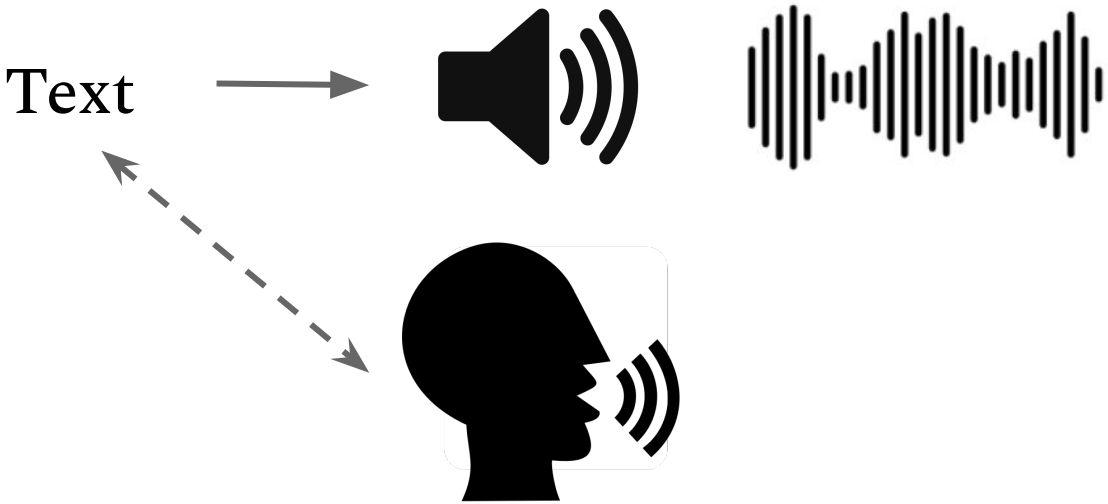
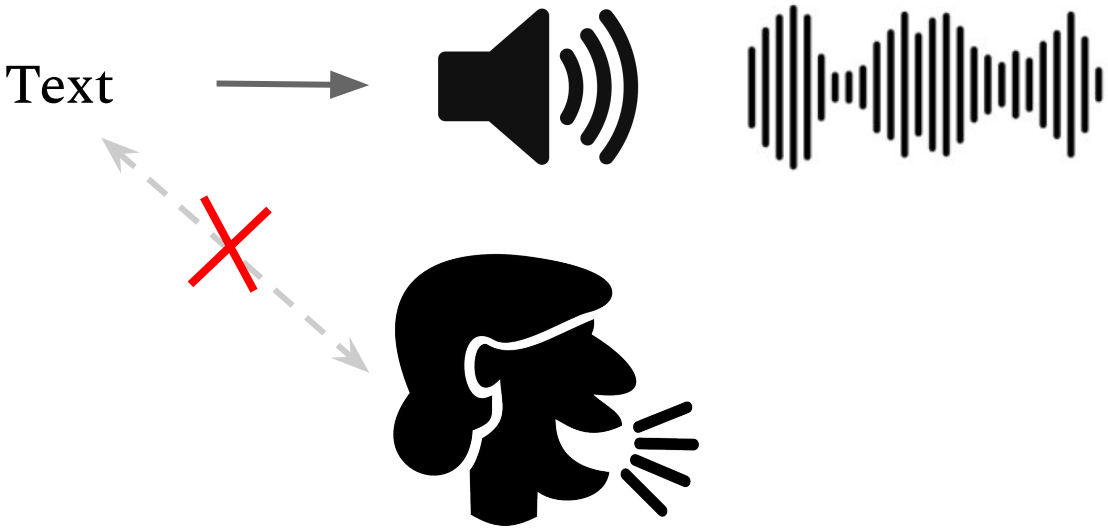
YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone

E. Casanova[1,2], J. Weber[2], C. Shulby[3], A. Candido Junior[4], E. Gölge[2], Moacir A. Ponti[1,5]

Zero-shot Text-to-Speech



Zero-shot Text-to-Speech

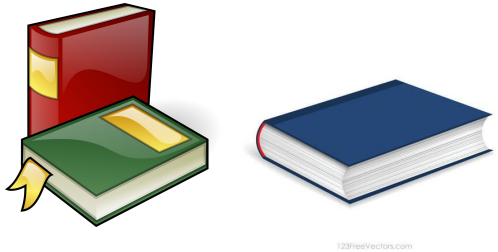


Resources and Performance Gap

English

Mandarin & a few others

Remaining languages



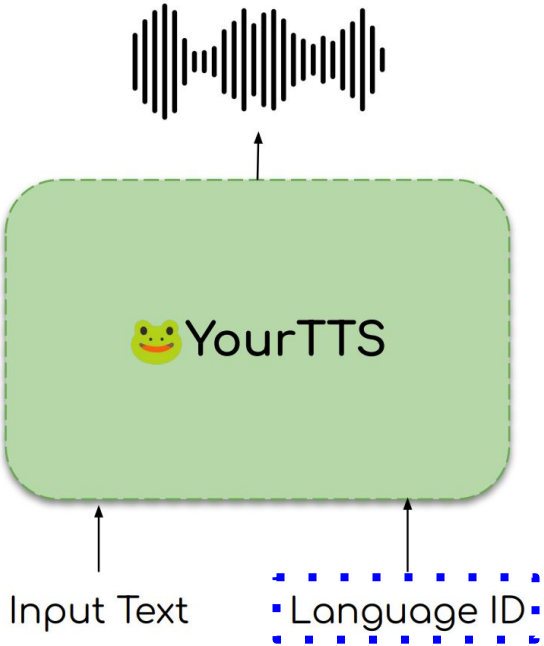
YourTTS, a multilingual model taking advantage of the high speaker count of English

