DiFlowDubber: Discrete Flow Matching for Automated Video Dubbing via Cross-Modal Alignment and Synchronization

Ngoc-Son Nguyen[1], Thanh V. T. Tran[1], Jeongsoo Choi[2], Hieu-Nghia Huynh-Nguyen[1], Truong-Son Hy[3], Van Nguyen[1]

[1] FPT Software AI Center, Vietnam, [2] KAIST, South Korea, [3] University of Alabama at Birmingham, USA

Contact me! Son Nguyen

Email: sonnn45@fpt.com

Welcome to reach out to me for further discussion and chat!

Personal Website | Linkedin

CVPR JUNE 3-7, 2026 DENVER COLORADO

arXiv | Demo Page | Code

Introduction

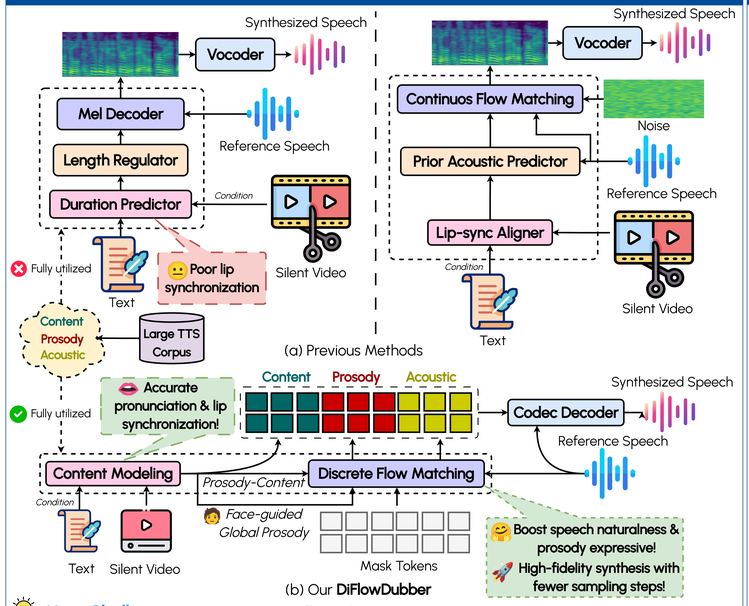


Figure 1. Motivation

Main Challenges

- **Knowledge transfer bottleneck**: Scarce dubbing data forces reliance on large-scale TTS coropa, yet conventional TTS-based methods:
  - Use integer-scaled duration predictors ⇒ Inaccurate lip synchronization
  - Weak utilization of TTS pre-training ⇒ Limited expressiveness, naturalness and pronunciation
- **Information gap**: Unified acoustic representations (e.g. mel-spectrograms) entangle speech attributes, hindering information gap between speech and silent video.

Key Contributions

- **Mitigating the information gap**: Factorized speech tokens modeling with discrete flow matching over prosody and acoustics for diversity, while content is modeled deterministically.
- **Scalable TTS knowledge transfer:** Better exploits large-scale TTS corpus characteristics (content, prosody, acoustic), reducing dubbing data dependence.
- **Accurate lip-sync & pronunciation:** A Content-Consistent Temporal Adaptation enforces precise cross-modal synchronization through a dual-alignment mechanism over text, video, and speech, while preserving content consistency with the target transcription.
- **Effective prosody guidance:** A FaPro maps facial expressions to global prosodic cues to guide expressive speech generation.

