



# Comparative study on multi-voice singing synthesize systems

Resna S.<sup>1</sup> and Rajeev Rajan<sup>2\*</sup>

<sup>1</sup>MultiMedia and Communication Vertical, Tata Elxsi, Technopark, Thiruvananthapuram, India

<sup>2</sup>College of Engineering, Trivandrum APJ Abdul Kalam Technological University, Trivandrum, India

\*Corresponding author: [rajeev@cet.ac.in](mailto:rajeev@cet.ac.in)

Accepted: 1<sup>st</sup> March 2023

OPEN ACCESS 

**Abstract:** In this paper, two multi-voice singing synthesis frameworks are compared. One proposed model consists of two blocks, namely, text-to-speech (TTS) converter and speech-to-singing (STS) converter. Synthesized speech is generated from lyrics for a target speaker's voice by TTS converter in the front-end. Later, a sung version is synthesized as per the given target-melody using encoder-decoder model in the STS module. We have compared our model with an existing multi-voice singing synthesis model, based on generative adversarial network (GAN) with phoneme synchronization information. The proposed system is systematically evaluated using subjective and objective tests. Three performance metrics, namely the mean opinion score (MOS), log spectral distance (LSD) have been analyzed as part of the study. Our study shows that the proposed model generates singing voices that adapt well to the target melody but the phonetic intelligibility is poor when compared to the baseline system.

