**My lips are concealed:**

**Audio-visual Speech Enhancement through obstructions**

Triantafyllos Afouras\(^1\), Joon Son Chung\(^1,2\), Andrew Zisserman\(^1\)

\(^1\)Visual Geometry Group, University of Oxford; \(^2\)Naver Corporation

---

**Motivation**

The task is isolating voices in videos of multiple simultaneous speakers. Several audio-visual methods that solve the problem effectively by conditioning on lip motion have been recently proposed [1,2,3].

- Models that fail when the lip region is occluded, e.g. by a microphone.

The input audio is often filtered out, resulting in silent output over the occluded frames.

**Synthetic training data**

- Generate thousands of synthetic noisy videos for training:
  - Mix interference voices into target audio to generate noisy audio.
  - Add artificial occlusions to speaker video in the form of random patches.
  - Datasets used for training: MV-LRS, LRS2-BBC, LRS3-TED

**Proposed solution**

- Train with artificial visual occlusions.
- Condition on both the speaker’s lip movements and a representation of their voice.
- Such models may fail when the lip region is occluded, e.g. by a microphone.

Add artificial occlusions to speaker video in the form of random patches.

**Proposed models**

- If not trained with occlusions, V-BLSTM performs bad for high occlusion levels.
- Proposed models outperform baseline when half or more of the visual input is occluded.
- If 20% or more of the frames are clean, self-enrollment performs better than reference enrollment.
- Both VoiceFilter and the proposed model perform well using enrollment signals from sources different from the target videos, although they have not been trained in this setting.

---

**Architecture**

- **Inputs:** noisy audio, video of lip region & optionally enrolment audio; **Output:** enhanced audio.
- **Model:** combination of audio-visual enhancement network [1] and VoiceFilter [4].
- **Speaker voice embeddings** are extracted with a network pre-trained for speaker verification.
- **The network predicts a soft-mask that filters the noisy magnitude STFT.**
- **The noisy phase is enhanced by a separate phase subnetwork [4].**

---

**Results**

- Enhance performance for varying levels of visual occlusion. 3 simultaneous speakers.
- Model Notations: VoiceFilter [2]; PIT: Permutation Invariant loss (audio-only). V-BLSTM: Baseline model that does not condition on speaker embeddings; T.Occ: Trained with occlusions; VS_Ref: Proposed model with provided enrollment signal; VS_Self: Proposed model with self-enrollment.

**Evaluation Protocol**

- **Reference sample**
  - Target Audio
  - Reference Audio
  - Loss
  - Training Occlusions
  - Speaker/Video
  - Synthetic Noisy Audio

- **Self-Enrollment:** If no enrollment audio is provided, the network is run twice:
  - First, a video-only pass produces preliminary enhanced magnitudes;
  - Those can then be used as input to the speaker embedding sub-network.
  - The network is run a second time, conditioning on both video and voice embeddings.

---

**References**

2. Ephrat et al. 2018, "Looking to Listen in the Cocktail party: A speaker-independent audio-visual model for speech separation."