# ASSIGNMENT 03 : AUDIO CLASSIFICATION USING SPECTROGRAMS

*Saptarshi Mandal*

Date : 22th Jan, 2024

## 1. PROBLEM STATEMENT

1. Build a 3-class vowel classifier to perform /a/ vs. /i/ vs. /u/ classification.

   (a) Use spectrogram of clean speech.

   (b) Use spectrogram of noisy speech.

   (c) Randomly select x% of rows of clean spectrogram and replace with uniformly distributed random noise for x = 1, 5, 10, 20, 50. Use the modified spectrogram.

   (d) Randomly select x% of columns of clean spectrogram and replace with uniformly distributed random noise for x = 1, 5, 10, 20, 50. Use the modified spectrogram.

   (e) Time-scale the clean speech with scale factor 1.2 and 0.8. Use the spectrogram of time-scaled speech.

   (f) Add reverberation of three types to the clean speech and use the reverberated spectrogram. Take any three RIR from the shared dataset. In each of the above cases, train your classifier with the train segments and report the classification accuracy on the test segments. Also report the confusion matrix in each of the above cases.

2. Repeat for 3-class consonant classification - /s/ vs. /sh/ vs. /f/.

All clean speech audio files are 2 sec segments of sustained phoneme utterances recorded from 30 subjects. The phonemes include three vowels - /a/, /i/, /u/ and three fricatives -/s/, /sh/, /f/. Multiple utterances of a phoneme are recorded by a subject. For each clean speech file, one among white noise, pink noise, babble noise and high-frequency radio channel noise is selected randomly and added to the clean speech at 5dB SNR (Signal to Noise Ratio) . This is how the Noisy Speech audio files are generated.

The Reverb File contains 12 different room impulse responses.

## 2. IMPLEMENTATION

Spectrogram is computed for speech audio signals and the Spectrograms are fed into a Convolutional Neural Network (CNN) for classification.

Spectrogram is calculated by calculating the Short-Time Fourier Transform (STFT) of the audio signal and taking the square of the modulus of the STFT. The STFT is calculated by the formula given in Eq.1

$$X[m,k] = \sum_{n=0}^{N-1} x[n]w[n-m]e^{-j2\pi kn/N} \tag{1}$$

For classification purposes the following CNN Architecture is designed:

1. Convolution Block 1
   - Convolution Layer 1 (64 kernels, 3x3 kernels)
   - Activation function: ReLU
   - Max pooling layer (2x2 pool size)
2. Convolution Block 2
   - Convolution Layer 1 (64 kernels, 3x3 kernels)
   - Activation function: ReLU
   - Max pooling layer (2x2 pool size)
3. Flattening
   - Reshape for input to fully connected layers (Vector Size [ 1x4096])
4. Fully connected layers:
   - FC layer 1(256 neurons)
   - Activation function: ReLU
   - FC layer 2 (3 neurons)
   - Activation function: SoftMax

The Result Obtained from all the cases for Vowel Classification is given in Table. 1

The Result Obtained from all the cases for Consonant Classification is given in Table. 2

| Dataset | Parameter | Test Accuracy |
|---|---|---|
| Clean Speech Spectrograms | NA | 99.72% |
| Noisy Speech Spectrograms | NA | 82.77% |
| Corrupted Rows in Clean Speech Spectrograms | 1% Corruption | 98.02% |
| | 5% Corruption | 94.35% |
| | 10% Corruption | 90.96% |
| | 20% Corruption | 79.94% |
| | 50% Corruption | 45.5% |
| Corrupted Columns in Clean Speech Spectrograms | 1% Corruption | 98.02% |
| | 5% Corruption | 92.93% |
| | 10% Corruption | 86.15% |
| | 20% Corruption | 88.13% |
| | 50% Corruption | 71.75% |
| Time Scaled Clean Speech Spectrograms | 1.2 Scale Factor | 98.2% |
| | 0.8 Scale Factor | 99.71% |
| Reverberation Added Clean Speech Spectrograms | smallroom1_near_anglb | 99.43% |
| | largeroom1_near_anglb | 99.15% |
| | mediumroom1_far_anglb | 99.71% |

