Evaluating Transformer-Based Architectures for Simultaneous Audio Speech Transcription and Background Audio Captioning

João Vitor R. da Silva[1], Francisco de Assis Boldt[2], Luis A. Souza Jr[1], Mariella Berger[2], Anselmo Frizera[1], Alberto F. De Souza[1], Thiago Oliveira-Santos[1], Claudine Badue[1]

[1]Laboratório de Computação de Alto Desempenho
Universidade Federal do Espírito Santo – Vitória – ES – Brazil

[2]Instituto Federal do Espírito Santo – ES – Brazil

joao.silva.17@edu.ufes.br, franciscoa@ifes.edu.br,
la.souza@inf.ufes.br, mariella.andrade@ifes.edu.br,
anselmo.frizera-neto@ufes.br, alberto@lcad.inf.ufes.br,
todsantos@inf.ufes.br, claudine@lcad.inf.ufes.br

Abstract. *This study evaluates transformer-based models for simultaneous speech transcription and background audio captioning in mixed audio scenarios. Using Whisper for speech and Prompteus for environmental sounds, the models are tested on the Clotho Voice dataset, which combines Portuguese speech (Common Voice 5.1) and environmental audio (Clotho 2.1). Results using WER and FENSE metrics show that each model performs well in its domain but degrades with overlapping signals. Whisper is robust to moderate noise, while Prompteus struggles with dominant speech. The findings highlight the need for hybrid approaches to enable reliable, context-aware audio processing in complex environments.*

Resumo. *Este estudo avalia modelos baseados na arquitetura de Redes Neurais Transformers para transcrição de fala e descrição de áudio de fundo simultâneas em cenários com sinais de áudio mistos. Utilizando o Whisper para fala e o Prompteus para sons ambientais, os modelos foram testados no conjunto de dados Clotho Voice, que combina fala em português (Common Voice 5.1) e sons ambientais (Clotho 2.1). Os resultados, obtidos por meio das métricas WER e FENSE, mostram que cada modelo apresenta bom desempenho em sua área de especialização, mas sofre degradação quando há sobreposição de sinais. O Whisper se mostra robusto a ruídos moderados, enquanto o Prompteus apresenta dificuldades quando a fala é dominante. Os achados destacam a necessidade de abordagens híbridas para viabilizar um processamento de áudio confiável e sensível ao contexto em ambientes complexos.*

1. Introduction

Generating natural language descriptions for audio clips represents a significant challenge in audio processing due to the need to capture diverse audio elements and their interactions, which often involve complex spatial and temporal dynamics not typically addressed in tasks like speech-to-text [Mei et al. 2023, Kim et al. 2019a]. Unlike speech-to-text, which focuses on transcribing spoken words, audio captioning aims

to describe the broader context of an audio clip, capturing prominent sounds, music, and environmental noise [Stöter et al. 2019, Gemmeke et al. 2017a, Lane et al. 2015]. This task has diverse practical applications, such as assisting individuals with hearing impairments, multimedia content retrieval, and audio analysis for security surveillance [Crocco et al. 2016, Czyżewski et al. 1998, Wang et al. 2000].

This research addresses a critical problem in the field of robotics applied to industrial environments: the reliable interpretation of voice commands in Portuguese amidst intense and varied background noise. In industrial settings, the simultaneous transcription of speech and detection of environmental sounds poses significant technical challenges, as noise can both interfere with speech clarity and provide crucial contextual information, such as warning alarms or machine failure signals. In this context, the study proposes a methodology based on Transformer architectures for the simultaneous processing of speech and background sounds, aiming to enable robots to accurately understand and respond to spoken commands even under adverse conditions. The approach seeks to balance two complementary aspects: the precision of speech transcription and the contextual interpretation of non-verbal sounds. Specifically, the research investigates the performance degradation of two advanced neural networks when processing composite audio inputs: Whisper, specialized in speech transcription, and Prompteus, designed for general audio captioning. Both models were evaluated for their ability to handle realistic mixtures of speech and noise, simulating typical industrial scenarios.

The subsequent sections delve into the literature on transformer-based models for speech transcription and background sound description (Section 2), outline the proposed methodology and dataset construction (Section 3), present the experimental results for both models (Section 4), and provide a comprehensive discussion of their performance and limitations (Section 5). Finally, the paper concludes with a summary of key findings and avenues for future research, highlighting the potential for further advancements in multimodal, real-time audio processing.

