Cross Lingual Style Transfer Using Multiscale Loss Function for Soliga: A Low Resource Tribal Language

Ashwini Dasare[1](✉), B. Lohith Reddy[1], A. Sai Chandra Koushik[1], B. Sai Raj[1], V. Krishna Sai Rohith[1], Satisha Basavaraju[2], and K.T. Deepak[1]

[1] Indian Institute of Information Technology, Dharwad, India
{aswhwini,19bcs014,19bcs006,19bcs061,19bcs016,deepak}@iiitdwd.ac.in
[2] Beltech AI Pvt. Ltd., Bangalore, India
sathisha.iitg@gmail.com

Abstract. Voice conversion is the art of mimicking different speaker voices and styles. In this paper, we present a cross-lingual speaker style adaptation based on a multi-scale loss function, using a deep learning framework for syntactically similar languages Kannada and Soliga, under a low resource setup. The existing speaker adaptation methods usually depend on monolingual data and cannot be directly adopted for cross-lingual data. The proposed method calculates multi-scale reconstruction loss on the generated mel-spectrogram with that of the original mel-spectrogram and adopts its weights based on the loss function for various scales. Extensive experimental results illustrate that the multi-scale reconstruction resulted in a significant reduction of generator noise compared to the baseline model and faithfully transfers Soliga speaker styles to Kannada speakers while retaining the linguistic aspects of Soliga.

Keywords: Low resource cross-lingual speech synthesis · Spoken language generation · Speaker adaptation · Voice conversion · Style transfer

1 Introduction

Tribal languages aid us in understanding our origin, the roots we evolved from, and human race capabilities. India has 22 official spoken languages and 1369 rationalized unofficial languages. About 197 languages of India are listed as endangered by UNESCO [12], and any language spoken by a population less than 10,000 people is considered a potentially endangered language by Govt of India [3]. Many tribal languages of India do not have literary tradition [4]. One such tribal language, Soliga is already on the verge of extinction as it has only a population of fewer than 40,000 people [14], located in Karnataka state. If we do not create digital platforms for languages to be used by respective communities, these languages may vanish in less than a decade [7]

A. Karpov et al. (Eds.): SPECOM 2023, LNAI 14339, pp. 182–194, 2023.
https://doi.org/10.1007/978-3-031-48312-7_15

Soliga language does not have a script, and to the best of our knowledge, no literature is available in Soliga. Therefore it can be considered as a "zero resource" language. Deep neural networks has been popularly used for TTS, but all of these models need transcripts for each speech file to train the models [10,15,17,19]. For zero-resource languages like Soliga, generating speech data from the written transcript is tedious. But when unlabeled data is used in Speech-to-speech based approaches [2,8], despite the significant improvement in the synthesized audio quality, these approaches tend to miss the prosodic aspects such as speech style, tone, volume, and pitch present in the sample of speech for a given speaker. To train the model for prosodic aspects, we require hundreds of hours of data [22] which is practically not feasible for low-resource languages, especially when the population is as small as Soliga. Finding people to get speech data is a huge time and effort-consuming process. However, cross-lingual style transfer approaches can address the issue of generating speech resources for low-resource languages. The generated speech utterances can be used for applications such as direct speech-to-speech translation, automatic speech recognition, and Speaker recognition and diarization. This can also serve as digital preservation of indigenous tribal languages.

The process of generating a speech sample of a source speaker to a different target speaker while retaining the linguistic content and speaker style characteristics of a source speaker is called Speech style transfer. This paper introduces a cross-lingual, speech-to-speech neural network to transfer the speech style across Soliga and Kannada languages. This work will show the importance of multi-scale reconstruction loss in a speech-style conversion task, thereby preventing the training objective from halting in the local minima.

1.1 Neural Style Transfer

Neural style transfer was introduced for image generation [5]. Since the speech waveform can also be represented in the form of an image in the frequency domain(mel-spectrogram), the same techniques of image style transfer can be applied to Speech data as well. This was shown by [21], in which the model synthesized spoken image descriptions directly without using any text or phonemes. Later using a similar approach of neural style transfer and GAN models, the Speech style transfer technique was proposed [1], which is the baseline model for our implementation. The baseline model uses L_1-loss computed at the input scale to penalize the reconstruction error between the same source and target mel-spectrogram images. However, penalizing the network for the errors at each scale of reconstruction may enable the decoder network to inject speaker-specific style information in the mel-spectrogram image. The overall pipeline is shown in Fig. 1.