**Table 1**: Test Accuracy for Vowels

## 3. EXPLANATION OF RESULTS

### 3.1. Vowel Classification

- **Clean Speech:** The CNN classifier achieves excellent performance on clean speech spectrograms, with a test accuracy of 99.72%. This indicates its ability to effectively extract and classify vowel features from clean audio signal spectrograms.
- **Noisy Speech:** The accuracy drops to 82.77% for noisy speech spectrograms, highlighting the negative impact of noise on classification performance.
- **Corrupted Rows:** The accuracy decreases gradually as more rows are corrupted with Gaussian noise. However, the classifier still maintains moderate accuracy even with significant corruption (50% of rows/columns), suggesting some robustness to noise.
- **Corrupted Columns:** The accuracy decreases gradually as more columns are corrupted with Gaussian noise. However Since, vowels are to some extent stationary audio signals so corruption in columns has less effect compared to rows since corruption in columns correspond to corruption in all frequency for a specific time location in the audio signal.
- **Effect of Noise Location:** The difference in accuracy between corrupting rows vs. columns is evident from the results, suggesting that the classifier is sensitive to the spatial distribution of noise within the spectrograms. It can be observed that the the classifier maintains better accuracy even with significant corruption when compared to corruption in rows.

  This can be explained by the fact that each row in a Spectrogram corresponds to a frequency band throughout the duration of the signal. Now since, the audio signals consist of vowels which are generally comprises of frequencies throughout its duration, so corruption in the rows significantly drops the classification accuracy.
- **Time-Scaling:** The accuracy remains high for time-scaled spectrograms (1.2x and 0.8x), indicating the classifier's ability to handle moderate variations in speaking rate. This suggests potential robustness to natural variations in speech speed.
- **Reverberation:**

  The accuracy remains high for reverb added audio spectrograms , indicating the classifier's ability to handle variations in reverberation added in audio signals. This suggests potential robustness to natural variations in room responses.

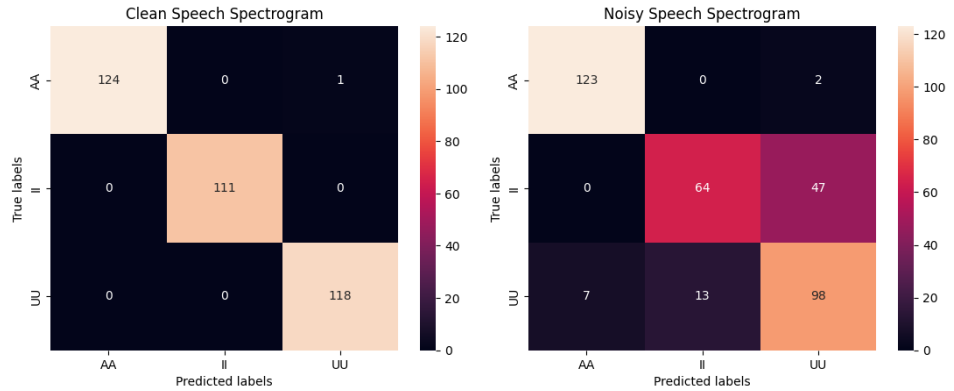
The Confusion Matrix for all vowel classifiers are given in Fig.1, Fig.2, Fig.3, Fig.4, Fig.5.

