Iterative Feature Normalization

Surabhi S Nath
2016271

Data Statistics

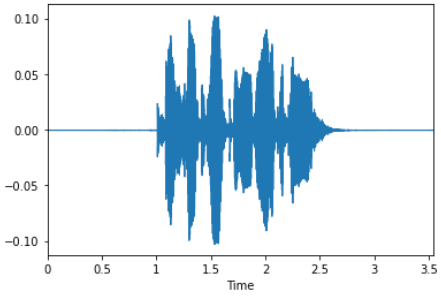
The dataset contains emotional audio data of 24 speakers. The 8 emotions included are Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Surprise, spoken as 2 sentences, “Kids are talking by the door” or “Dogs are sitting by the door.”

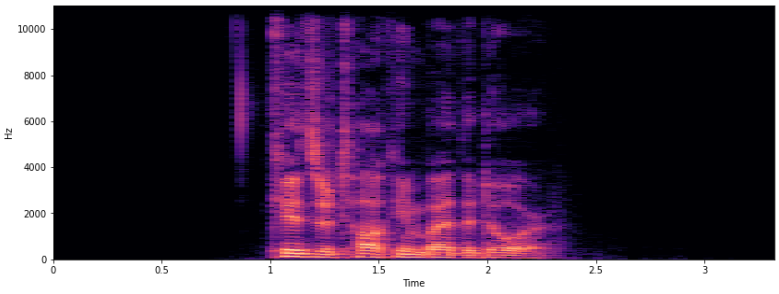
There are a total of 60 audio files per speaker which contains 4 of class Neutral and 8 of each of the other classes. The approximate duration of each sample is around 3s.

The emotions Calm, Happy, Sad, Angry, Fearful, Disgust and Surprise are grouped together as class Emotional, while the Neutral samples are class Neutral. The approach in this paper aims to enhance the performance of the binary classifier through iterative normalizations. The data is partitioned into Train and Test dataset, with speakers 1-18 under Train and 19-24 under Test.

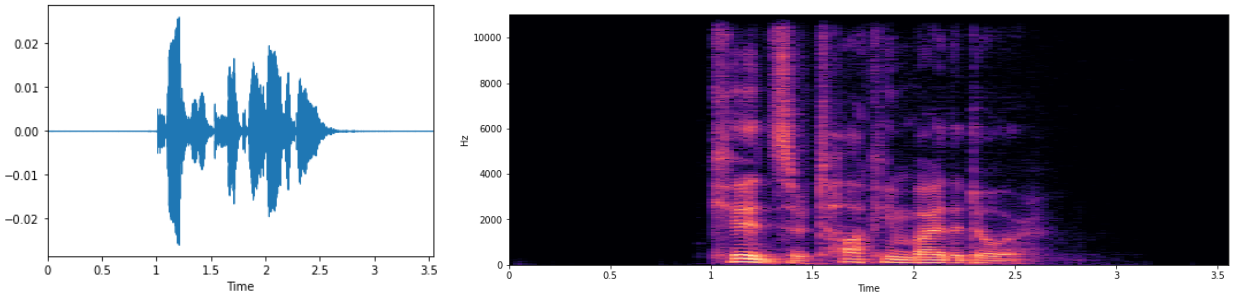
Shown below are sample Waveplots and Spectograms from the data:

Neutral Sample





Emotional Sample



Reference Neutral, Neutral, Emotional and Test Set

According to the paper, the reference neutral dataset is built using a reference database, different from the one used for the feature normalization. However, here, we have only one dataset from which a few neutral samples have been separated to form this neutral dataset.

While constructing the sample reference neutral dataset, it was ensured that one neutral audio clip of each participant has been used and also in a way to incorporate both the types of sentences. Thus, the size of the neutral dataset was 24 audio files.

This was used to calculate the F0_ref value, the numerator of the normalization factor, by taking the average of the mean value of all 24 F0 contours.

The remaining 3 neutral samples for the first 18 participants was taken under the neutral subset. This contains 54 audio files. The emotional samples of the first 18 participants was taken under the emotional set. This contains 1008 audio files. Finally, all audio files of participants 19-24 were taken into the test set. This set contains 360 audio files.

