**Visual Speech Recognition: A Deep Learning-Based Lip Reader**

***Nijat Zakarya***

***Azerbaijan Technical University, Baku, Azerbaijan***

[***zekeryenicat@gmail.com***](mailto:zekeryenicat@gmail.com)

**Key words:** CNNs, Bidirectional Long Short-Term Memory (BiLSTM), Loss function, CTC, Sequence-to-Sequence Modeling

**Problem Summary:**   
This project focuses on building a lip-reading system capable of transcribing spoken words from silent video footages using deep learning techniques. Traditional Automatic Speech Recognition (ASR) systems struggle in noisy environments or when audio data is missing. This system seeks to overcome those limitations by relying solely on visual input — specifically, the motion of lips — thereby enabling accessibility tools for the hearing impaired and robust transcription in audio-compromised settings.

**Technologies, Tools & Methodology:**

* **Programming Language:** Python
* **Frameworks:** TensorFlow, Keras
* **Libraries:** OpenCV (for image and video preprocessing), Matplotlib, ImageIO
* **Approach:**
  + Preprocessed videos of spoken words into sequences of grayscale mouth-region frames.
  + Used a Convolutional Neural Network (CNN) for spatial feature extraction and Bidirectional LSTM layers to model temporal dependencies.
  + Employed the **Connectionist Temporal Classification (CTC)** loss function to align input video frame sequences with output text sequences, inspired by the architecture proposed in *LipNet* (Assael et al., 2016).
  + Trained the model using datasets similar in structure to LRS or LJSpeech.

**Results & Novelty:**   
The model demonstrated the ability to predict character sequences from short silent video clips with promising accuracy. The novelty of the project lies in adapting an audio-based CTC-ASR model to function purely on visual input, a cross-domain application that required thoughtful preprocessing and architecture adaptation. Unlike typical ASR, this system does not require any audio, making it ideal for use cases such as noisy industrial environments, surveillance footage analysis, assistive communication tools for the deaf or camera investigations for police.