CS574 Assignment 2 Literature Survey

Speech separation using visual and speech cues

Group 15:

Shubham Goel	160101083
Shaurya Gomber	160101086
Archit Jugran	160101087
Rishabh Jain	160101088
Shashwat Jolly	160123036

Introduction

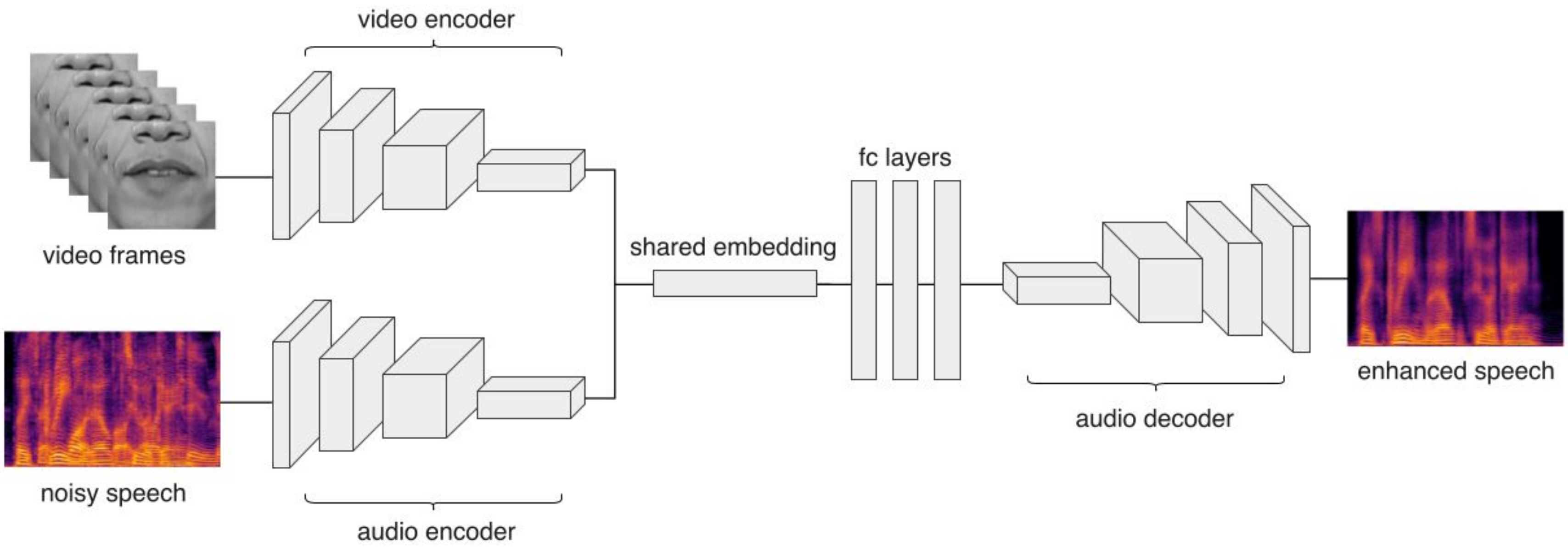
- The goal is to recover clean speech signals from a mixture of utterances
- Early solutions based on audio-only separation
- **Label Permutation Problem:** Associating each separated audio with its corresponding speaker
- Using both audio and visual cues gave much better solutions to the "Cocktail Party Problem"

Gabbay et al. (2017)

Key Points

- Introduced 'noise-invariant' training - speaker's voice added as background noise
- Forces the model to exploit visual features

