Multimodal Fusion Techniques to Enhance Voice Disorder Diagnoses

Multimodal Fusion Techniques to Enhance Voice Disorder Diagnoses

Qingqing Liu[1,*], Gabriele Ciravegna[1,*], Alkis Koudounas[1], Tania Cerquitelli[1] and Elena Baralis[1]

[1]Politecnico di Torino, Turin, Italy

Abstract

Voice disorders constitute a significant health concern, with an annual prevalence of approximately 7% among the adult population, adversely affecting patients' quality of life, encompassing both social and occupational functioning. Also, the majority of diagnostic methodologies continue to depend on invasive techniques, whereas non-invasive automated diagnostic approaches have not been extensively investigated yet. This study introduces a transformer-based method for detecting voice disorders aimed at enhancing detection efficacy through a multimodal fusion strategy. Specifically addressing two distinct types of voice recordings – extracted from sentences reading and vowels emissions -— we devised and assessed five multimodal fusion strategies across three stages: early, mid, and late. Our experimental findings indicate that the cross-attention mid-fusion method harnesses the benefits of both data types, and it achieves a detection accuracy of 0.885 and a macro F1 score of 0.843 on an internal dataset. These results represent an improvement of +.03 to +.06 in accuracy and +.02 to +.05 in macro F1 score when compared to unimodal models (trained on sentence or vowel data only). This study represents an advancement for an effective non-invasive detection of voice disorders and provides insights for clinical practice.

Keywords

Medical AI, Pathological voice disorder, Transformers, Modality Fusion, Multimodal learning,

1. Introduction

Voice disorders have a significant impact on people's lives, with 7% of adults suffering from them each year. They can lead to communication difficulties, reduced work productivity (7.4 work days lost per year on average), and even career changes, with 4% of patients reporting a career change due to voice problems [1, 2]. There are many types of voice disorders, including but not limited to murmurs, vocal cord dysfunction, and other voice problems caused by neurological diseases, and their early and accurate diagnosis is crucial for effective treatment [3].

Although traditional diagnostic techniques such as laryngoscopy and speech assessment are widely used clinically, they have significant limitations [4]. First, these diagnostic methods are very invasive and may cause discomfort to the patient, thus affecting the experience particularly for patients requiring several investigations and recurrent controls (e.g. cancer patients) [5]. Secondly, these technologies often rely on expensive equipment and highly specialized operators, which limits their accessibility in resource-poor settings. Thirdly, traditional methods rely on doctors' subjective judgments and suffer from subjective bias in evaluation results. Finally, these methods are mostly used for diagnosis when symptoms are evident rather than as proactive preventive screening tools, limiting their role in the early detection of voice disorders [6].

The development of artificial intelligence technology [7], especially the application of deep learning in the field of audio and sound processing, provides new possibilities for overcoming the above challenges [8]. By enabling automated, non-invasive, efficient diagnostics, deep learning methods can lower diagnostic costs and reduce the need for professionals, making detection more accessible and accurate. In addition, these technologies can be integrated into portable devices or mobile applications for active screening and real-time monitoring, providing a new solution for early detection and intervention of voice disorders.

Very recently, Transformer-based models [9, 10, 11] have been shown to be effective tools for the automatic detection of voice disorders. Their core advantage is the ability to capture long-term dependencies in time series data, which is crucial for analyzing complex speech patterns. Through the self-attention mechanism, transformers can not only efficiently process large-scale datasets, but also extract complex patterns that determine voice characteristics. However, this area still remains under-researched with several research questions that remain open due to the complexity and diversity of pathological voice features, which still remains an open issue. Indeed, doctors perform different patient voice assessments to assess different voice properties, such as requiring the patient to read pre-defined sentences and emitting sustained vowels.

This study addresses this challenge by proposing a multimodal approach to voice disorder detection. We leverage the strengths of the transformer architecture to analyze multimodal pathological speech data. Specifically, dealing with two different types of data, namely sentences and vowels only, we design a unified model to process them together. Three fusion strategies – early fusion, mid-level fusion, and late fusion – are investigated to effectively integrate cross-modal information.

We empirically demonstrate that mid-level fusion techniques are particularly suited for this task, demonstrating their ability to capture complementary features and improve detection performance. The cross-attention technique, in particular, achieves performance gains of +.03-.06 in accuracy and +.04-.05 in macro F1 compared to single-modality models, highlighting the potential of multimodal integration in enhancing detection performance. These findings highlight the feasibility of multi-modal transformer-based models in clinical applications and lay a solid foundation for further advancement of automatic voice disorder detection.

The rest of the paper is organized as follows. In Section 2 we first review the relevant research on voice disorder detection and analyze the main challenges of existing methods in application. Section 3 describes the proposed Transformer-based method in detail, focusing on different multimodal

Published in the Proceedings of the Workshops of the EDBT/ICDT 2025 Joint Conference (March 25-28, 2025), Barcelona, Spain, as part of the DARLI-AP workshop held in conjunction with the EDBT/ICDT 2025 conference.

