Submission from ZMM-TTS for Blizzard Challenge 2025

Jesujoba O. Alabi[1], *Cheng Gong*[2], *Erica Cooper*[3], *Yu Jiang*[4], *Dietrich Klakow*[1], *Junichi Yamagishi*[5]

[1]Saarland University, Saarland Informatics Campus, Germany
[2]Institute of Artificial Intelligence (TeleAI), China Telecom, China
[3]National Institute of Information and Communications Technology, Japan
[4]Tianjin University, China
[5]National Institute of Informatics, Japan
{jalabi,dietrich.klakow}@lsv.uni-saarland.de, jyamagis@nii.ac.jp

Abstract

We address the 2025 Blizzard Challenge for Bildts, a low-resource West Frisian language variety, using ZMM-TTS, a modular multilingual TTS model with separate text-to-vec and vec-to-waveform modules. We fine-tune the model on approximately 7 hours of Bildts data and explore the three input types supported by ZMM-TTS: raw characters, IPA characters, and phoneme representations. We also compare systems built by fine-tuning two pre-trained ZMM-TTS multilingual checkpoints. To enable multi-speaker synthesis for Bildts, we hypothesize that multilingual training will be beneficial; hence, we augment training with data from a selection of ZMM-TTS pre-training languages as well as geographically related languages (Dutch, German, English, French, Portuguese, and Spanish). Due to the lack of native speakers for evaluation, we rely primarily on objective metrics to select the final system.

Index Terms: Bildts, Text-to-speech, Low-resource, Multilingual

1. Introduction

Recent advancements in neural text-to-speech (TTS) systems have achieved remarkable performance in high-resource languages, enabling the generation of natural and intelligible speech from text [1, 2, 3, 4]. However, developing TTS models for low-resource languages remains a significant challenge due to the scarcity of high-quality audio and textual data [5], limiting accessibility and representation for many linguistic communities [6]. Considering these realities, we explore how high-quality TTS systems can be built with minimal reliance on extensive data or linguistic resources from the target language by leveraging multilingual frameworks and transfer learning, using datasets and pre-trained models that include multiple languages geographically close to the target.

In this work, we specifically address the 2025 Blizzard TTS challenge for Bildts, a low-resource West Frisian language variety spoken in the Netherlands, by leveraging ZMM-TTS [7], a multilingual, multi-speaker TTS model. ZMM-TTS consists of two independently trained modules: a text-to-discrete-representations model (txt2vec) and a discrete-representations-to-waveform model (vec2wav). Using adaptive fine-tuning, we customize the model for Bildts with a 7-hour dataset provided by the challenge committee. Given the limited availability of Bildts linguistic resources, we augment our training data with related and previously supported languages in the pre-trained ZMM-TTS, including Dutch, German, English, French, Portuguese, and Spanish. By combining data from these geographically and linguistically proximate languages, and comparing ZMM-TTS models trained on varying numbers of languages, we aim to optimize TTS performance for Bildts.

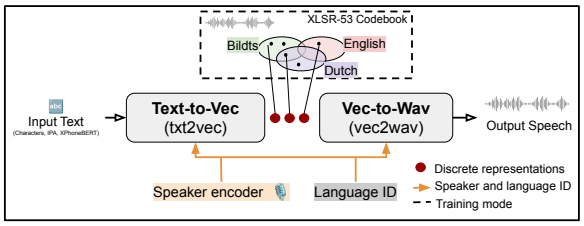


Figure 1: *Schematic overview of the ZMM-TTS.*

In our first experiment, we built three variants of the Bildts TTS system: a character-based model, an IPA-based model using an existing multilingual grapheme-to-phoneme (G2P) converter, and a model using a multilingual phoneme representation. As is common in low-resource settings, Bildts is not supported by either the multilingual G2P model or the multilingual phoneme representation model. In a second experiment, we evaluated the impact of different pre-trained models by building Bildts TTS systems based on two distinct ZMM-TTS checkpoints, and additionally trained variants that incorporated data from one or more geographically close languages with the aim of improving performance, particularly for zero-shot synthesis.

