X-Voice: Enabling Everyone to Speak 30 Languages via Zero-Shot Cross-Lingual Voice Cloning

Rixi Xu[1*], Qingyu Liu[1,3*], Haitao Li[2,6], Yushen Chen[1,2], Zhikang Niu[1,2], Yunting Yang[4], Jian Zhao[4], Ke Li[5], Berrak Sisman[3], Qinyuan Cheng[2,7], Xipeng Qiu[2,7], Kai Yu[1], Xie Chen[1,2†]

[1] X-LANCE Lab, Shanghai Jiao Tong University [2] Shanghai Innovation Institute [3] Center for Language and Speech Processing, Johns Hopkins University
[4] Geely [5] Dataocean AI [6] Zhejiang University [7] Fudan University

AT A GLANCE

- **0.4B** parameters
- **30** languages
- **420K-hour** multilingual corpus
- **Zero-shot** cross-lingual voice cloning
- **No prompt transcript** needed at inference
- **Open-source** resources

MOTIVATION

- Limited language coverage in existing multilingual TTS systems.
- Reliance on reference transcripts for cross-lingual voice cloning.
- Lack of large-scale open multilingual datasets and benchmarks.

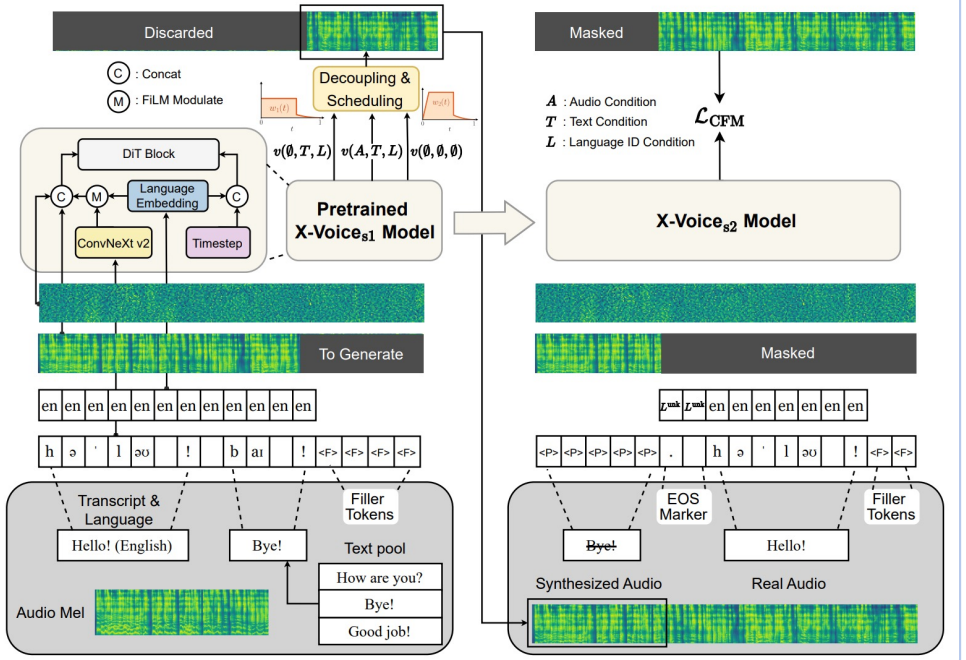
KEY CONTRIBUTIONS

1. Parameter-efficient multilingual foundation model for 30 languages.
2. Novel two-stage training pipeline for transcript-free voice cloning.
3. Open multilingual ecosystem with data, benchmark, and code.

OPEN-SOURCE RESOURCES

Source Code | Interactive Demo | Demo Page | Training Set

TWO-STAGE TRAINING PARADIGM



METHOD HIGHLIGNTS

1 Stage 1: Multilingual Foundation

Unified Representation
Use Pinyin / IPA to build a shared phonetic space across 30 languages

Dual-Level Language ID Injection
Inject Language ID at both textual and time levels to reduce accent leakage.

Decoupled & Scheduled CFG
Independently control acoustic and linguistic guidance for better similarity, intelligibility, and naturalness.

2 Stage 2: Transcript-Free Adaption

Real–Synthetic Pair Construction
Use X-Voice$_{s1}$ to synthesize speaker-consistent audio prompts from a high-quality multilingual subset.

SFT without Reference Text
Remove reference transcripts and reconstruct paired real speech from synthetic audio prompts, enabling transcript-free voice cloning.

