Tutorial on Multi-modal Learning

[Code & models]

A deep-dive into Speaker Separation problem


Sindhu B Hegde


Aditya Agarwal


Bipasha Sen


Rudrabha Mukhopadhyay


Seshadri Mazumder

IIIT Hyderabad

Contact: {sindhu.hegde, aditya.ag, bipasha.sen, radrabha.m}@research.iiit.ac.in

Motivation:
Isolating & Enhancing the Target Speaker

- **Multi-modal learning**: Engaging multiple streams/modalities to perform a desired task.
- In a **cocktail-party** like environment, separating a single speaker from other speakers can be an extremely important task.
    - **Example**: Understanding the target speaker's speech in news debates as shown below.

- In such challenging situations, using additional information from visual modality along with the audio stream proves to be beneficial.

Contact: {sindhu.hegde, aditya.ag, bipasha.sen, radrabha.m}@research.iiit.ac.in

Speaker Separation: Potential Applications

A. Debate denoising - let one person speak at a time!

B. Automatic transcriptions with multiple speakers (such as in meetings).

C. Controlled hearing aids - enhances the speech of target speaker in noisy environments.

D. Blind speech separation.

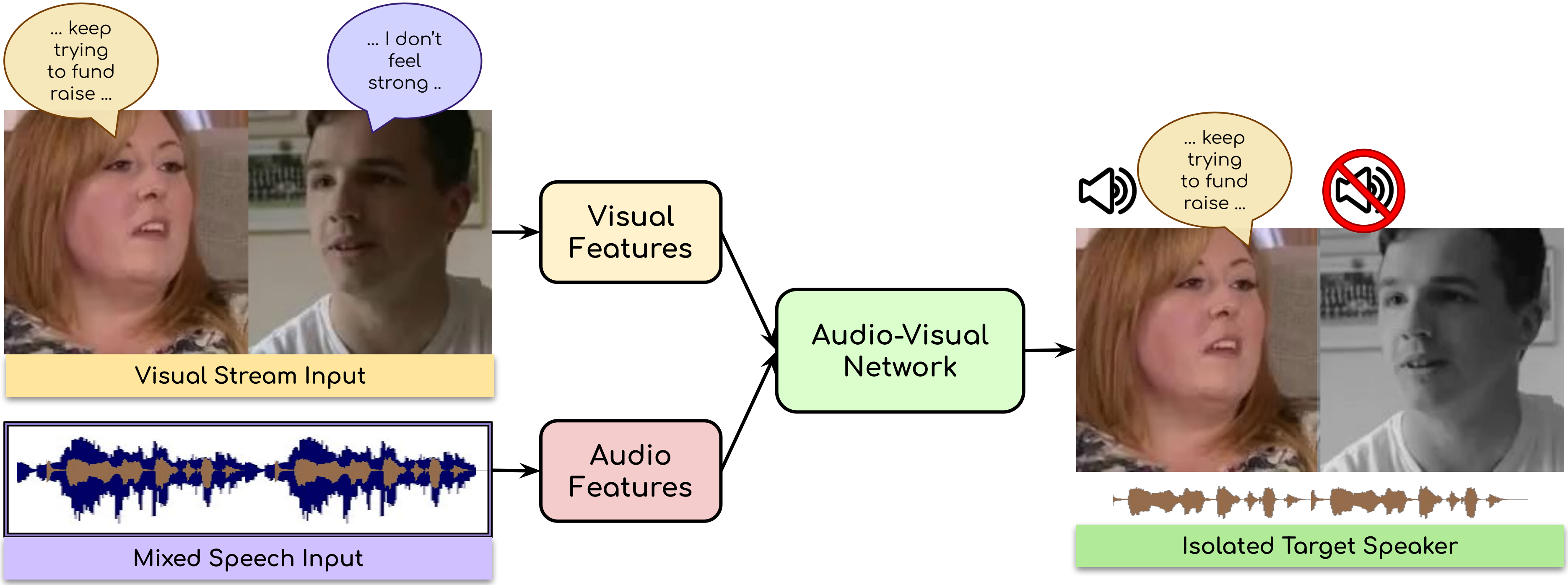
(a)

(b)

(c)

Contact: {sindhu.hegde, aditya.ag, bipasha.sen, radrabha.m}@research.iiit.ac.in

Audio-Visual Speaker Separation: Overview



Contact: {sindhu.hegde, aditya.ag, bipasha.sen, radrabha.m}@research.iiit.ac.in

Why do we need Visual Stream?

