Cross Lingual Style Transfer Using Multiscale Loss Function for a Low Resource Tribal Language

Anonymous submission to INTERSPEECH 2023

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.

Index Terms: Index Terms: Low resource Cross-lingual Speech synthesis, Spoken language generation, Speaker Adaptation, Voice con-version, 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[1], and any language spoken by a population less than 10,000 people is considered a potentially endangered language by Govt of India[2]. Many tribal languages of India do not have literary tradition[3]. One such tribal language, Soliga is already on the verge of extinction as it has only a population of fewer than 40,000 people [4], 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[5]

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. Speech synthesis that uses deep neural networks has been popularly used for Text to Speech (TTS), and all of these models need transcripts for each speech file to train the models[6, 7, 8, 9]. 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[10, 11], 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[12] which is practically not feasible for low-resource languages, especially when the population is as less 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 [13]. 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 [14], 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 [15], 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 Figure 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[16]. 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 Figure 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 sec-

tion 3.1.

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

2. Data preparation

Soliga does not have any written literature in the Soliga language. We can consider Soliga as a zero-resource language to create a speech database. So, We have created our data-set for Soliga and Kannada languages. It was carried out in two steps. The first was to create text sentences, and the second step was to record these read sentences by the tribal literate person. To create text sentences, we used the words which are commonly used in Indian villages as listed by "swadesh"[17]. Using these words, we created about 10,000 English sentences, which are about 3 to 15 words long. These English sentences were translated to Kannada since kananda is the contact language for Soliga tribe, the literate Soliga people are more comfortable with Kananda language than English. Hence they were asked to translate Kannada sentences into Soliga sentences, but since Soliga does not have its script, the Soliga sentences were written down in Kannada script. Then, we identified a female Soliga speaker with a good voice and diction for a studio-quality voice recording. We have designed a user interface(UI) to record speech data, which also records metadata information of the speaker like age, gender, education, and qualification, followed by user consent for donating their voice. The sampling rate is 44Khz at which the speech data is collected in a recording studio. The sentences are displayed on the UI window to be read by the speaker, and there is an option to listen to each recording and rerecord in case of any wrong utterances. Likewise, we have collected 5000 speech voices of Both Kannada and Soliga speakers for our experiments, which are about 5Hrs duration in both languages. Additionally, we used publicly available Kannada male voice data https://www.openslr.org/79/ to train our model to increase the number of speakers and incorporate different styles for Soliga to Kannada voice conversion.

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[18], 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 [19]

The neural style transfer model's basic theoretical framework employs a straightforward combination of encoder and decoder networks. The decoder performs the function of a generator by producing the target output. The encoder's function is to preserve the source speech's linguistic details and eliminates speaker identity-related information. The generator(decoder) job is to combine the input speech signal's content and style with the target voice to generate the mel-spectrogram of a target speech signal. The encoder used in the proposed model will generate a shared-latent space called "z" that captures only the content of each sample while discarding the identity of the original speaker. The encoder architecture from the base paper remains unchanged in the proposed model.

We have considered two generators for our model as we have only two speakers for two languages, i.e. Kannada and Soliga. G_1 and G_2 include the same layers as the encoder's, but in the opposite order. We have introduced two more layers to down-scaling and up-scaling experiments and four more layers to up-down-scaling of the generated mel-spectrogram and compared it with the original mel-spectrogram to calculate the multi-scale loss function. The base paper generator without this modification gives a noisy output. Our approach gives better results. To transfer the style from Soliga to Kananda and vice versa, As depicted in Fig. 1, the latent space for two given decoders is switched.

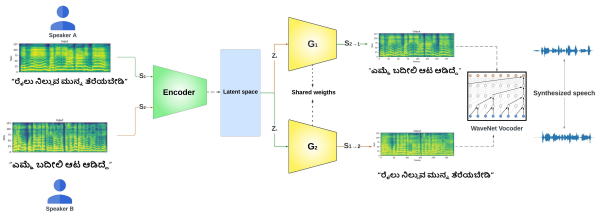

Figure 1: *The style transfer flow of Soliga and Kannada.*

3.1. Multi-scale Loss Function Experiments

In addition to conducting four experiments with the baseline model and multi-scale loss functions, we also incorporated the L_1 loss from the base paper. We then added additional multi-scale loss functions, including up-scaling, down-scaling, and the combination of up and down-scaling. After analyzing the results, we observed that the down-scaling experiment produced better intelligibility and voice quality than the baseline model. This suggests that incorporating a down-scaling multi-scale loss function can enhance the performance of the baseline model in terms of speech quality. The details are given below.

3.1.1. 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.

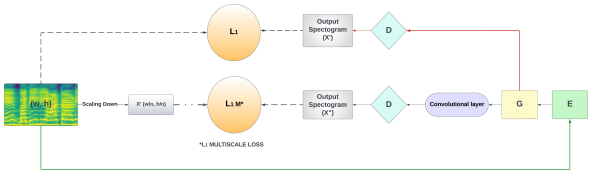

Figure 2: *Loss Function for Down-scaling.*

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

Here the GAN Loss of original input is calculated

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

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

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

3.1.2. *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.

