The 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023)

IJCAI/2023 MACAO

Learning to Speak from Text: Zero-Shot Multilingual Text-to-Speech with Unsupervised Text Pretraining

Takaaki Saeki[1], Soumi Maiti[2], Xinjian Li[2], Shinji Watanabe[2], Shinnosuke Takamichi[1], Hiroshi Saruwatari[1]

[1]The University of Tokyo, Japan
[2]Carnegie Mellon University, USA

Watanabe's Audio and Voice Lab

Multilingual Text-to-Speech (TTS)

TTS is widely used in AI voice user interfaces.

Recent neural TTS [Kim+21] achieves **human-like natural speech**.

Multilingual Text-to-Speech (TTS)

Only open to limited number of resource-rich languages.

- Requiring **paired speech-text data** with studio recording audio**.**
- **Hard to collect enough data** for low-resource languages.

Low-Resource Multilingual Text-to-Speech (TTS)

Low-resource TTS approaches to expand number of languages.

- Adapting multilingual model to low-resource language [Lee+18] [He+21]
- Using untranscribed speech for training [Zhang+20] [Ni+21]
- Joint semi-supervised learning with different types of data [Saeki+23]

Previous approaches heavily rely on speech recordings.

➢ **Often challenging to collect training data for target languages.**

Building TTS with Multilingual Text Data

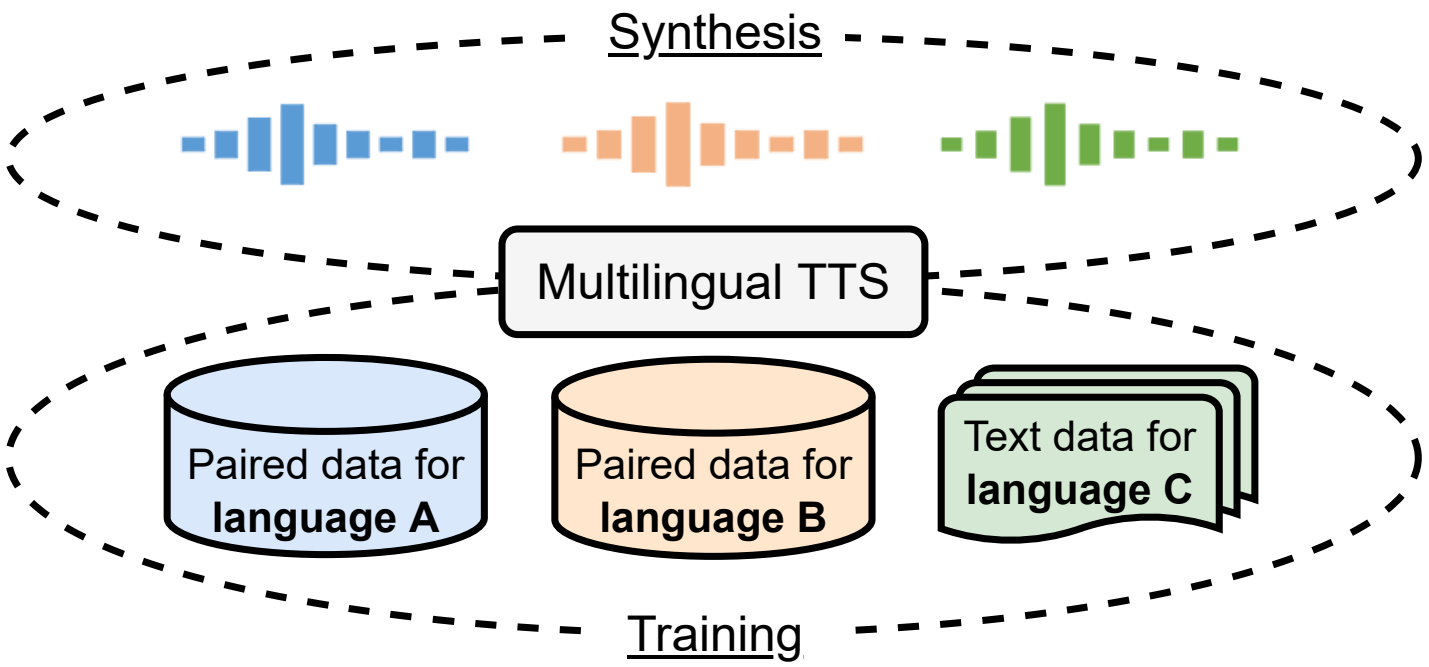
Text data is much easier to collect than paired speech-text data.

