RESEARCH ON TRANSFORMER MODEL AND ATTENTION MECHANISM FOR MONAURAL AUDIO-VISUAL SPEECH SEPARATION

Thinh Nguyen Hung[1, 2]

[1] University of Information Technology, Ho Chi Minh City, Vietnam

[2] Vietnam National University, Ho Chi Minh City, Vietnam

What ?

We propose a Transformer-based Audio-Visual Speech Separation framework designed to isolate target speech:

- Transformer Backbone: Replaces traditional CNNs to capture global context and long-term audio dependencies.
- Cross-Modal Attention: A novel fusion mechanism to dynamically align lip motion and facial attributes with the audio stream.

Why ?

- The Challenge: Audio-only models struggle in noisy, multi-talker environments.
- The Gap: Current CNN-based methods struggle with long sequence modeling.
- The Solution: Visual cues (faces/lips) provide essential priors to guide separation when audio is corrupted.
- Applications: Critical for improving hearing aids and video conferencing quality.

Overview

Visual Stream Encoder → Audio Stream Encoder → Cross-Modal Attention

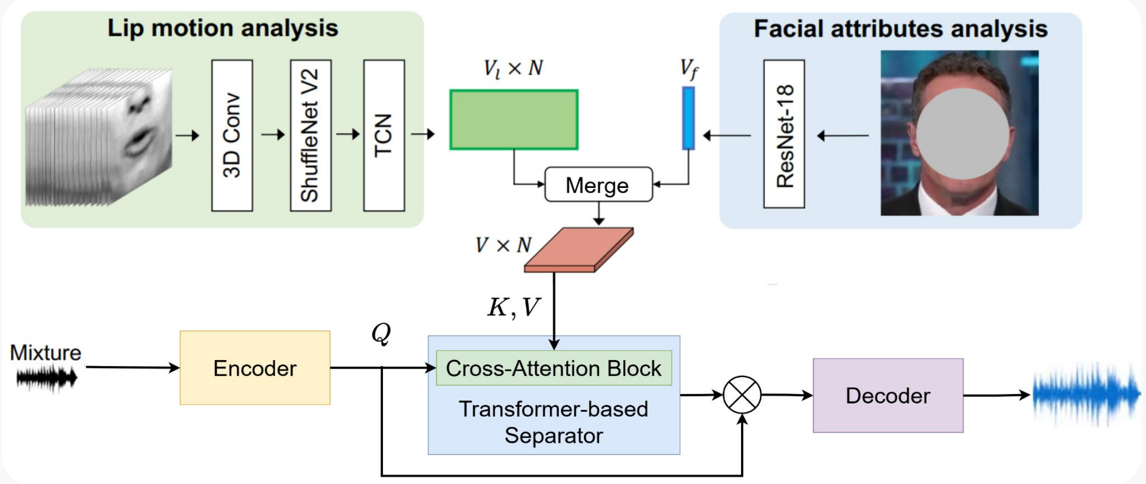

Figure 1. Proposed System Architecture.

Description

1. Visual Stream Encoder

- **Lip Motion Branch:** Extracts temporal features using 3D Conv, ShuffleNet V2, & TCN.
- **Facial Attributes Branch:** Encodes static speaker identity via ResNet-18.
- **Output:** Generates visual embeddings to guide the audio separation.

2. Audio Stream Encoder

- **Transformer Backbone:** Captures global context & long-term dependencies, overcoming CNN limits.
- **Semantic Encoding:** Encodes noisy mixture into high-level semantic features.
- **Attention Role:** Generates the Query to retrieve visual cues (Key, Value).

3. Cross-Modal Attention

- **Method:** Cross-Attention aligns Audio (Q) with Visual (K, V).
- **Function:** Dynamically filters noise using visual cues.
- **Result:** Robust separation even with occlusions.

Multi-task learning

- **Joint Learning:** Simultaneously optimizes for Speech Separation and Face-Voice Matching.
- **Embedding Alignment:** Forces separated voice and input face into a shared space to ensure identity consistency
- **Composite Loss:** Minimizes a total loss combining mask prediction and cross-modal constraints.

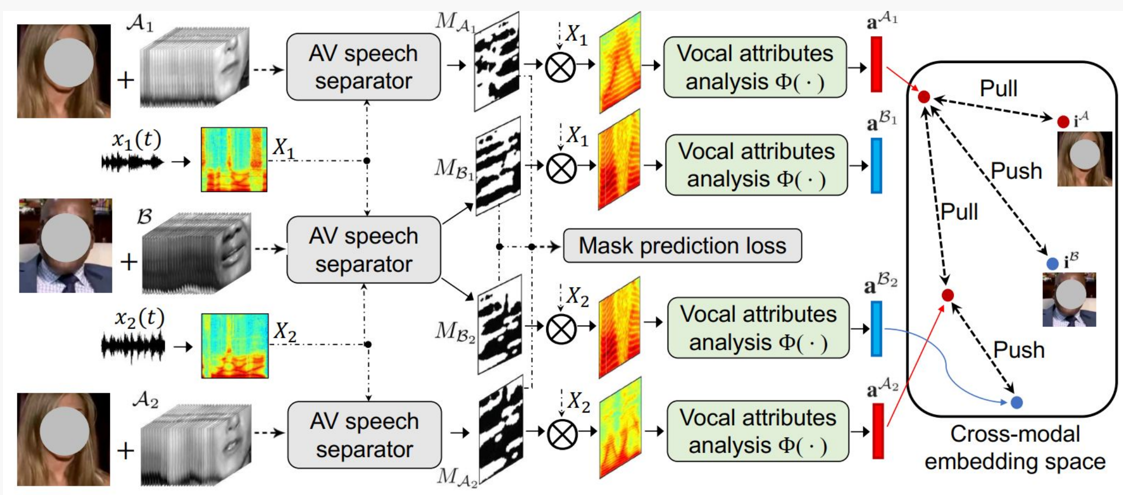

Figure 2. Multi-task learning framework

NII

Thinh Nguyen Hung–University of Information Technology, Ho Chi Minh City, Vietnam…

TEL : *******   Email : thinhnh.20@grad.uit.edu.vn**