Audio - Visual Speech Recognition

Abstract

In this survey, we review the paper ”End-to-End Audio-Visual Speech Recognition with Conformers” by Pingchuan Ma, Stavros Petridis, and Maja Pantic. The paper presents a hybrid CTC/Attention model incorporating a Convolution-Augmented Transformer (Conformer) trained on raw audio waveforms and video frames for Audio-Visual Speech Recognition (AVSR). This work provides substantial improvements over state-of-the-art methods, including results on the LRS2 and LRS3 datasets, using end-to-end training approaches.

1 Problem Statement

Audio-Visual Speech Recognition (AVSR) is the task of transcribing spoken language using both audio and visual inputs. The challenge is that traditional speech recognition systems struggle in noisy environments, where audio may be difficult to process accurately. The visual stream (lip movements) is unaffected by noise and can complement the audio stream. The goal is to improve speech recognition performance, particularly in noisy conditions, by using an end-to-end approach that processes raw audio waveforms and video frames.

2 Introduction

AVSR has garnered attention due to its robustness in environments with audio degradation. Previous methods rely on extracting features from audio and visual inputs separately before recognition. This two-stage approach has limitations, especially in noisy environments, where pre-extracted features may not fully represent the signal. This paper introduces an end-to-end model that processes raw pixels and waveforms, offering a more integrated and noise-robust solution. The proposed architecture also replaces recurrent networks with Conformers, which better handle long-term dependencies and temporal information.

3 Datasets

The study uses two publicly available datasets, LRS2 and LRS3:

- **LRS2 dataset:** Consists of 224.1 hours of video clips, collected from BBC programs. It contains 96,318 utterances for pre-training and 45,839 for training.
- **LRS3 dataset:** A larger dataset with 438.9 hours of TED and TEDx talks. It includes 118,516 utterances for pre-training and 31,982 for training-validation.

These datasets are challenging due to variations in head pose, illumination, and speaking style, making them ideal for testing AVSR models.

4 Related Work

1. **LipNet: End-to-End Sentence-Level Lipreading** - Assael et al. (2016)

Assael et al. introduced one of the first end-to-end models for visual speech recognition (VSR) using 3D Convolutions combined with Gated Recurrent Units (GRUs) to recognize visual speech from video sequences. Their model, called LipNet, was trained on the GRID corpus and predicted entire sequences of words rather than isolated phonemes or letters. This was an important shift in the field as it demonstrated the potential of deep learning to bypass traditional feature extraction stages. The success of this model laid the groundwork for future end-to-end architectures in both VSR and AVSR.

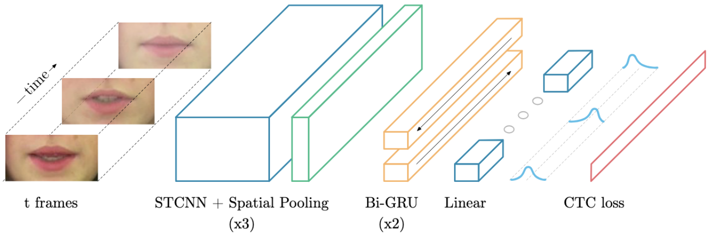
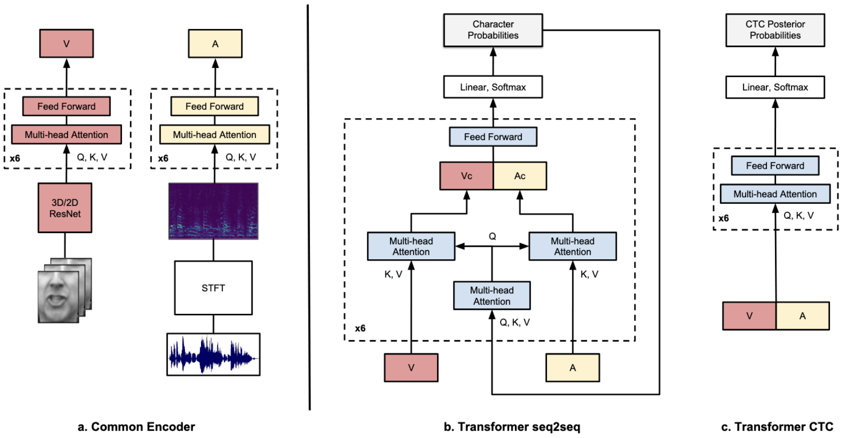


