Esmat Sahak, Maxx Wu, Amy Xin
Team Langang

ECE324 Project Final Report

Word Count: 1895

Introduction

In recent years, natural language processing (NLP) has become the forefront of machine learning research and innovation. NLP has accelerated progress in various areas such as automatic language identification (ALI). This project focuses on ALI, with the objective of constructing a classifier differentiating between English, Persian, and Mandarin audio samples.

ALI serves many practical purposes, particularly in speech recognition tasks and multilingual translation systems. For this reason, many large corporations such as Amazon, Google, and Apple, are investing resources into ALI research. For instance, in 2017, Amazon launched Amazon Transcribe, which implements ALI methods to convert speech into text. Given it's a relatively new field, this project will expose us to a relevant technical problem, making the learning experience more valuable.

Illustrations

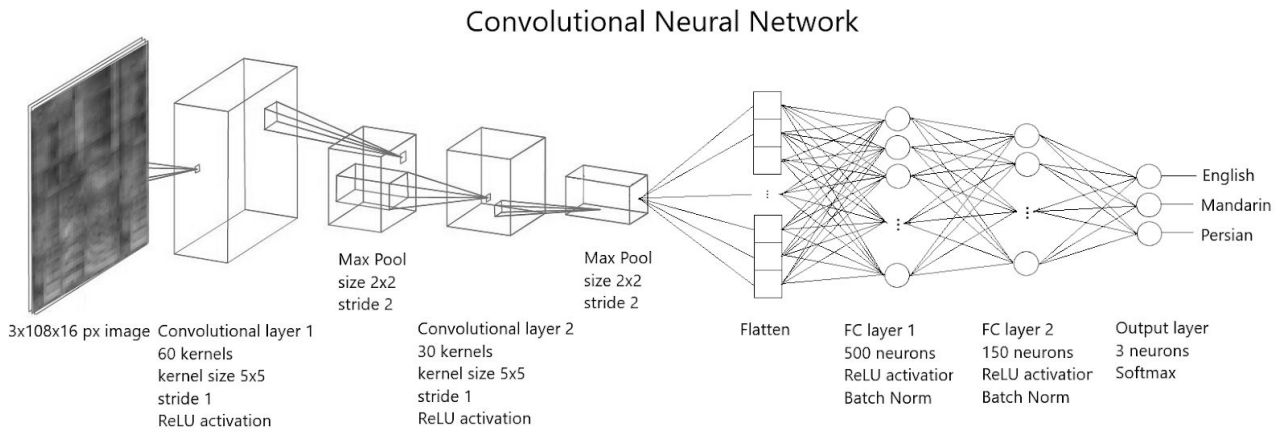

Figure 1. *CNN architecture illustration.*

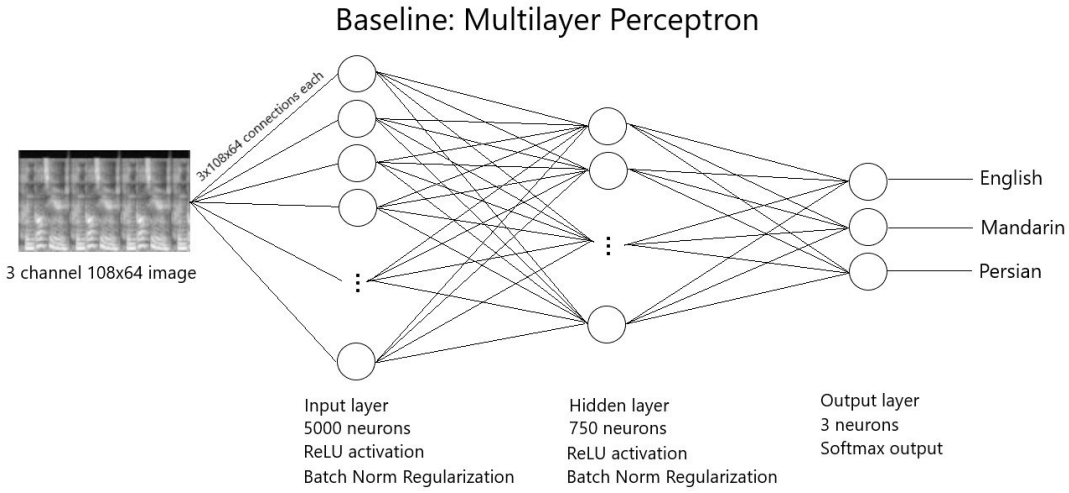

Figure 2. *Baseline architecture illustration.*

Esmat Sahak, Maxx Wu, Amy Xin
Team Langang

Background & Related Work

1) Stanford University in Spoken Language Classification [1]
Examines three machine learning models for ALI. Audio is divided into 25 ms frames and the cepstrum coefficients are extracted as "features" of the sample. The data is trained using Linear Support Vector Machine (SVM), Gaussian Mixture Model, and Multilayer Perceptron (MLP). The conclusion was that the neural network performed best achieving over 98% accuracy with 1000 samples per language.

2) Novetta in Language Identification [2]
This research uses the Residual Neural Network model for the image classification of log-Mel-spectrograms generated from raw audio. The network used pretrained ResNet50 architecture. During training, the learning rate is gradually increased then decreased linearly in each cycle of 8 epochs. The experiment tested two methods, the binary classification achieved 97% accuracy and the multiple language classification achieved 89% accuracy. Each method used 5000 training samples and 2000 validation samples.