*Corresponding author.

s315203@studenti.polito.it (Q. Liu); gabriele.ciravegna@polito.it (G. Ciravegna); alkis.koudounas@polito.it (A. Koudounas)

fusion strategies. In Section 4 we introduce the experimental setup and evaluation methods, while in Section 5 we provide a comprehensive analysis and interpretation of the experimental results. Finally, in Section 6 we summarize the significance of our research, explore potential limitations, and provide suggestions for future research directions.

2. Related work

This section reviews the relevant research on voice disorder detection and provides a theoretical basis for the tools and methods used in subsequent sections. The discussion focuses on the evolution of voice feature analysis techniques, the application of classifiers in detecting voice disorders, and the latest progress in data augmentation and fusion models.

2.1. Automatic Voice Disorder Detection Methods

Traditional voice disorder detection methods rely on artificial feature engineering, that is, extracting acoustic features such as Mel-frequency cepstral coefficients (MFCC), pitch jitter, and amplitude shimmer from speech signals [12, 13]. These features, rooted in digital signal processing and speech science, have long been the cornerstone of voice analysis. Using these manual features, researchers rely on shallow learning models such as support vector machines (SVMs) and multi-layer perceptrons (MLPs), which perform well in voice disorder detection problems in relatively simple or well-controlled environments [14, 15, 16]. However, the complexity of pathological voice features and the diversity of real-world scenarios have revealed the limitations of these traditional methods, particularly in terms of adaptability and generalization [17]

The advent of deep learning has transformed voice disorder detection, as it can automatically extract features from raw speech signals. Unlike traditional methods that rely on handcrafted features, deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) can learn more abstract and comprehensive feature representations directly from data [12, 16, 18, 19, 20]. CNNs excel at capturing local patterns, while RNNs excel at modeling temporally related patterns, making them more suitable for voice pathology analysis, particularly when employed together.

Recently, transformer-based architectures have made breakthroughs in automatic speech recognition and related tasks [21, 22, 23, 24, 25, 26]. These models use self-attention mechanisms to capture short and long-range dependencies at the same time, thus performing well in processing complex speech patterns [8, 11, 27]. Among them, Wav2Vec2's end-to-end modeling capability [27] combines a convolutional encoder for extracting potential speech representations, a transformer-based context network for capturing long-distance dependencies, and a quantization module for self-supervised learning, further simplifying the feature extraction process. This architecture enables the efficient and accurate analysis of voices under various conditions.

2.2. Multimodal fusion

In voice analysis, multi-modality refers to input data extracted from different data sources or forms of information [28]. For example, people chatting, singing, reading, or performing particular sound patterns are all typical modalities. The information provided by each modality may be different and complementary, and a single modality often cannot fully capture pathological features. Therefore, by fusing data from different modalities, we can have a more comprehensive understanding of the pathological condition, thus improving the accuracy and robustness of detection.

In multimodal fusion, there are three main strategies: early fusion, mid-level fusion, and late fusion [28, 29]. Early fusion combines features from different modalities into a vector before feeding them into a model. Mid-level fusion integrates data at an intermediate stage, allowing for more flexibility in capturing deeper correlations while maintaining some distinctions between modalities. Late fusion trains separate models for each modality and combines their predictions via an aggregation function such as average voting, weighted voting, or using a meta-classifier.

2.3. Shallow approach to Multi-modal Fusion for Voice Disorder detection

The research by Koudounas et al. [9] proposed an end-to-end method based on a transformer, which directly processes the original audio signal. To address the challenges posed by different recording types (such as sentence reading and sustained vowel utterances), they used a shallow mixture of experts (MoE) [30] framework to optimize the prediction alignment across recording types. Experimental results show that the method improves the single-modality approach in speech pathology detection and classification tasks, and achieves good performance on public and private datasets. However, this study mainly focuses on synthetic data and the MoE framework, and lacks in-depth exploration of multimodal fusion strategies.

Building on this, our study introduces a systematic study of multimodal fusion strategies in voice pathology detection. We focus on early, mid, and late fusion methods, especially mid and late fusion, because these two methods have greater flexibility and can capture deeper correlations between modalities. Compared with the method of [9], our study explores fusion strategies in more detail and demonstrates how mid-fusion strategies are the best multimodal approach in this domain to improve model generalization.

3. Method

This section outlines our contributions to multimodal fusion strategies, emphasizing the mathematical formulation of the problem and the model architecture. Specifically, we introduce early, mid-level, and late fusion strategies in transformer architecture that integrate multiple modalities for robust prediction.

In this study, we used two speech-based modalities to solve the voice pathology detection task, each capturing voice characteristics. The first modality x_1 represents the original features extracted from the sentence reading recording, while x_2 represents the features extracted from the second modality, the sustained vowel pronunciation recording. Given a multi-modal architecture f, we input the raw audio waveforms into the Wav2Vec2 model [27] to combine the feature extraction for the different modalities. The model then outputs the probabilities $\hat{y}$, which are used to produce the final classification result.