Figure 1: Architecture - LipNet

- **Input (t frames)**: The input to the model is a sequence of video frames (t frames), which captures the lip movements over time.
- **Spatiotemporal CNN (STCNN) + Spatial Pooling (x3)**: STCNN (Spatiotemporal Convolutional Neural Network) captures spatial features (from individual video frames) and temporal features (across the sequence of frames). This operation is performed three times with spatial max-pooling to downsample the data and retain important information.
- **Bidirectional GRU (Bi-GRU) (x2)**: Bi-GRU (Bidirectional Gated Recurrent Units) processes the sequence of lip movements both forward and backward, which helps in learning from both past and future context in the sequence. This is applied twice for better sequence learning.
- **Linear Transformation + Softmax**: After the GRU layers, a linear transformation is applied at each time step to reduce the data into a simpler form, which is followed by a softmax function that produces probabilities over the vocabulary (plus a CTC blank).
- **CTC Loss (Connectionist Temporal Classification)**: CTC Loss handles the variability in timing of the speech events, allowing the model to predict sequences (words or characters) even if the exact alignment between input frames and output labels is not known.

2. **Deep Audio-Visual Speech Recognition** - Afouras et al.(2018)

Afouras et al. developed one of the first transformer-based sequence-to-sequence models for AVSR. They used pre-computed visual features and log-Mel filter-bank features as inputs to the model, which significantly improved performance over prior approaches. The study demonstrated that deep neural networks could outperform traditional handcrafted feature extraction methods when applied in an end-to-end manner. However, their approach still relied on pre-computed features.



a. Common Encoder

b. Transformer seq2seq

c. Transformer CTC

Figure 2: Audio-visual speech recognition models

- **Visual Input:** The video sequence (lip movements or facial expressions) is processed by a *spatio-temporal ResNet* to extract features over space and time.
- **Audio Input:** The audio signal is converted into *spectrograms* (a visual representation of sound) by applying the *Short Time Fourier Transform (STFT)*.
- **Separate Encoders** Each modality (video and audio) is passed through its own *Transformer encoder* to capture complex features specific to the respective input.
- **TM-seq2seq Architecture** The model uses *Transformers* for both encoding and decoding.At each decoder layer: Video (V) and audio (A) encodings are attended to separately by *independent multi-head attention* modules.The context vectors V_c (for video) and A_c (for audio) generated by these attention mechanisms are concatenated channel-wise. The concatenated context vectors are then fed into the *feed-forward layers* to produce the output.
- **Attention Mechanism** The multi-head attention mechanism relies on the following components: **Key (K)**, **Value (V)**, and **Query (Q)** tensors are used to focus on different parts of the input data.In **self-attention layers**, we always have $Q = K = V$.In **encoder-decoder attention**, K and V represent the encodings (either from video or audio), while Q is the output from the previous decoding step, or the prediction from the previous layer.
- **TM-CTC Architecture** The TM-CTC model is built using stacks of *self-attention and feed-forward layers.*This model produces *CTC posterior probabilities* for each input frame.

3. **Attention is All You Need** - Vaswani et al. (2017)

Vaswani et al. introduced the Transformer architecture, which replaced traditional recurrent neural networks (RNNs) with self-attention mechanisms. This architecture improved training efficiency and achieved superior performance across a variety of natural language processing (NLP) tasks. The use of self-attention mechanisms enabled models to capture long-range dependencies in the input data more effectively.

