Event-Camera Visual Speech Recognition: DVS-Based Lip Reading with Deep Learning

Thenahandi De Silva, Dr Andrew Abel , Dr Laibing Jia
Department of Computer and Information Sciences , University of Strathclyde
thenahandi.de-silva@strath.ac.uk, andrew.abel@strath.ac.uk , l.jia@strath.ac.uk

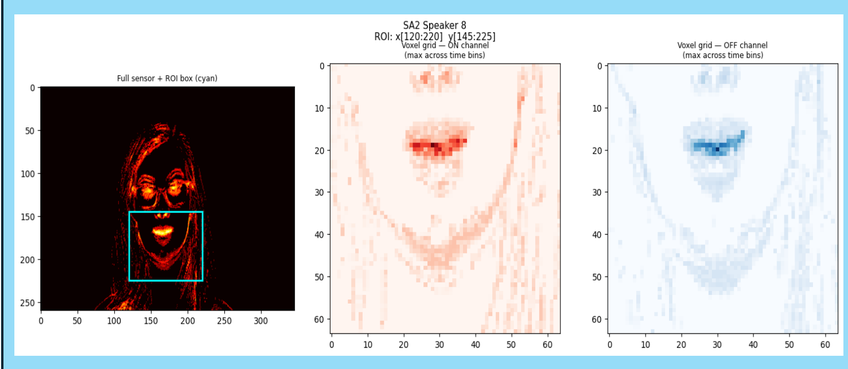
Dynamic Vision Sensor (Event Camera)

- Neuromorphic camera that mimics the function of the human retina.
- Doesn't capture full frames.
- Independent and asynchronous processing of each pixel,
- event encoded as a 4-tuple: (t,x,y,p).
  - T = timestamp (microseconds)
  - x and y are pixel coordinates,
  - p is the polarity,
- creating a sparse continuous stream of events

Visual Speech Recognition

- The general name of lip reading, is VSR.
- It works using no acoustic signal at all, but only visual information about lip movements.
- VSR is most useful in:
  - Environments with high noise levels
  - Silent environments
  - Hearing impairment communication
  - Private interfaces

Lip Region of Interest



- Fixed rectangular ROI: x = [120, 220), y = [145, 225)
- Spatial extent: 100x80 pixels
- Hand picked based on event density plots
- Encapsulates lip region for all 62 speakers
- Removes 70-85% of background events

