Audio Language Classifier: Model Metrics

Kirsten Regier

1 Gender classifier

1.1 Model structure

Two different models were trained and evaluated for the gender classifier. The structure of each model is shown in Table 1.

The models had identical input and output layers. The shape of the input was (32, 10, 128), which corresponds to the batch size (32) and the VGGish embedding (10, 128). Since the classification was binary, the output layer contained one node, with a sigmoid activation function.

The models differ in the number of hidden layers. Each hidden layer consisted of a dense layer followed by a dropout layer with a dropout rate of 50%. The first model contained a single hidden layer with 128 nodes. The second model contained two hidden layers, the first with 128 nodes, and the second with 64 nodes. Prior to being fed into the output layer, the output from the previous layer was flattened into a 1D array. The shape of the flattened array varied for each model, based on the number of nodes of the prior layer (128 vs 64).

Table 1: Structure of the two gender classifier models. Model 1 has one hidden layer with 128 nodes, while Model 2 has two hidden layers with 128 and 64 nodes, respectively.

Layer	Model 1	Model 2
Input	(32, 10, 128)	(32, 10, 128)
Dense	128 nodes	128 nodes
Dropout	50%	50 %
Dense		64 nodes
Dropout		50%
Flatten	(32, 1280)	(32, 640)
Output	(32, 1)	(32, 1)

1.2 Training time

Each model was set to run for 100 epochs, or until the validation loss stopped decreasing for 5 consecutive epochs. Model 1 trained for 8 epochs, with the lowest validation loss found at epoch 3. Model 2 trained for 7 epochs, with the lowest validation loss at epoch 2. The evaluation metrics below are based on the weights of the models with the lowest validation loss - after epochs 3 and 2, respectively.

1.3 Model evaluation

As shown in Table 2, both models performed quite well on the binary classification task, reaching an accuracy of 98.14% and 97.55%, respectively. Model 1 showed similar performance between both classes, with nearly equal numbers of speakers being misclassified from each class. Model 2

showed more discrepancies between the performance of the two classes. Male voices were identified correctly at a rate of 99.14%, while female voices were identified correctly only 96.5%. However, the precision of the predictions was the reverse - the precision of female voice prediction was 99.14%, while the precision of male voice prediction was 96.02%.

Table 2: Confusion Matrices for the Gender Classifier Models

Model 1		Predicted		Recall
		F	M	
Actual	F	590	10	0.9833
	M	12	572	0.9795
Precision		0.9801	0.9828	
Accuracy				0.9814

Model 2		Predicted		Recall
		F	M	
Actual	F	576	24	0.9650
	M	5	579	0.9914
Precision		0.9914	0.9602	
Accuracy				0.9755

Table 3: Gender Classifiers - Metrics summary

	Model 1	Model 2
loss	0.061640	0.076962
accuracy	0.981419	0.975507
precision	0.982818	0.960199
recall	0.979452	0.991438

Given the slightly better accuracy rates, and the more consistent values of precision and recall between the two classes, it seems that the simpler, one layer model performs slightly better than the more complex, two-layered model for general use cases.

2 Language classifier

2.1 Model structure

Three models were trained and evaluated for the language classifier. The structure of each model is shown in Table 4.

All of the models had identical input and output layers. The shape of the input was (32, 10, 128), which corresponds to the batch size (32) and the VGGish embedding (10, 128). There were 11 possible classes, so the output layer consisted of 11 nodes. A softmax activation function was used on the output layer, so that the output vector contained the probability that the speaker belonged to each of the output classes. The class with the highest probability was take to the be predicted class.

The models differ in the number of nodes and the number of hidden layers. Each hidden layer consisted of a dense layer followed by a dropout layer with a dropout rate of 50%. The first model contained a single hidden layer with 12 nodes, while the second model had a single layer with 128 nodes. The third model contained two hidden layers, the first with 128 nodes, and the second with 64 nodes. Prior to being fed into the output layer, the output from the previous layer was flattened into a 1D array. The shape of the flattened array varied for each model, based on the number of nodes of the prior layer (12 vs 128 vs 64).