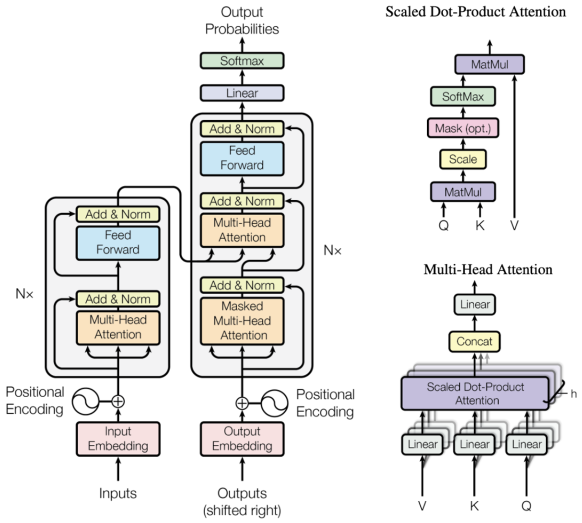


Figure 3: Transformer Model Architecture

- **Transformer Model Architecture :** Consists of an **encoder** and a **decoder**, both containing multi-head attention and feed-forward layers.Each layer is followed by *normalization* and *residual connections.* The **encoder** processes the input sequence, while the **decoder** generates output sequences based on shifted input. **Positional encodings** are added to embeddings to preserve sequence information, since Transformers do not have a built-in sense of order.

- **Scaled Dot-Product Attention :** Takes **Query (Q)**, **Key (K)**, and **Value (V)** matrices.Computes *attention scores* by performing dot products between Q and K. The dot products are scaled by the square root of the dimension of the keys, then passed through a **softmax** function to normalize. Optionally, a *mask* can be applied to prevent the model from attending to future positions (used in autoregressive models). The output is a *weighted sum of the values (V)*, where the weights are determined by the attention scores.

- **Multi-Head Attention :** Executes multiple parallel instances of **scaled dot-product attention** (referred to as *heads*). Each head has different learned linear projections for the **Q**, **K**, and **V** matrices. The results of all heads are *concatenated* and passed through a final *linear layer.* This mechanism allows the model to focus on different parts of the input simultaneously.

4. **Conformer: Convolution-Augmented Transformer for Speech Recognition** - Gulati et al. (2020)

Gulati et al. proposed the Conformer, which combines convolutional layers with Transformer architectures to enhance local feature extraction and global temporal modeling. The Conformer architecture was shown to significantly improve speech recognition tasks, particularly in noisy environments, by better capturing both short- and long-range dependencies in speech signals.

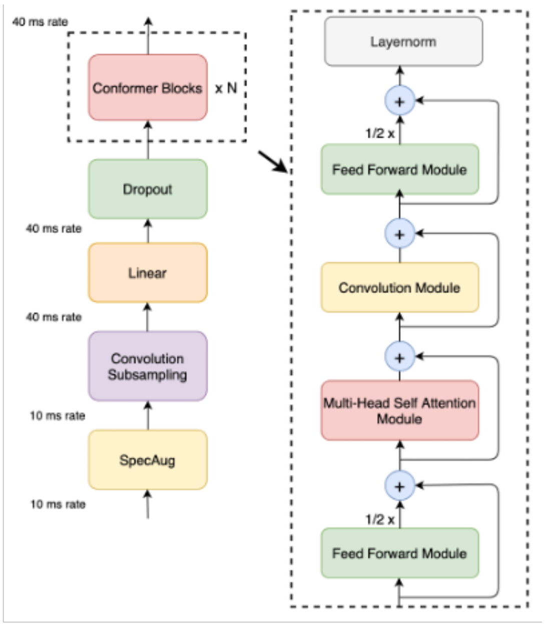


Figure 4: Convolution-Augmented Transformer

- **Conformer encoder model architecture :** Conformer comprises of two macaron-like feed forward layers with half step residual connections sandwiching the multi-headed self-attention and convolution modules. This is followed by a post layernorm.

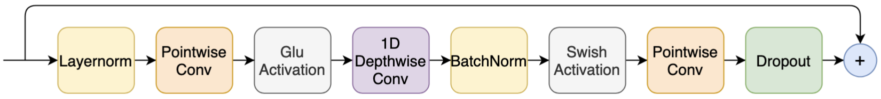


Figure 5: Convolution Module

- **Convolution module :** The convolution module contains a pointwise convolution with an expansion factor of 2 projecting the number of channels with a GLU activation layer, followed by a 1-D Depthwise convolution. The 1-D depthwise conv is followed by a Batchnorm and then a swish activation layer.

