Disentangling Overlapping Sources: Improving Vocal and Violin Source Separation in Carnatic Music

Adithi Shankar, Serafin Schweinitz, Genís Plaja-Roglans, Xavier Serra, Martín Rocamora
Music Technology Group, Universitat Pompeu Fabra. Barcelona, Spain
adithishankar.sivasankar@upf.edu

***Abstract*—Separating the individual elements in a music mixture is an important tool in computational musicology, allowing for an improved analysis of music repertoires. In the context of Carnatic music, this task remains a challenge given the suboptimal generalization of existing music source separation systems to this style. Although multi-stem Carnatic recordings exist, these are mostly collected from the mixing console in live performances. Therefore, there is an unintended presence of other sources in the background of the audio signal of an individual instrument. Another challenge for Carnatic music is the strong melodic correlation between the singing voice and the violin, two sources widely found in live performances of this repertoire. Existing strategies to address such problems struggle with source quality and only consider vocals. In this work, we propose to incorporate two components in the regular training scheme of a source separation network, namely a learned loss and a mixer model, to account for the source bleeding. We achieve improved separation while extending the separation targets to the violin, an important source in the repertoire, and therefore cover the separation of the most common melodic components in Carnatic Music. Code and models are available in `compiam`.**

***Index Terms*—Music Source Separation, Carnatic music, Source Bleeding, Violin Separation.**

I. Introduction

Music source separation (MSS) is concerned with automatically extracting individual elements in a musical mixture [1]. In a computational musicology context, MSS is an important tool to isolate certain sources in music signals for a more reliable analysis [2]. The state-of-the-art in source separation is currently led by neural networks that are trained with clean and aligned multi-stem datasets [3], which tend to include a very restricted music style distribution. Consequently, the publicly released models may not generalize to out-of-domain signals [4]. Moreover, these are typically restricted to a {*vocals, bass, drum, other*} music instrument arrangement.

In this work, we address the MSS problem for Carnatic music, an important music style originated in South India. Conveniently, a large corpus of music, metadata, and time-aligned annotations is available for Carnatic music [5]. Moreover, for a portion of this collection, multi-stem recordings are available for research purposes under the name of Saraga Carnatic dataset [6]. However, since Carnatic music is mostly enjoyed live, these tracks are directly recorded through the mixing console in live performances, which leads to *source bleeding* in the individual stems. We define source bleeding as the unintended presence of other sources in the background of the audio signal of an individual instrument. This kind of artifact hinders proper optimization of MSS networks [3].

Carnatic-tailored separation efforts have been done [4], aiming at outperforming the open systems in the literature that are generally used in the context of the computational analysis of Carnatic and Hindustani music [7], [8]. Several studies have pointed out the suboptimal generalization of openly available pre-trained models for Carnatic music examples [2], [4], [9], [10]. Works such as [4] propose different *bleeding-aware* techniques to improve over existing pre-trained models for singing voice separation (SVS). However, these still face two main challenges: **(1)** the separation quality is not leveling the latest in the literature, and **(2)** singing voice is the only source being currently separated. In certain music repertoires such as Carnatic music, instruments like the violin or mridangam (the main percussion instrument in the style) are crucial but have not been included in any separation attempt yet.

This work is a research effort toward bleeding-aware MSS focusing on two objectives: **(1)** improve the separation quality for the singing voice, and **(2)** extend the separation targets to violin, the most common melodic accompaniment instrument in Carnatic music, which normally has a prominent overlap with the singing voice, adding an extra layer of complexity to its analysis [11]. We first pre-train a separator network using the multi-stem data containing bleeding, leveraging the inherent knowledge of the Carnatic domain. The pre-trained separator is subsequently fine-tuned using an auxiliary loss based on the bleeding level in the pre-separated sources. The bleeding level is predicted using a bleeding estimator network which we pre-train using a small set of bleeding-free multi-stem data. However, this approach relies on the pre-trained separator to perform – even if constrained by the inherent limitations imposed by the source bleeding in the individual stems – an initial separation of the source.

While the approach based on the bleeding estimator loss shows positive impact for singing voice separation, preliminary experiments show that it is sufficiently effective for the violin source. The loudness differences between the vocal and violin stems may be a potential cause. In Carnatic music, the lead performer is typically the vocalist, and it tends to be accompanied by a violinist who closely follows its melody with slight variations, subtle delay, and lower loudness for aesthetic effect. This intricate interplay intertwines the vocal and violin, creating a complex challenge for analyzing this instrument.

