Exploring In-Context Learning Capabilities of ChatGPT for Pathological Speech Detection

Mahdi Amiri[1,2], *Hatef Otroshi Shahreza*[1], *Ina Kodrasi*[1]

[1]Idiap Research Institute, Switzerland
[2]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Email: {mahdi.amiri, hatef.otroshi, ina.kodrasi}@idiap.ch

Abstract

Automatic pathological speech detection approaches have shown promising results, gaining attention as potential diagnostic tools alongside costly traditional methods. Recently, it has been demonstrated that large language models (LLMs) can be leveraged for downstream tasks through few-shot in-context learning. In this paper, we investigate the use of multimodal LLMs, specifically ChatGPT-4o, for automatic pathological speech detection in a few-shot in-context learning setting. Experimental results demonstrate that this approach achieves competitive performance compared to state-of-the-art methods. To further understand its effectiveness, we conduct an ablation study to analyze the impact of different factors, such as input type and system prompts, on the final results. Our findings highlight the potential of multimodal LLMs for further exploration and advancement in automatic pathological speech detection.

1 Introduction

Pathological speech can result from neurological damage caused by conditions such as Cerebral Palsy, Amyotrophic Lateral Sclerosis, or Parkinson's disease. These disorders often lead to speech impairments such as dysarthria and apraxia of speech, which can significantly affect communication [1, 2]. Traditionally, speech and language pathologists conduct auditory-perceptual assessments to diagnose these conditions in a clinical setting, which is both costly and time-consuming. To reduce this burden on healthcare systems, researchers are actively developing automated methods for detecting pathological speech. Earlier approaches combined handcrafted acoustic features with traditional machine learning techniques [3–5]. With the remarkable success of deep learning (DL) in various fields [6, 7], efforts have increasingly shifted toward using DL-based approaches for automatic pathological speech detection [8–15].

With the rise of multimodal large language models (LLMs), new research directions are emerging that extend beyond traditional DL approaches. While these models were originally designed for natural language processing, they have been developed and extended for other domains [16–18]. Among these multimodal LLMs, ChatGPT-4o [19] stands out as one of the most advanced models, demonstrating exceptional capabilities in understanding and processing different modalities (such as text and vision) among different applications [20, 21]. Moreover, the performance of these models can be improved on downstream tasks through few-shot in-context learning [22]. In few-shot in context learning, the weights of the model remain unchanged, but the model is prompted with a few examples before being asked about the test query.

Given the promising in-context learning capabilities of ChatGPT, in this paper, we investigate the performance of the multimodal ChatGPT-4o for pathological speech detection in a few-shot in-context learning scenario. Ideally, one should directly analyze raw speech inputs in the context of pathological speech detection; however, GPT-4o does not support direct audio input. While GPT-4o-audio-preview supports direct audio input, its overall capabilities relative to GPT-4o remain unclear, as it is a preview model. Therefore, we focus on evaluating GPT-4o's ability to process short-time Fourier transform (STFT) magnitude spectrogram representations for pathological speech detection, making it the central objective of our paper. Nevertheless, to ensure completeness, we also provide an ablation study assessing the performance when using raw speech input and the GPT-4o-audio-preview model.

Experiments on the Noise Reduced UA-Speech database [23], which includes control and dysarthric speech from Cerebral Palsy patients, show promising results. Specifically, ChatGPT-4o achieves competitive performance compared to a state-of-the-art (SOTA) pathological speech detection approach that also operates on magnitude STFT spectrogram inputs [8, 24], despite having access to significantly less labeled training data. Notably, while the SOTA model is trained from scratch with more data [8, 24], it lacks the broad knowledge and pretraining of ChatGPT-4o. These findings highlight the potential of multimodal LLMs for further exploration and advancement in automatic pathological speech detection. To the best of our knowledge, this work is the first to employ multimodal LLMs for pathological speech detection. The key contributions of this paper are:

- We propose a method for utilizing multimodal LLMs in automatic pathological speech detection.
- We evaluate our proposed method against a SOTA baseline, demonstrating that it offers promising and competitive performance.
- We conduct an ablation study to further analyze the effect of different factors on the performance of the proposed method.

