An Initial Investigation of Language Adaptation for TTS Systems under Low-resource Scenarios

Cheng Gong[1,2], Erica Cooper[1], Xin Wang[2], Chunyu Qiang[1], Mengzhe Geng[3], Dan Wells[4], Longbiao Wang[1], Jianwu Dang[1], Marc Tessier[3], Aidan Pine[3], Korin Richmond[4], Junichi Yamagishi[2]

[1]Tianjin University, [2]National Institute of Informatics
[3]National Research Council Canada, [4]University of Edinburgh

INTERSPEECH 2024

Introduction

INTERSPEECH 2024

- Neural text-to-speech (TTS) models have made remarkable progress in many industrial applications and academic research.
- However, most previous multilingual and multi-speaker TTS models [1-3] are still limited in supporting a wide range of **languages and speakers**, as they require a large amount of high-quality training data.

Low-resource scenarios

- Thousands of languages globally
- Cost time and money

[1] Zhang et al., "Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning," in *Proc. Interspeech*. 2019.
[2] Badlani et al, "RAD-MMM: Multilingual multiaccented multi- speaker text to speech," in Proc. Interspeech, 2023.
[3] Casanova et al, "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone," in Proc. ICML, 2022.

Introduction

- With the advent of massively multilingual speech models like XLSR, using **self-supervised learning (SSL) speech representations** from multilingual models has become a promising solution for low-resource language speech processing tasks.

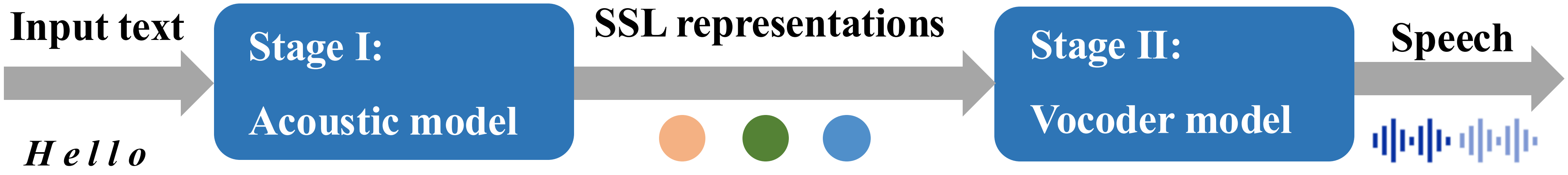


Fig.1 Speech synthesis using SSL representations [4,5]

1) Using SSL representations rather than Mel
2) Two stage pipeline

[4] H. Siuzdak, P. Dura, P. van Rijn, and N. Jacoby, "WavThruVec: Latent speech representation as intermediate features for neural speech synthesis," in Proc. Interspeech, 2022, pp. 833–837.
[5] Kharitonov et al., Speak, Read and Prompt: High-Fidelity Text-to-Speech with Minimal Supervision. Transactions of the Association for Computational Linguistics 2023.