Data and Data Processing

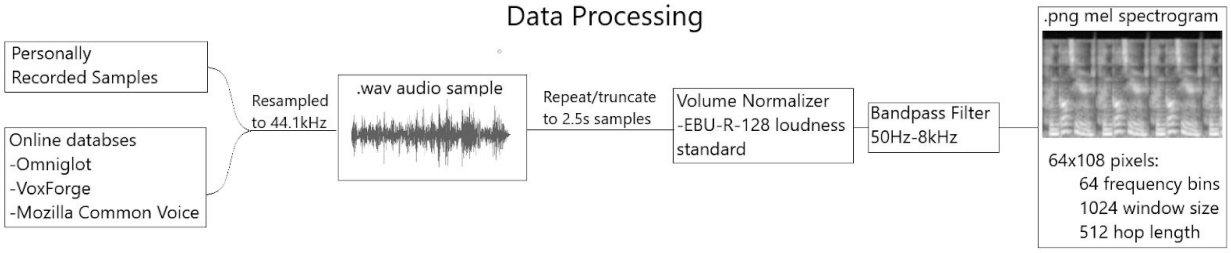

Figure 3. *Process of data collection and processing.*

Audio samples are collected by personal recordings of team members, friends, and family speaking a select set of 19 phrases per person. Audio is also collected from open source databases such as VoxForge, Omniglot, and Mozilla Common Voice. In total, there are 500 audio samples for each language. Each sample is 2.5s long and sampled to 44.1kHz. Any audio less than 2.5s is repeated until the targeted length, and any audio more than 2.5s are truncated.

To clean the data, the MATLAB Signal Analyzer app and Audio Toolkit are used. Audio samples are normalized by volume in accordance to the EBU-R-128 standard, which is an international loudness standard for television and radio programs. The standard defines the target loudness level as -23 LUFS [3][4]. This makes sure that all samples have consistent volume. Then, the audio signals are run through a bandpass filter in which we define to be between 50Hz - 8kHz. This frequency range is around the frequency range of human speech [5][6]. The bandpass filter is useful in mitigating interference of background noise in the audio.

Then, we build the mel spectrograms for each cleaned audio using the Python Librosa library. It is a Fast Fourier Transform of audio data from the time domain to the frequency spectrum. It uses the mel scale, which is a nonlinear transformation of the frequency scale like how humans perceive pitches[7]. Hence, it is appropriate in the context of language recognition. Based on previous works, we used a window size of 1024 to compute the frequency spectrum, 512 shifting length for overlapping with previous windows, and

64 frequency bins which dictate the height of our final image. As a result, we created 64 by 108 pixel images of mel spectrograms for our final datasets.