**Keywords:** multi-voice, encoder-decoder, generative adversarial network, song adaptation

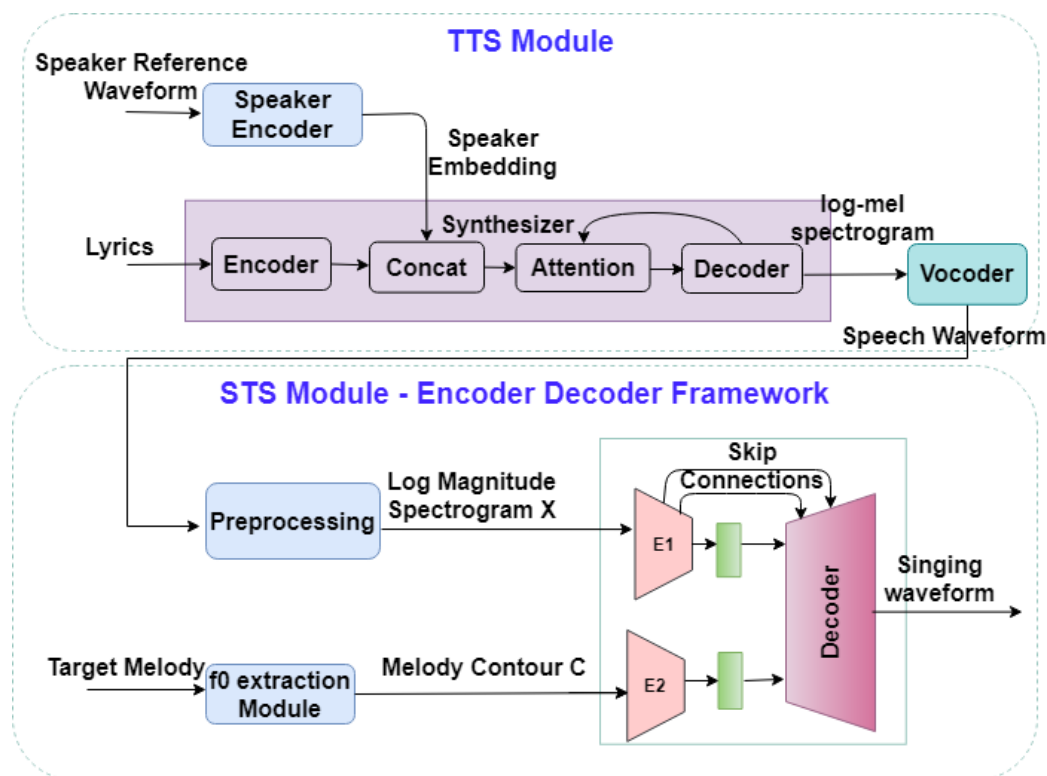
## Introduction

Music composers may wish to hear different melodic variations of their compositions to finalize the most attractive tunes. Besides, compositions may be sung in multiple singers to fix the best singing voice. Building systems to cater to these goals would result in interesting applications in the audio domain. Multi-voice singing synthesis has wide variety application in speech and music domains. This framework can be used in many deep learning applications as a n approach to data augmentation. A comparative study on two multi-voice singing framework is carried out in this paper. Multi-voice singing synthesis synthesize singing in target voice from lyrics. A target-speaker reference speech and target melody will be fed to the system as acoustic cues to compute the characteristics of the target speaker and melody -speaker's reference speech and a target melody.

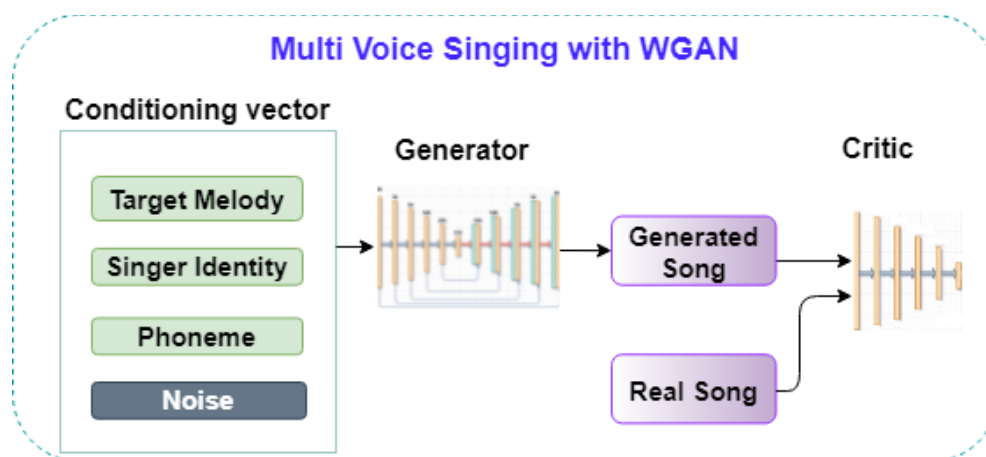
Singing voice synthesis has been studied in different aspects, including lyrics-to-singing alignment [1, 2], parametric synthesis [3], acoustic modeling [4], and

adversarial synthesis [5, 6]. Jinlong et al. [7] presented a lyrics-to-singing voice synthesis system with variable timbre based on Gaussian mixture model (GMM). TTS module converts text to speech followed by a melody control mechanism to synthesize song from speech. It is done by altering the acoustic parameters of speech. GMM-based singing voice morphing algorithm is employed to vary the timbre. Marc Freixes et al. [8] introduced a unit selection-based text-to-speech-and-singing synthesis framework, which integrates STS conversion to enable the generation of both speech and singing from an input text and a score.

DeepSinger, a multi-lingual multi-singer singing voice synthesis (SVS) system is also proposed using singing training data mined from music websites [9]. A lyrics-to-singing alignment model is designed to automatically extract the duration of each phoneme and further design a multi-lingual multi-singer singing model using feed-forward transformer and Griffin-Lim. The relationship between musical scores and their acoustic features was modeled to generate singing voice in [4]. Most of the



**Figure 2.** Proposed framework for multi-voice singing synthesis from lyrics (Model-1)



**Figure 1.** Multi-voice singing synthesis from lyrics based on WGAN (Model-2)

approaches in singing voice synthesis systems are mostly inspired by TTS and follow the basic components of TTS as building blocks [9]. In the proposed study, performance analysis is done on two models. We proposed a model (Model-1) by integrating multi speaker TTS synthesizer [10] and encoder-decoder framework [11] to synthesize singing voice. The second model (Model-2) is the state-of-the-art model proposed by [12]. Chandana et.al [12] employed WGAN approach for singing voice synthesis from a target melody.

