Spam user classification in crowdsourced speech data via Voice Activity Detection

Samrat Dutta - 193050026
Deepti Mittal - 193050025
Ankita Singh - 19305R002

Problem Statement

- Classification of audio files which are purely noise from the ones which have speech content as well
- Early detection of spam users in crowdsourced environments or remove non-speech clips
- Voice activity detection (VAD) refers to the task of determining whether a signal contains speech or not.
- Our classification problem : A binary decision to classify as spam or not spam

$$y^\star := \begin{cases} 0, & x \text{ is not speech}, \\ 1, & x \text{ is speech}. \end{cases}$$

Data-sets used

1. **Libri-speech dataset** : clear spoken english corpora
2. **QUT noise dataset** : different types of noise from CAFE, CAR, STREET etc
3. **CLAP Dataset** : Data from our project , further subdivided as:

   (i) Clean speech data : AIF >2

   (ii) Speech data with noise : AIF>0 + BIN >0

   (iii) CLAP noise (Trimming first 3 secs of all recordings where BIN>=2)

4. **Blind test set** : Speech data of all types combined with spam user data (obtained during one of our trials)

Methodology/Models used

1. Basic Multi-Layer Perceptron (MLP) Architecture
2. LSTM based architecture
3. CNN based architecture
4. Wav2vec based architecture/embeddings
5. DEMUCS combined with all of the above architectures

(Further details of all the architectures are mentioned in the report)

Experimentations

1. **Initial experiments :**

   -> MLP and MLP + DEMUCS used

   -> Metadata taken from all combinations of the data-sets and results noted

2. **Experiments with blind-set :**

   Spam user data combined with non-spam data (CLAP all types of speech)

   Spam user data + QUT noise (taken to see generalisation of models) combined with non-spam

   All architectures used and observations noted

Observations and Error analysis (Initial)

Data	Train Accuracy	Test accuracy	Comments
Libri speech + QUT noise (cafe noise only) (80:20 ratio of speech and noise) (Divided into test and train 5:1)	100.0	99.84	Speech is clean,noise signals and speech signals are easily distinguishable
Libri speech + QUT noise (cafe noise only) for training Marathi + QUT noise (cafe noise only) for testing	100	99.42	Noise domain is similar , hence high accuracies
Libri speech + QUT noise (cafe noise only) for training Marathi (CLAP data)+ noise (QUT all kinds of noise) for testing	100	98.10	We use different types of noises but yet the acoustics are distinct
Marathi (CLAP data) + noise (QUT all kinds) for training and testing	100	99.58	

Observations and Error analysis (Initial)

Libri speech + noise (QUT all kinds) for training Marathi+noise mixed and noise(cafe noise only) for testing	100	**56.75**	Clearly acoustics/domain of train and test are different and hence the numbers are bad
Marathi+noise mixed and only noise for both training and testing	99.62	96.96	
Libri speech + noise (QUT all kinds) for training DEMUCS applied on Marathi+noise mixed and noise(CAFE noise only) for testing	100	**78.45**	Compared to 56.75 in the testing results without DEMUCS, we see a significant improvement on passing Marathi+noise mixed speech over DEMUCS..
Libri speech + noise for training DEMUCS applied on Marathi+noise mixed and only noise denoised for testing	100	**53.64**	DEMUCS applied only on test data so non uniform data

Observations and Error analysis (Initial)

DEMUCS applied on both Libri speech + noise (CAFE noise) for training DEMUCS applied on Marathi+noise mixed and only noise denoised for testing	99.78	90.97	DEMUCS applied only on both train and test data, high accuracies because of noise signal being same.
Libri speech + noise (all kinds) for training CLAP noise (3sec trimmed) + clap marathi speech for testing	100	**40.46**	Clear domain mismatch for both speech and noise
CLAP noise (3sec trimmed) + clap marathi speech for training and testing (80/20 split)	99.26	98.52	

Observations on Blind test (Appendix)

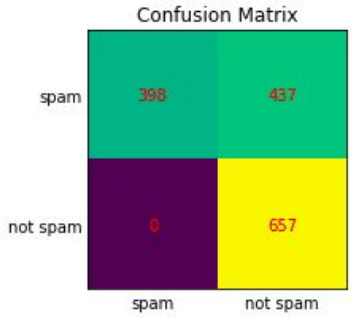
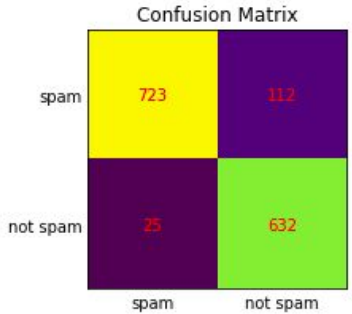
The training and testing was done for both **original** and **denoised** versions:

(1) **Train set** : All types of speech data from CLAP(Hindi, Marathi, Tamil ,Telugu) both noisy and clean along with 3 second noiseclips from background is noisy >=2 files