2 Related Works

2.1 DL-based Automatic Pathological Speech Detection

Traditionally, DL-based automatic pathological speech detection approaches use time-frequency input representations such as STFT [8], Mel-frequency cepstral coefficients [11, 12], or Mel spectrograms[10]. These representations are then processed with architectures like convolutional neural networks (CNNs)[8], recurrent neural networks [13], or autoencoders [10], to learn pathology-discriminant cues and perform automatic pathological speech detection. Moreover, with the success of self-supervised foundation models

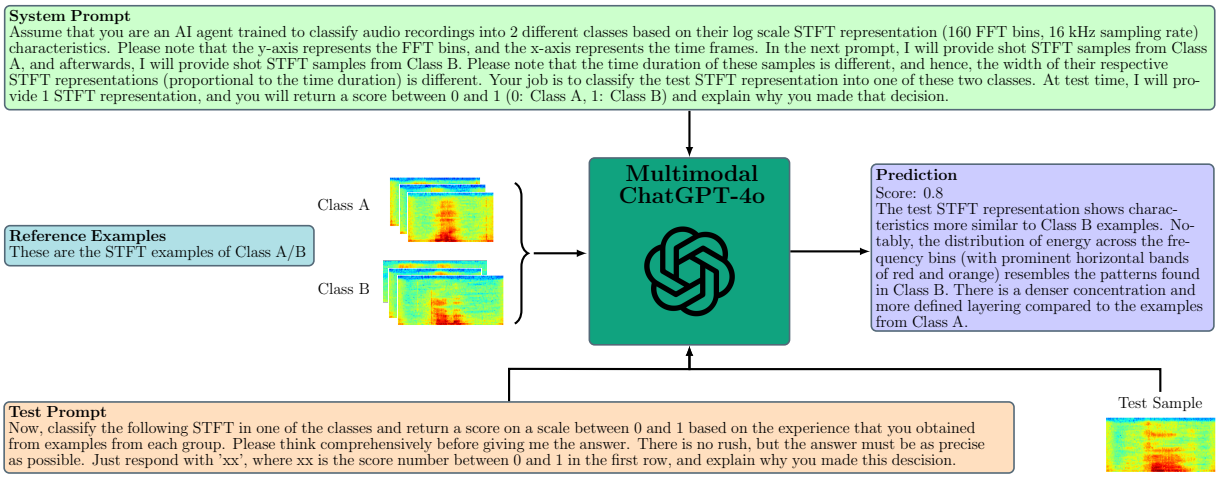

Figure 1: Schematic illustration of the proposed method. We set a system prompt that describes the classification task, input representation, and the number of reference samples per class. Then, several samples from each class are provided to the model, which is asked to classify the test sample based on them. In response, the model returns a classification score and explains the reasoning behind its decision.

like wav2vec 2.0 [25] for several downstream tasks [26], researchers have started leveraging embeddings obtained from these models and combining them with classifiers such as multi-layer perceptrons [24] for automatic pathological speech detection [27]. Although wav2vec 2.0-based approaches generally outperform those using time-frequency input representations, the CNN-based approach operating on STFT magnitude spectrograms has also demonstrated a strong performance [24].

2.2 Multimodal LLMs

Building on the strong performance of the Transformer architecture [6] across various natural language processing tasks, several studies have explored scaling training data and network structures to train LLMs. The pretrained LLMs could surpass SOTA models in downstream tasks for natural language understanding [22, 28]. Following the remarkable performance of LLMs, several models have been proposed that can also process other modalities (e.g., images, audio) in addition to text prompts [19, 29–31], enabling zero-shot and few-shot learning. Among these multimodal LLMs, GPT-4o [19] (also known as ChatGPT-4o) has achieved SOTA performance on various multimodal benchmarks. Consequently, we adopt this model for our study.

3 Proposed ChatGPT-4o-based Approach

Fig. 1 provides a schematic illustration of the proposed ChatGPT-4o-based approach for pathological speech detection. As depicted, the model is first provided with a system prompt describing the classification task, input representation, and number of reference samples. We then present the LLM with several reference samples (depending on the number of shots used) from both control and pathological speakers. Using this contextual knowledge, the model classifies a given test sample and generates an explanation for its decision.