Contribution

The first to explore a multilingual approach in ZS-TTS achieving state-of-the-art results +

- multilingual TTS
- zero shot TTS
- cross-lingual zero shot
- voice conversion

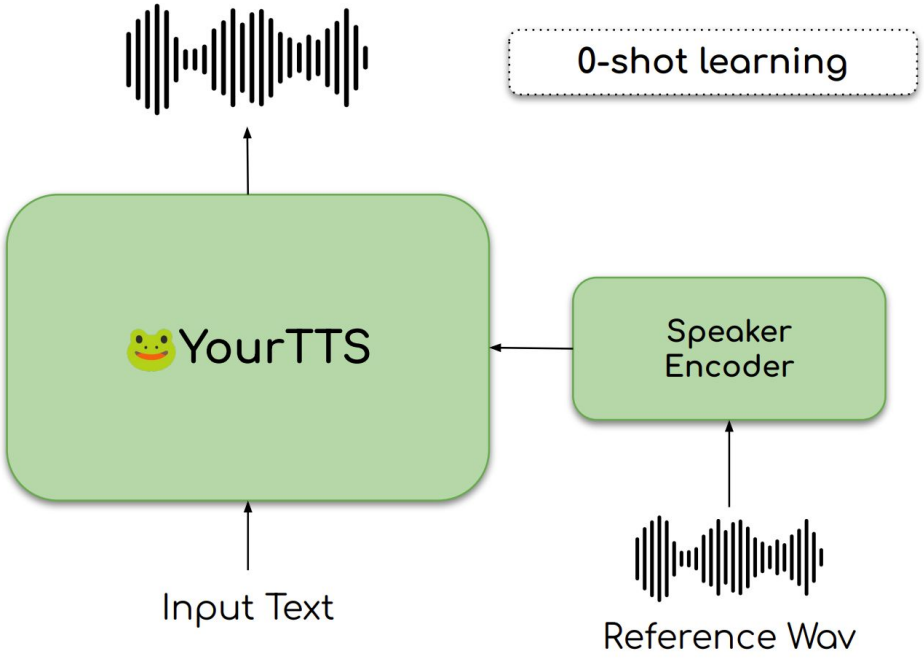
Multi-lingual TTS System



VCTK p228 EN

VCTK p228 FR

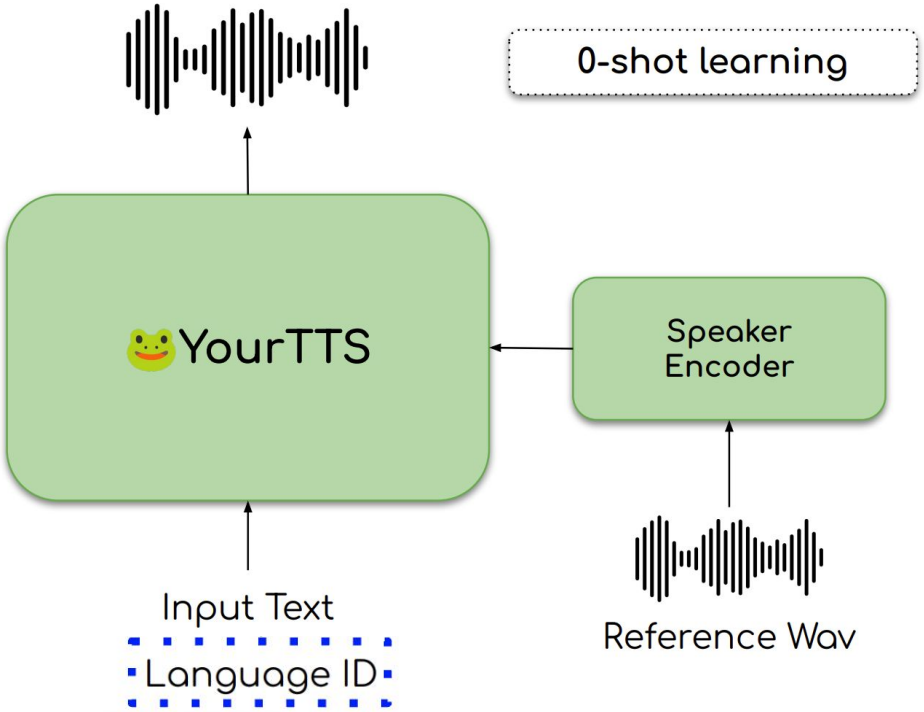
Zero Shot Learning



Julian reference

Julian zero-shot FR