Feature Extraction

The F0 contours were evaluated for each audio sample by using a correlation based frequency estimator. The audio sample was divided into windows of 40ms, with an overlap of 30ms. Thus, the number of windows per second was nearly 100 and for the audio file of around 3s, the length of the F0 contour was on average 300.

For each F0 contour, statistical features were extracted. These include:

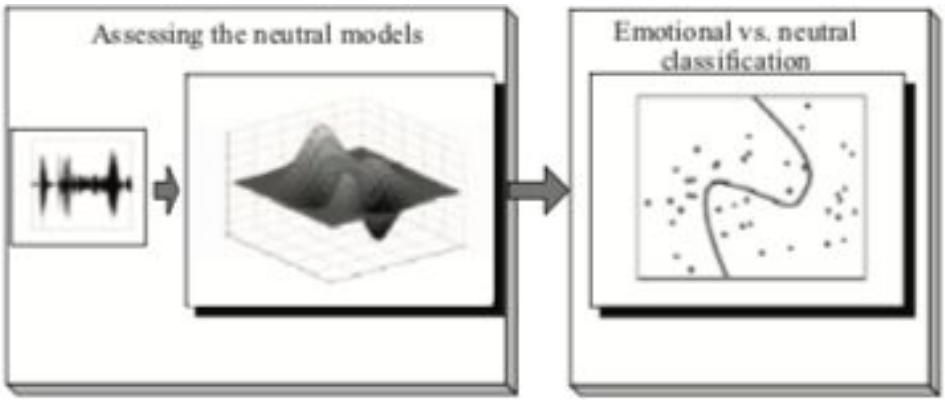
1. Mean
2. Standard Deviation
3. Median
4. Range
5. Q1
6. Q3
7. IQR

Hence, a 7 dimensional feature vector was obtained for each audio sample.

Classification Pipeline

The neutral reference dataset is used to train a GMM. One GMM was trained per feature, hence in all 7 GMMs were trained. The number of components for the GMM = K was set to 2. The emotional dataset was sampled to pick features equal to the number of neutral features. This is called bootstrapping. The number of bootstraps was set to 10 and later increased to 400. All results are an average score across bootstraps. Likelihood scores were obtained on the emotional + neutral dataset. As a result, a 7 dimensional vector of likelihoods was obtained for each feature from the neutral and emotional set. Here, an optional normalization step

was performed. Next, an LDA was trained on 70% of the combined data from neutral and emotional and 30% was reserved for validation. The trained LDA was tested on the entire neutral + emotional set and % change in labels in the 30% validation set were recorded.



Comparisons for 3 experiments have been made:

1. No Normalization
2. Optimal Normalization
3. Iterative Feature Normalization

1. For No Normalization, the optional normalization step is skipped.
2. For Optimal Normalization, the features are scaled by the optimal scaling factors, before the LDA training step. Here, optimal scaling factors are obtained by considering the true neutral samples per participant from the neutras set to get the denominator of the scaling factor.
3. For the Iterative Feature Normalization, the above pipeline is repeated until convergence, where convergence is defined as a 5% or lesser percentage change in labels. For the first iteration, no normalization is performed. For the subsequent iterations, scaling factors are calculated based on the samples classified as neutral, and the data is scaled each time. The values of scaling factors are in the range of 0 to 2.

Results

On num_bootstraps = 400

	No Normalization	Optimal Normalization	IFN
Average Val Accuracy	0.701	0.705	0.677
Average Test Accuracy	0.528	0.597	0.581

These values indicate that the performance when scaled with the optimal factors is higher than in the other 2 experiments for the Test set. This can be explained as follows: as the optimal scaling factors scale every speaker's data by a different factor, they are hence able to reduce speaker variability and increase detection performance by ensuring emotional vs neutral is the only distinguishable feature which is then classified using the 7 handcrafted speech features.

