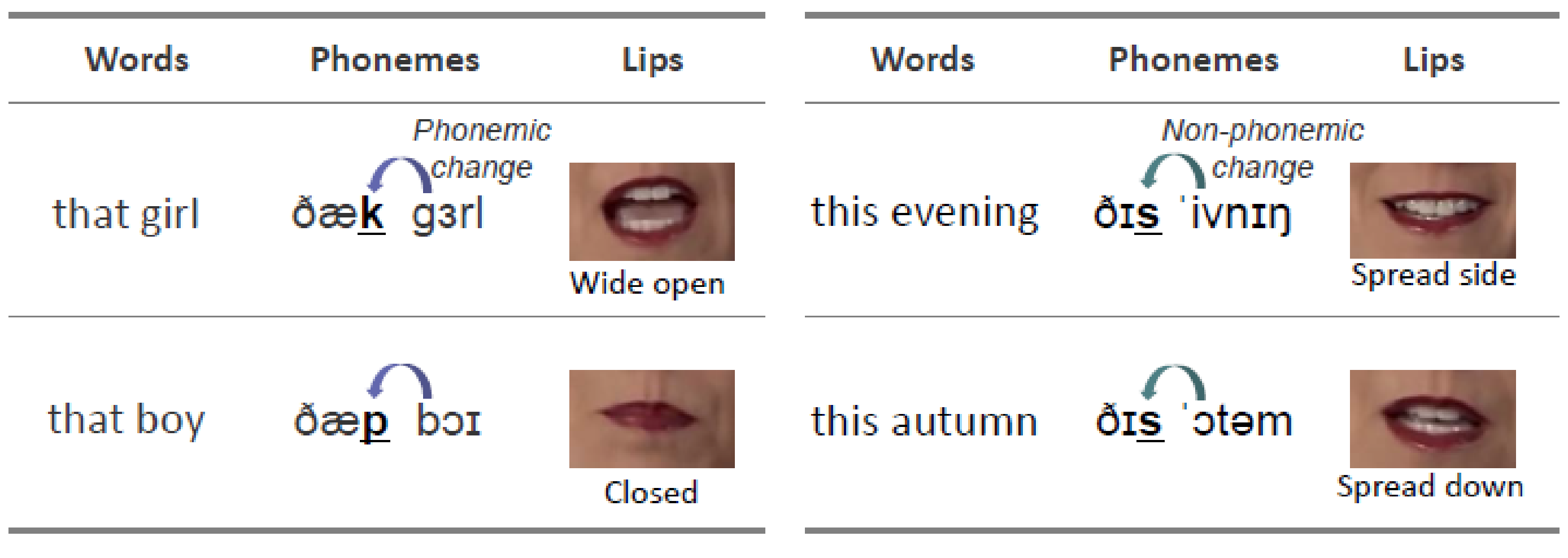
Exploring Phonetic Context-Aware Lip-Sync for Talking Face Generation

