Audio Classifier: Um Sistema Baseado em Whisper para Identificação de Sons Não Vocais

Rubem Valadares de Almeida
Departamento de Informática
Universidade Federal do Espírito Santo
Vitória, Brasil
rubem.almeida@edu.ufes.br

Raphael Machado Monteiro
Departamento de Informática
Universidade Federal do Espírito Santo
Vitória, Brasil
clickrapha@gmail.com

Guilherme C. D. Fernandes
Departamento de Informática
Universidade Federal do Espírito Santo
Vitória, Brasil
guilherme.c.fernandes@edu.ufes.br

***Abstract*—Este trabalho apresenta um sistema baseado no modelo Whisper da OpenAI para identificação e classificação de sons não vocais, como sirenes e tráfego de veículos. A metodologia empregada envolve técnicas de deep learning para processar sinais acústicos e transcrever eventos sonoros detectados. Experimentos com conjuntos de dados específicos foram conduzidos para avaliar a precisão do modelo e sua capacidade de generalização. Os resultados demonstram alto desempenho na classificação de sons ambientais, com acurácia superior a 98% para o modelo otimizado, destacando o potencial do sistema para aplicações em monitoramento urbano, acessibilidade e automação industrial.**

***Index Terms*—detecção de eventos acústicos, classificação de sons ambientais, aprendizado por transferência, modelo Whisper, análise de sons urbanos, áudio não-vocal, ajuste fino, análise multimodal de áudio**

I. Introdução

O reconhecimento de áudio tem aplicações diversas em assistentes virtuais, legendagem automática e acessibilidade. Entretanto, a maioria das soluções foca na transcrição de fala, negligenciando o potencial da classificação de sons não vocais e eventos sonoros específicos. Essa área emergente apresenta oportunidades relevantes em setores como segurança, monitoramento industrial e resposta automática de veículos autônomos.

Pesquisas recentes exploram a classificação de sons ambientais com desafios próprios. Mesaros et al. [7] destacam problemas na interpretação de sons com ambiguidade contextual, utilizando modelos híbridos para capturar padrões espectrais e temporais. Zhang et al. [8] demonstram que técnicas de deep learning alcançam desempenho superior na análise de sons complexos quando treinadas com grandes conjuntos de dados. Esses avanços são impulsionados por iniciativas como os DCASE Challenges [9], que promovem aplicações desde monitoramento de biodiversidade até diagnósticos médicos baseados em áudio.

O Audio Classifier utiliza o modelo Whisper da OpenAI [3] para identificar e classificar automaticamente sons não vocais. Originalmente projetado para transcrição de fala, o modelo foi adaptado para classificação de ruídos como sirenes e tráfego, empregando técnicas de aprendizado profundo para melhorar precisão e generalização. Todo o código-fonte e documentação estão disponíveis em repositório público[1], facilitando reprodução e extensão da pesquisa.

As aplicações do sistema são diversas. No setor de segurança, pode detectar tiros, explosões ou alarmes de emergência. Em acessibilidade, fornece descrições auditivas para pessoas com deficiência auditiva. Na indústria, auxilia na identificação de falhas mecânicas, reduzindo custos de manutenção preventiva. A automação da interpretação sonora permite analisar grandes volumes de dados minimizando supervisão humana.

II. Trabalhos Correlatos

Diversos modelos têm sido desenvolvidos para classificação de áudio no campo do aprendizado de máquina. O VGGish [1] extrai embeddings para classificação de sons em diferentes categorias, sendo aplicado em monitoramento ambiental e reconhecimento de sons urbanos. O YAMNet [2] utiliza o conjunto de dados AudioSet para classificar milhares de eventos acústicos com alta precisão.

Bases de dados específicas têm impulsionado o desenvolvimento na área. A UrbanSound8K [5] contém sons de ambientes urbanos como buzinas e sirenes, amplamente utilizados para treinar modelos de aprendizado profundo. A ESC-50 [6] disponibiliza sons ambientais de diferentes categorias, incluindo sons naturais, domésticos e de transporte.

O diferencial do Audio Classifier está na adaptação do Whisper, modelo originalmente treinado para reconhecimento de fala, para a classificação específica de sons não vocais. Esta abordagem agrega valor em aplicações de acessibilidade, segurança e monitoramento industrial, onde a precisão na identificação de eventos sonoros é crítica.

III. Metodologia

O desenvolvimento seguiu uma abordagem baseada em aprendizado profundo, combinando processamento de áudio e modelos de inteligência artificial para classificação automática de sons não vocais.