The baseline model does not give an intelligible voice for different sources and target languages. Therefore we introduced one more loss function called multi-scale L_1-loss. In computer vision models, it is proven that introducing multi-scale loss significantly improved the accuracy of the image reconstruction [6]. For our mel-spectrogram image, we have adopted this multi-scale reconstruction technique to estimate the loss function along with existing baseline loss functions. In the multi-scale approach, the combination of the individual losses at each scale is the total loss in the decoder, as shown in Fig. 2. Multi-scale reconstruction of the mel-spectrogram improves the performance of the discriminator. We experimented with different scales and discovered that the down-scaling approach works best for our model-detailed explanation in Sect. 3.1.

The overall pipeline is shown in Fig. 1, which is essentially carried out in two steps. In the first step, from a given input language i.e. Soliga, the target speaker's(Kannada) mel-spectrogram is generated, using VAE-GAN network, these spectrograms are fed to, the WaveNet-based vocoder to get the synthesized speech in the time domain. This approach was used in image-to-image style transfer model, and the same is adopted and applied to mel-spectrograms. We retain the base model single encoder structure to generalize and to make it more feasible for multiple target speakers. In this work, we introduce multi-scale loss estimation for mel-spectrograms generation.

The proposed method does not require parallel bilingual data and phoneme representation but only needs bilingual speech data without transcriptions to train a generator model. We claim that this works well for low-resource data for the same language, with different speaking styles. And also works well for languages that share similar syntactic structures. The languages we have chosen for style transfer and voice conversion have syntactically similar utterance patterns and Soliga has kannada word influence [13]. The training is performed in an end-to-end unsupervised manner without any transcript alignments between the input samples.

The paper organization is as follows, Sect. 2 gives data preparation details, Experimental details are given in Sect. 3, Sect. 4 presents Results and Discussion and Sect. 5 discusses concluding remarks and future work.

2 Data Preparation

In order to create a speech database for the Soliga and Kannada languages, we faced the challenge of Soliga being a zero-resource language with no written literature available. To overcome this hurdle, we followed a two-step approach: creating text sentences and recording these sentences with the help of a literate Soliga speaker. The data preparation process involved the following steps:

2.1 Text Sentence Creation

To generate text sentences, we utilized a list of commonly used words in Indian villages known as "swadesh" [20]. Using these words as a foundation, we constructed approximately 10,000 English sentences with lengths ranging from 3 to 15 words. These sentences served as the starting point for our data collection process. As Soliga resides in Karnataka, the contact language is Kannada, and the English sentences were first translated into Kannada, and in the second step, these Kananda sentences were translated into Soliga. Since Soliga does not have its own script, the translated Soliga sentences were written in the Kannada script. This approach ensured that the Soliga speakers, who were more comfortable with Kannada, could understand and read the sentences accurately. Likewise, 5000 parallel translated sentences were prepared for data recording.

2.2 Speech Recording

In the next phase, we sought a female Soliga speaker with a clear voice and good diction to perform studio-quality voice recordings. To facilitate the data collection process, we developed a user interface (UI) that allowed for the recording of speech data while capturing relevant metadata about the speaker, including age, gender, education, and qualifications. Additionally, the speakers were asked to provide their consent for donating their voice for research purposes. The recording sessions took place in a professional studio, with the speech data collected at a sampling rate of 44 kHz to ensure high-quality recordings. The UI displayed the sentences to be read by the speaker, providing an option to listen to each recording and allowing for rerecording in case of any errors or mistakes. Using this methodology, we collected approximately 5,000 utterances from both Kannada and Soliga speakers, resulting in a total duration of about 5 h of recorded speech data for each language.

Furthermore, to expand the number of speakers and incorporate different styles for Soliga to Kannada voice conversion, we augmented our dataset with publicly available Kannada male voice data obtained from sources such as Openslr [16]. By incorporating a diverse range of speakers and styles, we aimed to improve the robustness and flexibility of our model in performing Soliga to Kannada voice conversion.

