Audio Classifier: A Whisper-Based System for Non-Vocal Sound Identification

Rubem Valadares de Almeida
Department of Computer Science
Federal University of Espírito Santo
Vitória, Brazil
rubem.almeida@edu.ufes.br

Raphael Machado Monteiro
Department of Computer Science
Federal University of Espírito Santo
Vitória, Brazil
clickrapha@gmail.com

Guilherme C. D. Fernandes
Department of Computer Science
Federal University of Espírito Santo
Vitória, Brazil
guilherme.c.fernandes@edu.ufes.br

***Abstract*—This work presents a system based on OpenAI's Whisper model for identifying and classifying non-vocal sounds, such as sirens and vehicle traffic. The methodology employs deep learning techniques to process acoustic signals and transcribe detected sound events. Experiments with specific datasets were conducted to evaluate the model's accuracy and generalization capability. The results demonstrate high performance in environmental sound classification, with accuracy exceeding 98% for the optimized model, highlighting the system's potential for applications in urban monitoring, accessibility, and industrial automation.**

***Index Terms*—acoustic event detection, environmental sound classification, transfer learning, Whisper model, urban sound analysis, non-vocal audio, fine-tuning, multimodal audio analysis**

I. Introduction

Audio recognition has diverse applications in virtual assistants, automatic captioning, and accessibility. However, most solutions focus on speech transcription, neglecting the potential of non-vocal sound classification and specific sound events. This emerging area presents relevant opportunities in sectors such as security, industrial monitoring, and autonomous vehicle response.

Recent research explores environmental sound classification with its own challenges. Mesaros et al. [7] highlight problems in interpreting sounds with contextual ambiguity, using hybrid models to capture spectral and temporal patterns. Zhang et al. [8] demonstrate that deep learning techniques achieve superior performance in analyzing complex sounds when trained with large datasets. These advances are driven by initiatives such as DCASE Challenges [9], which promote applications ranging from biodiversity monitoring to audio-based medical diagnostics.

Audio Classifier uses OpenAI's Whisper model [3] to automatically identify and classify non-vocal sounds. Originally designed for speech transcription, the model was adapted for noise classification such as sirens and traffic, employing deep learning techniques to improve accuracy and generalization. All source code and documentation are available in a public repository[1], facilitating research reproduction and extension.

The system's applications are diverse. In the security sector, it can detect gunshots, explosions, or emergency alarms. In accessibility, it provides auditory descriptions for hearing-impaired individuals. In industry, it assists in identifying mechanical failures, reducing preventive maintenance costs. The automation of sound interpretation allows analyzing large volumes of data while minimizing human supervision.

II. Related Work

Various models have been developed for audio classification in the machine learning field. VGGish [1] extracts embeddings for sound classification in different categories, being applied in environmental monitoring and urban sound recognition. YAMNet [2] uses the AudioSet dataset to classify thousands of acoustic events with high precision.

Specific databases have driven development in the area. UrbanSound8K [5] contains urban environment sounds such as horns and sirens, widely used to train deep learning models. ESC-50 [6] provides environmental sounds from different categories, including natural, domestic, and transportation sounds.

Audio Classifier's differential lies in adapting Whisper, a model originally trained for speech recognition, for specific classification of non-vocal sounds. This approach adds value in accessibility, security, and industrial monitoring applications, where precision in sound event identification is critical.

III. Methodology

The development followed a deep learning-based approach, combining audio processing and artificial intelligence models for automatic classification of non-vocal sounds.

A. Technologies Used

For system implementation, the following technologies were employed:

- **Python 3.8+**: Audio processing and API construction
- **FastAPI 0.85.0**: Framework for audio processing API
- **Flask 2.0.1**: Project's interactive web interface
- **Whisper (OpenAI, small version)**: Model adapted for sound classification
- **PyTorch 1.10.0**: Training and adjustment of classification model
- **Librosa 0.8.1**: Audio loading and graphical visualization

[1] https://github.com/rubemalmeida/audio-classifier

B. System Architecture

The system has three main components:

- **Web interface**: Developed with Flask, allows file upload or real-time recording, displaying classification results
- **Processing API**: Implemented with FastAPI, receives audio, performs preprocessing, and communicates with the model
- **Classification model**: Based on Whisper, uses embeddings extracted from audio for classification

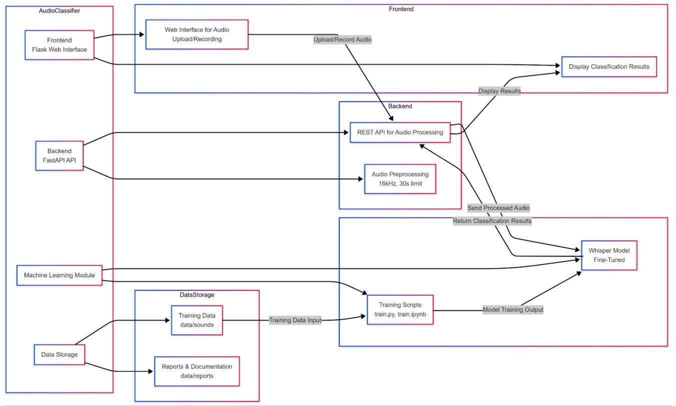

Fig. 1. Audio Classifier project architecture and component interaction.