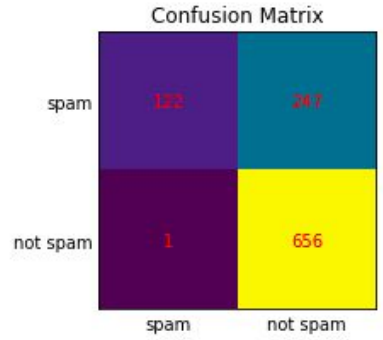
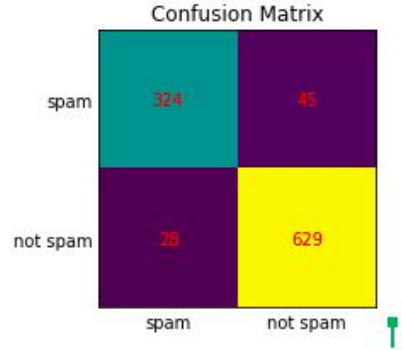
(2) **Test set**: All types of CLAP speech (we made sure test and train data were completely segregated) along with spam user clips + noisy clips and QUT noise (all kinds of noise)

(3) **Test set** : All types of CLAP speech (we made sure test and train data were completely segregated) along with spam user clips + noisy clips

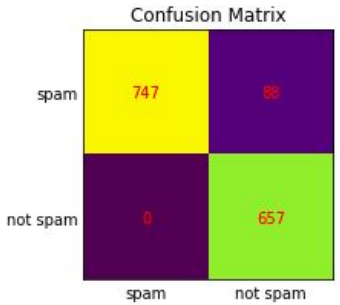
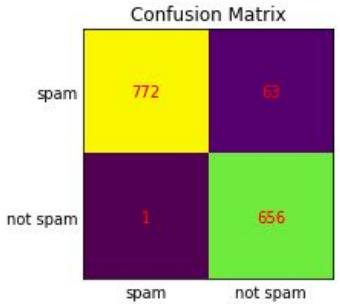
Observations and Error analysis (Blind-Test) -MLP

Data	Train	Test	Result Matrix
CLAP noise (3sec trimmed +BIN>=2) + clap (marathi+tamil+telugu+hindi) speech for train and All speech+spam to test (1) And (3) Normal (Final metadata)	95.78	70.71	
DENOISED : CLAP noise (3sec trimmed +BIN>=2) + clap (marathi+tamil+telugu+hindi) speech for train and All speech+spam to test (1) And (3) from denoised (Final metadata)	95.31	90.81	

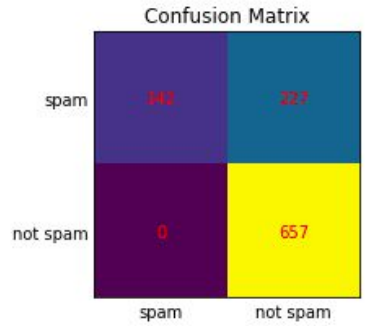
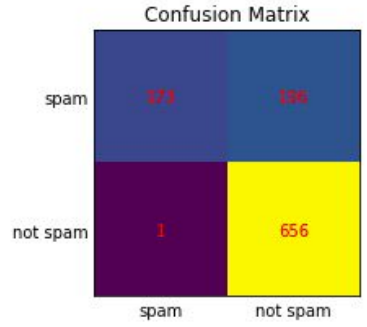
Observations and Error analysis (Blind-Test) -MLP

CLAP noise (3sec trimmed +noise clips +BIN>=2) + clap (marathi+tamil+telugu+hindi) speech for train and All speech+spam+librinoise to test (1) and (2) from NORMAL(final metadata)	95.78	75.82	
DENOISED : CLAP noise (3sec trimmed +BIN>=2) + clap (marathi+tamil+telugu+hindi) speech for train and All speech+spam+librinoise to test (1) And (2) from denoised	95.87	92.88	

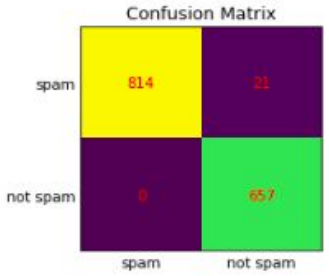
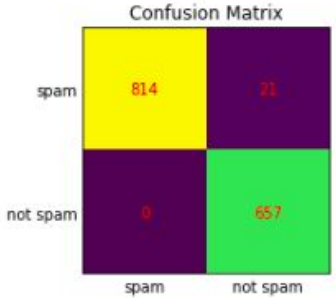
Observations and Error analysis (Blind-Test) -CNN

DATA	Train	Test	Result matrix
CLAP noise (3sec trimmed +BIN>=2) + clap (marathi+tamil+telugu+hindi) speech for train and All speech+spam to test (1) and(3) from Normal (Final metadata)	100.00	94.10	
DENOISED : CLAP noise (3sec trimmed +BIN>=2) + clap (marathi+tamil+telugu+hindi) speech for train and All speech+spam to test (1) And (3) from denoised (Final metadata)	99.62	95.71	