In summary, the data preparation process involved creating text sentences using a list of common words, translating them into Kannada for Soliga speakers, recording the sentences in a studio environment, and augmenting the dataset with additional Kannada male voice data. These steps ensured the availability of a comprehensive and diverse dataset for training our model.

3 Implementation

The Kannada speaker's voice is converted from the Soliga speaker's input signal while keeping the Soliga speaker's content and style intact. The representation of the mel-spectrogram is created from the input speech signal. Once the spectrogram is generated, it can be treated as a greyscale image, we can employ a neural style transfer model as mentioned in [11], and then the source mel-spectrogram can be converted to the target mel-spectrogram using the style of a target speaker. To convert this mel-spectrogram back into the time domain, we used the WaveNet vocoder [23]

The neural style transfer model used in this research employs a combination of encoder and decoder networks to achieve cross-lingual style transfer between

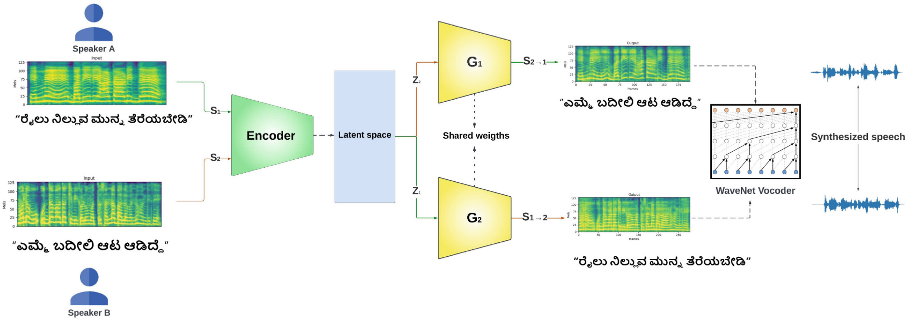

Fig. 1. The style transfer flow between Soliga and Kannada.

Soliga and Kannada languages. The decoder network functions as a generator, responsible for producing the target output speech. On the other hand, the encoder network aims to preserve the linguistic details of the source speech while eliminating speaker identity-related information.

In the proposed model, the encoder architecture remains unchanged from the base paper. Its role is to generate a shared-latent space called "z" that captures only the content of each sample while discarding the identity of the original speaker. By extracting the content information, the encoder helps facilitate the style transfer process. In our approach, we considered two generators, denoted as G_1 and G_2, to accommodate the two speakers for the Kannada and Soliga languages, respectively. Both G_1 and G_2 consist of the same layers as the encoder, but in the opposite order. This arrangement allows the generators to combine the input speech signal's content and style with the target voice, generating the mel-spectrogram of the target speech signal.

To improve the performance of the generator and address the issue of noisy output encountered in the base paper, we introduced additional layers to the generator architecture. Specifically, we incorporated two extra layers for down-scaling and up-scaling experiments. Additionally, we included four more layers for up-down-scaling of the generated mel-spectrogram. This modification enables a more effective transformation of the style while maintaining the integrity of the content. By comparing the generated mel-spectrogram with the original mel-spectrogram, we calculate the multi-scale loss function, which serves as a measure of the style transfer quality. The multi-scale loss function allows us to evaluate the performance of the generator at different levels of abstraction, capturing both local and global style transfer characteristics. In the context of transferring style between Soliga and Kannada languages, the latent space for the two given decoders is switched. This switch allows the generator to effectively capture the style attributes specific to each language and incorporate them into the generated target speech. The swapping of latent spaces facilitates the cross-lingual style transfer process and ensures that the resulting speech aligns with the desired linguistic and stylistic characteristics.

Figure 1 illustrates the architecture and process flow of the proposed model, highlighting the switching of latent spaces between the decoders for Soliga and Kannada languages. By incorporating these modifications to the generator architecture and leveraging the switched latent spaces, our approach improves the style transfer performance and yields more accurate and coherent results compared to the base paper.