Evaluating TTS quality for low-resource languages is particularly challenging due to the difficulty of recruiting native listeners. Consequently, our team, composed primarily of non-speakers of Bildts, relied mainly on objective evaluation metrics to select the final system for submission.

2. System Description

In this section, we describe our TTS model, which is built using ZMM-TTS [7] through adaptive fine-tuning. ZMM-TTS is a multilingual, multi-speaker TTS system composed of two independently trained components that can be combined together for end-to-end inference as shown in Figure 1. These components are: (1) a text-to-discrete-speech-representations module (txt2vec), and (2) a discrete-speech-representations-to-waveform module (vec2wav).

The txt2vec module follows an encoder-decoder architecture. It takes text input in the form of characters, IPA phonemes, or pre-trained phoneme representations and converts it into discrete speech units. The vec2wav module then transforms these discrete speech representations directly into audio waveforms. At training time, the discrete representations are obtained from the codebook of XLSR-53 [8], a multilingual speech encoder

10.21437/Blizzard.2025-4

Lang.	Gender	Source	#Spk	Dur (hrs)	# Sent
Bildt	Male	Blizzard	1	7.0	6,006
Dutch (nl)	Female	CML-TTS	5	5.0	2,027
	Male		1	1.0	
English (en)	Female	LibriTTS	28	7.0	10,805
	Male		28	7.0	
French (fr)	Female	CML-TTS	3	3.0	3,834
	Male		7	7.0	
German (de)	Female	CML-TTS	7	7.0	5,150
	Male		7	7.0	
Portuguese (pt)	Female	CML-TTS	2	2.0	2,015
	Male		4	4.0	
Spanish (es)	Female	CML-TTS	7	7.0	4,839
	Male		7	7.0	

Table 1: *Dataset statistics used in our experiments for training and evaluating the ZMM-TTS models.*

that includes several high-resource languages, including Dutch. However, Bildts is not covered. Despite this, existing work has shown that ZMM-TTS performs well for low-resource languages [7, 9].

2.1. Data

The challenge committee provided a 7-hour dataset of the Bildts language, consisting of recordings from a single male speaker. This dataset includes 6,023 sentences along with their corresponding transcriptions. Bildts, a low-resource language, can be categorized under the "Left-Behind" group in Joshi's classification [10], meaning it has minimal or no digital linguistic resources available online.

To address this limitation, we consider using data from other languages geographically close to Bildts, specifically Dutch and German, as well as those previously used in training the original ZMM-TTS models. In total, we collected additional data for six languages: Dutch, English, French, German, Portuguese, and Spanish. Due to the ease of data collection, we used English data from LibriTTS [11], and sourced Dutch, French, German, Portuguese, and Spanish data from CML-TTS [12]. Although Dutch was not included in the original training of ZMM-TTS, we incorporate it due to its geographic closeness to Bildts. The remaining five languages were part of the original ZMM-TTS training corpus.

2.2. Data Preprocessing

We followed data pre-processing pipeline similar to those in [7], including resampling all audio to a 16 kHz sampling rate and applying amplitude normalization using SV56 [13]. However, we included only audio clips shorter than 15 seconds in our training dataset. Hence, we filtered out 17 utterances (311.0 seconds) from the shared Bildts data and sampled 10 minutes for in-house evaluation.

Table 1 highlights the statistics of the selected data per language. For English, we selected 28 speakers, each with 10 minutes of speech, balanced across both genders. For the other languages, we aimed to extract 70 minutes of speech per speaker, targeting seven speakers per gender. However, not all languages had sufficient data to meet this requirement.

For the seen-speaker evaluation, we randomly selected 10 minutes of speech per language, except for English, where we selected only 5 minutes, from the data listed in Table 1. Hence, for all languages except English, each speaker contributes approximately one hour of speech to the training data. For the unseen-speaker (zero-shot) test, we collected 25 minutes of audio per speaker per gender, with a maximum of two speakers per gender. For English, we used a threshold of 12.5 minutes of speech and doubled the number of speakers.