### 3.2. Consonant Classification

- **Clean Speech :** The baseline accuracy of 74.38% for clean speech spectrograms is likely due to the model's ability to learn the inherent patterns in speech spectrograms that are relevant for consonant recognition.
- **Noisy Speech :** The slightly higher accuracy of 74.88% for noisy speech spectrograms compared to clean

| Dataset | Parameter | Test Accuracy |
|---|---|---|
| Clean Speech Spectrograms | NA | 74.38% |
| Noisy Speech Spectrograms | NA | 74.88% |
| Corrupted Rows in Clean Speech Spectrograms | 1% Corruption | 74.38% |
| | 5% Corruption | 53.20% |
| | 10% Corruption | 58.62% |
| | 20% Corruption | 39.9% |
| | 50% Corruption | 42.8% |
| Corrupted Columns in Clean Speech Spectrograms | 1% Corruption | 65.02% |
| | 5% Corruption | 64.53% |
| | 10% Corruption | 40.39% |
| | 20% Corruption | 57.63% |
| | 50% Corruption | 58.12% |
| Time Scaled Clean Speech Spectrograms | 1.2 Scale Factor | 68.47% |
| | 0.8 Scale Factor | 71.92% |
| Reverberation Added Clean Speech Spectrograms | smallroom1_far_angla | 67.48% |
| | smallroom1_far_anglb | 72.90% |
| | largeroom1_far_angla | 74.87% |

**Table 2**: Test Accuracy for Consonants




(a) Clean Speech Spectrogram

(b) Noisy Speech Spectrogram

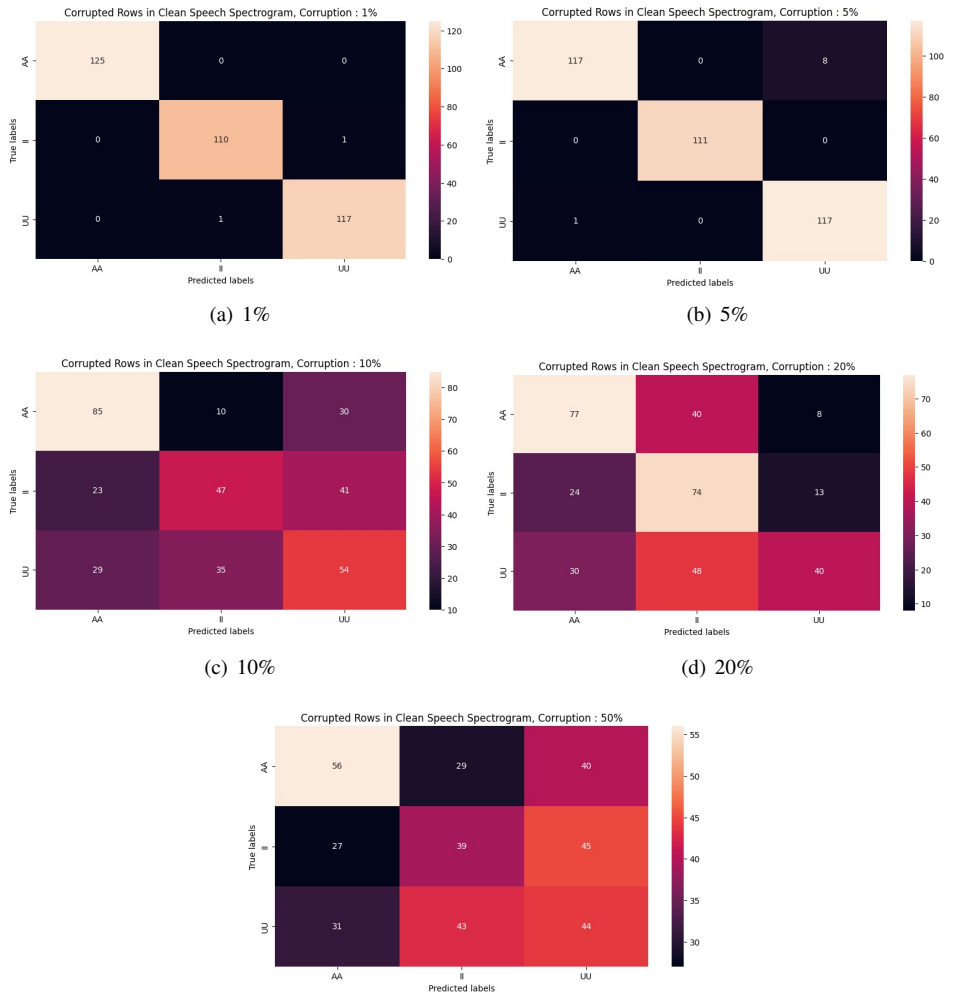
**Fig. 1**: Confusion Matrices for Clean and Noisy Speech Spectrograms for Vowels

