Evaluating Bias and Enhancing Fairness in Acoustic Diagnostic Models

Sihan. Xie, M.Eng. Candidate
Shanghaitech University, Shanghai, China

Abstract

With the development of deep learning, voice disorder diagnosis can automatically identify diseases from acoustic features, enabling efficient medical assistance with broad application potential. However, most existing studies focus primarily on overall model performance, while overlooking potential bias across demographic groups. To address this issue, this project employed two datasets—the Chinese EENT dataset from Fudan University and the German SVD dataset—and extracted pitch and mel spectrograms to support model training. A Gaussian Mixture Model (GMM) based on pitch features and a Convolutional Neural Network (CNN) based on Mel-spectrograms were developed and evaluated for model performance. Fairness evaluation reveals model biases across gender and age groups. To mitigate these biases, future work will incorporate multi-metric fairness evaluation, investigate the sources of bias, and explore mitigation strategies. This project aims to support the development of voice disorder diagnostic systems that are both accurate and fair across demographic groups.

Introduction

In the medical field, voice data has emerged as a promising modality for disease screening and monitoring due to its non-invasive nature, low cost, and remote accessibility. It can serve as a biomarker for conditions such as neurological and voice disorders, where symptoms manifest through speech changes. Traditional diagnostic methods rely on clinicians' auditory judgment or specialized equipment, which are often subjective and labor-intensive. Recent advances in deep learning have enabled the use of acoustic features—such as pitch, Mel-Frequency Cepstral Coefficients (MFCCs), and mel spectrograms—as input for disease classification tasks. This approach provides a new technical foundation for building efficient acoustic diagnostic models.

However, most existing studies focus primarily on improving overall model performance, such as accuracy, while paying little attention to fairness across demographic groups. Prior research has shown that AI models used in medical applications may suffer from algorithmic bias, potentially resulting in lower performance for certain populations[1]. This can lead to misdiagnosis and may further exacerbate inequalities in healthcare delivery. This issue is particularly critical in voice-based systems, as voice characteristics are inherently influenced by both gender and age.

To address this gap, the project aims to develop a deep learning-based acoustic diagnostic model, with a particular focus on evaluating its fairness across different gender and age groups. The project will also explore potential sources of bias and investigate strategies for mitigating them. An overview of the project workflow is illustrated in Figure 1. By emphasizing not only model performance but also fairness in diverse populations, this project seeks to contribute to the clinical usability and ethical reliability of diagnostic technologies. It offers both methodological innovation and practical significance.

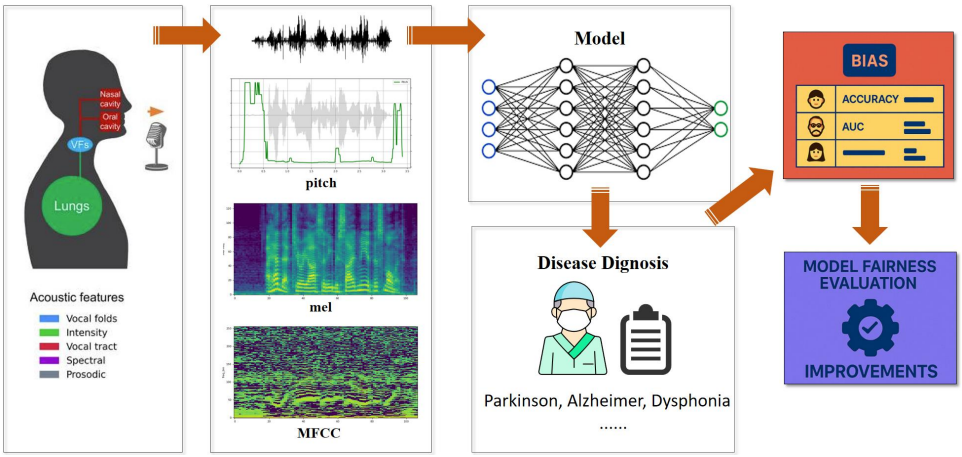

Figure 1. Overview of voice-based disease diagnosis and fairness evaluation framework.

Pilot Study

The pilot study employed two voice datasets: the Chinese EENT dataset and the German SVD dataset, both containing recordings from healthy individuals and patients with dysphonia.

The EENT dataset was provided by the Eye, Ear, Nose and Throat (EENT) Hospital of Fudan University and served as the main dataset for training, validation, and testing[2]. It includes recordings from 461 participants speaking standard Mandarin, focusing on the sustained vowels /a/ and /i/. The detailed subject characteristics on this dataset are summarized in Table 1. These statistics provide a foundation for subsequent bias analysis.

Table 1. Subject characteristics of the EENT dataset.