The paper is organized as follows; initially the model architectures are explained, followed by the assessment

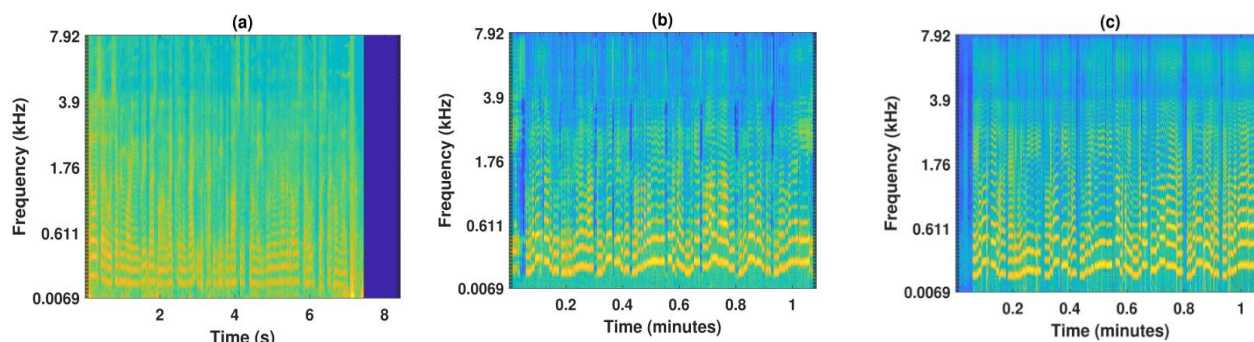
framework. Subsequently the results are analyzed. Finally, the paper concludes by giving inferences about the study.

## Model Architectures

The performance of two models, namely, Model-1 and Model-2 are investigated in detail using subjective and objective evaluations.

### Model-1

The block diagram of the Model-1 is shown in Figure



**Figure 3.** Mel-spectrograms (a) synthesized speech file, (b) generated music file: Model-1, (c) generated music file: Model-2

1. It consists of two modules namely, TTS (text-to-speech) module and STS (speech-to-singing) module. These standalone modules are available in the literature. We concatenated both modules to develop a new multi-voice singing synthesis model without the support of phoneme duration information of target singing.

#### *TTS Module*

TTS module [10] consists of three blocks, namely, recurrent speaker encoder, a sequence-to-sequence synthesizer and a vocoder as shown in Figure 1. A fixed dimensional d-vector is computed in the encoder. Speaker encoder is a speaker-discriminative framework trained on a speaker verification paradigm. Encoder maps a log-Mel spectrogram to d-vectors. 40 channel mel-spectrogram is processed by three stacked LSTMs. L2 normalization is applied at the output of the top layer to create the final embedding. Sequence-to-sequence synthesizer predicts a Mel-spectrogram from a sequence of grapheme or phoneme inputs, conditioned on the speaker embedding vector. The recurrent sequence-to-sequence with attention Tacotron 2 architecture is extended to support multiple speakers in the synthesizer. An embedding vector for the target speaker is concatenated with the synthesizer encoder output at each time step. The synthesizer is trained on pairs of text transcript and target audio. Vocoder converts the spectrogram into time-domain waveforms. Wavenet-based vocoder is used in our system [13]. Vocoder inverts synthesized mel-spectrograms emitted by the synthesis network into time-domain wave forms.

#### *STS Module*

An encoder-decoder framework [11] with spectra-to-spectra conversion is utilized for singing voice generation as illustrated in Figure 1. A vocal melody extractor [14], is used to extract melody contour from the inputted target melody, either humming or reference singing. Initially, silent frames are removed which aid the network to learn the alignment between speech and singing during training. The log magnitude spectrogram of the speech input is computed using a phase vocoder [15].

An encoder-decoder-based deep learning framework [11] produce two encodings, one for speech and another for the target melody obtained in the pre-processing stage. By using these encodings together, a sung version of the speech is produced using U-net [16] based network architecture. Finally, GriffinLim algorithm [17] is employed to reconstruct the waveform from the log magnitude-spectrogram.