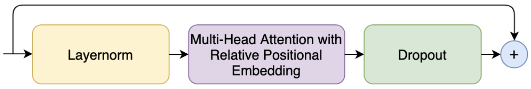


Figure 6: Multi-Headed Self-attention module

- **Self-attention module :** We use multi-headed self-attention with relative positional embedding in a pre-norm residual unit.

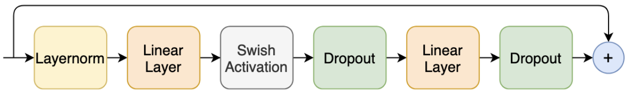


Figure 7: Feed forward module

- **Feed forward module :** The first linear layer uses an expansion factor of 4 and the second linear layer projects it back to the model dimension. We use swish activation and a pre-norm residual units in feed forward module.

5. **Lip Reading Sentences in the Wild** - Chung and Zisserman (2017)

Chung and Zisserman developed an attention-based sequence-to-sequence model for visual speech recognition "in the wild," which means the model was trained and tested on unconstrained, real-world video data. Their dataset (LRS2) became a benchmark for testing VSR models. The model used pre-extracted visual features and achieved significant improvements over earlier methods, marking a shift towards more robust VSR models.

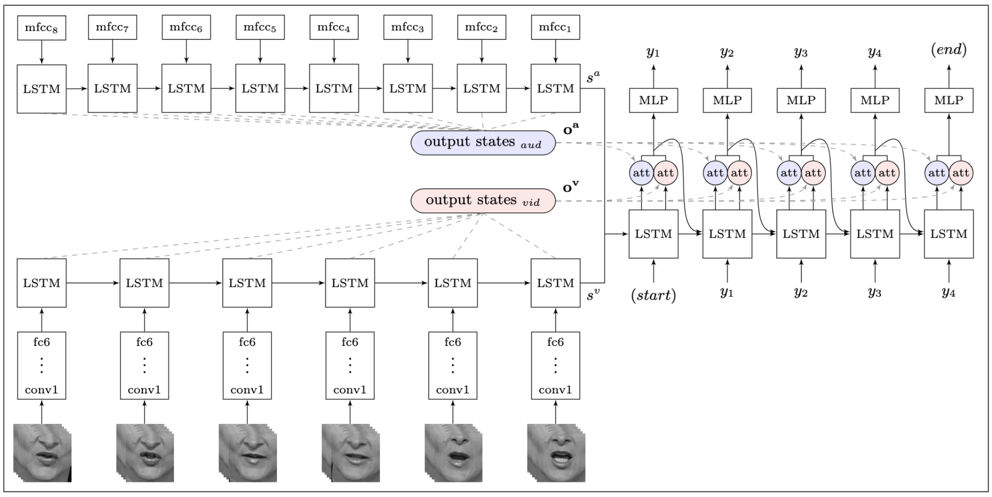


Figure 8: Lip Reading in the Wild

- **Watch, Listen, Attend and Spell architecture :** At each time step, the decoder outputs a character yi, as well as two attention vectors. The attention vectors are used to select the appropriate period of the input visual and audio sequences.

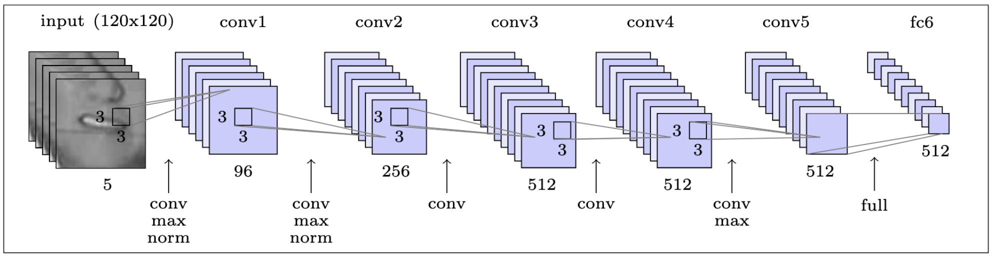


Figure 9: ConvNet

- **The ConvNet architecture :** The input is five gray level frames centered on the mouth region. The 512-dimensional fc6 vector forms the input to the LSTM.