Method

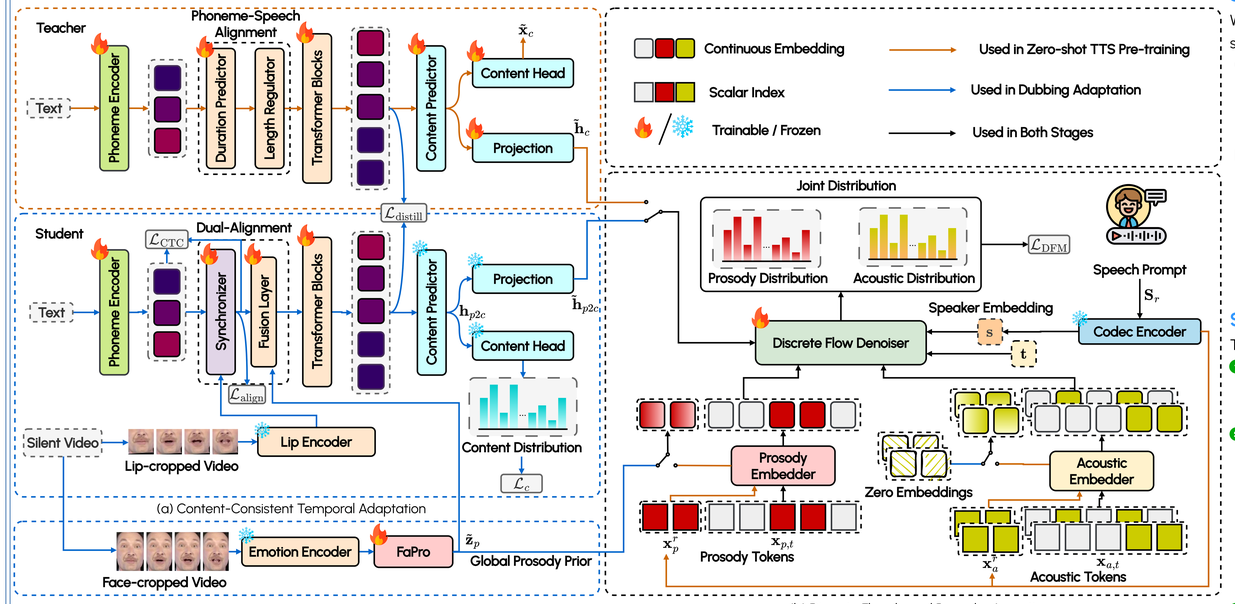


Figure 2. Detailed of DiFlowDubber Pipeline

Stage 1: Zero-shot TTS Pre-training

We use FACodec to decompose speech into three disentangled discrete token streams: prosody, content, and acoustic tokens.

Content Modeling:
- **Goal**: Linguistic structures
- **Condition**: Text
- Deterministic prediction

Prosody & Acoustic Modeling:
- **Goal**: Expressive and acoustic diversity
- **Condition**: Speaker identity, Prosody & Acoustic from reference speech, and Content latent
- Discrete Flow Matching backbone

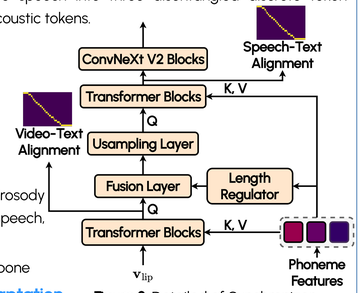


Figure 3. Detailed of Synchronizer

Stage 2: Video Dubbing Adaptation

Three modules adapt the pre-trained TTS model to the video domain:

1. Face-to-Prosody Mapper (Fapro)
   - **Goal**: Global prosody prior
   - **Condition**: Facial expressions
2. Content-Consistent Temporal Adaptation (CCTA)
   - **Goal**: Temporal lip synchronization and pronunciation consistency
   - Synchronizer Module (**Figure 3**)
     - **Goal**: Enforce precise cross-modal alignment between text, video, and speech via dual-alignment mechanism
     - **Condition**: Lip motions, text
   - Connectionist Temporal Classification (CTC) loss
   - Cosine similarity-based distillation loss
3. Prosody & Acoustic Modeling Adaptation:
   - **Goal**: Expressive and acoustic diversity
   - **Condition**: Global prosody prior (FaPro), prosody-to-content latent (CCTA)

