# Medical Dialogue Audio Transcription: Dataset and Benchmarking of ASR Models

**Aline E. Gassenn[1], Luis G. M. Andrade[2], Douglas Teodoro[3], Jose F. Rodrigues-Jr[1]**

[1] Institute of Mathematics and Computer Sciences
University of São Paulo (USP), São Carlos – SP – Brazil
[2] Clinical Hospital, Faculty of Medicine of Botucatu
São Paulo State University (UNESP), Botucatu – SP – Brazil
[3] University of Geneva – Geneva, Switzerland

{aline.gassenn,junio}@usp.br

douglas.teodoro@unige.ch, gustavo.modelli@unesp.br

***Abstract.*** *The development of Automatic Speech Recognition (ASR) technologies for healthcare applications is hindered by the limited availability of publicly accessible speech corpora that reflect both natural medical dialogues and the acoustic conditions typically found in clinical environments. In this study, we present the creation and characterization of MedDialogue-Audio, a new synthetic English-language corpus designed to address this gap. The dataset was derived from the MedDialog-EN transcription set and enriched through a multi-stage processing pipeline that involved text normalization with a large language model, speech synthesis, and the controlled addition of both white noise and hospital ambient sounds. We provide descriptive statistics for the corpus, which comprises more than 10,000 dialogues, as well as benchmarking results from leading ASR models. The experiments assess transcription performance across varying signal-to-noise ratios and establish baseline metrics to support future research in this field.*

## 1. Introduction

The use of artificial intelligence (AI) systems, such as Large Language Models (LLMs) and speech processing technologies, is playing an increasingly prominent role in healthcare applications. These solutions aim to optimize clinical documentation, support professional activities, and improve access to information [Arora et al. 2025, Lee et al. 2022]. In this context, Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) technologies have emerged as promising tools, particularly for automating records and documenting interactions between physicians and patients [Nurfadhilah et al. 2021].

The development and evaluation of such technologies, however, fundamentally depend on the availability of speech datasets that are representative of the target domain. In the healthcare field, the availability of public datasets remains notably limited, primarily due to ethical and legal privacy constraints and the high complexity associated with collecting and anonymizing clinical data [Le-Duc 2024]. As a result, many existing resources are access-restricted, such as the Brazilian Portuguese corpus described in [Gonçalves et al. 2024].

To mitigate this scarcity, the scientific community has invested in the creation of public corpora. Several initiatives focus on specific languages, including Vietnamese (VietMed [Le-Duc 2024]) and Indonesian (BPPT [Nurfadhilah et al. 2021]). For English, synthetic data generation has emerged as a strategy to overcome privacy barriers, as exemplified by United-MedASR [Banerjee et al. 2024]. Despite their scale, this and similar approaches present critical limitations: the audio samples often correspond to technical documents or isolated terms, failing to simulate conversational exchanges and typically lacking background noise.

More recently, MultiMed [Le-Duc et al. 2025] introduced a large-scale multilingual dataset for medical ASR, comprising real clinical conversations in five languages. The speech data were collected from publicly available medical videos on YouTube and manually curated to ensure transcription quality. However, its focus on multilingual, real-world recordings contrasts with our approach, which leverages synthetic generation to systematically simulate English-language medical dialogues under controlled noise conditions. Our method directly addresses the lack of publicly available English datasets with realistic conversational structure and ensures reproducibility for benchmarking purposes.

These gaps highlight two main challenges for the reliable deployment of ASR systems in real-world clinical settings. The first concerns the accurate transcription of specialized medical terminology, which extends beyond the vocabulary of general-purpose models [Lee et al. 2022]. The second refers to robustness against signal quality degradation caused by noise, a frequent occurrence in hospital and telemedicine environments. Together, these factors underscore the need for datasets that more realistically capture both the linguistic and acoustic conditions of these contexts.

To address these two challenges, this paper introduces MedDialogue-Audio, a new public corpus of English-language medical dialogues designed to support ASR research in clinical environments. The dataset was derived from the textual corpus MedDialog-EN [Zeng et al. 2020, Tang et al. 2023] through a structured pipeline that involved: (i) text preprocessing and linguistic correction of the source material; (ii) audio generation via speech synthesis; and (iii) controlled addition of acoustic noise at multiple intensity levels. The result is a resource that combines the linguistic variability of clinical dialogues with diverse acoustic conditions, enabling a more realistic assessment of ASR robustness. The dataset is publicly available at *https://huggingface.co/datasets/aline-gassenn/MedDialog-Audio*.

