{\text{AP}}]), \quad Z_{\text{Gating}} = G \odot [Z_{\text{SV}}^{\text{AP}}; Z_{\text{S}}^{\text{AP}}] \tag{4}$$

where W is a learnable matrix and σ is the sigmoid function. The gating allows the model to dynamically adjust the contribution of each source.

Feature-wise Linear Modulation (FiLM). FiLM enables bidirectional cross-source interactions by allowing each source to modulate the other [28]:

$$Z_{\text{S}}^{\text{FiLM}} = Z_{\text{S}}^{\text{AP}} \odot (1 + W_\gamma^s Z_{\text{SV}}^{\text{AP}}) + W_\beta^s Z_{\text{SV}}^{\text{AP}} \tag{5}$$

$$Z_{\text{SV}}^{\text{FiLM}} = Z_{\text{SV}}^{\text{AP}} \odot (1 + W_\gamma^v Z_{\text{S}}^{\text{AP}}) + W_\beta^v Z_{\text{S}}^{\text{AP}} \tag{6}$$

where $W_\gamma^s, W_\gamma^v, W_\beta^s, W_\beta^v$ are learnable parameters for scale and shift operations. The final representation combines both modulated features:

$$Z_{\text{FiLM}} = [Z_{\text{S}}^{\text{FiLM}}; Z_{\text{SV}}^{\text{FiLM}}] \tag{7}$$

This bidirectional modulation allows each source to influence the other through learned transformations.

Transformer Encoder (TE). We first concatenate the sequences along the time axis to form a combined sequence:

$$H_{\text{combined}} = [H_{\text{SV}}; H_{\text{S}}] \in \mathbb{R}^{(T_{\text{SV}}+T_{\text{S}}) \times d} \tag{8}$$

This sequence is processed through L transformer encoder layers, where self-attention mechanisms enable each time step from one source to interact with all the time steps from both sources. This fine-grained interaction preserves and leverages the temporal structure of both sources. The final representation is obtained through attention pooling:

$$Z_{\text{TE}} = \text{AP}(\text{TE}_L(H_{\text{combined}})) \in \mathbb{R}^d \tag{9}$$

2.3. Decision-Making Module

The output from any of the fusion strategies results in a final fused representation Z. This representation is passed to the decision-making module, which consists of a fully connected layer followed by a sigmoid activation function:

$$\hat{y} = \text{Sigmoid}(\text{FC}(Z)) \tag{10}$$

The output $\hat{y} \in [0, 1]$ represents the probability of the positive class (presence of voice pathology).

Table 1: ***Mean±std performance of different fusion approaches across three datasets**. Best results in **bold**, second-best underlined. Models use either LibriSpeech (`LS`) or AudioSet (`AS`) pre-training. →`Sent` indicates fine-tuning on read sentences, →`Vowel` on sustained vowels, →`Mix` on both sources. ❄ indicates frozen backbones. IFF fusion is implemented with TE.*

Model	# Params	SVD			AVFAD			IPV		
		Acc.	**F1**	**AUC**	**Acc.**	**F1**	**AUC**	**Acc.**	**F1**	**AUC**
Single-Source Baselines										
`LS→Sent`	94.64M	.873±.058	.849±.062	.850±.048	.872±.015	.871±.014	.877±.015	.875±.024	.870±.026	.847±.026
`LS→Vowel`	94.64M	.747±.075	.724±.074	.732±.084	.714±.051	.705±.061	.714±.061	.622±.064	.617±.062	.620±.062
`AS→Sent`	94.64M	.817±.060	.801±.061	.810±.052	.852±.045	.850±.047	.855±.050	.828±.039	.823±.038	.815±.044
`AS→Vowel`	94.64M	.779±.075	.747±.068	.760±.079	.798±.043	.758±.056	.756±.058	.676±.059	.649±.060	.665±.055
`LS→Mix`	94.64M	.780±.028	.791±.031	.765±.025	.827±.017	.826±.019	.827±.019	.840±.018	.831±.015	.831±.016
Waveform Concatenation (WC)										
`LS`	94.64M	.896±.054	.882±.053	.891±.056	.907±.054	.906±.055	.908±.052	.888±.066	.881±.067	.886±.060
`AS`	94.64M	.875±.063	.869±.061	.873±.068	.889±.042	.884±.045	.885±.049	.836±.031	.832±.032	.831±.037
Intermediate Feature Fusion (IFF)										
`LS+AS` ❄	17.98M	.826±.036	.809±.035	.832±.023	.833±.032	.831±.032	.834±.033	.813±.048	.806±.045	.809±.048
`LS+AS`	206.73M	**.958±.063**	**.953±.067**	**.958±.062**	**.962±.040**	**.962±.039**	**.963±.038**	**.939±.044**	**.931±.054**	**.936±.053**
Decision-Level Combination (DLC)										
`LS+LS`	189.28M	.885±.062	.863±.070	.860±.070	.881±.095	.874±.111	.879±.099	.882±.083	.873±.084	.862±.064
`AS+AS`	189.28M	.864±.072	.855±.076	.857±.075	.872±.034	.863±.034	.866±.035	.837±.061	.826±.095	.827±.102
`LS+AS`	189.28M	.898±.051	.884±.058	.896±.058	.888±.092	.887±.108	.889±.096	.896±.072	.877±.074	.882±.062