RESULTS

Language	WER↓ GT	Qwen3	LEMAS	MOSS	Fish	Omni	X-Voice$_{s1}$	X-Voice$_{s2}$	SIM-o↑ GT	Qwen3	LEMAS	MOSS	Fish	Omni	X-Voice$_{s1}$	X-Voice$_{s2}$
Asian Languages																
Chinese	2.41	**2.16**	6.07	2.91	2.57	2.23	2.86	2.87	0.723	0.728	0.655	0.722	0.686	**0.736**	0.698	0.700
Indonesian	3.67	-	4.40	-	6.11	**2.15**	2.98	2.53	0.670	-	0.604	-	0.596	**0.682**	0.644	0.651
Japanese	8.06	6.69	-	12.93	6.06	**5.98**	8.04	7.93	0.713	0.723	-	0.703	0.658	**0.724**	0.682	0.701
Korean	4.54	3.64	-	3.54	**1.37**	3.42	2.42	2.40	0.759	0.738	-	0.728	0.702	**0.748**	0.723	0.731
Thai	3.98	-	-	-	8.47	**2.90**	5.47	5.64	0.694	-	-	-	0.644	**0.698**	0.654	0.671
Vietnamese	3.37	-	4.97	-	29.31	4.38	4.25	**3.67**	0.741	-	0.642	-	0.696	**0.727**	0.677	0.686
European Languages Widely Used in TTS																
English	4.63	3.89	4.15	3.54	3.01	2.44	2.36	**2.29**	0.730	0.697	0.560	0.662	0.622	**0.719**	0.586	0.547
French	7.92	**6.65**	8.71	9.05	7.98	6.88	8.71	8.55	0.742	0.724	0.639	0.699	0.668	**0.740**	0.680	0.680
German	3.51	**2.64**	6.55	3.85	3.45	2.80	3.76	3.91	0.765	0.742	0.666	0.718	0.689	**0.763**	0.698	0.698
Italian	4.44	**2.93**	5.39	6.33	4.60	3.13	3.95	3.89	0.760	0.746	0.655	0.720	0.692	**0.757**	0.703	0.705
Portuguese	3.39	2.78	3.55	6.19	2.71	**2.11**	3.41	3.29	0.724	0.711	0.647	0.679	0.658	**0.720**	0.665	0.661
Russian	3.75	3.16	3.80	4.82	3.52	3.12	**2.68**	2.74	0.743	0.742	0.676	0.718	0.689	**0.744**	0.714	0.723
Spanish	3.32	**2.11**	4.60	3.88	2.90	2.28	2.83	2.89	0.762	0.749	0.667	0.726	0.694	**0.759**	0.695	0.693
Other European Languages																
Bulgarian	12.58	-	-	-	25.75	9.45	**9.27**	9.75	0.730	-	-	-	0.668	**0.721**	0.709	0.716
Croatian	11.33	-	-	-	11.42	**4.56**	4.81	4.84	0.813	-	-	-	0.744	**0.801**	0.782	0.790
Czech	8.01	-	-	12.02	12.70	**4.58**	4.96	4.84	0.721	-	-	0.692	0.644	**0.736**	0.702	0.706
Danish	13.26	-	-	19.67	25.93	**10.49**	12.53	12.16	0.702	-	-	0.653	0.613	**0.687**	0.669	0.676
Dutch	4.10	-	-	-	4.38	**2.18**	3.19	3.13	0.725	-	-	-	0.650	**0.713**	0.667	0.669
Estonian	18.15	-	-	-	28.12	13.11	11.52	**11.23**	0.776	-	-	-	0.713	**0.758**	0.733	0.732
Finnish	8.51	-	-	-	11.24	5.31	**4.41**	4.47	0.753	-	-	-	0.672	**0.754**	0.719	0.722
Greek	10.84	-	-	15.45	24.04	**8.89**	10.52	10.72	0.614	-	-	0.645	0.617	**0.704**	0.665	0.665
Hungarian	7.23	-	-	19.26	11.34	6.85	5.56	**5.42**	0.731	-	-	0.690	0.666	**0.733**	0.700	0.701
Latvian	11.39	-	-	-	25.24	8.83	**6.95**	7.13	0.715	-	-	-	0.647	**0.714**	0.692	0.694
Lithuanian	12.65	-	-	-	50.33	**11.73**	12.08	12.57	0.727	-	-	-	0.653	**0.727**	0.696	0.702
Maltese	76.06	-	-	-	80.43	70.93	**69.44**	68.11	0.705	-	-	-	0.611	**0.687**	0.653	0.663
Polish	5.12	-	-	8.92	6.12	**3.06**	3.30	3.71	0.726	-	-	0.688	0.657	**0.713**	0.679	0.683
Romanian	9.85	-	-	-	26.44	8.71	8.65	**8.43**	0.703	-	-	-	0.615	**0.708**	0.665	0.672
Slovak	12.67	-	-	-	18.76	**10.59**	10.52	11.04	0.699	-	-	-	0.620	**0.700**	0.670	0.676
Slovenian	12.21	-	-	-	15.57	8.07	**7.62**	7.73	0.683	-	-	-	0.606	**0.675**	0.645	0.653
Swedish	7.29	-	-	9.71	8.06	**5.06**	7.01	7.63	0.735	-	-	0.696	0.666	**0.734**	0.687	0.687

Intra-Lingual Result:
Robust performance across 30 languages, with competitive WER and speaker similarity;

X-Voice$_{s2}$ remains competitive without audio prompt transcripts.

Cross-Lingual Result:
Robust cross-lingual voice cloning across diverse language families, achieving the best or near-best WER in most language pairs.

Model	en→it	it→zh	zh→ru	ru→ko	ko→en	en→ko	ko→ru	ru→zh	zh→it	it→en
Qwen3-TTS	7.06	49.05	3.67	3.74	6.29	4.03	7.63	136.69	7.74	3.62
LEMAS-TTS	6.11	12.11	5.13	-	-	-	-	18.63	9.95	4.04
MOSS-TTS	7.21	11.64	5.13	3.64	9.16	4.06	10.23	7.52	5.72	12.4
Fish Audio S2	9.16	11.04	4.71	3.85	3.60	**1.71**	4.92	11.17	9.93	4.04
OmniVoice	**4.48**	7.58	3.94	5.36	3.56	3.14	11.56	4.38	4.76	2.44
X-Voice$_{s2}$	4.70	**3.11**	**2.85**	**3.00**	**2.15**	3.10	**2.58**	**3.22**	**3.91**	**2.31**

SJTU X-LANCE Lab is dedicated to advancing speech, audio-visual, and language intelligence through world-class technical research that can change the world.