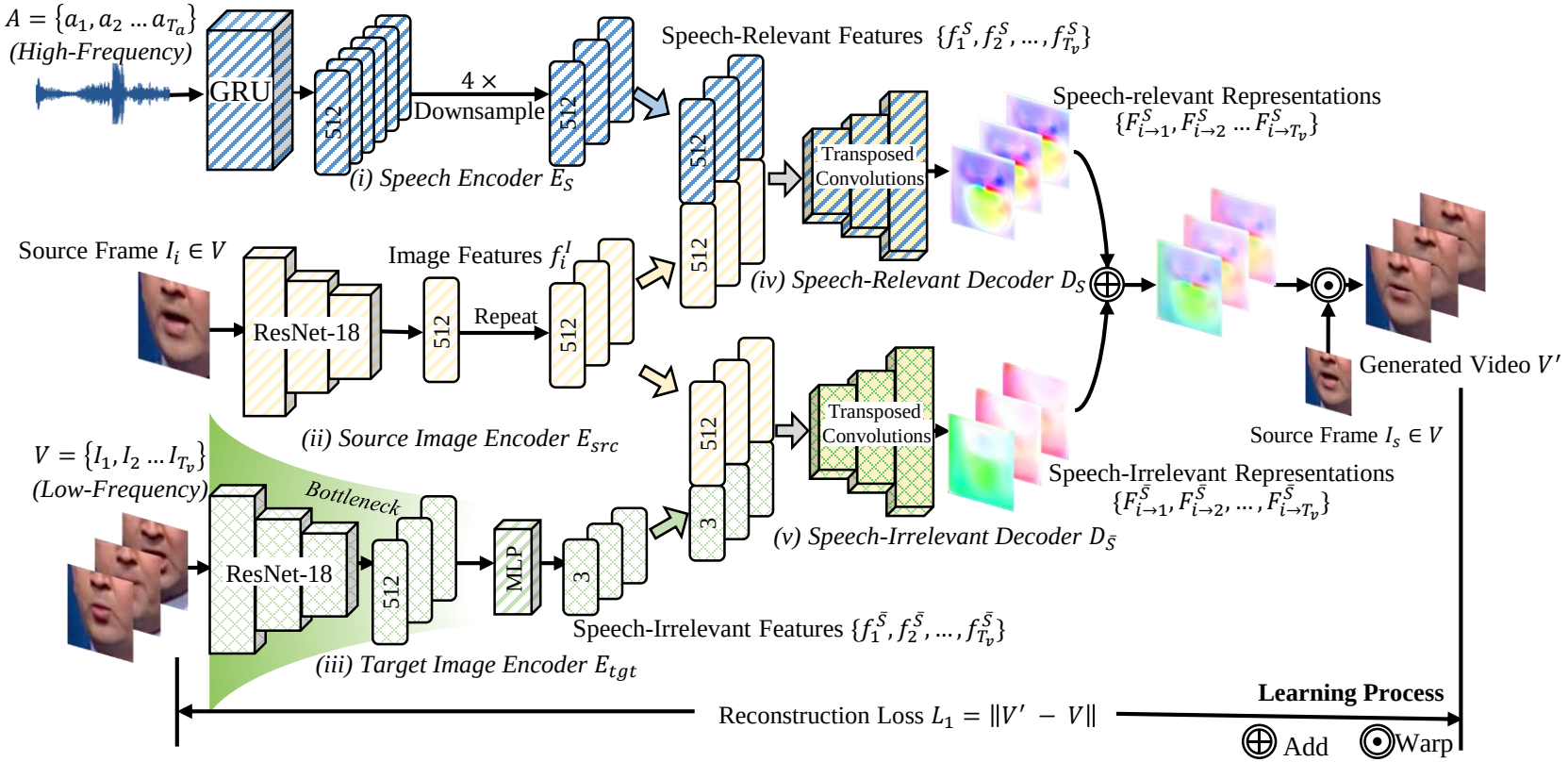
Audio-Guided Self-supervised Learning for Disentangled Visual Speech Representations

