Idiap kNN-TTS System for the Blizzard Challenge 2025

Enno Hermann[1], *Karl El Hajal*[1,2], *Ajinkya Kulkarni*[1], *Mathew Magimai.-Doss*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École polytechnique fédérale de Lausanne, Switzerland

{enno.hermann,karl.elhajal,ajinkya.kulkarni,mathew}@idiap.ch

Abstract

This paper describes Idiap's text-to-speech (TTS) system for the Blizzard Challenge 2025. The challenge targets synthesis for Bildts, a variety of Dutch with limited data availability, requiring effective low-resource approaches. We employ the kNN-TTS framework, training a Glow-TTS model to predict self-supervised speech representations instead of spectrograms on only the seven hours of provided Bildts data. Multi-speaker synthesis is achieved through kNN-based voice conversion using WavLM features and a pre-trained HiFi-GAN vocoder trained on English data. For improved zero-shot voice cloning with minimal reference audio, we augment target speaker data using OpenVoice conversion.

Index Terms: speech synthesis, voice cloning, low-resource

1. Introduction

The advancement of text-to-speech (TTS) synthesis has been marked by training datasets growing to tens of thousands [1, 2, 3] or even millions of hours of audio [4, 5], delivering highly natural and expressive output. However, this data-intensive paradigm is not applicable for low-resource languages, where such large speech corpora are not available or too expensive to create. In addition, the compute requirements for large-scale models are not accessible to everyone. This disparity between high- and low-resource languages represents a critical challenge in democratising speech technology.

The Blizzard Challenge 2025 addresses this with the aim of developing TTS models for Bildts, a variety of Dutch spoken in Het Bildt in the Dutch province of Friesland. The challenge provides seven hours of single-speaker training data and organises a subjective listening test to evaluate synthesis and zero-shot voice cloning performance of submitted systems.

Self-supervised learning (SSL) representations have opened new avenues for addressing data scarcity in speech synthesis. SSL models capture rich phonetic and acoustic representations that can potentially bridge the gap between high-resource foundation models and low-resource target applications. However, the effective integration of SSL representations into TTS frameworks for low-resource scenarios remains underexplored.

The kNN-TTS [6] framework represents a promising approach to this challenge by decoupling linguistic content modelling from speaker-specific voice characteristics. The architecture combines a text-to-SSL model that can be trained on a single speaker's data with kNN-based voice conversion (VC) [7] to enable multi-speaker synthesis. Although kNN-VC has demonstrated cross-lingual capabilities [8], kNN-TTS has so far only been evaluated on English [6].

A downside of kNN-VC and kNN-TTS is that at least 30 seconds of speech are required to provide sufficient phoneme coverage for voice cloning [7]. Many other recent systems claim to be able to clone from as little as 5–10 seconds [1, 2, 5].

This work investigates three primary research questions: (1) Can kNN-TTS maintain synthesis quality when trained exclusively on a small amount of Bildts data? (2) How effectively do SSL representations trained on high-resource languages transfer to low-resource dialect synthesis? (3) What strategies can overcome the reference audio limitations inherent in kNN-TTS for practical zero-shot applications?

To address these questions, we develop and evaluate a Glow-TTS-based [11] kNN-TTS system trained only on the provided Bildts corpus, while leveraging WavLM [9] representations and a HiFi-GAN [10] vocoder trained on larger, untranscribed English datasets. We further compare different approaches to reduce the reference audio requirements for zero-shot voice cloning since the challenge only provides a single sentence from each target speaker. We find that augmentation with OpenVoice conversion [19] produces the most natural and intelligible output, while maintaining reasonable speaker similarity.

The remainder of this paper is structured as follows: Section 2 provides technical background on the kNN-TTS framework. Section 3 details our experimental setup including dataset preparation and model configuration. Section 4 presents comprehensive subjective and objective evaluation results before Section 5 summarises and concludes the paper.

2. Background

kNN-VC [7] is a simple but effective any-to-any voice conversion method that relies on SSL representations where similar features also have similar phonetic properties. While an external SSL model and vocoder are required, the voice conversion itself is non-parametric.

