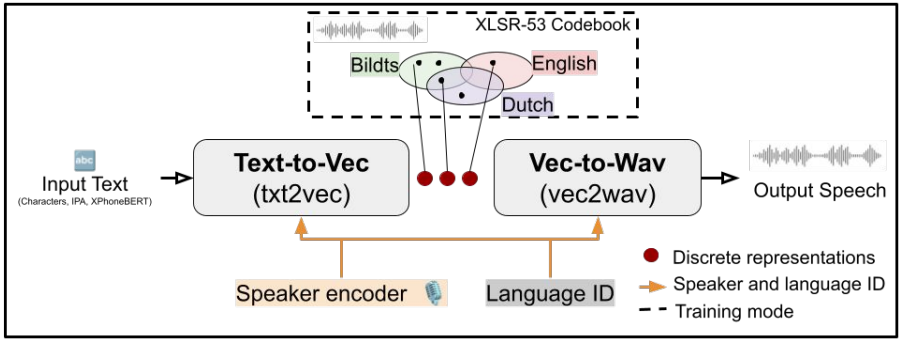
Submission from ZMM-TTS for Blizzard Challenge 2025

Jesujoba Alabi, Cheng Gong, Erica Cooper, Dietrich Klakow, Junichi Yamagishi

System Description

- Our system is built around ZMM-TTS
- ZMM-TTS is a modular system
  - **txt2vec:** text to discrete representations
  - **vec2wav:** discrete representation to waveform
  - both modules are independently trained
  - but combined for end-to-end inference
- txt2vec takes characters, IPA, or phoneme representations
- Discrete representations are extracted from XLSR-53's codebook



Schematic overview of the ZMM-TTS.

Dataset Used

- We got 7-hour speech dataset from the organizers
- To address the data sparsity
  - We considered languages included in the original training of ZMM-TTS
    - **English, French, German, Portuguese, and Spanish**
  - We include **Dutch** due to its linguistic proximity to Bildts
- For preprocessing: downsampling, amplitude normalization, speech not longer than 15s were used

Languages	Gender	Source	#Speakers	Dur (hrs)
Bildts	Male	Blizzard	1	7.0
Dutch	Female	CML-TTS	5	5.0
	Male		1	1.0
English	Female	LibriTTS	5	5.0
	Male		1	1.0
French	Female	CML-TTS	3	3.0
	Male		7	7.0
German	Female	CML-TTS	7	7.0
	Male		7	7.0
Portuguese	Female	CML-TTS	2	2.0
	Male		4	4.0
Spanish	Female	CML-TTS	7	2.0
	Male		7	7.0

Dataset statistics used in our experiments, including challenge and additional multilingual data.

Limitation of the Blizzard Lexicon

- The provided dictionary contains 7,756 entries
- The training corpus comprises 7,960 unique words
  - 6,296 are OOV
- We treated it as standard Dutch text
  - processed it directly using Epitran
- Compared the transcription Epitran with the reference annotations
  - phoneme error rate (PER) of 37.11%

Words	Epitran	Lexicon
later	laːtɛr	laːtər
nummer	nʏmmɛr	nömər
him	hɪm	hɪm
humor	hyːmər	hümɔr
dúdlik	dÝtlɪk	düdlək

The difference between the output of the official lexicon and the Epitran.

Implementation and Training

- We use the original implementation of ZMM-TTS
- Finetuned two variants
  - MLS - public model
    - trained on Multilingual LibriSpeech (MLS) corpus
  - GlobalPhone - private model
    - Trained on both MLS + GlobalPhone (GLB)
  - Both are originally multilingual, multispeaker TTS models
    - Covers English, French, German, Portuguese, Spanish, and Swedish
- Both modules of ZMM-TTS are finetuned for 50,000 steps
- Final checkpoint for inference

Experimental Setup

BH1 (MH1)

- What type of input is suitable for adaptation to Bildts?
- Does the base ZMM-TTS model used matter?

BZ1 (ZH1)

- Which improves Bildts TTS more: one close language or multiple?

