AI Music Detection using Deep Learning: Advancing Generalization and Analyzing Current Generators

Jiahui He[1], Tess Ivinjack[1], Navin Lo[1], Brian Lu[1], Nazhah Mir[1], Parker Reedy[1], Avani Tanna[1], Jazer Sibley-Schwartz[1], James O'Brien[2]
[1]University of California, Santa Barbara; [2]Sound Ethics

SOUND ETHICS

UC SANTA BARBARA | Data Science Initiative

Overview

- **Sound Ethics** is a non-profit company that is dedicated to safeguarding artists' rights and promoting ethical AI usage in the music industry
- The overarching goal of this project is to **gain valuable insights** into commercial AI music generators and to **develop a robust model**—built on current state-of-the-art architectures—capable of distinguishing between AI-generated and real music
- To achieve this, we expanded on the limited existing research by exploring diverse generator types and model architectures, enabling a deeper understanding of current limitations and advancing real-world model reliability

Datasets

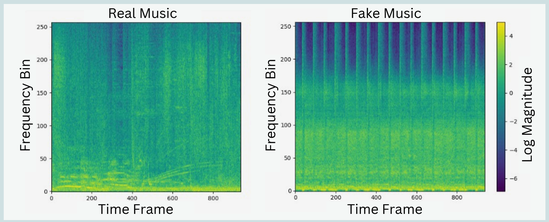

Fig 1. Spectrograms of a real song sample (left) and an AI generated song sample (right)

- **FakeMusicCaps dataset**: 5,521 YouTube scraped audio clips with captions fed through 5 text-to-music models to generate 27,605 AI counterparts
- **Sonics dataset**: 48,090 real songs scraped from YouTube and 49,074 fake songs generated from Suno and Udio using lyrics/music style prompts
- Fake music is made by modifying existing songs (**auto-encoder models**) or creating new songs from scratch using layered stems (**stem-based models**)

Methodology 1: Deezer

- First step toward music deepfake detection
- 6-Layer CNN using mel spectrograms
- **Strengths**: High binary accuracy in controlled settings. Detects decoder artifacts
- **Limitations**: Struggles with unseen generators. Also lacks robustness, calibration, and interpretability
- **Key Insight**: Complex models aren't enough, generalization must be tackled

Methodology 2: MusicCaps

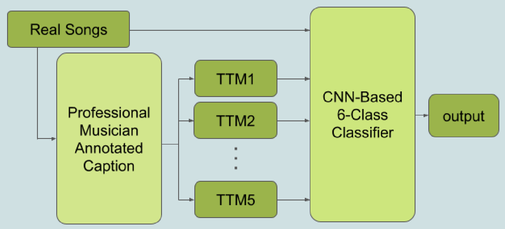

Fig 2. Pipeline of FakeMusicCaps' model

- Attribution and detection of music made by Text-To-Music generators
- ResNet18 + Spec (CNN) on 10s mel spectrograms
- **Strength**: Multi-Class prediction
- **Limitations**: Fails in open-set testing (misclassified unknown generators as real)
- **Key Insights**: Detection on open-source encoders work, but failed to address commercial generators

Methodology 3: SONICS

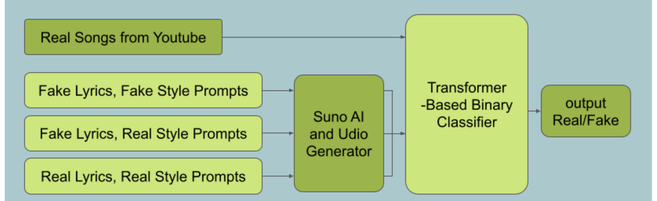

Fig 3. Pipeline of SONICS' model

- Detection of full-length end-to-end synthetic songs.
- Transformer based-model using spectro-temporal tokens. Efficient long-range audio modeling
- **Strengths**: Handles long durations, Detects long-term temporal structure in songs. Contains two models with differing input length; 5s focusing more on encoder artifacts and 120s focusing on song composition
- **Limitations**: Generalization is better than previous models but still inaccurate. Fails to detect hybrid songs where only some components are synthetic. Trained using 2 generators limiting diversity of training data
- **Our goals**: Improve generalization by fine-tuning with additional training data as well as gaining insights into untrained generators by analyzing whether their outputs share structural similarities with known generators