2. Related Work

Recent research has emphasized enhancing general audio captioning through Transformers, Large Language Models (LLMs), and their derivatives. Notable among these efforts is the study by Kadlčík et al. [Kadlčík et al. 2023], which employed a Whisper-based tool for general audio description. This research highlights the difficulty of collecting high-quality data for general audio detection while maintaining the refined capabilities of transformer-based architectures. Specialized models for environmental audio description may lose transcription capabilities inherited from their original Whisper checkpoints. Maintaining transcription capabilities is crucial for practical applications where both environmental context and accurate speech recognition are required, such as in industrial settings or assistive technologies. This trade-off impacts usability, as a loss of transcription ability may hinder the seamless integration of audio description into workflows reliant on both speech and environmental sound interpretation.

In contrast to image and video captioning, which have long been subjects of extensive research in the field of computer vision, the task of speech audio captioning has only begun to garner interest more recently. Studies in this area include those by Drossos et al. [Drossos et al. 2017], Kadlčík et al. [Kadlčík et al. 2023], and Mei et

al. [Mei et al. 2021], with a surge in attention starting in 2020 and a variety of methods proposed thereafter, as evidenced by works from Kim et al. [Kim et al. 2019b], Wu et al. [Wu et al. 2019], Ikawa et al. [Ikawa and Kashino 2019], and again Kadlčík et al. [Kadlčík et al. 2023]. Audio captioning is often approached as a sequence-to-sequence problem, where the prevalent methods employ an encoder-decoder architecture. In this setup, the decoder produces words based on the audio features that are extracted and processed by the encoder. The issue of captioning was initially tackled using recurrent neural networks equipped with attention mechanisms[Drossos et al. 2017, Kim et al. 2019b, Mei et al. 2021]; however, these architectures may exhibit limitations in modeling long-term temporal dependencies within audio signals.

Recent developments in Transformers, Large Language Models (LLMs), and their derivatives have been introduced to address the description of audio samples. Kadlčík et al. [Kadlčík et al. 2023] have finetuned a Whisper-based tool dedicated to the general task of audio description. To achieve this, the authors fine-tuned the Whisper-based model using three distinct sound-captioned datasets: Clotho [Drossos et al. 2020], AudioSet [Gemmeke et al. 2017b], and AudioCaps [Kim et al. 2019b]. Following the correct re-sampling and processing of the datasets, the authors succeeded in refining a model that is fully specialized in environmental and general audio descriptions, at the expense of the speech transcription features present in the original Whisper checkpoint before tuning.

2.1. Contributions

This study offers significant contributions to the field of audio processing with mixed inputs (speech and background sounds), particularly in noisy conditions and real-world scenarios. The main contributions are as follows: Development of the Clotho Voice dataset – a novel resource for mixed audio analysis: A new dataset was created through the controlled fusion of speech recordings from Common Voice 5.1 and environmental sounds from Clotho 2.1. This methodology enables the realistic simulation of scenarios where speech and background sounds coexist, providing a valuable resource for future research in audio captioning and simultaneous transcription tasks. Systematic evaluation of Transformer-based architectures: The study examines the performance of two state-of-the-art models—Whisper, designed for speech transcription, and Prompteus, tailored for general audio captioning—when exposed to composite audio inputs. Their strengths and limitations were analyzed across varying speech-to-noise ratios, offering insights into their robustness and suitability for different mixed audio scenarios.

2.2. Transformer

The Transformer architecture was proposed by Vaswani et al. in 2017, in the paper "Attention is All You Need". It revolutionized the way we handle sequence processing tasks, surpassing RNNs and LSTMs due to its efficiency and ability to manage long-range and complex dependencies. The attention mechanism and modular structure of the Transformer architecture make it ideal for a wide range of applications, from machine translation and text summarization to speech transcription and audio understanding. This architecture is composed of two main components: the Encoder and the Decoder, which work together to perform tasks such as translation, transcription, and text summarization. Furthermore, the core innovative concept of the Transformer is attention, specifically a mechanism known as Scaled Dot-Product Attention. The basic idea is that each word in a

sequence can attend to other words with varying degrees of importance when generating the output. Unlike RNN or LSTM networks, which process data sequentially, attention allows the model to evaluate all elements of the sequence simultaneously.