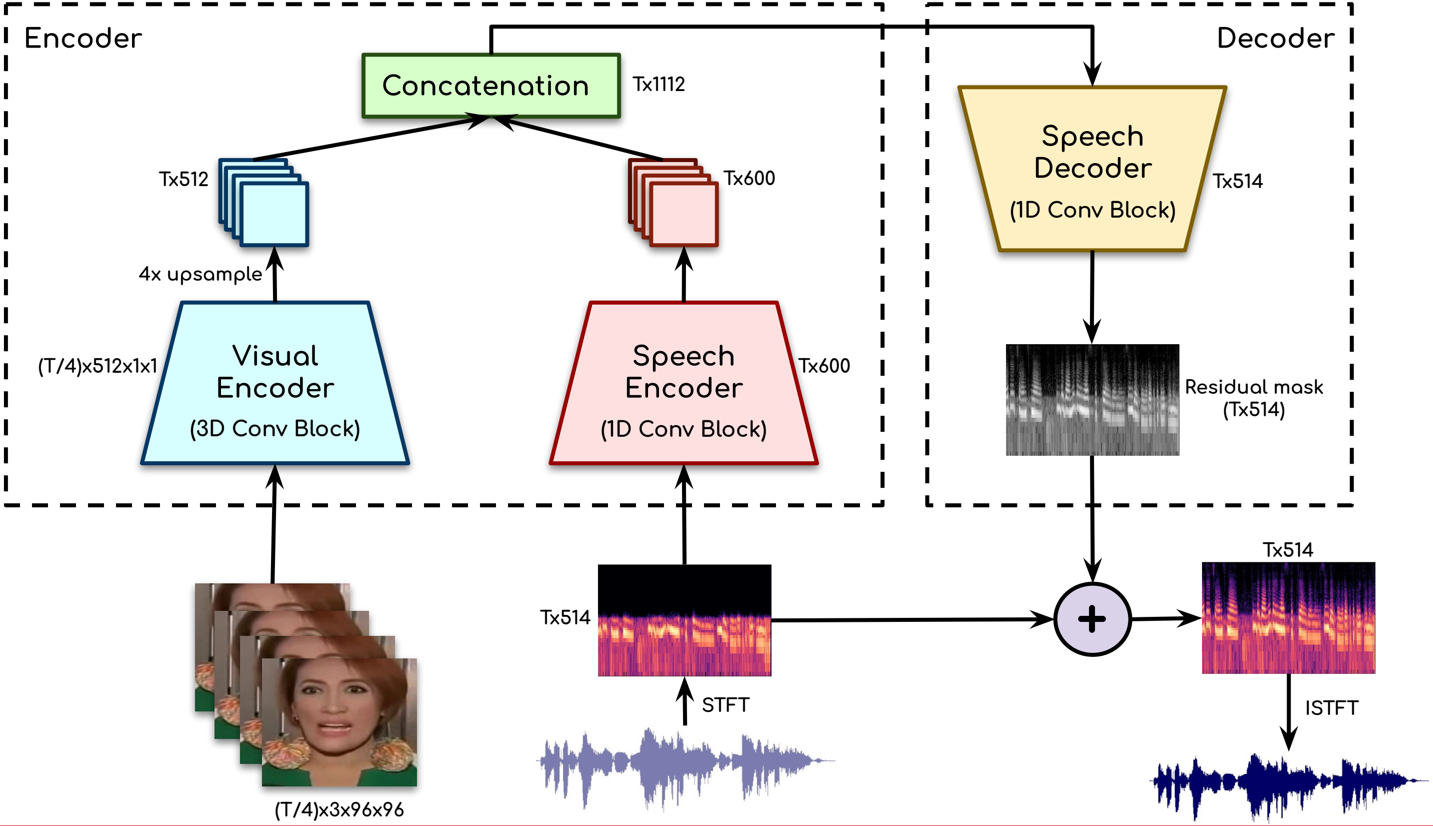
- The task of separating the speech can be done using the audio modality alone.
  - Very **hard to accomplish** this using solely the audio modality.
  - Audio alone falls short is bringing all the information.
  - **Permutation problem:** No easy way to associate each separated audio source with its corresponding speaker in the video (example - play this particular speaker)

Play the lady's voice -

- **Visual stream** along with the auditory input has proven to be extremely beneficial.
  - Visual stream allows us to "**focus**" the audio on the desired target speakers.
  - It also improves the overall **speaker separation performance**.