Introduction

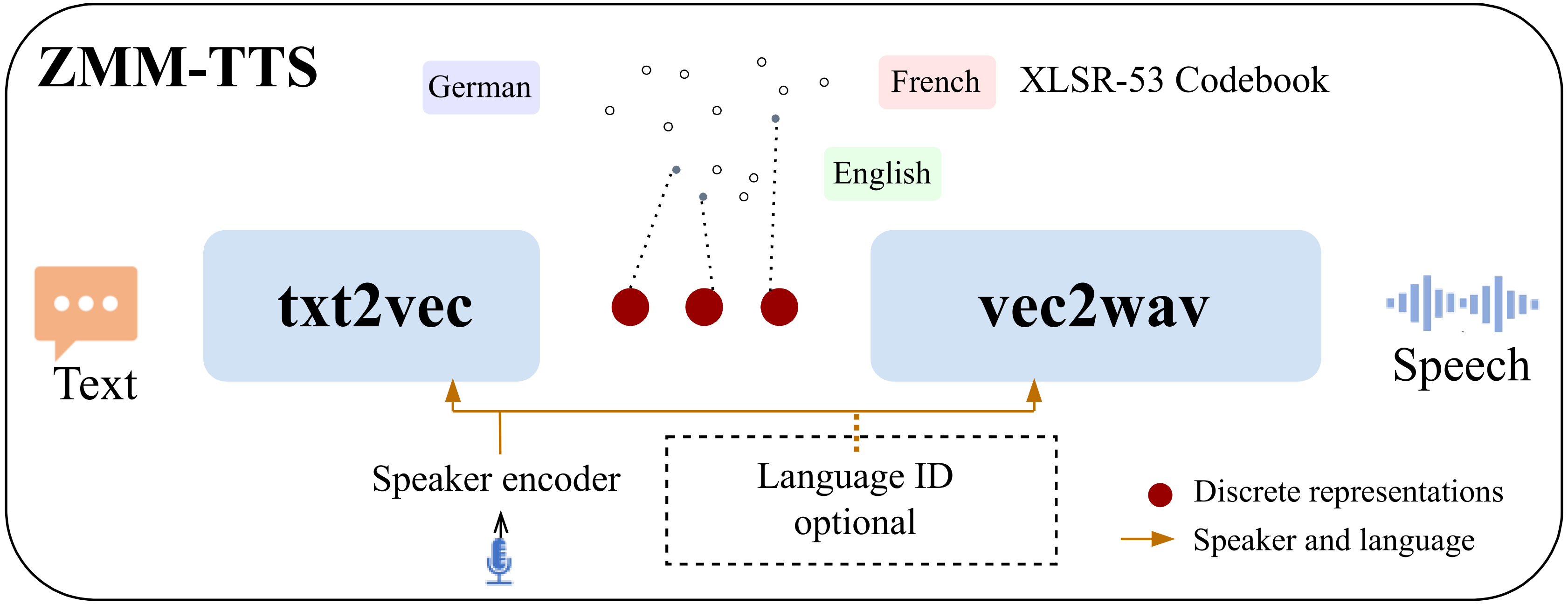
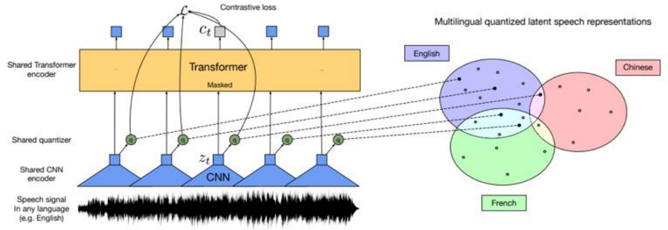


Fig.2 Overview of ZMM-TTS. [5]

- Multilingual audio-based model



- Multilingual text-based model
  - ✓ RoBERTa pre-training approach
  - ✓ 330M phoneme-level sentences
  - ✓ ~100 languages

1) Using XLSR-53 discrete representations
2) txt2vec: XPhoneBERT[6] (Pretrained phoneme representations), FastSpeech 2
3) Vec2wav (Check the paper)

[5] C. Gong, X. Wang, E. Cooper, D. Wells, L. Wang, J. Dang, K. Richmond, and J. Yamagishi, "ZMM-TTS: Zero-shot Mul- tilingual and Multispeaker Speech Synthesis Conditioned on Self-supervised Discrete Speech Representations, IEEE/ACM TASLP.

[6] L. T. Nguyen, et al., "XPhoneBERT: A Pre-trained Multilingual Model for Phoneme Representations for Text-to-Speech," in Proc. Interspeech, 2023, pp. 5506–5510.

Research questions

Considering there are **over 7,000 languages worldwide**, it is worthwhile to investigate the effectiveness of utilizing **self-supervised models** for achieving **low-resource speech synthesis** across various languages.