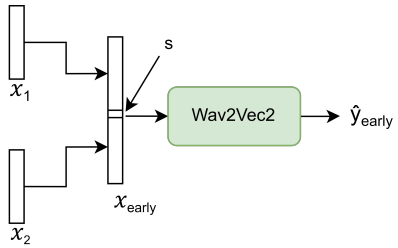

Figure 1: Diagram of the early-fusion.

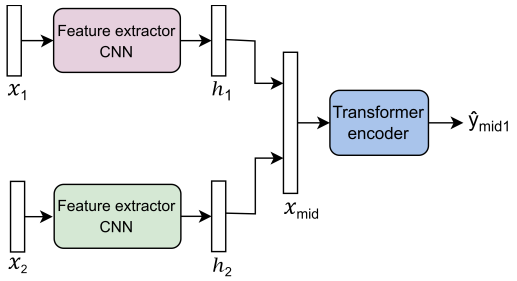

Figure 2: Mid-level fusion using concatenated embedding.

3.1. Early fusion strategies

The early fusion strategy connects the raw features of the two modalities into a unified input representation. In this stage, we first truncate all audio samples to a uniform length to standardize the input length and eliminate the bias caused by the difference in sample length. Then, we directly concatenate the raw audio features from the two modalities. Specifically, the raw audio features from the two modalities are concatenated in a fixed order: modality x_1 first, then modality x_2. To further distinguish the two of them, a 1-second silence (s) padding is inserted between the two, providing a clear boundary for the model (Eq.2). After concatenation, the generated unified features are fed as input into the pre-trained Wav2Vec2 model for prediction. Fig.1 visually depicts this process.

$$x_{early} = [x_1; s; x_2] \tag{1}$$

$$\hat{y} = \text{softmax}(f(x_{early})) \tag{2}$$

where:

- s: a 1-second silence padding between the two modalities.
- x_{early}: the concatenated feature vector after early fusion.
- The symbol [;]: the concatenation operation.
- $\hat{y}$: Predicted output probabilities which are produced by a softmax.

This method effectively captures modality-specific patterns from distinct modalities through simple and direct feature combinations.

3.2. Mid-level fusion strategies

In the mid-level fusion strategy, feature fusion is performed after CNN encoding but before the features are fed into the transformer encoder. This approach combines modality-specific features in a shared representation space, allowing the model to leverage interactions between modalities for more robust predictions. We will analyze two different fusion strategies: concatenated embedding and cross-attention.

3.2.1. Concatenated embeddings

In the concatenated embedding strategy, features are first extracted from each modality using a separate CNN layer and mapped to the same vector space (Eq.3). We thus decompose the network into the composition of two modules $f = g \circ e$, where e is the CNN-based feature extractor, while g represents the transformer encoder layers. After feature extraction, the extracted embeddings are normalized to ensure consistency across modalities, then concatenated as done in the early fusion approach (Eq.4), but at a deeper feature level, as shown in Figure 2. Finally, the combined feature vector goes through a dimension reduction layer to fit the input size of the subsequent transformer encoder.

$$h_1 = e_1(x_1), \quad h_2 = e_2(x_2) \tag{3}$$

$$x_{mid} = [h_1; h_2] \tag{4}$$

$$\hat{y} = \text{softmax}(g(x_{mid})) \tag{5}$$

where:

- h_1 and h_2: high-dimensional embeddings extracted from modalities x_1 and x_2 using CNN extractor, respectively.
- x_{mid}: Concatenated feature embeddings from both modalities.
- g represents the transformer encoder layers.

3.2.2. Cross-Attention

The cross-attention mechanism [11] dynamically captures interactions between modalities by computing attention weights based on the relationship between the Query (Q), Key (K), and Value (V) matrices. This allows the model to focus on important features across modalities.

First, given input feature matrices h_1 and h_2 of the two modalities, we generate Q, K, and V through linear transformation,

$$Q = h_1 W_Q, \quad K = h_2 W_K, \quad V = h_2 W_V \tag{6}$$

Here, W_Q, W_K, and W_V are learnable weight matrices for the query, key, and value, respectively. Next, we calculate the attention matrix A between the Query (Q) and the Key (K) by measuring their similarity, then normalized using softmax. The attention weight is used to perform a weighted sum of the Value V to generate output features O:

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) \tag{7}$$

$$O = AV \tag{8}$$

where:

- A is the general attention matrix.
- d_k is the dimension of the key, $\sqrt{d_k}$ is the normalization factor used for scaling.

As illustrated in Figure 3, cross-attention is computed in both directions to effectively capture interactions between the two modalities.

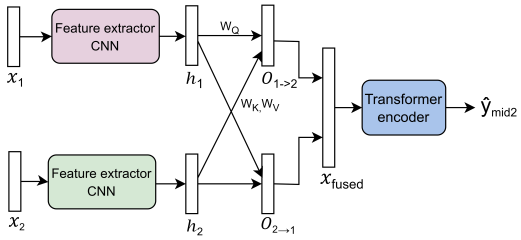

Figure 3: Mid fusion using cross-attention.

1. We use h_1 as the query and h_2 as the key and value to compute the attention (Eq.9).
2. We reverse the roles of the modalities and use h_2 as the query and h_1 as the key and value (Eq.10).