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

A. Tecnologias Utilizadas

Para a implementação do sistema, foram empregadas as seguintes tecnologias:

- **Python 3.8+**: Processamento de áudio e construção da API
- **FastAPI 0.85.0**: Framework para a API de processamento de áudio
- **Flask 2.0.1**: Interface web interativa do projeto
- **Whisper (OpenAI, versão small)**: Modelo adaptado para classificação de sons
- **PyTorch 1.10.0**: Treinamento e ajuste do modelo de classificação
- **Librosa 0.8.1**: Carregamento de áudio e visualização gráfica

B. Arquitetura do Sistema

O sistema possui três componentes principais:

- **Interface web**: Desenvolvida com Flask, permite upload de arquivos ou gravação em tempo real, exibindo resultados da classificação
- **API de processamento**: Implementada com FastAPI, recebe áudios, realiza pré-processamento e comunica-se com o modelo
- **Modelo de classificação**: Baseado no Whisper, utiliza embeddings extraídos dos áudios para classificação

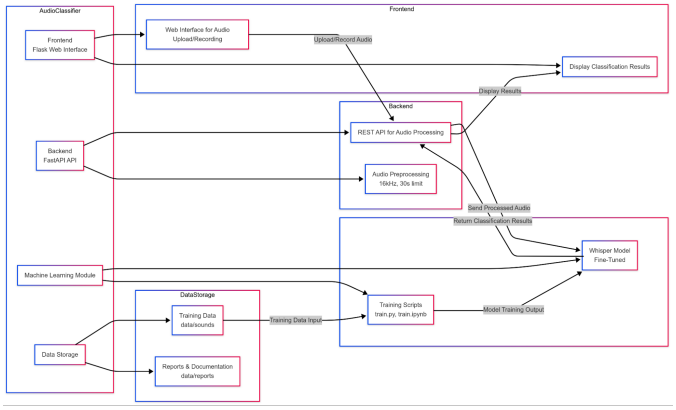

Fig. 1. Arquitetura do projeto Audio Classifier e interação dos componentes.

C. Fluxo de Classificação

O processo de classificação segue quatro etapas principais:

1) O usuário faz upload ou grava um áudio na interface web
2) O áudio é enviado para a API, onde passa por pré-processamento (remoção de ruídos e normalização)
3) O modelo Whisper processa o áudio e classifica o som
4) O resultado é retornado ao usuário na interface web

IV. Experimentos

Para validar a eficácia do Audio Classifier, foram realizados testes controlados com diferentes categorias de sons para avaliar a precisão na classificação.

A. Conjunto de Dados

Os experimentos utilizaram amostras de bases especializadas:

- SAMoSA: Conjunto de dados multimodais projetada para o reconhecimento de atividades humanas, combinando sensores de movimento (IMU) e áudio subamostrado. Base focada em sons de mobilidade urbana, incluindo veículos elétricos, combustão e híbridos [4]
- Veículos de emergência: Áudios de 3 segundos de sirenes de veículos de emergência (ambulância, bombeiro) e tráfego.
- Vehicle sound datasets: Conjunto com áudios rotulados de carros, caminhões, ônibus, trens, bicicletas, motocicletas, aviões e helicópteros.

Cada amostra foi processada pelo modelo e comparada com rótulos reais para calcular a precisão e o desempenho.

B. Configuração dos Testes

Os testes foram divididos em três categorias principais:

- Ruídos urbanos: Sons de tráfego, sirenes e motores
- Sons de veículos: Diferentes tipos de motores e ruídos mecânicos
- Eventos sonoros específicos: Freadas, buzinas e sons de fechamento de portas

Os experimentos foram executados em servidor com uma GPU, Nvidia GTX A3000, para acelerar a inferência do modelo Whisper.

V. Resultados Preliminares

O modelo apresentou desempenho satisfatório nos testes iniciais:

- Precisão média de 87% na classificação de ruídos urbanos e sons de veículos
- Classificações consistentes, com dificuldades apenas em áudios de baixa qualidade
- Latência média de 1,2 segundos por áudio, indicando boa eficiência
- Loss final de 0.24, sugerindo bom ajuste do modelo

TABLE I
Comparativo de Desempenho entre Modelos

Modelo	Acur.(%)	Perda	T.Tot.(m)	T/Ép.(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

Além das métricas de acurácia e perda apresentadas na Tabela I, o desempenho dos modelos também foi avaliado usando as taxas de erro WER (Word Error Rate) e CER (Character Error Rate), comumente utilizadas em tarefas de transcrição. O modelo Tiny apresentou WER de 0.639 e CER elevada de 22.916, indicando limitações significativas na transcrição precisa de sons ambientais. O modelo Base obteve resultados intermediários com WER de 0.622 e CER de 0.719, enquanto o modelo Small se destacou com valores

substancialmente menores: WER de 0.167 e CER de apenas 0.016. Estes resultados confirmam a superioridade do modelo Small não apenas em acurácia, mas também na qualidade da identificação dos sons não vocais, com taxas de erro consideravelmente inferiores aos modelos menores.

VI. Resultados

O SAMoSA [4], conjunto de dados multimodal que combina informações de áudio, sem o movimento, foi o primeiro utilizado para avaliar o modelo. Esta abordagem permitiu ao sistema atingir precisão média de 85% na classificação de ruídos urbanos, próximo aos 92,2% do SAMoSA original no reconhecimento de atividades. O que mostra que a abordagem especializada com o uso do IMU é mais efetiva para a detecção desses casos, mas o uso do whisper treinado com esse contexto ainda consegue atingir um resultado satisfatório com poucas épocas de treinamento.