2.3. Limitations of the Blizzard Lexicon

Although the organizers provided a lexicon, its limitations hinder its direct applicability in our systems. The provided dictionary contains 7,756 entries. In contrast, the training corpus comprises 7,960 unique words, of which 6,296 are not covered by the lexicon, highlighting a substantial number of out-of-vocabulary words. Therefore, instead of using the dictionary as a grapheme-to-phoneme (G2P) resource, we treated it as standard Dutch text and processed it directly using Epitran [14].[1] To qualitatively assess phoneme-level consistency, we randomly selected five words from the training corpus and compared the IPA transcriptions generated by Epitran-Dutch with those provided in the official Bildts dictionary. The results are presented in Table 2.

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

Table 2: *The difference between the output of the official lexicon and the Epitran.*

Furthermore, we compared the phoneme transcriptions generated by Epitran with the reference annotations for all words covered by the dictionary, resulting in a phoneme error rate (PER) of 37.11%. Hence, we did not use the provided lexicon in our models; instead, where necessary, we used the Dutch configuration of Epitran for Bildts.

2.4. Implementation and Training

For our experiments, we used the original implementation of ZMM-TTS[2] and fine-tuned two variants of the pre-trained ZMM-TTS models. These include the MLS and GlobalPhone variants, both of which were originally trained on English, French, German, Portuguese, Spanish, and Swedish, using data primarily from the Multilingual LibriSpeech (MLS) corpus [15] and a combination of MLS and GlobalPhone (GLB) [16], respectively. The MLS checkpoint is publicly available, while the GlobalPhone checkpoint is private.[3] We fine-tune each module of ZMM-TTS for 50,000 steps on our collected dataset and use character inputs unless otherwise stated. The final checkpoint was used at inference time for synthesizing speech.

3. Experiment

The challenge includes two tasks: (1) supervised TTS, and (2) zero-shot synthesis of sentences using audio references from speakers unseen during training. We submited outputs only for

[1] https://github.com/dmort27/epitran

[2] https://github.com/nii-yamagishilab/ZMM-TTS

[3] The authors shared the checkpoint with us.

the first task but describe our setup and research questions for both tasks in the following sections.

3.1. Supervised TTS [BH]

This supervised TTS task involves building a speech synthesizer from publicly available data. Because Bildts is low-resource and absent from the original ZMM-TTS training, we fine-tuned the model on the provided Bildts data to extend its coverage, focusing on three questions.

What type of input is suitable for adaptation to Bildts? The ZMM-TTS model supports three types of input representations: characters, IPA, and XPhoneBERT [17] phoneme representation. However, the provided lexicon is limited, and Bildts was not included in XPhoneBERT pretraining. Therefore, we focus on investigating which input representation works best for Bildts. Hence, we trained three variants of ZMM-TTS model for Bildts using the three input representations.

Does the base ZMM-TTS model used matter? The ZMM-TTS model was trained on two datasets, MLS and GlobalPhone, which include the same set of languages. Their differing sizes led us to investigate whether base-model choice impacts adaptation to Bildts.

3.2. Zero-shot TTS [BS]

This task involves developing a speech synthesis system capable of generating speech for unseen speakers of Bildts. No training data was available for these speakers, and only a short reference speech sample was provided at test time for voice synthesis. To address this challenge, we employ language augmentation, incorporating data from other languages during the continued pre-training of the ZMM-TTS model. Although ZMM-TTS was originally designed to support unseen speakers, we aimed to further improve its performance through multilingual training. We chose language augmentation due to its potential to improve performance in low-resource speech synthesis. Specifically, augmenting training data with geographically and linguistically related languages can support shared phonetic and acoustic features, enable improved representation learning, and enhance the model's ability to generalize to unseen speakers. Hence, we focused on answering the question:

Which improves Bildts TTS more: one close language or multiple? To investigate this question, we included all languages originally used in ZMM-TTS, except Swedish, for which no data were available. We trained six bilingual systems, each combining Bildts with one of the remaining languages. We also examined a multilingual setup that jointly incorporated all available languages during training.