For the ease of handling variable length speech signals, fully-convolutional architecture (1D) with GRU recurrent layers are used. Both time and frequency are down sampled by a factor of eight at encoder side and up sampled by a factor of eight at the decoder side. Skip connections between encoder E1 and decoder D are introduced, to control the gradient vanishing problem and to train deeper networks [11].

#### *Model-2*

GAN-based Model-2 is proposed in [12]. DCGAN-based Model2 architecture [18] is shown in Figure 2. Five convolution layers form the integral part of the encoder-decoder framework. Connections in the layers mimics the U-net architecture. The dependencies within in the block are modelled by a critic through analysing block of fixed length input. To ensure the dependence between consecutive blocks, overlap-add of consecutive blocks of output vocoder features are used, as described in [12]. Strided convolution is used for down sampling in the encoder and linear interpolation followed by normal convolution for up sampling is used in decoder [18]. The network process N-sized blocks of consecutive frames to produce same size of output. The WGAN training is done using the reconstruction loss, as specified in [12]. The inputs to this system consists of frame wise phoneme annotations, continuous fundamental frequency extracted using spectral autocorrelation (SAC) algorithm, singer identity that broadcast throughout the time dimension [19] and a noise vector. This input conditioning is similar to [20]. The WORLD vocoder is used [21] for acoustic modelling of the singing voice.

## Performance Evaluation

### Dataset

#### TTS

LibriSpeech data [22] is used for training and VoxCeleb1, and VoxCeleb2 [23] data is used for the validation for the encoder in WaveNet. LibriSpeech data is used for the evaluation of synthesizer and vocoder. Validation data is used to tune the pre-trained weights. Audio files of 2484 speakers with a duration of 820 hrs are part of LibriSpeech corpus. The audio files are sampled at sampling rate of 16kHz.

#### STS

We use the NUS-48E corpus [24], which consists of 48 popular English songs, sung by 12 singers both male and female. Each singer sings 4 different songs from a set of 20 songs, leading to a total of 169 min of recordings, with 25,474 phoneme annotations. We train the system using 10 out of 12 singers in NUS dataset. For testing we used two singers from NUS-48E corpus dataset and two singers never seen before.

## Experimental Setup

As we mentioned earlier, we have integrated two frameworks namely TTS and STS to obtain Model-1. Lyrics and target speakers voice are given as input to TTS framework. A speech is synthesized as output, like reading out the lyrics in target speaker's voice. This speech is fed to the STS encoder decoder model along with the target melody to generate the song. The output of Model-1 is compared with Model-2. Songs are synthesized in Model-1 and Model-2 by giving both singing and humming inputs as target melody. We synthesized songs with variable duration and maximum duration we tried is about 1 minute. The synthesized samples are shared at <https://rrs-mvs-official.github.io/SynthSamples/>

The performance of these models are evaluated using subjective and objective methodology. For objective evaluation, we need to have the ground truth songs sung by 4 singers, used for testing. For the singers in NUS dataset, ground truth versions of songs are available in the dataset itself. All the neural network architectures and audio processing framework for encoder-decoder framework are implemented using pytorch and librosa [16]. STFT is computed with 1,024-pt FFT size, 64ms window size, 16 milliseconds hop size and reused the phoneme dictionary in the dataset. Learning factor  $\lambda$  of 0.015 is chosen. Adam optimiser is adopted with initial learning rate 0.002. The network is trained for 14 epochs (1000 iterations each) with a batch size of 16.

A hop size of 5 ms is used for extracting the vocoder features and the conditioning for WGAN in Model-2. Block-size,  $N = 128$  frames, corresponding to 640 ms, a weight of recon = 0.0005 for Lrecon are used and trained the network for 3000 epochs. RMSProp is used for network optimization, with a learning rate of 0.0001. The mel-spectrograms of synthesized speech file, generated music files from Model-1 and Model-2 are shown in Figure 3.

## Evaluation Methodology

Subjective and objective evaluation has been carried out to assess the efficacy of the models. Subjective evaluation is conducted using a perception test. The audio files will be assessed by listening to the files by evaluators. Objective evaluation computes parameters from the synthesized files and evaluate to measure the quality of files.