Finally, the outputs of the cross-attentions from both directions, $O_{1\to2}$ and $O_{2\to1}$, are concatenated to form a unified representation, x_{fused}, as shown in Eq.11.

This concatenation helps to merge the information from both modalities in a unified feature space. The fused features are then processed through a shared fusion layer before passing them to a transformer encoder for deeper feature extraction and ultimately classification.

$$O_{1\to2} = A_{1\to2}V_2 = \text{CrossAttention}(h_1, h_2) \tag{9}$$

$$O_{2\to1} = A_{2\to1}V_1 = \text{CrossAttention}(h_2, h_1) \tag{10}$$

$$x_{\text{fused}} = [O_{1\to2}; O_{2\to1}] \tag{11}$$

$$\hat{y} = \text{softmax}(g(x_{fused}) \tag{12}$$

where:

- Arrows represent the direction of attention.

3.3. Late fusion strategies

While mid-level fusion captures fine-grained interactions between modality-specific embeddings, late fusion is performed at the decision level, allowing each modality to be optimized independently and then integrated into a unified prediction. This approach allows each model to focus on its specific modality before being integrated, although it doubles the size of the final model. Two late fusion techniques are employed in our study.

Simple average In this approach, the outputs of the two models, $\hat{y}_1$ and $\hat{y}_2$, are combined by taking their simple average, as illustrated in the top part of Figure 4. This strategy assumes that both models contribute equally to the final prediction. The combined output $\hat{y}_{\text{late1}}$ is computed as follows:

$$\hat{y}_{\text{late1}} = \frac{1}{2}(\hat{y}_1 + \hat{y}_2) \tag{13}$$

where $\hat{y}_1$ and $\hat{y}_2$ are the probability distributions produced by the two individual models.

This fusion method is simple and it is computationally efficient as it avoids any extra parameters.

Mixture of Expert As a second late fusion strategy, we employ a shallow mixture of experts (MoE) to combine the outputs of two independent models and improve the overall performance of the system. Unlike the simple averaging method, this approach assigns weights to each model's predictions based on how relevant they are to the final output.

As shown in Figure 4, we use a simple multi-layer perceptron (MLP) configured with a single hidden layer to predict weights to combine the outputs of each model. The input layer of the MLP is a probabilistic concatenation of the two modalities ($\hat{y}_1$, $\hat{y}_2$), and the output layer applies a softmax function to ensure that the sum of all model weights is 1 (Eq.14). During inference, the final prediction is computed using the weights to combine the contributions of both models (Eq. 15). This approach improves the system's performance on unseen data while maintaining a simple architecture.

$$[w_1, w_2] = \text{softmax}(\text{MLP}([\hat{y}_1; \hat{y}_2])) \tag{14}$$

$$\hat{y}_{\text{late2}} = w_1 \cdot \hat{y}_{1,\text{test}} + w_2 \cdot \hat{y}_{2,\text{test}} \tag{15}$$

Here:

- $[w_1, w_2]$: Weights learned from the concatenated outputs $\hat{y}_1$ and $\hat{y}_2$ on the validation set.
- $\hat{y}_{1,\text{test}}$, $\hat{y}_{2,\text{test}}$: Predicted probabilities from the two models on the test set.

4. Results

This section provides an overview of the datasets and preprocessing methods used in our experiments, followed by a detailed description of the training setup to ensure reproducibility.

All experiments were conducted in a cloud-based environment equipped with a Tesla P100-PCIE-16GB GPU[1].

Details of the software environment can be found in the project repository[2].

4.1. Dataset

IPV The Italian Pathological Voice (IPV) dataset is a novel and diverse resource designed specifically for voice pathology research, currently unpublished and introduced in [9]. Collected from participants in Italian otolaryngology and voice therapy clinics, the dataset includes both healthy individuals and patients with varying degrees of voice disorders. All recordings were conducted under strict standardization protocols in quiet environments, ensuring high-quality samples with a signal-to-noise ratio exceeding 30 dB and a fixed microphone distance of 30 cm.

The dataset comprises two modalities: sustained phonation of the vowel /a/ (SV) and reading of five phonetically balanced sentences (CS) adapted from the Italian version of CAPE-V [31]. Each sample includes detailed health condition notes and diagnoses from experienced physicians. Table 1 provides a detailed summary of the dataset characteristics, including sample distribution, record length, and modal information.