3.3. Evaluation

To evaluate the synthesized speech, we primarily used objective metrics, including speaker verification, word error rate (WER), and mel cepstral distance (MCD). For speaker verification, we extract speaker embeddings using UniSpeech-SAT-Large [18, 19] and computed cosine similarity. For MCD, we used the tool described in [20]. Also, we included automatic mean opinion scores such as UTMOS [21], and NISQA-MOS [22]. To calculate WER, we transcribed the speech using Whisper-Large-v3[4][5] and compared the transcriptions against the ground truth text. Lastly, in a few cases, we conducted qualitative evaluations with a native speaker on a small sample of utterances.

[4] `https://huggingface.co/openai/whisper-large-v3`

[5] We provided Whisper with the language names as generation parameter.

4. Results and Discussion

4.1. Supervised TTS: Input and Backbone Comparison

To evaluate the models trained on Bildts data, we compared the outputs of the MLS-based systems using three input variants: characters, IPA, and XPhoneBERT. As non-speakers, we listened to several examples from the test split but observed no clear differences among the variants. Furthermore, we compared models trained on character-based inputs using both the MLS and GLB pre-trained ZMM-TTS models. Our evaluation as non-speakers again indicates no significant differences in output quality. To ensure a more accurate assessment, we also shared the same samples with a native speaker at the end of the challenge. We provide a summary of their response in Section 4.6.

4.2. Supervised TTS: Bilingual vs Multilingual Augmentation

Table 3 presents metrics comparing the performance of ZMM-TTS trained on Bildts paired with another language, as well as in a multilingual configuration, where L1 corresponds to the language in each row. For both Bildts and the six paired languages, we observe that the speaker-similarity score (SEC) for synthesized Bildts speech remains nearly constant, ranging from 0.92 to 0.93, with the Dutch-paired model achieving the highest SEC at 0.93. In contrast, the SEC for synthesized speech in the other languages ranges from 0.95 to 0.98, consistently higher than for Bildts.

The Dutch model also yields the lowest MCD for Bildts at 6.39, followed closely by the Portuguese model at 6.40, whereas the remaining models exceed 6.50. For the other individual languages, MCD values range from 6.21 to 6.99, with English reaching the highest value at 6.99. However, when examining objective quality metrics such as UTMOS and NISQA-MOS, the bilingual English model performs best on Bildts. English also achieves the highest MOS scores among the languages, likely due to the backbone models being English-centric. The multilingual model exhibits similar patterns to the bilingual systems, producing comparable results and showing no consistent advantage on any metric.

The varying results on the other languages were also not informative for selecting a single best model. For example, we found the WER on English to be extremely high, exceeding 300%. From our observations, this is due to overgeneration and repetition of words and phrases during transcription by Whisper. This is likely caused by quality issues in the large English dataset, which included more speakers but less data per speaker in our TTS adaptation dataset.

Overall, no single metric consistently favors any specific language when paired alongside Bildts during training. These mixed results highlight the difficulty using currently existing objective metric in evaluating synthesized speech in low-resource languages. As non-speakers, we listened to a few sample cases and found no significant differences in output quality.

4.3. Zero-shot TTS

Table 4 show the metrics obtained when evaluating TTS models for in zero-shot setting in both bilingual and multilingual

	Languages	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	$0.92_{0.04}$	$0.98_{0.02}$	18.54	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	$3.21_{0.43}$	$2.54_{0.46}$	$4.02_{0.43}$	$3.72_{0.53}$
	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	$\mathbf{0.92}_{0.05}$	$0.97_{0.02}$	11.04	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		$0.98_{0.03}$	17.55		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		$0.96_{0.03}$	32.33		7.36		$2.96_{0.37}$		$3.72_{0.54}$
	pt		$0.98_{0.02}$	14.38		6.28		$2.81_{0.47}$		$3.84_{0.49}$

