Exploring Cross-Lingual Voice Conversion Methods for Anonymizing Low-Resource Text-to-Speech

Shenran Wang[1]
shenranw@cs.ubc.ca

Aidan Pine[2]
aidan.pine@nrc-cnrc.gc.ca

Mengzhe Geng[2]
mengzhe.geng@nrc-cnrc.gc.ca

[1]Department of Computer Science, University of British Columbia
[2]Digital Technologies Research Centre, National Research Council Canada

Abstract

We describe and compare multiple approaches for using voice conversion techniques to mask speaker identities in low-resource text-to-speech. We build and evaluate speaker-anonymized text-to-speech systems for two Canadian Indigenous languages, nêhiyawêwin and SENĆOŦEN, and show that cross-lingual speaker transfer via multilingual training with English data produces the most consistent results across both languages. Our research also underscores the need for better evaluation metrics tailored to cross-lingual voice conversion. Our code can be found at https://github.com/EveryVoiceTTS/Speaker_Anonymization_StyleTTS2

1 Introduction

Modern speech synthesis is being used in a growing number of applications. Early neural systems like Tacotron2 (Shen et al., 2018) and FastSpeech2 (Chien et al., 2021) transformed the speech synthesis landscape. But, with the increased level of naturalness brought by neural TTS systems, there has been growing concern about the extent to which these modern TTS systems clone a person's voice, and what the privacy and ethical considerations of this might be (Hutiri et al., 2024).

The work in this paper is related to the Speech Generation for Indigenous Language Education project (Pine et al., 2025), which is focused on building TTS systems to support Indigenous language education in Canada. In many Indigenous language communities, the people who have created recordings are well known to the community. Creating generative speech models capable of saying *anything* presents a number of unique risks to those individuals. For example, if the model makes mistakes, their reputation as a language expert might be in jeopardy; if the model is misused for illegal or nefarious purposes, the person's likeness might lend credibility to those attacks and make people more likely to believe a scam attack. One possible mitigation for these risks lies in the ability to mask or anonymize the identity of the speaker. Anonymization is by no means a comprehensive solution for all risks posed by neural TTS models, but it would mitigate some of the potential harms for the individuals who contribute their voices to these types of projects.

The goal of this paper is to investigate a variety of voice conversion strategies that conceal the original identity of the speaker(s) whose recordings were used during training while maintaining high-quality synthesized speech with respect to naturalness. We use StyleTTS2 (Li et al., 2023) for our experiments since it has been demonstrated to be highly effective at few-shot cross-lingual speaker adaptation (Li et al., 2023; Pine, 2025). We separate our experiments into three distinct categories: **(1)** voice conversion techniques that modify pretrained StyleTTS2 models without any fine-tuning (§3.2); **(2)** off-the-shelf voice conversion models (§3.3); and **(3)** cross-lingual speaker transfer from a higher-resource language seen during training (§3.4).

2 Data

Languages We focus on two languages: nêhiyawêwin (CRK) and SENĆOŦEN (STR)[1]. The first language, nêhiyawêwin, is an Algonquin language also known as Plains Cree spoken throughout a vast area primarily in the prairies of Alberta and Saskatchewan. The second language, SENĆOŦEN, is a Salish language spoken in W̱SÁNEĆ territory on the Saanich peninsula. It is notable for TTS for having a significantly large consonant inventory.

Data partition To create our multilingual datasets, we mix the nêhiyawêwin and SENĆOŦ

[1]Abbreviations adopted from ISO639-3 (https://iso639-3.sil.org/code_tables/639/data)

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 225–231

Split	# Utterances	Duration (hours)	# Speakers
Total			
Train	43801	64.58	257
Validation	300	0.37	81
Test	300	0.38	85
SENĆOŦEN			
Train	6931	3.94	2
Validation	100	0.05	2
Test	100	0.06	2
nêhiyawêwin			
Train	5516	8.17	8
Validation	100	0.16	7
Test	100	0.15	7
LibriTTS-R			
Train	31358	52.69	247
Validation	100	0.16	72
Test	100	0.18	76

Table 1: Summary statistics of data we use for this project. For nêhiyawêwin, there is a speaker who only has two samples. We only included those two samples in the training set.

