# Comparative Analysis of Hate Classification in Audio: A Study of Waveform vs. Spectrogram-based Approaches

Damian Kopp, Luca Engel, Nino Gerber

**Group 45**

## Problem definition

- **Objective:** Compare hate speech classification using raw waveforms and spectrogram-based approaches in audio data.
- **Hypothesis:** Spectrogram-based models will outperform raw waveform models.
- **Data Source:** Text-to-speech synthesized audio from social media comments.

## Key Related Works

•**Traditional ML**: Logistic Regression, Random Forest, SVM for audio hate speech detection.
•**wav2vec2**: Meta's model for learning speech representations.
•**AST**: MIT's attention-based model for audio classification.
•**Coqui-TTS**: Open-source Text-to-Speech generator.

## Method

- **Models:**
  - **Wav2vec2:** Fine-tuned on raw audio waveforms.
  - **AST:** Fine-tuned on spectrograms treated as images.
  - **DistilBERT:** Fine-tuned on text data for comparison.
- **Preprocessing:** Text extraction, hate speech score assignment, text-to-speech synthesis using Coqui-TTS**.**
- **Training Configurations:** Varied learning rates, batch sizes, and gradient accumulation steps due to hardware constraints

| Parameter | wav2vec2 | AST | DistilBERT | Parameter |
|---|---|---|---|---|
| Learning Rate | 4E-5 | 3E-5 | 8E-5 | Learning Rate |
| Batch Size | 32 | 16 | 32 | Batch Size |
| Gradient Accumulation Steps (GAS) | 4 | 2 | 4 | Gradient Accumulation Steps (GAS) |

## Dataset(s)

•**Source**: UC Berkeley D-Lab "Measuring Hate Speech".
•**Details**: 39,565 samples, balanced for hate and non-hate speech, synthesized for audio.

## Validation

- **Models Used:** wav2vec2 (raw waveforms), AST (spectrograms), DistilBERT (text).
- **Key Metrics:**

-Validation Set

| Model | Loss | Accuracy | Recall | Precision | F1 Score |
|---|---|---|---|---|---|
| Wav2vec2 | 0.656 | 0.622 | 0.785 | 0.599 | 0.680 |
| AST | 0.631 | 0.649 | 0.837 | 0.614 | 0.708 |
| DistilBERT | 0.998 | 0.774 | 0.812 | 0.753 | 0.781 |

-Test Set

| Model | Loss | Accuracy | Recall | Precision | F1 Score |
|---|---|---|---|---|---|
| Wav2vec2 | 0.660 | 0.619 | 0.782 | 0.594 | 0.676 |
| AST | 0.644 | 0.632 | 0.819 | 0.600 | 0.693 |
| DistilBERT | 1.064 | 0.754 | 0.793 | 0.741 | 0.766 |

## Limitations

•**Synthetic Data**: May not fully capture nuances of real human speech.
•**Computational Constraints**: Affected model performance due to reduced batch sizes and steps.

## Conclusion

- **Findings:** Audio classification is effective for hate speech detection.
- **Performance:** AST slightly better than wav2vec2; both close to DistilBERT.
- **Implications:** Potential for further research with larger, diverse, human speech datasets.

**References**

[1] Y. Gong, Y. Chung, and J. R. Glass, “AST: Audio Spectrogram Transformer,” CoRR, vol. abs/2104.01778, 2021.
[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Advances in Neural Information Processing Systems (NeurIPS), 2020.
[3] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," CoRR, vol. abs/1910.01108, 2019.