Evaluation

- Mostly objective metrics
  - Speaker verification, WER, MCD
- Speaker verification - cosine similarity on representation from UniSpeech-SAT-Large
- WER - transcription using Whisper-Large-v3
- Automatic mean opinion scores
  - UTMOS, NISQA-MOS
- we setup a subjective evaluations with a native speaker on a small sample of utterances
  - On our test set and final test set

Result - BH1 (MH1)

What type of input is suitable for adaptation to Bildts?

- we listened to the outputs of character, IPA and XPhoneBERT inputs
- they all sound similar

Does the base ZMM-TTS model used matter?

- we listened to the outputs of character models of MLS and GLB
- they all sound similar too

We choose character input and used MLS for other experiments

Result - BZ1 (ZH1)

- Bilingual: We trained ZMM-TTS on the combination of Bildts and any of the other languages, six pairs
- Multilingual: We trained on all the languages
- Tested on both seen and unseen speakers

Seen Speaker Result

		SEC		WER	MCD		UTMOS		NISQA-MOS	
		Bildts	L1	L1	Bildts	L1	Bildts	L1	Bildts	L1
Bilingual	de	$0.92_{0.05}$	$0.97_{0.02}$	11.21	6.57	6.69	$3.28_{0.43}$	$2.96_{0.35}$	$4.12_{0.45}$	$4.15_{0.58}$
	en	$0.92_{0.05}$	$0.95_{0.04}$	397.52	6.50	6.99	$3.33_{0.37}$	$3.77_{0.44}$	$4.21_{0.42}$	$4.33_{0.52}$
	es	$0.92_{0.04}$	$0.98_{0.02}$	17.02	6.55	6.90	$3.28_{0.41}$	$2.57_{0.35}$	$4.11_{0.39}$	$4.14_{0.63}$
	fr	$0.92_{0.05}$	$0.98_{0.02}$	13.42	6.50	6.99	$3.20_{0.48}$	$2.72_{0.45}$	$4.02_{0.50}$	$4.21_{0.44}$
	nl	$\mathbf{0.93}_{0.04}$	$0.96_{0.03}$	27.02	**6.39**	6.85	$\mathbf{3.21}_{0.43}$	$2.54_{0.46}$	$\mathbf{4.02}_{0.43}$	$3.72_{0.53}$
	pt	$0.92_{0.05}$	$0.98_{0.02}$	15.02	6.40	6.21	$3.29_{0.39}$	$2.77_{0.43}$	$4.10_{0.41}$	$3.93_{0.58}$
Multilingual	de	$0.92_{0.05}$	$0.97_{0.02}$	11.05	6.74	6.80	$3.34_{0.35}$	$3.01_{0.38}$	$4.08_{0.53}$	$4.18_{0.57}$
	en		$0.94_{0.04}$	441.99		7.36		$3.80_{0.39}$		$4.34_{0.48}$
	es		$0.98_{0.03}$	18.87		6.99		$2.60_{0.36}$		$4.15_{0.66}$
	fr		$0.97_{0.02}$	12.75		7.02		$2.77_{0.44}$		$4.23_{0.45}$
	nl		$0.96_{0.03}$	27.02		7.36		$2.96_{0.37}$		$3.72_{0.54}$
	pt		$0.98_{0.02}$	14.40		6.28		$2.81_{0.47}$		$3.84_{0.49}$