Se Jin Park Minsu Kim Jeongsoo Choi Yong Man Ro

Integrated Vision Language Lab., KAIST, South Korea

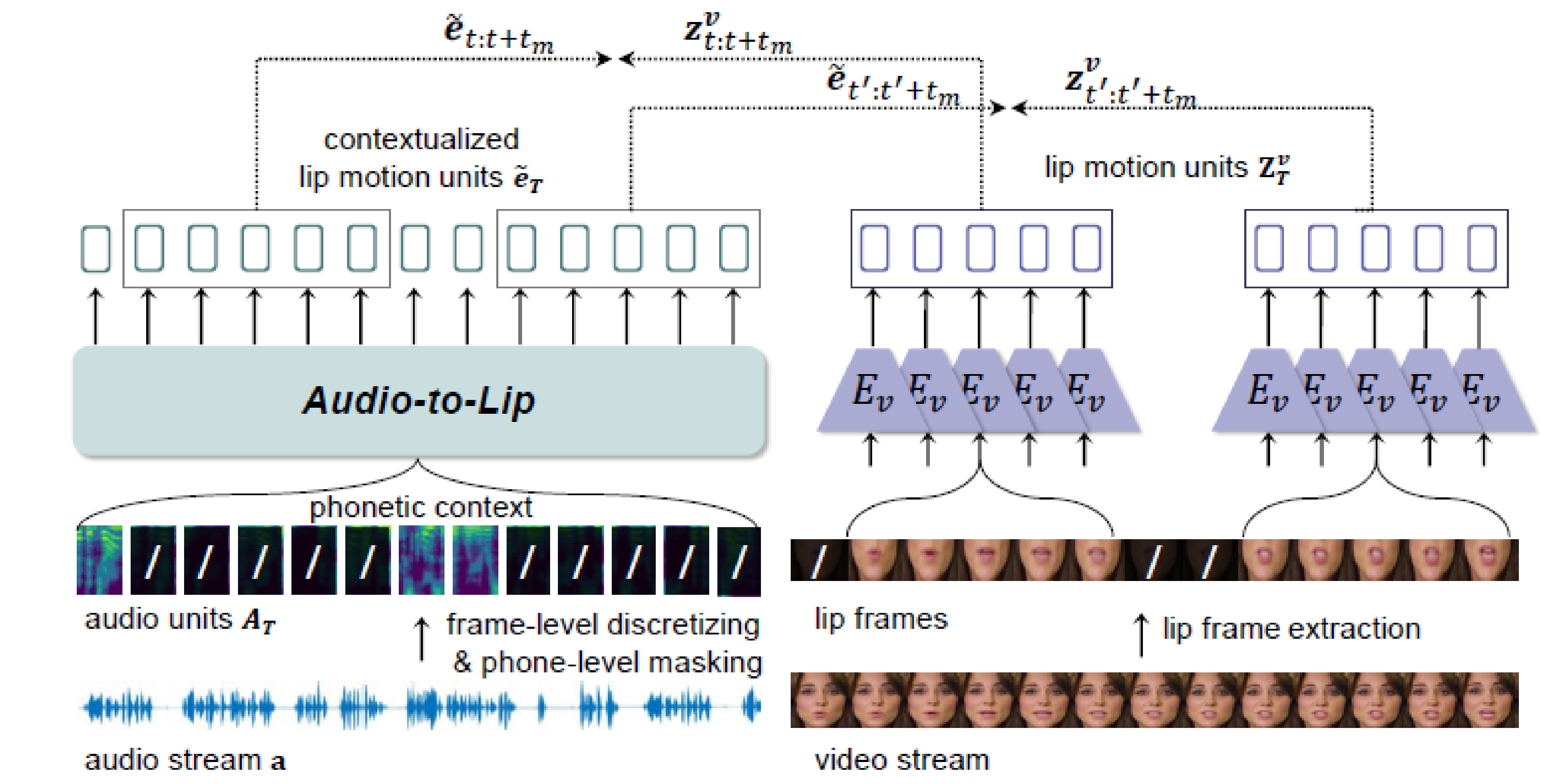
INTRODUCTION

- Talking face generation aims to generate face videoes with accurate synchronization between the lip motion and the driving audio input.
- Due to *co-articulation*, where articulation of the current speech segment changes due to the neighboring speech, the visual articulatory movements including the lips are affected by the neighboring phones (i.e. either interchanged or natural blend in).
- The paper investigates phonetic context in lip motion for talking face generation to improve spatio-temporal alignment of the lip motion.

