

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	Model 2		Predicted		Recall
		F	M				F	M	
Actual	F	590	10	0.9833	Actual	F	576	24	0.9650
	M	12	572	0.9795		M	5	579	0.9914
Precision		0.9801	0.9828		Precision		0.9914	0.9602	
Accuracy				0.9814	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.

Model 1 has similar rates of mis-categorizing both male and female voices. About the same number of male voices are identified as female as female voices that are identified as male. Likewise, roughly the same percentage of voices identified as female are actually female and identified as male are actually male. (Roughly equal sensitivity, specificity and precision, roughly equal positive predictive value and negative predicted value).

Model 2 has higher sensitivity for male voices and lower specificity for female voices. If you expect most of the callers to be male, if the cost is low for misidentifying females as males,

Model 2 - better sensitivity (recall) for male voices [but worse specificity for female voices] - identifies male voices as male most of the time; better precision for female voices - fewer predicted female voices were not female. Model 2 has high sensitivity for male voices, but lower specificity of female voices. Higher precision for female voices (more cases predicted female are actually female, fewer cases predicted male are actually male.) More likely to categorize a female voice as male than a male voice as female. Model 2 rarely mis-categorizes a male voice as female, but more frequently mis-categorizes a female voice as male. If the model predicts female, the voice is more likely to be female; if the model predicts male, there is slightly less of a chance that the voice is actually male. However, if the voice is actually male, it is more likely that it will be categorized correctly.

2 Language classifier

2.1 Model structure

Several different 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 taken to be the 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	12 nodes	128 nodes	128 nodes
Dropout	50%	50%	50 %
Dense			64 nodes
Dropout			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, and a final accuracy on the test dataset of 23.30%.

Model 2 trained for 8 epochs, with the lowest validation loss at epoch 3, and a final accuracy of 23.86%.

Model 3 trained for 11 epochs with the lowest validation loss at epoch 6, and final accuracy of 25.00%.

For Model 1, the evaluation metrics reported below are based on the model weights after epoch 7 (the best weights), whereas for Models 2 & 3, the metrics are based on the model weights after the final training epoch.

2.3 Model evaluation

2.3.1 Baseline calculations

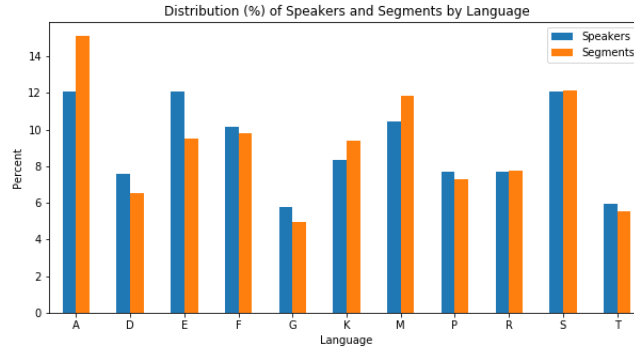


Figure 1: Distribution of Speakers and Segments by Language

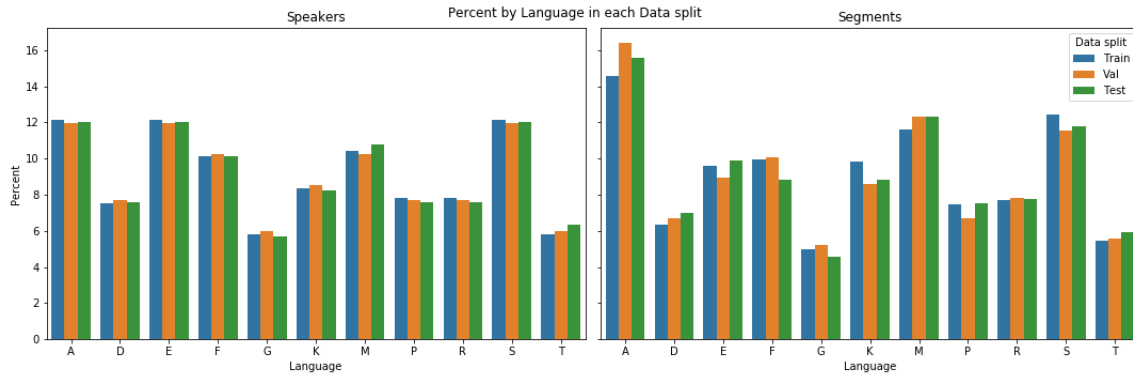


Figure 2: Distribution of Speakers and Segments by Language

2.4 Model Metrics

The confusion matrix for the Model 3 predictions is shown in Table 6. Not surprisingly, the class with the highest correct predictions was the majority class, Arabic, at 28. Mandarin and Dutch also had relatively high correct predictions at 18 and 17, respectively. German, the smallest class, and Russian showed no correct predictions, while Turkish and Portuguese each had a single correct prediction.

Behind Arabic, Mandarin and Spanish contributed about the same number of segments to each of the training sets (12.344656% and 12.676056 %, respectively). While there were more predictions for Mandarin (73) than Spanish (27), and more correctly classified Mandarin samples (18) than Spanish (6), the precision rates for the two classes was similar at .25 for Mandarin and .22 for Spanish.

Dutch is really interesting - there were relatively few Dutch samples, but Dutch was predicted

Table 5: Language Classifiers - Metrics summary 2

	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 micro	0.232955	0.238636	0.25
F1 macro	0.16999	0.195512	0.191103
F1 weighted	0.20067	0.228178	0.224565

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

3rd most often (behind Arabic and Mandarin), and was also third highest in number of correct predictions.

Other "common" confusions: Korean as Arabic - 12 Arabic as Dutch - 10 Arabic as Mandarin - 12

Related languages - that might be similarly confused English vs Dutch - 2 D as E, 6 E as D Eng vs German - 0 E as G, 4 G as E Dutch vs German - 0 D as G, 1 G as D

Span vs port. 2 S as P; 1 P as S 7 P as F; 6 F as P 4 S as F; 4 F as S

2.4.1 Classification report

Model metrics for all models

15.11% of segments were Arabic - majority class in segment data.

Next was Spanish as 12.167, then Mandarin at 11.84, French as 9.82, and English at 9.5.

English has 12% of speakers, but only 9.5% of segments. (Not too surprising, given speaking rate for L2 tends to be slower).

Table 7: Metrics (precision, recall and F1 score) by language class - new

Language	precision	recall	f1-score	support
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
Rrench	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