Aiming at addressing the limitation of the bleeding estimator loss for the violin, we propose to use an alternative system composed of **(1)** two separator networks for the vocal and the violin stem respectively, and **(2)** a *mixer* network [12], which is aimed at exploiting the knowledge between tracks to reduce the bleeding in the outputs of the separators. The system is trained using only the multi-stem signals with bleeding, therefore no clean multi-stem recordings are needed. Moreover, no pre-training is required. Our evaluation experiments indicate that the learned loss approach outperforms existing baselines for vocal separation, while the mixer model allows satisfactory separation of the violin stem. The code and pre-trained models are made available for reproducibility. Additionally, we make an out-of-the-box implementation available through the `compIAM` Python library.

II. Method

A. Baselines

Training an MSS model using multi-stem data with bleeding has a limited achievable performance imposed by the intrinsic bleeding in the data. Fine-tuning pre-trained separation models using these

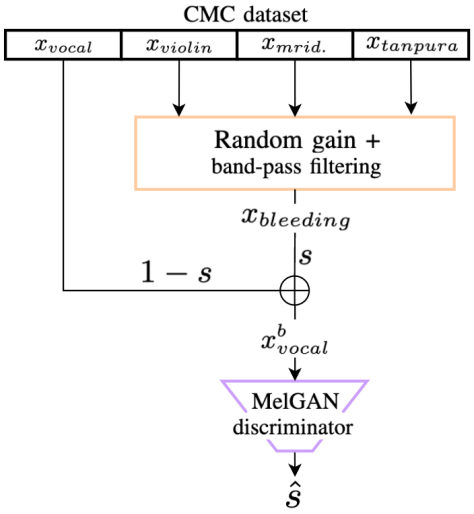

Fig. 1. Training process of the bleeding estimator. First, we process and combine the non-vocal stems. We sample a bleeding level s, and create the artificial sample of vocals with bleeding. Having ground-truth s at hand, we optimize the MelGAN Discriminator to estimate the bleeding level.

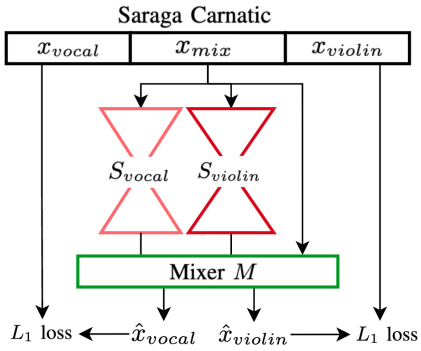

Fig. 2. Training process of the mixer model. The mixture x_{mix} is the input of the separators S_{vocal} and S_{violin}, and also of the mixer M, which refines the pre-separated outputs of both S models, reducing the induced source bleeding.

datasets has also been explored [13], however, in this case, the bleeding in the ground truth still poses a challenge and the separated sources lack quality and cleanliness from interferences. These experiments suggest that, in order to take complete advantage of the multi-stem data with bleeding, we need strategies to identify the presence of bleeding and suppress it.

Existing works have attempted to use the Saraga dataset, aiming at capturing the instrument timbres and practices of Carnatic Music from the audio signals, while exploring methods to overcome the bleeding problem. [4] proposes a training strategy inspired by generative cold diffusion to train a separator network using Saraga Carnatic, while detecting the spectrogram bins corresponding to the bleeding. Although improving on interference removal, said strategy has a negative impact on the quality of the separated vocals, especially on the high-frequency end and for effects such as reverberation.

To explore the actual effect of training a network to perform MSS using data with bleeding and establish a baseline for the proposed bleeding-aware techniques, we rely on a base separator network, denoted S, which in this work is a TFC-TDF-Net [12]. This convolutional U-Net downsamples the two-dimensional spectrogram of the mixture audio with 5 blocks of 2D convolutions, batch normalization and a ReLU activation functions. After each downsampling unit, a TFC-TDF block is applied. These combine Time-Frequency Convolutions (TFC) with Time-Domain Filters (TDF) to effectively capture both spectral and temporal features. This architecture is selected due to its demonstrated stability, rapid convergence, and outstanding vocal separation quality.