speech spectrograms might be due to the model learning to compensate for the noise, potentially generalizing better to unseen noisy data.

- **Corrupted Rows:** Corrupting rows in clean speech spectrograms had a significant impact on accuracy, with accuracy decreasing as the percentage of corruption increased. This is likely because corrupting rows disrupts the temporal structure of the spectrograms, which is important for consonant recognition.
- **Corrupted Columns :**

  Corrupting columns in clean speech spectrograms had a less severe impact on accuracy compared to corrupting rows. This is likely because columns represent frequency bands, and the model might be able to learn from the remaining non-corrupted columns.

- **Time-Scaling :** Time scaling the clean speech spectrograms resulted in a decrease in accuracy for both scaling factors. This suggests that the model is sensitive to the temporal characteristics of the spectrograms and that significant changes to the timescale can negatively impact performance.
- **Reverberation :** Adding reverberation to clean speech spectrograms had mixed results, with some reverberation conditions (smallroom1_far_angla) leading to lower accuracy and others (smallroom1_far_anglb, largeroom1_far_angla) having minimal impact. This suggests that the model's performance under reverberation depends on the specific characteristics of the




(a) 1% (b) 5%

(c) 10% (d) 20%

(e) 50%

**Fig. 2**: Confusion Matrices for Corrupted Rows (varying percentage of corruption) of Clean Speech Spectrograms for Vowels

reverberation.

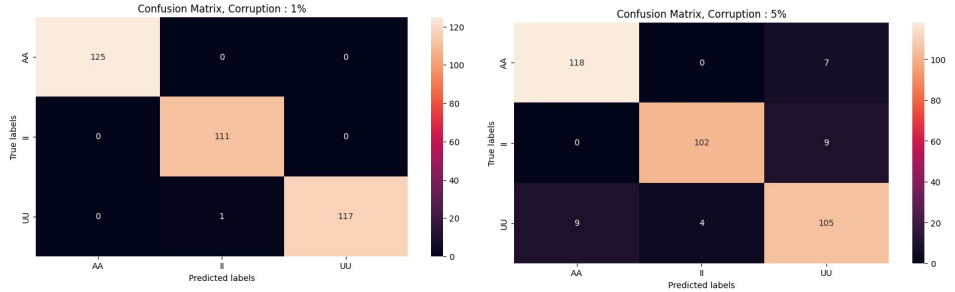
The Confusion Matrix for all consonant classifiers are given in Fig.6, Fig.7, Fig.8, Fig.9, Fig.10.

### 3.3. Vowels vs Consonants

In this section we look at the comparison of the results for consonants and vowels, along with explanations and interpretations:

**Key Observations:**

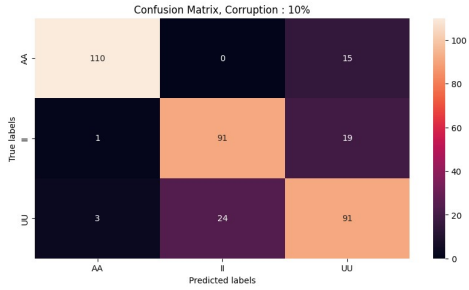
- **Generally higher accuracy for vowels:** The model consistently achieves higher accuracy for vowels across all conditions, indicating that vowels are inherently easier to classify from spectrograms than consonants. This is likely due to their more steady-state stationary nature and clearer spectral patterns.
- **Robustness to noise for consonants:** The model's accuracy for consonants is slightly higher with noisy speech spectrograms compared to clean ones. This suggests that the model might have learned to filter out noise to some extent, potentially improving its generalization to unseen noisy data.
- **Sensitivity to row corruption for consonants:** Consonant accuracy degrades more severely with row corruption than with column corruption. This highlights the importance of the temporal structure in spectrograms for consonant recognition, as rows represent time slices.
- **Less sensitivity to time scaling for vowels:** Vowels are less affected by time scaling compared to consonants, suggesting that their spectral characteristics are more dominant for classification than precise temporal