This approach enables a broad range of analyses, which we explore in an ablation study. In Section 5, we evaluate the performance by framing the task specified in the system prompt as either a general audio classification task (cf. Fig. 1 and Section 5.1) or a dysarthria classification task (cf. Section 5.2). Additionally, we investigate the effect of requesting both a classification score between 0 and 1 (0: Controls, 1: Patients with Cerebral Palsy) along with an explanation (cf. Fig. 1 and Section 5.1) versus requesting only a classification score (cf. Section 5.3). Finally, we analyze the influence of the input representation on performance by evaluating the ChatGPT-4o-audio-preview model with raw speech input (cf. Section 5.4).

4 Experimental Settings

4.1 Database

We use the Noise Reduced UA-Speech Dysarthria Dataset [23][1], which is a denoised version of the UA-Speech [32] dataset. The dataset includes recordings from 16 speakers with Cerebral Palsy (4 females, 11 males) and 13 control speakers (4 females, 9 males). Each speaker utters various common words (CW), uncommon words (UW), commands (C), letters (L), and digits (D). The recordings are acquired using a 7-channel microphone array with a sampling frequency of 44.1 kHz. Recordings are downsampled to 16 kHz. For the following experiments, we use recordings from the arbitrarily selected 5-th channel.

Prior research [12] has shown that the UA-Speech dataset exhibits significant differences between control and pathological recordings due to variations in recording setups and noise conditions. These differences are easier for DL-based approaches to learn than pathology-discriminant cues [12], yielding an unconventionally high accuracy of these approaches on this dataset. Although we use the denoised version of the UA-Speech dataset, non-pathology-related differences persist, as enhancement methods introduce distortions that depend on noise conditions. As a result, the Noise Reduced UA-Speech dataset is also not optimal for automatic pathological speech detection experiments. However, due to licensing restrictions, it is, to our knowledge, the only dataset we can redistribute to third parties (i.e.,

[1] Under Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license.

Table 1: Speaker-level accuracy of the proposed ChatGPT-4o-based model and the CNN-based baseline. The ChatGPT-4o model is evaluated under different few-shot scenarios. For a fair comparison, in addition to the baseline model trained on the entire training set, we also consider baseline models trained only on data from the same speakers included in the few-shot setup.

Model	ChatGPT-4o-based Network			CNN-based baseline			
Shot/Speaker number	1 Shot	3 Shot	5 Shot	2 spk	6 spk	10 spk	Full Data
Accuracy (%)	70.2 ± 3.4	82.1 ± 0.0	85.7 ± 0.0	84.5 ± 1.6	88.1 ± 1.6	90.5 ± 1.6	95.2 ± 1.6

upload to ChatGPT-4o). For this reason, we have chosen to use it in this paper despite its limitations.

4.2 CNN Baseline Model

As previously mentioned, this paper focuses on analyzing ChatGPT-4o's ability to process STFT magnitude spectrograms. For a fair comparison, we use a SOTA CNN model that also operates on STFT magnitude spectrograms as our baseline [8]. Since the CNN model accepts only fixed-size inputs, we split each utterance into 500 ms segments with an overlap of 250 ms. The STFT of these segments is computed using a 10 ms Hanning window without overlap and the logarithm of the STFT magnitude is used as input representation. Input representations are normalized through a LayerNorm function ($\mu = 0, \sigma = 1$). The CNN architecture we use is adopted from [8].

For training and evaluation, we use a leave-one-speaker-out approach. In each fold, one speaker is used as a test speaker, while the remaining speakers are divided into training and validation sets with a $9:1$ ratio. The model is trained using the Adam optimizer with a learning rate of 0.001 and a weight decay of 5×10^{-3}.

4.3 ChatGPT-4o Setup

For the experiments conducted with ChatGPT-4o, we used OpenAI's official API. To use ChatGPT-4o in an in-context learning scenario, we first define a system message that outlines the classification task, input representation, and number of reference samples per class. Next, we present the reference samples for each group based on the selected number of shots for in-context learning. Finally, we provide the test sample to the model.