- **No need to collect studio recording audio**.
- **No need of preprocessing** to align speech and text.
- **Free from sensitive speaker-related information**.

Building TTS with Multilingual Text Data

Goal: Building TTS for languages with only textual resources.

Open-up TTS systems to much more languages.



Zero-Shot TTS with Unspoken Text Data

Strong zero-shot cross-lingual transferability of multilingual BERT [Devlin+19] in natural language processing tasks [Pires+19].

We investigate cross-lingual transfer of multilingual LM for TTS.

Zero-Shot TTS with Unspoken Text Data

Strong zero-shot cross-lingual transferability of multilingual BERT [Devlin+19] in natural language processing tasks [Pires+19].

We investigate cross-lingual transfer of multilingual LM for TTS.

Contributions:

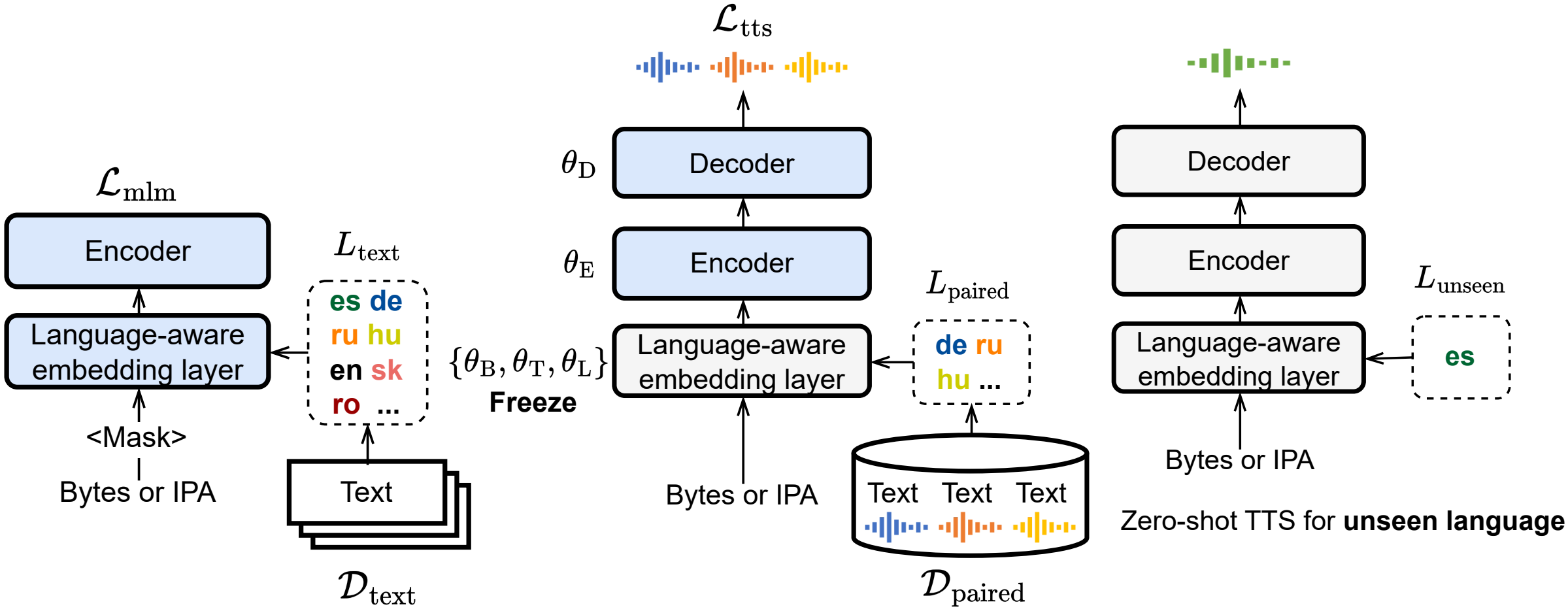
- ❑ **Propose zero-shot TTS from text data, achieving high intelligibility.**
- ❑ **Improve multilingual TTS without per-language pronunciation knowledge.**
- ❑ **Conducted comprehensive ablation studies.**