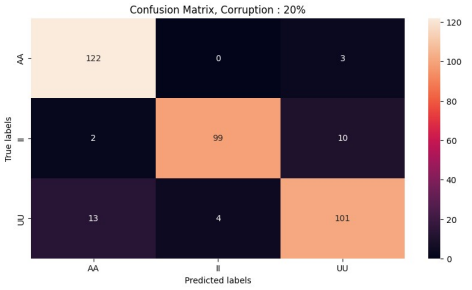

(a) 1%



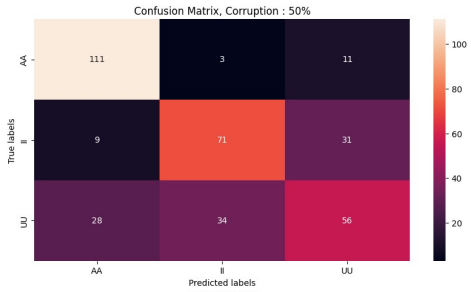

(b) 5%




(c) 10%

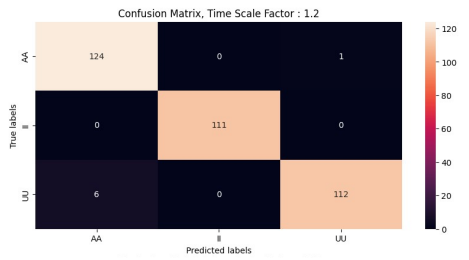



(d) 20%

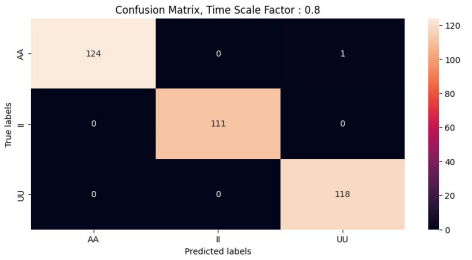



(e) 50%

**Fig. 3**: Confusion Matrices for Corrupted Rows (varying percentage of corruption) of Clean Speech Spectrograms for Vowels