3. Methodology

To investigate the performance of two distinct neural networks—one specialized in speech sounds and another focused on non-speech sounds—when processing mixed audio containing speech and other sounds, we constructed the Clotho Voice dataset. This dataset is composed of various combinations of speech sounds from the Common Voice 5.1 dataset [Ardila et al. 2020] and general sounds from the Clotho 2.1 dataset [Drossos et al. 2020]. This section details the protocol established to evaluate how each network's performance deteriorates when faced with mixed audio inputs. The overall procedure is summarized in Figure 1. Each step is further described in the following subsections.

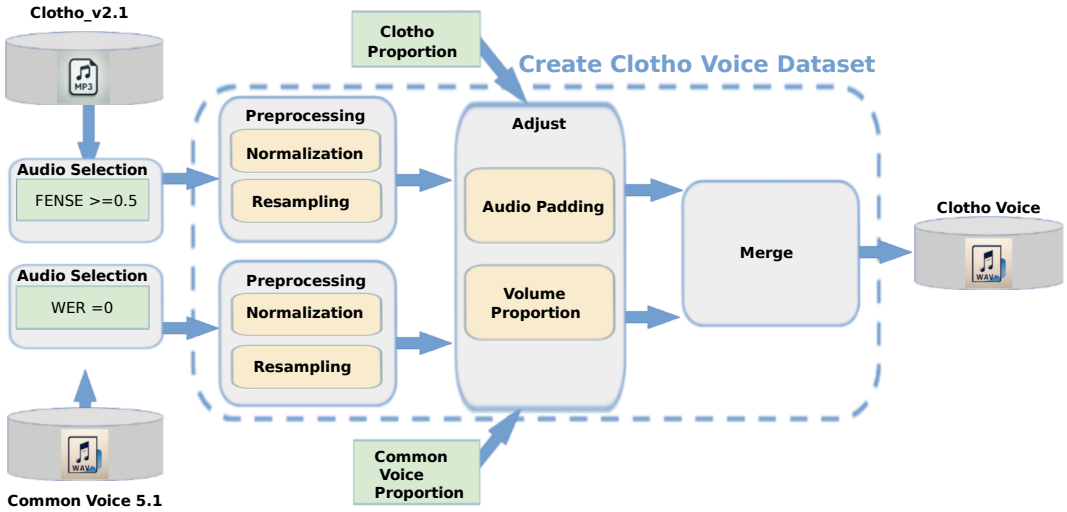

Figure 1. Clotho Voice Dataset Creation: Steps for creating the Clotho Voice Dataset, including (i) selecting the raw Clotho v2.1 and Common Voice v5.1 datasets, (ii) preprocessing them, (iii) applying audio padding and volume proportion adjustments, and (iv) merging the audio samples.

3.1. Datasets

The Clotho v2.1 dataset [Drossos et al. 2020] is a benchmark dataset designed to facilitate research in audio captioning, a task that involves generating descriptive textual captions for audio recordings. It is an updated version of the original Clotho dataset, containing diverse audio clips sourced from the Freesound platform, with durations ranging from 15 to 30 seconds. It contains $6,974$ general audio samples. Each audio clip in Clotho v2.1 is paired with five human-annotated captions, providing a rich variety of descriptive perspectives for the same auditory scene. The captions range from 8 to 20 words in length. The dataset is notable for its focus on real-world audio complexity, encompassing diverse environmental sounds, human activities, and instrumentals. This dataset is widely used to train and evaluate machine learning models in understanding and describing audio content, pushing the boundaries of multimodal AI systems. The Clotho dataset is available in the WAV (Waveform Audio File) format, ensuring high-fidelity audio quality.

The Common Voice v5.1 dataset [Ardila et al. 2020], developed by Mozilla, is a large-scale, multilingual dataset designed to advance research in automatic speech recognition and other speech-related technologies. This dataset is part of the Common Voice project, which crowdsources speech recordings from volunteers worldwide to create an open and diverse speech corpus. Version 5.1 features over 7,300 hours of recorded audio in 54 languages, including underrepresented languages. Each recording is paired with a corresponding written text, facilitating supervised learning tasks. The dataset is freely accessible and is used by researchers and developers to train, evaluate, and improve speech recognition models, as well as to support linguistic studies and the development of voice technology for low-resource languages.