B. A fine-tuning loss based on the bleeding level

Aiming at leveraging the inherent knowledge in the multi-stem data with bleeding, we first pre-train a separator model S using solely these recordings, followed by a fine-tuning stage using an auxiliary loss that penalizes the presence of bleeding. This loss term is computed by a side network, denoted *bleeding estimator*, which we separately train to estimate the level of bleeding in an audio signal. Preliminary experiments show that pre-training a separator model S using data with bleeding may be able to preserve the target source while partially removing the rest [4], letting us assume that the distinction between the sources has been learned to an extent that may be advantageous.

In speech denoising and enhancement, learned loss functions [14] modeling speech quality metrics, e.g. PESQ or STOI, have helped improve separation performance [15]. In [16], the quality metrics are directly predicted from a corrupted audio file without the need for a clean reference track. During training, different kinds of noise are added to the clean speech signal, and PESQ and STOI metrics are calculated for the clean voice. A dense convolutional network is trained to predict these metrics only from the noisy audio signal.

To address the problem of learning the source bleeding, we propose a training process that reformulates the source separation training objective. First, the model minimizes the L1 distance to the reference tracks affected by bleeding until the loss stabilizes. Then, we use a pre-trained bleeding estimator model as a learned loss function for fine-tuning. This bleeding estimator predicts the amount of bleeding as a scalar between 0 (*no bleeding*) and 1 (*only bleeding*) directly from audio and can, therefore, be used as a loss function that does not depend on clean reference tracks [16].

Following this approach, we train a bleeding estimator on artificially simulated bleeding derived from a second, smaller, clean multi-stem dataset. In Fig. 1 this process is depicted for the vocal source. During training, a bleeding level $s \in [0, 1]$ is sampled for every training step. We sample the instrumentation stems of a clean multi-stem recording and process these using random gain and band-pass filtering, following the SDX bleeding challenge data generation [3]. We denote the resulting signal $x_{bleeding}$. We normalize in loudness the vocal signal, denoted x_{vocal}, and also the $x_{bleeding}$, and mix them together regarding the bleeding level s:

$$x^b_{vocal} = (1-s) \times \text{norm}(x_{vocal}) + s \times \text{norm}(x_{bleeding}) \quad (1)$$

where x_{vocal} is the vocal signal, $x_{bleeding}$ is the bleeding instrumental mix, and s is the bleeding factor (see Fig. 1). To estimate the bleeding factor s, we employ the multi-level waveform discriminator architecture from MelGAN [17]. This process may be applied to other sources by permuting the signals that are mixed to compute $x_{bleeding}$.

C. Multi-source mixer model

In the context of SVS for Carnatic music, one of the major challenges is the presence of violin accompaniment. The violin closely follows the vocalist throughout the performance, often mirroring the same fundamental frequency This constant overlap in pitch creates significant difficulties in isolating the vocal signal from the violin accompaniment. To address this challenge, we aim at isolating both

TABLE I
PERFORMANCE METRICS FOR VARIOUS SEPARATION MODELS ON VOCAL AND VIOLIN AUDIO STEMS.

Model	Vocals			Violin		
	SI-SDR	SIR	SAR	SI-SDR	SIR	SAR
ColdDiffSep [4]	4.03	-2.31	1.20	✗	✗	✗
TFC-TDF Net	6.45	**15.40**	5.60	-0.43	-7.97	-3.91
TFC-TDF Net with Learned Loss, CMC	**7.92**	4.15	0.00	✗	✗	✗

the vocal and violin stems together. This approach not only aims to improve SVS but also facilitates a more in-depth analysis of the violin, recognizing its vital role as a melodic accompaniment in Carnatic music.

We first propose reproducing the training pipeline in Sec. II-B for the violin stem. However, as seen in Tab. I, the baseline for Carnatic violin separation is not capable of learning the task as successfully. This may be due to the loudness difference. For instance, the average signal power of the vocal stem within all songs containing violin in the *CMC* dataset (see Sec. III-A2) is ≈ 2.8 times higher than the power of the violin stem. The bleeding estimator is effective when the pre-trained model achieves partial separation of the violin with some bleeding. However, when the model fails to perform meaningful separation, the bleeding estimator becomes ineffective.

To overcome this limitation, we hypothesize that sharing information between sources, and especially singing voice and violin, may disentangle these intertwining sources and potentially reduce the bleeding in the separated signals. We explore the potential of the so-called *mixer* network proposed in [12] as a disentangler and bleeding suppressor.