Outline

- ❑ Background
- ❑ Method
- ❑ Experiment
- ❑ Conclusions

Overview of Our Framework

Transformer-based multilingual TTS model using text-only and paired data.

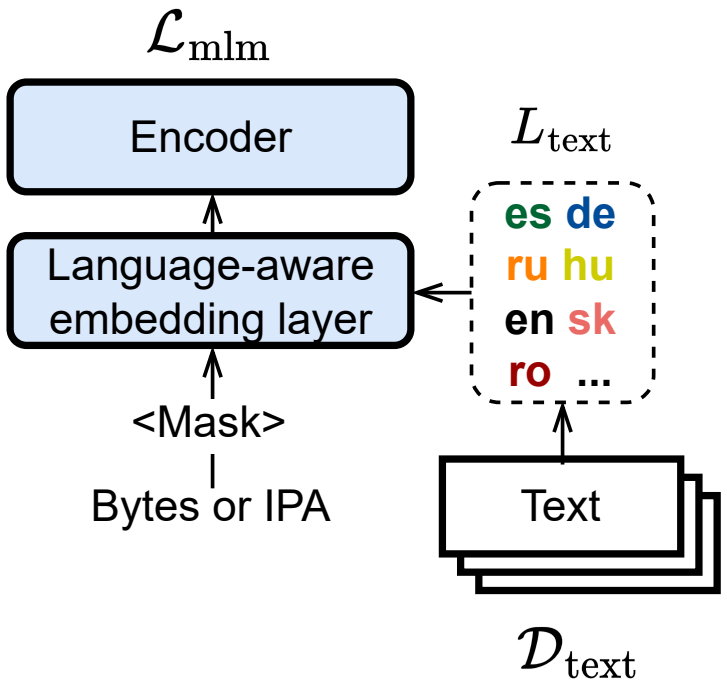


(a) Unsupervised multilingual text pretraining (b) Supervised learning with paired data (c) Inference

Unsupervised Text Pretraining

MLM pretraining with multilingual text including many languages

- ❑ **Language-agnostic tokens (Bytes and IPA*)** for TTS
- ❑ **Language-aware embedding layer** to inject language info.

