CS574 Assignment 3 Implementation

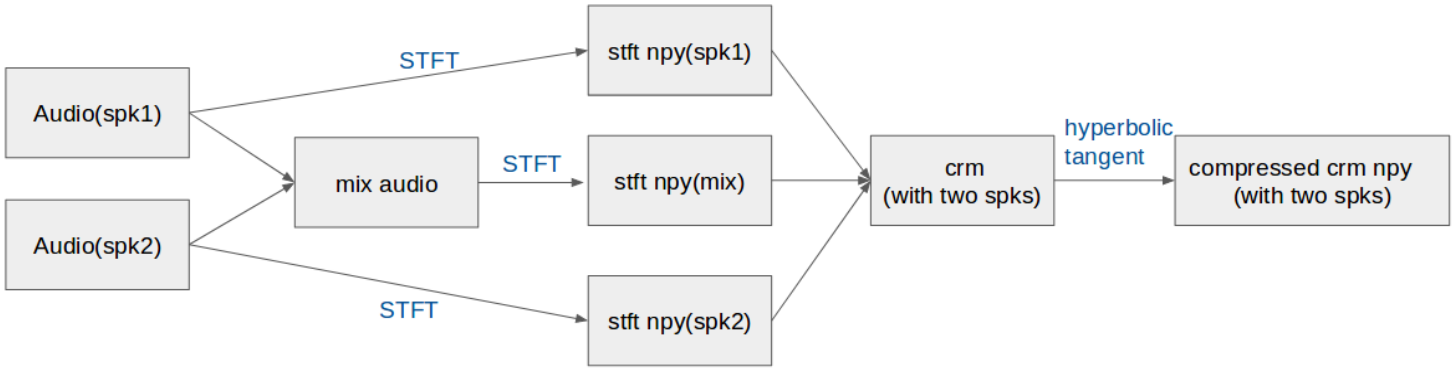
Speech separation using visual and speech cues

Group 15:	**Shubham Goel**	**160101083**
	Shaurya Gomber	**160101086**
	Archit Jugran	**160101087**
	Rishabh Jain	**160101088**
	Shashwat Jolly	**160123036**

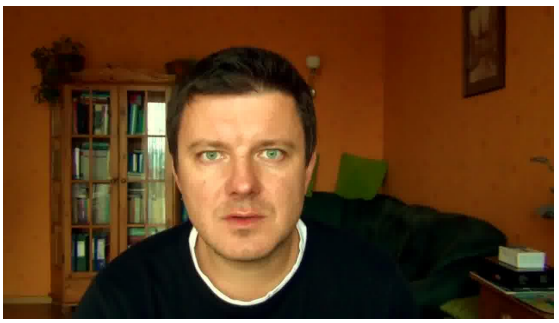
Ephrat et al. [1]

The speciality of this paper was that it created the AVSpeech database and used it for its training and testing. This is a large-scale dataset (4700 hours) taken from a total of 290k YouTube videos. All the clips had a single person in frame and there were minimal interfering noises.

An audio downloader file was used to download the audios related to the videos. Then it was preprocessed as follows:



Similarly, a video downloader file was used to download the videos (3s clips) from the dataset. To preprocess them, these steps were taken:



After all this, we implemented the model given in the paper and as we were not able to get a pre-trained model, we trained it using 200 randomly selected videos from AVSpeech dataset (200 because of resource and time limitations) and then used the model to enhance the videos present in the testing set.

Results:

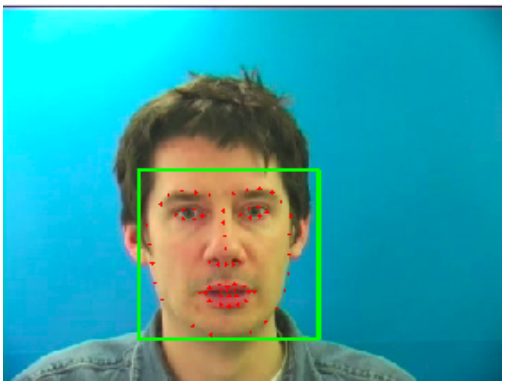
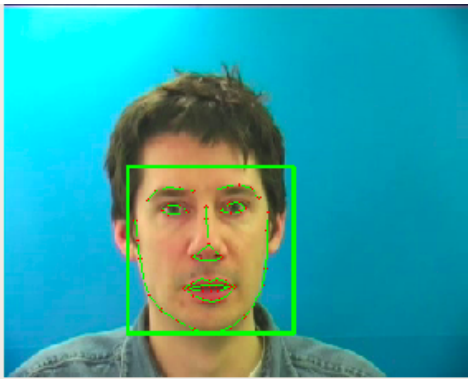
The enhanced audios were heard and were found to be close to the target audio with less interference. To quantify this, we used Short Term Objective Intelligibility (stoi) measures to measure the correlation (similarity) between the target audio and the mixed audio enhanced using above model. Mean correlation of 76.83% was observed in the testing set (50 videos).

Morrone et al. [2]

This paper used the existing GRID corpus to train its model. It stands out from other papers in that it uses stacked BLSTMs and difference of consecutive video frames (to simulate motion).

Audio samples are downloaded from the online available GRID database and are mixed to get the training, testing and validation data. Then, power-law compressed spectrograms are generated for all the generated mixed audio samples.

Dlib face detector and landmark extractor is used for getting the face and landmark (68 major) of the target speaker in the videos and stored in txt format.

After this, the model was implemented as indicated in the paper and as we didn't get a pre-trained model, we trained it using around 50 videos (less due to the resource limitation) selected from the GRID corpus.

Results:

The enhanced audios were somewhat clearer and near to the target audio but will improve once more data is used for training. Mean correlation of 45% was measured on testing set as calculated by stoi mentioned above. SNR and SDR comparisons were as follows:

```
Enhanced SDR: -2.14732 [0.95558]
Enhanced SIR: -2.04359 [0.96672]
Enhanced SAR: 18.51551 [1.25167]
Enhanced L2 (spectrogram): 29624.25391 [6319.35107]
Enhanced SNR: -1.18832 [0.52067]

Mixed SDR: -2.71063 [0.74176]
Mixed SIR: -2.69710 [0.74531]
Mixed SAR: 27.74200 [2.61929]
Mixed L2 (spectrogram): 56321.25781 [8779.49805]
Mixed SNR: -3.46551 [0.74705]
```

The results will get better once more data is used for training.

Limitations and Improvements

1. **Noise-invariant training** (used by Gabbay et. Al. [3]), which involves adding synthetic background noise taken from the voice of the target speaker to the training videos, was not used in both these papers. This can be beneficial as it forces our model to exploit visual features also while enhancing the speech as only audio features can not separate 2 utterances of the same speaker and can be easily implemented by mixing voice of target speaker as noise in training samples.

2. Both these papers only enhance the amplitude of the audio wave and uses the noisy phase as output phase. Approaches without above consideration works well for high SNR videos but as SNR decreases, the noisy phase become a bad approximation of ground truth value. **Predicting true phase using visual features** can give better results and can be implemented as done by Afouras et al. [4].

3. Both these paper use the short time Fourier transform(STFT) for the audio preprocessing. We can consider using Fractional Fourier transform as it has better performance regarding the frequency-time resolution.

4. The model used by Morrone et al. [2] is trained using the GRID Corpus in this paper. But Ephrat et al. [1] gave the speech processing community, the AVspeech database which is better than GRID corpus as it has audios in many languages, the target speaker is mostly facing the camera and the audio is interference free. This AVspeech can be used to train this model and might give better results.

5. Ephrat et al. [1] uses static video frames for training. But intuitively, the difference between frames is a better indication of motion and face movement than the frame themselves and so the difference of consecutive video frames can be used for training.

References

[1] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. 2018. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation. ACM Trans. Graph. 37, 4, Article 112 (August 2018), 11 pages. https://doi.org/10.1145/3197517.3201357

[2] G. Morrone, L. Pasa, V. Tikhanoff, S. Bergamaschi, L. Fadiga,and L. Badino, "Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments," arXiv preprint arXiv:1811.02480, 2018.

[3] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. 2017. Visual Speech Enhancement using Noise-Invariant Training. arXiv preprint arXiv:1711.08789 (2017).

[4] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in arXiv:1804.04121, 2018.