IMVC 2024

LipVoicer: Generating Speech from Silent Videos Guided by Lip Reading

,Yochai Yemini, Aviv Shamsian, Lior Bracha
Sharon Gannot and Ethan Fetaya

Bar-Ilan University
אוניברסיטת בר־אילן

Problem Definition

Given a soundless video of a person talking, generate the missing speech as accurately as possible.

Requirements

- Intelligibility.
- Naturalness.
- Synchronization with lip motion.
- Alignment with the speaker’s characteristics (age, gender etc.).
- Ambiguities inherent in lip motion - several phonemes can be attributed to the same lip movement sequence.

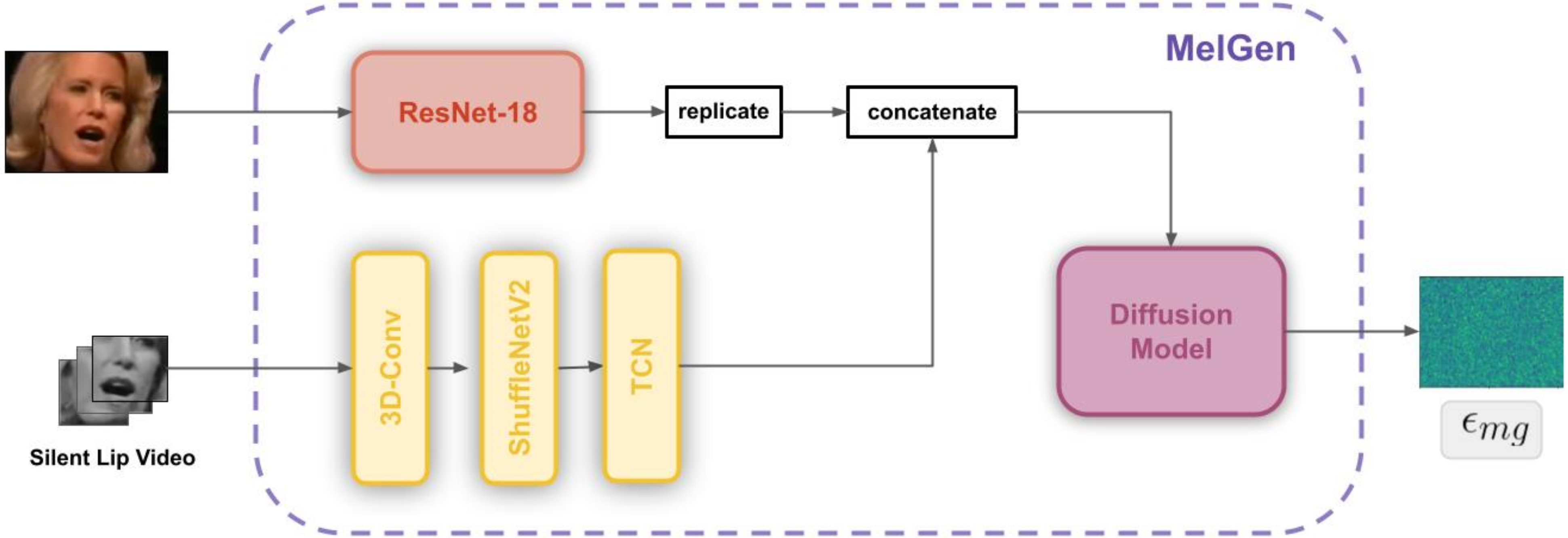
LipVoicer

Our method comprises three main components:

1. **MelGen** – a diffusion model that generates mel-spectograms from the silent video
2. A pre-trained **lip-reading network**.
3. An **Automatic speech recognition** (ASR) system