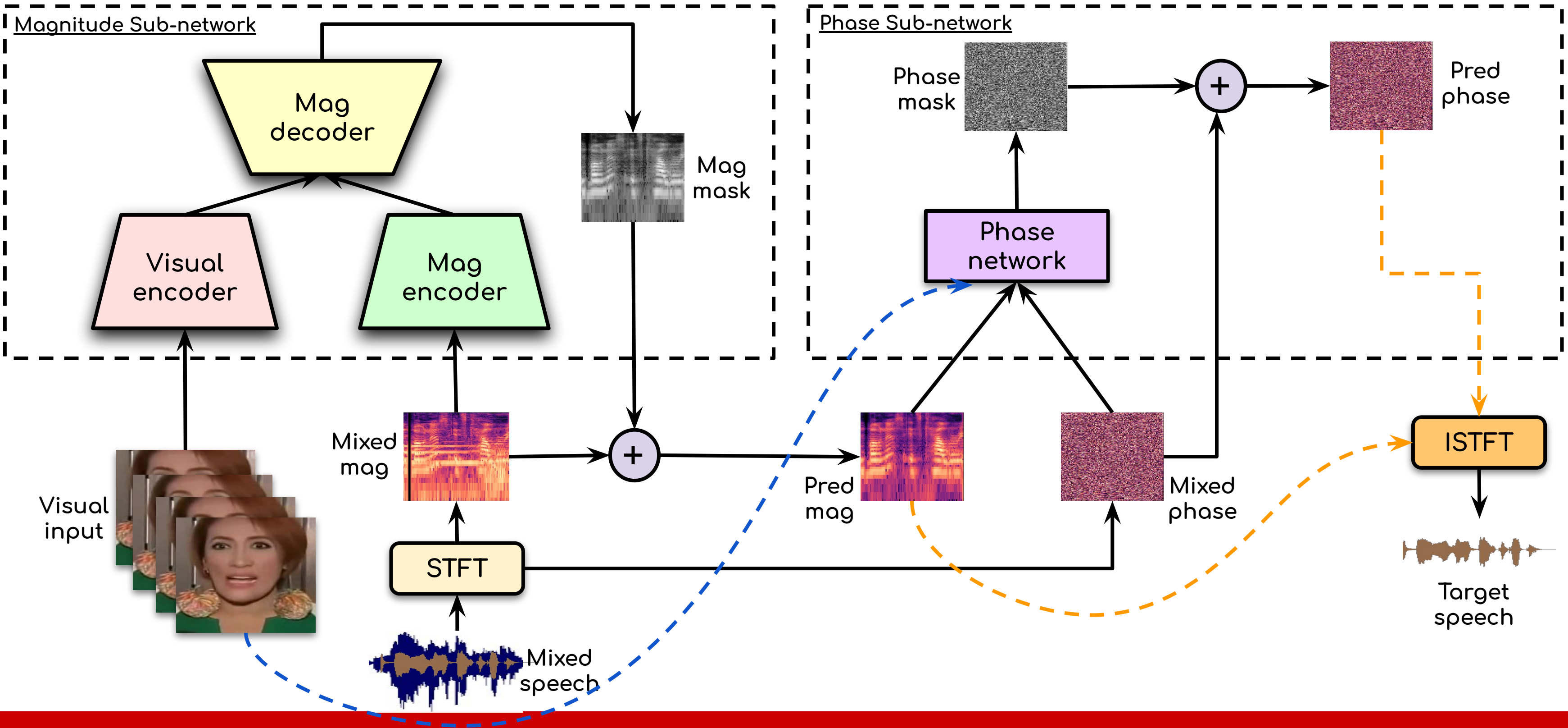
Contact: {sindhu.hegde, aditya.ag, bipasha.sen, radrabha.m}@research.iiit.ac.in

Audio-Visual Network: Architecture Overview



Contact: {sindhu.hegde, aditya.ag, bipasha.sen, radrabha.m}@research.iiit.ac.in

Audio-Visual Network: Detailed Architecture



Contact: {sindhu.hegde, aditya.ag, bipasha.sen, radrabha.m}@research.iiit.ac.in

Audio-Visual Network: Representations

- **Audio-Visual network**: Takes both the visual stream and the mixed auditory stream as the input and generates the isolated speech for the target speaker.

- **Audio representation**:
  - Extract **linear spectrogram** using short-time Fourier transform (***STFT***) from 1-second segment of mixed speech input.
  - Decompose the complex time-frequency representation ($T \times 257$) into **magnitude** and the **phase** components, and normalize them between [0, 1].
  - The mag and the phase components, each of dimension ($T \times 257$) act as input to the respective magnitude and phase encoder networks.