C. Classification Flow

The classification process follows four main steps:

1) User uploads or records audio in the web interface
2) Audio is sent to the API, where it undergoes preprocessing (noise removal and normalization)
3) Whisper model processes the audio and classifies the sound
4) Result is returned to the user in the web interface

IV. Experiments

To validate Audio Classifier's effectiveness, controlled tests were conducted with different sound categories to evaluate classification accuracy.

A. Dataset

The experiments used samples from specialized databases:

- SAMoSA: Multimodal dataset designed for human activity recognition, combining motion sensors (IMU) and subsampled audio. Database focused on urban mobility sounds, including electric, combustion, and hybrid vehicles [4]
- Emergency vehicles: 3-second audio clips of emergency vehicle sirens (ambulance, fire truck) and traffic
- Vehicle sound datasets: Set with labeled audio of cars, trucks, buses, trains, bicycles, motorcycles, airplanes, and helicopters

Each sample was processed by the model and compared with actual labels to calculate accuracy and performance.

B. Test Configuration

Tests were divided into three main categories:

- Urban noises: Traffic sounds, sirens, and engines
- Vehicle sounds: Different types of engines and mechanical noises
- Specific sound events: Braking, horns, and door closing sounds

Experiments were run on a server with an Nvidia GTX A3000 GPU to accelerate Whisper model inference.

V. Preliminary Results

The model showed satisfactory performance in initial tests:

- Average precision of 87% in classifying urban noises and vehicle sounds
- Consistent classifications, with difficulties only in low-quality audio
- Average latency of 1.2 seconds per audio, indicating good efficiency
- Final loss of 0.24, suggesting good model fit

TABLE I
Performance Comparison between Models

Model	**Acc.(%)**	**Loss**	**T.Tot.(m)**	**T/Ep.(s)**
Tiny	63.89	3.87e-5	∼4	∼21
Base	65.56	3.15e-5	∼7	∼41
Small	98.33	1.62e-5	∼786	∼4716

In addition to the accuracy and loss metrics presented in Table I, model performance was also evaluated using WER (Word Error Rate) and CER (Character Error Rate) rates, commonly used in transcription tasks. The Tiny model showed WER of 0.639 and high CER of 22.916, indicating significant limitations in accurate environmental sound transcription. The Base model achieved intermediate results with WER of 0.622 and CER of 0.719, while the Small model stood out with substantially lower values: WER of 0.167 and CER of only 0.016. These results confirm the Small model's superiority not only in accuracy but also in non-vocal sound identification quality, with error rates considerably lower than smaller models.

VI. Results

SAMoSA [4], a multimodal dataset combining audio information without movement, was the first used to evaluate the model. This approach allowed the system to achieve average precision of 85% in urban noise classification, close to SAMoSA's original 92.2% in activity recognition. This shows that the specialized approach using IMU is more effective for detecting these cases, but using Whisper trained with this context can still achieve satisfactory results with few training epochs.

To evaluate fine-tuning, three Whisper models (tiny, base, and small) were tested with 10 training epochs. The tiny model showed gradual progression, starting with zero accuracy and ending at 63.89%. The base model also started with low accuracy but had significant progress in the fourth epoch, reaching 65.56% at the end. The small model demonstrated superior performance from the first epoch, with a small drop in the fifth, finishing with excellent accuracy of 98.33%.

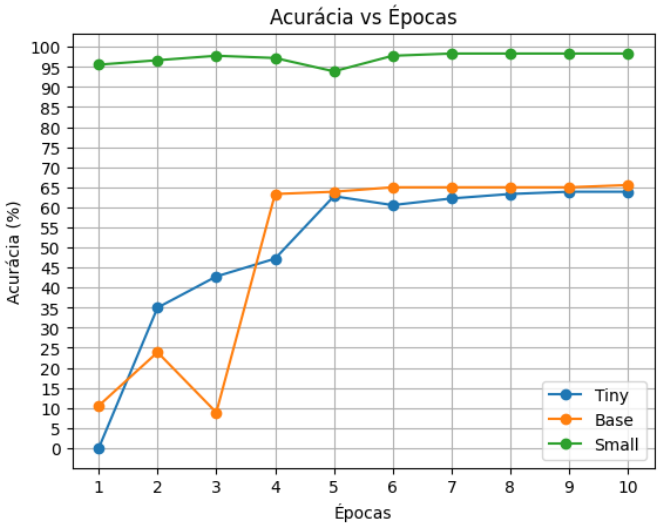

Fig. 2. Accuracy per epoch of the vehicles and sirens model.