- Which one is more effective fine-tuning approach, **paired-data** or **audio-only data**?
- How do the **size** of fine-tuning datasets and the **number of speakers** included affect the final performance of speech synthesis?
- How does the performance of language adaptation vary across **different evaluation metrics**?

Datasets

INTERSPEECH 2024

Pre-trained data

Tab.1: The detail of pre-trained data.

Lang	#Spk	Dur(h)	#Sent
eng	91 (46F, 45M)	23.93	6,941
fra	91(45F, 46M)	24.17	9,436
der	91(46F,45M)	24.39	10,468
por	91(46F,45M)	18.91	7,389
spa	91(45F,46M)	20.46	6,374
swe	91(45F,46M)	18.89	9,642
Total	546	130.75	50,250

- 6 Indo-European languages（~130h)
- Language-gender-speaker balance
- From MLS[1], GloablPhone[2], CSS10[3], LJS[4], NST[5]

Fine-tune data

Tab.2: 12 low-resource languages.

Bulgarian (bul)	Croatian (hrv)	Czech (ces)	Dutch (nld)
Italian (ita)	Japanese (jpn)	Korean (kor)	Chinese (cmn)
Polish (pol)	Russian (rus)	Turkish (tur)	Vietnamese (vie)

Tab.3: Fine-tuning data set configurations. S, M, and L denote small, medium, large.

Name	S1	S2	S3	S4	M1	M2	M3	M4	L1	L2	L3	L4
Spk	2	4	10	20	2	4	10	20	2	4	10	20
Utt	12	6	2	1	25	12	5	2	50	25	10	5
Total	24	24	20	20	50	48	48	40	100	100	100	100

- **12** languages, **12** different fine-tuning **data size**
- **European/Central Asian/East Asian** language
- Maximum **100 sentences**, **20 speakers**
- **Tonal/non-tonal** language

[1] https://www.openslr.org/94/
[2] T. Schultz et.al, "GlobalPhone: A multilingual text & speech database in 20 languages," ICASSP 2023.
[3] https://github.com/Kyubyong/css10
[4] https://keithito.com/LJ-Speech-Dataset/
[5] https://huggingface.co/datasets/jimregan/nst_swedish_tts

Experiments

Fine-tune method

- ◻ **Paired-data fine-tuning**. We used paired data {text, audio} and performed fine-tuning on both the **txt2vec and vec2wav** models.
- ◻ **Audio-only fine-tuning.** We used audio-only data for fine-tuning the **vec2wav model,** and during testing, txt2vec processes the input in a zero-shot manner.
- ◻ Zero-shot. **Without employing any data** for fine-tuning, both txt2vec and vec2wav were directly tested on zero-shot inference.

- Total of **25 configurations**, including **12 (data sizes)** × **2 (paired and unpaired)** fine-tuned methods with limited data and one **zero-shot model.**
- {S1, S2, · · · , L4} represents **audio-only** fine-tuning, while **{S1′, S2′, • • • , L4′}** represents **paired-data fine-tuning**. In the subsequent sections, we use **0** to represent **zero-shot** inference.

Experiments

INTERSPEECH 2024

Evaluation metrics

✓ **Character error rate (CER)**

- We synthesized **100** sentences for **each language** and computed the CER between the input text and the **ASR-produced** (Whisper[1]) transcripts.

✓ **Language identification probability (LI)**

- Whisper will also recognize the **probability** that these utterances belong to the target language.

✓ **Speaker Encoder Cosine Similarity (SECS)**

- Speaker embeddings of two audio samples extracted through Resemblyzer[2].
- For each language, we use the **same 2 (1 female, 1 male) seen speakers** and **4 (2 female, 2 male) unseen speakers** for the speaker similarity test, and **three sentences** for each speaker.