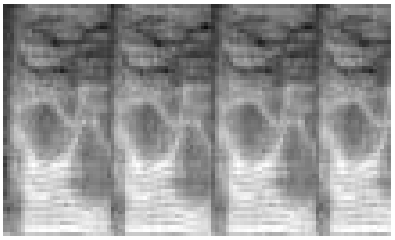

Figure 4. *English mel spectrogram*

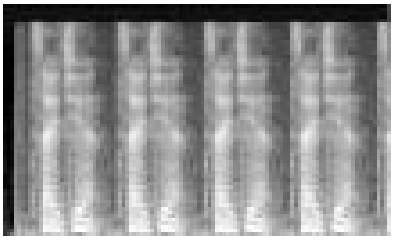

Figure 5. *Mandarin mel spectrogram*

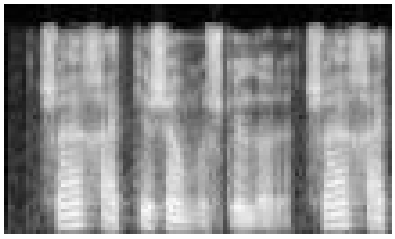

Figure 6. *Persian mel spectrogram*

Finally, the data is split to 60% training, 20% validation, and 20% test sets. Each dataset has an equal distribution of each language and the same speakers do not appear in multiple datasets to encourage generalization of our models.

The following shows the distribution of sex for the data collected. Due to the limitations of the available online data, which consists of voluntary audio submissions, the distribution between male and female voices could not be balanced.

Table 1. *Distribution of sex from online data.*

	Male	Female
English	77.05%	22.95%
Mandarin	83.61%	16.39%
Persian	91.57%	8.43%

Table 2. *Distribution of sex from personal data.*

	Male	Female
English	66.67%	33.33%
Mandarin	64.19%	35.81%
Persian	66.67%	33.33%

Architecture

The final model is a Convolutional Neural Network consisting of 2 convolutional layers and 3 fully connected layers, as depicted in Figure 1. The hyperparameters associated with the architecture were derived from an automated grid search over a range of sensible values, with the best results summarized in Table 3.

Table 3. *Best CNN model hyperparameters derived from grid search.*

Convolutional layer 1 kernel number	60
Convolutional layer 2 kernel number	30
Fully connected layer 1 neuron number	500
Fully connected layer 2 neuron number	150
Kernel size	5x5
Max pool	2x2, stride 2

Other hyperparameters associated with the training of the model were tested and adjusted manually, with the best results summarized in Table 4.

Table 4. *Best CNN model hyperparameters derived from manual adjustment.*

Learning rate	0.001
Epoch number	44
Optimizer	torch.optim.Adam
Loss Function	torch.nn.CrossEntropyLoss
Activation Function	torch.nn.functional.relu
Batch size	20

During training, a dropout layer was used immediately after the convolutional layers with a dropout probability of 0.5. Finally, batch norm regularization was used on the first two fully connected layers and a softmax output function was used on the output layer.

Baseline Model

The baseline model was selected to be a MLP consisting of one hidden layer; specific details regarding its architecture are highlighted in Figure 2. This is an appropriate baseline for comparison to the CNN model, as one can determine the effectiveness of the convolutional layers which precede the fully connected layers. A manual hyperparameter search was conducted in an effort to optimize accuracy. Hyperparameters are summarized in Table 5.

Table 5. *Best baseline model hyperparameters.*

Input layer neurons	5000
Hidden layer neurons	750
Learning rate	0.001
Epoch number	23
Optimizer	torch.optim.Adam
Loss function	torch.nn.MSELoss
Activation function	torch.nn.functional.relu
Batch size	20

Quantitative Results

The training loss and accuracy plots for the MLP are shown in Figure 7.

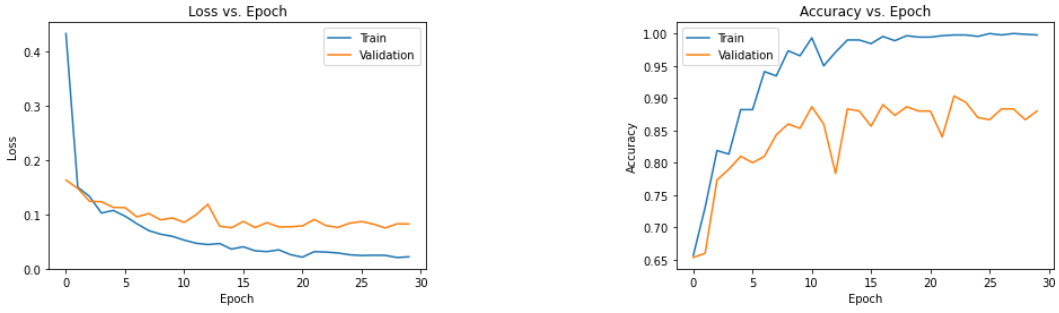

Figure 7. *MLP training loss and accuracy plots.*

The MLP successfully reduces the loss and achieves 100% training accuracy given sufficient epochs, suggesting a functioning model and viable dataset. The accuracy plot shows signs of slight overfitting as the validation accuracy levels off around 90%. Potential solutions for this include increasing the dataset, adding regularization, further hyperparameter tuning, and reducing the model size.