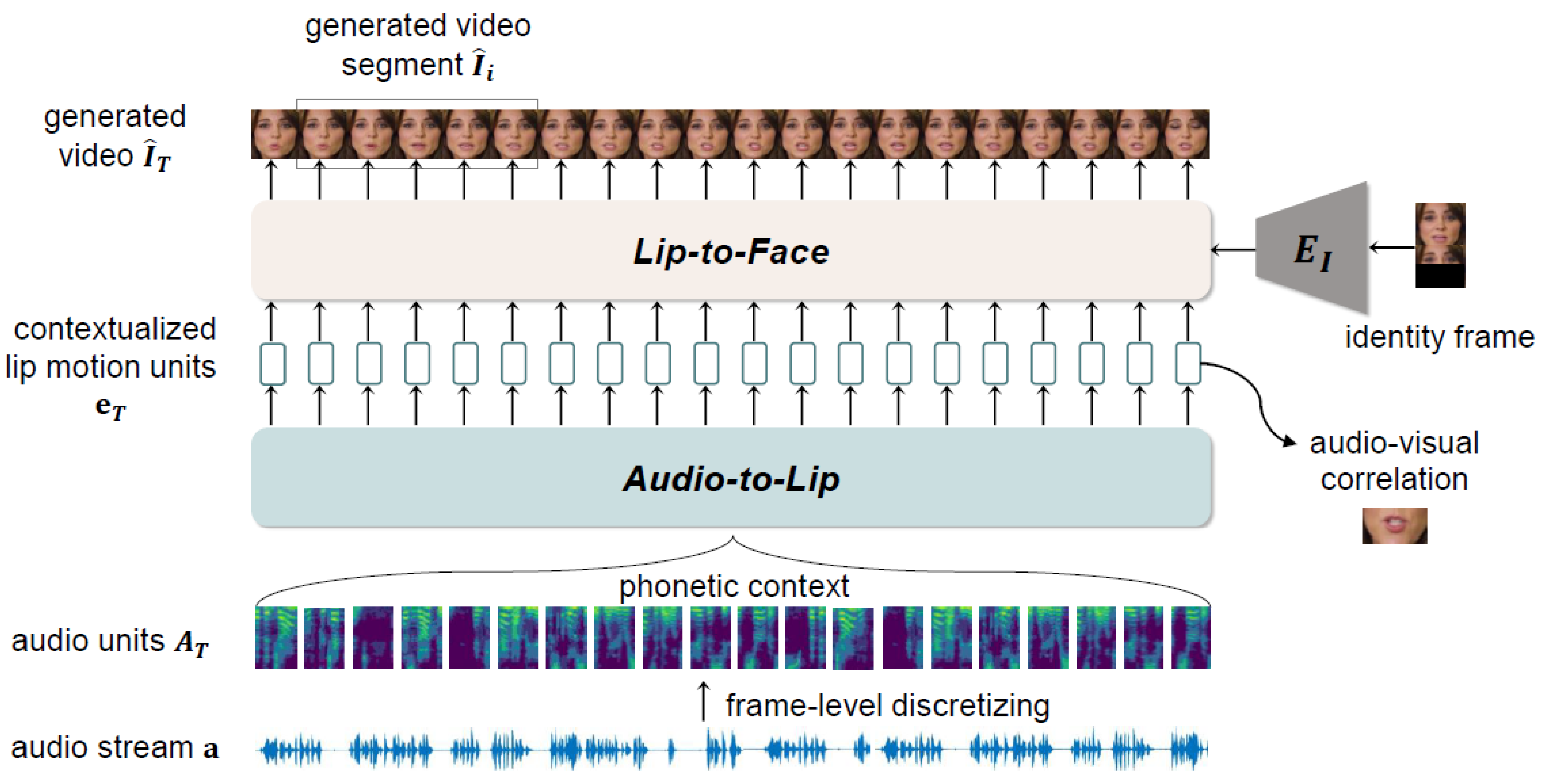


(a) that / ðæt (b) this / ðɪs

METHODOLOGY



Context-Aware Lip-Sync (CALS) Framework



Audio-to-Lip Module

- The Audio-to-Lip module, f_θ, translates audio to visual lip units, and associates phonetic context while establishing the audio-lip correlation.
- Motivated by the masked prediction in context learning, we process the audio into frame-level mel-spectrograms, $\mathbf{A}_T$, and corrupt the audio as : $\tilde{\mathbf{A}}_T = r(\mathbf{A}_T, M)$.
- The Audio-to-Lip module translates the corrupted audio units into contextualized lip motion units, $\tilde{\mathbf{e}}_T = f_\theta(\tilde{\mathbf{A}}_T)$, which is guided to predict the corresponding ground truth lip motion units, $\boldsymbol{Z}_T^v$, of the masked timestep as : $\mathcal{L}_{a2l} = \sum_{t\in M}(\mathbf{z}_t^v - \tilde{\mathbf{e}}_t)^2$.
