NGHIÊN CỨU MÔ HÌNH TRANSFORMER VÀ CƠ CHẾ ATTENTION CHO BÀI TOÁN TÁCH GIỌNG NÓI ĐƠN KÊNH ĐA PHƯƠNG THỨC

Nguyễn Hùng Thịnh - 250201030

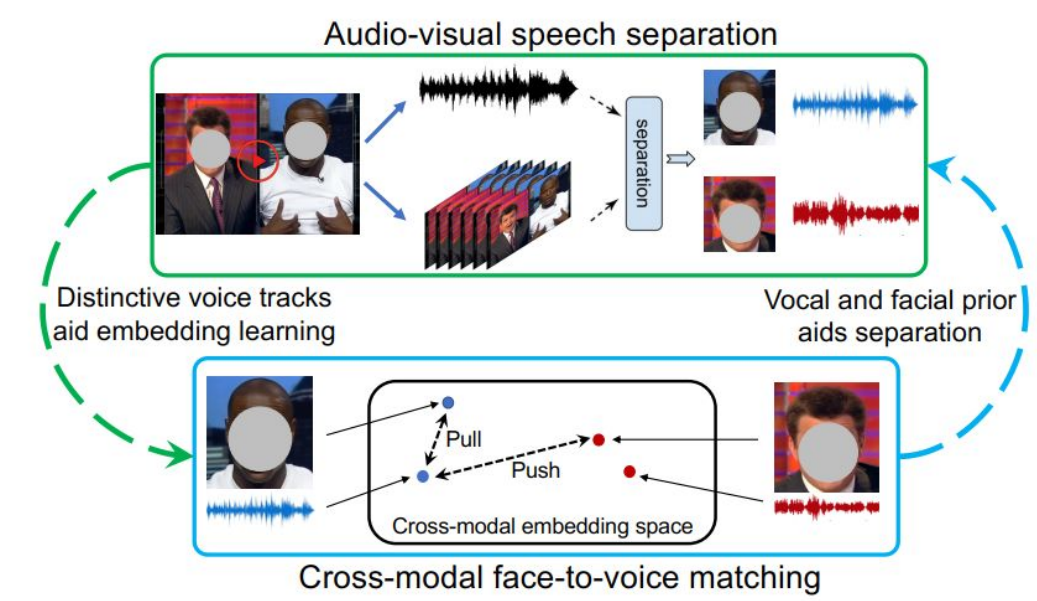
Thông tin chung

- Lớp: CS2205.CH201
- Link Github: thinhGithub/CS2205.CH201-250201030
- Link YouTube video: thinhYoutube/CS2205.CH201-250201030

Nguyễn Hùng Thịnh
250201030

Giới thiệu

- **Bài toán (Problem): Tách giọng nói đơn kênh đa phương thức**
  - Tách giọng mục tiêu trong môi trường ồn ào và nhiều người nói
  - Sử dụng kết hợp dữ liệu Âm thanh (Audio) và Hình ảnh (Visual)



Giới thiệu

- **Tính thời sự:** Phương pháp CNN hạn chế trong xử lý chuỗi dài; xu hướng hiện nay là áp dụng Transformer nắm bắt ngữ cảnh toàn cục
- **Bài toán tính toán (Computational Problem):**
  - **Input:** Video hỗn hợp *V* chứa âm thanh trộn lẫn *x(t)* (nhiều người nói + nhiễu) và hình ảnh khuôn mặt người nói
  - **Output:** Tín hiệu âm thanh sạch tương ứng với từng người nói
- **Ứng dụng:** Hỗ trợ thiết bị trợ thính, cải thiện chất lượng hội thoại video (Zoom/Teams), trích xuất thông tin video thực tế.