3.1 Multi-scale Loss Function Experiments

In this section, we describe the experiments conducted to evaluate the effectiveness of the multi-scale loss functions in improving the performance of the baseline model. In addition to the L_1 loss used in the base paper, we introduced several multi-scale loss functions, including up-scaling, down-scaling, and a combination of up and down-scaling. The purpose of these experiments was to investigate how different scaling operations applied to the generated mel-spectrograms can impact the intelligibility and voice quality of the transferred speech. By analyzing the results, we aimed to identify the most effective approach for enhancing the performance of the baseline model.

In the first experiment, we applied a down-scaling operation to the generated mel-spectrograms. This down-scaling process involved reducing the resolution of the spectrogram while preserving its content and style information. The motivation behind this experiment was to examine whether reducing the resolution could lead to improved speech quality and intelligibility. Next, we conducted an up-scaling experiment, where we increased the resolution of the generated mel-spectrograms. This operation aimed to enhance the fine-grained details in the transferred speech, potentially improving the overall quality and fidelity. Finally, we combined the up-scaling and down-scaling operations in a single experiment. This experiment sought to leverage the benefits of both scaling approaches simultaneously, allowing for a more comprehensive evaluation of their impact on the transferred speech.

After analyzing the results of these experiments, we observed that the down-scaling experiment yielded better intelligibility and voice quality compared to the baseline model. The incorporation of the down-scaling multi-scale loss function appeared to enhance the performance of the baseline model in terms of speech quality. The down-scaling operation likely facilitated a more compact representation of the style and content information in the mel-spectrograms, resulting in clearer and more intelligible speech. By reducing the resolution, the down-scaling operation may have removed some unnecessary noise or artifacts present in the baseline model's output.

While the up-scaling and combined scaling experiments did not demonstrate significant improvements over the baseline model, they provided valuable insights into the effects of scaling operations on the style transfer process. These experiments highlighted the importance of finding the right balance between preserving the content and style information and enhancing the finer details in the transferred speech. Overall, the incorporation of a down-scaling multi-scale loss function proved to be an effective enhancement to the baseline model, leading to improved speech quality and intelligibility. This finding suggests that carefully

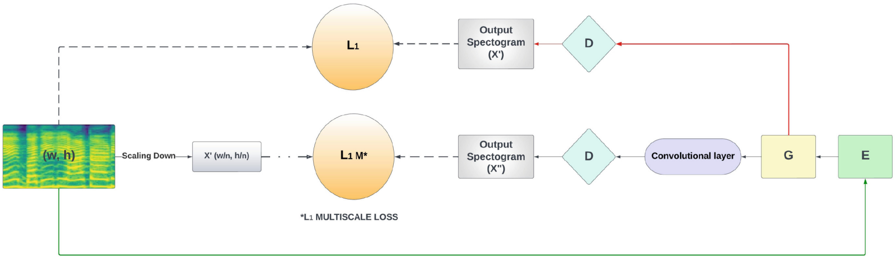

Fig. 2. Loss Function for Down-scaling. Here E is Encoder, G is generator and D is Discriminator

designed scaling operations can contribute to the success of neural style transfer for speech synthesis applications.

In the following sections, we present detailed analyses and discussions of the results obtained from these multi-scale loss function experiments, providing insights into the strengths and limitations of each approach.

Loss of Down-Scale. To improve the efficiency of the discriminator, we added two additional scales to its input by down-scaling the generator output to half and quarter of the original size. We calculated the GAN Loss for each scale and added them to the existing loss calculated from the original mel-spectrogram.

$$L_{GAN} = W_1 \times L_{GAN_1} + W_2 \times L_{GAN_2} + W_4 \times L_{GAN_4} \tag{1}$$

Here the GAN Loss of original input is calculated

$$\begin{aligned} L_{GAN_1} = & \sum_i E_{S_i \sim P_{S_i}} [\log\ D_i\left(S_i\right)] \\ & + \sum_{i,j} E\left(S_{j \to i \mid\ z_j}\right) [\log\left(1 - D_i\left(S(x)\right)\right] \end{aligned} \tag{2}$$

The following equation calculates the loss of down-scaling the generator output to half (n = 2) and quarter (n = 4) of the original size.