We previously extended this approach to TTS [6] by modifying conventional TTS architectures, for example Glow-TTS [11] or Grad-TTS [12], to predict SSL features instead of Mel spectrograms. Crucially, this text-to-SSL model is the only component that is trained on transcribed audio data, which may be from a single speaker. Multi-speaker output is achieved by combining this model with kNN-VC.

As illustrated in Figure 1, the kNN-TTS pipeline operates as follows: (1) a text-to-SSL model generates source speaker features from input text; (2) a kNN retrieval algorithm matches these features to units in a target speaker database containing SSL features; and (3) a pre-trained vocoder synthesises the final waveform.

SSL encoder: We need an intermediate audio representation that meets the following criteria: (1) it should encode both linguistic and speaker-specific information; (2) features close in the embedding space should exhibit similar phonetic properties

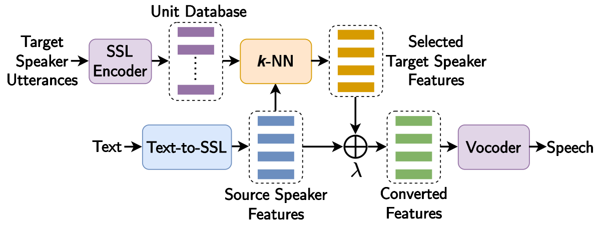

Figure 1: *kNN-TTS framework overview [6]. Only the Text-to-SSL model is trained on transcribed audio. The SSL encoder, vocoder are pre-trained on untranscribed multi-speaker data, and the kNN algorithm is non-parametric.*

while preserving speaker identity; and (3) it should be possible to decode the features back to waveform. Recent works show that SSL models encode speech into such representations [13].

Text-to-SSL: We train a Text-to-SSL model that generates corresponding SSL features from a given text input. Notably, this is the only component of our framework that requires audio data paired with text transcriptions for training. It is possible to train this model on the speech of a single speaker.

kNN Retrieval: To synthesise speech in a target speaker's voice, units (or frames) from the target speaker unit database are selected to replace corresponding frames from the source speaker features. The selection is done by comparing source and target frames using a linear distance metric. This results in selected target speaker features that maintain the phonetic information while replacing the voice attributes with those of the target speaker. Due to the nature of this process, the target audio should provide sufficient phoneme coverage to avoid artefacts. The selected target speaker features are linearly interpolated with the source speaker features to obtain the converted features:

$$y_{\text{converted}} = \lambda\, y_{\text{selected}} + (1-\lambda)\, y_{\text{source}} \qquad (1)$$

Finally, a pre-trained vocoder decodes the converted features back into a speech waveform.

3. Experimental setup

3.1. Datasets

For the Blizzard Challenge 2025, 7 hours of Bildts data from a single male speaker are provided for training.[1] It is sampled at 48 kHz and has transcriptions and text-normalised transcriptions in Praat TextGrid format. We train on the latter and also do not apply any text normalisation at test time. We split the training data into individual utterances based on the provided segment start and end timestamps and downsample to 16 kHz to match the sampling rate of WavLM. While a pronunciation dictionary for training grapheme-to-phoneme conversion models is also provided, we train grapheme-based models for simplicity.

For more detailed evaluations, we additionally train English models on the LJSpeech [14] and LibriTTS-R [15] datasets. These are trained with Espeak phonemes for consistency with previous work [6].

3.2. TTS and vocoder models

SSL encoder: We select the pre-trained WavLM-Large checkpoint[2] [9] for its effective audio reconstruction capabilities, obtained through training with masked speech denoising and prediction tasks on 94k hours of untranscribed English speech. Consistent with previous works [6, 7], we choose features from the 6th layer, which encode both phonetic and speaker characteristics [16]. These representations can be pre-extracted and cached before training and inference, eliminating the need to load WavLM during either process if the target speaker is known.