- The ground truth lip motion units, $\boldsymbol{Z}_T^v$, are extracted from a visual encoder pretrained using contrastive learning between video frames and corresponding audio frames.

Lip-to-Face Module

- The Lip-to-Face module, f_ϕ, synthesizes face frames with the context-aware lip motion units drawn from the Audio-to-Lip module as : $\hat{\mathbf{I}}_T = f_\phi(\mathbf{e}_T, \mathbf{f}_T^I)$.
- The identity frame, $\mathbf{f}_T^I$, is a random reference frame concatenated with a pose-prior (target face with lower-half masked).
- Since the lip motion units, $\mathbf{e}_T$, have attended to every other phones in the context, the generated dynamics are more temporally stable and consistent.

Total Loss

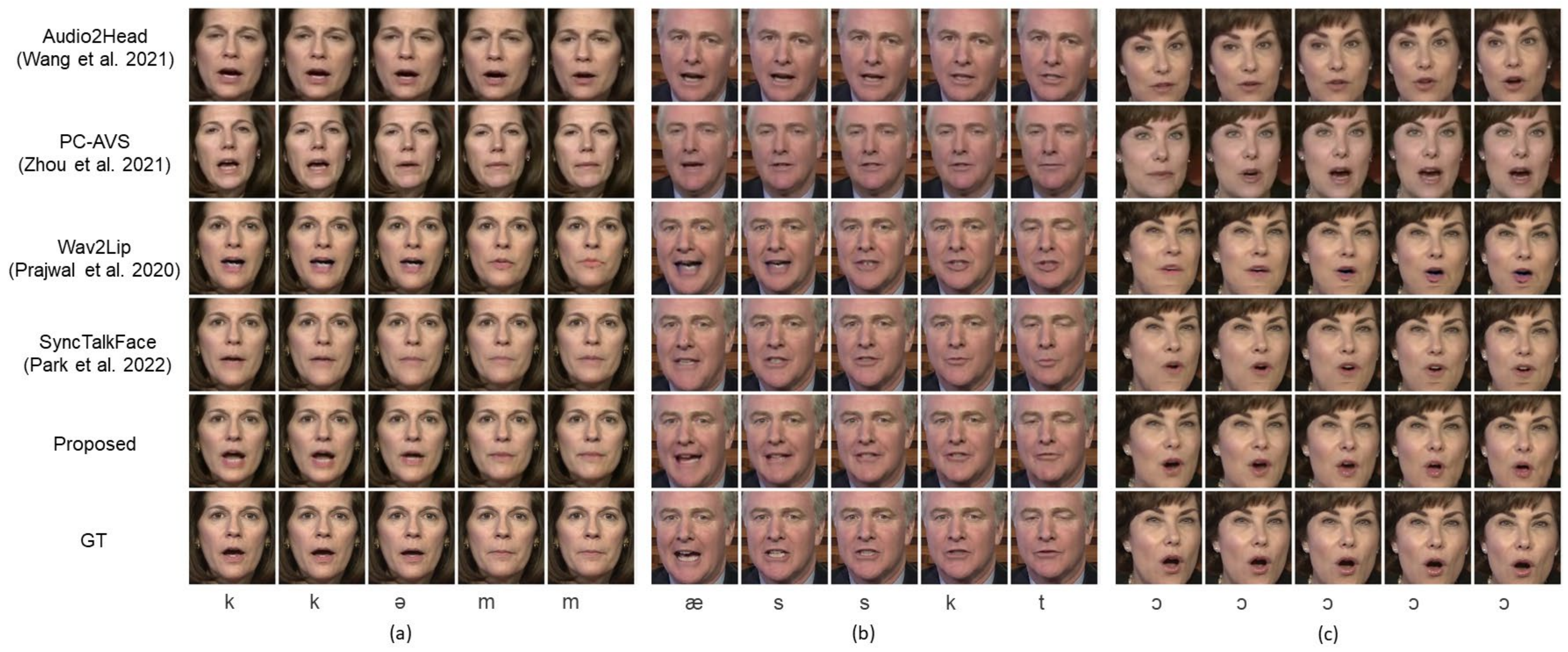
$$L = \lambda_1 L_{recon} + \lambda_2 L_{gan} + \lambda_3 L_{sync},$$