Table 3: *Evaluation result for seen speakers.*

	Lang.	SEC	WER	MCD	UTMOS	NISQA-MOS
Bilingual	de	$0.94_{0.04}$	67.18	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	$0.95_{0.03}$	126.21	9.53	$2.98_{0.37}$	$4.10_{0.48}$
	fr	$0.94_{0.05}$	97.11	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	$0.88_{0.05}$	164.14	9.56	$2.64_{0.44}$	$3.75_{0.50}$
Multilingual	de	$0.94_{0.04}$	67.25	7.29	$3.34_{0.35}$	$4.23_{0.38}$
	en	$0.93_{0.04}$	86.6	7.23	$3.82_{0.37}$	$4.36_{0.38}$
	es	$0.95_{0.04}$	120.84	9.32	$3.06_{0.35}$	$4.10_{0.50}$
	fr	$0.94_{0.05}$	115.57	8.54	$2.81_{0.47}$	$4.00_{0.46}$
	nl	$0.95_{0.04}$	26.18	8.02	$2.88_{0.38}$	$4.40_{0.25}$
	pt	$0.90_{0.05}$	112.93	9.50	$2.71_{0.47}$	$4.08_{0.53}$

Table 4: *Evaluation result for unseen speakers.*

settings across six languages. Since we do not have evaluation data for unseen speaker evaluation in Bildts, we focused only on these six languages. Our results show little difference in performance between the bilingual and multilingual models. However, when comparing unseen speakers to seen speakers, we observed a decrease in SEC and an increase in MCD. Meanwhile, UTMOS and NISQA-MOS scores remained similar to those of seen speakers, which could suggest that overall speech quality is maintained.

4.4. Final Submission

For the first subtask, we submitted outputs from our multilingual system, as we expected it to be more robust. To generate these, we randomly selected a single speaker embedding from the training split of the Bildts data and used it for all synthesis. Since we did not have a number-to-word converter for Bildts, we used the publicly available tool num2words[6] to convert numbers into Dutch, the closest available language. All texts were then synthesized using our TTS model. As the audio was originally synthesized at 16 kHz, we upsampled it to 48 kHz using AudioSR [23].

However, for the zero-shot task, the organizers provided reference audio for six speakers. We did not submit results for this task because, upon listening to the synthesized speech, it resembled the original training speaker more than the target voices. We hypothesize that this may be due to language-speaker entanglement or interference. This issue could potentially be addressed with more Bildts data, either additional hours of speech, a greater number of speakers, or an improved fine-tuning and adaptation regime. Future work should explore this further, particularly within the ZMM-TTS architecture.

[6] https://github.com/savoirfairelinux/num2words

4.5. Our System vs. Other Submitted Systems

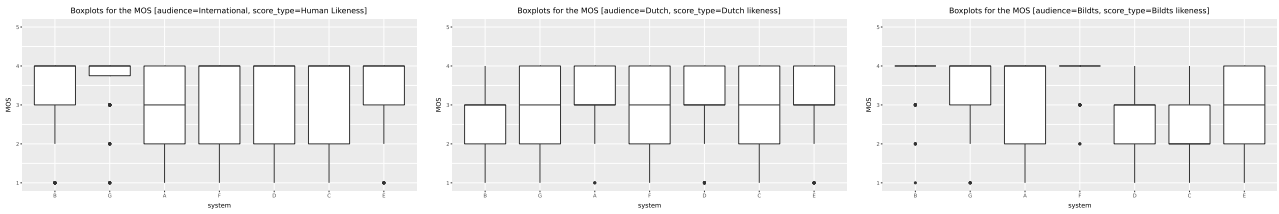
Based on the human evaluation results provided by the organizers for the supervised TTS setup, we observed that while the synthesized speech from our multilingual system sounds human-like and Dutch-like, it was not rated as Bildts-like by the evaluators. Figure 2 illustrates these findings by comparing our system (labeled C) with other submissions, showing ratings from international, Dutch, and Bildts speakers across human-likeness, Dutch-likeness, and Bildts-likeness.

