Leveraging Image Based Machine Learning Techniques to Determine the Predominant Spoken Language

Kavya Jain[1], Nicolas Echevarrieta Catalan[1], Alicia Bilbao Martinez[1], Laura Vitale[2], Daniel S. Messinger[2], Vanessa Aguiar-Pulido[1]

[1]Department of Computer Science, University of Miami, Coral Gables, FL, USA

[2]Department of Psychology, University of Miami, Coral Gables, FL, USA

MIAMI

INTRODUCTION

Language plays a critical role in communication and education, especially in multicultural societies like that of Miami, where children from diverse linguistic backgrounds coexist and mix languages in elementary school settings. Understanding the spoken language of students is essential for effective classroom support. Past studies that aim to solve similar problems use Hidden Markov Models and Deep Learning to identify spoken language with audio recognition[1], and Deep Learning via image-based spectrogram analysis for instrument detection in polyphonic music[2].

The **objective** of this project is to develop a machine learning method to accurately predict the predominant language of speech of audio files by transforming audio into spectrograms and leveraging image-based classification machine learning techniques (Figure 1).

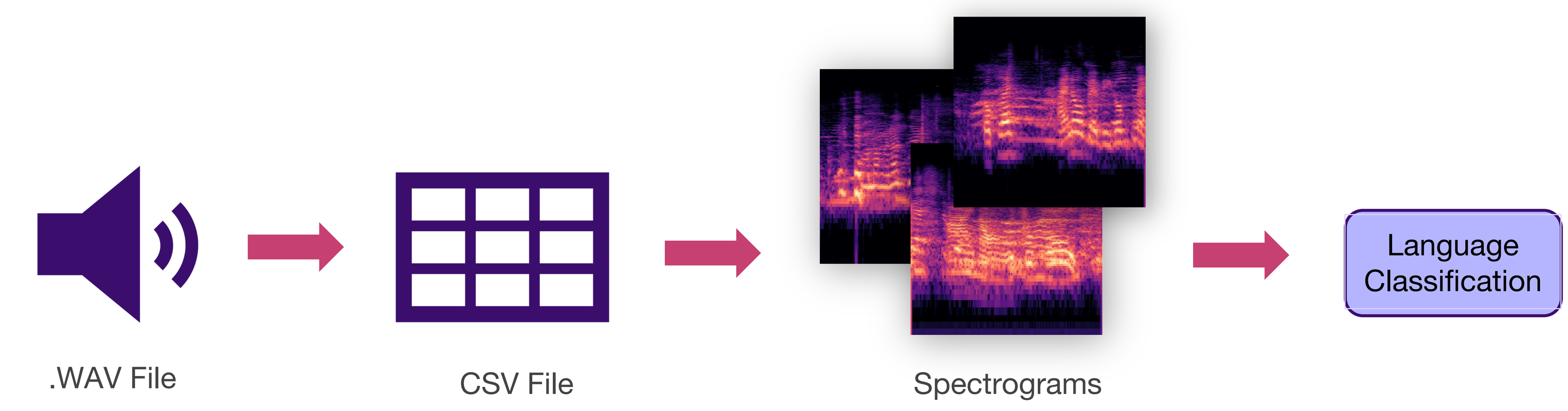

Figure 1 | **General approach for converting audio files to images**

MATERIALS AND METHODS

<u>Creating an Original Dataset</u>

1. Data Collection: Obtained audio recordings from elementary schools in Miami that consist of spoken language from children. Ensured that the recordings include a mixture of Spanish and English speech to reflect the linguistic diversity in the schools.
2. Annotation: We carefully listened to each audio clip and manually annotated the predominant language spoken in each sentence in a corresponding CSV file.
3. Preprocessing: The audio file was first sliced into segments as determined by the data stored in the CSV. To maintain homogeny of data, sentences less than 2 seconds were discarded, and sentences greater than two seconds were converted to 2 second segments.
4. Spectrogram Generation: Transformed each audio file segment into a spectrogram representation utilizing a signal processing library called Librosa[4]. A spectrogram therefore converted the audio signal into an image.