EN data with the 53 hours from the LibriTTS-R-train-clean-100 partition (Koizumi et al., 2023) which intersects with the phonemized LibriTTS filelist provided by StyleTTS2. We randomly sample 100 samples for validation and 100 samples for testing for each dataset and reserve the rest for training. The duration of each split is listed in Table 1. Note that in our evaluations, we only use the nêhiyawêwin and SENĆOŦEN test sets, as we only care about anonymizing speech in these languages. All speakers in the test set of nêhiyawêwin and SENĆOŦEN are exposed to the model during training.

3 Experiments

An important feature of StyleTTS2 that is relevant to our experiments is that it uses self-supervision to learn style and speaker characteristics from audio during training. During inference, reference audio must be provided to the model to guide the style and speaker quality of the generated audio.

3.1 No Conversion Baseline

We begin by not changing the voice of any speaker. On a monolingual model trained with the nêhiyawêwin or SENĆOŦEN dataset, we take a random sample from each nêhiyawêwin or SENĆOŦEN speaker as style reference audio to generate the corresponding style embeddings to guide the speech generation of that speaker.

3.2 Inference-Time Experiments

We also conduct experiments that can be accomplished without any training solely by modifying StyleTTS2's style embeddings: the embeddings learned by the acoustic and prosodic style encoders. On a baseline model trained with only the nêhiyawêwin dataset, we perform the following modifications:

Average for each speaker, we take all his/her samples to generate a list of style embeddings, then take the mean of these style embeddings.

Gender Steering we compute the mean style embeddings for male and female speakers and take their difference ($\bar{female} - \bar{male}$). At inference, we add this offset to male embeddings (and vice versa) to shift styles between genders.

Gender Average we use the average embedding of male speakers to replace the style embeddings for male data points, and vice versa for female data points.

Noise we add Gaussian noise to the style embeddings of each speaker.

Custom we randomly pick 5 male speakers and 5 female speakers from LibriTTS-R to extract style embeddings for guiding the generation of the monolingual model. Results are taken as the average of all 10 runs.

For the SENĆOŦEN model, we only conduct baseline and custom settings.

3.3 Voice Conversion Experiments

Additionally, we conduct voice conversion experiments with SeedVC (Liu, 2024). Using four pretrained SeedVC models from Liu (2024) (v1-base, v1-small, v1-xlsr, v2)[2], we perform voice conversion on the no-conversion generations of the nêhiyawêwin model, with similar procedure as the above **Custom** setting. For SeedVC v1, we use the default configurations. Since SeedVC v2 also reports to convert accents, we experiment with the accent-related hyperparameters. Details can be found in Appendix A. For SeedVC v2, we experiment with either not converting style, or $\alpha\%$ of similarity to the reference speech.

3.4 Training-Time Experiments

Other than the above inference time experiments, we also train multilingual models from scratch. The goal is to train a StyleTTS2 model that has been exposed to many more speakers than the low-resource baseline models. Then, at inference, we use a sample from the LibriTTS-R dataset as the