The remainder of this paper is organized as follows: Section 2 describes the source textual corpus; Section 3 details the methodological process for dataset construction; Section 4 presents a quantitative analysis of the generated data; Section 5 discusses a benchmarking case study; Section 6 provides concluding remarks; and finally, Section 7 outlines the limitations of the models used in the dataset construction process.

## 2. Med-Dialog

The MedDialog corpus, introduced by [Zeng et al. 2020], is a large-scale dataset that plays a pivotal role in research on dialogue systems within the healthcare domain. This study is based on its English-language version, MedDialog-EN, which consists of approximately 260,000 dialogues, totaling 510,000 utterances and 44.53 million tokens. The data, collected from telehealth platforms such as *icliniq.com* and *healthcaremagic.com*

between 2008 and 2020, span 96 medical specialties and are structured to detail the patient's condition, followed by the subsequent conversational turns.

A later version of this corpus, focused on enhancing terminological accuracy, was presented by [Tang et al. 2023]. This enriched version applies methodologies for the identification and contextualization of medical terms to improve semantic consistency and serves as the starting point for the methodology adopted in this work. The decision to use this preprocessed corpus is grounded in the need for a text base with high clinical fidelity, aiming to minimize the propagation of semantic ambiguities into the subsequent speech synthesis stage.

## 3. Methodology

The methodology for constructing the MedDialogue-Audio dataset followed a sequential processing pipeline, divided into two main stages: the preprocessing and enrichment of the source textual corpus, and the subsequent generation of the audio corpus with noise addition. The complete workflow of this pipeline is illustrated in Figure 1, and each stage is detailed in the following subsections.

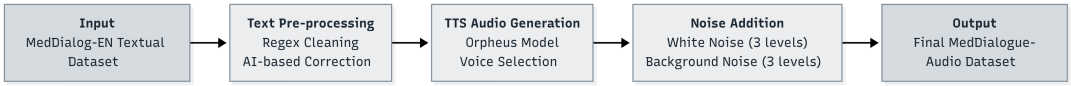



**Figure 1. Workflow of the methodological pipeline for creating the MedDialogue-Audio dataset.**

### 3.1. Data Collection and Preprocessing

The starting point for this methodology was the MedDialog-EN corpus, specifically its version with terminological enrichment [Tang et al. 2023]. The first processing step involved reconstructing the conversational structure by concatenating patient utterances (source.txt) and doctor utterances (target.txt) into complete dialogue interactions. A subsequent text normalization phase was performed using regular expressions to remove artifacts that could introduce inconsistencies in speech synthesis, such as URLs, email addresses, and textual metadata from the source platforms.

The enrichment phase was mediated by the *gpt-4o-mini* language model, which was guided by the prompt shown in Figure 2. Aimed at maximizing both the intelligibility and naturalness of the audio, the model was tasked with semantic cleaning and canonical normalization of the text. These operations included correcting grammatical and technical inaccuracies, as well as expanding abbreviations and numerals. Additionally, to enable the selection of synthesis voices compatible with the speaker profile, the model performed demographic attribute extraction for the patient (gender and age group). The output of the entire process was consolidated into a JSON object, ensuring both structural integrity and interoperability of the generated data.

### 3.2. Audio Synthesis and Acoustic Augmentation

The synthesis of textual dialogues into audio was performed using the Orpheus TTS model [Canopyai 2025], specifically the *canopylabs/orpheus-3b-0.1-ft* variant. The selection of this model was based on its Speech-LLM architecture, pretrained on a corpus

```
system_msg = (
    """
    Below is a conversation between a patient and a doctor. The conversation is
        structured as:
    "patient: [text]. doctor: [text]".

    Please perform the following tasks:

    1. Correct any grammatical errors, typos, unit measurement mistakes, and
        inaccuracies in technical nomenclature.
    2. Replace all abbreviations with their full, original words.
    3. Extract and return the corrected text separately for the patient and the doctor.
       Use the keys "return_text_patient" for the patient's text and "return_text_doctor
           " for the doctor's text.
    4. Determine the patient's gender, choosing one of the following classes: "male",  "
        female", or "neutral".
    5. Determine the patient's age group, selecting one of the following classes: "child
        ", "teenage", "adult", "elderly", or "neutral".

    Return your output as a JSON object with the keys:
    - "return_text_patient": the corrected text of the patient,
    - "return_text_doctor": the corrected text of the doctor,
    - "return_gender": the identified gender,
    - "return_age": the identified age group.

    Ensure your output is in valid JSON format.
    """
)
```

**Figure 2. Prompt submitted to the model for the correction and classification task.**

exceeding 100,000 hours of audio, which endows it with the capability to generate speech with high fidelity and natural prosody. Additionally, its permissive Apache-2.0 license was a key practical criterion, ensuring broad usability and enabling redistribution of the derived dataset.