$$\begin{aligned} L_{GAN_n} = & \sum_i E_{S_i \sim P_{S_i}} [\log\ D_i\left(S_i\right)] \\ & + \sum_{i,j} E\left(S_{j \to i \mid\ z_j}\right) [\log\left(1 - D_i\left(S(x/n)\right)\right)] \end{aligned} \tag{3}$$

Loss of Up-Scale. In this loss, we add two additional scales to its input by up-scaling the generator output to double and quadrupling the original size. We calculated the GAN Loss for each scale and added them to the existing loss calculated from the original mel-spectrogram (Fig. 3).

$$L_{GAN} = W_1 \times L_{GAN_1} + W_2 \times L_{GAN_2} + W_4 \times L_{GAN_4} \tag{4}$$

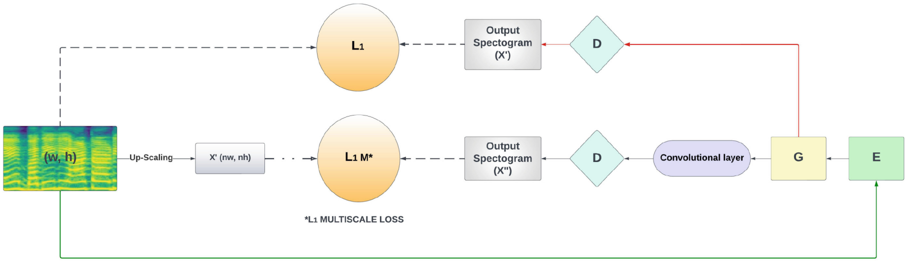

Fig. 3. Loss Function for Down-scaling. Here E is Encoder, G is generator and D is Discriminator

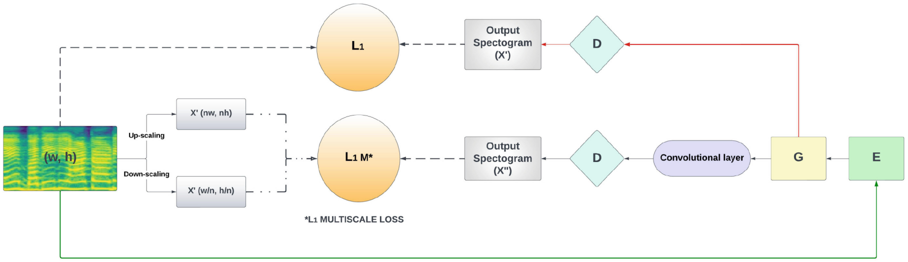

Fig. 4. Loss Function for Up and Down-scaling.

Here the GAN Loss of the original input is calculated,

$$\begin{aligned} L_{GAN_1} = {} & \sum_i E_{S_i \sim P_{S_i}} [\log\ D_i\left(S_i\right)] \\ & + \sum_{i,j} E\left(S_{j \to i\,|\,z_j}\right) [\log\left(1 - D_i\left(S(x)\right)\right] \end{aligned} \tag{5}$$

The following equation calculates the loss of Up-scaling the generator output to double (n = 2) and quadruple (n = 4) of the original size.

$$\begin{aligned} L_{GAN_n} = {} & \sum_i E_{S_i \sim P_{S_i}} [\log\ D_i\left(S_i\right)] \\ & + \sum_{i,j} E\left(S_{j \to i\,|\,z_j}\right) [\log\left(1 - D_i\left(S(nx)\right)\right] \end{aligned} \tag{6}$$

Combined Loss of Up-Scale and Down-Scale. In this case, for calculating loss, the discriminator in a Generative Adversarial Network (GAN), we propose a multi-scale loss approach. This approach involves adding four additional scales to the input of the discriminator. We achieve this by down-sampling and up-sampling the generator output to half and a quarter of the original size and double and four times, respectively (Fig. 3).

For each of these scales, we calculate the GAN Loss and add it to the existing loss calculated from the original image. The multi-scale loss is calculated as the weighted sum of these individual scale losses. The weight of each scale is determined by its size relative to the original image.

$$L_{GAN} = \sum_{i=1}^{5} W_i \times L_{GAN_i} \tag{7}$$

where W_i is the weight for each scale and L_{GAN_1} is the GAN Loss for original input from Eq. (1), L_{GAN_2} , L_{GAN_3} are the GAN Losses for Down-scale input as shown in equation(3) and L_{GAN_4}, L_{GAN_5} are the GAN Losses for Up-scale input as shown in Eq. (6).