$$\mathcal{L}_{sync} = d_{av} + d_{vv}, \quad d_{vv} = -\frac{1}{S}\sum_i^S \log \frac{\exp\left(d_{cos}\left(\mathcal{F}_v(\hat{\mathbf{I}}_i), \mathcal{F}_v(\mathbf{I}_i)\right)\right)}{\sum_j^S \exp\left(d_{cos}\left(\mathcal{F}_v(\hat{\mathbf{I}}_i), \mathcal{F}_v(\mathbf{I}_j)\right)\right)}$$

EXPERIMENTS

Table 3. Quantitative comparison with state-of-the-art methods on LRW, LRS2 and HDTF.

	LRW					LRS2					HDTF				
Method	PSNR	SSIM	LMD	LSE-D	LSE-C	PSNR	SSIM	LMD	LSE-D	LSE-C	PSNR	SSIM	LMD	LSE-D	LSE-C
Ground Truth	N/A	1.000	0.000	6.968	6.876	N/A	1.000	0.000	6.259	8.247	N/A	1.000	0.000	7.508	7.128
Audio2Head [36]	28.578	0.385	2.654	8.935	3.487	28.726	0.395	2.088	8.518	5.393	29.449	0.602	2.304	7.207	7.681
PC-AVS [37]	30.257	0.746	1.989	6.502	7.438	29.736	0.688	1.590	6.560	7.770	29.864	0.709	1.950	7.758	6.588
Wav2Lip [15]	31.831	0.882	1.437	6.617	7.237	31.182	0.841	1.519	5.895	8.795	32.354	0.873	1.595	7.272	7.343
SyncTalkFace [16]	32.887	0.894	1.322	7.023	6.591	32.327	**0.881**	1.069	6.350	7.929	32.682	0.883	1.381	7.931	6.406
Proposed	**33.219**	**0.900**	**1.183**	**6.432**	**7.463**	**32.603**	0.876	**1.056**	**5.337**	**9.225**	**32.992**	**0.895**	**1.373**	**6.850**	**8.185**


Audio2Head (Wang et al. 2021)
PC-AVS (Zhou et al. 2021)
Wav2Lip (Prajwal et al. 2020)
SyncTalkFace (Park et al. 2022)
Proposed
GT
k
k
ə
m
m
æ
s
s
k
t
ɔ
ɔ
ɔ
ɔ
ɔ

(a) (b) (c)

Table 1. Ablation study on the proposed method on LRS2.