3.2. Evaluation Metrics

The Word Error Rate (WER) [Morris et al. 2004] is a well-known and applied metric to evaluate the performance of an automatic speech recognition system. Derived from the Levenshtein distance, the WER performs at the word level instead of the phoneme level, being a valuable metric that first aligns the recognized word sequence with the reference (spoken) word sequence using dynamic string alignment. Further, the correlation between perplexity and word error rate is calculated to ensure the theory of the power law. Such a metric has been employed to evaluate Clotho Voice dataset due to its broad application in the evaluation of speech transcriptions conducted on Open AI's Whisper [Radford et al. 2023] models.

The FENSE metric is a novel evaluation framework specifically designed to assess the quality of audio captions. Unlike traditional metrics, such as BLEU or CIDEr, which were initially developed for image or machine translation tasks, FENSE incorporates semantic similarity and robustness to erroneous descriptions, aligning better with human judgments in the audio domain. It combines the capabilities of Sentence-BERT for capturing semantic similarities and introduces an Error Detector that penalizes inaccurate or misleading captions. This combination enables FENSE to outperform traditional metrics in accurately reflecting the quality of audio captions, as shown by its superior performance on benchmark datasets like AudioCaps-Eval and Clotho-Eval [Zhou et al. 2021].

3.3. Audio selection

Our primary objective is to examine how network performance degrades when audio signals are combined. To achieve this, we excluded audio samples in which the transcription or audio captioning networks inherently performed poorly. Accordingly, audio samples from each dataset were chosen based on specific selection criteria.

For the Common Voice 5.1 dataset, only the Portuguese audio samples that were perfectly transcribed by the Whisper transcription network (WER $= 0$) were retained. This choice ensures that any observed performance loss stems from mixing audio signals. Hence, we selected the $43,902$ Portuguese speech recordings from the Common Voice 5.1 dataset, ran them through the Whisper model, and calculated the WER metric. Only the $32,559$ audio samples with WER $= 0$ were included in the Clotho Voice dataset.

Selecting audio samples from the Clotho v2.1 dataset posed a greater challenge, because various descriptions can sufficiently match the same audio, yet each audio in Clotho provides only five reference captions. For this step, we adopted the FENSE metric,

which is reportedly more suitable for audio captioning. We evaluated $6,974$ audio samples from the Clotho v2.1 dataset using our chosen audio captioning network (Prompteus) and calculated the FENSE metric for each. Although the FENSE metric ranges from zero to one, the Prompteus captions did not reach a score of one. Therefore, we selected a few random samples from different metric strata—namely FENSE > 0.8, $0.6 <$ FENSE < 0.8, $0.5 <$ FENSE < 0.6, and $0.4 <$ FENSE < 0.5—for subjective evaluation.

For the subjective evaluation, we randomly chose five audio samples from each stratum. Two human annotators listened to each audio and created a caption. Then, these human-generated captions were compared to the captions produced by the audio captioning network, and both were subsequently compared with the five reference captions from the Clotho v2.1 dataset. We observed that audio samples yielding FENSE scores above 0.5 had reasonably accurate captions. Therefore, a FENSE threshold of 0.5 was established for the Clotho v2.1 samples.

3.4. Preprocessing

To exert more control over the audio mixing process and to systematically investigate performance degradation as a function of the proportion of speech versus non-speech content, the selected audio samples underwent normalization and resampling. For each file, the absolute maximum amplitude was first identified, and all values in that file were scaled by this maximum, yielding amplitude ranges from -1 to 1. This normalization aligns the volume levels across all audio samples, thereby enabling consistent comparisons and combinations.

3.5. Audio mixing

Since the Common Voice dataset yielded more selected samples, it served as the base dataset. For each audio sample from Common Voice, we randomly paired one audio sample from Clotho. Consequently, certain Clotho audio samples may appear multiple times in the Clotho Voice dataset. Because the audio files typically differed in duration, the shorter audio was padded with zeros at its endpoint. Next, we parameterized the relative volumes of each dataset. We adjusted each audio sample by a volume coefficient and mixed them, producing five sub-datasets:

- **Clotho-100-Voice-0**: The Clotho audio was multiplied by one, and the Common Voice audio was multiplied by zero. This sub-dataset serves to assess the effect of preprocessing on the performance of the audio captioning network.
- **Clotho-75-Voice-25**: The Clotho audio was multiplied by 0.75, and the Common Voice audio was multiplied by 0.25. This setup enables analysis of how minimal speech content can degrade audio captioning performance, while also evaluating the transcription network's robustness to substantial background noise.
- **Clotho-50-Voice-50**: Both Clotho and Common Voice audio samples were each multiplied by 0.5. This configuration reflects a scenario where speech and background noise are equally prominent, representing an average case of performance degradation for both networks.
- **Clotho-25-Voice-75**: The Clotho audio was multiplied by 0.25, and the Common Voice audio was multiplied by 0.75. This sub-dataset tests how a relatively quiet noise signal and dominant speech signal may degrade the audio captioning network. It also reveals the impact of a small amount of background noise on the transcription network.

- **Clotho-0-Voice-100**: The Clotho audio was multiplied by zero, and the Common Voice audio was multiplied by one. This sub-dataset investigates how preprocessing alone influences the transcription network's performance.

Together, these five sub-datasets encompass a broad range of mixing ratios, providing a comprehensive platform for examining performance degradation in both audio captioning and transcription networks.

4. Results

In this section, we report the performance outcomes of both Whisper and Prompteus when exposed to different levels of speech and background content in the Clotho Voice dataset.

4.1. Evaluating Whisper's Transcription Robustness Under Background Interference

Figure 2 illustrates the Word Error Rate (WER) obtained by the Whisper model when transcribing each subdataset of the Clotho Voice dataset, highlighting how increasing amounts of non-speech background sounds affect speech detection accuracy. At one extreme, Clotho-100-Voice-0 (all background, zero speech) exhibits the highest WER, indicating a substantial mismatch between Whisper's speech-focused architecture and purely non-speech inputs. As background content becomes progressively mixed with speech—moving from Clotho-75-Voice-25 to Clotho-50-Voice-50 to Clotho-25-Voice-75—there is a marked decrease in WER, reflecting the model's ability to better discern speech when the signal-to-noise ratio improves. Finally, Clotho-0-Voice-100 (pure speech) yields an almost negligible WER, confirming that Whisper excels when background interference is minimal or absent.

Figure 3 presents a chart showing the number of audio files that achieved perfect transcription (WER $= 0$, shown in green) versus those with transcription errors (WER > 0, shown in red) across the five Clotho Voice sub-datasets. Notably, Clotho-100-Voice-0—composed primarily of background sounds—features no samples with WER = 0, underscoring the speech-centric nature of Whisper's architecture. As the proportion of speech gradually increases in Clotho-75-Voice-25, Clotho-50-Voice-50, and Clotho-25-Voice-75, the number of perfectly transcribed files (green bars) grows steadily. The subdataset containing only speech (Clotho-0-Voice-100) stands out with the highest count of flawlessly transcribed samples, showcasing the model's strong performance when minimal background interference is present. However, it is noteworthy that 222 samples in the Clotho-0-Voice-100 sub-dataset did not achieve WER $= 0$, indicating that the preprocessing applied to the original audio still impacts the network's performance.

Overall, these findings demonstrate that Whisper is well-suited for scenarios with predominantly speech-based inputs but exhibits notable performance drops when non-speech components increase. By systematically analyzing each Clotho Voice sub-dataset, we have highlighted the trade-offs involved in applying a speech-centric model to mixed audio conditions. These insights lay the groundwork for future investigations aimed at further enhancing transcription robustness, either through improved preprocessing or targeted retraining strategies that account for varying levels of background interference.

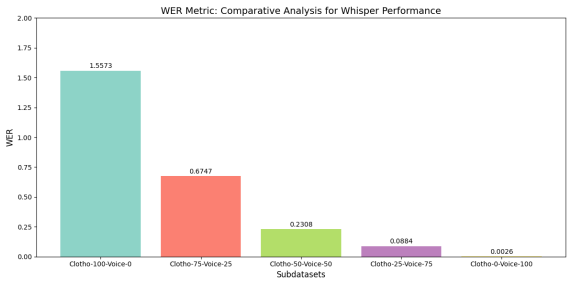

Figure 2. Speech detection evaluation on Clotho Voice dataset.