Similarly, international speakers rated the synthesized speech as competitive with other systems in terms of overall quality. However, for appropriateness, Dutch speakers rated it as moderate, while Bildts speakers rated it low, indicating that the system is not competitive in appropriateness for Bildts speakers. Furthermore, Figure 4 shows the top 12 words most frequently mispronounced, as rated by Bildts speakers, across the seven systems. As can be seen, our system is among the models with the highest number of these mispronunciations, which may explain the lower ratings we received from the Bildts speakers. Since we are non-speakers and lack additional information from the organizers, we are unable to provide further interpretation of these results.

4.6. Qualitative Human Evaluation

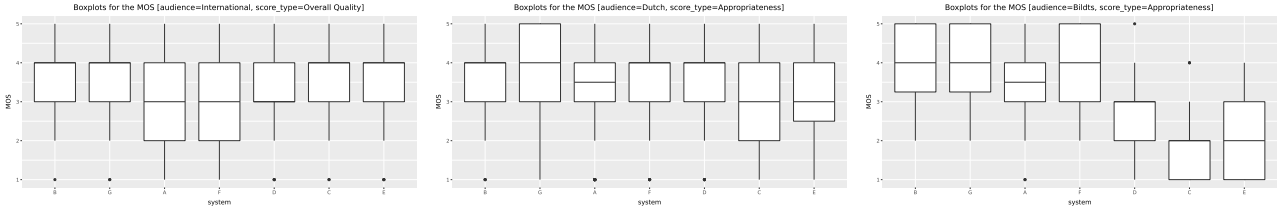
To better understand the performance of our system relative to other submissions, we conducted a small-scale qualitative human evaluation to collect additional feedback and more thoroughly assess the models we trained.

The human evaluation was carried out in two parts. In the first part (part A), the evaluator was asked to rate the naturalness and pronunciation of synthesized speech by comparing multiple examples and selecting the best; it was possible to select more than one. Five examples were selected from our test split of the 7-hour Bildts dataset provided to us. For the comparison of input types, characters, IPA, or XPhoneBERT, we provided synthesized speech at both 16 kHz and 48 kHz.



(a) *Human-likeness (International)* (b) *Dutch-likeness* (c) *Bildts-likeness*

Figure 2: *Human-likeness of the synthensized speech.*



(a) *Overall quality (International)* (b) *Appropriateness (Dutch)* (c) *Appropriateness (Bildts)*

Figure 3: *Overall quality and of appropriateness synthesized speech.*

At 16 kHz, the evaluator rated the IPA-based synthesis as more natural in 4 out of 5 examples and XPhoneBERT as more natural in 2 out of 5; similar trends were observed for pronunciation. At 48 kHz, XPhoneBERT was rated best in 3 out of 5 examples, characters were rated better in 1, and in one case no system was preferred.

Although this evaluation was small, it suggests that when adapting pre-trained models such as ZMM-TTS to low-resource languages, using representations such as phonemes leveraging multilingual tools and models such as Epitran and XPhoneBERT may be more effective than using characters. We do not have a complete explanation for the inconsistency between the ratings at 16 kHz and 48 kHz; however, these differences may reflect the effects of super-resolution on low-resource languages and should be further investigated.

We also compared synthesized speech from three character-based models: (1) a model trained only on the provided Bildts data, (2) a bilingual model trained on Bildts and Dutch, and (3) a multilingual model, all upsampled to 48 kHz. Using the same five examples as before, the evaluator rated naturalness and pronunciation across systems. The results showed no clear winner: the Bildts-only model was rated better once, the bilingual model once, the multilingual model once, and all the three systems were chosen equally good in two cases.

In the second part of the evaluation, the evaluator rated five examples from our final outputs submitted to the organizers for both naturalness and pronunciation, on a scale from 1 to 5 (with 5 being the best and 1 the worst). The results aligned with the organizers' evaluation of Bildts-likeness and overall quality. Our outputs were mostly rated 1 or 2, indicating poor quality; on average, naturalness scored 1.4 and pronunciation 1.2. These results highlight pronunciation errors as a major weakness in the synthesized speech. Future work will be needed to better understand the influence of multilingual training on low-resource languages like Bildts during adaptation.