The generation process began with the segmentation of dialogues into units of up to 60 words, an optimization intended to maintain synthesis quality. The audio segments for each speaker were then concatenated separately, resulting in the creation of two distinct *.wav* files per dialogue: one for the patient and another for the doctor. Voice assignment was parameterized to ensure both diversity and clarity: for patients, selection was guided by the inferred gender, using a set of five female voices (tara, leah, jess, mia, zoe) and three male voices (leo, dan, zac). For doctors, voice selection was made from the same overall pool, with the restriction that it differed from the voice assigned to the patient, thereby ensuring clear speaker differentiation. This process was applied to 10,534 dialogues, resulting in a corpus comprising 21,068 individual audio files.

To enhance the dataset's robustness and its applicability in real-world scenarios, an acoustic augmentation step was implemented, in which each original audio file generated six noisy variations. Two noise modalities were employed, with controlled intensity levels to achieve predefined Signal-to-Noise Ratio (SNR) targets:

- White Noise: Added at three levels, corresponding to 2%, 6%, and 10% of the original audio signal amplitude.
- Background Noise: Introduced at three levels (20%, 40%, and 60% of the signal amplitude), using samples from the *Hospital Ambient Noise Dataset* [Ali and Shuvo 2021, Ali et al. 2023]. To ensure variability, a different

noise sample was selected at each level. The dataset contains diverse hospital-related sounds, including background child crying, door movements, conversations in reception and waiting areas, and equipment operation noises.

The combination of high-fidelity synthesis with controlled acoustic augmentation is a cornerstone of this work. The former ensures linguistic content clarity, while the latter introduces systematic signal degradations, enabling the dataset to serve as a benchmark for evaluating speech systems under adverse and clinically representative conditions.

### 3.3. Data Description

The dataset was constructed through speech synthesis from a preprocessed textual corpus, followed by acoustic augmentation. The dataset comprises 10,534 dialogues. Since each dialogue results in one audio file for the patient and another for the doctor, and each of these has six noise-augmented variations, the corpus totals 147,476 audio files. This multi-faceted structure enables the evaluation of models under various controlled acoustic conditions.

The organization of the audio files follows a systematic naming convention designed to facilitate the identification of the source and characteristics of each segment:

`[DIALOGUE_ID]_[SPEAKER][AUDIO_TYPE][NOISE_LEVEL].wav`

Where:

- DIALOGUE_ID: A unique numeric identifier for each dialogue.
- SPEAKER: Indicates the speaker, with `1` for the patient and `2` for the doctor.
- AUDIO_TYPE: Characterizes the nature of the audio: `o` for original (noise-free), `w` for white noise, and `b` for hospital background noise.
- NOISE_LEVEL: Indicates the noise intensity level, expressed as a percentage. For original audio (`o`), this field takes the value `00`, indicating no added noise. For white noise samples (`w`), the levels are 2%, 6%, or 10%. For hospital background noise (`b`), the levels correspond to 20%, 40%, or 60%.

In addition to the audio corpus, the repository includes a comprehensive metadata file, *metadata.csv*, which documents the properties of the noise-free audio samples. Each row in this file corresponds to one original recording and contains both acoustic descriptors and the associated transcription. These metadata entries are shared across all augmented versions derived from the same original audio.

The metadata file includes the following columns:

- filename: The exact name of the audio file.
- duration_s: Duration of the recording, in seconds.
- mean_rms_energy: Mean root-mean-square energy of the signal.
- mean_f0_hz: Mean fundamental frequency (F0), in Hertz.
- mean_spectral_centroid_hz: Mean spectral centroid, in Hertz.
- hnr_db: Harmonic-to-noise ratio (HNR), expressed in decibels.
- transcription: Canonically normalized textual transcription of the utterance.