Class	Diagnosis	Number			Age (mean ± SD) (years)
		Female	Male	Total	
Normal		143	80	223	39.3 ± 11.9
Dysphonia		178	60	238	40 ± 13.5
	Functional dysphonia	25	6	31	
	Glottal incompetence	30	7	37	
	Nodules	23	0	23	
	Polyps	34	4	38	
	UVFP	38	23	61	
	Sulcus	28	20	48	
Total		321	140	461	39.6 ± 12.7

The SVD dataset (Saarbruecken Voice Database) is a publicly available German voice database, used in the pilot study as an independent external test set. After excluding low-quality recordings, the final SVD dataset contains 200 participants, including 100 healthy individuals and 100 patients with dysphonia. As with the EENT dataset, only the vowels /a/ and /i/ were used for model training and evaluation.

To ensure consistent model training and reliable feature extraction, the pilot study implemented a structured preprocessing pipeline. Voiced segments were first identified using a silence detection algorithm, and unstable portions at the beginning and end were trimmed to retain the central stable region. Each audio clip was then split into non-overlapping 1.5-second segments and normalized in duration through padding or truncation. Loudness normalization was applied where necessary to reduce variability from recording conditions. Two types of acoustic features were extracted: mel spectrograms and pitch. Both features were standardized using z-score normalization and finally saved as .pkl files, organized by language, vowel type, and feature type for use in subsequent model development.

Based on the extracted features, this pilot study constructed two classification models to compare their performance. The first is a traditional machine learning model—Gaussian Mixture Model (GMM), which is primarily built using the pitch as an acoustic feature. The second is a deep learning model based on a two-dimensional Convolutional Neural Network (CNN), which takes mel-spectrograms as input features. The CNN model consists of four convolutional modules and two fully connected layers. Each convolutional module includes a two-dimensional convolutional layer and a max-pooling layer, which are used to extract frequency-domain and time-domain features, respectively, while downsampling to retain key information. The architecture of the CNN is illustrated in Figure 2. The entire network was implemented using the TensorFlow framework, with the Adam optimizer and cross-entropy loss function employed during training.

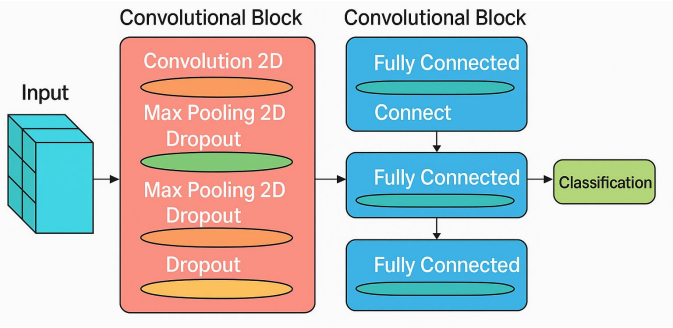

Figure 2. Architecture of the 2D CNN model.

To evaluate models, the pilot study used random sampling to split the EENT dataset, with details in Table 2. To prevent data leakage, all audio samples from the same participant were placed in one set, ensuring no participant appears in both sets. The vowel categories for each participant were also kept consistent across sets. Five-fold cross-validation was used on the development set during training. The final model was then evaluated on the independent test set, including the EENT test set and external SVD test set, to assess generalization.

Table 2. Allocation of subjects and standardized audio segments.

	Development set (80%)			**Test set (20%)**			**Total**		
	Normal	**Dysphonia**	**Total**	**Normal**	**Dysphonia**	**Total**	**Normal**	**Dysphonia**	**Total**
Participants	178	190	368	45	48	93	223	238	461
Segments /a/	1152	989	2141	291	260	551	1443	1249	2692
Segments /i/	1322	981	2303	320	237	557	1642	1218	2860

Results showed that the CNN model outperformed the GMM model on both test sets. On the EENT test set, the CNN achieved an accuracy of 97%, compared to 72% for GMM. On the SVD test set, the CNN reached 87%, while GMM reached 73%. Figure 3 further confirmed these results with ROC curves and AUC values, demonstrating the superior generalization of the CNN model in cross-language acoustic diagnosis.

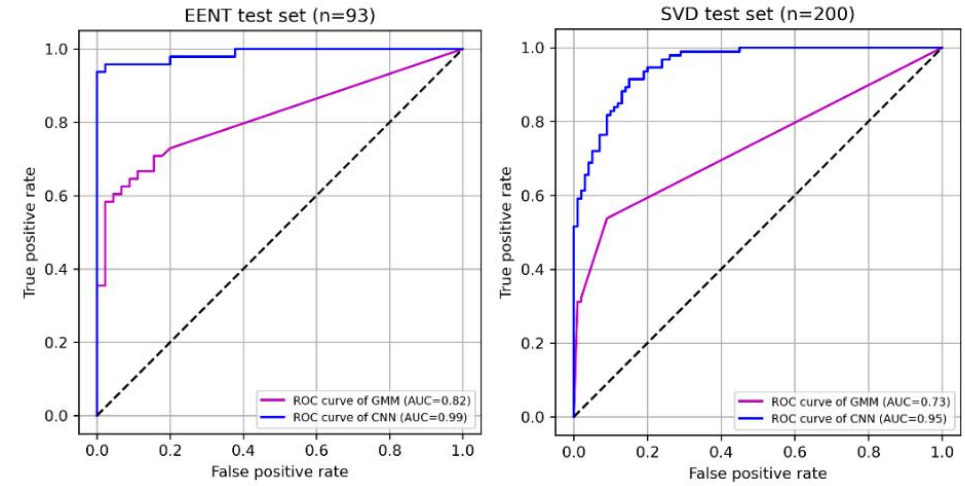

Figure 3. ROC and AUC of CNN and GMM on EENT and SVD test sets.