3. Experimental Setup

To ensure reproducibility and fair comparison, in the following, we detail our experimental setup across datasets, training, data processing, and evaluation procedures[1].

3.1. Datasets

We evaluate our approach on three datasets that contain both sentence readings and sustained vowel emissions recorded under controlled conditions.

SVD. The Saarbruecken Voice Database contains German voice recordings from 2,043 subjects (687 healthy and 1,356 pathological). Each recording session includes sustained vowels at different pitch levels and a sentence-reading task. For consistency with other datasets, we only use normal pitch vowels and sentence readings.

AVFAD. The Advanced Voice Function Assessment Database includes Portuguese recordings from 709 subjects (346 healthy and 363 pathological). Each subject performs multiple tasks, including sustained vowels `/a/`, `/e/`, `/o/`, and reading six phonetically balanced sentences. We randomly select one of the six sentences and one of the vowels for each subject.

IPV. The Italian Pathological Voice is a dataset including recordings from 513 subjects (173 healthy and 340 pathological). Voice samples were recorded under standardized conditions with a 30 cm mouth-to-microphone distance and ambient noise below 30 dB. Each subject recorded a sustained vowel `/a/` and five phonetically balanced sentences. We randomly select one sentence and the vowel for each speaker.

3.2. Implementation Details

Training Protocol and Data Processing. We perform 10-fold cross-validation across all experiments, ensuring speaker-independent splits. Each fold maintains the same healthy-to-pathological ratio as the original dataset. All audio files are resampled to 16 kHz and normalized to zero mean and unit variance. For consistent processing, recordings are padded or truncated to a fixed length of 5.0 seconds. We use an AdamW optimizer with 5e-5 learning rate and 0.01 weight decay. Training runs for 10 epochs with early stopping (patience=5) on validation loss. We use binary cross-entropy loss and batch size 64. Training was performed on a single NVIDIA A100 80GB GPU. For the TE fusion strategy, we set $L = 2$ transformer encoder layers, which provides an effective balance between model complexity and representational power.

Data Augmentation. We implement separate augmentation strategies for sentences and sustained vowels to preserve their distinct characteristics. For sentence readings, we apply augmentation with 25% probability, including noise addition (SNR between 0-30dB), speed perturbation (0.75x to 1.25x), pitch shifting (±4 semitones), and their combinations. For sustained vowels, we apply augmentation with a lower probability of 10% to preserve their core vocal characteristics. This empirically-determined approach aims to balance data diversity with signal integrity, which is particularly important for sustained vowels where stable phonation is essential for pathology detection.

Single-Source Baselines. We implement single-source baselines using HuBERT models pre-trained on either LibriSpeech (LS) or AudioSet (AS) as they both were proven effective in previous work [6]. Each baseline uses the same architecture as our multi-source models but processes only one recording type, enabling fair comparison of the multi-source advantage. An additional baseline involves a single HuBERT (LS) model trained on a mix of sentences and vowels.

Model Configuration and Evaluation. Our transformer backbone uses HuBERT in its base configuration with 94.64M parameters. For single-source baselines, we evaluate both pretrained models on each type of recording and the HuBERT (LS) model trained on the mix of sources. In the IFF strategy, we experiment with both frozen backbones (❄ – 17.98M trainable parameters) and fine-tuned backbones (206.73M parameters). We apply the fusion strategy extracting representation from the 5th layer of the SSL backbones – see Section 4.2 for detailed ablation studies. For DLC, we train two LS- or AS- backbone

[1] `https://github.com/koudounasalkis/MVP`

Table 2: ***A comparison of IFF fusion strategies,*** *AUC scores. Best results in **bold** second-best underlined.*

Method	SVD	AVFAD	IPV
`Concat`	.918±.060	.920±.044	.915±.050
`AP`	.948±.060	.955±.040	.929±.047
`TE`	**.958±.062**	**.963±.038**	**.936±.053**
`Gating`	.947±.068	.956±.045	.926±.049
`FiLM`	.951±.062	.961±.041	.934±.053