✓ **UT-MOS**

- We employed automatic MOS **(UT-MOS[3])** prediction model to assess naturalness.
- UT-MOS, CER, and LI were measured on the **same test set.**

[1]https://github.com/openai/whisper
[2]https://github.com/resemble-ai/Resemblyzer
[3]https://github.com/sarulab-speech/UTMOS22

Experiments

INTERSPEECH 2024

Language similarity analysis

- Inspired by the use of angular similarity (calculable from cosine similarity) between two languages' vectors of **phone frequencies** to measure **the similarity between their phone systems** [7-8], we followed this method to analyze the **phonetic similarity** between **12 adaptation languages** and **six pretraining languages** in our study.

$$S_{A,B} = 1 - \frac{2}{\pi} \arccos\left(\frac{\mathrm{PF}_A^\top \mathrm{PF}_B}{||\mathrm{PF}_A|| \, ||\mathrm{PF}_B||}\right)$$

- For language A, we extracted its **phone set (IPA) through CharsiuG2P** and then computed its vector of **phone frequencies PF_A.**
- Angular Similarity of Phone Frequencies **(ASPF)**
- The value of ASPF S ranges **from 0 to 1.**

[7] P. Do et al., "Text-to-speech for under-resourced languages: Phoneme mapping and source language selection in transfer learning," in Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under- Resourced Languages, 2022, pp. 16–22.
[8] ——, "Strategies in Transfer Learning for Low-Resource Speech Synthesis: Phone Mapping, Features Input, and Source Language Selection," in SSW 2023.

Experiments

Result 1: Impact of language variation

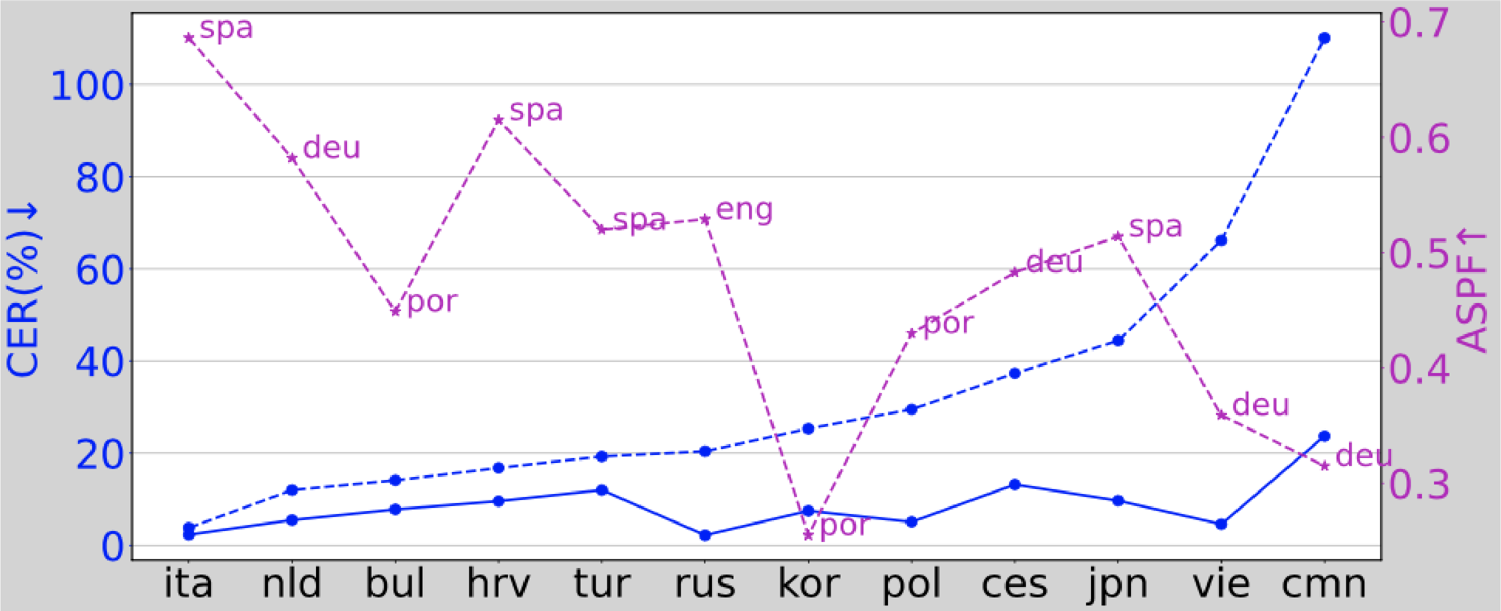


Fig.3: **CER** and **ASPF** values for different languages.

