Audio-Visual Processing in SPRING

Sharon Gannot

February 21, 2024

WP3: Robust Audio-Visual Perception of Humans

Task T3.1: Audio-visual speaker detection & tracking.

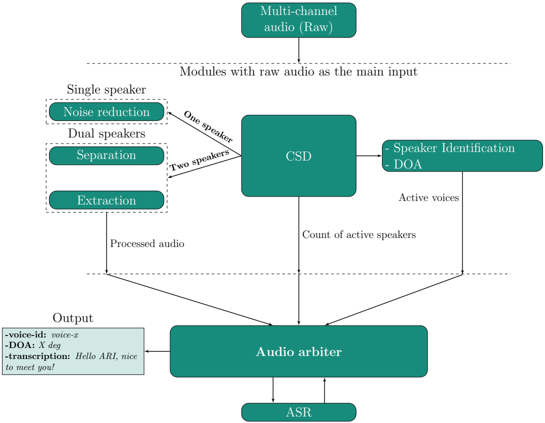
Task T3.2: Extraction of desired sources (static robot).

Task T3.3: Extraction of desired sources (moving robot).

Modules

1. Noise reduction based on mixture of deep experts (MoDE) algorithm
2. Narrowband noise reduction
3. Single microphone source separation and VAD
4. Single microphone speaker extraction and dereverberation
5. Speaker identification using voice embedding with ECAPA2
6. Classification of audio activity patterns (concurrent speaker detector)
7. Multi-person visual tracking based on FairMOT (detector + Kalman filter) inc. fish-eye camera correction
8. Audio DOA Est. (GCC-PHAT)
9. Late DOA Audio-Video fusion

Boxes...



Challenges

Today... Audio-less audio-video processing

LipVoicer:
Generating Speech from Silent Videos Guided by Lip Reading

Accepted to ICLR, 2024

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

Bar-Ilan University, Israel

Table of Contents

Background

Lip-to-Speech

- Given a soundless video of a person talking, generate the missing speech as accurately as possible.
- Such a task may occur when the speech signal is completely obfuscated due to background noises.

Challenges

Requires the generated speech to satisfy multiple criteria

- Intelligibility.
- Synchronization with lip motion.
- Naturalness.
- Alignment with the speaker's characteristics such as age, gender, accent, and more.
- Ambiguities inherent in lip motion - several phonemes can be attributed to the same lip movement sequence.

LipVoicer: Highlights

Concept

- We use a diffusion model to generate mel-spectrograms for the silent video.
- In addition to the given video, it leverages lip-reading to facilitate generation.
- A neural vocoder is utilised for generating the raw audio.

Driving Ideas

- The diffusion model captures the dynamics and characteristics of the speaker.
- The textual modality alleviates the lip motion ambiguity.

Diffusion Models

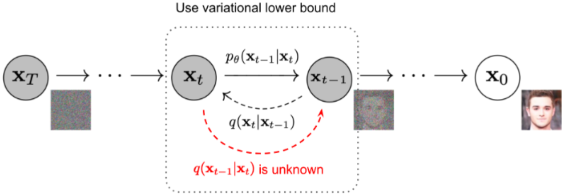


Fig. 2. The Markov chain of forward (reverse) diffusion process of generating a sample by slowly adding (removing) noise. (Image source: Ho et al. 2020 with a few additional annotations)

- Diffusion models are the reversal of a gradual noising process.
- x_0 - sample from a data distribution.
- x_t for $t \in [1, T]$ obtained by gradually adding noise, starting from x_0.
- Noise is applied so that each instance is noisier than the previous.
- x_T can be seen as a sample from a predefined noise distribution.

Forward Process

- When a Gaussian noise is applied

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$$

- $\beta_t \in [0,1]$ for $t \in [1,T]$ selected such that $x_T \sim \mathcal{N}(x_T; \mathbf{0}, \mathbf{I})$.
- According to this choice

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I})$$

- $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.
- Therefore, β_t must be chosen so that $\bar{\alpha}_T = \prod_{s=1}^{T} \alpha_s \approx 0$.