6. **GFNet: Gated Fusion Network for Video Saliency Prediction** - Wu et al. (2023)

Wu et al. propose GFNet, a model aimed at enhancing video saliency prediction (VSP) by addressing the issue of spatiotemporal feature dilution in 3D convolutional encoder-decoder structures. The core innovation in GFNet is a gated fusion (GF) module that allows the model to selectively preserve important details, resulting in more accurate identification of salient regions in video frames.

- **Architecture :** Utilizes a fully convolutional 3D encoder-decoder architecture based on the S3D network backbone, pre-trained on an action classification dataset.

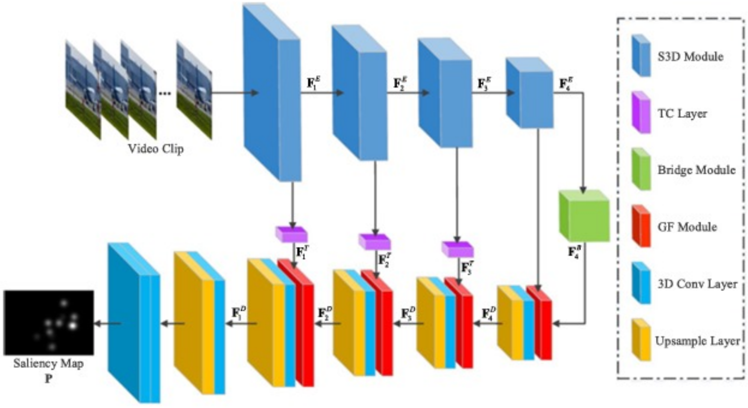


Figure 10: Overall architecture of GFNet

- **Encoder :** Includes multiple levels of feature extraction, using convolutional layers to capture spatial and temporal details.
- **Bridge Module :** Applies 3D atrous convolutions with various dilation rates to enhance high-level feature maps, broadening the receptive field and capturing multiscale global context information.

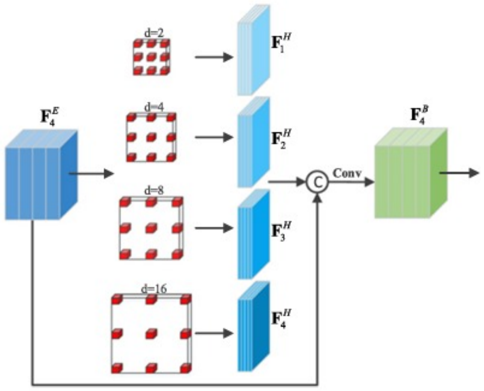
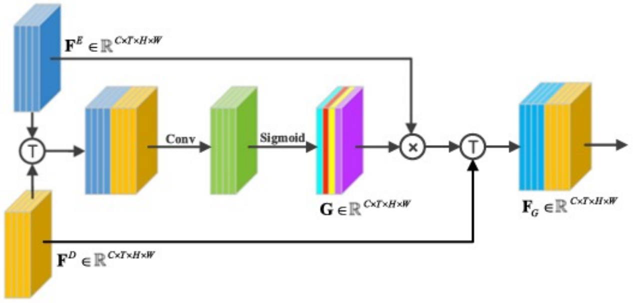


Figure 11: Structure of bridge module

- **Gated Fusion (GF) Module :** A novel module introduced between each encoder and decoder block. It selectively weights encoder features before combining them with decoder features, focusing on salient areas by controlling information flow. The GF module applies convolutional and nonlinear operations, followed by a sigmoid activation, to generate a gate that weights the encoder features in temporal, spatial, and channel dimensions.