We define multiple separators S_{src} as defined in Sec. II-A, with one dedicated to each melodic source we target, namely vocals, denoted S_{vocal}, and violin, denoted S_{violin}. Each separator receives a musical mixture x_{mix} as input and outputs the corresponding separated source. However, in this case, a single separator S does not have information about the other sources, as it solely targets a single source from the mixture individually. Let M be the mixer model, which consists of a single and trainable linear layer. Then, the entire system operates as follows:

$$\hat{x}_{vocal}, \hat{x}_{violin} = M(x_{mix}, S_{vocal}(x_{mix}), S_{violin}(x_{mix})) \quad (2)$$

Therefore, the mixer M has 2 extra input channels when operating on stereo, and 1 extra input channel when the data is monophonic. In other words, M takes, as input, the pre-separated sources $\hat{x}_{vocal}$ and $\hat{x}_{violin}$, and the corresponding mixture x_{mix}, and refines the separations by relying on the shared information between sources. We use only multi-stem data with bleeding for this training process. Note again that theoretically, the glass ceiling of both S models is the corresponding source with the bleeding as noticeable as in the ground-truth data.

III. EXPERIMENTS

A. Datasets

1) Saraga Carnatic: To our best knowledge, Saraga Carnatic [6], and its audiovisual counterpart [13] are the largest openly available research collections for Carnatic Music, containing ≈ 60 h of multi-stem data, including vocals, violin, mridangam, and occasionally ghatam. These data have been collected in live performances hence the individual tracks are not completely clean, but there is source leakage or bleeding in the background, which makes these recordings not ideal for training MSS models. For this work, we rely on Saraga Carnatic [6] only, pursuing consistency with the training data of the compared models. However, we hypothesize that including the Audiovisual counterpart would contribute to a better performance. We use the 168 multi-stem recordings in Saraga Carnatic, which totals ≈21 hours of music.

2) Carnatic Multi-stem Clean (CMC): This private collection of 58 tracks amounts to ≈5 h of multi-stem Carnatic music recordings, including lead vocals, violin, mridangam, tanpura, and any additional instruments that may be featured. Notably, each instrument is recorded separately, ensuring that no bleed occurs between tracks.

3) Sanidha: This is an open dataset of ≈ 8 h of clean multi-stem Carnatic recordings including 5 concerts [18]. It also provides video recordings for each performer. Similarly to CMC, given the limited amount of recording time, and diversity of artists and recording setups, we hypothesize that these data may not be sufficient for a supervised training of a MSS model. However, it is a valuable resource for evaluating the separation systems, therefore we rely on Sanidha as an additional testing set.

B. Experimental setup

1) Bleeding Estimator Fine-tuning: We train a TFC-TDF U-Net with pairs of mixture audio signals and the corresponding target source, namely the vocal or the violin stem with bleeding. The model has 10.2M trainable parameters. We compute L1 loss between the separations and the target signals that include bleeding. We train the model using a learning rate of 0.0004 and RMSprop optimizer until the training process converges after 300k iterations. The loss estimator is trained on source stems containing artificially added bleeding labeled with the bleeding level, as described in Sec II-B.

The model comprises three identical convolutional neural networks that operate on the input signal at the original sampling rate of 24kHz, as well as downsampled versions at 12kHz and 6kHz. Each CNN contains 6 layers of 1D-convolutions with kernel sizes: $(41, 41, 41, 41, 5, 3)$, channels: $(16, 64, 256, 1024, 1024, 1024)$ and strides: $(4, 4, 4, 4, 2, 2)$. The convolutional layers are followed by an average pooling operation with strides of 2. Additionally, a LeakyRELU activation function follows the first 4 convolution layers. A final average-pooling operation combines the predictions of the 3 models. To constrain the output within the range [0,1] we apply a sigmoid activation function. The model has 16.9M parameters. We compute L2 loss between predictions and ground truths and train until convergence for 20k steps using Adam optimizer and a learning rate of 0.0001. We fine-tune the preceding experiment using our loss estimator trained on 58 tracks of the CMC dataset. We use RMSprop optimizer with a learning rate of 0.000001 and fine-tune for 1k iterations. Finally, we also explore low-data resource scenarios by training the bleeding estimator with just 15 and 5 tracks and fine-tuning the proceeding experiment using these bleeding estimators.

TABLE II
COMPARISON, ON TWO DIFFERENT TEST SETS, OF THE PRE-TRAINED TFC-TDF NET USING SARAGA OPTIONALLY FINE-TUNED USING THE BLEEDING ESTIMATOR LEARNED LOSS, WHICH IS TRAINED USING DIFFERENT SIZES OF CLEAN DATASETS (5, 15, AND 58).