Finally, when asked to compare the quality of Bildts speech synthesized from our in-house evaluation split (used in part A) of the provided dataset versus the final test set provided by the organizers, the evaluator noted: "The clarity and articulation of part A is better and less dull." However, across both parts, there were several cases of mispronunciation, poor articulation, and incomplete pronunciations, which negatively impacted our rankings, particularly those conducted by the native Bildts speaker.

5. Conclusion

In this paper, we present details of our submission to the Blizzard Challenge 2025, specifically for the supervised task [BH1 (MH1)], which focuses on Bildts, a low-resource language. Our submission is based on a modular architecture and a pretrained TTS model, ZMM-TTS. By adaptively fine-tuning a pretrained ZMM-TTS model, we explored several relevant aspects, including the choice of input format, the backbone model, and the effects of training on Bildts using either a closely related single language or a multilingual setup.

We compared the final outputs using several metrics. While we were unable to identify a single metric that reliably captured overall performance, we submitted our multilingual system. According to human evaluation experiments conducted by the challenge organizers, the system was rated as human-like and Dutch-like, but not Bildts-like, highlighting issues with its appropriateness for the target language.

6. Acknowledgements

We would like to thank Mr. Sytse Buwalda for providing valuable feedback on the speech samples that were shared. Jesujoba Alabi was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

7. References

[1] A. van den Oord *et al.*, "WaveNet: a generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, 2016.

[2] J. Shen *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," *ICASSP*, 2018.

[3] J. Kong *et al.*, "HiFi-GAN: Generative adversarial networks for

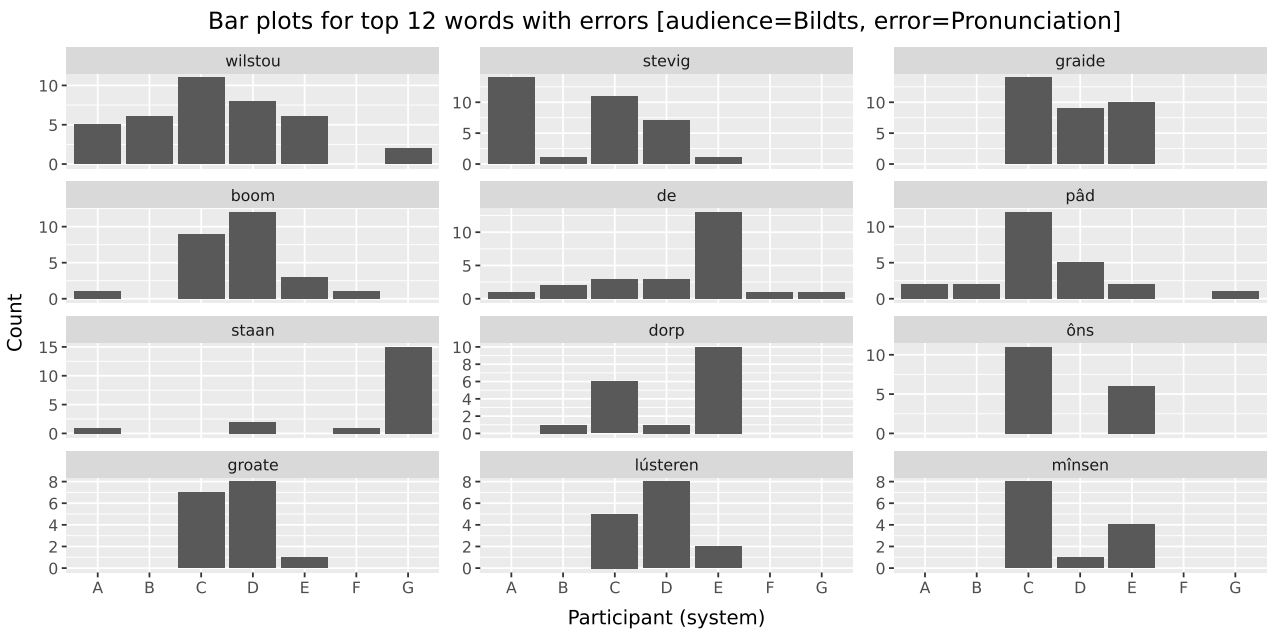


Figure 4: *Bar plots for top 12 words with errors.*

efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, 2020.