### Objective Evaluation

We used Log spectral distance (LSD) for objective evaluation. It is computed by averaging the Euclidean distance between true and synthesized log-spectrogram frame over time for frequencies between 100 Hz to 3.5 kHz. LSD is defined as:

$$LSD = (1/2\pi) \sqrt{\int_{-\pi}^{\pi} (10 \log_{10} \left( \frac{P(w)}{Q(w)} \right)^2 dw} \quad (1)$$

Where,  $P(w)$ ,  $Q(w)$  represent power spectra of true and synthesized audio files, respectively.

### Subjective Evaluation

16 subjects evaluated the quality by playing the audio files. All the listeners are presented with target lyrics, target voices and target melody (both singing and humming) and synthesized songs using Model-1 and Model-2. We computed four perceptual metrics as follows;

#### Adaptation of song to target melody:

This metric is used to measure how good the generated song matches to the target melody. It measures the drift in the melody from the target

**Singing quality:** This metric focus on the quality of singing by considering the noise degradation and breaks in the singing. It is not mandatory that synthesized files are having the singing quality. It may end up in speech like sounds. These factors will be assessed by this metric.

**Phoneme quality of song:** Intelligibility is an important factor which affect the quality of the synthesized files. Phonetic quality is assessed by this metric. Phonemic variation may deteriorate the quality of synthesized voice.

**Voice adaptation of target singer:** Our task also evaluates whether the singing voice matches with that of target. Even though the voice characteristics of the target speaker learned, synthesized voice may vary from the characteristics of the target voice. Thus, evaluators are directed to evaluate the voice adaptation in of the generated voice. Five choicer are given from very low (1) to very high (5). These grades are later converted to a numerical score as 5 (Very high), 4 (high), 3 (medium), 2 (low), 1(very low).

## Results and Analysis

### Objective Evaluation

As mentioned earlier, we evaluated LSD scores for the models under study. We measured the LSD scores in two models, Model-1 and Model-2 with target melody as singing inputs as well as humming inputs. The results are tabulated in Table 1.

#### LSD

From Table 1, it is worth noting that LSD of 9.98dB is reported for Model-2 as compared to 14.62 dB for Model-1 in the case of singing melody input. The best system is the scheme with low LSD. The significant margin shows that Model-2 matches well to the naturalness of true audio files. The trend is same for the case of humming voice too. The scores for the test-audio files for both singing and humming are plotted in Figure 5 (upper pane). It can be seen that lowest LSD is reported for Model-2 for all the test cases.

**Table 1.** Objective evaluation metric LSD of Model-1 and Model-2.

Melody Input		Methods	
		Model-1	Model-2
Singing voice	LSD(dB)	14.62	<b>9.98</b>
Humming voice	LSD(dB)	15.45	<b>10.48</b>

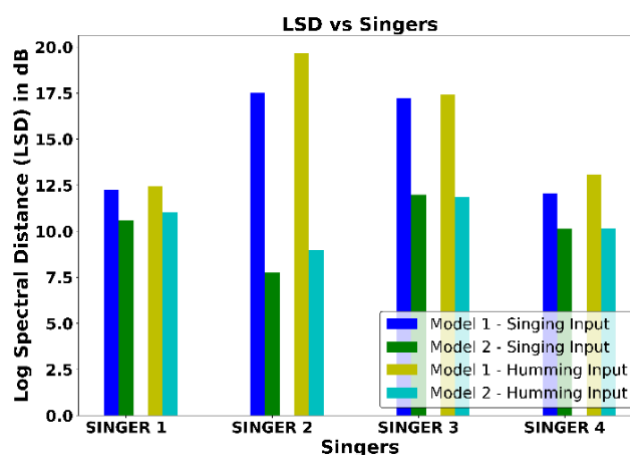
### Subjective Evaluation

Subjective evaluation scoring is performed using graphical user interface (GUI). The listeners were given guidelines and sufficient time to record their feed-backs in the interface. Listeners evaluated the quality of audio by considering four criteria, namely, adherence to target

melody, singing quality, phoneme clarity and voice quality. MOS scores are calculated for all the four criteria with four different singers each having five singing inputs and five humming inputs. Equal weight is given to all the evaluation-metric specific MOS (weight of 0.25). Total MOS is tabulated in Table 2 for both singing and humming. It can be observed that Model-2 performs better than Model-1. MOS scores obtained from 16 listeners for four criteria are shown in Figure 4. It is worth noting that MOS score is better for Model-2 in all evaluation metrics. It shows that overall quality of the audio samples synthesized by Model-2 dominates the files generated by Model-1.