This metadata enables downstream tasks such as acoustic analysis, supervised model training, and benchmarking of ASR systems under both clean and adverse audio conditions, thereby enhancing the utility of the corpus for experimental reproducibility and robust system evaluation.

## 4. Exploratory Analysis of the Dataset

This section presents the empirical results of the study, organized into two complementary analyses: the quantification of the impact of text processing and the characterization of the properties of the final acoustic corpus, considering the noise-free version.

### 4.1. Text Augmentation Validation

The magnitude of textual modifications introduced by the enrichment pipeline was quantified using the Levenshtein distance, a metric commonly employed in sequence comparison tasks that indicates the minimum number of insertions, deletions or substitutions necessary to transform one string into another [Devatine and Abraham 2024]. Patient utterances presented a higher mean distance of 80.69, while doctor utterances showed a mean distance of 70.02. This difference reflects the greater presence of orthographic errors and informal constructions in the original patient texts. It is important to note that these modifications were applied to a textual corpus originally sourced from telemedicine platforms, not to transcribed speech, and were designed to improve orthographic consistency and ensure compatibility with text-to-speech systems, while preserving the natural variability characteristic of spontaneous language.

The effectiveness of the spelling correction process was quantified based on the variation in the frequency of errors identified using the PySpellChecker library [Norvig 2025], which performs verification against a standard English dictionary. This tool compares each term in the text against its lexical database and counts as a spelling error any word not recognized. The results, presented in Figure 3, indicate a 68.7% reduction in the number of errors in patient texts (from 26,117 to 8,176) and a 61.0% reduction in doctor texts (from 21,775 to 8,495). The residual volume of errors is largely attributable to the use of highly specialized domain vocabulary, including technical jargon and proper names, which are not covered by the reference spell-checking dictionary used in the correction process.

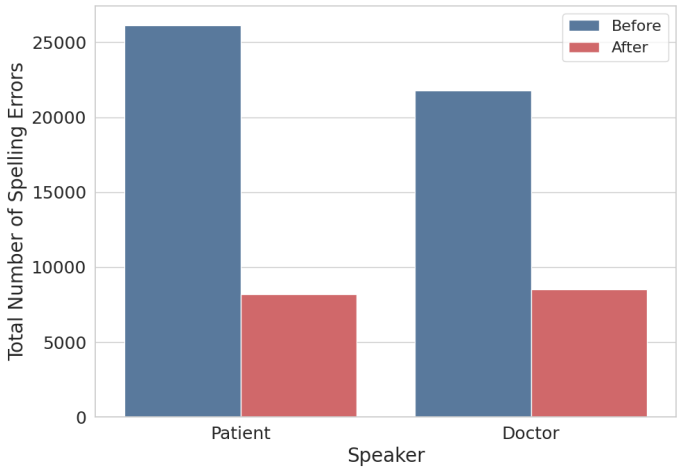



**Figure 3. Number of spelling errors identified before and after the application of the text enrichment pipeline.**

### 4.2. Acoustic Corpus Characterization

The audio corpus, in its original noise-free version, totals 136.68 hours, covering 21,068 utterances. Table 1 presents a summary of the main descriptive statistics for this set.

The complexity of the corpus is most evident in the distribution profile of the Fundamental Frequency (F0), which exhibits a distinctly bimodal pattern (Figure 4a). Density peaks centered around approximately 140 Hz and 190 Hz are consistent with the presence of one male vocal group and one female vocal group, a direct consequence of the synthesis methodology employed. This multimodality also extends to other metrics: both RMS Energy (Figure 4c) and the Harmonics-to-Noise Ratio (HNR) (Figure 4d) display bimodal distributions, indicating the existence of multiple intensity profiles and distinct phonation quality clusters (ranging from breathy to clear voice characteristics), respectively.

In contrast to this heterogeneity, other metrics point to an underlying consistency in the synthesis process. The Spectral Centroid shows a unimodal profile with a peak around 2100 Hz, suggesting a stable overall spectral ”brightness”. Similarly, utterance duration (Figure 4b) follows a unimodal distribution with pronounced positive skewness (peak at approximately 19 seconds), indicating a central tendency in dialogue length, despite the presence of a long tail of extended interactions.