For selecting the reference samples, we strive to maintain a balanced distribution. We ensure that the number of speakers from each gender is equal in both classes and select identical utterances for each group. For instance, if a male speaker utters a specific common word in the control group, there will also be a male speaker uttering the same word in the pathological group. For the test samples, we randomly select two utterances from each of the five groups of recordings (i.e., CW, UW, C, D, and L), resulting in a total of 10 test samples. In this setup, each test speaker is evaluated separately, using reference samples from (a subset of) the remaining speakers.

4.4 Performance Evaluation

To evaluate the performance of the considered approaches, we consider the speaker-level accuracy. For the baseline model, we compute soft labels by passing the network's output through a softmax function for all segments belonging to each speaker. The speaker-level decision is then made through soft voting based on the scores of all these segments. For the ChatGPT-4o-based approach, the speaker-level decision is made through soft voting of the classification scores obtained for of all utterances belonging to the speaker.

Since ChatGPT-4o is non-deterministic, we repeat each experiment three times on the same data and report the mean and standard deviation of the results. For the baseline model, we train all networks three times with different random seed initializations, and similarly report the mean and standard deviation of the accuracy.

5 Experimental Results

5.1 Performance of Proposed and Baseline Models

In the following, the performance of the proposed ChatGPT-4o-based model is compared to the performance of the SOTA CNN-based baseline. To this end, we perform in-context learning for the ChatGPT-4o-based model with different number of shots, i.e., 1, 3, and 5. We frame the task as a general audio classification task (cf. System Prompt in Fig. 1) and request both a classification score and an explanation of the decision for the test sample (cf. Test Prompt in Fig. 1). To ensure a fair comparison to the baseline CNN model, we consider training different CNN models for each shot by using only data from the reference speakers selected for the ChatGPT-4o-based model.[2] For completeness, we also consider the results obtained when all the training data is used for the CNN model.

Table 1 presents the obtained results. It can be observed that as expected, both models show a performance improvement as they are exposed to more training data. More importantly, the ChatGPT-4o-based model demonstrates promising results compared to the SOTA baseline. In the following subsections, we analyze the impact of different settings on the performance of the ChatGPT-4o-based model.

5.2 Impact of System Prompt

The system prompt defines the model's behavior by specifying the task and instructing it to classify test samples based on reference samples. As previously mentioned, there are two primary ways to define the task in the system prompt: (i) as an audio classification task without further details as in Section 5.1 or (ii) as a dysarthria classification task, where the model is provided with patient characteristics and asked to classify based on both the reference samples and the given description. In describing dysarthria, we highlight key symptoms such as articulation deficiencies, vowel

[2] Please note that these CNN models use all the data available from the reference speakers, not just the reference samples used for the ChatGPT-4o-based model. This is necessary to provide sufficient training data for the CNN model.

Table 2: Impact of different settings on the performance of the proposed approach under different few-shot scenarios.

Accuracy (%)	1-Shot	3-Shot	5-Shot
Setting from Section 5.1	70.2 ± 3.4	82.1 ± 0.0	85.7 ± 0.0
Dysarthria-specific prompt	60.7 ± 2.9	69.0 ± 1.6	76.2 ± 1.6
Non-detailed response	64.3 ± 0.0	75.0 ± 0.0	79.8 ± 3.4
Raw speech input	52.4 ± 6.7	60.7 ± 5.8	67.9 ± 2.9

distortions, reduced loudness variation, hypernasality, and syllabification issues [1]. To examine the impact of the system prompt on performance, we repeat the ChatGPT-4o-based experiments from Section 5.1 using a system prompt that includes a description of dysarthria.