Insights

Our project consists of two main outcomes:

- Insights, where we applied existing methods to analyze a broad range of commercial AI music generators; and
- Results, where we addressed the generalization challenge in the state-of-the-art model, SONICS

For **Insights**, we tested both the FakeMusicCaps and SONICS models on the 9 most commonly used AI music generators (Table 1):

Generator	SONICS 5s (%Fake)	SONICS 120s (%Fake)	FakeMusicCaps (Majority Generator)
Suno V4	**85.7**	**85.7**	Inconclusive
Udio	**27.3**	**100.0**	Inconclusive
Mureka	**90.0**	**100.0**	Inconclusive
Tad.ai	10.0	40.0	**Mustango**
Soundverse	15.0	90.0	Inconclusive
Mubert	14.3	57.1	Inconclusive
Jenmusic	9.1	54.5	Inconclusive
Beatoven	0.0	70.0	**MusicGen**
Boomy	**0.0**	**0.0**	Inconclusive

Table 1. Testing results with FakeMusicCaps and SONICS on commercial generators

- The 120-second model outperforms the 5-second model in overall accuracy, as expected. Additionally, Suno and Udio, being part of the training set, naturally exhibit higher classification accuracy
- Udio may have **modified its decoder** artifacts to evade detection, as suggested by its notably lower 5-second model accuracy
- Mureka's results closely resemble those of Suno, raising the possibility that **Mureka leverages Suno's backend**
- Boomy scored 0% on both models, indicating it may rely on **stem-based composition** rather than an encoder-based generation approach
- FakeMusicCaps yielded inconclusive results, likely due to its **emphasis on open-source encoders** rather than commercial-generation methods
- Notable trends in the classifications of Tad.ai and Beatoven suggest their encoders may be derived from, or influenced by, corresponding open-source models

Results

For our model improvement result, we observe the following **Results** (Table 2):

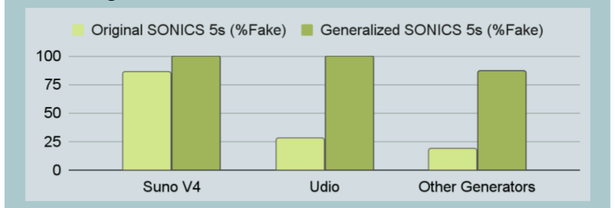

Fig 4. Results from tuning SONICS model with expanded training data

- The generalized model showed improved detection of fake songs across a wider range of generators.
- The generalized model also accounted for suspected artifact changes in Udio.
- For unseen generators (Others), accuracy rose from 18.4% to 86.8%, indicating strong generalization.

Future Steps

Data Diversity & Augmentation

- Expand AI music generator coverage and music dataset to tackle the generalization issue
- Apply advanced audio augmentation techniques (e.g., pitch shifting, time stretching) to increase input variability and robustness

Granular Classification & Analysis

- Perform subclassification based on varying degrees of "fakeness" (vocals, instrumentals, etc.) once a sufficiently diverse sample set is available

Deployment & Practical Integration

- Implement the next state-of-the-art model and transition it into a functional, real-world application

Acknowledgements and References

Special thanks to our mentors, **Avani Tanna** and **Jazer Sibley-Schwartz,** our advisor **James O'Brien**, and **Sound Ethics** for their support and guidance throughout the process

Afchar, D., Meseguer-Brocal, G., & Hennequin, R. (2024). Detecting music deepfakes is easy but actually hard.
Comanducci, L., Bestagini, P., & Tubaro, S. (2024). FakeMusicCaps: A dataset for detection and attribution of synthetic music generated via text-to-music models
Rahman, M. A., Hakim, Z. I. A., Sarker, N. H., Paul, B., & Fattah, S. A. (2024). SONICS: Synthetic Or Not – Identifying Counterfeit Songs