Text-to-SSL: Following [6], we train a Glow-TTS model with the Coqui TTS toolkit[3] for 600k steps at batch size 32 on the Bildts data. Glow-TTS is based on a non-autoregressive architecture with a transformer-based text encoder, a duration predictor, and a flow-based decoder [17]. We maintain the default configurations and cost functions for training, only adjusting the output dimension to 1024 channels to match the WavLM-Large features instead of Mel spectrograms. The training recipe and data preparation scripts are available on Github. [4]

For a more comprehensive evaluation, we additionally train three English models: (1) *EN-7h* trained on a 7-hour subset of LJSpeech to match the amount of Bildts training data; (2) *EN-24h* trained on the full 24 hours of LJSpeech; (3) *EN-124h*, a multi-speaker model trained on LJSpeech and the train-clean-100 subset of LibriTTS-R. The latter is trained for 1.2M steps to adjust for the larger amount of data.

kNN Retrieval: For each WavLM source frame, we compute its cosine distance with every target speaker frame within the unit database. We then select the k closest units, and average them with uniform weighting. Similar to [7], we use $k = 4$ which was determined to be suitable across different amounts of target audio.

Vocoder: We use a pre-trained HiFi-GAN V1 [10] model trained to reconstruct 16 kHz waveforms from WavLM-Large layer 6 features. The model checkpoint, sourced from [7], was trained on the LibriSpeech train-clean-100 set, consisting of 100 hours of English speech from 251 speakers [18]. We use the prematched version, where kNN regression was also applied at training time to better match how it is used for inference.

3.3. Zero-shot voice cloning

The provided evaluation data only has one sentence (7–12 seconds) of reference speech for each of the six target speakers for voice cloning. However, due to the nature of the retrieval algorithm, kNN-VC and kNN-TTS require 30 seconds to 5 minutes of reference data to provide sufficient phoneme coverage and produce high-quality outputs [7]. We investigate different ways to generate additional reference data to mitigate this:

- Synthesise additional sentences in Dutch or English from the *North Wind and the Sun* passage with XTTS [1] and the provided reference audio as target speaker. Dutch is the most closely related to Bildts, so should provide at least a baseline phoneme coverage.
- Convert the Bildts training data to the target speaker with the OpenVoice v2 conversion model [19].
- Replace the kNN retrieval algorithm with MKL-VC, a recently proposed factorised optimal transport formulation [20]. It claims to only require 5 seconds of reference speech by matching distributions instead of performing frame selection. We choose $K = 256$ for increased speaker similarity at the cost of intelligibility.

We also evaluate these methods on the English 7-hour

[1] https://zenodo.org/records/14792457

[2] https://github.com/microsoft/unilm/tree/master/wavlm

[3] https://github.com/idiap/coqui-ai-TTS

[4] https://github.com/idiap/knn-tts/tree/blizzard2025

model, with one sentence from each of the 39 speakers from the LibriTTS-R test-clean subset as reference audio.

3.4. Evaluation

3.4.1. Subjective evaluation

A comprehensive subjective listening test was organised by the Blizzard Challenge 2025. The evaluation was conducted online with three listener groups: Bildts speakers, Dutch/Frisian speakers, and international speakers. 7 teams participated in the hub task (BH1) and 5 teams also in the optional zero-shot task (BS1). Our submission is denoted as system D and highlighted with an asterisk in figures. In this paper, we only present the most relevant results of the evaluation.

For the listening test, utterances had to be synthesised at the sentence, paragraph and full text level. Our system was trained on single sentences only.

Descriptive statistics are provided for each evaluation metric. Wilcoxon signed-rank tests with Bonferroni correction were performed for pairwise system comparisons to determine statistically significant differences ($p < 0.05$).

3.4.2. Objective evaluation

We additionally conduct our own objective evaluations for model selection and analysis using the VERSA toolkit [21]. We measure naturalness with UTMOSv2 [22], speaker embedding cosine similarity with ESPnet's `voxcelebs12_rawnet3` model [23] and word and character error rate (WER/CER) with the Whisper `large` model [24]. We do not evaluate intelligibility for Bildts due to lack of a suitable speech recognition model. For Bildts, we only include the individual sentences in the objective evaluation. For English, we synthesise the same 240 sentences from the *devtest* subset of FLORES+ [25] as XTTS [1].