Architecture



Results

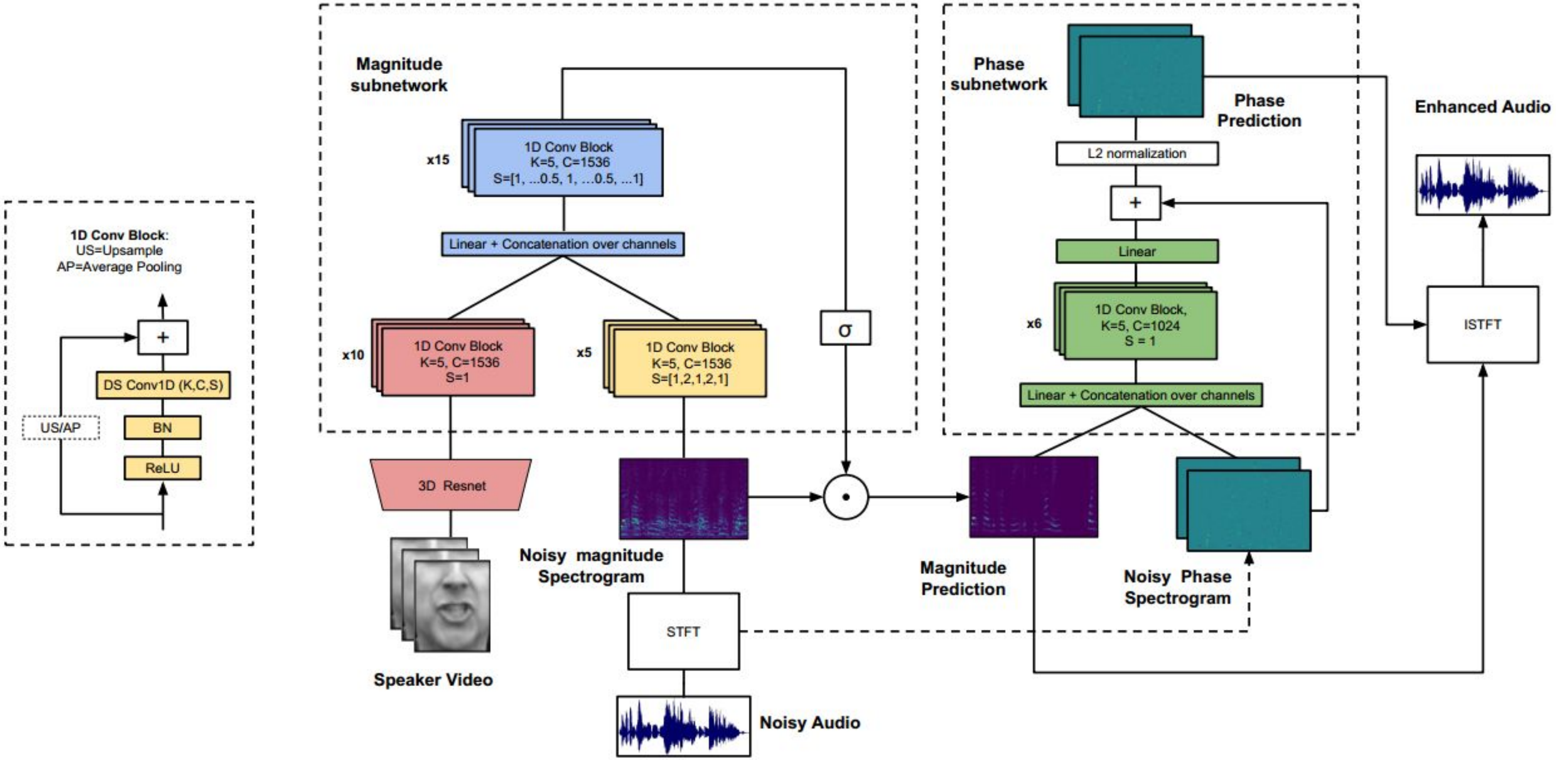
- Better than audio-only approaches, especially in self mixtures
- When model is trained without self mixtures, it is not capable of separating speech samples of the same voice
- Test data on GRID (in SNR[db]):
  - Without self: 2.81
  - With self: **4.05**

Afouras et al. (2017)

Key Points

- Output is spectrogram mask instead of the actual separated speech
- Takes the noisy phase into consideration along with the noisy magnitude
- Magnitude-only approaches work fine for high SNR inputs but phase considerations become important as SNR decreases