Dalu FENG, Shuang YANG, Shiguang SHAN, Xilin CHEN

Frontiers of Computer Science, DOI: 10.1007/s11704-024-3787-8

Problems & Ideas

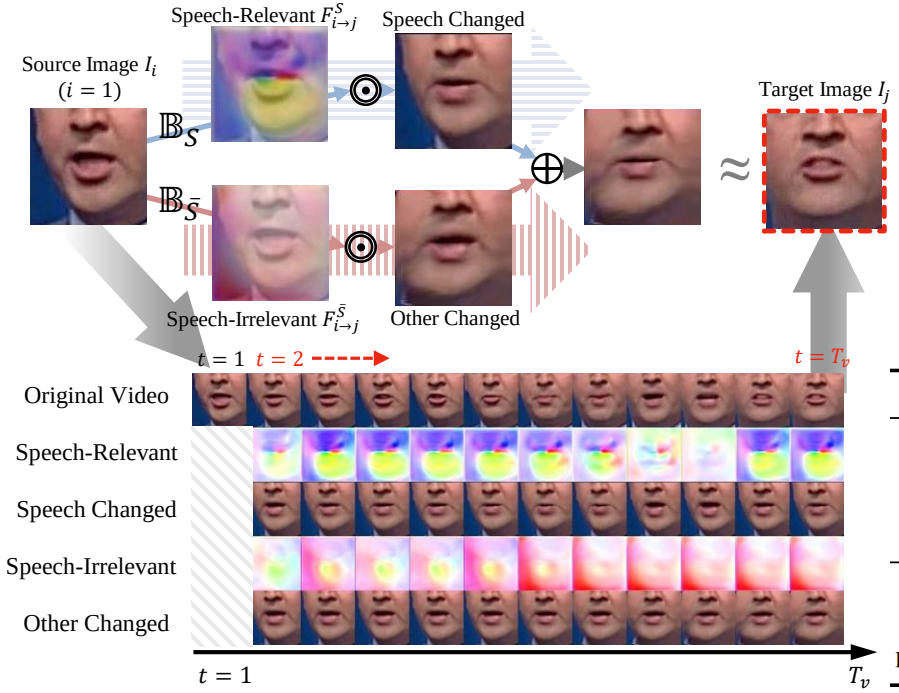
- Learning discriminative visual speech representations
  - Important for speech-related tasks (e.g. lip reading, speech recognition, etc.)
  - The key difficulties lies in addressing speech-irrelevant factors
- Ideas: Disentangling speech-relevant and speech-irrelevant facial movements by their frequency variations



Introducing the high-frequency audio signal to learn the speech-relevant cues and an information bottleneck to restrict the capacity from acquiring high-frequency and fine-grained speech-relevant information.

Main Contributions

- Contributions
  - A novel mechanism to learn disentangled visual speech representations by leveraging the intrinsic properties of the frequency variations between the speech-relevant and speech-irrelevant facial movements
  - A new method to disentangle the speech-relevant and speech-irrelevant representations by introducing the guidance of high-frequency audio signals and a special information bottleneck respectively
  - Empirical studies show the model's capability of learning disentangled visual speech representations and its potential for downstream speech-related tasks



Results of our method when taking a fixed frame as the source.

Method	Accuracy (%)↑
Ma et al. [3] (Ensemble)	88.6
Ma et al. [4]	88.4
Koumparoulis et al. [5]	89.5
Ours Baseline $\mathcal{M}_{original}$	84.5
Speech-relevant Deformation Flow F'^S_{adj} learned w/o $\mathbb{B}_S$ ($\mathcal{M}'_{F^S_{adj}}$)	89.5
Speech-relevant Deformation Flow F^S_{adj} learned w $\mathbb{B}_S$ ($\mathcal{M}_{F^S_{adj}}$)	91.4
Flow-distilled VSR Model ($\mathcal{M}'_{original}$)	85.5

Comparison with the state-of-the-art methods on LRW.

Method	Training Data	Total Hours (h)	WER (%) ↓
Ma et al. [7]	LRW, LRS2, LRS3	804	49.2
Ma et al. [8]	LRW, LRS3, AVSpeech	1495	25.5
Ma et al. [9]	LRW, LRS2, LRS3, VoxCeleb2, AVSpeech	3448	14.6
Ma et al. [7] (Reproduce)	LRS2	195	49.9
Ours baseline $\mathcal{M}_{original}$	LRS2	195	44.8
Flow-based VSR $\mathcal{M}_{F^S_{adj}}$			22.1
Flow-distilled VSR $\mathcal{M}'_{original}$			41.8

Comparison with the state-of-the-art methods on LRS2-BBC.