<u>Approach</u>

- The “original dataset” described was utilized as input to a Convolutional Neural Network (CNN). CNNs are a type of artificial neural network utilized to analyze spatial data, particularly images, that have local dependencies.
- We divided the full dataset into train (80%), test (10%), and validation (10%) and employed stratified K-fold cross-validation.
- We used ResNet18[3], which consists of 18 total layers, including 13 convolutional layers, 5 max-pooling layers, and 3 fully connected layers, to classify the images.
- We employed transfer learning with the ResNet18 model trained on the ImageNet dataset as our source model. Transfer learning allows incorporating prior knowledge by initializing a model with the majority of parameters from an already trained model (Figure 2)., thus requiring less data to train.

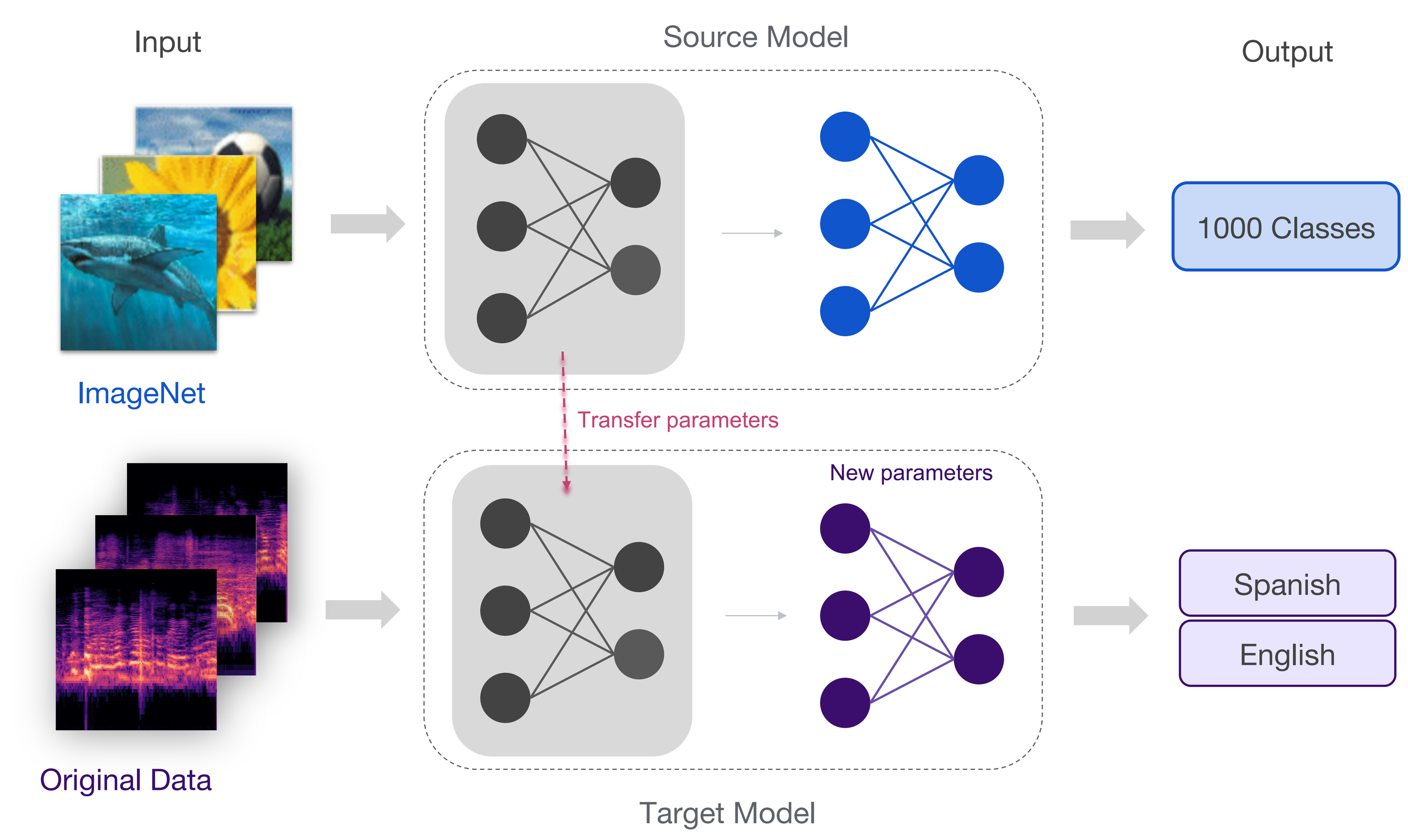

Figure 2 | **Transfer learning with ResNet18 trained on ImageNet**

PRELIMINARY RESULTS

A first model was implemented and tested on the data collected so far as proof-of-concept. Preliminary results of the model's performance are shown in Figure 3.

a)

Accuracy	F1	ROC_AUC	PR_AUC
0.64091204	0.68856683	0.65168073	0.74990324

b)

		Predicted Values	
		Negative	Positive
Actual Values	Negative	60.6	70.6
	Positive	18.6	98.6

Figure 3 | **Performance of the proof-of-concept classifier**
a. | Performance metrics for the ResNet18 proof-of-concept model
b. | Confusion Matrix for the data set

CONCLUSIONS

- We devised a novel approach that utilizes mixed language audio clips for language identification. We implemented a first proof-of-concept classifier based on ResNet-18.
- Instead of using the audio files directly as input, we proposed transforming audio into spectrograms and applying image-based classification techniques using Convolutional Neural Networks (CNNs).
- Our proof-of-concept model achieved 64% cross-validation accuracy when employing the collected data set.
- The results suggest that this could be an effective strategy in predicting the predominant language of speech in elementary school settings with linguistic diversity, such as those in Miami.

FUTURE WORK

- <u>Perform more comprehensive testing:</u> Carry out hyperparameter optimization and assess the performance of all possible combinations to select the best configuration possible.