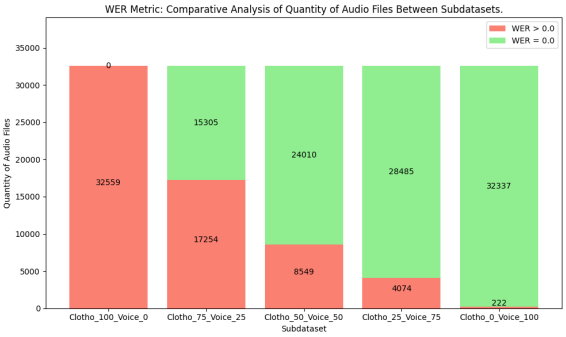

Figure 3. Number of samples that achieved WER $= 0$ and WER > 0.

4.2. Prompteus's Background Sound Detection Under Mixed Speech Conditions

Figure 4 presents the FENSE scores obtained by Prompteus for background sound detection across the Clotho Voice sub-datasets. The sub-dataset comprising only background audio (Clotho-100-Voice-0) attains the highest score (0.66), indicating that Prompteus effectively captures background phenomena when no speech content is present. As speech is incrementally introduced into the audio mixtures (Clotho-75-Voice-25 through Clotho-25-Voice-75), the FENSE metric progressively decreases, suggesting that overlapping speech signals impede the model's ability to discern background events. Finally, the nearly negligible score observed in the all-speech sub-dataset (Clotho-0-Voice-100) confirms that Prompteus's architecture, tuned for background detection, is markedly limited when no clear environmental cues are present in the input signal.

Figure 5 provides a chart depicting how many audio files in each Clotho Voice sub-dataset yielded FENSE scores above (green) or below (red) the 0.5 threshold. In Clotho-100-Voice-0, which contains only background sounds, most samples exceed the FENSE 0.5 threshold, indicating Prompteus's effectiveness in describing environmental audio. As speech content increases across Clotho-75-Voice-25 through Clotho-25-Voice-75, the proportion of samples with high FENSE scores steadily decreases, reflecting the growing difficulty of accurately characterizing background events amid overlapping speech. Finally, in Clotho-0-Voice-100, which is entirely speech-based, only a small number of files achieve FENSE scores above 0.5, underscoring Prompteus's reduced capacity for capturing background elements when foreground speech dominates the input.

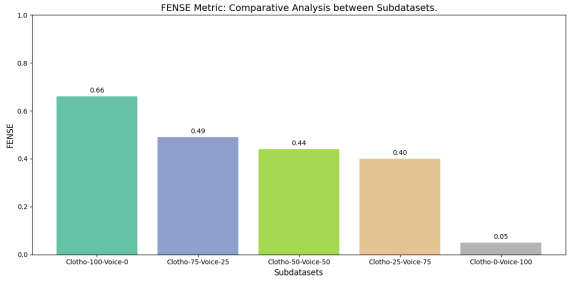

Figure 4. Background sound detection evaluation on Clotho Voice dataset.

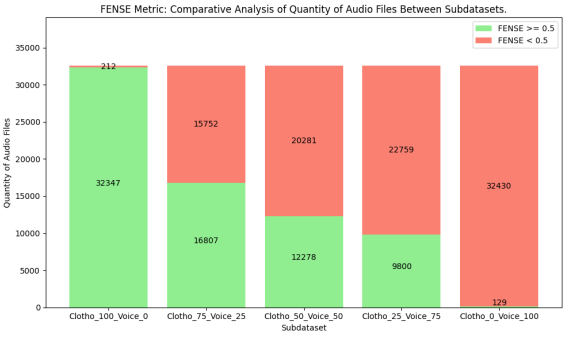

Figure 5. Number of samples with FENSE ≥ 0 and FENSE < 0.

It is noteworthy that 212 samples in the Clotho-100-Voice-0 sub-dataset, which contains purely background audio, obtained FENSE scores below 0.5. These instances indicate that the preprocessing steps applied to the original audio may have adversely affected borderline samples, particularly those with inherently low signal quality or ambiguous acoustic cues. As a result, the Prompteus model struggled to accurately characterize environmental components in a subset of these processed background recordings. Interestingly, 129 samples in the Clotho-0-Voice-100 sub-dataset, which is entirely speech-based, achieved FENSE scores exceeding 0.5. Closer inspection of these audio files revealed that they indeed featured people speaking, leading the reference captions to describe speech events within the recordings. Because these references coincidentally aligned with Prompteus's capacity to detect environmental cues, the model was able to produce sufficiently accurate descriptions, thereby resulting in high FENSE values even in a nominally non-background dataset.