Cross-lingual Zero-Shot

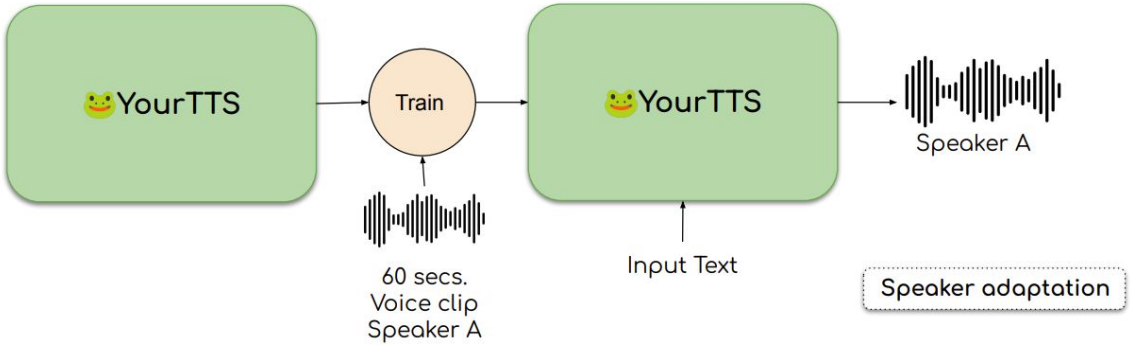


Julian reference

Julian zero-shot FR

Julian zero-shot EN

Finetuning a TTS model with a short sample

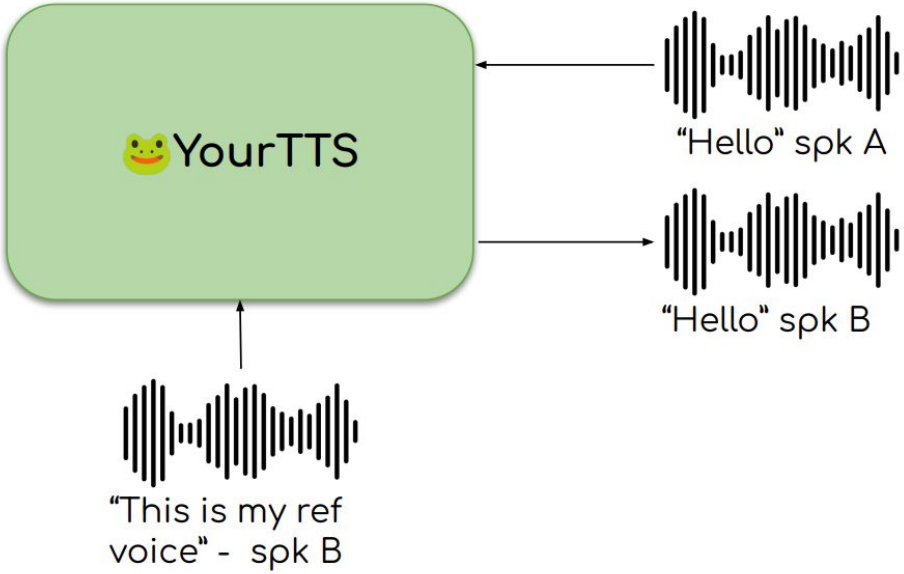


Chris reference

Chris Zero-Shot

Chris Fine-Tuned

Voice conversion



p228 reference (B)

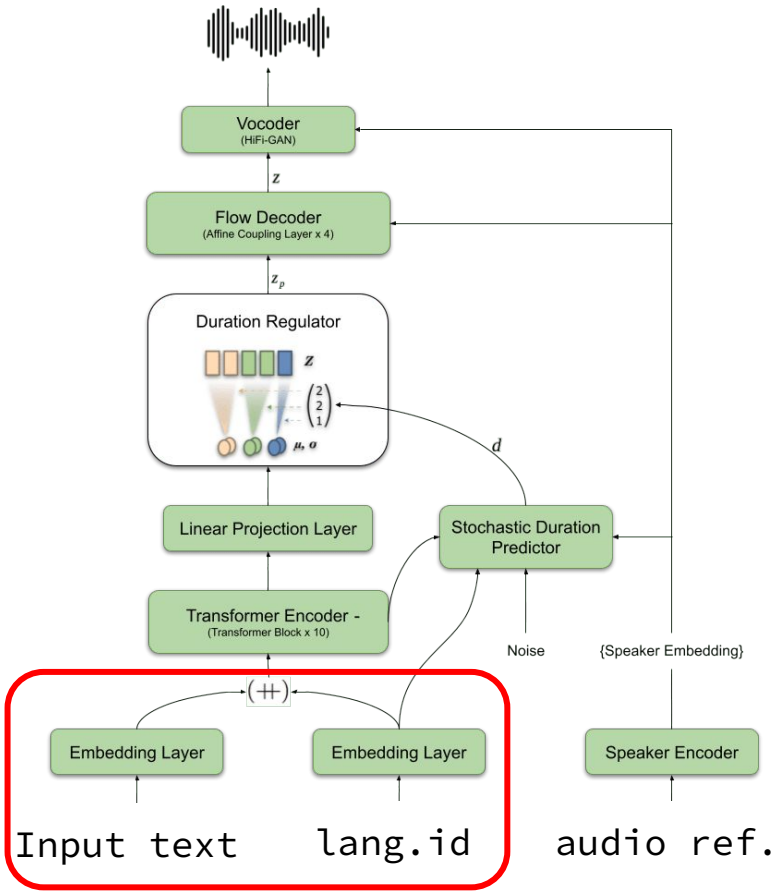
p226 driving audio (A)