Unseen Speaker Result

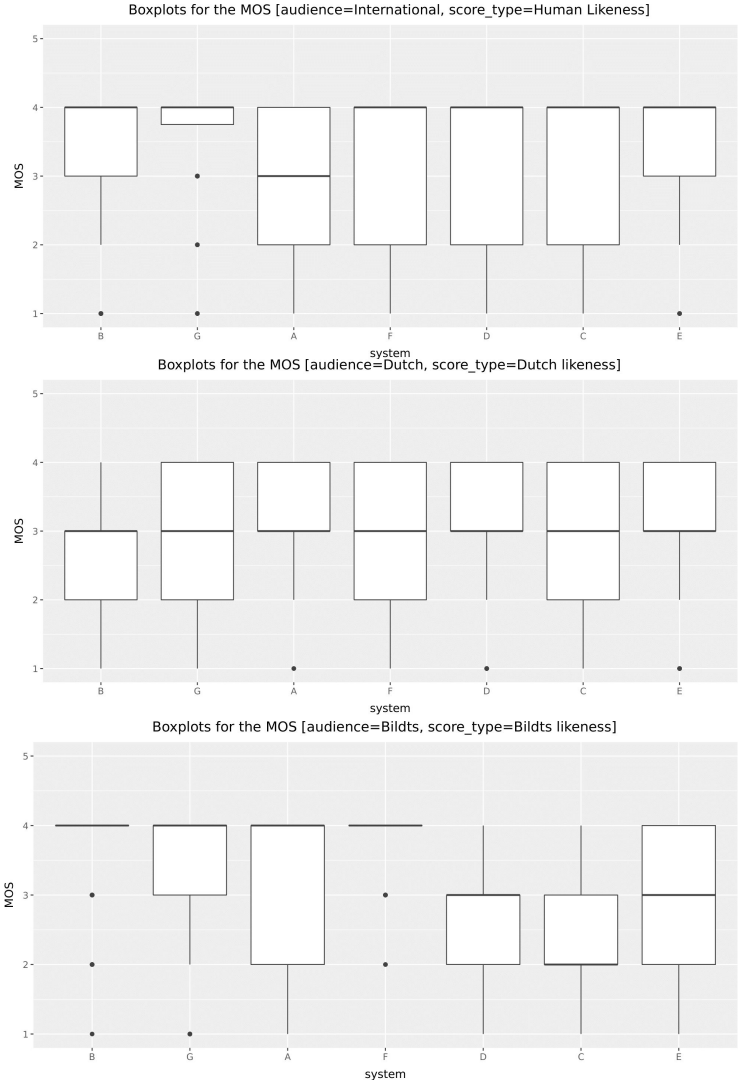
Languages		SEC	WER	MCD	UTMOS	NISQA-MOS
Bilingual	de	$0.94_{0.04}$	66.55	7.32	$3.31_{0.36}$	$4.16_{0.37}$
	en	$0.93_{0.04}$	96.62	7.17	$3.86_{0.38}$	$4.42_{0.38}$
	es	$0.95_{0.03}$	138.71	9.53	$2.98_{0.37}$	$4.10_{0.48}$
	fr	$0.94_{0.05}$	92.84	8.59	$2.73_{0.49}$	$3.90_{0.48}$
	nl	$0.93_{0.05}$	32.59	7.81	$2.44_{0.37}$	$3.92_{0.51}$
	pt	$0.88_{0.05}$	157.81	9.56	$2.64_{0.44}$	$3.75_{0.50}$
Multilingual	de	$0.97_{0.02}$	65.88	7.29	$3.34_{0.35}$	$4.23_{0.38}$
	en	$0.93_{0.04}$	86.60	7.23	$3.82_{0.37}$	$4.36_{0.38}$
	es	$0.95_{0.04}$	130.60	9.32	$3.06_{0.35}$	$4.10_{0.50}$
	fr	$0.94_{0.05}$	106.25	8.54	$2.81_{0.47}$	$4.00_{0.46}$
	nl	$0.95_{0.04}$	32.59	8.02	$2.88_{0.38}$	$4.40_{0.25}$
	pt	$0.90_{0.05}$	102.19	9.50	$2.71_{0.47}$	$4.08_{0.53}$

Final Submission

- For BH1 (MH1), supervised TTS, we submitted our multilingual system
- We randomly selected a single speaker embedding from the training split of the Bildts data and used it for all synthesis
- We converted number to words
  - num2words for Dutch
- We upsampled synthesized audio to 48,000 Hz using AudioSR
- **Our system is C**
- For BZ1 (ZH1), zero-shot TTS, we did submit our system
  - synthesized speech sound to the male speaker in the training data
  - We hypothesize it is as a result of language-speaker entanglement