**Table 2.** Mean opinion score for Model-1 and Model-2. Total MOS is computed as weighted combination of four metrics.

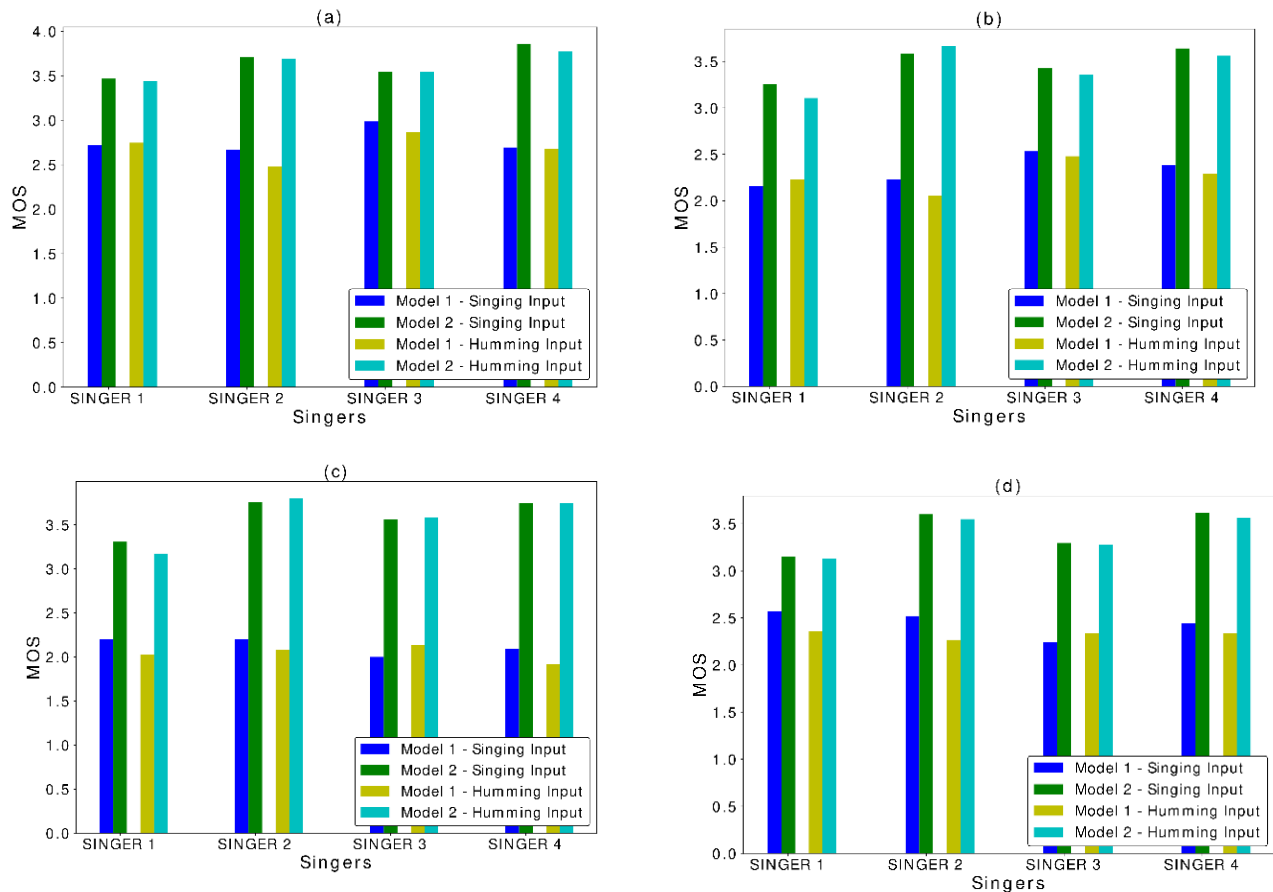
Melody Input	Methods	
	Model -1	Model-2
Singing voice	2.42	<b>3.53</b>
Humming voice	2.33	<b>3.50</b>