The loss function employed by Variational Autoencoder (VAE) models aims to minimize the dissimilarity between the encoded distribution of input data and a prior distribution. This loss function comprises of two terms: the first one is the Kullback-Leibler (KL) divergence between the encoded and prior distributions, while the second term is the negative log-likelihood of the reconstructed input data.

$$\begin{aligned} L_{VAE} = \lambda_4 \sum_i D_{KL}\left(q\left(z_i \mid S_i\right) \parallel p\left(z\right)\right) \\ - \sum_i E_{z_i \sim q\left(\, z_i \mid S_i\right)}\left[\log p\, G_i\left(S_i \mid z_i\right)\right] \end{aligned} \tag{8}$$

The loss function is used in models that aim to generate new content from existing content. The first term in the equation is the KL divergence between the distribution of the latent codes for the source and target mel-spectrograms. The second term is the negative log-likelihood of the target mel-spectrogram given the source mel-spectrogram and its corresponding latent code.

$$\begin{aligned} L_{CC} = \lambda_4 \sum_{i,j} D_{KL}\left(q\left(z_i \mid S_{i\rightarrow j}\right) \parallel p\left(z\right)\right) \\ - \sum_{i,j} E_{z_i \sim q\left(\, z_j \mid S_{i\rightarrow j}\right)}\left[\log\, p\, G_i\left(S_i \mid z_j\right)\right] \end{aligned} \tag{9}$$

Cycle consistency loss we have retained as it is from the baseline model. The regularization parameters in the objective functions, use $\lambda_1 = 100$, $\lambda_2 = 10$, $\lambda_3 = 10$, and $\lambda_4 = e^{-3}$.The regularization parameters are given these values to emphasize the loss from reconstruction in L_{VAE} than the other loss terms.

We have experimented with different values of weights given to the multi-scale loss function. The values $w_1 = 0.5$, $w_2 = 0.25$, and $w_3 = 0.25$ gave better results. We choose these values to give more weight to the original scale compared to other scales. In addition, the WaveNet vocoder is trained independently using the mel-spectrograms generated by both G_1 and G_2 as inputs, while the original waveform of each speaker was used as the reference to compare the utterance and style of the target.

Table 1. Comparison of D and G loss for different models

Soliga-Kannada Style Transfer	D Loss	G Loss
Baseline-model	0.326	32.097
Down-scale model	**0.268**	**24.658**
Up-scale model	0.266	26.616
Up and Down-scale model	0.257	27.236

4 Results and Discussion

The loss reduction of our model compared to the baseline model is represented in Table 1. The generator loss was calculated using the Kannada-Soliga dataset, in both cases, our proposed method error is lesser than the baseline model as shown in the Table 1. Though we have experimented with different loss functions, i.e. by up-scaling and combining up-scaling and down-scaling, the generator gave less error for the down-scaling experiment. This could be because error propagation is high when you upscale the generated mel-spectrograms, the error gets added to the up-scaled mel-spectrogram as well, and when you downscale the mel-spectrogram the error will be reduced to down-scaled mel-spectrograms. The samples of Soliga and Kannada style transfer can be found on Results page [18].

We have visualized the source speaker and target speaker-specific feature embedding and found that the source features and style transferred features in the target cluster together in latent space. That gives the overall performance of our model in terms of the naturalness of the original and synthetic style aspects, as shown in Fig. 5. "K" stands for Kannada, "S" for Soliga "S-K" stands for Soliga sentence and style in the Kannada speaker's voice., "K-S" Kannada sentence and style in Soliga speaker's voice. Essentially the feature embeddings of "S" and "K-S" cluster together, and "K" and "S-K" should cluster together for good quality style transfer. Compared to the baseline model all three models proposed have better clustering of feature embeddings.