**Table 1. Descriptive statistics of acoustic metrics for the original (noise-free) audio corpus.**

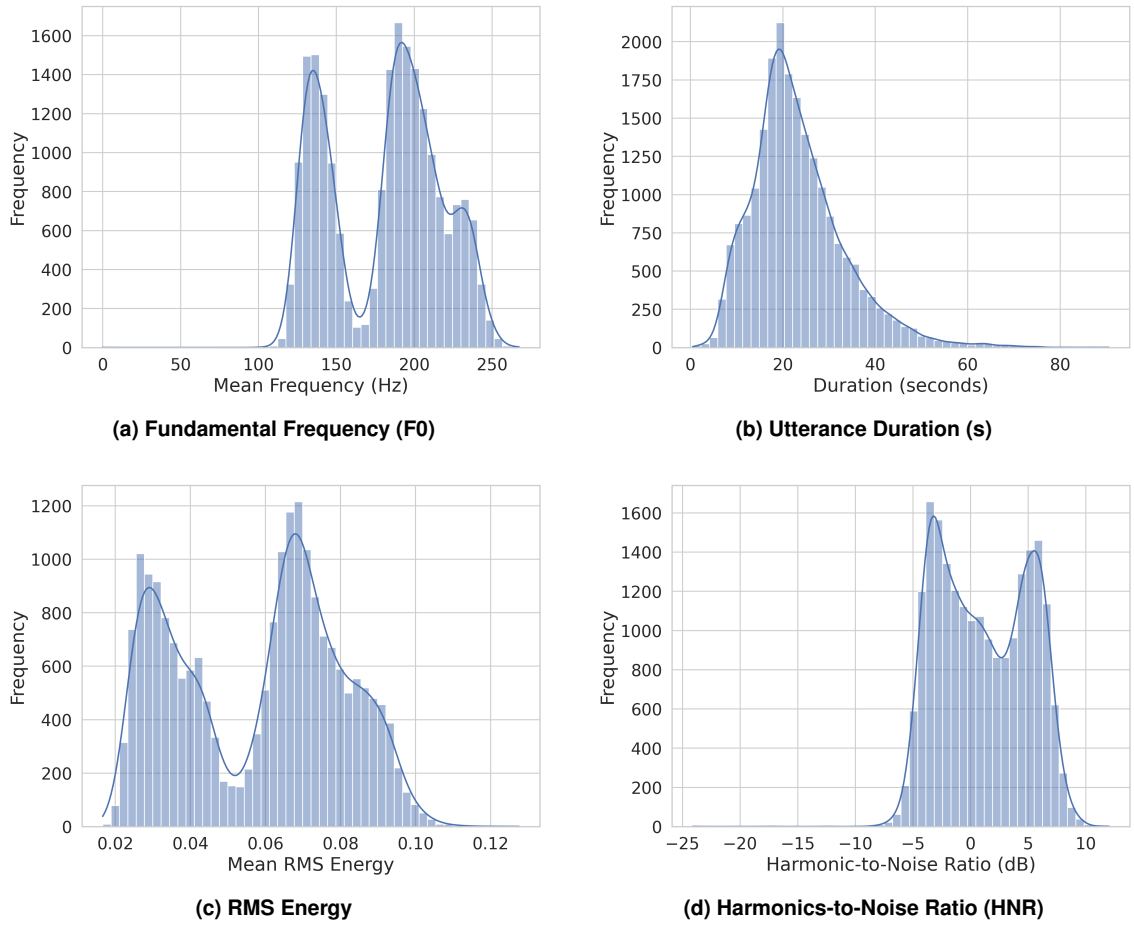
| Metric | Mean | Standard Deviation | Min | Max |
|---|---|---|---|---|
| Duration (s) | 23.31 | 10.11 | 0.51 | 90.54 |
| RMS Energy | 0.058 | 0.022 | 0.017 | 0.128 |
| Fundamental Frequency (Hz) | 181.52 | 30.44 | 108.07 | 267.14 |
| Spectral Centroid (Hz) | 2228.07 | 287.71 | 1407.62 | 3956.16 |
| HNR (dB) | 1.06 | 3.89 | -24.11 | 12.01 |

## 5. ASR Model Benchmarking

To demonstrate the applicability of MedDialogue-Audio and to establish a quantitative reference point, a benchmarking case study was conducted. The aim of this analysis is not to provide an exhaustive evaluation of Automatic Speech Recognition (ASR) models, but rather to use them as diagnostic tools to characterize the complexity of the corpus, with particular attention to the impact of noise variations on transcription performance.