- The **blue** dashed line represents **the best CER performance** achievable by synthesized audio from **25 configurations**, while the **solid line** represents the CER performance of **natural audio**.
- The **purple** dashed line represents the **ASPF value most similar** to the 6 pre-trained languages and its corresponding language.

- **European** vs. **East Asian** languages
  - Ita, nld, and bul CER<20%
  - cmn, vie, and jpn CER>40%
- Correlations between **CER and ASPF**
  - PCC r=-0.630, p=0.028
  - ita spa high similarity (ASPF 0.686)
- Correlations between **GT and syn**
  - PCC r=0.715, p=0.008
  - Chinese

Experiments

INTERSPEECH 2024

Result 2：Impact of finetuning configurations

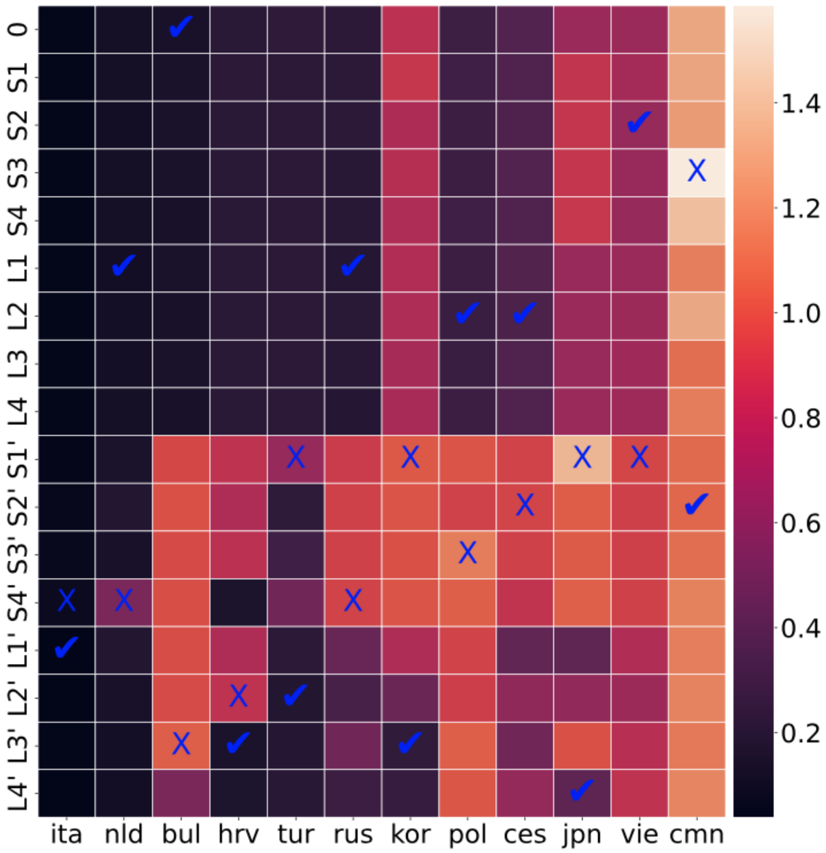


Fig.4: CER results for different languages under various fine-tuning methods. ✓ represents the best result for each language, while an × indicates the worst result.