[2] https://github.com/Plachtaa/seed-vc

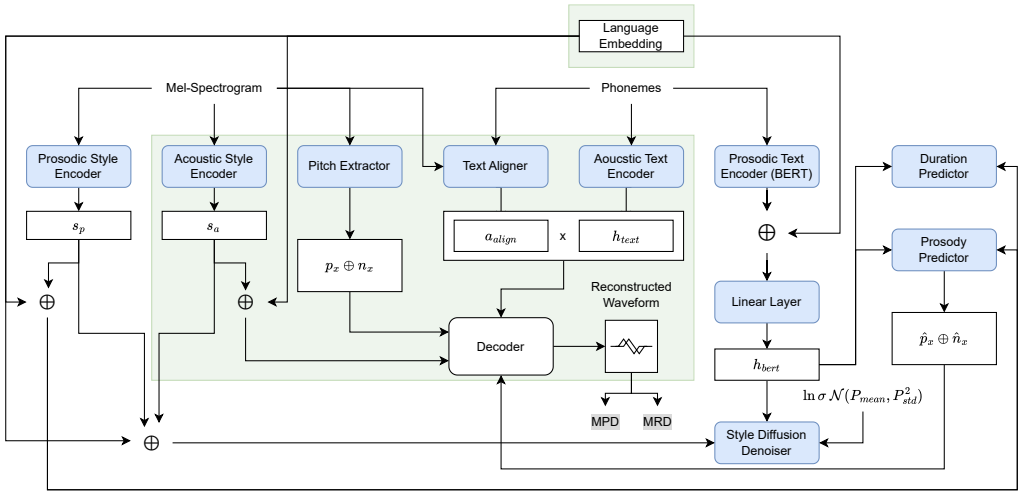

Figure 1: Architecture of the multilingual model with language embedding. We optimize only the modules inside the green box in the first stage, then all modules jointly in the second stage (details in Appendix B). We add language embeddings to the style encoders, the prosodic text encoder, and the text encoder to facilitate multilingual training. Other modules of the original StyleTTS2 architecture are not changed.

style reference to convert the voice of the target language. To disentangle language information from the StyleTTS2 style encoder module, we adapt the model to include language embeddings. The model design with language embedding is presented in Figure 1. Training details can be found in Appendix B. We train and include the results for both variants, one with language embeddings and one without. Like the 'Custom' setting described in §3.2, we convert using 10 random speakers of LibriTTS-R, 5 male and 5 female, and take the average of their calculated style embeddings. We train a total of 6 multilingual models with 3 combinations of datasets: nêhiyawêwin + LibriTTS-R (CRK-ENG), SENĆOŦEN + LibriTTS-R (STR-ENG), and nêhiyawêwin + SENĆOŦEN + LibriTTS-R (CRK-STR-ENG).

4 Results

Metrics Standard objective evaluation metrics like PESQ and SI-SDR require comparing to a ground truth waveform. However, since we have produced speech in another voice, these standard metrics cannot be applied. We report the quality results with predicted PESQ, and SI-SDR using torchaudio-squim (Kumar et al., 2023), and MOS prediction using UTMOS (Saeki et al., 2022), as we were unable to perform human evaluation due to the difficulty of human evaluation in low-resource settings (Pine et al., 2022). For evaluating voice similarity, we use Resemblyzer [3] to generate a pairwise similarity score between the generated audio and the ground truths. We then take the mean of all pairwise scores as a measure of speaker embedding cosine similarity (SECS). In addition to SECS, we also want to evaluate our models in terms of accent, since high SECS may be heavily influenced by speaker timbre and provide insufficient evidence for evaluating accent; this is particularly important in evaluating cross-lingual voice conversion using model variants that have no language embedding. To this end, we use Massive Multilingual Speech - Finetuned LID (MMS-LID) (Pratap et al., 2024) to evaluate accent by assessing the confidence with which MMS-LID predicts the audio is nêhiyawêwin (CRK-Conf). Since there is no existing language classifier that is trained on SENĆOŦEN, we only evaluate accent on nêhiyawêwin.

Discussion We present our results for nêhiyawêwin and SENĆOŦEN in Tables 2 and 3. Additional results are available in Appendix C. We may first observe from Table 2 that almost all inference-time experiments perform similarly, yielding high naturalness and high speaker similarity (i.e., poor anonymization), despite the different strategies we use. The two exceptions to this are the 'average' and 'custom' inference-time strategies, which have both the highest MOS scores and lowest SECS scores among all inference-time