For the Val set, the value under optimal normalization is very similar to the unnormalized average accuracy value. Further, the performance of no normalization is better than IFN on average. This could be because the GMM is trained on unscaled neutral samples of the same dataset, it is assigning a higher likelihood for the unscaled neural samples as compared to the scaled neutral samples. This problem can be solved when the GMM is trained using a different reference dataset as done in the paper. The less difference between no normalization and optimal normalization can be explained by a similar argument.

For the test set however, no scaling is possible since we do not have factors for participants in the test set who are new unseen speakers. Also, unlike the problem above, since the speakers are unseen in the train set, the GMM cannot simply assign/low likelihoods based on similarity Hence, the difference in the test set is due to the differences in training of the LDA classifier. We see that the performance is better with IFN as compared to no normalization. This indicated that the LDA trained on normalized features is able to classify better as compared to an LDA trained on unnormalized features.

Plots

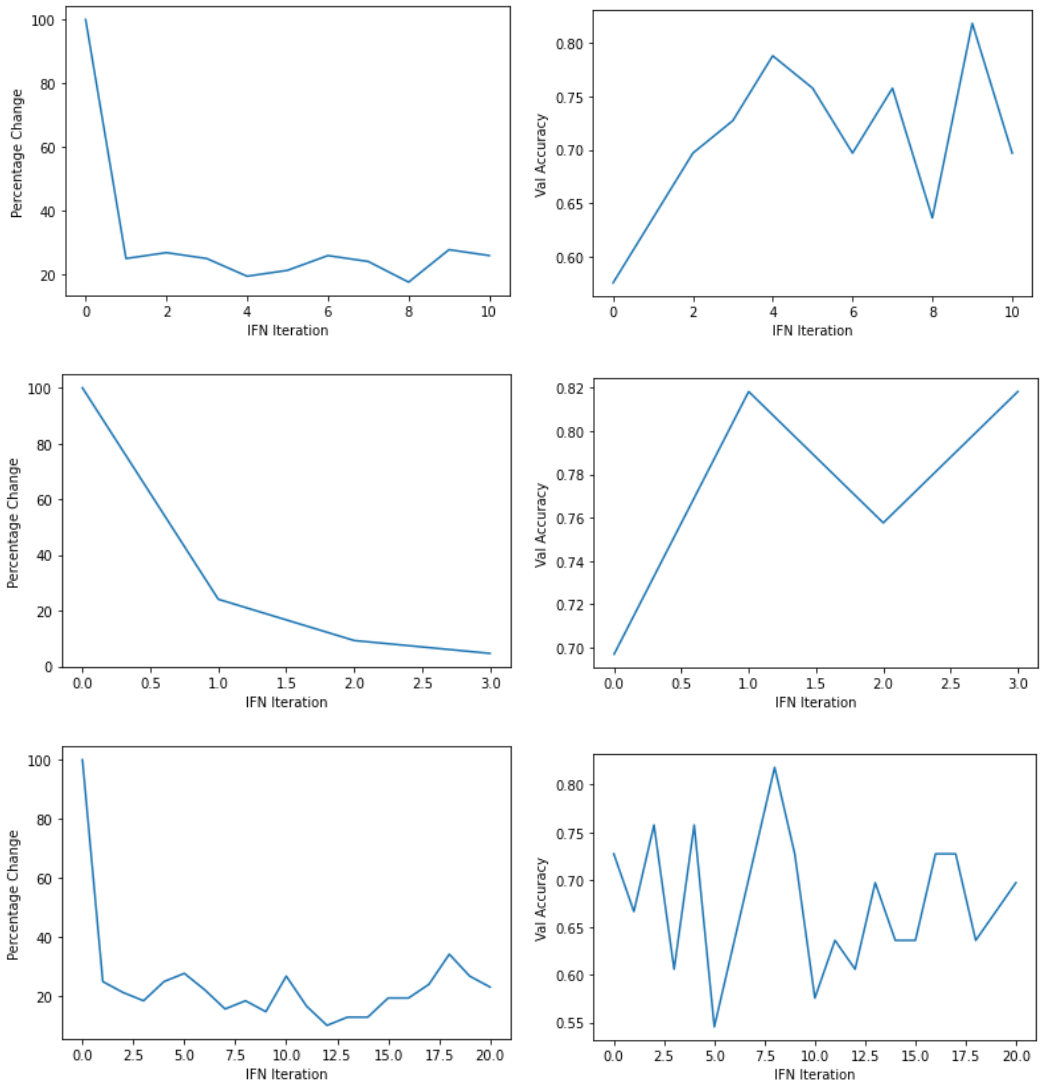
Across Iterations

Bootstrap 1

```
-------------- IFN Iteration 0 -------------
TRAIN ACC = 0.72
VAL ACC = 0.7272727272727273
PERCENT CHANGE = 100.0
-------------- IFN Iteration 1 -------------
{0: 0.9759453881623769, 1: 0.8168647668679816,