5. Discussion

In this section, we present a comprehensive discussion of how both the Whisper and Prompteus models respond to varying degrees of speech and background sounds in the evaluated audio samples. We investigate both qualitative and quantitative aspects, emphasizing how the tasks of transcription and audio captioning degrade under conditions of mixed or mismatched input signals.

5.1. Qualitative Evaluation of Speech and Background Detection

As expected, the greater the proportion of merged audio portions fed into Whisper and Prompteus, the larger the degradation in their respective performances. Such behavior arises because the evaluated models were initially tuned to yield optimal performance under specific conditions, i.e., Whisper focuses on speech-based audio, whereas Prompteus was designed to capture background phenomena. When audio containing both speech and non-speech sounds is supplied to these models, a degree of misinterpretation ensues. Whisper attempts to convert non-speech effects into words, and Prompteus attempts to describe not only the background effects but also interprets that "someone is speaking" in the audio.

Interestingly, Whisper's degradation for the Clotho-50-Voice-50 subset is more similar to Clotho-0-Voice-100 than to Clotho-75-Voice-25. This finding suggests that Whisper is robust to background effects that can mask speech in the input audio. Even with half of the audio consisting of background sounds, Whisper can still handle the speech component and produce a reasonable transcription, as indicated by the WER results shown in Figure 2.

A similar pattern emerges when evaluating the Clotho-Voice dataset on the Prompteus model. A progressive and smooth degradation is observed from Clotho-100-Voice-0 to Clotho-25-Voice-75. However, performance drops considerably from Clotho-25-Voice-75 to Clotho-0-Voice-100 (which is essentially the Common Voice subset). Again, Prompteus, derived from a Whisper-based architecture, retains its ability to detect sound effects even when additional speech "clouds" the background audio. Figure 4 illustrates this behavior, highlighting the potential of Prompteus for background-detection tasks in the presence of non-background phenomena in the input samples.

Depending on the proportion of speech and background effects in the merged input, both Whisper and Prompteus can detect their respective targets—speech and background phenomena—in a complementary manner. Together, they can fully describe the content of each audio in a closed-caption style, with an acceptable degradation rate, as shown in the Results section.

5.2. Ablation Study: Analyzing the Impact of Audio Merging on Model Performance

After evaluating the complete Clotho-Voice dataset, as presented in Figures 2 and 4, we conducted an ablation study to examine how the merging process contributes to performance degradation in Whisper and Prompteus. For this analysis, five distinct pairs of tracks from Clotho and Common Voice were randomly selected, merged, and then evaluated using both Whisper and Prompteus. When examining the most relevant Clotho-Voice subset (the 50/50 mixture), several notable outcomes were observed. Whisper, even when faced with background sounds from Clotho, remains robust enough to detect speech in the merged input, achieving the same WER as the original pre-merging audio. However, if the Clotho sample in the mixture also contains speech, Whisper attempts to transcribe it, thereby substantially influencing the WER metric. Because Whisper expects speech-based signals, introducing additional speech to the input can compromise the transcription outcome.

A similar evaluation on Prompteus, where predictions for five merged samples (50/50 ratio) were compared with their original Clotho recordings, revealed instances where certain Clotho samples did not contain background phenomena. Conversely, some Common Voice samples did include background sounds (which ideally should not occur). When merged, these unexpected background sounds led to descriptions of the Clotho-Voice mixture that were markedly different from the original Clotho sample. Nevertheless, Prompteus is robust enough to recognize any effect in the input, although the added audio elements can drastically alter the descriptions compared to the corresponding original files in the Clotho dataset. In addition, Prompteus often detects human speech in the merged samples, showing high sensitivity to speaker gender. However, changes in phrasing or tense—e.g., “A woman is speaking.” vs. “A woman speaks.” or “A girl spoke.” and so forth—can adversely affect comparison metrics that rely on exact word matching and positioning. Although the overall semantic context may remain intact, minor textual variations can lead to discrepancies in these quantitative metrics.