		Sanidha			**CMC**		
Learned Loss	No. of Tracks	SI-SDR	SIR	SAR	SI-SDR	SIR	SAR
✗	✗	5.37	7.81	-0.24	6.45	**15.40**	**5.60**
✓	58	**6.70**	14.75	-0.25	**7.92**	4.15	0.00
✓	15	6.40	9.37	-0.84	6.81	**8.65**	0.00
✓	5	6.19	**15.74**	**-0.15**	7.08	4.68	-0.01

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT MODELS FOR VIOLIN SEPARATION.

	Violin		
Model	SI-SDR	SIR	SAR
TFC-TDF Net	-0.43	-7.97	-3.91
TFC-TDF Net with Mixer	**3.05**	**12.33**	**3.09**

2) Mixer model: We train the multi-source mixer model on the Saraga dataset [6] for 750k iterations. In each training step, we sum the instrumental stems in a Saraga recording and mix them to the *vocal* stem with a decibel ratio, randomly sampled between -5 and 5. We normalize the input mixture and scale the loudness of the violin and vocal stems, which are used as targets accordingly. We compute L1 loss on the separated vocal and violin against the corresponding targets. This loss is optimized using RMSprop and a learning rate of 0.0002, as suggested in [12]. The model has 20.4M parameters.

IV. RESULTS

See Tab. I for a report of the experiments on the bleeding estimator learned loss. The system in [4] serves as a baseline model for SVS in this evaluation, trained solely on vocal data from the Saraga dataset. As a result, it achieves an SI-SDR of 4.03 dB for vocals, while the SIR and SAR scores are -2.31 dB and 1.20 dB respectively. To the best of our knowledge, there is currently no dedicated violin separation model in the literature that could serve as a baseline for comparison. The baseline TFC-TDF Net model shows a moderate improvement to separate the vocals, achieving an SI-SDR of 6.45 dB, an SIR of 15.40 dB, and an SAR of 5.60 dB. Fine-tuning the pre-trained separation with the learned loss trained on the complete CMC dataset boosts the SI-SDR to 7.92 dB for the vocal stem.

In Tab. II we report an ablation study on the bleeding estimator loss. We report low-resource experiments testing the dependence of the system on clean multi-stem data. Moreover, we compute the metrics on the Sanidha dataset, in order to neglect the bias that may arise from evaluating the system on the same data that is used to train the bleeding estimator. The difference in SI-SDR performance between the 5 and the 58-track bleeding estimator is of ≈ 0.5 dB for the Sanidha dataset, and ≈ 0.8 dB for CMC, which suggests that the proposed approach provides a notable improvement under limited data conditions. This is a convenient property in the context of MSS, given the notable complexity of compiling clean multi-stem dataset, especially for repertoires that are generally enjoyed live. The dataset shift does not imply an important performance loss, in fact, we observe an improvement on interference removal.

The baseline TFC-TDF Net performance on the violin stem is notably poor, with a negative SI-SDR (-0.43 dB) and SIR (-7.97 dB). These results indicate high interference and distortion in the separated violin track. Therefore, we conclude that the trained TFC-TDF Net struggles to discriminate between vocals and violin, particularly in isolating the violin stem. In this scenario where the pre-trained separator network is not capable of performing a satisfactory preliminary separation, fine-tuning with the bleeding estimator becomes ineffective.

The trained mixer model allows the separation of the violin, achieving an SI-SDR of 3.05 dB, as seen in Tab. III. These results suggest that the mixer M may rely on the shared information between individual sources to reduce the bleeding and improve the source quality for the violin, substantially improving the baseline. Finally, as suggested by the SIR results, the separations of the violin are considerably clean from interferences. See our online demo page for separation examples and relevant links.[1]

V. CONCLUSION

This work explores bleeding-aware techniques for music source separation in Carnatic music, specifically designed to address the challenges posed by overlapping melodic sources in multi-stem live recordings. By enabling the separation of both vocals and violin—two fundamental and often intertwining melodic components in Carnatic repertoire—this research extends the limitations of existing separation frameworks, which have traditionally focused only on vocal source separation. We propose a bleeding estimator loss for fine-tuning a pre-trained separator. Our findings demonstrate that the bleeding estimator, when fine-tuned with a limited number of training tracks, significantly enhances vocal separation, providing a marked improvement over baseline models. However, while the bleeding estimator effectively addresses vocal separation, the challenge of isolating the violin stem remains. In this regard, we explore the potential of using a Mixer model for separation. By leveraging shared information between the vocal and violin sources, the mixer model enables better separation of both stems, particularly the violin, which is otherwise difficult to disentangle from the vocals due to their overlapping melodic content. This model provides a substantial improvement in the separation of the violin. By informally listening to separation examples we confirm the obtained metrics, especially the capability of the model to remove interferences from other sources. To further investigate these perceptual improvements, we plan to conduct more rigorous listening tests, which will provide a deeper understanding of the separation quality, especially for the violin stem.