- **Visual representation**:
  - The corresponding visual 1-second of frames are extracted (25 frames).
  - The resized frames ($96 \times 96 \times 3$) act as input to the visual encoder.

Contact: {sindhu.hegde, aditya.ag, bipasha.sen, radrabha.m}@research.iiit.ac.in

Audio-Visual Network: Training details

- <u>Magnitude Sub-network:</u>
  - Visual Encoder:
    - Processes the input images using a stack of residual *2D*-convolution blocks and generates a **visual embedding** for each frame (*T'x512*) where *T'*=25 frames.
    - The output of the visual encoder module is **up-sampled** 4× to match the spectrogram temporal dimension (*Tx512*) where *T*=100.
  - Mag Encoder:
    - Processes the input mixed mag representation (*T x 257*) using a stack of *1D*-convolution blocks with residual connections.
    - Convolutions are performed along the **temporal dimension**, by considering the frequency component of the input spectrograms as channels (*Tx600*).
  - Mag Decoder:
    - Concatenate the learned features of each stream along the channels (*Tx1112*).
    - Processes the **fused representation** using a stack of residual *1D*-convolution blocks.
    - **Output**: A **magnitude mask** (Tx257) that is **added** to input magnitude followed by a sigmoid activation to generate the **enhanced magnitude spectrogram** output (*Tx257*).

Contact: {sindhu.hegde, aditya.ag, bipasha.sen, radrabha.m}@research.iiit.ac.in

Audio-Visual Network: Training details

- <u>Phase Sub-network:</u>
    - Concatenate the predicted magnitude ($Tx257$), visual embeddings ($Tx512$) and the input mixed phase ($Tx257$) representations along the channels ($Tx1026$).
    - The phase network processed the fused representation using a stack of residual 1D convolution layers.
    - Output: A **residual phase mask** ($Tx257$) that is **added** to the input phase followed by a sigmoid activation to generate the **enhanced phase spectrogram** output ($Tx257$).

- The **enhanced speech output** is obtained by computing the **inverse-STFT** (*ISTFT*) from the magnitude and phase predictions.

- <u>Losses:</u>
    - Magnitude prediction: **L1** loss
    - Phase prediction: **Cosine** similarity
    - **Total loss** = Mag loss + Phase loss

Contact: {sindhu.hegde, aditya.ag, bipasha.sen, radrabha.m}@research.iiit.ac.in

Dataset and Experimental setup

- VoxCeleb2 dataset:
  - A large-scale talking-face video dataset containing **celebrity** videos.
  - Contains over 1 million utterances for 6,112 celebrities.
  - A challenging dataset that spans a wide variety of identities, languages, and face poses.

Fig.: Dataset samples.

Table.: Statistics of the VoxCeleb2 dataset.

	Train	Test
# speakers	5,994	118
# videos	145,569	4,911

Contact: {sindhu.hegde, aditya.ag, bipasha.sen, radrabha.m}@research.iiit.ac.in

Qualitative Results

Contact: {sindhu.hegde, aditya.ag, bipasha.sen, radrabha.m}@research.iiit.ac.in

Qualitative Results

Contact: {sindhu.hegde, aditya.ag, bipasha.sen, radrabha.m}@research.iiit.ac.in

Qualitative Results

Contact: {sindhu.hegde, aditya.ag, bipasha.sen, radrabha.m}@research.iiit.ac.in

Q&A Break

Time for interaction

Contact: {sindhu.hegde, aditya.ag, bipasha.sen, radrabha.m}@research.iiit.ac.in

Time for Code Walk-through!

- **Repository:** https://github.com/Sindhu-Hegde/speaker-separation
  - Clone and star the repo 😄

- The repo has the **complete train and test codes** along with the pre-trained model for the task of speaker separation.
  - A **demo** inference file (collab notebook) is also provided.

Related works:

1. **The Conversation: Deep Audio-Visual Speech Enhancement.**Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, In *Interspeech 2018*.
2. **Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation**. Ephrat, A., Inbar Mosseri, Oran Lang, Tali Dekel, K. Wilson, Avinatan Hassidim, W. Freeman and Michael Rubinstein, In *ACM Transactions on Graphics (ToG) 2018*.

Contact: {sindhu.hegde, aditya.ag, bipasha.sen, radrabha.m}@research.iiit.ac.in