6. Conclusions

Overall, this study has demonstrated the feasibility and challenges of jointly handling speech transcription and background sound detection using two specialized transformer-based models, Whisper and Prompteus. Through systematic experimentation with the Clotho Voice dataset, which merges speech from Common Voice and general sounds from Clotho, we showed that both models exhibit high performance within their specialized domains but experience predictable degradation when confronted with significant overlap between speech and non-speech signals. We further highlighted that preprocessing steps, such as audio normalization and mixing, can have non-negligible effects on model performance, particularly for borderline examples.

This investigation underscores the need for versatile approaches capable of handling diverse audio tasks under real-world conditions. In industrial environments or similarly noisy settings, the ability to accurately capture both speech and contextually relevant background sounds remains essential for safety, automation, and robust multimodal interfaces. Our findings suggest that future work should focus on enhancing model robustness through more comprehensive training strategies, including targeted data augmentation for challenging acoustic conditions and hybrid architectures that integrate joint representations of speech and environmental cues. Moreover, progress in general sound captioning would greatly benefit from the creation of a large-scale, dataset that encompasses a broader array of environmental and non-speech events. Ultimately, this line of research paves the way for more intelligent, context-aware systems that can leverage both transcription and audio captioning to better understand and interact with complex auditory environments.

Acknowlegments

This study was financed in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil); and Fundação de Amparo à Pesquisa do Espírito Santo (FAPES, Brazil) - grants 2021-07KJ2, 2022-NGKM5, 2022-CD0RQ and 193/2022. Os autores agradecem a FAPES/UnAC (Nº FAPES 1228/2022 P 2022-CD0RQ, Nº SIAFEM 2022-CD0RQ) pelo apoio financeiro dado por meio do Sistema UniversidaES.

References

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., and Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222.

Crocco, M., Cristani, M., Trucco, A., and Murino, V. (2016). Audio surveillance: A systematic review. *ACM Computing Surveys (CSUR)*, 48(4):1–46.

Czyżewski, A., Skarzynaski, H., Kostek, B., and Geremek, A. (1998). Multimedia technology for hearing impaired people. *1998 IEEE Second Workshop on Multimedia Signal Processing (Cat. No.98EX175)*, pages 181–186.

Drossos, K., Adavanne, S., and Virtanen, T. (2017). Automated audio captioning with recurrent neural networks. *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 374–378.

Drossos, K., Lipping, S., and Virtanen, T. (2020). Clotho: an audio captioning dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740.

Gemmeke, J., Ellis, D., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017a). Audio set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017b). Audio set: An ontology and human-labeled dataset

for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Ikawa, S. and Kashino, K. (2019). Neural audio captioning based on conditional sequence-to-sequence model. In *Workshop on Detection and Classification of Acoustic Scenes and Events*.

Kadlčík, M., Hájek, A., Kieslich, J., and Winiecki, R. (2023). A whisper transformer for audio captioning trained with synthetic captions and transfer learning.

Kim, C. D., Kim, B., Lee, H., and Kim, G. (2019a). Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132.

Kim, C. D., Kim, B., Lee, H., and Kim, G. (2019b). AudioCaps: Generating captions for audios in the wild. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, Minneapolis, Minnesota. Association for Computational Linguistics.

Lane, N., Georgiev, P., and Qendro, L. (2015). Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*.

Mei, X., Liu, X., Huang, Q., Plumbley, M. D., and Wang, W. (2021). Audio captioning transformer.

Mei, X., Meng, C., Liu, H., Kong, Q., Ko, T., Zhao, C., Plumbley, M., Zou, Y., and Wang, W. (2023). Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354.

Morris, A. C., Maier, V., and Green, P. D. (2004). From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Interspeech*, pages 2765–2768.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Stöter, F.-R., Uhlich, S., Liutkus, A., and Mitsufuji, Y. (2019). Open-unmix - a reference implementation for music source separation. *J. Open Source Softw.*, 4:1667.

Wang, Y., Liu, Z., and Huang, J. (2000). Multimedia content analysis-using both audio and visual clues. *IEEE Signal Process. Mag.*, 17:12–36.

Wu, M., Dinkel, H., and Yu, K. (2019). Audio caption: Listen and tell. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 830–834. IEEE.

Zhou, Z., Zhang, Z., Xu, X., Xie, Z., Wu, M., and Zhu, K. Q. (2021). Can audio captions be evaluated with image caption metrics? *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 981–985.