p226 converted into p228

YourTTS inference

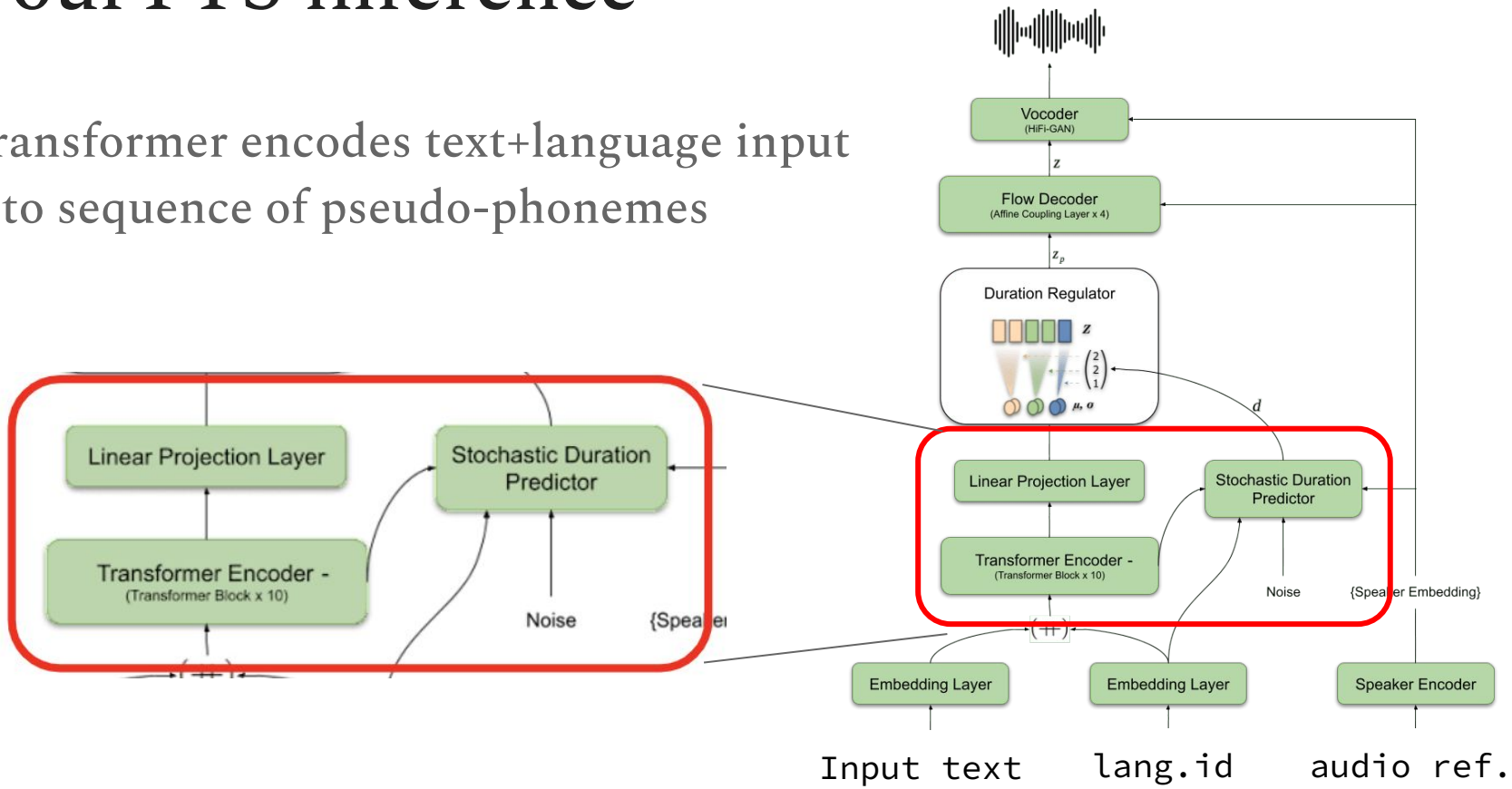
Language id concatenated to every char embedding:

- *allows for code switching*

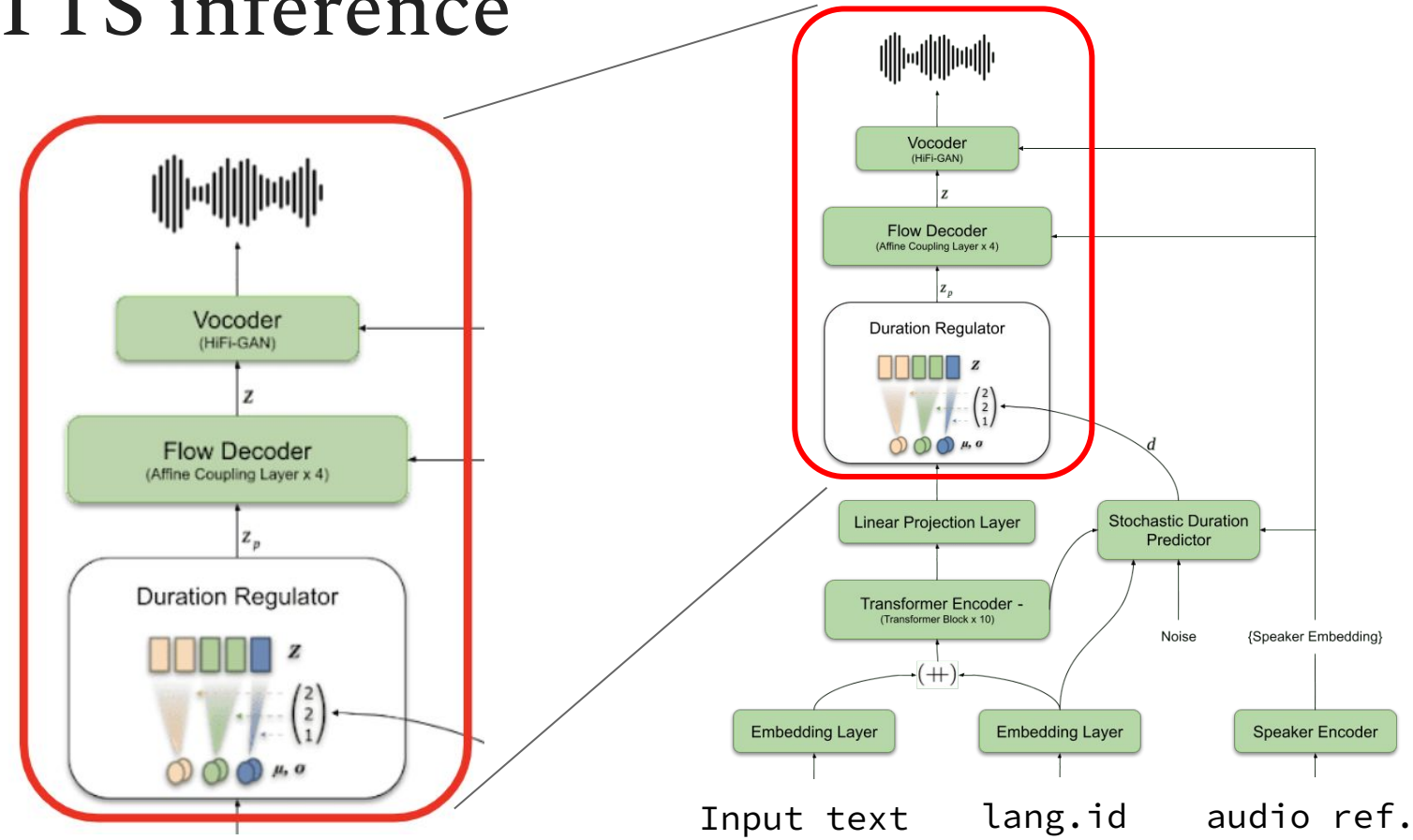


YourTTS inference

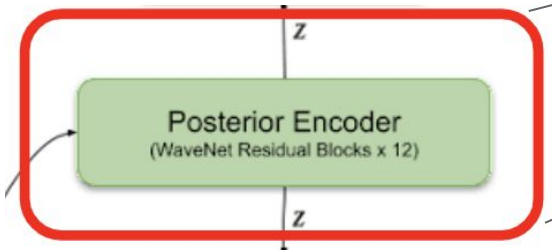
Transformer encodes text+language input into sequence of pseudo-phonemes