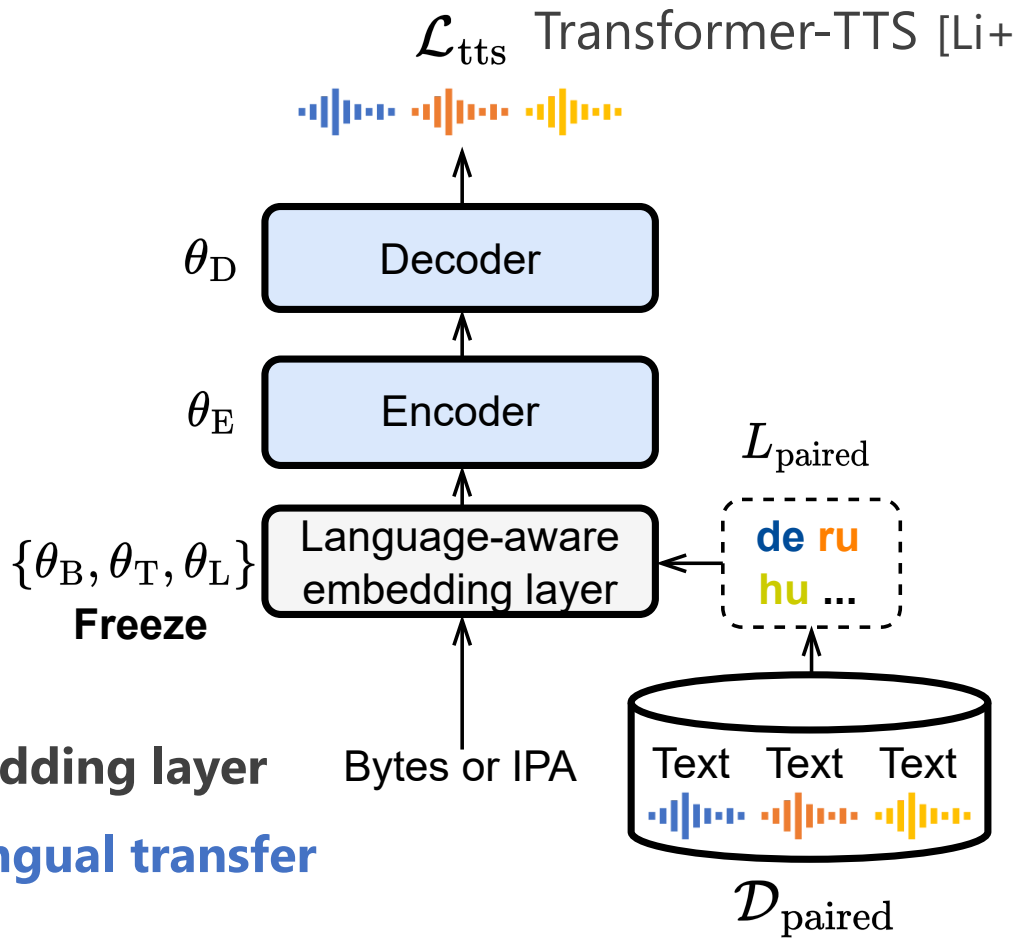


(a) Unsupervised multilingual text pretraining

*IPA: Using International Phonetic Alphabet Symbols

Overview of Our Framework

Supervised learning with paired data including a few languages.



(b) Supervised learning with paired data

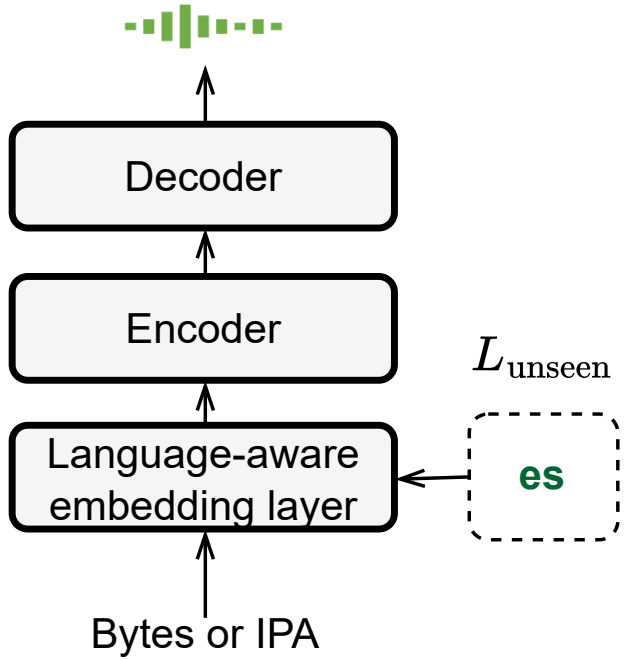
Frozen language-aware embedding layer

Facilitating zero-shot cross-lingual transfer

Overview of Our Framework

Zero-shot TTS by using *unseen* language IDs

Using language IDs only included in text data, not in paired data.



Zero-shot TTS for **unseen language**

(c) Inference

Outline

- ❑ Background
- ❑ Related Work
- ❑ Method
- ❑ Experiment
- ❑ Conclusions

Languages and Dataset

Text data:19 European languages.

Paired data: 7 European languages.