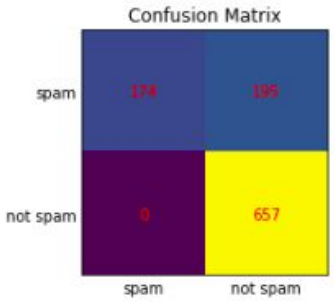
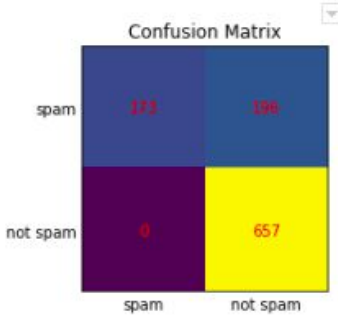
Observations and Error analysis (Blind-Test) -CNN

CLAP noise (3sec trimmed +noise clips +BIN>=2) + clap (marathi+tamil+telugu+hindi) speech for train and All speech+spam+librinoise to test (1) And (2) from NORMAL(final metadata)	100.00	77.88	
DENOISED : CLAP noise (3sec trimmed +BIN>=2) + clap (marathi+tamil+telugu+hindi) speech for train and All speech+spam+librinoise to test (1) And (2) from denoised	99.83	80.80	

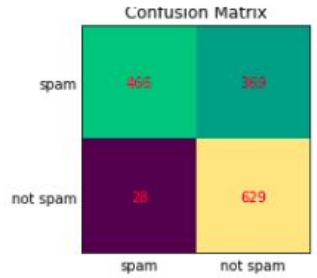
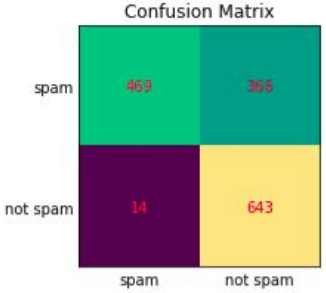
Observations and Error analysis (Blind-Test) - LSTM

CLAP noise (3sec trimmed +BIN>=2) + clap (marathi+tamil+telugu+hindi) speech for train and All speech+noise+spam to test	100	98.59	
DENOISED : CLAP noise (3sec trimmed +BIN>=2) + clap (marathi+tamil+telugu+hindi) speech for train and All speech+noise+spam to test	100	98.59	

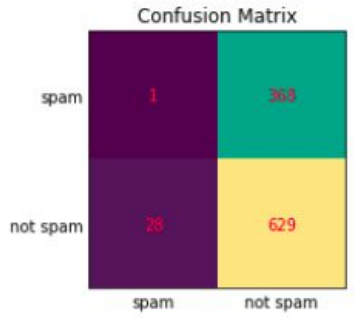
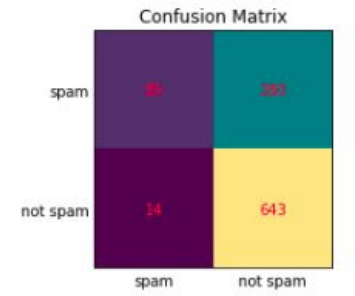
Observations and Error analysis (Blind-Test) - LSTM

CLAP noise (3sec trimmed +BIN>=2) + clap (marathi+tamil+telugu+hindi) speech for train and All speech+spam+librinoise to test	100	80.99	
DENOISED: CLAP noise (3sec trimmed +BIN>=2) + clap (marathi+tamil+telugu+hindi) speech for train and All speech+spam+librinoise to test	100	80.89	

Observations and Error analysis (Blind-Test) - Wav2vec

Data	Train	Test	Result Matrix
CLAP noise (3sec trimmed +BIN>=2) + clap (marathi+tamil+telugu+hindi) speech for train and All speech+noise+spam to test	97.88	74.33	
DENOISED : CLAP noise (3sec trimmed +BIN>=2) + clap (marathi+tamil+telugu+hindi) speech for train and All speech+noise+spam to test	97.95	74.50	

Observations and Error analysis (Blind-Test) - Wav2vec

CLAP noise (3sec trimmed +BIN>=2) + clap (marathi+tamil+telugu+hindi) speech for train and All speech+spam+librinoise to test	97.88	62.50	
DENOISED: CLAP noise (3sec trimmed +BIN>=2) + clap (marathi+tamil+telugu+hindi) speech for train and All speech+spam+librinoise to test	97.95	72.15	

Analysis and results

1. MLP+DEMUCS model generalises better for new domains
2. LSTM based architecture performs best among all other models for CLAP speech data and spam segregation with an accuracy of 98.59
3. DEMUCS helps improve MLP models but has little or no contributions for CNN/Bi-LSTM models
4. CNN and LSTM based models have no misclassified non-spam -> spam examples, better for real life scenarios
5. Wav2vec embeddings with a LSTM based classifier did not perform well due to lesser amount of data.

REAL TIME DEMO TO BE SHOWN