Cross-Lingual F5-TTS: Towards Language-Agnostic Voice Cloning and Speech Synthesis

Qingyu Liu, Yushen Chen, Zhikang Niu, Chunhui Wang, Yunting Yang,
Bowen Zhang, Jian Zhao, Pengcheng Zhu, Kai Yu, Xie Chen

1. Highlights

- Removes F5-TTS's dependence on audio prompt transcripts via MMS forced alignment
- Language-agnostic duration estimation from audio prompt pace
- Maintains competitive intra-lingual voice cloning performance compared with F5-TTS
- Enables ***cross-lingual cloning*** using audio prompts from ***unseen languages***

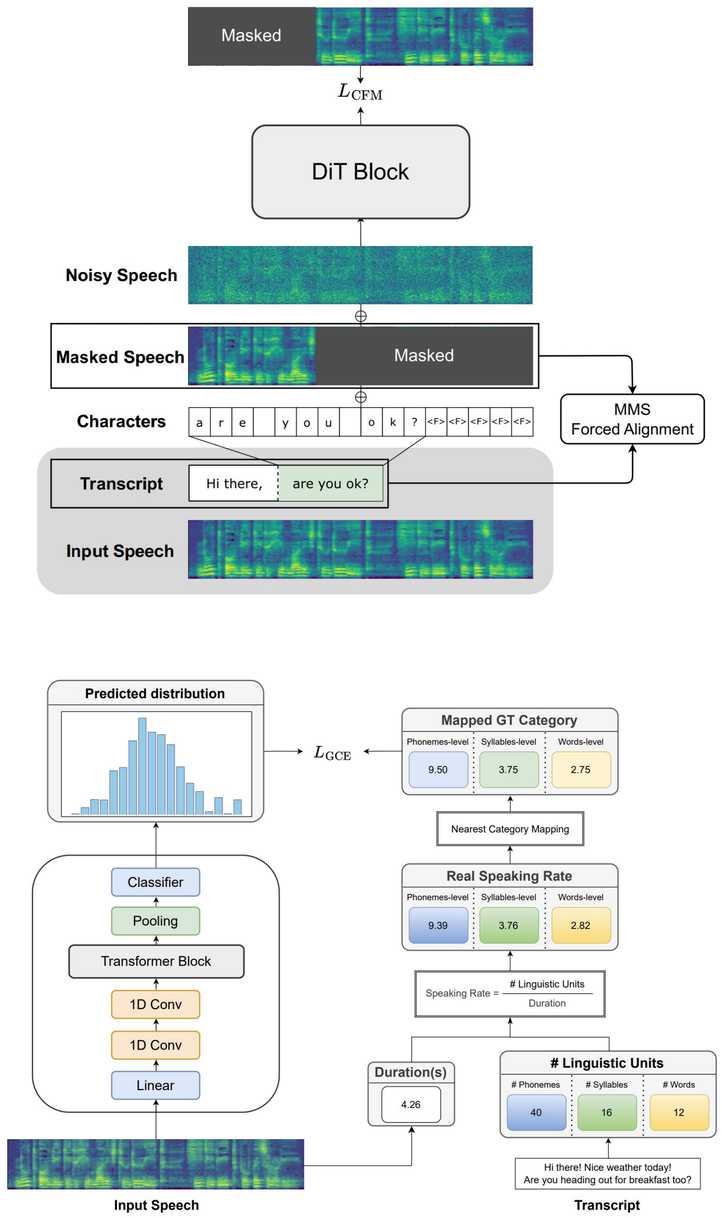
2. Motivation & Solution

Existing NAR TTS systems require audio prompt transcripts, which creates inherent challenges when reference transcripts are unavailable.

Solution:

- Preprocess audio with ***MMS forced alignment*** to extract word boundaries, enabling transcript-free training.
- Train ***speaking rate predictors*** to estimate duration directly from audio prompt to predict the duration of the target utterance

3. Methods



4. Results

System	Duration Method	WER(%)↓	SIM-o↑	UTMOS↑
LibriSpeech-PC *test-clean*				
Baseline	Length-ratio	2.205	**0.668**	3.797
CL-F5	**M1**	**2.079**	0.663	3.884
CL-F5	**M2**	2.120	0.658	**3.892**
CL-F5	**M3**	2.894	0.652	3.855
SeedTTS *test-en*				
Baseline	Length-ratio	1.545	**0.676**	3.581
CL-F5	**M1**	**1.513**	0.662	**3.629**
CL-F5	**M2**	1.594	0.660	3.625
CL-F5	**M3**	2.009	0.646	3.593
SeedTTS *test-zh*				
Baseline	Length-ratio	**1.475**	0.762	2.898
CL-F5	**M1**	1.605	0.759	**2.913**
CL-F5	**M2**	1.481	**0.764**	2.887
CL-F5	**M3**	1.616	0.763	2.889

ID	System	MAE(s)↓	MRE(%)↓
Librispeech-PC *test-clean*			
M1	Phonemes-level predictor	0.759	**11.932**
M2	Syllables-level predictor	**0.757**	11.945
M3	Words-level predictor	1.171	18.406
SeedTTS *test-en*			
M1	Phonemes-level predictor	**0.637**	**15.017**
M2	Syllables-level predictor	0.704	16.497
M3	Words-level predictor	0.886	20.031
SeedTTS *test-zh*			
M1	Phonemes-level predictor	0.845	14.469
M2	Syllables-level predictor	**0.783**	**13.771**
M3	Words-level predictor	0.908	16.156