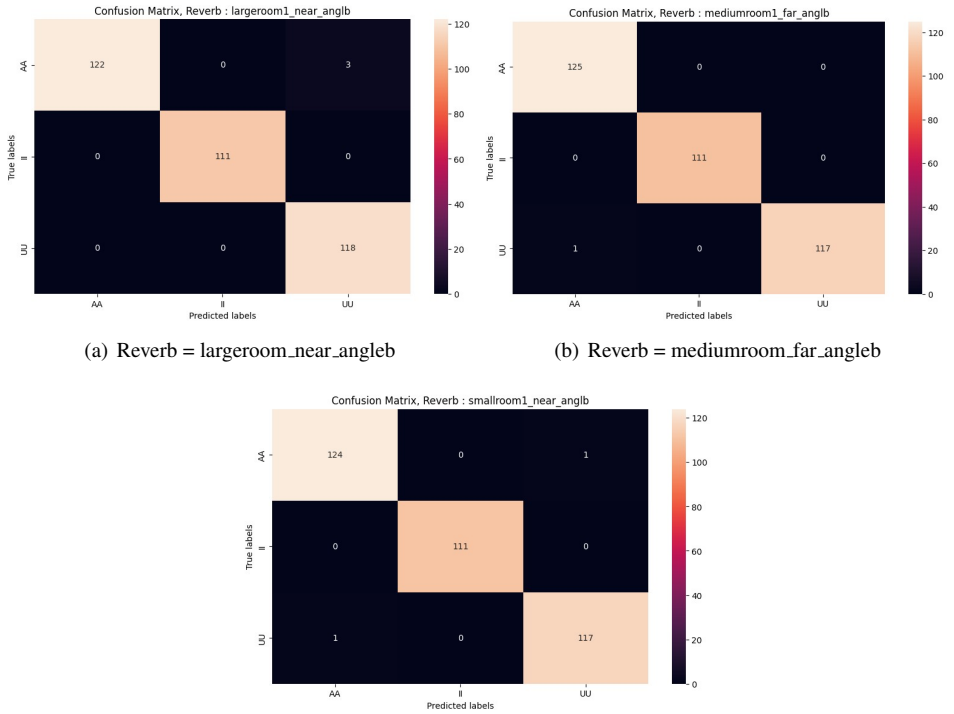

(a) Scale = 1.2




(b) Scale = 0.8

**Fig. 4**: Confusion Matrices for Time Scaled (varying scales) of Clean Speech Spectrograms for Vowels

patterns.

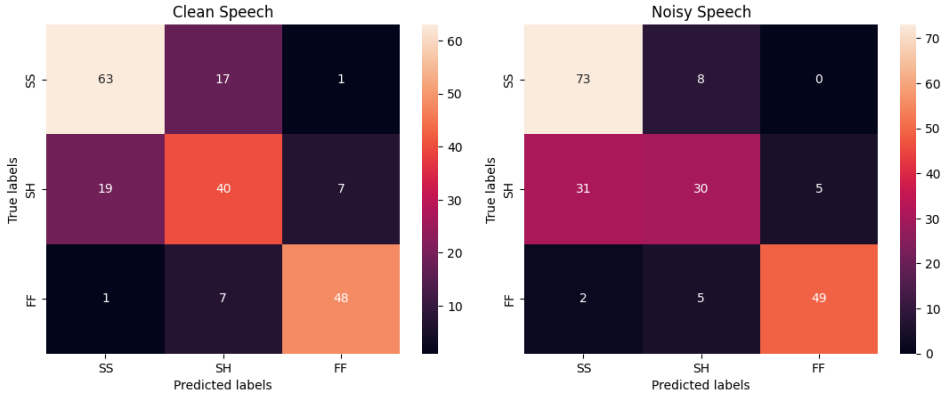



(a) Reverb = largeroom_near_angleb

(b) Reverb = mediumroom_far_angleb

(c) Reverb = smallroom_near_angleb

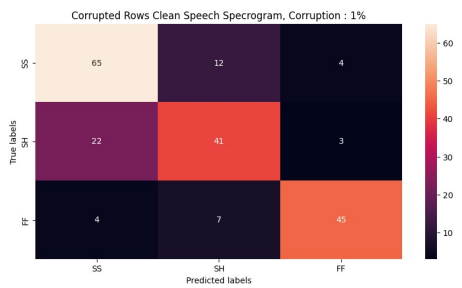
**Fig. 5**: Confusion Matrices for Reverb added Clean Speech Spectrograms for Vowels



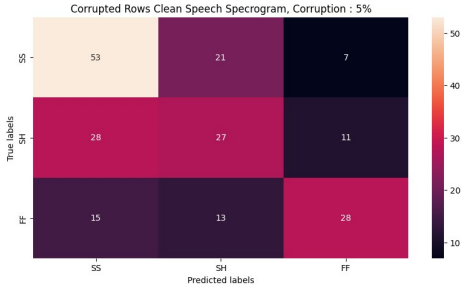

(a) Clean Speech Spectrogram

(b) Noisy Speech Spectrogram

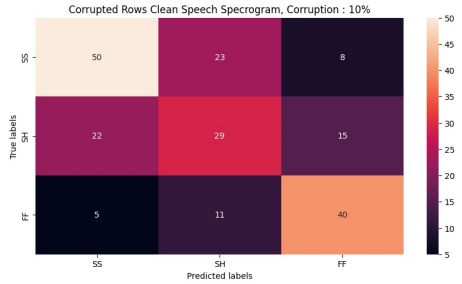
**Fig. 6**: Confusion Matrices for Clean and Noisy Speech Spectrograms for Consonants