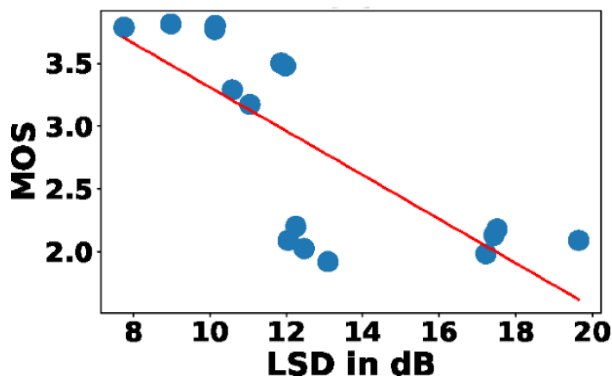
**Figure 5.** The objective evaluation metrics LSD

Also, we analysed the correlation between objective and subjective evaluations to validate the effectiveness of the evaluation procedure. As shown in Figure 6, LSD and phonetic quality shows a desired negative correlation. As LSD increases, phonetic quality decreases. The correlation study shows the effectiveness of the tool in the performance analysis. On comparing both models, melody transfer is more or less same, but the phonetic intelligibility is poor in Model-1. This is due to the absence of phonetic alignment information in Model-1.





**Figure 4.** MOS results (a) Song adaptation (b) Quality of singing (c) Phoneme quality of songs (d) Voice adaptation.



**Figure 6.** Correlation between MOS (Phoneme quality of songs) and LSD.

## Conclusion

The paper compares two models for multi-voice singing synthesis. The proposed model synthesizes singing voice without any phonetic alignment details. In addition, the models provide song in target melody fed to the system. The baseline model is the WGAN based multi-voice-singing-synthesis approach. We examined the

performance using subjective and objective parameters. Since WGAN model used phonetic alignment information for song generation, it had good phonetic intelligibility compared to our proposed model. As a future work, we are planning to enhance the phonetic intelligibility of our model by employing style transfer techniques.

## References

- [1] C. Gupta, H. Li and M. Goto, "Deep Learning Approaches in Topics of Singing Information Processing," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2422-2451, 2022. <https://doi.org/10.1109/TASLP.2022.3190732>
- [2] Hiromasa Fujihara, Masataka Goto, J.O., Okuno, H.G.: Lyric- synchronizer: Automatic synchronization system between musical audio signals and lyrics. *IEEE Journal of Selected Topics in Signal Processing* 5, 1252--1261 (2011). <https://doi.org/10.1109/JSTSP.2011.2159577>
- [3] J. Kim, H. Choi, J. Park, S. Kim, J. Kim, and M. Hahn., "Korean singing voice synthesis system based on an LSTM

recurrent neural network,” in Proc. of Interspeech, pp. 1551–1555, 2018.

<https://doi.org/10.21437/Interspeech.2018-1575>

[4] M. Nishimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “Singing voice synthesis based on deep neural networks.” in Proc. of Interspeech., pp. 2478–2482, 2016. <https://doi.org/10.21437/Interspeech.2016-1027>

[5] Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “Singing voice synthesis based on generative adversarial networks.” in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6955–6959, 2019. <https://doi.org/10.1109/ICASSP.2019.8683154>

[6] Lee, J., Choi, H.S., Jeon, C.B., Koo, J., Lee, K.: “Adversarial trained end-to-end Korean singing voice synthesis system”. arXiv preprint arXiv:1908.01919 (2019). <https://doi.org/10.21437/Interspeech.2019-1722>

[7] J. Li, H. Yang, W. Zhang, and L. Cai, “A lyrics to singing voice synthesis system with variable timbre,” Communications in Computer and Information Science., pp. 186–193, 2011. [https://doi.org/10.1007/978-3-642-23220-6\\_23](https://doi.org/10.1007/978-3-642-23220-6_23)