$F^E \in \mathbb{R}^{C\times T\times H\times W}$
T
Conv
Sigmoid
×
T
$G \in \mathbb{R}^{C\times T\times H\times W}$
$F_G \in \mathbb{R}^{C\times T\times H\times W}$
$F^D \in \mathbb{R}^{C\times T\times H\times W}$

Figure 12: Structure of Gated Fusion Module

- **Decoder :** Combines fused features with 3D convolutions and upsampling to refine spatial details, ultimately producing a saliency map.

7. **Gated Multimodal Unit (GMU) for Multimodal Learning** - Arevalo et al. (2017)

Arevalo et al. introduce the Gated Multimodal Unit (GMU), a neural architecture developed to improve multimodal learning by dynamically controlling the influence of different data sources, such as text and images. The GMU employs gating mechanisms to create optimal representations from multiple modalities, achieving high performance in tasks like movie genre classification.

- **Architecture :** The GMU is an internal neural network component for fusing different modalities (e.g., text and images), inspired by gating mechanisms in recurrent neural networks like LSTMs and GRUs.

- **Gated Multimodal Unit (GMU) :** For each modality (e.g., text and visual data), a separate gate is applied. This gate learns to control the contribution of each modality to the unit's output based on relevance to the specific task.

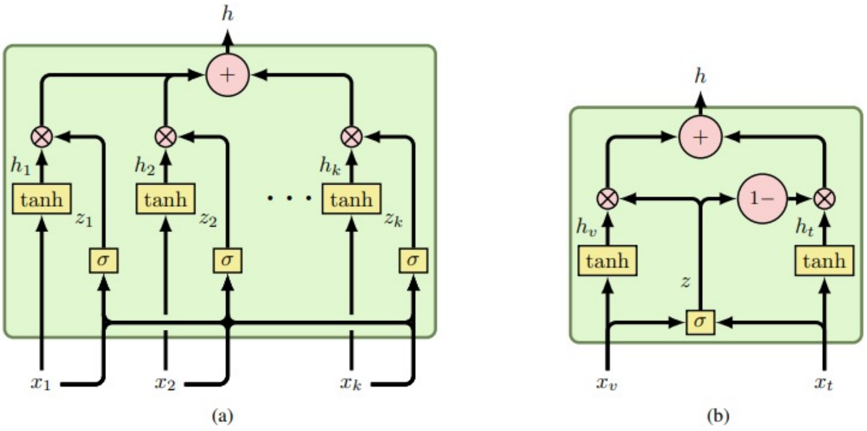


Figure 13: Illustration of gated units. a) The proposed model to use with more than two modalities. b) A simplification for the bimodal approach.

- **Fusion Mechanism :** Each input modality feeds a neuron with a tanh activation function, encoding its feature representation. The GMU then uses sigmoid-activated gates to weight and combine these features, adjusting the impact of each modality depending on the input data to optimize the final fused representation for tasks like genre classification.
- **Classifier :** After the GMU combines features, the final fused output is passed through a neural network classifier, which uses fully connected layers to perform tasks like multilabel classification. The architecture also incorporates maxout activations for the fully connected layers, improving the network's ability to approximate complex functions.

5 Outcomes

- We tested our model on a sample video with excellent lighting conditions, ensuring clear visibility of the region around the mouth. This setup allowed us to assess the model's accuracy under optimal visual conditions. Given the modularity of our model, we evaluated performance across different modality configurations (audio-only, video-only, and audio-visual) to analyze the effect of each modality independently and in combination.
- **Audio-Only Inference:** With only the audio modality, the model outputted the transcript: 'COMPLETELY UNCONSTRAINED ENVIRONMENTS WHERE WE HAVE LARGE CHANGES IN CATHOLES AND'. While close to the actual text, minor errors in word recognition (e.g., "CATHOLES" instead of "HEADPHONES") suggest some limitation in accurately distinguishing phonetically similar sounds in isolated audio processing.
- **Video-Only Inference:** With only the video modality, the model outputted the transcript: 'COMPLETELY CONCENTRATED ENVIRONMENTS WHERE WE