Main Quantitative Results

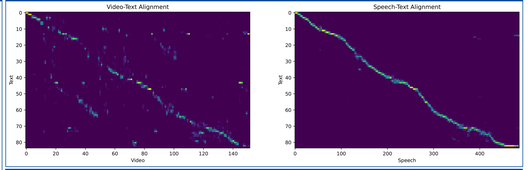
Dataset	Model	Setting 1.0 LSE-C ↑	LSE-D ↓	WER ↓	SECS ↑	UTMOS ↑	Setting 2.0 LSE-C ↑	LSE-D ↓	WER ↓	SECS ↑	MOS-S ↑	UTMOS ↑
Chem	Ground Truth	8.12	6.59	3.85	100.00	4.18	8.12	6.59	3.85	100.00	4.34 ± 0.18	4.18
	V2C-Net [1] CVPR'22	1.97	12.17	90.47	51.52	1.81	1.82	12.09	94.59	44.19	-	1.76
	HPMDubbing [2] CVPR'23	7.85	7.19	16.05	85.09	2.16	3.98	9.50	29.82	73.55	-	2.01
	StyleDubber [3] ACL'24	3.87	10.92	13.14	87.72	3.14	3.74	11.00	14.18	82.07	4.02 ± 0.24	3.04
	Speaker2Dubber [5] ACM MM'24	3.76	10.56	16.98	74.73	3.61	3.45	11.17	18.10	69.28	3.69 ± 0.25	3.64
	ProDubber [6] CVPR'25	2.58	12.54	**9.45**	72.13	3.85	2.78	12.14	11.69	50.85	3.48 ± 0.26	3.76
	EmoDubber [4] CVPR'25	8.11	6.92	11.72	**90.62**	3.82	8.09	6.96	12.81	**85.06**	4.14 ± 0.23	3.75
	DiFlowDubber (Ours)	**8.31**	**6.73**	9.65	84.59	**4.02**	**8.37**	**6.70**	**11.12**	73.93	**4.18 ± 0.18**	**4.03**
GRID	Ground Truth	7.13	6.78	22.41	100.00	3.94	7.13	6.78	22.41	100.00	4.58 ± 0.13	3.94
	V2C-Net [1] CVPR'22	5.59	9.52	47.82	80.98	2.41	5.34	9.76	49.09	71.51	-	2.40
	HPMDubbing [2] CVPR'23	5.76	9.13	45.51	85.11	2.14	5.82	9.10	44.15	71.99	-	2.11
	StyleDubber [3] ACL'24	6.12	9.03	18.88	93.79	3.73	6.09	9.08	19.58	**86.67**	3.83 ± 0.24	3.71
	Speaker2Dubber [5] ACM MM'24	5.27	9.84	17.07	**94.50**	3.69	5.19	9.93	17.42	85.76	3.57 ± 0.26	3.73
	ProDubber [6] CVPR'25	5.23	9.59	18.60	89.03	3.87	5.56	9.37	19.17	81.72	4.04 ± 0.16	3.86
	EmoDubber [4] CVPR'25	7.12	6.82	18.53	92.22	3.83	7.10	6.89	19.75	86.02	4.33 ± 0.17	3.81
	DiFlowDubber (Ours)	**7.32**	**6.73**	**16.79**	82.52	**3.95**	**7.28**	**6.78**	**16.21**	71.63	**4.38 ± 0.18**	**3.93**

Runtime Quality Analysis

Model	#Params	NFE	RTF ↓	LSE-C ↑	LSE-D ↓	WER ↓	SECS ↑	UTMOS ↑
ProDubber	200M	5	0.08	2.58	12.54	9.45	72.13	3.85
EmoDubber	116M	10	0.05	8.11	6.92	11.72	90.62	3.82
DiFlowDubber	250M	1	0.03	8.25	6.78	10.72	83.63	2.82
		2	0.04	8.27	6.74	10.99	83.34	2.79
		4	0.04	8.32	6.73	10.79	83.82	3.50
		8	0.05	8.29	6.73	10.28	84.50	3.85
		10	0.06	8.30	6.72	10.82	83.73	3.88
		16	0.08	8.27	6.75	11.02	83.83	3.94
		32	0.12	8.34	6.69	10.90	84.58	4.00
		64	0.22	8.32	6.70	10.79	84.07	4.03
		128	0.40	8.31	6.73	9.65	84.59	4.02

Ours achieves competitive performance with comparable inference speed despite a ~2× larger parameter count.

Alignment Visualization



Takeaways

- **DFM** enables high-quality speech generation with fewer sampling steps
- **Synchronizer** improves lip-sync through dual-alignment machanism.
- **Fully utilizing speech characteristics from large-scale TTS corpora** improves speech quality.