- **Paired-data vs audio-only fine-tuning**
  - Finetuning **both of txt2vec and vec2wav** is not always the best option
  - 9 languages obtained the **poorest CER** results with **20 paired data**
  - Overfitting
- **Audio-only fine-tuning vs. zero-shot**
  - Not much difference
  - Influenced more by training **text** than audio
- **Datasize impact**
  - Increasing the amount of fine-tuning data
  - 9 languages obtained the **best CER** results with **100** samples (audio-only or paired data)

Experiments

INTERSPEECH 2024

Result 3：Results across different metrics

Tab.5: Results of different metrics. Bold indicates the best result for each language under different fine-tune configurations.

Metrics	Lang	Fine-tuning configurations									GT
		Zero-shot	Audio-only				Paired data				
		0	$L1$	$L2$	$L3$	$L4$	$L1'$	$L2'$	$L3'$	$L4'$	
CER (%) ↓	ita	4.6	4.6	4.7	4.6	4.8	**3.8**	**3.8**	4.7	4.3	2.3
	jpn	67.9	67.0	67.4	67.2	67.8	45.4	64.5	97.6	**44.4**	9.7
	tur	24.4	23.5	22.7	23.1	23.1	23.4	**19.3**	20.7	21.5	12.0
LI (%) ↑	ita	90.6	91.6	92.2	92.0	92.9	96.7	96.6	95.8	**97.0**	98.1
	jpn	41.2	59.1	60.5	61.4	62.0	85.0	92.2	19.7	**88.6**	97.7
	tur	26.1	21.6	25.9	25.7	26.2	66.9	**77.0**	74.9	75.3	98.4
Seen SECS ↑	ita	0.898	0.924	0.926	0.913	0.901	0.948	**0.951**	0.930	0.904	0.999
	jpn	0.849	0.921	0.896	0.889	0.881	**0.942**	0.917	0.799	0.906	0.999
	tur	0.909	**0.949**	0.942	0.936	0.929	**0.949**	0.937	0.933	0.916	0.999
Unseen SECS ↑	ita	**0.842**	0.828	0.841	0.840	0.832	0.803	0.819	0.831	0.832	0.999
	jpn	0.853	0.829	0.858	**0.867**	0.846	0.807	0.851	0.773	0.847	0.999
	tur	0.876	0.840	0.864	0.862	0.859	0.845	**0.877**	0.872	0.873	0.999
UT-MOS ↑	ita	**3.359**	3.227	3.235	3.194	3.227	3.120	3.145	3.065	3.121	3.111
	jpn	**3.286**	3.136	3.100	3.150	3.178	2.977	3.078	1.840	3.050	3.109
	tur	**2.975**	2.812	2.882	2.899	2.896	2.723	2.824	2.851	2.896	2.971

- **Speaker similarity**
  - Seen vs. unseen
  - Increasing the number of utterances from the target speaker
  - zero-shot vs fine-tuning on unseen
- **Language identification**
  - Correlation between CER and LI
  - Paired-data fine-tuning improve LI
- **Predicted MOS (UT-MOS)**
  - Highest MOS values on zero-shot
  - Bad performance on multilingual

Summary

INTERSPEECH 2024

- This paper explores the language adaptation ability of **ZMM- TTS**, an **SSL-based** multilingual speech synthesis system.
- Experiments on **12 languages** with various **fine-tuning configurations** reveal the impact of **phonetic similarity** and language category on adaptation performance.
- Additionally, we find that the **fine-tuning dataset size** and **speaker diversity** influence adaptability.
- Surprisingly, using **paired data** for fine-tuning is not always optimal compared to **audio-only data.**
- Beyond speech intelligibility, our analysis covers **speaker similarity, language identification**, and **predicted MOS.**

Demo

INTERSPEECH 2024

Lang	bul	hrv	ces	nld	ita	jpn	kor	cmn	pol	rus	tur	vie
GT												
Syn												
Fine-tune method	0	L3′	L2	L1	L1′	L4′	L3′	S2′	L2	L1	L2′	S2

More demo

Code

Thanks for listening

Q&A

More demo

Code