Audio Preprocessing To ensure the consistency of audio duration and facilitate comparison, we cropped the samples in the datasets to fixed lengths: CS samples were cropped to 19 seconds, and SV samples were cropped to 18 seconds. These lengths are designed to cover approximately 90% of the samples in each modality, ensuring that most voice

[1]We gratefully acknowledge the computational resources provided by Kaggle (https://www.kaggle.com/) for this research. We also appreciate the early-stage support from HPC@Polito (http://www.hpc.polito.it).
[2]Github repository: github.com/multimodal_pathologies_prediction

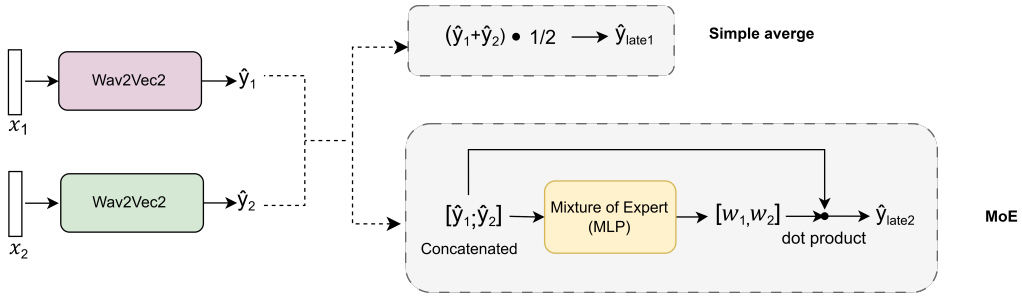

Figure 4: Two late fusion strategies. The upper right part is the simple average method, and the lower right part is the MoE method.

Table 1
The table summarizes the characteristics of the dataset. Healthy and Pathological represent the number of healthy and diseased samples, respectively. CS indicates the number of sentence reading samples, and SV represents the number of syllable articulation samples. $T(s)$ denotes the average duration of the audio samples in seconds.

	Healthy	pathological	CS	SV	T(s)
IPV	362	672	517	517	12.95

information is preserved for effective model training while reducing the impact of outlier samples that are too long. For samples shorter than the fixed lengths, zero-padding was applied to extend them to the required duration.

Then, the audio data was standardized using the predefined processor provided by the Wav2Vec2 framework. The processor first resamples the audio data to 16kHz to ensure compatibility with the framework, and reduce computational overhead. Then the converted feature representation can not only effectively capture the key information in the speech signal, but also provide consistent and efficient input features for the model to support subsequent training tasks.

In order to avoid issues with the imbalance of pathological voice data (healthy samples are less than pathological samples), a stratified sampling method was used in the data division process to ensure proportional representation of healthy and pathological samples across all splits. We divided the data into training, validation, and test sets in a ratio of 8:1:1 to ensure fair and reproducible evaluations. The test set was first separated using a fixed random seed. Subsequently, the training and validation sets were further split using three different random seeds to create multiple splits. The final results are calculated by averaging the performance metrics over these splits to ensure the robustness and reliability of the evaluation.

4.2. Baselines

To verify the effectiveness of our proposed method and provide a comparison, we designed a series of traditional baseline models, including the classic multi-layer perceptron (MLP) and a lightweight convolutional neural network (MobileNetV2 [32]) based on transfer learning. These baseline models are trained based on traditional audio features to evaluate the performance of different model architectures. In contrast, the unimodal model based on the Wav2Vec2 processor directly processes the audio waveform to extract features, reflecting the advantages of end-to-end methods.

In the feature extraction process of the baseline model, the audio data is uniformly sampled to 16kHz and truncated to a fixed maximum duration to ensure sample consistency. We extract 40-dimensional MFCC features through librosa, transpose them into a time-step sequence form, and uniformly zero-fill the feature sequence. At the same time, a padding mask is generated to distinguish between real data and padding parts. The following is the specific design of the two baseline models.

MLP is designed with two fully connected layers containing 50 hidden units, using the ReLU activation function to extract high-dimensional features, aggregating the time dimension information through the global average pooling layer, and finally performing binary classification through the Softmax output layer. The training process uses the Adam optimizer with a learning rate of 0.01, a batch size of 16, and an early stopping strategy to prevent overfitting.

2D-CNN The audio features are converted to 2D images by repeating a single channel to RGB three channels to fit the input requirements of the pre-trained model. We load the pre-trained weights (ImageNet [33]) of MobileNetV2 [32], remove the top classification head, and add a global average pooling layer, a 512-unit fully connected layer, and a Softmax classification layer. Dropout is added to the top network to improve generalization, and the pre-trained feature extraction part is fine-tuned. Two fine-tuning strategies are used: full fine-tuning and head-only fine-tuning. In full fine-tuning, all layers of MobileNetV2 are updated during training to maximize performance optimization; while in head-only fine-tuning, only the newly added classification head is trained, while the pre-trained feature extraction layer is frozen to retain the common features learned from ImageNet. The training hyperparameters of both strategies are consistent with the MLP model.

4.3. Training Procedure

Our method is based on a pre-trained Wav2Vec2.0 model (trained on the LibriSpeech 960-hour dataset) and evaluates three fusion strategies on the IPV dataset: early fusion, mid-level fusion, and late fusion.

Early fusion is accomplished by directly concatenating the original audio of CS and SV, and adding 1 second of silence (38 seconds) after the total length of the audio to avoid feature loss. The concatenation is performed on the same individual. The concatenated audio signals are uniformly processed in a Wav2Vec2.0 processor to ensure consistency in feature extraction. Mid-level fusion is based on 2 fine-tuned Wav2Vec2.0 models, and global feature modeling is achieved through a shared Transformer encoder (initialized

Table 2
Performance Comparison of Single-Modality Baselines and Dual-Modality Fusion Methods. CS refers to a single modality with sentence reading, SV to a single modality with vowel pronunciation. Values spanning both columns refers to modality fusion methods. Bold values indicate the best performance for a given metric.