Architecture



Results

- Better results including phase considerations

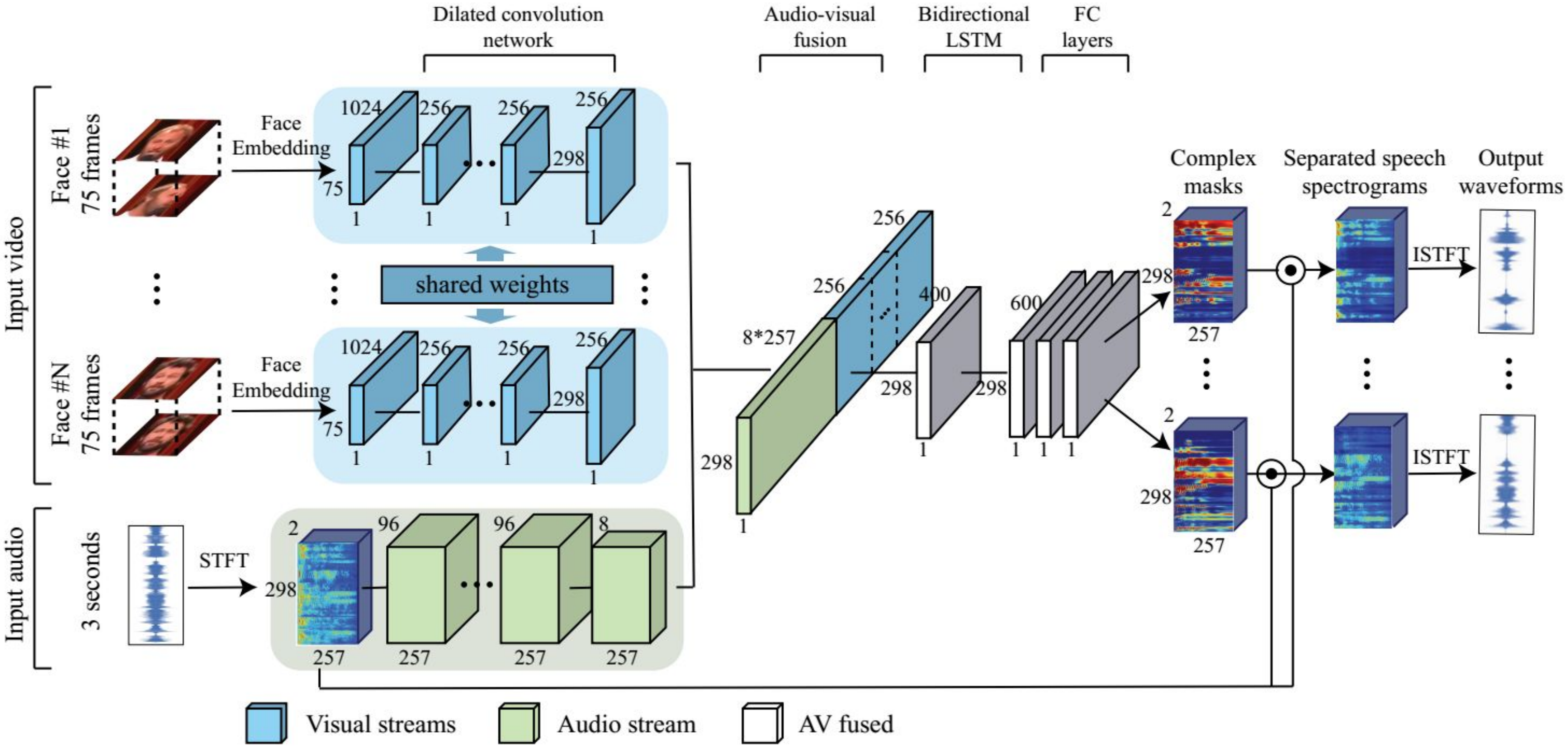
Magnitude	Phase	SDR	PESQ
Pr	GT	10.30	3.02
Pr	Mix	6.71	2.59
Pr	Pr	**7.91**	**2.67**

Ephrat et al. (2018)

Key Points

- Uses off-the-shelf Google Cloud Vision API to detect faces
- Uses BLSTM to utilise both past and future context information
- Trained on a new, large-scale audio-visual dataset, AVSpeech, carefully collected and processed by Google

Architecture



Results

- Results are consistently better than earlier works because of the high quality large dataset and more modern techniques such as BLSTM and dilated CNNs.

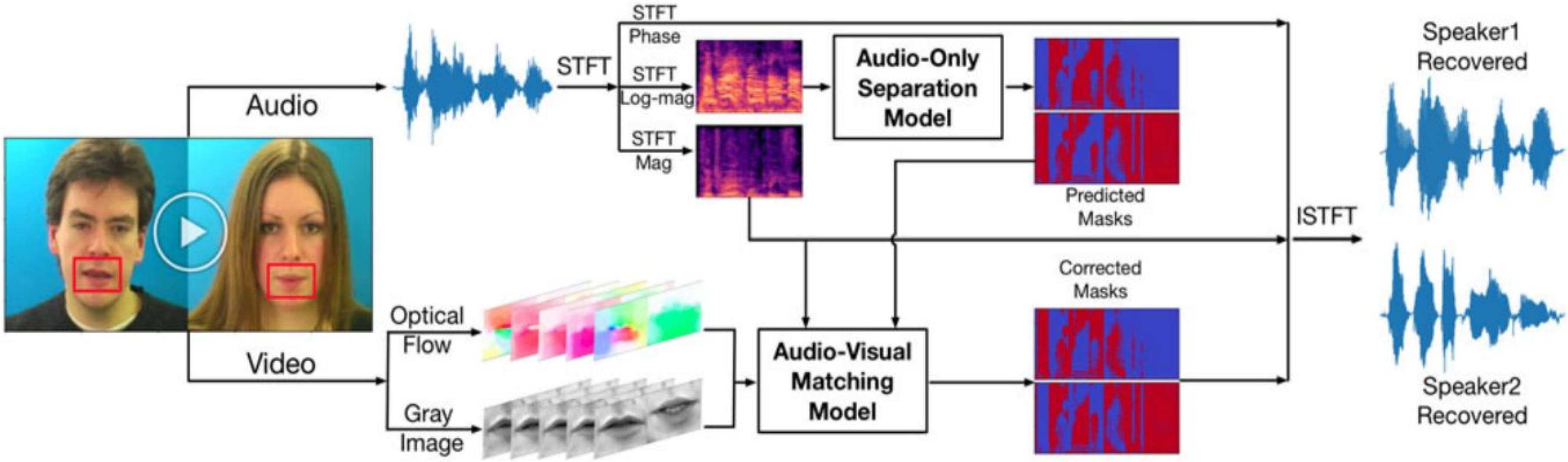
(on TCD-TIMIT)	**SDR**	**PESQ.**
Gabbey et al.	0.4	2.03
Ephrat et al.	4.1	2.42