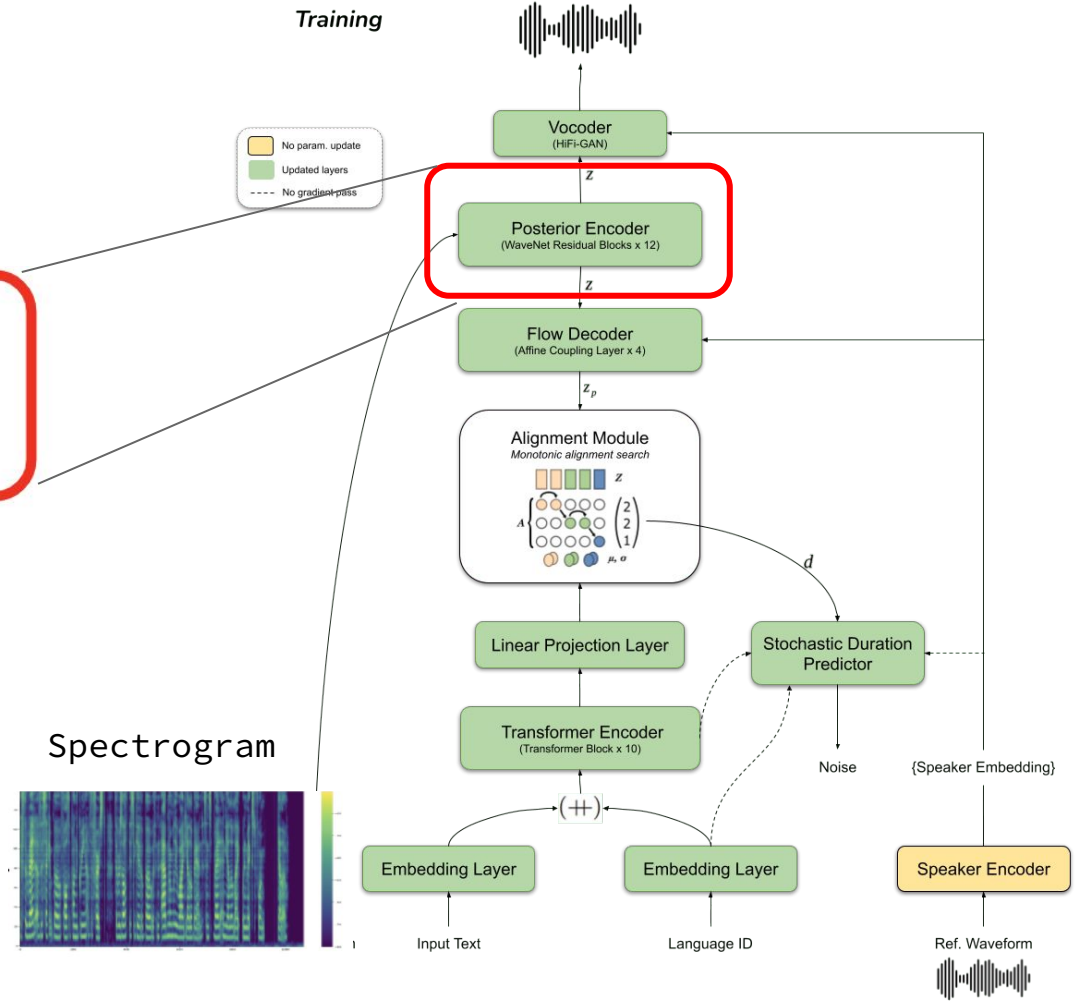
YourTTS inference



Training



Latent representation ***z***

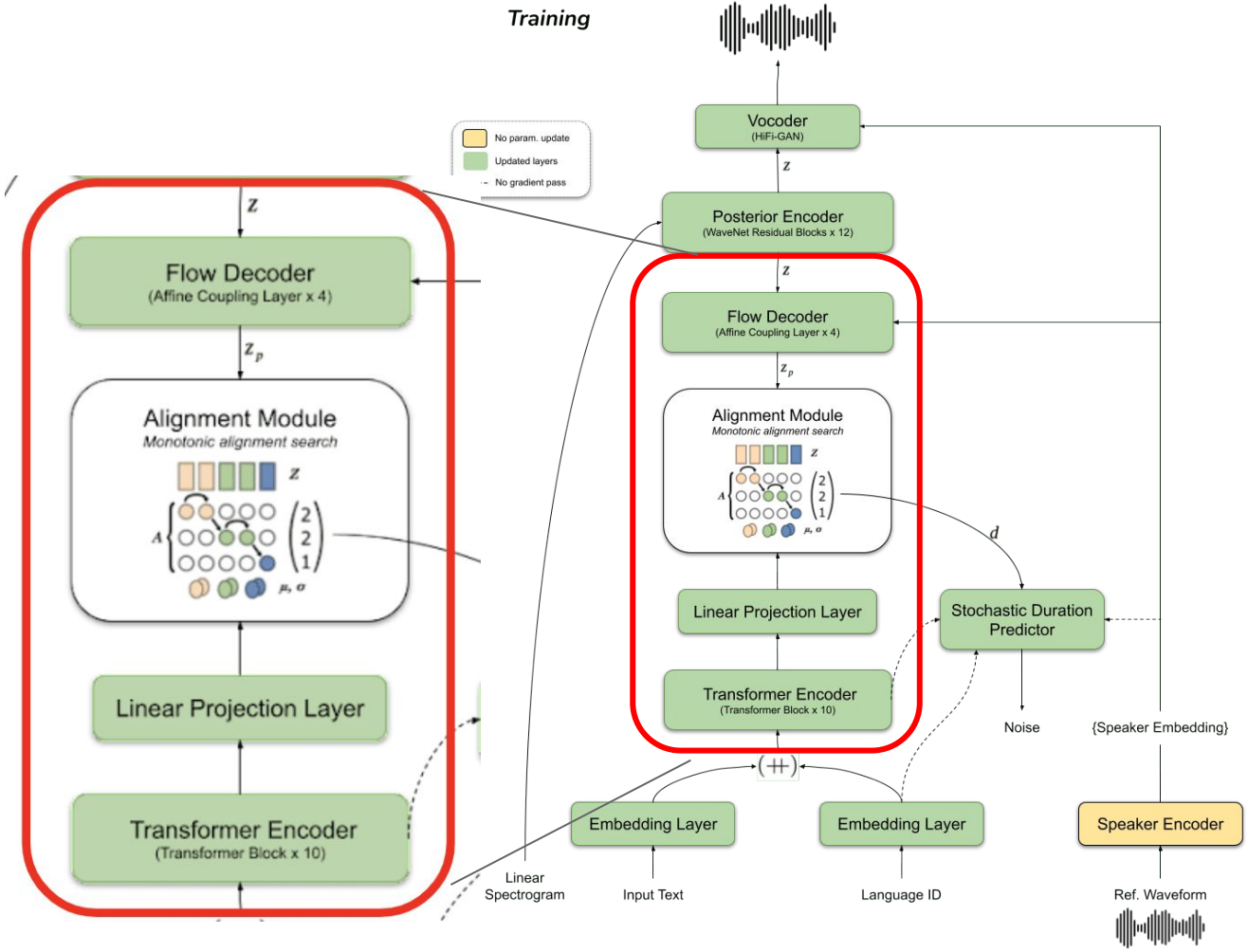


Training

z_p

vs

pseudo-phonemes



Loss Functions

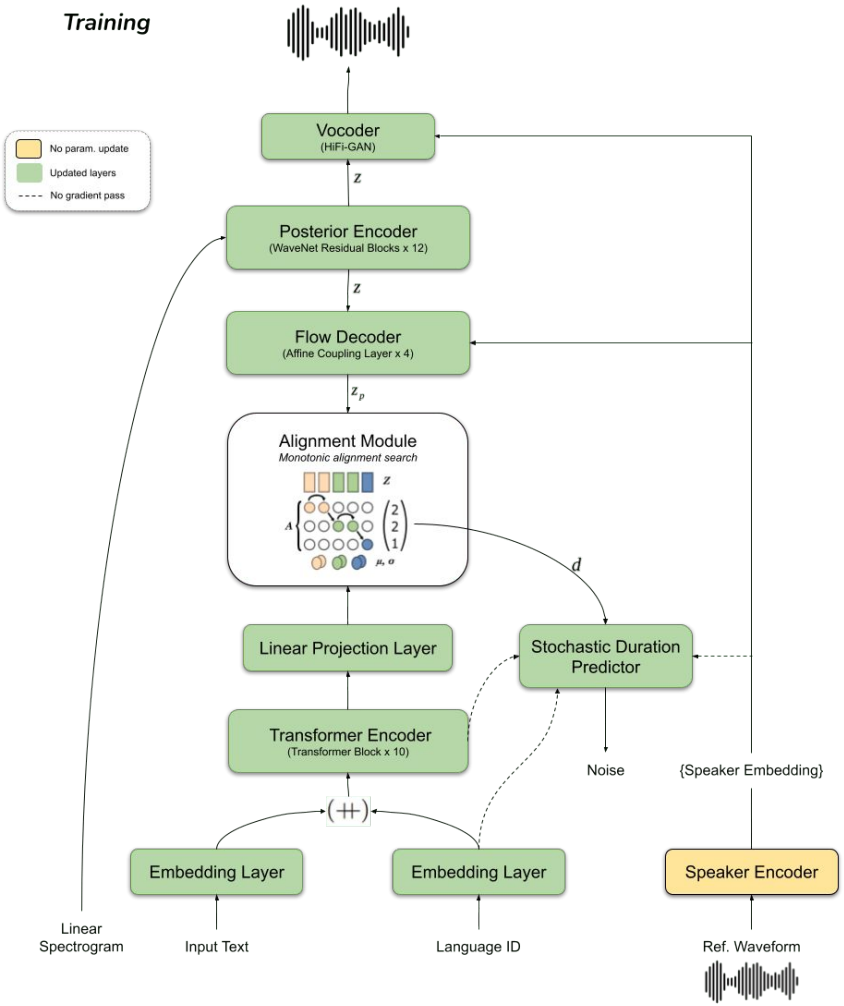
Speaker Embedding pretraining:

- Softmax + Prototype Angular with VoxCeleb dataset

TTS Model:

- Speaker Consistency Loss

  Cosine similarity between ground truth and generated audio embeddings



Portuguese

O modelo me ouviu falando apenas em Português. Mas, com YourTTS, sei falar também em Inglês e Francês.

English

The model heard me speaking only Portuguese. But with YourTTS I can also speak English and French

French

Le modèle m'a entendu parler uniquement en portugais. Mais avec YourTTS, je peux aussi parler en anglais et en français.

Results

Experiments

1. VCTK dataset (98 speakers);
2. VCTK and TTS-Portuguese Corpus (1 speaker);
3. VCTK, TTS-Portuguese and M-AILABS french dataset (5 speakers);
4. VCTK + TTS-Portuguese + M-AILABS + LibriTTS (1151 speakers).

Experiments Setup

- All experiments were implemented using coqui TTS : github.com/coqui-ai/TTS
  an open source TTS framework.