4. Results

Section 4.1 presents our results on the single speaker hub task (BH1) and section 4.2 on the optional zero-shot voice cloning task (BS1).

4.1. Single speaker TTS (BH1)

The aim of this task was to build a voice from the provided 7 hours of Bildts speech, using only publicly available data. We did not train on any additional transcribed data, although the WavLM and vocoder models were trained on untranscribed English speech.

4.1.1. Objective evaluation

We first evaluate the effect of the reduced training data on kNN-TTS in a low-resource setting. The original kNN-TTS model [6] was trained on 24 hours of English speech. Table 1 shows a slight decrease in intelligibility while maintaining naturalness if we reduce the data to 7 hours, matching the available data for Bildts.

Adding another 100 hours from other speakers actually leads to worse results when evaluated only on the LJSpeech speaker. We note that training on multiple speakers is not required because the kNN retrieval algorithm enables voice cloning even when trained on a single speaker. We also apply kNN retrieval for single-speaker synthesis for all models in Table 1 and they all achieve comparable speaker similarity. These results indicate that more Bildts data would only lead to marginal improvements.

Table 1: *Single-speaker objective evaluation results measuring naturalness (UTMOSv2), speaker similarity (SSIM) and intelligibility (WER/CER).*

Model	UTMOSv2 ↑	SSIM ↑	WER ↓	CER ↓
Bildts-7h	3.26	0.49	-	-
EN-7h	3.86	0.54	4.09	1.70
EN-24h	3.80	0.53	3.41	1.31
EN-124h	3.56	0.50	4.63	1.97

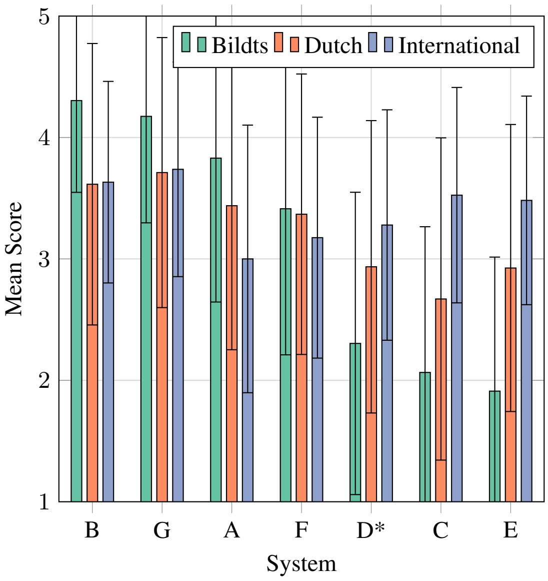

Figure 2: *Speech quality evaluation across different audiences using 5-point scales. Bildts speakers (green) and Dutch speakers (orange) rated naturalness given audio and text, while international speakers (blue) rated overall quality from the audio alone. Higher scores indicate better quality. Our system is always indicated with an asterisk (D*).*

4.1.2. Subjective evaluation

The main task evaluates synthesised Bildts speech using various perceptual measures across different listener populations: 28 Bildts (green), 100 Dutch (orange) and 50 English (blue) speakers.

Naturalness and Quality Evaluation: Figure 2 shows naturalness ratings by Bildts and Dutch listeners (5-point MOS scale: 1=Completely unnatural to 5=Completely natural) and overall quality ratings by international listeners (5-point scale: 1=Bad to 5=Excellent). While international listeners do not observe significant differences in overall quality, Bildts and Dutch speakers find our proposed system (D*) less natural than most others.

Language Variety Evaluation: Figure 3 shows how different audiences perceive the dialectal characteristics of synthesised speech. Bildts speakers rated Bildts likeness (4-point scale: 1=Not Bildts to 4=Bildts), while Dutch speakers rated Dutch likeness (4-point inverted scale: 1=Dutch to 4=Not standard Dutch). Although outputs of other systems are judged to be closer to Bildts, interestingly Dutch speakers find ours to be among those that sound the least like standard Dutch. A possible explanation could be that other systems also included Dutch training data.