Reverse Process

- If the forward process can be reversed, we can create a true sample x_0 from Gaussian noise.
- Any intermediate step x_t can be sampled given a noise sample $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$$
- x_0 can be backtraced through
$$x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon)$$
- The reverse process is also Markovian.
- However, $q(x_{t-1}|x_t)$ is intractable.
- It can be shown that $q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0, t), \tilde{\beta}_t\mathbf{I})$ is tractable.

Reverse Process

- The reversed denoising process is parameterized with a neural network
$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1};\ \mu_\theta(x_t, t), \sigma_t^2\mathbf{I})$$
- Training this model is done by sampling a random $t \in [1, T]$ and minimizing the loss L_t
$$L_t = D_{\mathrm{KL}}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))$$
- The loss function can be simplified to

$$L_t = ||\epsilon - \epsilon_\theta(\underbrace{\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon}_{x_t}, t)||^2$$

Halfway Summary

Denoising diffusion probabilistic model (DDPM)

Algorithm 1 Training

repeat
  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
  $t \sim \mathrm{Uniform}(\{1, \ldots, T\})$
  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Take gradient descent step on
    $\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$
until converged

Algorithm 2 Sampling

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \ldots, 1$ **do**
  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
end for
return $\mathbf{x}_0$

Fig. 4. The training and sampling algorithms in DDPM (Image source: Ho et al. 2020)

Guidance

- One key feature in many diffusion models is the use of guidance for conditional generation.
- Guidance enables us to “guide” our iterative inference process to generate outputs that are more faithful to our conditioning information.
- For example, in text-to-image, it helps enforce that the generated images match the prompt text.
- Two main guidance types: with or without a classifier

Classifier Guidance

- Assume we wish to sample from $q(\mathbf{x}_t|\mathbf{c})$.
  - $\mathbf{x}_t$ - our sample at the current iteration.
  - $\mathbf{c}$ - some conditioning.
  - $p(\mathbf{c}|\mathbf{x}_t)$ - a pre-trained classifier.
- Our goal is to generate $\mathbf{x}_{t-1}$ that has the right context $\mathbf{c}$.

Bottom Line

- The diffusion model returns $\epsilon_\theta(\mathbf{x}_t, t)$.
- Classifier guidance alters the noise term that will be used for the update to

$$\hat{\epsilon} = \epsilon_\theta(\mathbf{x}_t, t) - \omega_1\sqrt{1-\bar{\alpha}_t}\nabla_{\mathbf{x}_t}\log p(\mathbf{c}|\mathbf{x}_t)$$

- ω_1 is a hyperparameter that controls the degree of guidance.

Classifier-Free Guidance

Motivation

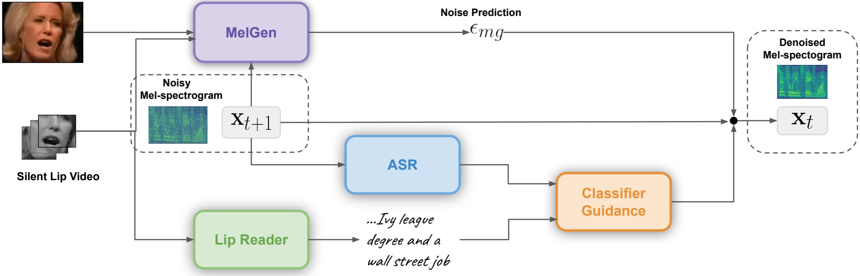
Remove the dependence on an existing classifier.

- In classifier-free guidance, we make two noise predictions
  - $\epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t)$ - with the conditioning context information.
  - $\epsilon_\theta(\mathbf{x}_t, \varnothing, t)$ - no conditioning.
- We then use $\hat{\epsilon} = \epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t) + \omega_2(\epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t) - \epsilon_\theta(\mathbf{x}_t, \varnothing, t))$.
- The hyperparameter ω_2 controls the guidance strength.

LipVoicer

Goal

Given a silent talking-face video $\mathcal{V}$, generate a mel-spectrogram that corresponds to a high likelihood underlying speech signal.



LipVoicer (Cont.)