[8] M. Freixes, F. Alias, and J. C. Carrie, “A unit selection text-to-speech-and-singing synthesis framework from neutral speech: proof of concept,” EURASIP Journal on Audio, Speech, and Music Processing, vol. 2019, pp. 1–14, 2019. <https://doi.org/10.1186/s13636-019-0163-y>

[9] Y. Ren, X. Tan, T. Qin, J. Luan, Z. Zhao, and T.-Y. Liu, “Deepsinger: Singing voice synthesis with data mined from the web,” in Proc. of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining, pp. 1979–1989, 2020. <https://doi.org/10.1145/3394486.3403249>

[10] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Lopez-Moreno, and Y. Wu, “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” Advances in Neural Information Processing Systems 31 (2018), 4485–4495.

[11] J. Parekh, P. Rao, and Y. H. Yang, “Speech-to-singing conversion in an encoder-decoder framework,” in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 261–265, 2020. <https://doi.org/10.1109/ICASSP40776.2020.9054473>

[12] P. Chandna, M. Blaauw, J. Bonada, and E. Gomez, “WGANsing: A multi-voice singing voice synthesizer based on the wasserstein-GAN,” in Proc. of 27th European Signal Processing Conference, pp. 1–5, 2019. <https://doi.org/10.23919/EUSIPCO.2019.8903099>

[13] Resna, S., Rajan, R. “Multi-Voice Singing Synthesis from Lyrics”. Circuits Syst Signal Process 42, 307–321 2023. <https://doi.org/10.1007/s00034-022-02122-3>

[14] L. Su, “Vocal melody extraction using patch-based CNN,” in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 371–375, 2018. <https://doi.org/10.1109/ICASSP.2018.8462420>

[15] B. McFee, C. Raffel, D. Liang, D. Ellis, M. Mcvitar, E. Battenberg, and O. Nieto, “Librosa: Audio and music signal analysis in python,” in Proc. of 14th Python in Science Conference, pp. 18–24, 01 2015. <https://doi.org/10.25080/Majors-7b98e3ed-003>

[16] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234–241. Springer International Publishing, 2015. [https://doi.org/10.1007/978-3-319-24574-4\\_28](https://doi.org/10.1007/978-3-319-24574-4_28)

[17] D. Griffin and Jae Lim, “Signal estimation from modified short-time Fourier transform,” in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 2, pp. 236–243, April 1984. <https://doi.org/10.1109/TASSP.1984.1164317>

[18] Radford, A., Metz, L., Chintala, S.: “Unsupervised representation learning with deep convolutional generative adversarial networks.”, in Proc. of 4th International Conference on Learning Representations, ICLR, 2016.

[19] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: Wavenet: A generative model for raw audio. In: Arxiv (2016).

[20] Blaauw, Merlijn, and Jordi Bonada. “A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs” *Applied Sciences* 7, no. 12: 1313, 2017. <https://doi.org/10.3390/app7121313>

[21] Morise, M., Yokomori, F., Ozawa, K.: World: A vocoder-based high-quality speech synthesis system for real-time applications. IEICE Transactions on Information and Systems E99.D, 1877–1884 (2016).

<https://doi.org/10.1587/transinf.2015EDP7457>

[22] A. Pandey and D. Wang, "On Cross-Corpus Generalization of Deep Learning Based Speech Enhancement," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2489-2499, 2020.

<https://doi.org/10.1109/TASLP.2020.3016487>

[23] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech Language*, vol. 60, pp. 1010–27, 2020.

<https://doi.org/10.1016/j.csl.2019.101027>


[24] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, "The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech," in *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–9, 2013.

<https://doi.org/10.1109/APSIPA.2013.6694316>

---

**Publisher:** Chinese Institute of Automation Engineers (CIAE)

**ISSN:** 2223-9766 (Online)

 **Copyright:** The Author(s). This is an open access article distributed under the terms of the [Creative Commons Attribution License \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are cited.