International Listener Evaluations: Figure 4 shows com-

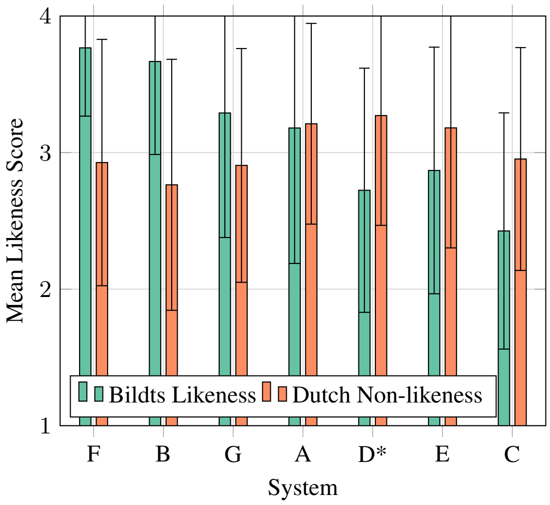

Figure 3: *Language variety evaluation across audiences using 4-point scales. Bildts speakers (green) rated Bildts likeness (1=Not Bildts to 4=Bildts), while Dutch speakers (orange) rated Dutch likeness (1=Dutch to 4=Not standard Dutch, inverted to show non-Dutch-likeness). Higher scores might indicate stronger dialectal characteristics.*

prehensive perceptual evaluation by fluent English speakers using audio-only presentations. These listeners had diverse Bildts proficiency, with 43 having no knowledge, 13 claiming native proficiency, and 13 having passive knowledge. Three dimensions were evaluated: listening effort, voice pleasantness, and human-likeness. Among these, only the difference to system C in terms of voice pleasantness is statistically significant in the pairwise comparisons.

4.2. Zero-shot voice cloning (BS1)

The zero-shot task evaluates voice cloning capabilities on six unseen Bildts speakers given one sentence of reference audio.

4.2.1. Objective evaluation

Table 2 compares different approaches for improved zero-shot voice cloning with kNN-TTS. Bildts results are on the official evaluation data and for English we synthesise 100 sentences from FLORES+ with the EN-7h model for each of the 39 speakers of the LibriTTS-R test-clean subset.

We find that replacing kNN-VC with MKL-VC [20] does not preserve the speaker identity very well and also decreases intelligibility and naturalness. Generating additional Dutch or English reference audio with XTTS performs similarly to just using the original reference sentence in terms of naturalness and speaker similarity. As expected, for English we observe that XTTS and OpenVoice augmentation significantly improve intelligibility, even outperforming the top-line that uses all available reference audio for the speaker instead of just a single sentence. While converting the training data to the target speaker with OpenVoice as additional reference audio results in slightly lower speaker similarity, it produces the most natural and most intelligible output. We therefore choose OpenVoice augmentation for our final submission.