Table 2 presents the obtained results using a system prompt that describes dysarthria (Dysarthria-specific prompt). For ease of comparison, it also includes the ChatGPT-4o-based results from Section 5.1, where the system prompt frames the task as general audio classification. The results show that using a system prompt specifically describing dysarthria leads to a performance degradation, regardless of the number of shots considered. As described in Section 4.1, we expect the denoised UA-Speech database to contain pathology-unrelated differences between the two groups of speakers that are considerably easier to learn than pathology-discriminant cues. We believe that when a description of the pathology is included in the system prompt, the ChatGPT-based model shifts its focus to these characteristics, rather than relying on spurious pathology-unrelated cues. Therefore, we hypothesize that this degradation in performance is due to the model focusing on genuine pathological cues rather than unintended artifacts. Further investigation is needed to fully understand the network's decision-making process in this context.

5.3 Impact of Non-detailed Response

At test time, we prompt the LLM to classify the test STFT representation into one of the predefined classes. This can be done in two ways: (i) requesting only a classification score, or (ii) asking the model to provide both a score and an explanation for its decision as in Section 5.1. To examine the impact of this choice on performance, we repeat the ChatGPT-4o-based experiments from Section 5.1 requesting only a classification score. The results for this setting are reported in Table 2 (Non-detailed response). We observe a degradation in the performance of the proposed method when it is asked to return only a classification score, compared to when it also provides an explanation for its prediction. This is because the model performs better when it follows a step-by-step reasoning process to generate its response. Similar findings have been reported in previous studies, where the chain-of-thought prompting technique has been shown to improve the performance of LLMs [33].

5.4 Impact of Raw Speech Input

To examine the impact of different input representations on performance, we repeat the experiments from Section 5.1 using raw speech input and the ChatGPT-4o-Audio-Preview model. The results are presented in Table 2 (Raw speech input). As observed, there is a degradation in performance compared to when the STFT representation is used as input. While raw speech inherently contains more information than STFT, we conclude that ChatGPT-4o's vision capabilities are more advanced than its audio processing abilities, resulting in a better performance with STFT input.

6 Future Work

Our experiments demonstrate the in-context learning capabilities of ChatGPT-4o for automatic pathological speech detection. There are several directions that can be explored to further improve the performance of these approaches. First, in this study, the in-context learning samples were selected randomly from the dataset. Although our method achieved performance comparable to SOTA approaches, we believe that a more deliberate selection of in-context samples could close the remaining performance gap and potentially achieve SOTA results.

Additionally, ChatGPT-4o provides explanations alongside classification outputs. These explanations have the potential to enhance the interpretability of pathological speech detection models. Therefore, further investigation is warranted to better understand the model's decision-making process in this context.

7 Conclusion

In this study, we explored the in-context learning capabilities of ChatGPT-4o for automatic pathological speech detection. We evaluated our approach across different shot settings and compared it to a SOTA CNN-based method on the UA-Speech dataset. Our results demonstrate that the proposed method achieves promising and competitive performance, while also providing explainability by detailing the reasoning behind its decisions.

To further analyze our method, we conducted an ablation study. We observed that while ChatGPT-4o is inherently non-deterministic, its performance stabilizes when exposed to more samples, resulting in consistent outputs in higher-shot scenarios. Additionally, we found that explicitly describing the task in the system prompt led to a slight performance degradation, likely because the model focused more on actual pathological cues rather than spurious ones. Finally, our results indicate that requesting an explanation for the model's classification leads to improved accuracy. This aligns with previous findings, where chain-of-thought prompting has been shown to enhance the performance of LLMs.

Our promising results suggest that further research should explore this direction to fully leverage the potential of LLMs for pathological speech detection.

8 Acknowledgments

This work was supported by the Swiss National Science Foundation project no CRSII5_202228 on "Characterisation of motor speech disorders and processes".

References

[1] J. R. Duffy *et al.*, *Motor speech disorders: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2012.

[2] W. Ziegler, I. Aichert, and A. Staiger, "Apraxia of speech: Concepts and controversies," *Journal of speech, language, and hearing research*, vol. 55, no. 5, pp. S1485–S1501, 2012.

[3] I. Kodrasi, M. Pernon, M. Laganaro, and H. Bourlard, "Automatic discrimination of apraxia of speech and dysarthria using a minimalistic set of handcrafted features.," in *Proc. Annual Conference of the International Speech Communication Association*, pp. 4991–4995, Oct. 2020.