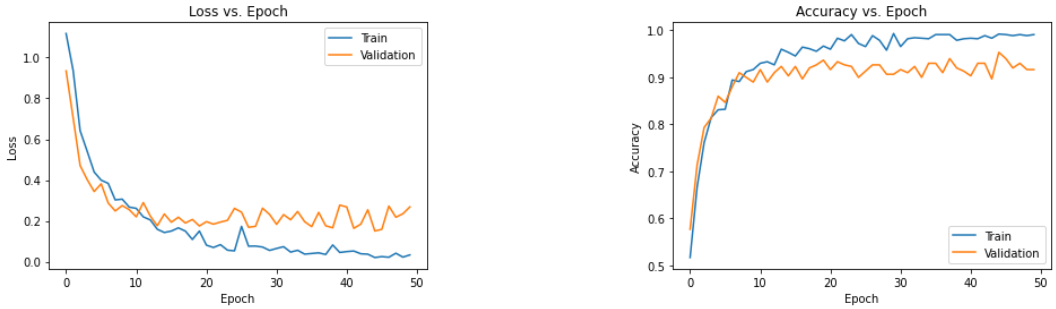

Figure 8. *CNN training loss and accuracy plots.*

The CNN accuracy curve is representative of a good fit, with a validation accuracy just below the training accuracy. The performance of the best MLP and CNN models are summarized in Table 6.

Table 6. *Best MLP and CNN quantitative results.*

Metric	MLP	CNN
Training Accuracy	1.0	1.0
Validation Accuracy	0.90	0.95
Test Accuracy	**0.84**	**0.91**
Parameter Size	**409.89MB**	**18.34MB**
Training Time	21min	17min

The most significant measures for evaluating model performance are the test accuracies and parameter size. The CNN achieved a test accuracy of 0.91 which is higher than the 0.84 achieved by the MLP. Furthermore, the size of the CNN model is ~20x smaller than the MLP. In both of these aspects, the CNN outperforms the MLP model.

Qualitative Results

Figure 9 shows the confusion matrix for the model predictions on the 300 sample test dataset. The model makes the correct prediction with over 90% accuracy, with Mandarin having the most false predictions among the three languages.

		Predictions		
		P	E	M
Truth	P	92	3	5
	E	1	95	4
	M	5	9	86

Figure 9. *Confusion matrix from test dataset.*

Figure 10 shows a sample of 5 model predictions all with label 0 (English). The second to fifth samples were correctly predicted with fairly high “probability”. The first sample was falsely predicted as class 2 (Persian). The corresponding input spectrogram for the first sample is shown in Figure 11.

In general, most predictions were of high probability which is expected as the chosen languages are from different language families. Intuitively, the distinctness of the languages should make the classification task easier which results in high probability predictions. The falsely predicted sample was a spectrogram with much more dark pixels compared to others. The corresponding audio sample had a period of silence at the start of the clip which is likely the cause of this. A data processing step to remove silences from audio clips can be implemented to address this source of error.

```
[3.6035e-04, 1.2703e-02, 9.8694e-01],
[7.9277e-01, 2.0675e-01, 4.7844e-04],
[9.9998e-01, 1.1644e-05, 1.0227e-05],
[8.3526e-01, 4.4846e-03, 1.6025e-01],
[9.9990e-01, 1.0096e-04, 4.5328e-08],
```

Figure 10. *Sample outputs with label 0.*

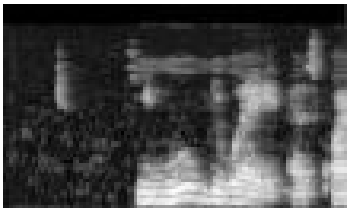

Figure 11. *Spectrogram of false prediction.*

Discussion and Learnings

Overall, the CNN model performed better than the baseline given its higher accuracy and lower memory requirement.As an extension for learning, the CNN model was used to make predictions for samples of different languages that were not trained for. 10 samples each of German, Arabic, and Japanese were taken from the Omniglot database and processed the same way. The results are summarized in the confusion matrix in Figure 12. The model predicted German as English all 10 times and never falsely predicted Arabic or Japanese as English. Even though the sample size is too small to draw concrete conclusions, this does show that some languages share many similarities. It also suggests that the current model can be used in transfer learning to train for German audio samples.