- Audio samples and checkpoints of all experiments are available on:
  github.com/Edresson/YourTTS

ZS-TTS results

	VCTK			LibriTTS			MLS-PT		
Exp.	**SECS**	**MOS**	**Sim-MOS**	**SECS**	**MOS**	**Sim-MOS**	**SECS**	**MOS**	**Sim-MOS**
Ground Truth	0.824	4.26±0.04	4.19±0.06	0.931	4.22±0.05	4.22±0.06	0.9018	4.61±0.05	4.41±0.05
Attentron ZS	(0.731)	(3.86±0.05)	(3.30 ±0.06)	–	–	–	–	–	–
SC-GlowTTS	(0.804)	(3.78±0.07)	(3.99±0.07)	–	–	–	–	–	–
Exp. 1	**0.864**	4.21±0.04	4.16±0.05	0.754	**4.25±0.05**	3.98±0.07	–	–	–
Exp. 1 + SCL	0.861	4.20±0.05	4.13±0.06	0.765	4.21±0.04	4.05±0.07	–	–	–
Exp. 2	0.857	**4.24±0.04**	4.15±0.06	0.762	4.22±0.05	4.01±0.07	0.740	3.96±0.08	3.02±0.1
Exp. 2 + SCL	**0.864**	4.19±0.05	**4.17±0.06**	0.773	4.23±0.05	4.01±0.07	0.745	4.09±0.07	2.98±0.1
Exp. 3	0.851	4.21±0.04	4.10±0.06	0.761	4.21±0.04	4.01±0.05	0.761	4.01±0.08	**3.19±0.1**
Exp. 3 + SCL	0.855	4.22±0.05	4.06±0.06	0.778	4.17±0.05	3.98±0.07	0.766	**4.11±0.07**	3.17±0.1
Exp. 4 + SCL	0.843	4.23±0.05	4.10±0.06	**0.856**	4.18±0.05	**4.07±0.07**	**0.798**	3.97±0.08	3.07±0.1

- SOTA results in similarity and speech quality for unseen speakers
- Produce female voice in PT without seen female voice during training

Zero-shot voice conversion results

REF/TAR	M-M		M-F		F-F		F-M		ALL	
	MOS	SIM-MOS	MOS	SIM-MOS	MOS	SIM-MOS	MOS	SIM-MOS	MOS	SIM-MOS
EN-EN	4.22±0.10	4.15±0.12	4.14±0.09	4.11±0.12	4.16±0.12	3.96±0.15	4.26±0.09	4.05±0.11	4.20±0.05	4.07±0.06
PT-PT	3.84 ± 0.18	3.80 ± 0.15	3.46 ± 0.10	3.12 ± 0.17	3.66 ± 0.2	3.35 ± 0.19	3.67 ± 0.16	3.54 ± 0.16	3.64 ± 0.09	3.43 ± 0.09
EN-PT	4.17±0.09	3.68 ± 0.10	4.24±0.08	3.54 ± 0.11	4.14±0.09	3.58 ± 0.12	4.12±0.10	3.58 ± 0.11	4.17±0.04	3.59 ± 0.05
PT-EN	3.62 ± 0.16	3.8 ± 0.10	2.95 ± 0.2	3.67 ± 0.11	3.51 ± 0.18	3.63 ± 0.11	3.47 ± 0.18	3.57 ± 0.11	3.40 ± 0.09	3.67 ± 0.05

- Intra-lingual results comparable to SOTA in VCTK
- Cross-lingual results EN2PT similar as PT2PT

Speaker Adaptation results

	Sex	Dur. (Sam.)	Mode	SECS	MOS	Sim-MOS
EN	M	61s (15)	GT	0.875	4.17±0.09	**4.08±0.13**
			ZS	0.851	4.11±0.07	4.04±0.09
			FT	**0.880**	4.17±0.07	**4.08±0.09**
	F	44s (11)	GT	0.894	4.25±0.11	**4.17±0.13**
			ZS	0.814	4.12±0.08	4.11±0.08
			FT	**0.896**	4.10±0.08	**4.17±0.08**
PT	M	31s (7)	GT	0.880	4.76±0.12	**4.31±0.14**
			ZS	0.817	4.03±0.11	3.35±0.12
			FT	**0.915**	3.74±0.12	4.19±0.07
	F	20s (5)	GT	0.873	4.62±0.19	**4.65±0.14**
			ZS	0.743	3.59±0.13	2.77±0.15
			FT	**0.930**	3.48±0.13	4.43±0.06

- good results with 1 min of speech, presenting naturalness over zero-shot
- 44 seconds or less reduces the quality when compared to the zero-shot.

Limitations and room for improvement

- Accent supressing
- Monotonic tones for long sentences
- Unnatural speeds for some speaker/language combinations
- Reduce audio artefacts

Possibilities

- Capture different accents and particularities
- Learn from few speakers of (very) low resource languages
  - indigenous languages
  - dialects
- Learn speech of persons that may lose their voices

Thanks

Center for Artificial Intelligence

CEIA
CENTRO DE EXCELÊNCIA EM IA

FAPESP

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

coqui

DEFINED.AI

ICMC USP
SÃO CARLOS

mercado libre