MelGen is a model that we train, the other two are used only at inference time

LipVoicer: MelGen

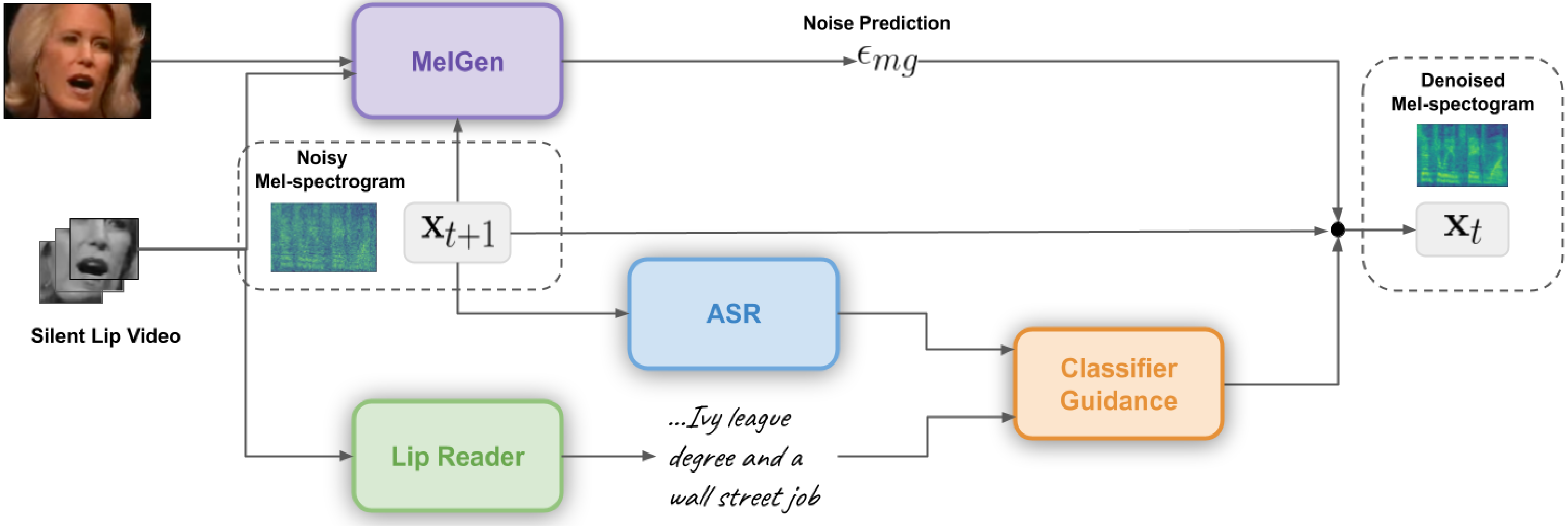


$$\epsilon_{mg}(\mathbf{x}_t, \mathcal{V}_L, \mathcal{I}, \omega_1) = (1+\omega_1)\epsilon_\theta(\mathbf{x}_t, \mathcal{V}_L, \mathcal{I}) - \omega_1\epsilon_\theta(\mathbf{x}_t, \varnothing_L, \varnothing_I)$$

The diffusion model is conditioned using classifier-free guidance

If We Just Use MelGen

LipVoicer: Full Scheme (Inference)



$$\hat{\epsilon} = \epsilon_{mg}(\mathbf{x}_t, \mathcal{V}_L, \mathcal{I}, \omega_1) - \omega_2\sqrt{1-\bar{\alpha}_t}\nabla_{\mathbf{x}_t}\log p(t_{LR}|\mathbf{x}_t)$$

Results

LipVoicer (ours)

ground-truth

SVTS

LipVoicer (ours)

ground-truth

VCA-GAN

Quantitative Results

- Evaluated on the LRS2 and LRS3 datasets
- English language
- Thousands of different speakers
- Large vocabularies

Human Listening Score (MOS)

	Intelligibility	Naturalness	Quality	Synchronization
GT	4.33 ± 0.04	4.43 ± 0.04	4.34 ± 0.04	4.39 ± 0.04
LIP2SPEECH (Kim et al., 2023)	2.07 ± 0.08	1.98 ± 0.08	1.93 ± 0.08	2.66 ± 0.10
VCA-GAN (Kim et al., 2021)	1.77 ± 0.08	1.85 ± 0.09	1.77 ± 0.08	2.34 ± 0.09
LIPVOICER (OURS)	$\mathbf{3.53} \pm \mathbf{0.07}$	$\mathbf{3.54} \pm \mathbf{0.08}$	$\mathbf{3.69} \pm \mathbf{0.08}$	$\mathbf{3.82} \pm \mathbf{0.07}$

Table 1: LRS2 Human evaluation (MOS).

	Intelligibility	Naturalness	Quality	Synchronization
GT	4.38 ± 0.03	4.45 ± 0.03	4.42 ± 0.03	4.36 ± 0.03
LIP2SPEECH (Kim et al., 2023)	2.21 ± 0.08	2.20 ± 0.09	2.01 ± 0.07	2.69 ± 0.08
SVTS (de Mira et al., 2022)	2.17 ± 0.08	2.15 ± 0.09	1.99 ± 0.07	2.71 ± 0.09
VCA-GAN (Kim et al., 2021)	2.19 ± 0.08	2.20 ± 0.09	2.08 ± 0.08	2.71 ± 0.08
LIPVOICER (OURS)	$\mathbf{3.44} \pm \mathbf{0.07}$	$\mathbf{3.52} \pm \mathbf{0.07}$	$\mathbf{3.42} \pm \mathbf{0.08}$	$\mathbf{3.56} \pm \mathbf{0.07}$

Table 2: LRS3 Human evaluation (MOS).

Objective Measures

	WER ↓	STOI-Net ↑	DNSMOS ↑	LSE-C ↑	LSE-D ↓
GT	1.5%	0.91	3.14	6.840	7.194
LIP2SPEECH	51.4%	0.70	2.37	**6.815**	**7.370**
VCA-GAN	100.7%	0.51	2.26	3.369	10.703
LIPVOICER (OURS)	**17.8%**	**0.91**	**2.89**	6.600	7.840

Table 3: Performance comparison between LipVoicer and the baselines on LRS2.

	WER ↓	STOI-Net ↑	DNSMOS ↑	LSE-C ↑	LSE-D ↓
GT	1.0%	0.93	3.30	6.880	7.638
LIP2SPEECH	57.4%	0.67	2.36	5.231	8.832
SVTS	82.4%	0.65	2.42	6.018	8.290
VCA-GAN	90.6%	0.63	2.27	5.255	8.913
LIPVOICER (OURS)	**21.4%**	**0.92**	**3.11**	**6.239**	**8.266**

Table 4: Performance comparison between LipVoicer and the baselines on LRS3.

Ablation – Lip-Reader

LR	LR WER	WER ↓	STOI-Net ↑	DNSMOS ↑	LSE-C ↑	LSE-D ↓
GT	0%	5.4%	0.92	3.10	6.257	8.220
Ma et al. (2023)	19.1%	21.4%	0.92	3.11	6.239	8.266
Ma et al. (2022)	32.3%	38.1%	0.92	3.09	6.053	8.362

Table 7: Ablation study for the choice of the lip reading accuracy, as evaluated on LRS3. LR signifies lip-reader.

Thank you

Yochai Yemini

yochai.yemini@biu.ac.il

Code is publicly available