We have also conducted subjective evaluations on multi-scale models for intelligibility and style transfer tasks, and Mean Opinion Score(MOS) was taken by 10 Soliga speakers and 10 Kannada speakers to assess the quality of synthesized speech based on the parameters like intelligibility, style, and accent, on a scale of 1 to 5, where 1 being the poor quality and 5 being the best quality score. It was found that the Downscale model outperformed the baseline model, as shown in Table 2. This matches with down-scale Generator loss of Table 1.

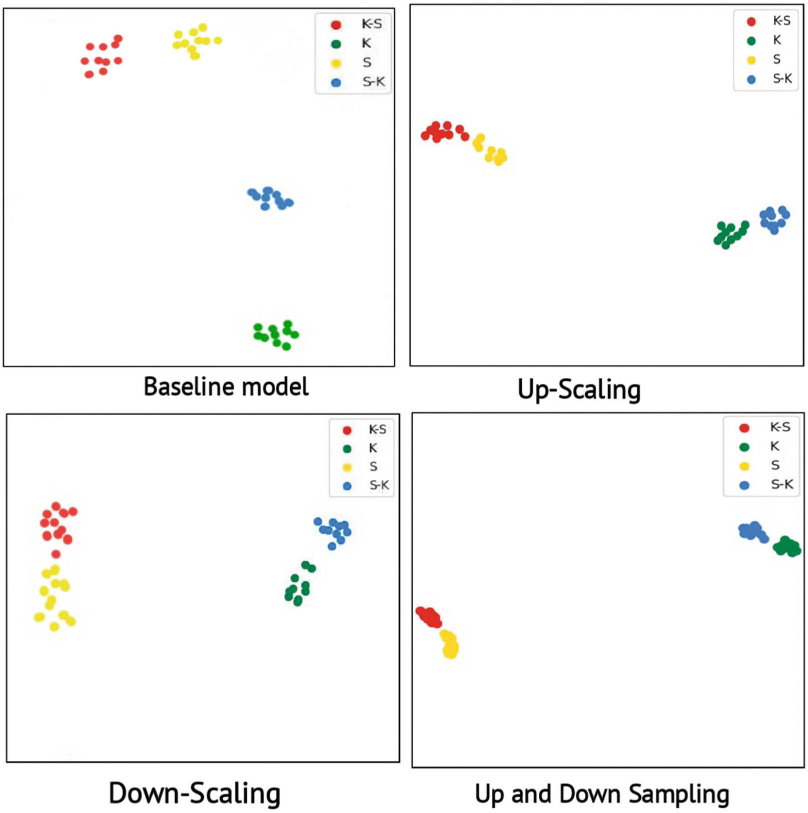

Fig. 5. Features embedding visualization for each speaker on the original and synthesized speech [9]

Table 2. Comparison of Mean Opinion Score (MOS)

Kannada-Soliga	MOS
Baseline-model	3.02
Upscale-model	3.32
Downscale-model	**3.99**
Up and down scale-model	3.50

5 Conclusion and Future Work

In this project, our main objective was to propose a technique for cross-lingual speaker style transfer between Kannada and Soliga, two languages that share similar syntax. We emphasized the significance of utilizing a multi-scale loss in deep neural networks to enable the learning and incorporation of subject-specific style into identity-independent feature embeddings. Through our experiments, we successfully demonstrated that the integration of multi-scale loss in deep neural networks effectively reduced generator noise and facilitated the faithful transfer of Soliga speaker styles to Kannada speakers, while preserving the original speech's content and style. Notably, our model exhibited promising results in voice conversion between different genders. However, further improvements are required to enhance voice conversion within the same gender.

In future research, there is an interesting avenue to explore regarding cross-linguistic speech style transfer between low-resource tribal languages and Indic languages. This area of study presents an opportunity to investigate the adaptation of speech styles across language boundaries, particularly in the context of languages with limited resources and the diverse landscape of Indic languages. By delving into this domain, researchers can uncover novel techniques and approaches for facilitating the transfer of speech styles, which can have a range of practical applications. For instance, it can contribute to the development of speech-to-speech translation systems that seamlessly adapt speech styles, enhance automatic speech recognition systems for low-resource languages, and even improve speaker recognition and diarization technologies for diverse linguistic communities.

