FLASHFORMER: MÔ HÌNH TÁCH GIỌNG NÓI ĐƠN KÊNH TỐI ƯU TỐC ĐỘ VÀ HIỆU SUẤT

Nguyễn Đức Đạt - 250201006

Thông tin chung

Lớp: CS2205.CH201

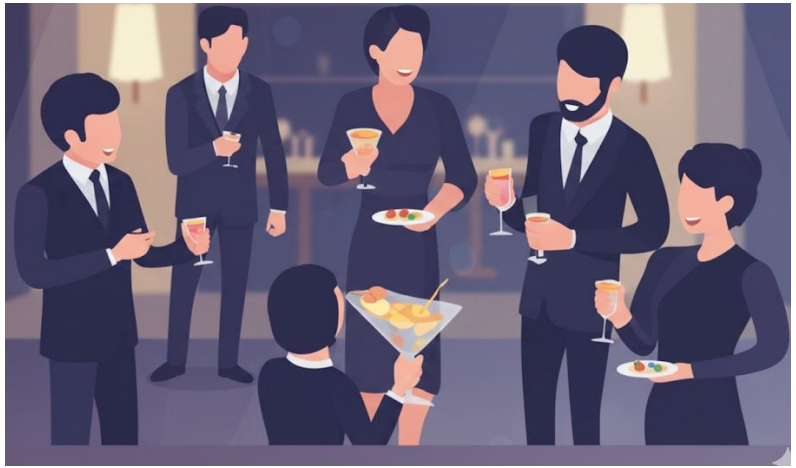
Họ và tên: Nguyễn Đức Đạt - 250201006

Github (https://github.com/nddat1811/CS2205.CH201/)

Youtube: (https://youtu.be/_p5t0YEh90A)

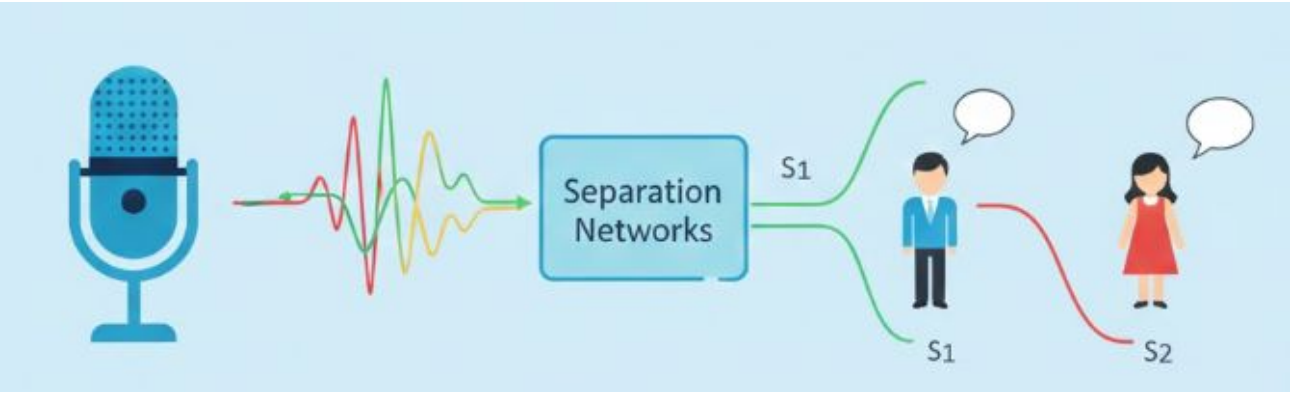
Tóm tắt

- **Thách thức**: Nghe rõ giọng một người trong môi trường “tiệc cocktail”
- **Hạn chế**: Các mô hình hiện nay tiêu tốn nhiều tài nguyên với đoạn âm thanh dài
- **Mô hình FlashFormer:** Tối ưu phần cứng giúp tăng tốc độ xử lý.



Giới thiệu

- **Tách giọng nói đơn kênh** là bài toán cốt lõi trong xử lý tiếng nói, ứng dụng rộng rãi trong trợ lý ảo, hội nghị trực tuyến và thiết bị thông minh.
- Các mô hình hiện đại đạt độ chính xác cao nhưng **chi phí tính toán lớn**, khó triển khai thời gian thực
- Transformer mô hình hóa ngữ cảnh tốt nhưng **độ phức tạp cao với chuỗi dài**
- Nghiên cứu đề xuất **FlashFormer** nhằm cân bằng giữa **chất lượng tách giọng và tốc độ xử lý**



Mục tiêu

01

- Phân tích và đánh giá các hạn chế của các mô hình hiện đại

02

- Xây dựng kiến trúc lai Transformer–FSMN tích hợp Flash Attention

03

- Huấn luyện và kiểm chứng hiệu năng mô hình

Nội dung và Phương pháp

Khảo sát các mô hình tách giọng đơn kênh (Mossformer2, DPRNN,..)