		Predictions		
		P	E	M
Truth	A	2	0	8
	G	0	10	0
	J	4	0	6

Figure 12. *Confusion matrix for new language samples.*

This project demonstrated the versatility of image classification, as it was implemented in a NLP context. This introduced a new perspective when considering classification problems - if data can be mapped to images, then a CNN can be used to train the model. In fact, there exist standard methods to convert non-image data into image data such as DeepInsight, which uses dimensionality reduction techniques to map data points to pixels [8].

To improve the model, the following steps can be taken:

1. Collect more data with a more diversified set of speakers. This helps generalize the neural network by accommodating for a wider range of voices.

2. Ensure audio samples are over 3 seconds in length to avoid repeating audio samples. Repetition results in more effort required for manual preprocessing and may bias the neural network into associating some importance to repetition.
3. Explore different spectrogram parameters including sampling rate, length of FFT window, and hop length.

Ethical Framework

Stakeholder	**Principles/Stakeholder Interest**
Team Langang (from the viewpoint of the project as a learning experience)	**Beneficence** ● Members of the project team learn about principles in machine learning and data science. **Justice** ● Equal distribution of work among team members.
Teaching Team	**Beneficence** ● Members of the teaching team provide the resources and knowledge for students to successfully engage in machine learning techniques and understanding.
Speakers (those providing the data)	**Autonomy** ● Speakers must provide data with consent and choose to do so from their own free will. **Non-maleficence** ● Speakers should remain anonymous if they choose so
Translators	**Justice** ● ALI is usually embedded as part of larger translation systems which has the potential of automating one's employment.
Users (those who might use the model to make inferences)	**Non-maleficence** ● When it comes to classifying the language spoken for minority groups or those with accents, misclassification can be discriminating. We want the models to be able to generalize for all kinds of speech.
Environment	**Non-maleficence** ● Expending computational resources on training has environmental consequences. Engineers should also strive to utilize efficient processes to minimize resources.

References

[1] De Mori, M. Faizullah-Khan, C. Holt and S. Pruisken, Spoken Language Classification. Stanford University, 2012. [Online]. Available: http://cs229.stanford.edu/proj2012/DeMoriFaizullahKhanHoltPruisken-SpokenLanguageClassification.pdf [Accessed: 06- Dec- 2020].
[2] S. Revay, "Deep Learning for Language Identification on Audio Signals", Novetta, 2019. [Online]. Available: https://www.novetta.com/2019/05/lifas/. [Accessed: 06- Dec- 2020].
[3]"Loudness Normalization in Accordance with EBU R 128 Standard- MATLAB & Simulink", Mathworks.com. [Online]. Available: https://www.mathworks.com/help/audio/ug/loudness-normalization-in-accordance-with-ebu-r-128-standard.html. [Accessed: 06- Dec- 2020].
[4]"What is LUFS and Why Should I Care", 2020. [Online]. Available: https://www.sweetwater.com/insync/what-is-lufs-and-why-should-i-care/. [Accessed: 06- Dec- 2020].
[5]"The Human Voice and the Frequency Range", Accusonus Blog, 2020. [Online]. Available: https://blog.accusonus.com/pro-audio-production/human-voice-frequency-range/. [Accessed: 06- Dec- 2020].
[6]"human speech noise filter", Signal Processing Stack Exchange, 2012. [Online]. Available: https://dsp.stackexchange.com/questions/2993/human-speech-noise-filter. [Accessed: 06- Dec- 2020].
[7] L. Roberts, "Understanding the Mel Spectrogram", Medium, 2020. [Online]. Available: https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53. [Accessed: 06- Dec- 2020].
[8]A. Ye, "Turning Non-Image Data into Images for Classification is Surprisingly Effective", towards data science, 2020. [Online]. Available: https://towardsdatascience.com/turning-non-image-data-into-images-for-classification-is-surprisingly-effective-70ce82cfee27. [Accessed: 06- Dec- 2020].

Permissions

	Esmat Sahak	Maxx Wu	Amy Xin
Post Video	Yes	Yes	Yes
Post Report	Yes	Yes	Yes
Post Source Code	Yes	Yes	Yes