- <u>Test other CNNs as basis for the classifier:</u> Test a more comprehensive set of CNNs for the classifier component of the proposed approach.
- <u>Expand the size of the dataset:</u> We are currently collecting more data to increase the size of the input dataset. Utilizing a larger dataset should enhance the model's performance and generalization capabilities.
- <u>Expand the diversity the dataset:</u> Using audio recordings from different schools, age groups, demographics, linguistic accents, and areas of Miami would allow for a more robust model.
- <u>Multilingual support:</u> Expanding the dataset to include other languages spoken in the area: Haitian Creole and Brazilian Portuguese, for instance.
- <u>Remove noise:</u> Utilize the *Librosa*[4] library to clean the data from noise in the preprocessing stage to improve data quality, thus potentially impacting the overall accuracy of the model.
- <u>Integration with other classroom data and models</u>: Investigate combining with the related *Machine Learning Classification Characterization of Autism Disorder in Preschool Aged Children* project to provide more data and support to educators.

ACKNOWLEDGEMENTS

Support for the program REU Site: Scientific Computing for Structure in Big or Complex Datasets is by the NSF grant CNS-1949972, program solicitation NSF 13-582, Research Experiences for Undergraduates (REU).

National Science Foundation

python

PyTorch

REFERENCES

1. Gazeau, Valentin, and Cihan Varol. 2018. ‘Automatic Spoken Language Recognition with Neural Networks’, International Journal of Information Technology and Computer Science, 10.8: 11–17 https://doi.org/10.5815/ijitcs.2018.08.02
2. Gururani, Siddharth, Cameron Summers and Alexander Lerch. 2018. ‘Assessment of Student Music Performances Using Deep Neural Networks’, Applied Sciences, 8.4: 507 https://doi.org/10.3390/app8040507
3. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. ‘Deep Residual Learning for Image Recognition’, ArXiv.org https://arxiv.org/abs/1512.03385
4. McFee, Brian, Matt McVicar, Daniel Faronbi, Iran Roman, Matan Gover, and others. 2023. ‘Librosa/Librosa: 0.10.0.Post2’, Zenodo https://doi.org/10.5281/zenodo.7746972
5. Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, and others. 2019. ‘PyTorch: An Imperative Style, High-Performance Deep Learning Library’, ArXiv.org https://arxiv.org/abs/1912.01703