Table 4: Structure of the language classifier models. Model 1 has one hidden layer with 12 nodes, Model 2 has one hidden layer with 128 nodes, and Model 3 has two hidden layers with 128 and 64 nodes, respectively.

Layer	Model 1	Model 2	Model 3
Input	(32, 10, 128)	(32, 10, 128)	(32, 10, 128)
Dense Dropout	12 nodes 50%	128 nodes 50%	128 nodes 50 %
Dense Dropout			64 nodes 50%
Flatten	(32, 120)	(32, 1280)	(32, 640)
Output	(32, 11)	(32, 11)	(32, 11)

2.2 Training time

Each model was set to run for 100 epochs, or until the validation loss stopped decreasing for 5 consecutive epochs. Model 1 trained for 19 epochs with lowest validation loss at epoch 15, Model 2 trained for 8 epochs, with the lowest validation loss at epoch 3, and Model 3 trained for 11 epochs with the lowest validation loss at epoch 6.The metrics below are based on the final trained models, with weights from the last trained epoch.

2.3 Model evaluation

The dataset contained many more speakers of English, Spanish and Arabic than speakers of other languages. To balance the classes for the Language Classifier, the number of speakers for these languages was downsampled to 75, which meant that each of these language classes contained 12.0773% of the total number of speakers in the dataset. German had the fewest number of speakers (36), which comprised 5.797% of the data. The distribution of the number of **speakers** sampled per language class is shown in blue in Figure 1. The speakers of each language were split into training, validation and testing sets, so the distribution of speakers in each data split reflects the overall distribution of speakers.

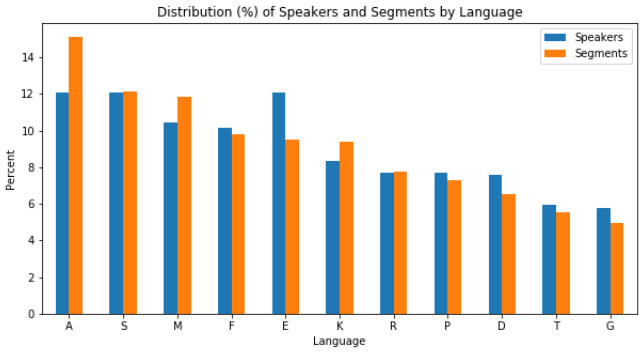

Figure 1: Distribution of Speakers and Segments by Language

After the speakers were assigned to the data splits, the audio files for all speakers were seg-

mented, and the segments in the training and validation splits were augmented with noise. Since the number of segments per speaker was dependent on the length of the original audio file, which was not determined before the speakers were split, the number of segments per language and per data split were not identical to the distribution of the speakers.

While the distribution of speakers and segments for most of the languages remains similar (seen by comparing the heights of the blue and orange bars in Figure 1), the distribution of Arabic and English segments changed dramatically. While there were equal numbers of Arabic and English speakers (75 per language, or 12.07% each), there were 380 Arabic segments and 239 English segments. Thus, the distribution of Arabic segments grew to 15.109344% of the segments, while the distribution of the English segments shrank to 9.502982% of segments, moving English behind Spanish, Mandarin and French in the percentage of distribution of the segments.

2.4 Model Metrics

A naive classifier that always predicted the majority class (Arabic) would have an accuracy of 15%. All three of the models improved upon this baseline accuracy rate, as shown in the second row of Table 5. While Model 1 has the highest precision rate, it has the lowest accuracy, recall and F1 scores. Model 2 has the highest recall and F1 scores, but also the lowest precision. The most complicated model, Model 3, has the highest overall accuracy, and has F1 scores only slightly lower than Model 2.