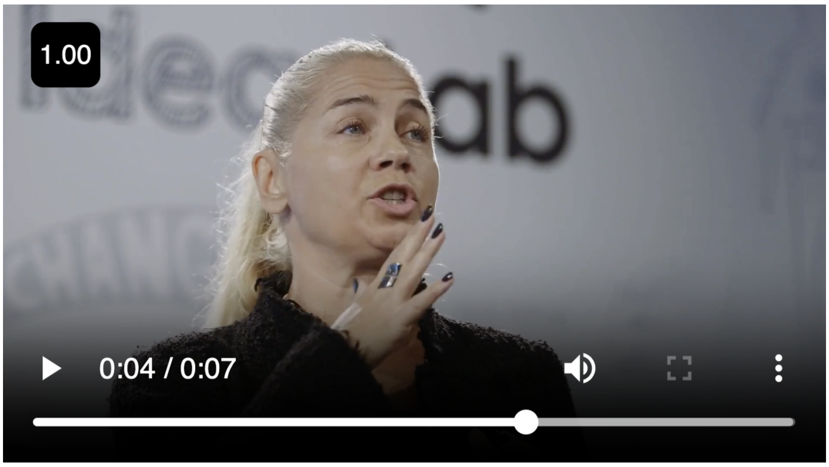


Figure 14: Sample video for inference. **Actual transcript:** ’COMPLETELY UNCONSTRAINED ENVIRONMENTS WHERE WE HAVE LARGE CHANGES IN HEADPHONES AND’

HAVE LARGE CHANGES IN GET POSTS AND’. Errors here, such as ”CONCENTRATED” and ”GET POSTS,” indicate challenges in interpreting visual cues alone, especially for words with subtle mouth movements or homophenes (visually similar lip movements for different sounds).

- **Audio-Visual Inference:** With both audio and video modalities, the model produced the transcript: ’COMPLETELY UNCONSTRAINED ENVIRONMENTS WHERE WE HAVE LARGE CHANGES IN GET POLES’. The combined modality reduced some errors seen in singular modalities (e.g., ”UNCONSTRAINED” was correctly inferred). However, minor inaccuracies remain in the final phrase (”GET POLES” instead of ”HEADPHONES”), suggesting that even with multimodal data, some complex or phonetically challenging words require further refinement.

6 Summary

- **End-to-End Processing:** The papers we reviewed explore models that process both audio and visual inputs either separately or together. They work directly with raw data, like audio waveforms and video frames, or with pre-extracted features such as spectrograms and visual embeddings. This approach improves the model’s ability to handle noisy environments where audio alone might not be reliable.

- **Hybrid Model Architectures:** A common theme across the models is the combination of CNNs, RNNs, and Transformers. These models use convolutional layers to capture local features, while Transformers help in understanding long-term dependencies and patterns in speech across time.

- **Advances in Attention Mechanisms:** Attention mechanisms, especially multihead attention, are a game-changer. They help the models focus on the most relevant parts of the audio and visual inputs, significantly boosting the accuracy of speech recognition.

- **Conformer Architecture:** A major innovation is the Conformer architecture, which blends convolutional layers with Transformer elements. This combination helps the model learn both local and global patterns, leading to much better performance, especially in speech recognition tasks.

- **Datasets and Real-World Applications:** The use of large, real-world datasets like LRS2 and LRS3 shows that these models are designed to handle real-world challenges. They deal with varying head poses, lighting conditions, and speaking styles, which is crucial for achieving state-of-the-art results in Audio-Visual Speech Recognition (AVSR).

7 References

- **Main paper:** End-to-end Audio-visual Speech Recognition with Conformers
- **Related works papers:**
  - LipNet: End-to-End Sentence-Level Lipreading
  - Deep Audio-Visual Speech Recognition
  - Attention Is All You Need
  - Conformer: Convolution-augmented Transformer for Speech Recognition
  - Lip Reading Sentences in the Wild
  - GFNet: Gated Fusion Network for Video Saliency Prediction
  - Gated Multimodal Unit (GMU) for Multimodal Learning