Proposed Method								
Baseline	A2L	d_{av}	d_{vv}	PSNR	SSIM	LMD	LSE-D	LSE-C
✓	✗	✗	✗	31.182	0.841	1.519	5.895	8.795
✓	✓	✗	✗	32.419	0.870	1.311	5.623	9.144
✓	✓	✓	✗	32.501	0.867	1.064	**5.204**	**9.421**
✓	✓	✓	✓	**32.603**	**0.876**	**1.056**	5.337	9.225

Table 2. Human evaluation by MOS with 95% confidence interval

Method	Visual Quality	Lip-Sync Quality	Realness
Ground Truth	4.607 ± 0.080	4.687 ± 0.071	4.627 ± 0.081
Audio2Head [36]	2.761 ± 0.111	2.721 ± 0.134	2.458 ± 0.126
PC-AVS [37]	2.567 ± 0.093	3.109 ± 0.110	2.458 ± 0.103
Wav2Lip [15]	2.975 ± 0.093	3.557 ± 0.110	3.109 ± 0.103
SyncTalkFace [16]	3.333 ± 0.102	3.761 ± 0.100	3.502 ± 0.102
Proposed	**3.761 ± 0.086**	**4.119 ± 0.088**	**3.940 ± 0.081**

Implementation Details

- We train and test on LRW, LRS2, and HDTF. Videos are processed into face crops of 128x128 in 25fps. The audio is processed into frame-level mel-spectrograms with window size of 400 and hop size of 160 in 100fps.

Results

- We analyzed the extent to which the phonetic context assists in lip synchronization, complementing the missing audio, and verified the effective window size to be approximately 1.2 s.