Para avaliar o fine-tuning, foram testados três modelos do Whisper (tiny, base e small) com 10 épocas de treinamento. O modelo tiny apresentou progressão gradual, iniciando com acurácia zero e finalizando com 63.89%. O base iniciou também com baixa acurácia, mas teve progresso significativo na quarta época, atingindo 65.56% ao final. O modelo small demonstrou desempenho superior desde a primeira época, com pequena queda na quinta, finalizando com excelente acurácia de 98.33%.

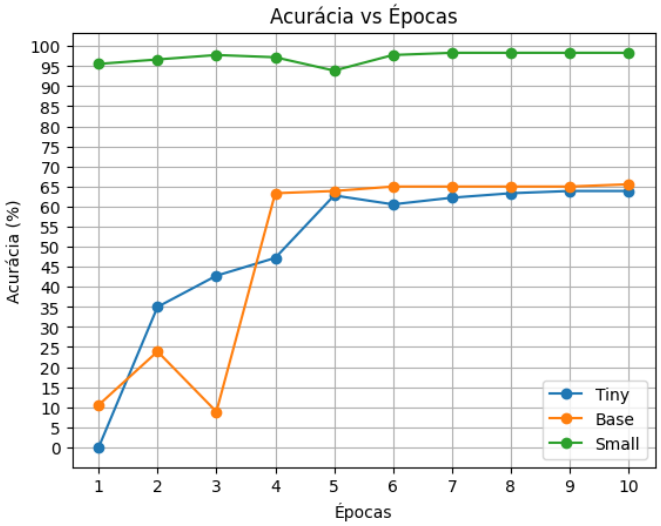

Fig. 2. Acurácia por época do modelo de veículos e sirenes.

O modelo small apresentou menor perda durante o processo de *fine-tuning*, indicando melhor ajuste aos dados de treinamento. Os modelos maiores demonstraram capacidade superior na minimização de erro, enquanto os menores apresentaram mais dificuldade para reduzir a *loss*, mesmo com maior tempo de treinamento.

Com os treinos foi possível constatar que o uso de diferentes *labels* pode impactar o resultado final do *fine-tuning* do modelo. O que aponta que utilizar rótulos diferentes podem

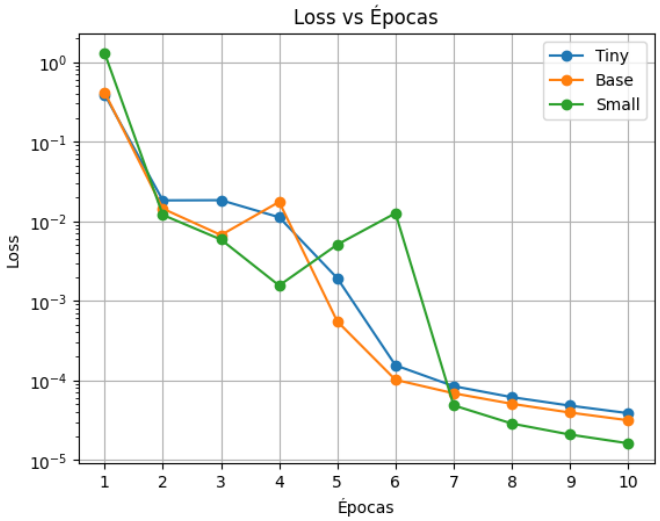

Fig. 3. Loss por época do modelo de veículos e sirenes.

auxiliar na maximização da qualidade do modelo quanto a identificação de sons externos.

Como o dataset de classes de veículos era muito grande, foram utilizadas 300 amostras por classe, o treinamento mostrou o progresso de acurácia foi constante, iniciando em 80% e atingindo 97,5% ao final. A convergência do modelo foi rápida, com queda drástica da perda a partir da segunda época, atingindo 0.0033 ao final, indicando adaptação eficiente aos dados do dataset.

VII. Conclusão

O Audio Classifier demonstrou eficácia na classificação de sons não vocais, atingindo precisão superior a 85% para ruídos urbanos, com o modelo Whisper Small alcançando impressionantes 98,33% nos testes. Os experimentos evidenciaram o potencial do sistema para aplicações em monitoramento ambiental, segurança pública e acessibilidade.

Desafios como sobreposição de sons e qualidade do áudio persistem, afetando a precisão em cenários complexos. A influência da qualidade do áudio de entrada sobre o desempenho indica a necessidade de estratégias mais robustas para lidar com ruídos de fundo e distorções.

Trabalhos futuros podem focar em:

- Otimizações na arquitetura do modelo para melhor desempenho em áudios com sobreposição
- Expansão da base de dados para incluir maior diversidade de sons ambientais
- Implementação de técnicas avançadas de pré-processamento para melhorar robustez
- Experimentação com modelos Whisper de maior capacidade (medium e large)
- Integração com sistemas de alerta em tempo real para aplicações práticas
- Implementação de técnicas que ajudem o modelo atingir a máxima acurácia, como a validação cruzada

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/