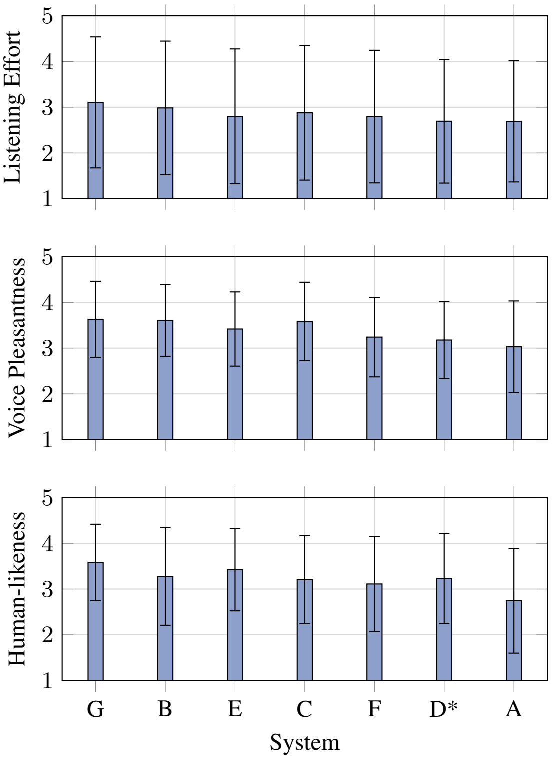

Figure 4: *Fluent English listener ratings using 5-point scales for listening effort (1=No meaning understood to 5=Complete relaxation possible), voice pleasantness (1=Very unpleasant to 5=Very pleasant), and human-likeness (1=Machine-like to 5=Human-like). Higher scores indicate better performance for all metrics.*

4.2.2. Subjective evaluation

Generated samples for one male (HJH) and one female (RH) speaker were used in the subjective listening test. 5 systems participated in this task, with ground truth references included as controls. Listeners evaluated Bildts likeness and speaker similarity.

Bildts Likeness: 22 native or fluent Bildts speakers evaluated how close the synthesised speech is to Bildts using the same 4-point scale as the main task. Figure 5 shows that our system

Table 2: *Zero-shot voice cloning objective evaluation results for Bildts/EN-7h, measuring naturalness (UTMOSv2), speaker similarity (SSIM) and intelligibility (WER/CER). Outputs were generated with only the original reference audio, using MKL-VC instead of kNN-VC, or reference audio augmented with XTTS or OpenVoice. For EN-7h we also compute a top-line performance with all available reference data instead of just a single sentence.*

Approach	UTMOSv2 ↑	SSIM ↑	WER ↓	CER ↓
Original	2.71 / 3.38	**0.37** / **0.44**	- / 12.1	- / 6.5
XTTS	2.85 / 3.54	0.36 / **0.44**	- / 5.0	- / 2.1
OpenVoice	**3.13** / **3.77**	0.17 / 0.28	- / **3.6**	- / **1.4**
MKL-VC	2.12 / 3.23	0.08 / 0.30	- / 16.3	- / 12.1
Top-line	- / 3.68	- / 0.42	- / 6.6	- / 3.6

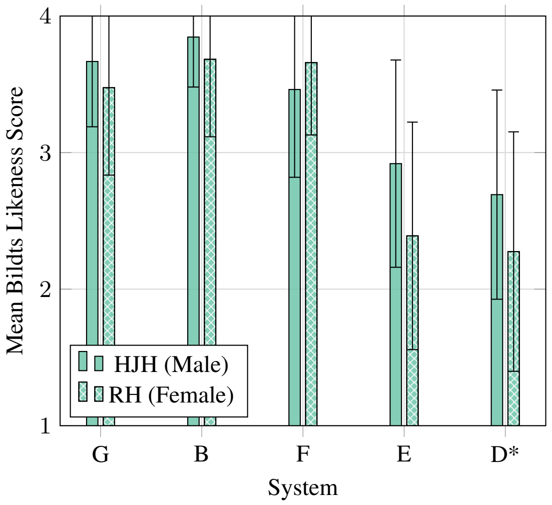

Figure 5: *Bildts likeness evaluation by native Bildts speakers for HJH (Male, solid bars) and RH (Female, hatched bars) reference speakers. Scale: 1=Not Bildts, 2=Somewhat Bildts, 3=Mostly Bildts, 4=Bildts.*

(D*) and system E are judged less Bildts-like than the others.

Speaker Similarity (MUSHRA-like): 60 Dutch and 50 international listeners compared synthesised speech to reference speakers using a continuous 0–100 scale (0–19=Different speaker, 20–39=Probably different, 40–59=Similar speaker, 60–79=Probably same, 80–100=Same speaker). Listeners first familiarised themselves with 3 ground-truth samples, then rated 4 non-identified samples including 1 hidden reference and 5 TTS systems. Listeners who failed to give the hidden reference sample a score of 100 were excluded from the results. Text was not provided for this evaluation. As Figure 6 illustrates, together with system F, our kNN-TTS model is worse at matching the speaker identity than others. To a certain degree this is expected because kNN-TTS is not optimised for voice cloning from such short reference audio and as shown earlier, the OpenVoice augmentation also reduces speaker similarity.