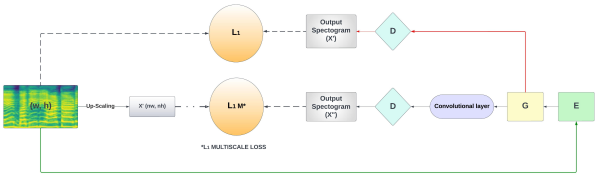

Figure 3: *Loss Function for Up-scaling.*

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

Here the GAN Loss of the original input is calculated,

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

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

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

3.1.3. *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.

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.

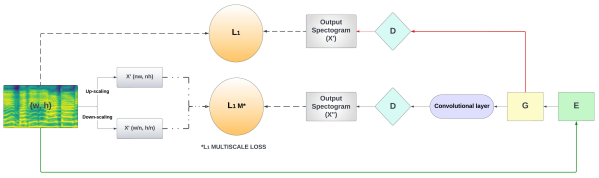

Figure 4: *Loss Function for Up and Down-scaling.*

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

Where W_i is the weight for each scale and L_{GAN_1} is the GAN Loss for original input from equation(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 equation(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.

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

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.

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

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.

4. Results and Discussion

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

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 reduced the error considerably. 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 `https://style-transfer-five.vercel.app/`

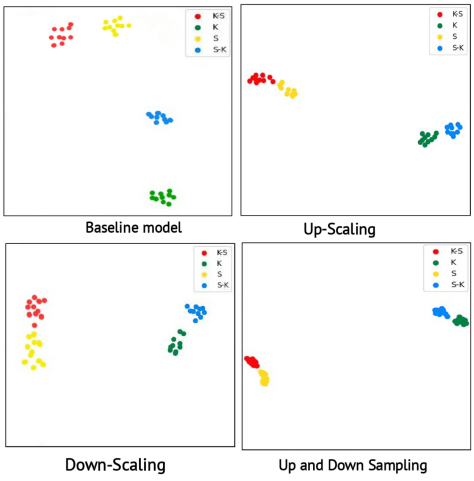

Figure 5: *Features embedding visualization for each speaker on the original and synthesized speech [20]*

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 figure 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.

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

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. 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.

5. Conclusion and Future work

In this paper, we proposed a cross-lingual speaker style transfer between two syntactically similar languages Kannada and Soliga. We have shown that multi-scale loss is vital in enabling the existing deep neural networks to learn and inject subject-specific style into the identity-independent feature embeddings. Our experimental results show that the introduction of multi-scale loss in the deep neural networks reduced generator noise and faithfully transferred Soliga speaker styles to Kannada speaker while simultaneously maintaining the content and style of the original speech. The proposed model shows promising results on voice conversion between different gender. However, there is a lot of scope for improvement in voice conversion between the same gender. In future work, we explore the possibilities of cross-linguistic speech style transfer between low-resource tribal languages and Indic languages. We believe our contribution may help explore various possibilities of cross-linguistic speech style transfer between low-resource and Indic languages.

6. References

[1] C. Moseley, *The UNESCO atlas of the world's languages in danger: Context and process*. World Oral Literature Project, 2012.

[2] C. Chandramouli and R. General, "Census of india 2011," *Provisional Population Totals. New Delhi: Government of India*, pp. 409–413, 2011.

[3] A. Dasare, K. Deepak, M. Prasanna, and K. S. Vijaya, "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)*. IEEE, 2022, pp. 1–6.

[4] S. Nautiyal, C. Rajasekaran, and N. Varsha, "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-International Journal of Phytomedicines and Related Industries*, vol. 6, no. 4, pp. 254–271, 2014.

[5] T. Haokip, "Artificial intelligence and endangered languages," *Available at SSRN 4212504*, 2022.

[6] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[7] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017.

[8] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[9] Z. Kons, S. Shechtman, A. Sorin, C. Rabinovitz, and R. Hoory, "High quality, lightweight and adaptable tts using lpcnet," *arXiv preprint arXiv:1905.00590*, 2019.

[10] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanevsky, and Y. Jia, "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.

[11] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," *arXiv preprint arXiv:1904.06037*, 2019.

[12] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[13] G. Ghiasi, H. Lee, M. Kudlur, V. Dumoulin, and J. Shlens, "Exploring the structure of a real-time, arbitrary neural artistic stylization network," *arXiv preprint arXiv:1705.06830*, 2017.

[14] X. Wang, S. Feng, J. Zhu, M. Hasegawa-Johnson, and O. Scharenborg, "Show and speak: Directly synthesize spoken description of images," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4190–4194.

[15] E. A. AlBadawy and S. Lyu, "Voice conversion using speech-to-speech neuro-style transfer." in *Interspeech*, 2020, pp. 4726–4730.

[16] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.

[17] M. Swadesh, "Lexico-statistic dating of prehistoric ethnic contacts: with special reference to north american indians and eskimos," *Proceedings of the American philosophical society*, vol. 96, no. 4, pp. 452–463, 1952.

[18] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *Advances in neural information processing systems*, vol. 30, 2017.

[19] R. Yamamoto, "Wavenet vocoder," https://github.com/r9y9/wavenet_vocoder, 2018.

[20] C. Joly, "Real-time voice cloning," https://doi.org/10.5281/zenodo.1472609, 2018.