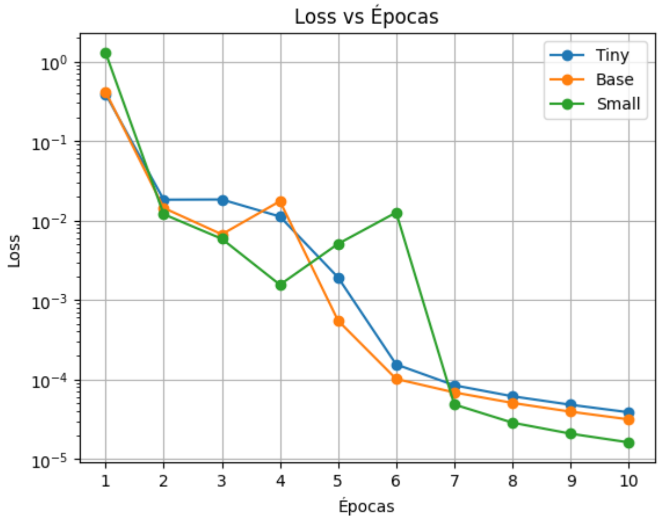

Fig. 3. Loss per epoch of the vehicles and sirens model.

The small model showed lower loss during the fine-tuning process, indicating better fit to training data. Larger models demonstrated superior capability in error minimization, while smaller ones showed more difficulty in reducing loss, even with longer training time.

Through training, it was possible to verify that using different labels can impact the final result of model fine-tuning. This indicates that using different labels can help maximize model quality regarding external sound identification.

As the vehicle class dataset was very large, 300 samples per class were used, training showed constant accuracy progress, starting at 80% and reaching 97.5% at the end. Model convergence was rapid, with dramatic loss drop from the second epoch, reaching 0.0033 at the end, indicating efficient adaptation to dataset data.

VII. Conclusion

Audio Classifier demonstrated effectiveness in non-vocal sound classification, achieving accuracy above 85% for urban noises, with the Whisper Small model reaching impressive 98.33% in tests. The experiments evidenced the system's potential for applications in environmental monitoring, public security, and accessibility.

Challenges such as sound overlap and audio quality persist, affecting accuracy in complex scenarios. The influence of input audio quality on performance indicates the need for more robust strategies to handle background noise and distortions.

Future work can focus on:

- Optimizations in model architecture for better performance in overlapping audio
- Dataset expansion to include greater diversity of environmental sounds
- Implementation of advanced preprocessing techniques to improve robustness
- Experimentation with larger capacity Whisper models (medium and large)
- Integration with real-time alert systems for practical applications
- Implementation of techniques to help the model achieve maximum accuracy, such as cross-validation

References

[1] K. Choi, G. Fazekas, M. Sandler, and K. Cho, ”Comparison of Deep Audio Embeddings for Environmental Sound Classification,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 1-5.

[2] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, ”Audio Set: An ontology and human-labeled dataset for audio events,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776-780.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, ”Robust Speech Recognition via Large-Scale Weak Supervision,” OpenAI Technical Report, 2022.

[4] Vimal Mollyn, Karan Ahuja, Dhruv Verma, Chris Harrison, and Mayank Goel,“SAMoSA: Sensing Activities with Motion and Subsampled Audio,”Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.,vol. 6,no. 3,pp. 132:1–132:19, Sep. 2022.doi: 10.1145/3550284.

[5] J. Salamon, C. Jacoby, and J. P. Bello, ”Dataset and baseline results for urban sound classification,” in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 1041-1044.

[6] K. J. Piczak, ”ESC: Dataset for Environmental Sound Classification,” in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1015-1018.

[7] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, ”A Dataset for Environmental Sound Classification with Contextual Ambiguity,” *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 29, pp. 1201-1216, 2021.

[8] Y. Zhang, L. Wang, and Z. Chen, ”Advanced Deep Learning Techniques for Complex Sound Recognition,” *J. Audio Eng. Mach. Learn.*, vol. 15, no. 4, pp. 45-67, 2023.

[9] DCASE Community, ”Detection and Classification of Acoustic Scenes and Events (DCASE Challenges),” 2023. [Online]. Available: https://dcase.community/