Furthermore, exploring cross-linguistic speech style transfer in low-resource contexts can provide insights into the challenges faced by endangered languages and contribute to efforts aimed at their preservation and revitalization. It can help bridge the gap between low-resource tribal languages and the broader ecosystem of Indic languages, fostering better communication and understanding.

Overall, the exploration of cross-linguistic speech style transfer between low-resource tribal languages and Indic languages holds significant promise for future research. It offers an opportunity to better understand the dynamics of speech styles across languages, preserve endangered languages, and develop innovative technologies to bridge linguistic divides.

References

1. AlBadawy, E.A., Lyu, S.: Voice conversion using speech-to-speech neuro-style transfer. In: Interspeech, pp. 4726–4730 (2020)
2. Biadsy, F., Weiss, R.J., Moreno, P.J., Kanevsky, D., Jia, Y.: Parrotron: an end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation. arXiv preprint arXiv:1904.04169 (2019)
3. Chandramouli, C., General, R.: Census of india 2011. Provisional Population Totals. New Delhi: Government of India, pp. 409–413 (2011)
4. Dasare, A., Deepak, K., Prasanna, M., Vijaya, K.S.: Text to speech system for lambani-a zero resource, tribal language of india. In: 2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), pp. 1–6. IEEE (2022)
5. Ghiasi, G., Lee, H., Kudlur, M., Dumoulin, V., Shlens, J.: Exploring the structure of a real-time, arbitrary neural artistic stylization network. arXiv preprint arXiv:1705.06830 (2017)

6. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3828–3838 (2019)
7. Haokip, T.: Artificial intelligence and endangered languages. Available at SSRN 4212504 (2022)
8. Jia, Y., et al.: Direct speech-to-speech translation with a sequence-to-sequence model. arXiv preprint arXiv:1904.06037 (2019)
9. Joly, C.: Real-time voice cloning (2018). https://doi.org/10.5281/zenodo.1472609
10. Kons, Z., Shechtman, S., Sorin, A., Rabinovitz, C., Hoory, R.: High quality, lightweight and adaptable tts using lpcnet. arXiv preprint arXiv:1905.00590 (2019)
11. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Conference on Neural Information Processing Systems (NIPS), pp. 1–9 (2017)
12. Moseley, C.: The UNESCO atlas of the world's languages in danger: Context and process. World Oral Literature Project (2012)
13. Nag, S.: Early reading in kannada: the pace of acquisition of orthographic knowledge and phonemic awareness. J. Res. Reading **30**(1), 7–22 (2007)
14. Nautiyal, S., Rajasekaran, C., Varsha, N.: Cross-cultural ecological knowledge related to the use of plant biodiversity in the traditional health care systems in biligiriranga-swamy temple tiger reserve, karnataka. Medicinal Plants-Inter. J. Phytomed. Related Indus. **6**(4), 254–271 (2014)
15. Oord, A.v.d., et al.: Wavenet: a generative model for raw audio. arXiv preprint arXiv:1609.03499 (2016)
16. Openslr. https://www.openslr.org/79/, (Accessed 29 Sept 2023)
17. Ping, W., et al.: Deep voice 3: scaling text-to-speech with convolutional sequence learning. arXiv preprint arXiv:1710.07654 (2017)
18. Results page. https://style-transfer-five.vercel.app/, (Accessed 29th Sept 2023)
19. Shen, J., et al.: Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4779–4783. IEEE (2018)
20. Swadesh, M.: Lexico-statistic dating of prehistoric ethnic contacts: with special reference to north American indians and eskimos. Proc. Am. Philos. Soc. **96**(4), 452–463 (1952)
21. Wang, X., Feng, S., Zhu, J., Hasegawa-Johnson, M., Scharenborg, O.: Show and speak: directly synthesize spoken description of images. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4190–4194. IEEE (2021)
22. Wang, Y., et al.: Style tokens: unsupervised style modeling, control and transfer in end-to-end speech synthesis. In: International Conference on Machine Learning, pp. 5180–5189. PMLR (2018)
23. Yamamoto, R.: Wavenet vocoder (2018). https://github.com/r9y9/wavenet_vocoder