Decoding Narrator Identity: Machine Learning Approaches to Audiobook Classification

Hannah Nguyen, Minh Le, Khoa Ho - Data Analytics Program - Research Advisor: Dr. Matthew Lavin

INTRODUCTION

Scan Me

- **The digital revolution:** The shift from print to digital platforms in literature consumption
- **Optimize audio analysis:** Machine-learning for predicting audiobook narrators' identities and styles.
- **Project scope:** Use feature extraction and Convolutional Neural Networks (CNNs) model to categorize narrators' identities based on their audio.

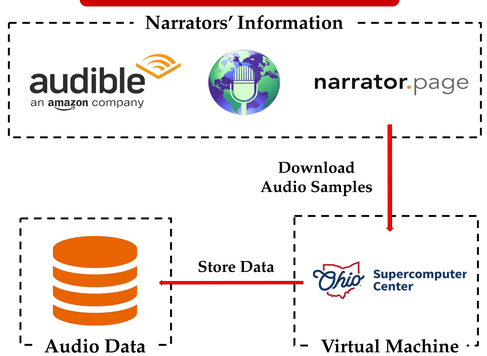
DATA COLLECTION

Exploration

- **Convenience Sample Dataset:** 38 sound files for initial exploration.
- **Each entry includes key metadata:** author, title, narrator, category, etc.

Book	Narrator	Category
The Dutch House	Tom Hanks	Celebrity
Harry Potter and the Goblet of Fire	Stephen Fry	Impersonating
Gilead	Tim Jerome	Elderly
Snow Crash	Jonathan Davis	Middle age
The Dharma Bums	Ethan Hawke	US Accent

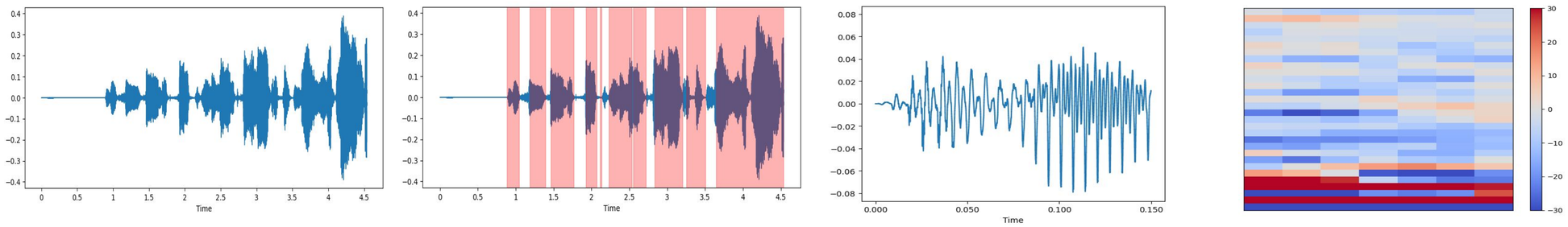
Automation



- Audio and narrator data is taken from Audible, NarratorList.com and narrator.page.
- We utilized the Ohio Supercomputing Center (OSC) to reduce pipeline runtime.
- Data stored for processing before model input.

DATA PROCESSING

1. Original Excerpt Audio Signal → 2. Used Energy & Zero crossing rate (ZCR) threshold method to detect the voiced and unvoiced parts → 3. Segmented audio into 0.15 second time windows → 4. Converted each segment into Mel-Frequency Cepstral Coefficients (MFCC) images



MODEL IMPLEMENTATION

Illustration of the input image and its pixel representation

The kernel's weights are displayed on a grid, with black squares representing zero weights and white squares representing ones

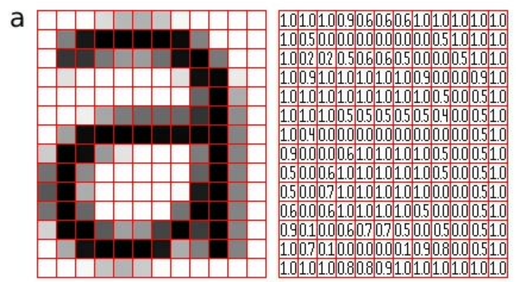
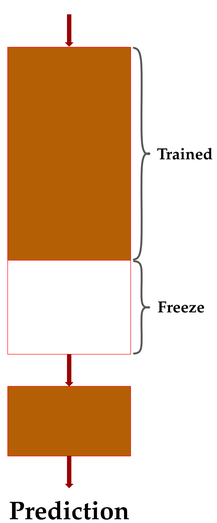


Fig 1. From Jain, R. (2018). Convolutional neural networks explained. Towards Data Science.



General Structures:

- Accepts 224 x 224 RGB images as input.
- Uses pre-trained model for feature extraction.
- Employs dropout for regularization.
- Outputs a single probability score for classification.

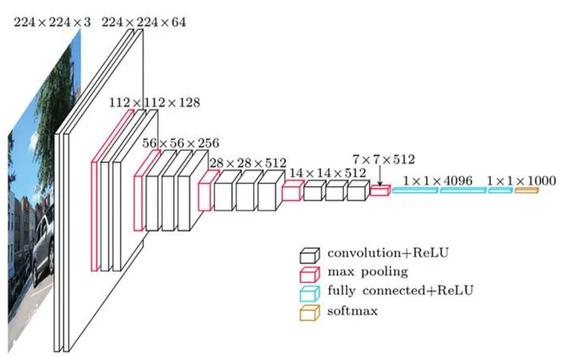


Fig 2. Architecture of the VGG-19. From Masood, D. (2020). Pre-train CNN architectures: Designs, performance analysis, and comparison. Medium.

RESULTS

Optimizer	Early stopping	Dense Layer Activation	Per Segment Accuracy	Per Audio Accuracy*
Adam	No	Sigmoid	0.66	0.67
Adam	Yes	Sigmoid	0.61	0.73
Adam	No	Relu	0.46	0.4
Adam	Yes	Relu	0.6	0.64
RMSprop	No	Sigmoid	0.64	0.64
RMSprop	Yes	Sigmoid	0.64	0.64

*Note: Per audio result were created by choosing the higher total number of segments resulted as male or female

Key Findings:

Best Segment Accuracy	Adam optimizer + Sigmoid activation (66%)
Best Full Audio Accuracy	Adam optimizer + Sigmoid activation + early stopping (73%)
Optimizer Comparison	RMSprop delivered consistent results but didn't outperform Adam.

DISCUSSION

- **Potential Limitations:** Model types, small sample size, sample bias, segmentation method.
- **Future Directions:** Since the accuracies of the current approach are not noticeably high among variety of settings, we suggested 2 main targets: Implementation of different feature extractions or models.

FEATURE

- Apply other data preprocessing methods
- Segment based on words - Vosk Model
- Use model as feature extractor

MODEL

- Add more layers to the current model
- New Pre-trained models: YAMNet, Vggish, OpenL3
- Self-trained model

ACKNOWLEDGMENTS

We would like to express our gratitude to *The Reid and Polly Anderson Endowment*, *The Laurie Bukovac and David Hodgson Endowed Fund*, and *The Denison University Research Foundation* for their support in facilitating this research.

WORKS CITED

tinyurl.com/citation-anderson