[3] http://github.com/resemble-ai/Resemblyzer

Model	Setting	PESQ↑	SI-SDR↑	MOS↑	SECS↓	CRK-Conf↑
-	Ground Truth	3.58	25.48	3.27	1.00	0.53
CRK	No Conversion	3.83	27.28	3.43	0.87	0.67
Inference-time experiments						
CRK	Average	**4.09**	**30.40**	4.02	0.64	**0.70**
CRK	Steering Male to Female	3.74	25.92	3.73	0.81	0.61
CRK	Steering Female to Male	3.82	27.11	3.70	0.85	0.63
CRK	Gender Average	4.01	28.43	3.89	0.72	0.64
CRK	Noise	3.78	26.74	3.76	0.86	0.69
CRK	Custom	3.91	27.56	**4.06**	**0.63**	0.62
Voice conversion experiments						
CRK + SeedVC-v1-base	Default	2.46	7.99	2.34	0.54	0.12
CRK + SeedVC-v1-xlsr	Default	**3.71**	**25.62**	**3.98**	0.58	**0.51**
CRK + SeedVC-v1-small	Default	3.45	19.76	3.76	0.55	0.25
CRK + SeedVC-v2	Similarity = 0	3.33	19.00	3.84	0.53	0.02
CRK + SeedVC-v2	Similarity = 0.3	3.32	18.96	3.79	0.53	0.02
CRK + SeedVC-v2	Similarity = 0.5	3.29	18.98	3.77	**0.52**	0.03
CRK + SeedVC-v2	Similarity = 1	3.23	18.07	3.64	0.53	0.02
CRK + SeedVC-v2	Not converting style	3.23	17.98	3.67	0.54	0.17
Train-time experiments - with language embedding						
CRK-ENG	Custom	**3.92**	**29.57**	**4.03**	0.63	**0.67**
CRK-STR-ENG	Custom	3.92	28.81	3.76	**0.61**	0.55
Train-time experiments - without language embedding						
CRK-ENG	Custom	3.91	**30.03**	**4.25**	**0.52**	0.003
CRK-STR-ENG	Custom	**4.05**	29.98	4.23	0.54	**0.06**

Table 2: Evaluation results for nêhiyawêwin experiments. The best performances are highlighted for each subcategory. Lower SECS values indicate better performance, as the objective is speaker anonymization rather than speaker similarity.

Model	Setting	PESQ↑	SI-SDR↑	MOS↑	SECS↓
-	Ground Truth	3.29	23.05	3.17	1.00
STR	No Conversion	3.40	23.75	3.21	0.87
Inference-time experiments					
STR	Custom	3.55	24.64	3.29	0.74
Train-time experiments - with language embedding					
STR-ENG	Custom	3.53	24.50	**3.43**	**0.69**
CRK-STR-ENG	Custom	**3.57**	**25.16**	3.40	0.69
Train-time experiments - without language embedding					
STR-ENG	Custom	**3.82**	**28.26**	**3.93**	**0.56**
CRK-STR-ENG	Custom	3.87	28.37	3.87	0.56

Table 3: Evaluation results of SENĆOŦEN experiments. The best performances are highlighted for each subcategory.

experiments. However, the 'custom' setting on our SENĆOŦEN model does not perform as well, which indicates that these inference-time approaches might be less robust either across languages or the number of speakers in a dataset.

The voice conversion experiments achieved lower speaker similarity (i.e., better anonymization), but in exchange for worse overall quality.

While models trained without language embeddings gave slightly better measures in terms of quality and similarity compared to models with language embeddings, we find them having a strong English accent due to the entanglement of speaker and language. This is quantified by the substantially lower MMS-LID confidence scores (CRK-Conf) for models without language embeddings. In Table 3, we observe that when mixed with LibriTTS-R, multilingual models were able to produce speech with much lower similarity compared to the monolingual model. We also impressionistically find that models without language embeddings produced SENĆOŦEN speech in a strong English accent.

In general, we find our trained models with language embeddings to be the most robust solution. Models without language embeddings always produce speech with a strong English accent; inference-time experiments worked well on the nêhiyawêwin model but not on the SENĆOŦEN model; and off-the-shelf voice conversion systems sacrifice too much quality for lower similarity. Pretraining with more languages is not always better, however, as we see that the CRK-ENG model performs better than the CRK-STR-ENG across most metrics.