Table 5: Language Classifiers - Metrics summary

	Language 1	Language 2	Language 3
Loss	2.25485	2.323	2.20684
Accuracy	0.232955	0.238636	0.25
Precision	0.625	0.5	0.605263
Recall	0.0142045	0.0738636	0.0653
F1 Macro	0.16999	0.195512	0.191103
F1 Weighted	0.20067	0.228178	0.224565

The confusion matrix for the Model 3 predictions is shown in Table 6, with the precision and recall rates by class in Table 7. Not surprisingly, the class with the highest number of correct predictions was the majority class, Arabic, with 24/55 segments correctly predicted (recall of 0.43.63%) and 24/52 predictions being correct (precision of 0.46.15%). Dutch, Spanish, English, and Mandarin had 12-15 correct predictions per class. German, the smallest class, along with Russian and Portuguese, had no correct predictions, while Turkish has a single correct prediction.

Behind Arabic, Spanish and Mandarin contributed about the same number of segments to each of the training sets (12.167% and11.84 %, respectively). Interestingly, the model made 63 predictions for Spanish and 62 predictions for Mandarin, which is more than it predicted the majority class. However, since there were fewer correct predictions for Mandarin (15) and Spanish (12), the precision rates for these classes is roughly half that of Arabic.

While Dutch contributed only 6.520875% of the segments overall, almost half of its testing files were correctly predicted (12/25), making it the class with the highest recall (48%).

Table 6: Confusion matrix for Model 3 predictions. The bold numbers on the diagonal represent correct predictions.

lang	R	A	T	K	G	D	S	F	E	P	M	Segments
Russian	**0**	2	1	4	0	1	7	4	1	1	7	28
Arabic	0	**24**	0	4	1	7	2	2	4	0	11	55
Turkish	0	1	**1**	4	0	3	4	0	2	1	5	21
Korean	1	10	0	**6**	0	3	2	2	4	2	1	31
German	0	0	0	1	**0**	0	5	1	5	2	1	15
Dutch	0	4	0	0	0	**12**	3	1	3	1	1	25
Spanish	1	2	0	6	2	1	**12**	3	8	0	7	42
French	1	3	0	4	1	2	10	**5**	1	1	4	32
English	0	1	0	3	0	5	2	2	**13**	0	4	30
Portuguese	0	3	0	3	1	0	7	5	3	**0**	6	28
Mandarin	0	2	0	8	2	4	9	1	2	2	**15**	45
Total	3	52	2	43	7	38	63	26	46	10	62	**352**

Table 7: Model 3 Metrics by Language

Language	Precision	Recall	F1-score	Samples
Russian	0.00	0.00	0.00	28
Arabic	0.46	0.44	0.45	55
Turkish	0.50	0.05	0.09	21
Korean	0.14	0.19	0.16	31
German	0.00	0.00	0.00	15
Dutch	0.32	0.48	0.38	25
Spanish	0.19	0.29	0.23	42
French	0.19	0.16	0.17	32
English	0.28	0.43	0.34	30
Portuguese	0.00	0.00	0.00	28
Mandarin	0.24	0.33	0.28	45
accuracy			0.25	352
macro avg	0.21	0.22	0.19	352
weighted avg	0.23	0.25	0.22	352

2.5 Comments

One question about the language classifier was whether it would identify similarities between languages with historical similarities. For example, Spanish, French and Portuguese are all descendants of Latin, and are more closely related to each other than to other languages in this database, so it would not be surprising for them to be confused with each other more often than with other, unrelated languages. However, the language classifier showed only weak connections between some related languages.

For example, Portuguese samples were misclassified as French 5 times and Spanish 7 times, but were identified as Mandarin 6 times, and misclassified as Arabic, Korean and English 3 times each. French samples were misclassified as Spanish 10 times, but was more likely to be classified as Korean and Mandarin (4 times each) than Portuguese (1). Spanish samples were only misclassified as French 3 times, but as Mandarin 7 times and Korean 6 times, even though Mandarin and Korean are not related to Spanish. Of the samples that were misclassified as Spanish, 10 were actually French and 7 were Portuguese, but 9 were Mandarin and 7 were Russian. So, while there are weak patterns of misclassification between the Romance languages, Mandarin and Korean also show similar levels of misclassification with Spanish, French and Portuguese.