[4] I. Kodrasi and H. Bourlard, "Spectro-temporal sparsity characterization for dysarthric speech detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1210–1222, June 2020.

[5] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Subspace-based learning for automatic dysarthric speech detection," *IEEE Signal Processing Letters*, vol. 28, pp. 96–100, Jan. 2020.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[7] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, (Vancouver, Canada), pp. 6645–6649, May 2013.

[8] P. Janbakhshi and I. Kodrasi, "Experimental investigation on STFT phase representations for deep learning-based dysarthric speech detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, (Philadelphia, USA), pp. 6477–6481, May 2022.

[9] P. Janbakhshi and I. Kodrasi, "Supervised speech representation learning for Parkinson's disease classification," in *Proc. ITG Conference on Speech Communication*, (Kiel, Germany), pp. 154–158, Sept. 2021.

[10] P. Janbakhshi and I. Kodrasi, "Adversarial-free speaker identity-invariant representation learning for automatic dysarthric speech classification," in *Proc. Annual Conference of the International Speech Communication Association*, (Incheon, Korea), pp. 2138–2142, Sept. 2022.

[11] K. L. Kadi, S. A. Selouani, B. Boudraa, and M. Boudraa, "Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge," *Biocybernetics and Biomedical Engineering*, vol. 36, pp. 233–247, Jan. 2016.

[12] G. Schu, P. Janbakhshi, and I. Kodrasi, "On using the UA-Speech and TORGO databases to validate automatic dysarthric speech classification approaches," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[13] J. Millet and N. Zeghidour, "Learning to detect dysarthria from raw speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, (Brighton, UK), pp. 5831–5835, May 2019.

[14] M. Amiri and I. Kodrasi, "Test-time adaptation for automatic pathological speech detection in noisy environments," in *2024 32nd European Signal Processing Conference (EUSIPCO)*, pp. 86–90, IEEE, 2024.

[15] M. Amiri and I. Kodrasi, "Suppressing noise disparity in training data for automatic pathological speech detection," in *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 110–114, 2024.

[16] S. Latif, M. Shoukat, F. Shamshad, M. Usama, Y. Ren, H. CuayÃ¡huitl, W. Wang, X. Zhang, R. Togneri, E. Cambria, and B. W. Schuller, "Sparks of large audio models: A survey and outlook," 2023.

[17] M. Awais, M. Naseer, S. Khan, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, and F. S. Khan, "Foundation models defining a new era in vision: a survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[18] H. O. Shahreza and S. Marcel, "Foundation models and biometrics: A survey and outlook," *Authorea Preprints*, 2025.

[19] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[20] M. He and P. N. Garner, "Can chatgpt detect intent? evaluating large language models for spoken language understanding," *arXiv preprint arXiv:2305.13512*, 2023.

[21] A. Komaty, H. O. Shahreza, A. George, and S. Marcel, "Exploring ChatGPT for Face Presentation Attack Detection in Zero and Few-Shot in-Context Learning," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025.

[22] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[23] "Noise reduced uaspeech dysarthria dataset." https://www.kaggle.com/datasets/aryashah2k/noise-reduced-uaspeech-dysarthria-dataset/.

[24] M. Amiri and I. Kodrasi, "Adversarial robustness analysis in automatic pathological speech detection approaches," in *Proc. Annual Conference of the International Speech Communication, Rhodes Islands, Greece*, pp. 1415–1419, 2024.

[25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Annual Conference on Neural Information Processing Systems*, (Virtual), pp. 12449–12460, Dec. 2020.

[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, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, pp. 1194–1198, 2021.

[27] F. Javanmardi, S. Tirronen, M. Kodali, S. R. Kadiri, and P. Alku, "Wav2vec-based detection and severity level classification of dysarthria from speech," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, 2019.

[29] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[30] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23716–23736, 2022.

[31] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, M. Zejun, and C. Zhang, "Salmonn: Towards generic hearing abilities for large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[32] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Interspeech 2008*, pp. 1741–1744, 2008.

[33] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.