5 Conclusion

This paper explored cross-lingual voice conversion methods as a way of masking speaker identities in low-resource TTS systems, with a focus on Indigenous languages such as nêhiyawêwin and SENĆOŦEN. We evaluated inference-time and training-time approaches using StyleTTS2, finding that inference-time methods, while capable of producing high-quality speech, struggled with fully masking speaker identities in SENĆOŦEN. Training-

time methods, especially those with language embeddings, showed improved anonymization by reducing speaker similarity while preserving naturalness and intelligibility for both languages, though challenges like accent entanglement persisted. Our findings underscore the need for better evaluation metrics tailored to cross-lingual voice conversion and highlight future directions, including conducting human evaluations, and exploring model architectures and training routines that balance privacy with speech quality.

Limitations

Due to the limited number of speakers, we were unable to perform human evaluations on our models. We also adopted predicted evaluation metrics from pretrained models, which could contain bias or errors.

Acknowledgments

We would like to thank and acknowledge the W̱SÁNEĆ School Board and University nuhelot'įne thaiyots'į nistameyimâkanak Blue Quills for their work with us in the SGILE project that led to this research.

Ethical Statements

The model we used in our research, StyleTTS2, is a zero-shot TTS model that has the potential for misuse and deception. To tackle these potential challenges and also to adhere to our existing research protocols for working with Indigenous communities, we will not disclose our model weights to the public. The models remain in the sole ownership of the communities we work with.

References

Chung-Ming Chien, Jheng-Hao Lin, Chien-yu Huang, Po-chun Hsu, and Hung-yi Lee. 2021. Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8588–8592.

Wiebke Hutiri, Orestis Papakyriakopoulos, and Alice Xiang. 2024. Not my voice! a taxonomy of ethical and safety harms of speech generators. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 359–376, New York, NY, USA. Association for Computing Machinery.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. 2023. Libritts-r: A restored multi-speaker text-to-speech corpus. In *Interspeech 2023*, pages 5496–5500.

Anurag Kumar, Ke Tan, Zhaoheng Ni, Pranay Manocha, Xiaohui Zhang, Ethan Henderson, and Buye Xu. 2023. Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. 2023. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 19594–19621. Curran Associates, Inc.

Songting Liu. 2024. Zero-shot voice conversion with diffusion transformers. *Preprint*, arXiv:2411.09943.

Aidan Pine. 2025. The NRCC Submission to the Blizzard Challenge 2025. In *The Blizzard Challenge 2025*, pages 24–30.

Aidan Pine, Erica Cooper, David Guzmán, Eric Joanis, Anna Kazantseva, Ross Krekoski, Roland Kuhn, Samuel Larkin, Patrick Littell, Delaney Lothian, Akwiratékha' Martin, Korin Richmond, Marc Tessier, Cassia Valentini-Botinhao, Dan Wells, and Junichi Yamagishi. 2025. Speech generation for indigenous language education. *Computer Speech Language*, 90:101723.

Aidan Pine, Dan Wells, Nathan Brinklow, Patrick Littell, and Korin Richmond. 2022. Requirements and motivations of low-resource speech synthesis for language revitalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7346–7359, Dublin, Ireland. Association for Computational Linguistics.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling Speech Technology to 1,000+ Languages. *Journal of Machine Learning Research*, 25(97):1–52.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. In *Interspeech 2022*, pages 4521–4525.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and

Yonghui Wu. 2018. Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions. pages 4779–4783.

A Voice Conversion Experiment Details

We present the configurations of voice conversion models in Tables 4 and 5. According to the authors of SeedVC, SeedVC-v1-base is for singing voice conversion, for which it is best to set `diffusion-steps` to 30 to 50.

Config	SeedVC-v1-base	SeedVC-v1-xlsr	SeedVC-v1-small
`diffusion-steps`	50	30	30
`length-adjust`	1.0	1.0	1.0
`inference-cfg-rate`	0.7	0.7	0.7
`f0-condition`	False	False	False
`auto-f0-adjust`	False	False	False
`semi-tone-shift`	0.0	0.0	0.0

Table 4: Configurations for SeedVC-v1's default settings.