In addition to evaluating the diagnostic performance, it is also important to consider a critical question: is the model fair across different population groups? To this end, the pilot study conducted a preliminary bias analysis, focusing on the model's performance across different age and gender groups. At this stage, the bias analysis was limited to the GMM model, while the fairness evaluation of the CNN model will be carried out in extended project.
Specifically, participants were divided into three age groups: young (under 35), middle-aged (35 – 50), and old (over 50). Grouping was also performed based on gender. The classification accuracy of the model was compared across these subgroups, as shown in Table 3.

Table 3. Accuracy of the GMM model in different age and gender groups.

Group		**Participant Number**	**Accuracy**	
			EENT test set	**SVD test set**
Age split	Young (0,35)	185	61%	56%
	Middle-aged [35,50)	181	60%	58%
	Old [50,+)	95	54%	69%
Gender split	Female	321	57%	58%
	Male	140	63%	64%

In the EENT test set, the older group showed lower accuracy compared to the young and middle-aged groups. In contrast, the older group achieved the highest accuracy in the German SVD test set. These findings suggest that the model's performance disparities may not be solely attributed to age, but also influenced by language-specific factors. In terms of gender, Female participants consistently showed lower accuracy than males across both test sets, indicating a potential systematic gender bias.
In summary, the current model demonstrates systematic gender bias, highlighting the importance of identifying its sources and developing mitigation strategies.

Extended Project

The Pilot Study observed notable model's bias. Therefore, the extended project will focus on the systematic evaluation of model bias, analysis of its underlying causes, and the development of mitigation strategies—shifting from bias evaluation to fairness enhancement.
To begin with, a set of fairness metrics will be introduced to enable a more comprehensive evaluation of model performance across demographic groups. These include the True Positive Rate (TPR) Gap, which measures disparities in sensitivity between groups; Demographic Parity, which assesses whether predictions are independent of group membership; and Equalized Odds, which requires both true positive rates and false positive rates to be similar across groups. These metrics go beyond traditional accuracy measures by capturing fairness from multiple perspectives.
To better understand the sources of bias, model interpretability techniques, particularly SHAP value analysis, will be applied to examine whether the model disproportionately relies on specific acoustic features. Additionally, statistical comparisons of feature distributions across demographic groups will be conducted to explore whether imbalance in input representations contributes to observed disparities.
Based on the results of bias detection and source analysis, mitigation strategies will be developed at both the data and model levels. At the data level, the VOICED dataset[3]—which includes explicit age and gender labels—will be incorporated to improve the model's ability to learn from group-specific characteristics. During training, sample reweighting will be employed to assign greater importance to underrepresented or underperforming groups, thereby reducing bias in model learning.
It is noteworthy that this project also faces potential risks, particularly model overfitting in small demographic subgroups. To mitigate this, the CNN model incorporated Dropout and Early Stopping strategies to improve generalization. Furthermore, five-fold cross-validation was used to enhance the reliability of evaluation outcomes. Final model performance will be assessed on other datasets to validate the effectiveness and cross-lingual generalizability of the proposed fairness-enhancing methods.

Timeline

The planned tasks and corresponding timeline are presented in the Table 4.

Table 4. Extended project timeline.

Period	Task
5.19 - 5.26	Successfully run fairness evaluation code for the CNN. Use statistical methods to assess the fairness of both GMM and CNN.
5.27 - 6.2	Investigate sources of bias and explore mitigation strategies. Expand the dataset for disease diagnosis and evaluate bias.
6.3 - 6.6	Conduct fairness evaluations through multiple runs. Organize the results, and prepare for the defense.
6.7 - 6.12	Organize data and code, and complete the final report.

References

1. Wan Z, Guo Y, Bao S, Wang Q, Malin BA. Evaluating sex and age biases in multimodal large language models for skin disease identification from dermatoscopic images. Health Data Science. 2025 Apr 1;5:0256.
2. Chen Z, Zhu P, Qiu W, Guo J, Li Y. Deep learning in automatic detection of dysphonia: Comparing acoustic features and developing a generalizable framework. International Journal of Language & Communication Disorders. 2023 Mar;58(2):279-94.
3. Cesari U, De Pietro G, Marciano E, Niri C, Sannino G, Verde L. A new database of healthy and pathological voices. Computers & Electrical Engineering. 2018 May 1;68:310-21.