5. Conclusions

This paper presents Idiap's kNN-TTS system for the Blizzard Challenge 2025 to synthesise Bildts speech from limited training data. The kNN-TTS framework proves viable for low-resource synthesis with only 7 hours of transcribed training data, leveraging SSL representations to transfer knowledge from larger English datasets. It achieves middle-tier performance among participating systems. While OpenVoice augmentation improves naturalness and intelligibility for zero-shot voice cloning from a single sentence, speaker similarity remains challenging with minimal reference audio, highlighting the trade-off between data requirements and cloning quality.

Future work should explore more effective reference audio augmentation strategies and investigate the application of kNN-TTS to other low-resource settings. Cross-lingual performance could be improved by adopting multilingually trained SSL models and vocoders.

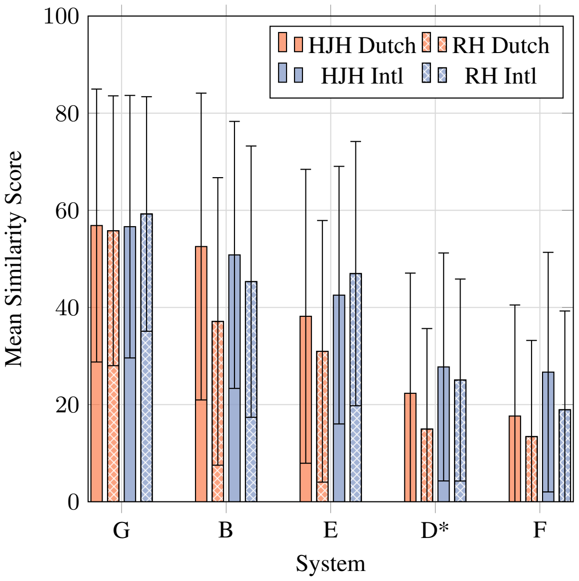

Figure 6: *MUSHRA similarity mean scores and standard deviation comparing HJH (Male, solid bars) and RH (Female, hatched bars) reference speakers across Dutch (orange) and international listeners (blue).*

6. Acknowledgements

This work was partially supported by the Swiss National Science Foundation grant agreement no. 219726 on "Pathological Speech Synthesis (PaSS)" and the Innosuisse flagship grant agreement no. PFFS-21-47 on "Inclusive Information and Communication Technologies (IICT)".

7. References

[1] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi *et al.*, "XTTS: a massively multilingual zero-shot text-to-speech model," in *Proc. Interspeech*, 2024, pp. 4978–4982.

[2] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers," 2023. [Online]. Available: https://arxiv.org/abs/2301.02111

[3] Z. Ye, X. Zhu, C.-M. Chan, X. Wang, X. Tan, J. Lei, Y. Peng, H. Liu, Y. Jin, Z. Dai, H. Lin, J. Chen, X. Du, L. Xue, Y. Chen, Z. Li, L. Xie, Q. Kong, Y. Guo, and W. Xue, "Llasa: Scaling train-time and inference-time compute for Llama-based speech synthesis," 2025. [Online]. Available: https://arxiv.org/abs/2502.04128

[4] O. Atamanenko, A. Chalova, J. Coombes, N. Cope, P. Dang, Z. Deng, J. Du, M. Ermolenko, F. Fan, Y. Feng, C. Fichter, P. Filimonov, L. Fischer, K. Gibbs, V. Gusarova, P. Karpik, A. A. Kottner, I. Lee, O. Louie, J. Mai, M. Mamontov, S. Mao, N. Morshed, I. Poletaev, F. Radu, D. Semernia, E. Shingarev, V. Sivaraja, P. Skirko, R. Takhautdinov, R. Villahermosa, and J. Wang, "TTS-1 technical report," Inworld, Tech. Rep., 2025. [Online]. Available: https://arxiv.org/abs/2507.21138