Xây dựng kiến trúc FlashFormer bằng Pytorch

Huấn luyện mô hình trên bộ dữ liệu WSJ0-2mix, Libri2Mix, LibriheavyMix

Đánh giá dựa trên các chỉ số SI-SDR, SI-SDRi

So sánh hiệu năng với các phương pháp khác

Ứng dụng vào thực tế

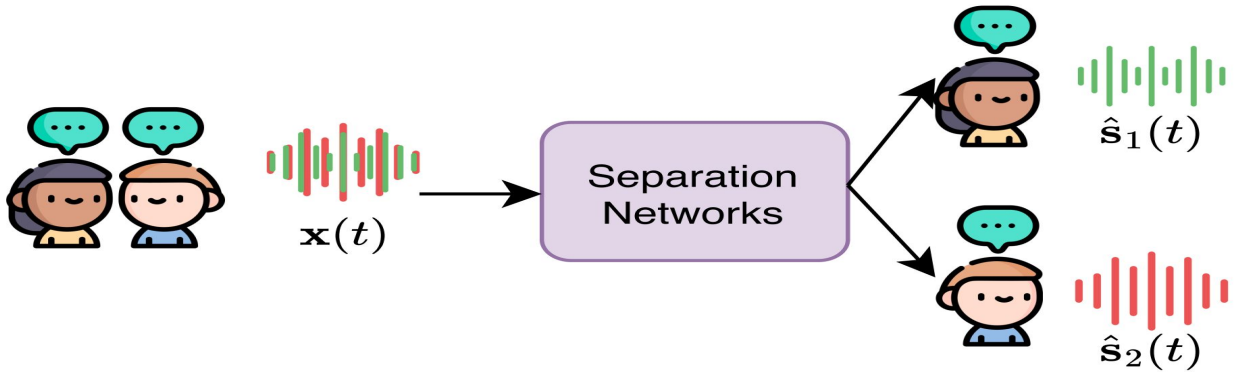
Kết quả dự kiến

Xây dựng thành công mô hình FlashFormer

Tăng tốc độ xử lý, hiệu năng tốt hơn

Hiệu quả tốt hơn trên dữ liệu tín hiệu dài

Ứng dụng kết quả đạt được vào môi trường thực tế



Tổng hợp

Vấn đề	Giải pháp đề xuất	Kết quả dự kiến	Ứng dụng thực tế
Mô hình mạnh nhưng **tốn tài nguyên**, khó thời gian thực	**FlashFormer** (Transformer + FSMN + Flash Attention)	**Tăng tốc suy luận**, giảm bộ nhớ, **chất lượng tốt hơn**	Hội nghị trực tuyến, trợ lý ảo, thiết bị IoT thời gian thực

Tài liệu tham khảo

[1] Yi Luo and Nima Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," 2019

[2] Yi Luo, Zhuo Chen, and Takuya Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," 2020

[3] Shengkui Zhao, Yukun Ma, Chongjia Ni, Chong Zhang, Hao Wang, Trung Hieu Nguyen, Kun Zhou, Jia Qi Yip, Dianwen Ng, and Bin Ma, "MossFormer2: Combining Transformer and RNN-Free Recurrent Network for Enhanced Time-Domain Monaural Speech Separation," 2024

[4] Shiliang Zhang, Cong Liu, Hui Jiang, Si Wei, Lirong Dai, and Yu Hu, "Feedforward Sequential Memory Networks: A New Structure to Learn Long-term Dependency," 2015

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, and Łukasz Kaiser, "Attention Is All You Need," 2017

[6] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra and Christopher Ré, "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness," 2022

[7] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge and Emmanuel Vincent, "LibriMix: An Open-Source Dataset for Generalizable Speech Separation," 2020

[8] Zengrui Jin, Yifan Yang, Mohan Shi, Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Liyong Guo, Lingwei Meng, Long Lin, Yong Xu, Shi-Xiong Zhang, Daniel Povey, "LibriheavyMix: A 20,000-Hour Dataset for Single-Channel Reverberant Multi-Talker Speech Separation, ASR and Speaker Diarization," 2024