[4] D. Lim *et al.*, "JETS: Jointly training FastSpeech2 and HiFi-GAN for end to end text to speech," in *Interspeech*, 2022.

[5] T. Saeki *et al.*, "Learning to speak from text: Zero-shot multilingual text-to-speech with unsupervised text pretraining," in *32nd International Joint Conference on Artificial Intelligence, IJCAI 2023*. International Joint Conferences on Artificial Intelligence, 2023, pp. 5179–5187.

[6] P. Baljekar and A. W. Black, "Utterance selection techniques for TTS systems using found speech," *9th ISCA Speech Synthesis Workshop*, 2016.

[7] C. Gong *et al.*, "Zmm-tts: Zero-shot multilingual and multispeaker speech synthesis conditioned on self-supervised discrete speech representations," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 32, p. 4036–4051, Sep. 2024. [Online]. Available: https://doi.org/10.1109/TASLP.2024.3451951

[8] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning for speech recognition," in *Interspeech 2021*, 2021, pp. 2426–2430.

[9] C. Gong *et al.*, "Libritts: A corpus derived from librispeech for text-to-speech," in *Interspeech 2024*, 2024, pp. 4963–4967.

[10] P. Joshi *et al.*, "The state and fate of linguistic diversity and inclusion in the NLP world," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky *et al.*, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 6282–6293. [Online]. Available: https://aclanthology.org/2020.acl-main.560/

[11] H. Zen *et al.*, "Libritts: A corpus derived from librispeech for text-to-speech," in *Interspeech 2019*, 2019, pp. 1526–1530.

[12] F. S. Oliveira *et al.*, "Cml-tts: A multilingual dataset for speech synthesis in low-resource language," in *Text, Speech, and Dialogue*, K. Ekštein *et al.*, Eds. Cham: Springer Nature Switzerland, 2023, pp. 188–199.

[13] International Telecommunication Union, "Recommendation g.191: Software tools and audio coding standardization," https://www.itu.int/rec/T-REC-P.56/en, Nov. 2005, [Online; accessed 2025-08-28].

[14] D. R. Mortensen *et al.*, "Epitran: Precision G2P for many languages," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. C. C. chair) *et al.*, Eds. Paris, France: European Language Resources Association (ELRA), May 2018.

[15] V. Pratap *et al.*, "Mls: A large-scale multilingual dataset for speech research," in *Interspeech 2020*, 2020, pp. 2757–2761.

[16] T. Schultz *et al.*, "Globalphone: A multilingual text & speech database in 20 languages," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8126–8130.

[17] L. The Nguyen *et al.*, "Xphonebert: A pre-trained multilingual model for phoneme representations for text-to-speech," in *Interspeech 2023*, 2023, pp. 5506–5510.

[18] C. Wang *et al.*, "Unispeech: Unified speech representation learning with labeled and unlabeled data," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 10 937–10 947. [Online]. Available: http://proceedings.mlr.press/v139/wang21y.html

[19] S. Chen *et al.*, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," 2021.

[20] S. Taubert and J. Sternkopf, "mel-cepstral-distance," Apr. 2025. [Online]. Available: https://github.com/stefantaubert/mel-cepstral-distance

[21] Takaaki Saeki and Detai Xin and Wataru Nakata and Tomoki Koriyama and Shinnosuke Takamichi and Hiroshi Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Interspeech 2022*, 2022, pp. 4521–4525.

[22] G. Mittag and S. Möller, "Deep learning based assessment of synthetic speech naturalness," in *Interspeech 2020*, 2020, pp. 1748–1752.

[23] H. Liu *et al.*, "Audiosr: Versatile audio super-resolution at scale," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1076–1080.