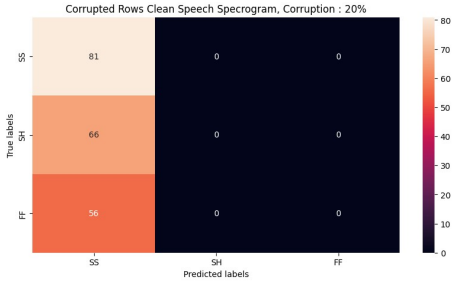

(a) 1%



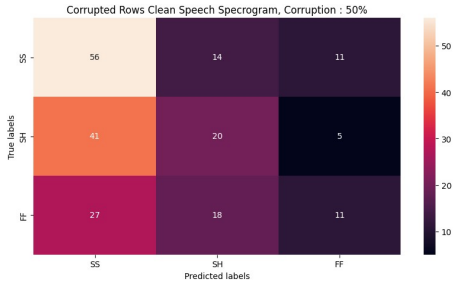

(b) 5%




(c) 10%

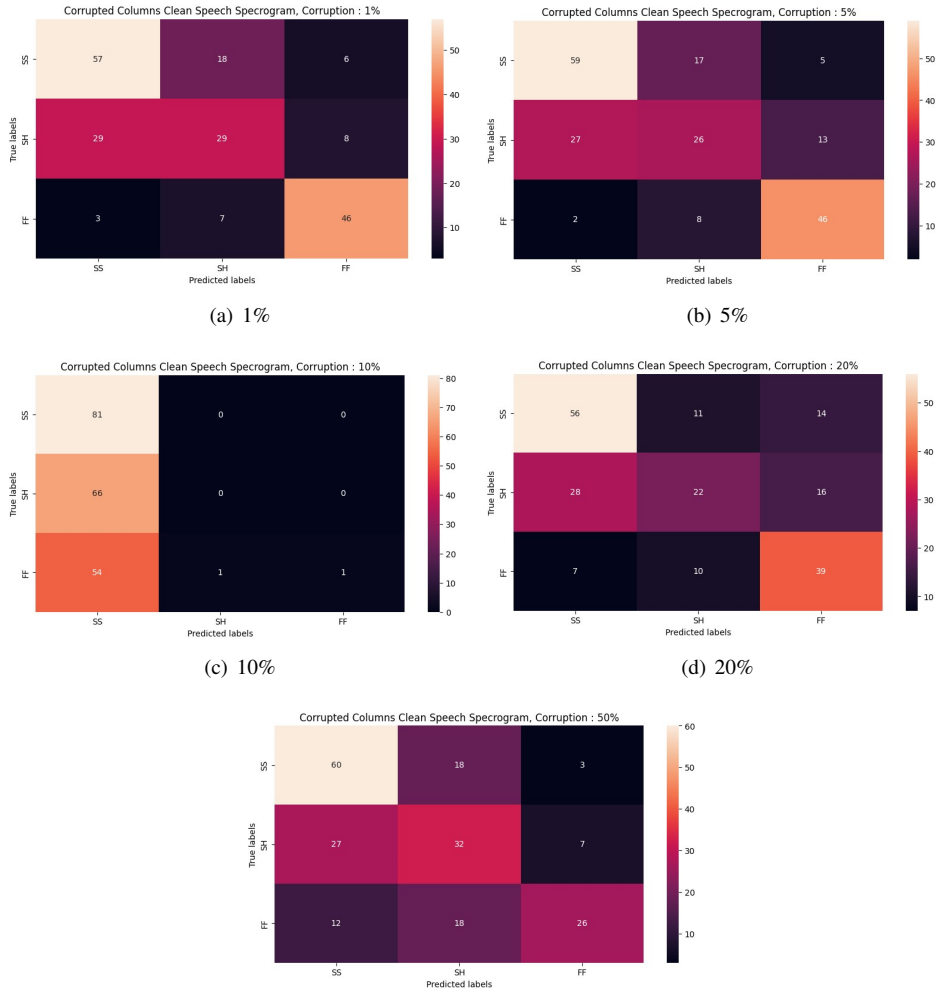



(d) 20%




(e) 50%

**Fig. 7**: Confusion Matrices for Corrupted Rows (varying percentage of corruption) of Clean Speech Spectrograms for Consonants




(a) 1%

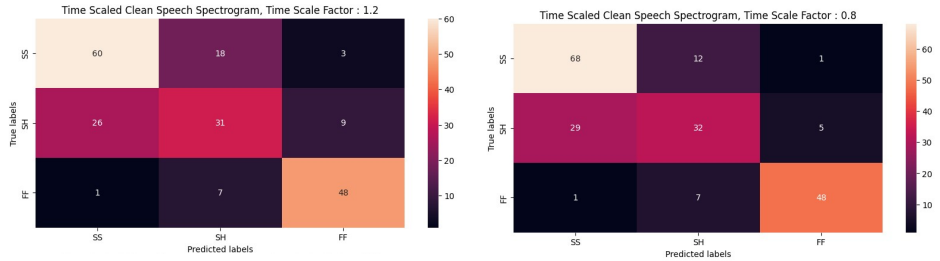
(b) 5%

(c) 10%

(d) 20%

(e) 50%

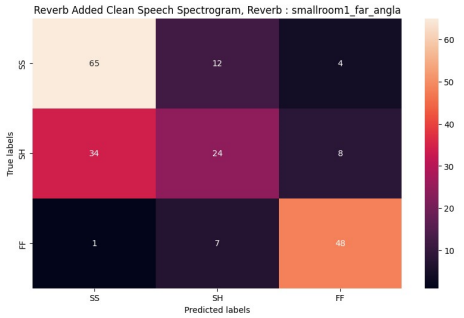