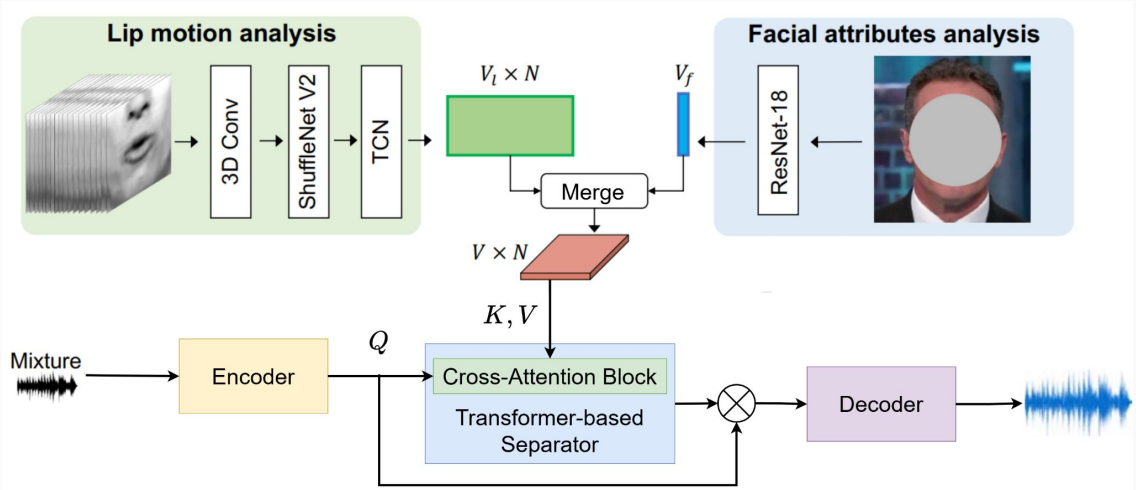
Mục tiêu

- **1. Khắc phục hạn chế của CNN truyền thống:** Áp dụng Transformer để mô hình hóa sự phụ thuộc dài hạn của tín hiệu âm thanh.
- **2. Tối ưu hóa kết hợp (Fusion):** Xây dựng Cross-Modal Attention liên kết đặc trưng hình ảnh và âm thanh, thay thế phép nối (Concat).
- **3. Đạt hiệu suất SOTA:** so sánh với baseline VisualVoice [1] và các phương pháp Audio-only

[1] Ruohan Gao, Kristen Grauman: VisualVoice: Audio-Visual Speech Separation with Cross-Modal Consistency. CVPR 2021: 15490–15500

Nội dung và Phương pháp

Quy trình thực hiện theo các bước:



- **Bước 1:** Tiền xử lý dữ liệu
- **Bước 2:** Xây dựng luồng xử lý
  - Visual Stream
  - Audio Stream
- **Bước 3:** Cơ chế Fusion
  - Cross-Modal Attention
  - Attention(Q, K, V): với Q là Audio, K, V là Visual.
- **Bước 4:** Huấn luyện & Tối ưu hóa (Optimization)

$$L_{total} = L_{mask} + \lambda_1 L_{cross-modal} + \lambda_2 L_{consistency}$$

Kết quả dự kiến

- **Dữ liệu kiểm thử (Tiếng Anh):** VoxCeleb2, LRS2,...
- **Định lượng (Quantitative Metrics):**
  - **Chỉ số SDR:** Kỳ vọng đạt mức > 11dB (trên tập VoxCeleb2), cao hơn mức ~10.2dB của VisualVoice.
  - **Chỉ số PESQ:** Kỳ vọng cải thiện > 0.2 điểm so với baseline.
- **So sánh (Comparison):**
  - Chứng minh sự vượt trội so với nhóm phương pháp Audio-only
  - Chứng minh sự vượt trội so với VisualVoice gốc (về khả năng xử lý khi lip-motion bị khuất/mờ) nhờ cơ chế Attention.

Tài liệu tham khảo

- [1] Ruohan Gao, Kristen Grauman: VisualVoice: Audio-Visual Speech Separation with Cross-Modal Consistency. CVPR 2021: 15490–15500
- [2] Shengkui Zhao, Zexu Pan, Bin Ma: ClearerVoice-Studio: Bridging Advanced Speech Processing Research and Practical Deployment. arXiv:2506.19398, 2025
- [3] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, Jianyuan Zhong: Attention Is All You Need in Speech Separation. ICASSP 2021: 21–25
- [4] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, Abdelrahman Mohamed: Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction. arXiv:2201.02184, 2022
- [5] Joon Son Chung, Arsha Nagrani, Andrew Zisserman: VoxCeleb2: Deep Speaker Recognition. arXiv:1806.05622, 2018