Languages	Code	Text-only data	Paired data	
			Text	Audio
Seen languages for evaluation L_{seen}				
German	de	359MB	0.73MB	16.13h
French	fr	372MB	0.94MB	19.15h
Dutch	nl	336MB	0.75MB	14.10h
Finnish	fi	308MB	0.47MB	21.36h
Hungarian	hu	104MB	0.51MB	10.53h
Russian	ru	4.9MB	1.5MB	10.00h
Greek	el	0.39MB	0.39MB	4.13h
Unseen language for evaluation L_{unseen}				
Spanish	es	345MB	0.0MB (1.2MB)	0.00h (23.81h)

Languages and Dataset

Chose **Spanish** as an unseen language for main evaluation.

* To ensure enough human evaluators.

Languages	Code	Text-only data	Paired data	
			Text	Audio
Seen languages for evaluation L_{seen}				
German	de	359MB	0.73MB	16.13h
French	fr	372MB	0.94MB	19.15h
Dutch	nl	336MB	0.75MB	14.10h
Finnish	fi	308MB	0.47MB	21.36h
Hungarian	hu	104MB	0.51MB	10.53h
Russian	ru	4.9MB	1.5MB	10.00h
Greek	el	0.39MB	0.39MB	4.13h
Unseen language for evaluation L_{unseen}				
Spanish	es	345MB	0.0MB (1.2MB)	0.00h (23.81h)

Ablation Models

Methods

- ***Baseline***: **without** unsupervised text pretraining.
- ***Proposed***: **with** unsupervised text pretraining.
- ***Oracle***: **Using paired data** for the target language.

Token types

- **Bytes:** Without language-specific knowledge.
- **IPA:** Using per-language pronunciation dictionary.

Evaluation Metrics

Evaluation metrics

- ❑ Mel Cepstral Distortion (**MCD**) [Fukada+1992]
- ❑ Character Error Rates (**CER**) computed by Whisper [Radford+22]
- ❑ Automatic Mean opinion scores (**AMOS**)
- ❑ Subjective mean opinion scores (**MOS**)

Results (Unseen Language)

Oracle: Using unseen language during training

	Spanish	
	MCD (↓)	CER (↓)
Ground-truth	-	2.71
Oracle (Bytes)	8.65	10.70
Oracle (IPA)	6.20	5.32
Baseline (IPA)	10.75	44.75
Proposed (Bytes)	9.05	18.27
Proposed (IPA)	9.44	11.69

Results (Unseen Language)

Oracle: Using unseen language during training

	Spanish	
	MCD (↓)	CER (↓)
Ground-truth	-	2.71
Oracle (Bytes)	8.65	10.70
Oracle (IPA)	6.20	5.32
Baseline (IPA)	10.75	44.75
Proposed (Bytes)	**9.05**	18.27
Proposed (IPA)	9.44	**11.69**

Proposed **achieved intelligible** (< 20% CER) **zero-shot TTS compared with** ***Baseline***

Results (Unseen Language)

Oracle: Using unseen language during training

	Spanish	
	MCD (↓)	CER (↓)
Ground-truth	-	2.71
Oracle (Bytes)	8.65	10.70
Oracle (IPA)	**6.20**	**5.32**
Baseline (IPA)	10.75	44.75
Proposed (Bytes)	9.05	18.27
Proposed (IPA)	9.44	11.69

Proposed **was still worse than** ***Oracle*** **(Phones) but comparable to** ***Oracle*** **(Bytes)**

Demo

Se me representaba el sonido de las campanas de la iglesia, tocadas por los cuatro muchachos o por el ingrato padre.

	Spanish	
	Sample	CER (↓)
Ground-truth		2.71
Oracle (IPA)		5.32
Baseline (Bytes)		66.45
Proposed (Bytes)		**18.27**

Summary

Background

- Need to reduce cost of data collection of neural multilingual TTS

Method

- **Multilingual unsupervised text pretraining**
- Zero-shot TTS from unseen language

Results

- Achieved highly intelligible (CER < 12%) zero-shot TTS
- Observed **language dependency**

Future work

- Need to improve **naturalness and prosody**
- Need to develop a method that **works well for many languages**

Paper

More audio samples