Modality	Method	Accuracy		Macro F1	
		CS	SV	CS	SV
Single	MLP	.801±.011	.750±.057	.767±.022	.686±.053
	2D-CNN (Train all layers)	.667±.011	.673±.000	.400±.004	.402±.000
	2D-CNN (Fine-tune classify head)	.789±.019	.782±.048	.765±.021	.723±.063
	Wav2Vec2	.859±.029	.827±.000	.837±.038	.793±.000
Multi	Early Fusion	.859±.011		.829±.016	
	Mid (Concatenated Embeddings)	.878±.011		.838±.014	
	Mid (Cross Attention)	**.885±.000**		.843±.005	
	Late (Simple Average)	.852±.022		.824±.027	
	Late (MoE)	.872±.011		**.857±.012**	

with pre-trained Wav2Vec2 parameters). The first method directly concatenates CS and SV extracted features, and the second achieves feature interaction through a bidirectional cross-modal attention mechanism. The number of attention heads is set to 4. Late fusion utilizes fine-tuned CS and SV models to generate the final classification results by combining the probabilities from both modalities, either through simple averaging or a shallow MOE (an MLP with 10 hidden nodes) that determines modality weighting based on the probabilities from the training and validation sets.

All experiments above were completed within 50 training rounds (epochs), and using fixed random seed to ensure the reproducibility of the results. The AdamW optimizer (weight decay = 0.01) was used for all experiments. A linear learning rate scheduler is used to optimize the learning rate adjustment. The scheduler reduced the learning rate linearly over the total number of training steps, with no warm-up steps. Initial learning rates are optimized by manual adjustment, using 1e-5 for single modality and concatenated fusion and 6e-6 for cross-attention fusion. To address class imbalance, a weighted cross-entropy loss function was applied, with class weights computed based on the training dataset's label distribution. The batch size was set to 8, and an early stopping strategy with the patience of 10 epochs was used to terminate training when the validation performance plateaued. More experimental details and hyperparameter configurations can be found in the GitHub repository of the article.

4.4. Evaluation Metrics

To evaluate the performance of the model in the voice disorder detection task, we used two key metrics:

Accuracy Accuracy measures the proportion of correctly predicted samples to the total number of samples, providing an overall assessment of classification performance:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Samples}} \tag{16}$$

While accuracy is a useful general metric, it can be less informative in imbalanced datasets.

Macro F1-Score To better evaluate performance across imbalanced classes, we adopted the macro-average F1 score, which calculates the F1 score for each class and then averages them:

$$F1 = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

$$\text{Macro F1-Score} = \frac{1}{N} \sum_{i=1}^{N} F1_i \tag{18}$$

Here, N is the total number of categories. Macro F1-Score gives equal importance to all classes, making it particularly suitable for tasks with class imbalance, such as voice disorder detection.

5. Discussion

In this section, we analyze and interpret the experimental results by focusing on two key aspects: comparing baseline models to assess their effectiveness as reference, and evaluating different fusion models to explore their ability to integrate multimodal information and improve generalization to unseen data. By systematically studying these factors, we aim to highlight the strengths and limitations of the proposed approach and provide insights for future improvements.

Benchmark comparison Table 2 presents a comparison of the performance of unimodal baseline models for voice disorder detection on the IPV dataset.

As expected, Wav2Vec2 achieved the best results among the four baseline models, with accuracy of .859 and .827 in CS and SV modes, respectively, and .837 and .793 for F1 macro, respectively. The superior performance of Wav2Vec2 underscores the benefits of self-supervised pre-training on large-scale audio data. This means that the model does not need to be trained from scratch, but through pre-training and transfer learning capabilities, it can have audio features with good generalization capabilities, even with a small amount of labeled data. Moreover, it benefits of the attention mechanism which better extract relevant features from long sequence of data.

The MLP model performs well in CS mode with an accuracy of .801 and F1 Macro of .767, but drops to .750 and .686 in SV mode, highlighting its limitations in capturing complex audio features with limited contextual information. Compared to MLP, our method improves F1 Macro by +.07 in CS mode and +.10-.11 in SV mode, with corresponding accuracy improvements of +.05-.06 and +.07-.08.

For the 2D-CNN, fully fine-tuning all layers leads to poor performance (.667 and .673 accuracy in CS and SV modes, respectively; .400 and .402 F1 Macro), likely due to the disruption of pre-trained features. Fine-tuning only the classification head improves the performance to .789 and .782 accuracy in CS and SV modes, and .765 and .723 F1 Macro,

respectively. However, our method performs better than the fine-tuned 2D-CNN, with +.07-.08 improvement in F1 Macro, +.07-.08 and +.04-.05 improvement in accuracy in CS and SV modes.

The above results show that fine-tuning the pre-trained Wav2Vec2 model is an effective solution for small dataset tasks, it highlights the necessity of carefully designed optimization methods.