TRAIN ACC = 0.7066666666666667
VAL ACC = 0.7575757575757576
PERCENT CHANGE = 22.22222222222222
-------------- IFN Iteration 2 -------------
{0: 1.0413492935243005, 1: 1.0000000000000002,
TRAIN ACC = 0.6933333333333334
VAL ACC = 0.8181818181818182
PERCENT CHANGE = 12.037037037037036
-------------- IFN Iteration 3 -------------
{0: 0.9999999999999998, 1: 1.0, 2: 1.0, 3: 1.1
TRAIN ACC = 0.7066666666666667
VAL ACC = 0.8484848484848485
PERCENT CHANGE = 7.407407407407407
-------------- IFN Iteration 4 -------------
{0: 1.0, 1: 1.0, 2: 1.0, 3: 0.866849128679013,
TRAIN ACC = 0.68
VAL ACC = 0.8484848484848485
PERCENT CHANGE = 5.555555555555555
```

Bootstrap 2

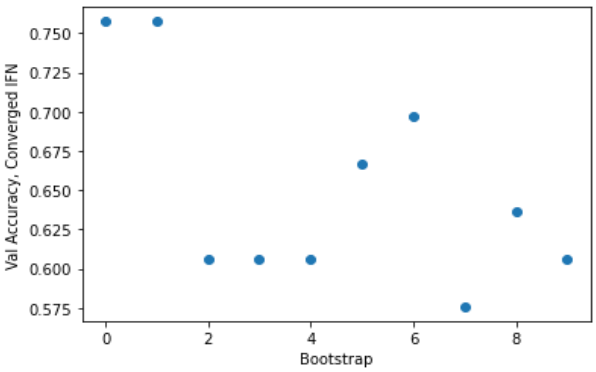
```
-------------- IFN Iteration 0 -------------
TRAIN ACC = 0.6
VAL ACC = 0.7575757575757576
PERCENT CHANGE = 100.0
-------------- IFN Iteration 1 -------------
TRAIN ACC = 0.68
VAL ACC = 0.7272727272727273
PERCENT CHANGE = 25.0
-------------- IFN Iteration 2 -------------
TRAIN ACC = 0.6666666666666666
VAL ACC = 0.696969696969697
PERCENT CHANGE = 12.962962962962964
-------------- IFN Iteration 3 -------------
TRAIN ACC = 0.6666666666666666
VAL ACC = 0.7272727272727273
PERCENT CHANGE = 23.14814814814815
-------------- IFN Iteration 4 -------------
TRAIN ACC = 0.6533333333333333
VAL ACC = 0.6666666666666666
PERCENT CHANGE = 17.59259259259259
-------------- IFN Iteration 5 -------------
TRAIN ACC = 0.6933333333333334
VAL ACC = 0.6060606060606061
PERCENT CHANGE = 23.14814814814815
```