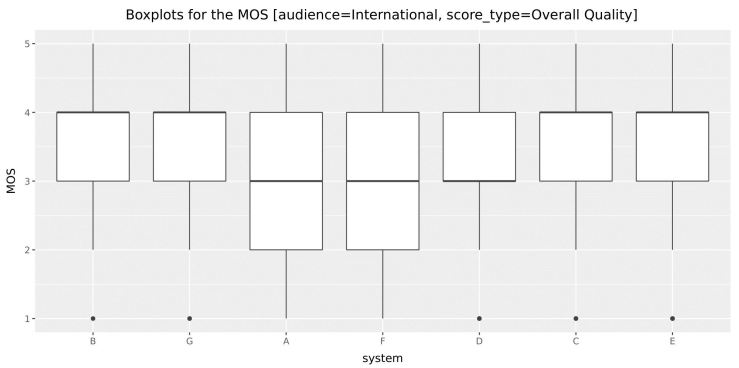
ZMM-TTS vs Other Systems

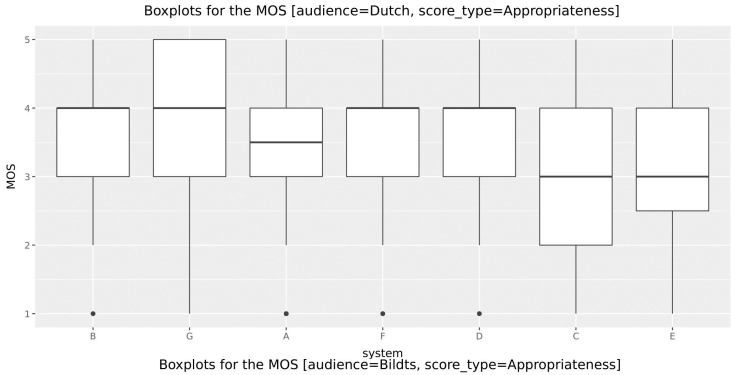
- For Human likeness and, and Dutch likeness, ZMM-TTS is comparable to other systems on average
- It is not competitive for Bildts likeness

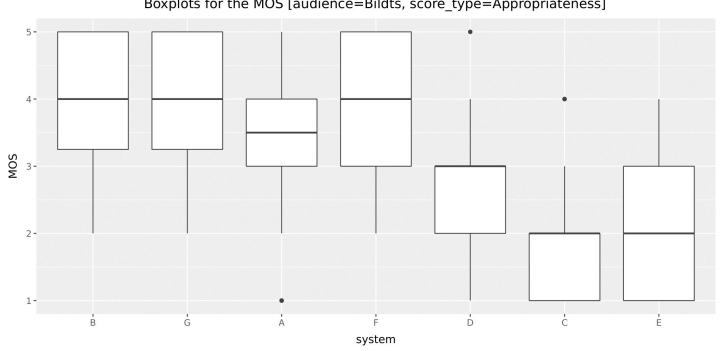


ZMM-TTS vs Other Systems

- For overall quality, international speakers rated the synthesized speech as competitive
- For appropriateness, Dutch speakers rated the synthesized speech as medium
- Bildts speakers rated it as bad







Based on our External Human Evaluator

- They were provided five examples across the evaluations
- No winner between text, IPA, and XPhoneBERT as input for txt2vec
- Our submitted outputs are rated as bad and poor
- We wait to get their explanation or justification for camera-ready

Sample Speech Submitted

- 'Here Jezus...', saai Sierd, sachys, maar niet soa sachys dat de man 't niet hore kon.
- De deuren fan de kafees stonnen open om de wêrmte bútten te houwen.
- D'r gebeurde trouwens niet alleen 'n prot an 'e Bildtse kust, maar seker ok in 't gemeentehuus fan 't Bildt.

They all obtained bad ratings (naturalness and appropriateness) from the human evaluator

Overall

- We present ZMM-TTS, our approach for the challenge
- Focused on answering some questions while trying to make ZMM-TTS work for Blizzard
- For zero-shot TTS, we observed speaker-language entanglement
- Our supervised TTS outputs, are Human-like, Dutch-like, but not Bildt-like

Thank you!