The experiments were conducted on a 10% subset of the corpus, covering both the original audio files and all their noisy versions. This material was processed by three state-of-the-art ASR systems, all used in their pretrained configurations (zero-shot scenario). The selected models represent different learning paradigms: the supervised encoder-decoder architecture Whisper (base) [Radford et al. 2023], and two self-supervised approaches, Wav2Vec 2.0 (base-960h) [Baevski et al. 2020] and HuBERT (large-ls960-ft) [Hsu et al. 2021].

The quality of the transcriptions was assessed using three complementary metrics. Lexical accuracy was measured by the Word Error Rate (WER), where lower values indicate better performance. The ability to recognize domain-specific vocabulary was evaluated using the Medical Term Recognition Accuracy (MTRA). Finally, semantic fidelity between the hypothesis and the reference was quantified using the BERTScore (F1), which reflects semantic similarity.




**(a) Fundamental Frequency (F0)** **(b) Utterance Duration (s)**

**(c) RMS Energy** **(d) Harmonics-to-Noise Ratio (HNR)**

**Figure 4. Distribution profiles for key acoustic metrics. Figures (a), (c), and (d) highlight the dataset's heterogeneity through multimodal distributions, while Figure (b) illustrates the skewed distribution of utterance durations.**

Figures 5 to 7 present the results obtained for each metric, discriminated across the different noise conditions considered.

The results demonstrate a negative correlation between noise level and model performance across all evaluated systems, with substantial variations in degradation magnitude among the different architectures. The lexical accuracy analysis (Figure 5) reveals a robustness hierarchy, where the Whisper model consistently outperforms others across all test conditions, showing a gradual performance decline as noise intensity increases. In contrast, Wav2Vec 2.0, while initially competitive, suffers a sharp degradation under more severe noise levels.

This disparity between models becomes even more pronounced when analyzing domain-specific term recognition (Figure 7). The Whisper MTRA shows high resilience, whereas Wav2Vec 2.0 exhibits an abrupt decline, reaching a residual accuracy of only 2.5%. The qualitative microanalysis in Table 2 illustrates this phenomenon, highlighting severe lexical failures by Wav2Vec 2.0 in transcribing technical terms. The semantic similarity analysis (Figure 6) reinforces these trends.

These results establish a set of baselines that not only validate the utility of MedDialogue-Audio for investigations into ASR robustness but also quantify the inherent

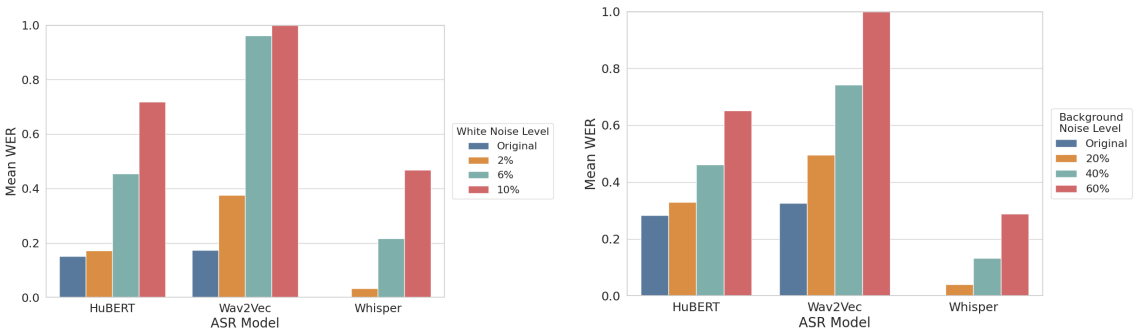



**Figure 5. Word Error Rate (WER) under White Noise (left) and Background Noise (right) conditions. Notably, for the Whisper model, the WER for the original audio was so low that it was not visually perceptible on the adopted graph scale, resulting in the apparent absence of the corresponding bar.**