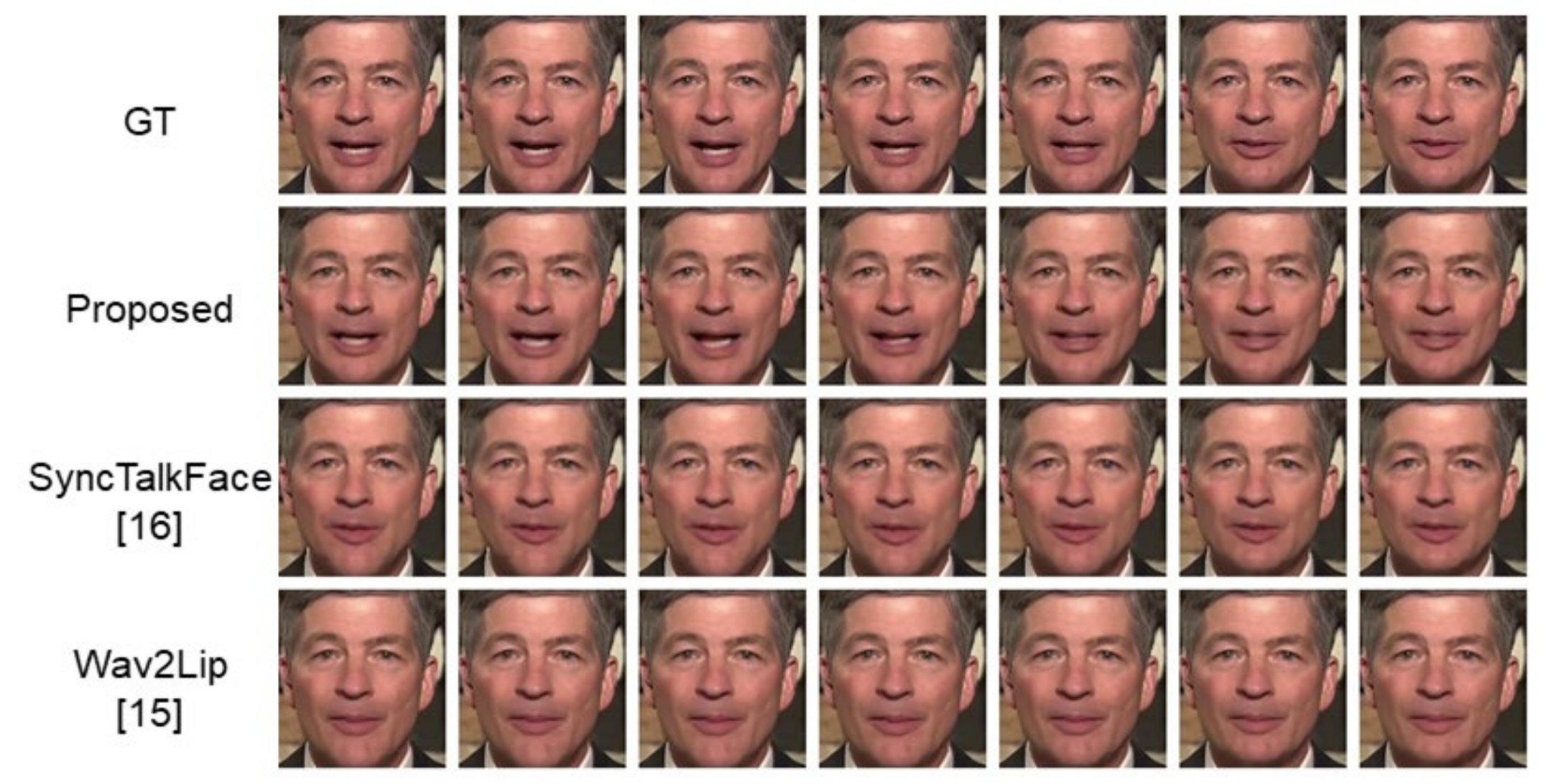


Fig. 2. Generation of frames with corresponding audio time-steps masked out. Please zoom in to see in detail.

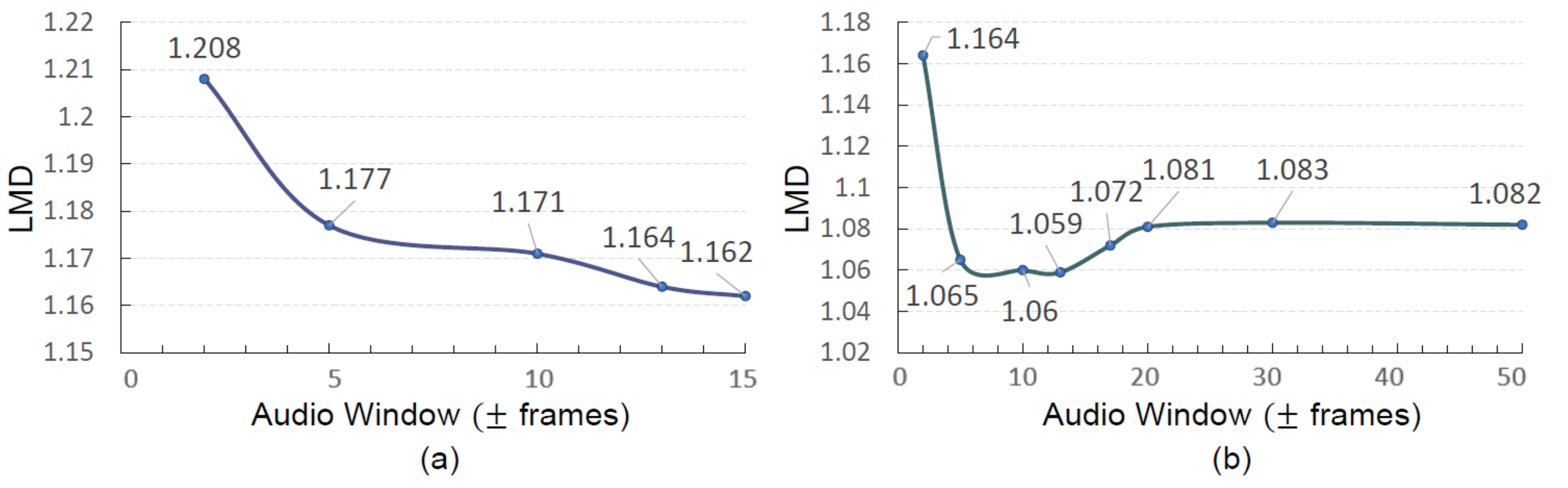


Fig. 3. LMD of the middle frame with varying audio window size on (a) LRW and (b) LRS2.