Table 3: ***Ablation study on backbone models,*** *AUC scores.* `H`*=HuBERT,* `v2v`*=voc2vec,* `LS`*=LibriSpeech,* `AS`*=AudioSet. Best results in **bold** second-best underlined.*

Sentence	Vowel	SVD	AVFAD	IPV
`H-LS`	`H-LS`	.942±.059	.951±.042	.917±.067
`H-AS`	`H-AS`	.935±.072	.939±.042	.901±.048
`H-LS`	`H-AS`	**.958±.062**	**.963±.038**	**.936±.053**
`H-LS`	`v2v`	.953±.056	.958±.044	.929±.044

models separately on sentences and vowels or a combined approach using both specialized backbones (LS+AS). In all cases, the number of trainable parameters is 189.28M. We report accuracy, macro F1 score, and AUC-ROC averaged across folds with standard deviations to evaluate model performance.

4. Experimental Analysis

Table 1 demonstrates that our multi-source approach significantly outperforms single-source baselines across all datasets. The IFF-TE method with fine-tuned backbones achieves the highest AUC scores: 95.8% (SVD), 96.3% (AVFAD), and 93.6% (IPV). This represents a 10-13% improvement over the best single-source baseline, showing the clear advantage of the multi-source approach. We separately investigate the impact of IFF fusion strategies in Section 4.1.

When focusing on the single-source baselines, the results confirm previous findings [6]: models consistently perform better on sentence readings compared to sustained vowels. Sentence readings may contain richer diagnostic information, possibly because they capture both phonation quality and dynamic speech characteristics. However, training a single model on a mixture of sources proves ineffective, This is likely because the model cannot adapt to the diversity of recording types, leading to weak overall performance. In contrast, the significant performance improvement observed with our multi-source approach highlights the complementary value of sustained vowels. IFF also outperforms other fusion strategies such as WC and DLC. Interestingly, WC performs better than DLC in two out of three datasets while requiring only a single model, actually halving the number of parameters. Even in simpler settings, concurrent access to both sources allows WC to learn cross-source patterns, supporting the value of joint analysis. DLC exhibits high standard deviations, likely due to the limited amount of data, especially when performing 10-fold cross-validation. An important finding also emerges from the resource-constrained version of IFF. Although the best results are achieved by fine-tuning both backbones (206.73M parameters), the frozen variant (❄) still outperforms four of five single-source baselines with less than 20% of their trainable parameters. This highlights the benefits of multi-source analysis even in resource-constrained scenarios where fine-tuning large models may not be feasible.

Table 4: ***Ablation study on feature extraction layer depth***, *AUC scores. Best results in **bold** second-best underlined.*

Layer	SVD	AVFAD	IPV
`4th`	.944±.054	.945±.043	.924±.048
`5th`	**.958±.062**	**.963±.038**	**.936±.053**
`6th`	.950±.055	.943±.046	.923±.060
`7th`	.948±.066	.954±.042	.929±.061
`Last`	.954±.061	.958±.037	.932±.072
`Weighted`	.953±.059	.957±.042	.932±.057

4.1. Fusion Strategies

In Table 2 we analyze the impact of different IFF strategies for cross-source learning. The Transformer Encoder (TE) achieves the best overall performance, particularly on SVD and AVFAD, likely due to its ability to correlate temporal relationships between sources. While FiLM and AP provide strong alternatives, the significant performance gap between learned fusion strategies and simple concatenation (up to 4.7% on AVFAD) highlights the importance of modeling fine-grained interactions between sources. Effective voice pathology detection requires careful modeling of how different vocal characteristics manifest across both sustained and dynamic speech patterns. The results suggest these characteristics in isolation are insufficient; their correlations provide crucial diagnostic information.

4.2. Ablation Studies

To validate our choices, we conduct specific ablation studies examining: (i) the impact of model architecture and pre-training data sources, (ii) the optimal layer for feature extraction.

Impact of model architectures and pre-training sources. Table 3 reveals the role of the model architecture and specialized pre-training for each source. The combination of HuBERT models pre-trained on LibriSpeech (LS) for sentences and on AudioSet (AS) for sustained vowels consistently outperforms other configurations across all datasets. This aligns with recent works [6] showing that sentence analysis benefits from pre-training on speech data (LS), while sustained vowel analysis benefits from exposure to diverse acoustic events (AS). To process sustained vowels, we also evaluate voc2vec [23] which mimics the wav2vec 2.0 [29] architecture and it is pre-trained on non-speech vocalizations. It shows competitive performance without surpassing the HuBERT LS-AS combination. Specialized pre-training can thus improve performance, but the acoustic event coverage in AudioSet provides more effective representations to capture pathological voice characteristics.