Fusion strategy vs. single modality performance The fusion method shows an advantage over the single modality by effectively combining the complementary information of CS and SV inputs. In particular, as shown in Table 2, the proposed mid-level fusion pipeline shows significant improvements over single modality models. Concatenated Embeddings improves accuracy by +.02 and macro F1 by +.001 on the CS model, and by +.05 and +.04 on the SV model, respectively. Cross Attention performs even better, with accuracy and F1 gains of +.03 and +.006 on the CS model, and +.06 and +.05 on the SV model for accuracy and macro F1, respectively. These results highlight the benefits of leveraging complementary information from multiple modalities.

When compared to other fusion strategies, instead, mid-level fusion consistently outperforms both early and late fusion methods. The cross-attention method achieves the best results with .885 accuracy and .843 macro F1, which is +.02-.03 in accuracy and +.01-.02 in macro F1 compared with early fusion. Similarly, it achieves +.01-.04 improvement in accuracy and +.01-.03 improvement in macro F1 compared to late fusion strategies such as Mixture of Experts (MoE). These results demonstrate the effectiveness of dynamically capturing inter-modality dependencies during feature integration.

Compared with early fusion that concatenates raw features, the proposed mid-level fusion method can model complex inter-dependencies, leading to robust feature representation. In contrast, late fusion methods, while simpler to implement, operate at the decision level and cannot fully exploit the interactions between modalities.

In summary, our proposed mid-level fusion strategy, especially the cross-attention strategy, achieves the best performance among all methods. The results show that it is able to dynamically integrate complementary modality information, leading to significant improvements in accuracy and macro F1 performance.

6. Conclusion

This study investigates the effectiveness of various models, and fusion methods for speech impairment detection using unimodal and multimodal approaches. We leverage end-to-end pre-trained models Wav2Vec2, which is once again proven to be an effective model for solving audio tasks, even with a limited dataset size. This not only reduces the steps of manual feature extraction but also enables robust features to be extracted from audio data through self-supervised pre-training, showing good generalization ability.

Among multimodal methods, our experiments show that mid-level fusion strategies, especially the cross-attention mechanism, outperform early and late fusion techniques. The cross-attention mechanism dynamically captures fine-grained inter-modal dependencies, leading to the highest performance. In contrast, early fusion methods, while beneficial for capturing joint features from the beginning, may lack flexibility in handling complex interactions between modalities. This often leads to inferior performance compared to mid-level fusion. Late fusion methods are easier to implement but have limited capabilities in modeling complex feature interactions and a higher number of parameters.

These findings provide valuable insights into the design of voice disorder detection systems, especially with regard to their potential applications in clinical diagnosis and health monitoring.

Future Work Although this study provides valuable insights, there are still some limitations and directions for improvement.

First, the experiments are limited to a specific dataset, IPV, which contains two homogeneous audio modalities and cannot cover a wider range of scenarios. Future work can explore larger and more diverse datasets, including datasets collected in realistic noisy environments, or cross-lingual datasets to evaluate the reliability of the model in the real world. In addition, future work can integrate other medical modalities (e.g. laryngoscope images + audio samples), to expand audio beyond the audio domain for more comprehensive voice disorder detection. Second, the current study only focuses on voice disorder detection tasks. In the future, it can be expanded to multi-classification tasks to more comprehensively evaluate the effectiveness of the model in practical applications, especially in the classification of different types of pathologies that are at the root of the voice disorder.

Third, we only used the wav2vec2 model for feature extraction and multi-modal fusion, and did not compare it on other advanced Tansformer models (e.g. Hubert, WavLM, etc). Future work can explore and evaluate their effectiveness in the medical voice pathology analysis of these models.

Data augmentation techniques can also be combined to enhance the generalization ability of the model, so as to maintain excellent performance in more diverse application scenarios.

By addressing these limitations, future research can build on this study to develop more powerful, efficient, and scalable voice disorder detection solutions, thereby bringing greater social and technological impact for practical applications.

References

[1] N. Bhattacharyya, The prevalence of voice problems among adults in the united states, The Laryngoscope 124 (2014) 2359–2362.

[2] N. Roy, R. M. Merrill, S. D. Gray, E. M. Smith, Voice disorders in the general population: prevalence, risk factors, and occupational impact, The Laryngoscope 115 (2005) 1988–1995.

[3] C. L. Payten, G. Chiapello, K. A. Weir, C. J. Madill, Frameworks, terminology and definitions used for the classification of voice disorders: a scoping review, Journal of Voice (2022).

[4] P. Daraei, C. R. Villari, A. D. Rubin, A. T. Hillel, E. R. Hapner, A. M. Klein, M. M. Johns, The role of laryngoscopy in the diagnosis of spasmodic dysphonia, JAMA Otolaryngology–Head & Neck Surgery 140 (2014) 228–232.

[5] G. Ciravegna, A. Koudounas, M. Fantini, T. Cerquitelli, E. Baralis, E. Crosetti, G. Succo, Non-invasive ai-powered diagnostics: The case of voice-disorder detection-vision paper, EDBT/ICDT Workshop 2348 (2024).