**Fig. 8**: Confusion Matrices for Corrupted Columns (varying percentage of corruption) of Clean Speech Spectrograms for Consonants



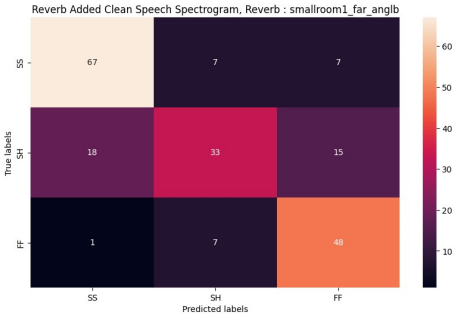

(a) Scale = 1.2

(b) Scale = 0.8

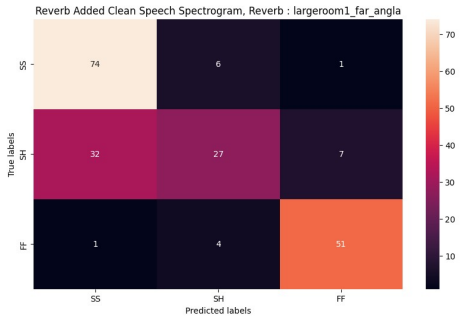
**Fig. 9**: Confusion Matrices for Time Scaled (varying scale) Clean Speech Spectrograms for Consonants




(a) Reverb = smallroom_far_anglea




(b) Reverb = smallroom_far_angleb




(c) Reverb = largeroom_far_anglea

**Fig. 10**: Confusion Matrices for Reverb added Clean Speech Spectrograms for Consonants