Optimal Representations. The results in Table 4 provide insights into the ideal point of feature extraction. Although performance remains relatively stable across different layers, feature extraction at the 5th layer consistently gives optimal results. Mid-level representations provide the best balance between preserving source-specific characteristics and enabling effective cross-source integration [30]. Learned weighted sum across all layers shows competitive but not superior performance, proving that it may need more data for optimal weighting.

5. Conclusions

This paper presented MVP, a novel multi-source approach for voice pathology detection that effectively combines sustained vowels and sentence readings. Our experimental results across three languages demonstrate that intermediate feature fusion with transformers consistently outperforms single-source methods, achieving up to 13% AUC improvement.

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), is partially supported by the "D.A.R.E. – Digital Lifelong Prevention" project (code: PNC0000002, CUP: B53C22006450001), co-funded by the Italian Complementary National Plan PNC-I.1 Research initiatives for innovative technologies and pathways in the health and welfare sector (D.D. 931 of 06/06/2022), and partially supported by the European Union – Next Generation EU under the National Recovery and Resilience Plan (PNRR) – M4 C2, Investment 1.1: Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN) - PRIN 2022 - "SHAPE-AD" (CUP: J53D23007240008). 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. Thibeault, R. A. Parsa, S. D. Gray, and E. M. Smith, "Prevalence of voice disorders in teachers and the general population," *Journal of Speech, Language, and Hearing Research*, 2004.

[2] 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.

[3] M. Fantini, G. Ciravegna, A. Koudounas, T. Cerquitelli, E. Baralis, G. Succo, and E. Crosetti, "The rapidly evolving scenario of acoustic voice analysis in otolaryngology," *Cureus*, vol. 16, no. 11, 2024.

[4] S. Akbulut, A. am Zehnhoff-Dinnesen, F. de Jong, M. Echternach, U. Eysholdt, M. Fuchs, T. Hacki, K. Izdebski, A. Keilmann, P. Kummer *et al.*, "Basics of voice disorders," in *Phoniatrics I: Fundamentals–Voice Disorders–Disorders of Language and Hearing Development*. Springer, 2019, pp. 193–238.

[5] S. M. Cohen, J. Kim, N. Roy, and M. Courey, "Delayed otolaryngology referral for voice disorders increases health care costs," *The American Journal of Medicine*, 2015.

[6] A. Koudounas, G. Ciravegna, M. Fantini, E. Crosetti, G. Succo, T. Cerquitelli, and E. Baralis, "Voice disorder analysis: a transformer-based approach," in *Interspeech 2024*, 2024, pp. 3040–3044.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] U. K. Lilhore *et al.*, "Hybrid cnn-lstm model with efficient hyperparameter tuning for prediction of parkinson's disease," *Scientific Reports*, 2023.

[12] 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*, 2022.

[13] 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.

[14] 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.

[15] 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.

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

[17] M. L. Quatra, M. F. Turco, T. Svendsen, G. Salvi, J. R. Orozco-Arroyave, and S. M. Siniscalchi, "Exploiting foundation models and speech enhancement for parkinson's disease detection from speech in real-world operative conditions," *Interspeech*, 2024.

[18] M. La Quatra, J. R. Orozco-Arroyave, and M. S. Siniscalchi, "Bilingual dual-head deep model for parkinson's disease detection from speech," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[19] S.-H. Lee, J.-F. Yu, T.-J. Fang, and G.-S. Lee, "Vocal fold nodules: A disorder of phonation organs or auditory feedback?" *Clinical Otolaryngology*, 2019.

[20] G. Schlotthauer, M. E. Torres, and M. C. Jackson-Menaldi, "A pattern recognition approach to spasmodic dysphonia and muscle tension dysphonia automatic classification," *Journal of voice*, 2010.

[21] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, 2021.

[22] M. La Quatra, A. Koudounas, L. Vaiani, E. Baralis, L. Cagliero, P. Garza, and S. M. Siniscalchi, "Benchmarking representations for speech, music, and acoustic events," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 505–509.

[23] A. Koudounas, M. La Quatra, S. M. Siniscalchi, and E. Baralis, "voc2vec: A foundation model for non-verbal vocalization," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

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

[25] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*. IEEE, 2017.

[26] M. La Quatra, N. D. Cilia, V. Conti, S. Sorce, G. Garraffa, and V. M. Salerno, "Vision-language multimodal fusion in dermatological disease classification," in *Pattern Recognition. ICPR 2024 International Workshops and Challenges*. Cham: Springer Nature Switzerland, 2025, pp. 211–225.

[27] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017.

[28] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *AAAI*, 2018.

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

[30] T. Nguyen, C. Fredouille, A. Ghio, M. Balaguer, and V. Woisard, "Exploring asr-based wav2vec2 for automated speech disorder assessment: Insights and analysis," in *SLT*. IEEE, 2024.