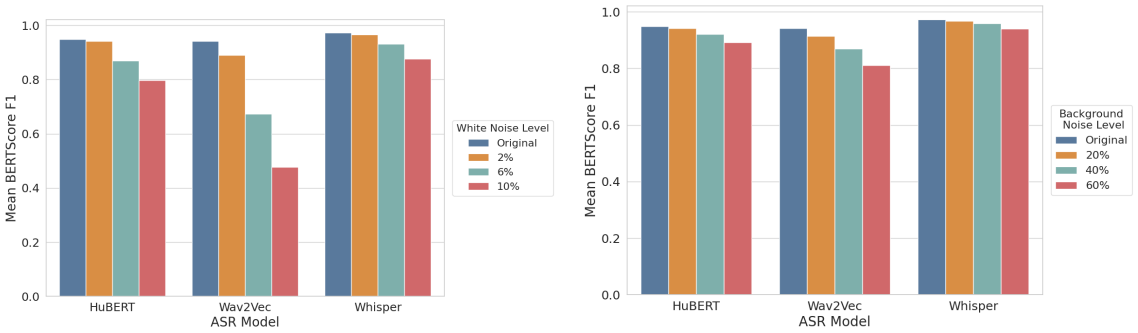



**Figure 6. Semantic similarity measured by BERTScore (F1) under White Noise (left) and Background Noise (right) conditions.**

challenges of speech processing in the medical domain under adverse acoustic conditions.

## 6. Conclusion

This work introduced MedDialogue-Audio, a synthetic public corpus of English-language medical dialogues designed to address the shortage of resources for research in Automatic Speech Recognition (ASR) in noisy clinical contexts. The corpus was derived from MedDialog-EN through a pipeline that integrates linguistic enrichment using an LLM, high-fidelity speech synthesis, and data augmentation with controlled acoustic noise.

The pipeline validation demonstrated a substantial reduction in errors in the source text. Subsequent benchmarking experiments established baseline performances for prominent ASR models, revealing a direct correlation between noise intensity and performance degradation. These results not only quantify the robustness of current architectures but also demonstrate the utility of the corpus in simulating the challenges of real clinical environments.

By being made publicly available, MedDialogue-Audio offers a new resource for the training and rigorous evaluation of speech technologies. Future research directions include expanding the dataset and applying it to the fine-tuning of ASR models, with the goal of improving automatic transcription accuracy in the healthcare domain.

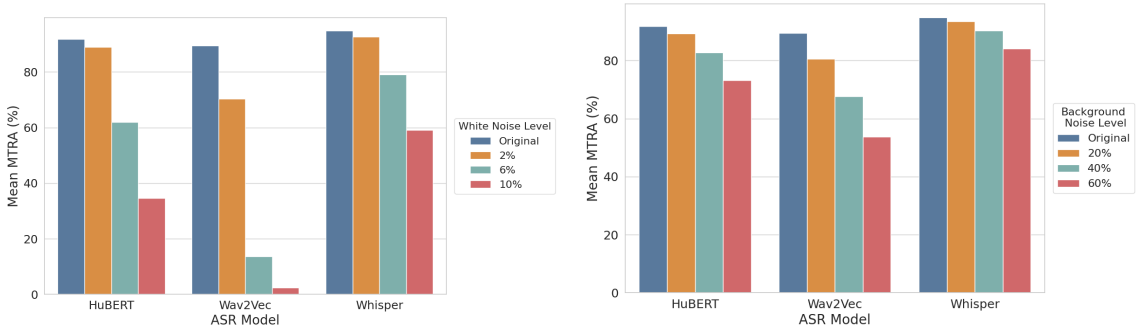



**Figure 7. Medical Term Recognition Accuracy (MTRA) under White Noise (left) and Background Noise (right) conditions.**

**Table 2. Example of ASR transcription for a medical statement. Differences from the original text are underlined.**

| Source | Transcription |
|---|---|
| Original | You may have a renal stone or dysfunctional uterine bleeding. Please get an abdominal ultrasound and a routine urine test, and let me know the results. |
| Whisper | you may have a renal stone or dysfunctional uterine bleeding. please get an abdominal ultrasound and a routine urine test. and let me know the results. |
| Wav2Vec 2.0 | you may have a reedal stone or disfunctional uteran bleeding_ please get an adominable ultresound in a routine earan test_ and let me know the results_ |
| HuBERT | you may have a renal stone or disfunctional uterin bleeding_ please get an abdominal ultra sound and a routine urin test and let me know the results_ |

## 7. Limitations

Although MedDialogue-Audio represents a relevant contribution to ASR research in clinical scenarios, it presents some limitations that must be considered. As a synthetic dataset, it is subject to biases arising from the language model *gpt-4o-mini* used in the text normalization stage. The inference of demographic attributes such as gender and age group, based solely on textual input (see Figure 2), may reproduce stereotypical associations implicitly present in the model's training data.