B Training Details

StyleTTS2 adopts a two-stage training where they first pretrain the acoustic module, then jointly train the entire model. All first-stage trainings are done on two A100-40GB GPUs, and all second-stage trainings are done on one A100-40GB GPU. Hyperparameters and optimization details are listed in Table 6.

C Additional Results

We trained a 32-component PCA and a logistic regression to classify the gender of speakers for the nêhiyawêwin model, using the style embeddings, in hopes of finding certain dimensions that are the most important to the style by looking at the highest-valued dimensions in PCA's first component and logistic regression's weight. We selected the top 32 dimensions of the entire style embedding or only the acoustic style embedding from the PCA and logistic regression to perform all experiments listed in §3.2. Results are listed in Table 7. In general, we find the results here being similar to all other inference-time methods.

Config	Not Converting Style	Similarity=0	Similarity=0.3	Similarity=0.5	Similarity=1
`intelligibility-cfg-rate`	0.7	0.7	0.7	0.7	0.7
`similarity-cfg-rate`	0.7	0.0	0.3	0.5	1.0
`top-p`	0.9	0.9	0.9	0.9	0.9
`temperature`	1.0	1.0	1.0	1.0	1.0
`repitition-penalty`	1.0	1.0	1.0	1.0	1.0
`convert-style`	False	True	True	True	True
`anonymization-only`	False	False	False	False	False

Table 5: Configurations for SeedVC-v2's settings.

Model	# Parameters	Optimizer	Learning Rate (Model / BERT / Acoustic Module)	Epochs (1st stage / 2nd stage)	Batch Size	Language Embedding Size
CRK	190,454,891	Adam	1e-4 / 1e-5 / 1e-5	150 / 100	4	-
STR	190,454,891	Adam	1e-4 / 1e-5 / 1e-5	150 / 100	4	-
CRK-ENG	198,065,195	Adam	1e-4 / 1e-5 / 1e-5	150 / 100	4	64
CRK-ENG (no language embedding)	190,454,891	Adam	1e-4 / 1e-5 / 1e-5	150 / 100	4	-
CRK-STR-ENG	198,065,259	Adam	1e-4 / 1e-5 / 1e-5	150 / 100	4	64
CRK-STR-ENG (no language embedding)	190,454,891	Adam	1e-4 / 1e-5 / 1e-5	150 / 100	4	-
STR-ENG	198,065,195	Adam	1e-4 / 1e-5 / 1e-5	150 / 100	4	64
STR-ENG (no language embedding)	190,454,891	Adam	1e-4 / 1e-5 / 1e-5	150 / 100	4	-

Table 6: Training Details

Setting	PESQ↑	SI-SDR↑	MOS↑	SECS↓	CRK-Conf↑
PCA-top32/Average	3.78	26.40	3.73	0.85	0.67
PCA-top32/Steering Male to Female	3.85	27.50	3.80	0.87	0.68
PCA-top32/Steering Female to Male	3.82	26.98	3.74	0.87	0.67
PCA-top32/Gender Average	3.80	26.24	3.73	0.86	0.66
PCA-top32/Noise	3.84	26.70	3.77	0.87	**0.68**
PCA-top32/Custom	**3.90**	**28.14**	4.05	0.63	0.62
LR-top32/Average	3.86	26.81	3.75	0.85	0.68
LR-top32/Steering Male to Female	3.82	27.07	3.79	0.87	0.66
LR-top32/Steering Female to Male	3.87	27.11	3.75	0.87	0.66
LR-top32/Gender Average	3.82	26.47	3.78	0.86	0.67
LR-top32/Noise	3.79	26.38	3.74	0.87	0.66
LR-top32/Custom	3.87	27.67	**4.06**	**0.63**	0.62

Table 7: Additional evaluation results of the nêhiyawêwin model experiments. The best performances are highlighted.