[5] KimiTeam, D. Ding, Z. Ju, Y. Leng, S. Liu, T. Liu, Z. Shang, K. Shen, W. Song, X. Tan, H. Tang, Z. Wang, C. Wei, Y. Xin, X. Xu, J. Yu, Y. Zhang, X. Zhou, Y. Charles, J. Chen, Y. Chen, Y. Du, W. He, Z. Hu, G. Lai, Q. Li, Y. Liu, W. Sun, J. Wang, Y. Wang, Y. Wu, Y. Wu, D. Yang, H. Yang, Y. Yang, Z. Yang, A. Yin, R. Yuan, Y. Zhang, and Z. Zhou, "Kimi-Audio technical report," Kimi, Tech. Rep., 2025. [Online]. Available: https://arxiv.org/abs/2504.18425

[6] K. E. Hajal, A. Kulkarni, E. Hermann, and M. Magimai Doss, "kNN retrieval for simple and effective zero-shot multi-speaker text-to-speech," in *Proc. NAACL*, 2025, pp. 778–786.

[7] M. Baas, B. van Niekerk, and H. Kamper, "Voice conversion with just nearest neighbors," in *Proc. Interspeech*, 2023, pp. 2053–2057.

[8] M. Baas and H. Kamper, "Voice conversion for stuttered speech, instruments, unseen languages and textually described voices," in *Proc. Artificial Intelligence Research*, 2023, pp. 136–150.

[9] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1505–1518, 2022.

[10] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, 2020, pp. 17 022–17 033.

[11] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: a generative flow for text-to-speech via monotonic alignment search," in *Proc. NeurIPS*, 2020.

[12] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A diffusion probabilistic model for text-to-speech," in *Proc. ICML*, 2021.

[13] E. Dunbar, N. Hamilakis, and E. Dupoux, "Self-supervised language learning from raw audio: Lessons from the zero resource speech challenge," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1211–1226, 2022.

[14] K. Ito and L. Johnson, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[15] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, and A. Bapna, "Libritts-r: A restored multi-speaker text-to-speech corpus," in *Proc. Interspeech*, 2023, pp. 5496–5500.

[16] S. Wang, G. E. Henter, J. Gustafson, and E. Szekely, "On the Use of Self-Supervised Speech Representations in Spontaneous Speech Synthesis," in *Proc. 12th ISCA Speech Synthesis Workshop (SSW)*, 2023.

[17] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Proc. NeurIPS*, 2018.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[19] Z. Qin, W. Zhao, X. Yu, and X. Sun, "OpenVoice: Versatile instant voice cloning," Myshell.ai, Tech. Rep., 2023. [Online]. Available: https://arxiv.org/abs/2312.01479

[20] A. Lobashev, A. Yermekova, and M. Larchenko, "Training-free voice conversion with factorized optimal transport," in *Proc. Interspeech*, 2025, pp. 1373–1377.

[21] J. Shi, H. jin Shim, J. Tian, S. Arora, H. Wu, D. Petermann, J. Q. Yip, Y. Zhang, Y. Tang, W. Zhang, D. S. Alharthi, Y. Huang, K. Saito, J. Han, Y. Zhao, C. Donahue, and S. Watanabe, "VERSA: A versatile evaluation toolkit for speech, audio, and music," in *Proc. NAACL*, 2025, pp. 191–209.

[22] K. Baba, W. Nakata, Y. Saito, and H. Saruwatari, "The T05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech," in *Proc. SLT*, 2024, pp. 818–824.

[23] J. weon Jung, W. Zhang, J. Shi, Z. Aldeneh, T. Higuchi, A. Gichamba, B.-J. Theobald, A. Hussen Abdelaziz, and S. Watanabe, "ESPnet-SPK: full pipeline speaker embedding toolkit with reproducible recipes, self-supervised front-ends, and off-the-shelf models," in *Proc. Interspeech*, 2024, pp. 4278–4282.

[24] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023.

[25] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard *et al.*, "No language left behind: Scaling human-centered machine translation," *arXiv:2207.04672*, 2022.