The text-to-speech system adopted, Orpheus TTS, while offering high-fidelity audio generation, does not support expressive prosody. As a result, the synthesized audio lacks vocal modulations associated with emotional or physical states such as pain, crying, or empathy. In addition, the limited number of available speaker voices constrains the acoustic variability of the corpus.

Another relevant limitation concerns the absence of human validation. All evaluations were carried out using automatic ASR models. Manual verification of transcriptions, demographic labels, and semantic coherence was not performed due to constraints related to time and availability of expert annotators. Nonetheless, human evaluation remains fundamental for assessing the realism and reliability of synthetic corpora and should be

incorporated in future iterations of this work.

Future directions include the integration of human validation procedures, the annotation of medical specialties and diagnostic categories, and the application of the corpus to the development of summarization models for automatic generation of structured clinical records.

## Acknowledgements

This study was partially financed by the Sao Paulo Research Foundation (FAPESP – grants 2024/04761-0, 19/07665-4, 2024/13328-9, and 23/18026-8), the National Research Council (CNPq 307946/2021-5, and 304805/2025-4), and the Coordination for Higher Education Personnel Improvement (CAPES – grant 001).

## References

Ali, S. N. and Shuvo, S. B. (2021). Hospital ambient noise dataset.

Ali, S. N., Shuvo, S. B., Al-Manzo, M. I. S., Hasan, A., and Hasan, T. (2023). An end-to-end deep learning framework for real-time denoising of heart sounds for cardiac disease detection in unseen noise. *IEEE Access*, 11:87887–87901.

Arora, R. K., Wei, J., Hicks, R. S., Bowman, P., Quiñonero-Candela, J., Tsimpourlas, F., Sharman, M., Shah, M., Vallone, A., Beutel, A., Heidecke, J., and Singhal, K. (2025). Healthbench: Evaluating large language models towards improved human health.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 12449–12460. Curran Associates, Inc.

Banerjee, S., Agarwal, A., and Ghosh, P. (2024). High-precision medical speech recognition through synthetic data and semantic correction: United-medasr. *arXiv preprint arXiv:2412.00055*.

Canopyai (2025). Canopyai/orpheus-tts: Towards human-sounding speech.

Devatine, N. and Abraham, L. (2024). Assessing human editing effort on llm-generated texts via compression-based edit distance. *arXiv preprint arXiv:2412.17321*.

Gonçalves, Y. T., Alves, J. V. B., Sá, B. A. D., da Silva, L. N., de Macedo, J. A. F., and da Silva, T. L. C. (2024). Speech recognition models in assisting medical history. In *Proceedings of the 39th Brazilian Symposium on Databases (SBBD)*, pages 485–497, Florianópolis, SC, Brazil.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Le-Duc, K. (2024). VietMed: A dataset and benchmark for automatic speech recognition of Vietnamese in the medical domain. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17365–17370, Torino, Italia. ELRA and ICCL.

Le-Duc, K., Phan, P., Pham, T.-H., Tat, B. P., Ngo, M.-H., Nguyen-Tang, T., and Hy, T.-S. (2025). MultiMed: Multilingual medical speech recognition via attention encoder decoder. In Rehm, G. and Li, Y., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1113–1150, Vienna, Austria. Association for Computational Linguistics.

Lee, S.-H., Park, J., Yang, K., Min, J., and Choi, J. (2022). Accuracy of cloud-based speech recognition open application programming interface for medical terms of korean. *Journal of Korean Medical Science*, 37(18).

Norvig, P. (2025). Pyspellchecker: Pure python spell checking library.

Nurfadhilah, E., Jarin, A., Ruslana Aini, L., Pebiana, S., Santosa, A., Teduh Uliniansyah, M., Butarbutar, E., Desiani, and Gunarso (2021). Evaluating the bppt medical speech corpus for an asr medical record transcription system. In *2021 9th International Conference on Information and Communication Technology (ICoICT)*, pages 657–661.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 28492–28518. PMLR.

Tang, C., Zhang, H., Loakman, T., Lin, C., and Guerin, F. (2023). Terminology-aware medical dialogue generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zeng, G., Yang, W., Ju, Z., Yang, Y., Wang, S., Zhang, R., Zhou, M., Zeng, J., Dong, X., Zhang, R., Fang, H., Zhu, P., Chen, S., and Xie, P. (2020). MedDialog: Large-scale medical dialogue datasets. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.