[6] M. Fantini, G. Ciravegna, A. Koudounas, T. Cerquitelli, E. Baralis, G. Succo, E. Crosetti, The rapidly evolving scenario of acoustic voice analysis in otolaryngology, Cureus 16 (2024) e73491.

[7] P. Rajpurkar, E. Chen, O. Banerjee, E. J. Topol, Ai in health and medicine, Nature medicine 28 (2022) 31–38.

[8] S. Schneider, A. Baevski, R. Collobert, M. Auli, wav2vec: Unsupervised pre-training for speech recognition, arXiv preprint arXiv:1904.05862 (2019).

[9] A. Koudounas, G. Ciravegna, M. Fantini, E. Crosetti, G. Succo, T. Cerquitelli, E. Baralis, Voice disorder analysis: a transformer-based approach, in: Interspeech 2024, 2024, pp. 3040–3044. doi:10.21437/Interspeech.2024-1122.

[10] M. La Quatra, M. F. Turco, T. Svendsen, G. Salvi, J. R. Orozco-Arroyave, S. M. Siniscalchi, Exploiting foundation models and speech enhancement for parkinson's disease detection from speech in real-world operative conditions, in: Interspeech 2024, 2024, pp. 1405–1409. doi:10.21437/Interspeech.2024-522.

[11] A. Vaswani, Attention is all you need, Advances in Neural Information Processing Systems (2017).

[12] X. Peng, H. Xu, J. Liu, J. Wang, C. He, Voice disorder classification using convolutional neural network based on deep transfer learning, Scientific Reports 13 (2023) 7264.

[13] L. W. Lopes, L. B. Simões, J. D. da Silva, D. da Silva Evangelista, A. C. d. N. e Ugulino, P. O. C. Silva, V. J. D. Vieira, Accuracy of acoustic analysis measurements in the evaluation of patients with different laryngeal diagnoses, Journal of voice 31 (2017) 382–e15.

[14] M. Alhussein, G. Muhammad, Automatic voice pathology monitoring using parallel deep models for smart healthcare, Ieee Access 7 (2019) 46474–46479.

[15] P. H. Leung, K. T. Chui, K. Lo, P. O. de Pablos, A support vector machine–based voice disorders detection using human voice signal, in: Artificial Intelligence and Big Data Analytics for Smart Healthcare, Elsevier, 2021, pp. 197–208.

[16] X. Peng, H. Xu, J. Liu, J. Wang, C. He, Voice disorder classification using convolutional neural network based on deep transfer learning, Scientific Reports 13 (2023) 7264.

[17] U. K. Lilhore, S. Dalal, N. Faujdar, M. Margala, P. Chakrabarti, T. Chakrabarti, S. Simaiya, P. Kumar, P. Thangaraju, H. Velmurugan, Hybrid cnn-lstm model with efficient hyperparameter tuning for prediction of parkinson's disease, Scientific Reports 13 (2023) 14605.

[18] A. S. Almasoud, T. A. E. Eisa, F. N. Al-Wesabi, A. Elsafi, M. Al Duhayyim, I. Yaseen, M. A. Hamza, A. Motwakel, Parkinson's detection using rnn-graph-lstm with optimization based on speech signals, Comput. Mater. Contin 72 (2022) 872–886.

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

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

[21] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al., Wavlm: Large-scale self-supervised pre-training for full stack speech processing, IEEE Journal of Selected Topics in Signal Processing 16 (2022) 1505–1518.

[22] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, Hubert: Self-supervised speech representation learning by masked prediction of hidden units, IEEE/ACM transactions on audio, speech, and language processing 29 (2021) 3451–3460.

[23] A. Koudounas, E. Pastor, G. Attanasio, L. de Alfaro, E. Baralis, Prioritizing data acquisition for end-to-end speech model improvement, in: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 7000–7004. doi:10.1109/ICASSP48485.2024.10446326.

[24] A. Koudounas, M. La Quatra, S. M. Siniscalchi, 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.

[25] M. La Quatra, A. Koudounas, L. Vaiani, E. Baralis, P. Garza, L. Cagliero, 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.

[26] 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, H. yi Lee, Superb: Speech processing universal performance benchmark, in: Interspeech 2021, 2021, pp. 1194–1198. doi:10.21437/Interspeech.2021-1775.

[27] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.

[28] S. R. Stahlschmidt, B. Ulfenborg, J. Synnergren, Multimodal deep learning for biomedical data fusion: a review, Briefings in Bioinformatics 23 (2022) bbab569.

[29] L. Ilias, D. Askounis, J. Psarras, Detecting dementia from speech and transcripts using transformers, Computer Speech & Language 79 (2023) 101485.

[30] R. Gupta, K. Audhkhasi, S. Narayanan, A mixture of experts approach towards intelligibility classification of pathological speech, in: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2015, pp. 1986–1990.

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

[32] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, L. Chen, Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation, CoRR abs/1801.04381 (2018). URL: http://arxiv.org/abs/1801.04381. arXiv:1801.04381.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.