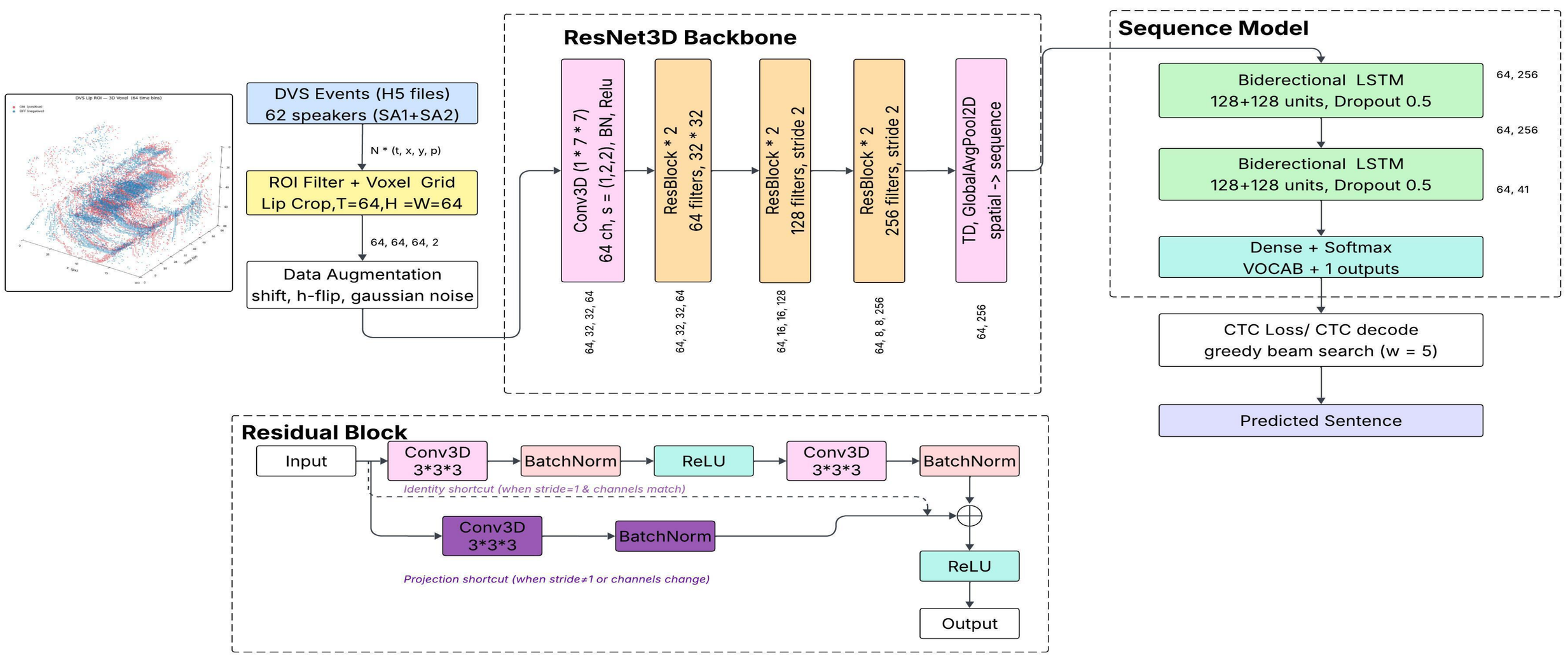


Fig. 1. Proposed DVS-LipNet architecture for event-based visual speech recognition, illustrating the end-to-end pipeline from raw DVS event streams

Proposed Framework

- Hybrid ResNet3D + BiLSTM framework for DVS-based Visual Speech Recognition
- Dataset: TCD-TIMIT – 62 speakers, sentences SA1 & SA2
- DVS Simulation: RGB videos transferred into asynchronous events using v2e simulator
- ROI Filtering: static lip crop, with 70-85% background events filtered
- Voxel Encoding: Dual-polarity spatio-temporal grids (T=64, H=64, W=64, 2)
- Feature Extraction: ResNet3D backbone, comprises 3 residual block groups
- Sequence Modelling: two stacked BiLSTM layers (128+128 units)
- Training:CTC loss (no frame-level alignment needed)
- Inference: Beam search decoding (width 5) used; WER and CER evaluation measures used

Advantages of Proposed Framework

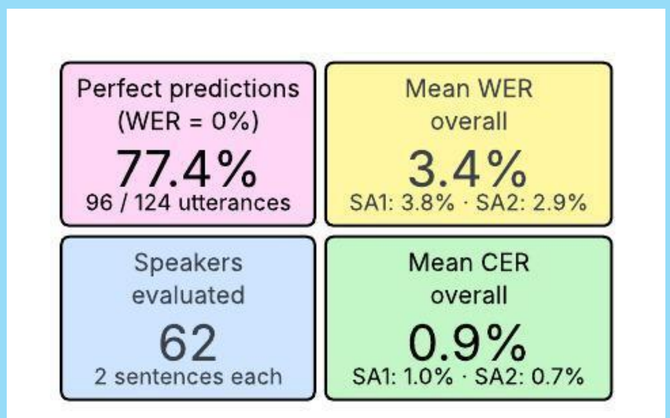
Goal

- To recognize visual speech from DVS event streams in the absence of an audio stream.
- To provide the first benchmark of sentence-level lip reading on the TCD-TIMIT corpus using neuromorphic event streams.

Benefits

- Microsecond temporal resolution- This enables the precise capture of fast lip movements and co-articulation which can be easily missed with conventional frame cameras.
- Motion-sensitive encoding- By using a dual-polarity (ON/OFF) voxel representation, we only retain movement data and disregard up to 85% of spurious data.
- No frame-level synchronization necessary- Training is done on sequence-level transcripts using CTC loss.
- Effective modelling of temporal sequences - Using a BiLSTM allows us to model long range temporal co-articulation dependencies

Results



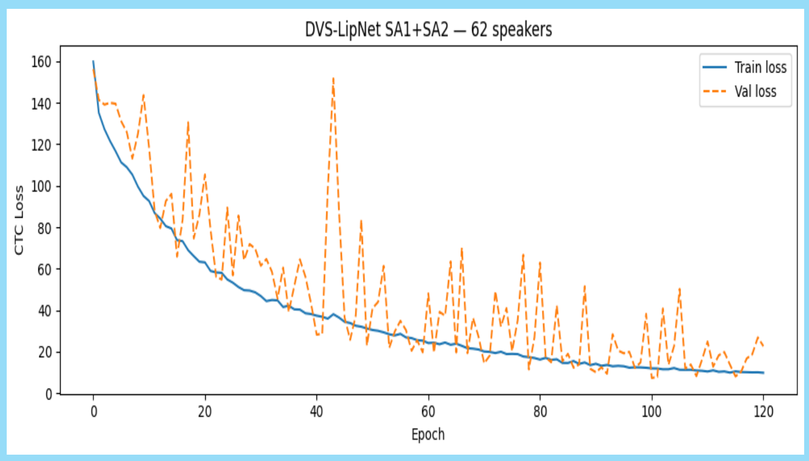


Fig. 2. CTC training and validation loss curves for the proposed DVS-LipNet framework over 120 epochs across 62 speakers (SA1 and SA2).