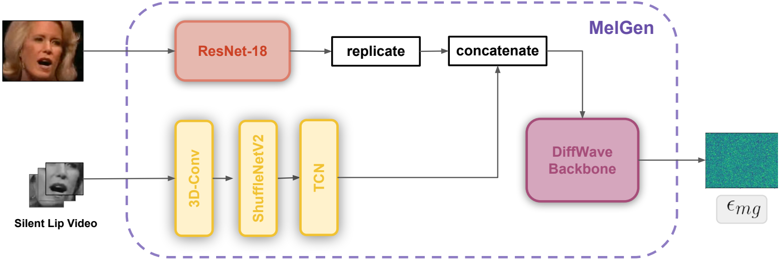
The proposed method comprises three main components

1. A mel-spectrogram generator (MelGen) which is trained to create a mel-spectrogram image from $\mathcal{V}$.
2. A pre-trained lip-reading network that predicts, at inference time, the most likely text from the silent video.
3. An Automatic speech recognition (ASR) system that anchors the mel-spectrogram recovered by MelGen to the text predicted by the lip-reader.

MelGen

- MelGen is a conditional diffusion model that we train to generate a mel-spectrogram waveform $\mathbf{x}$ conditioned on the video $\mathcal{V}$.
- We use a DiffWave residual backbone for MelGen.
- The representation of $\mathcal{V}$ should encapsulate all the needed information to generate the mel-spectrogram.
- It should also be cost-effective.

MelGen (Cont.)


MelGen
ResNet-18
replicate
concatenate
3D-Conv
ShuffleNetV2
TCN
DiffWave Backbone
Silent Lip Video
ϵ_{mg}

MelGen (Cont.)

$\mathcal{V}$ is replaced by a greyscale mouth crop region video $\mathcal{V}_L$ and a randomly chosen a single full-face image $\mathcal{I}_F$.

Feature Extraction

- For $\mathcal{I}_F$, the face embedding $\mathbf{f} \in \mathbb{R}^{D_f}$ is computed using ResNet-18.
- $\mathcal{V}_L$ is encoded using a lip-reading architecture, resulting in the lip video embedding $\mathbf{m} \in \mathbb{R}^{N \times D_m}$ (N - #frames).

A DDPM is trained to generate the mel-spectrogram conditioned on the video embedding $\mathbf{v}$ following the classifier-free mechanism

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

where ω_1 is a hyperparameter.

Text Guidance

Motivation

- The text modality can make MelGen robust to scenarios characterized by an unconstrained vocabulary.
- Syllables uttered in a silent talking-face video can be ambiguous.
- May consequently lead to an incoherent reconstructed speech.
- A pre-trained lip-reading network can be harnessed to ground the generated mel-spectrogram to the predicted text.

- One could add *text* as a global conditioning, similar to $\mathcal{I}_F$.
  - ✗ Ignores the temporal information in the text.
- One could also try to align the text and the video
  - ✗ Complicated process.

Text Guidance (Cont.)

Proposed Solution

- At inference time, we employ text guidance by harnessing the classifier guidance approach.
- Circumvents the challenge of aligning text with video content.
- Using a powerful ASR model, we can compute $\nabla_{\mathbf{x}} \log p(t_{LR}|\mathbf{x})$ needed for guidance.
- t_{LR} - the text predicted by a lip-reader.

Full Scheme

The inferred noise $\hat{\epsilon}$ used in the inference update step of the diffusion model is:

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

- Modified by both classifier guidance and classifier-free guidance.
- $\mathbf{x}_t$ - the mel-spectrogram at time step t of the diffusion inference process.
- ω_2 is a hyperparameter.
- An ASR is utilised rather than audio-video ASR, to encourage the model to focus on audio generation.

Results

Datasets

- We specifically select in-the-wild datasets, LRS2 and LRS3.
- Variations in lighting conditions, speaker characteristics, speaking styles, and speaker-camera alignment.
- LRS2
  - Videos of British English.
  - Contains roughly 142,000 training videos of
  - Amounts to 220 hours of speech by various speakers.
  - In the test set, there are 1,243 videos.
- LRS3
  - Train set: 9,000 different speakers, 151,000 videos, 430 hours of speech videos.
  - There are 1,452 videos in the test split.
  - English, but with different accents including non-native ones.

Results: Mean-Opinion-Score

	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).

Results: Objective Metrics

	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**

Video Samples

https://lipvoicer.github.io