Lu et al. (2018)

Key Points

- Audio-only method used to separate the input audio into clean speech signals
- Solves the label permutation problem using visual features
- **Modular:** Any audio-only method can be combined with the audio-visual training model

Architecture



Results

- Works well for same-gender voice mixtures, where audio-only methods suffer as the vocal characteristics are similar

(on GRID dataset)	SDR		
	Female - Female	**Male - Male**	**Overall**
No visual cues	6.23	6.45	7.89
With visual cues	8.40	7.02	8.64

- Improvement due to the audio-visual model becomes more pronounced as the SDR of the input signal becomes low

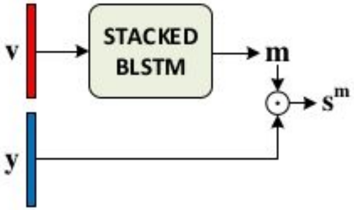
Morrone et al. (2019)

Key Points

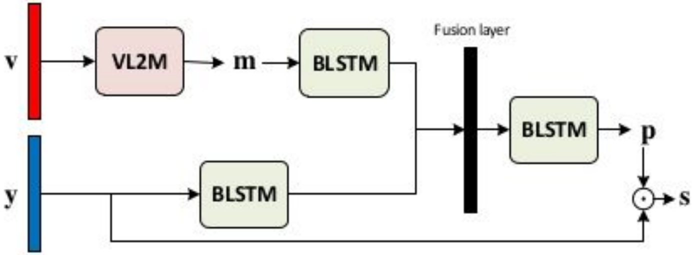
- Uses difference in consecutive video frames as input
- Differences are a better indicator of movement, which in turn are better indicator of utterances
- Uses LSTMs and BLSTMs which allow the model to retain context information in both audio and video

Architecture

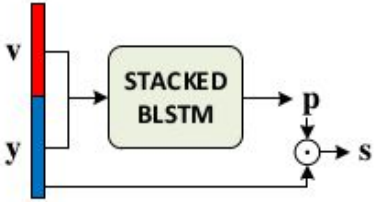
v: video input **y**: noisy spectrogram $\mathbf{s}^{\mathrm{m}}$: clean spectrogram TBM **s**: clean spectrogram IAM **m**: TBM **p**: IAM



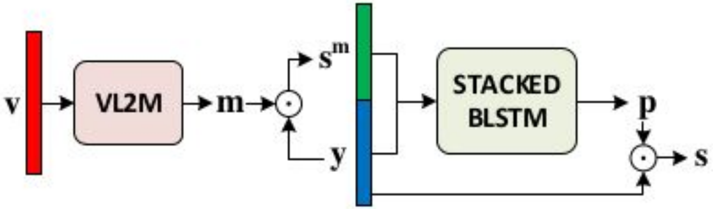
(a) VL2M



(b) VL2M_ref



(c) Audio-Visual concat



(d) Audio-Visual concat-ref

Results

- VL2M performs the worst, indicating that good predictions require acoustic context too, along with the visual context

(on GRID dataset)	**SDR**	**PESQ**
VL2M	3.02	1.81
VL2M_ref	6.52	2.53
AV concat	7.37	2.65
AV c-ref	8.05	2.70

- AV c-ref model shows it is better to refine an estimated/predicted spectrogram rather than refining the estimated mask

Conclusions

- Predictions using spectrogram masks as the pre-final output stage are usually better than direct predictions, except in the case of low SNR
- Modern approaches increasingly use more complex BLSTMs
    - **Not 'real-time'** as future context is required