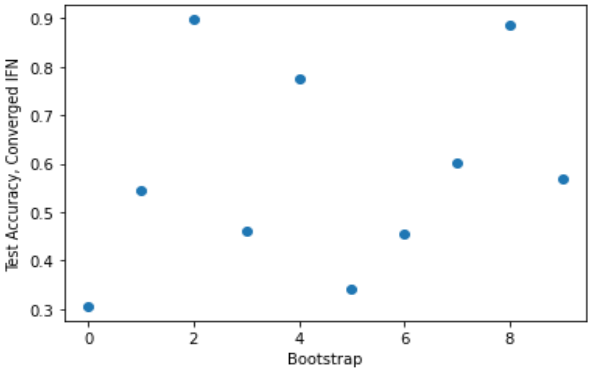


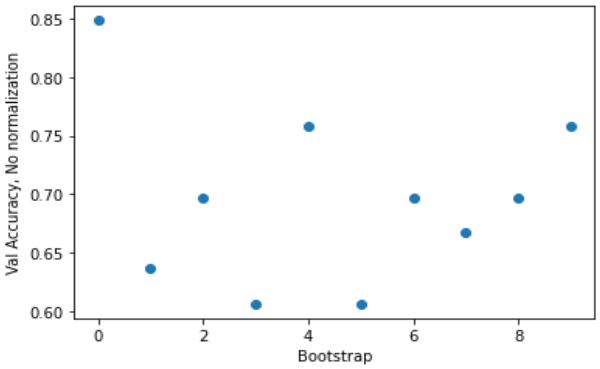
These plots and output lines show the fluctuations in validation accuracies within a bootstrap across iterations. In the output log for Bootstrap 1, it can be seen that the val accuracy is increasing with iterations, while in Bootstrap 2, it seems to be declining. Similarly in the plots, we see very varying values of validation accuracy. This could be attributed to unreliable scaling factors biased due to the GMM being trained on the same dataset. The percent changes, on the other hand, consistently decrease across iterations. This could mean that with repeated normalization, the fluctuation in label changes is reduced which can be graphically visualized as the

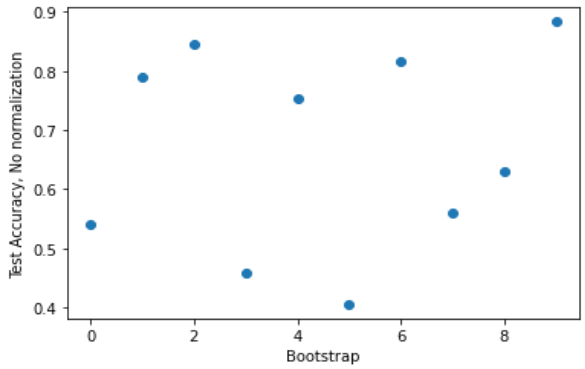
multiplied factor causing a change in the same direction with respect to the LDA decision boundary.

Across Bootstraps - Num_bootstraps = 10










```
Avergae Val Accuracy, Converged IFN 0.7515151515151517
Avergae Val Accuracy, No Normalization 0.7272727272727274
Avergae Test Accuracy, Converged IFN 0.4733333333333333
Avergae Test Accuracy, No Normalization 0.5416666666666666

Average Val Accuracy, Optimal Normalization 0.7181818181818183
Average Test Accuracy, Optimal Normalization 0.701388888888889

-------------------------------------------------------------------------------------

Avergae Val Accuracy, Converged IFN 0.6515151515151516
Avergae Val Accuracy, No Normalization 0.696969696969697
Avergae Test Accuracy, Converged IFN 0.5838888888888889
Avergae Test Accuracy, No Normalization 0.6683333333333333

Average Val Accuracy, Optimal Normalization 0.6515151515151516
Average Test Accuracy, Optimal Normalization 0.4716666666666667
```

We see high variability in accuracy values in different bootstraps. This could be a result of coincidental sampling to incorporate only a few speaker samples or miss out on a few speaker samples leading to unreliable scaling factors. Further, due to the differences in trends across separate executions of 10 bootstraps, the number of bootstraps were increased to 400 to report the final results given previously.

Conclusion

In conclusion, the IFN technique has the potential to enhance performance as evidenced in the case of the test accuracies. However, this enhancement can only be ascertained when there are no inherent dependencies in the reference neutral dataset with the other dataset as in the paper. The paper employed 3 datasets and also employed a mix of statistical and voiced features. Hence, a direct comment on agreement or disagreement cannot be made given the large number of differences in the approaches. However, the applicability and potential of such normalization can be understood through this experiment.