[1] https://mtg.github.io/violin-vocal-sep/

REFERENCES

[1] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Proc. of the 18th Int. Society for Music Information Retrieval Conf. (ISMIR),* Suzhou, China, 2017, pp. 745–751.

[2] T. Nuttall, G. Plaja-Roglans, L. Pearson, and X. Serra, "In search of sañcāras: tradition-informed repeated melodic pattern recognition in carnatic music," in *Proc. of the 23rd Int. Society for Music Information Retrieval Conf. (ISMIR),* Bengaluru, India, 2022, pp. 337–344.

[3] G. Fabbro, S. Uhlich, C.-H. Lai, W. Choi, M. Martínez-Ramírez, W. Liao, I. Gadelha, G. Ramos, E. Hsu, H. Rodrigues *et al.*, "The Sound Demixing Challenge 2023: Music Demixing Track," 2023. [Online]. Available: http://arxiv.org/abs/2308.06979

[4] G. Plaja-Roglans, M. Miron, A. Shankar, and X. Serra, "Carnatic singing voice separation using cold diffusion on training data with bleeding," in *24th Int. Society for Music Information Retrieval Conf. (ISMIR),* Milano, Italy, 2023.

[5] A. Srinivasamurthy, G. K. Koduri, S. Gulati, V. Ishwar, and X. Serra, "Corpora for music information research in indian art music," in *Proc. of the Int. Computer Music Conf. (ICMC),* Athens, Greece, 2014.

[6] A. Srinivasamurthy, S. Gulati, R. C. Repetto, and X. Serra, "Saraga: Open datasets for research on indian art music," *Empirical Musicology Review*, vol. 16, no. 1, pp. 85–98, 2021.

[7] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, pp. 1–4, 2020.

[8] A. Défossez, "Hybrid spectrogram and waveform source separation," 2021. [Online]. Available: http://arxiv.org/abs/2111.03600

[9] M. Clayton, P. Rao, N. N. Shikarpur, S. Roychowdhury, and J. Li, "Raga classification from vocal performances using multimodal analysis." in *Proc. of the 23rd Int. Society for Music Information Retrieval (ISMIR),* Bengaluru, India, 2022, pp. 283–290.

[10] T. Nuttall, G. Plaja-Roglans, L. Pearson, and X. Serra, "The matrix profile for motif discovery in audio-an example application in carnatic music," in *Int. Symposium on Computer Music Multidisciplinary Research,* Tokyo, Japan, 2021, pp. 228–237.

[11] G. Plaja-Roglans, T. Nuttall, L. Pearson, X. Serra, and M. Miron, "Repertoire-specific vocal pitch data generation for improved melodic analysis of carnatic music," *Transactions of the Int. Society for Music Information Retrieval*, vol. 6, no. 1, pp. 13–26, 2023.

[12] M. Kim, W. Choi, J. Chung, D. Lee, and S. Jung, "KUIELab-MDX-Net: a two-stream neural network for music demixing," 2021. [Online]. Available: http://arxiv.org/abs/2111.12203

[13] A. Shankar, G. Plaja-Roglans, T. Nuttall, M. Rocamora, and X. Serra, "Saraga audiovisual: a large multimodal open data collection for the analysis of carnatic music, San Francisco, United States," in *Proc. of the 25th Int. Society for Music Information Retrieval Conf.*, 2024. [Online]. Available: http://hdl.handle.net/10230/68399

[14] S.-W. Fu, C.-F. Liao, and Y. Tsao, "Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality," *IEEE Signal Processing Letters*, vol. 27, p. 26–30, 2020. [Online]. Available: http://dx.doi.org/10.1109/LSP.2019.2953810

[15] H. Z. Xin Bai, Xueliang Zhang and H. Huang, "Perceptual loss function for speech enhancement based on generative adversarial learning," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022.

[16] Y. L. Wuxuan Gong, Jing Wang and H. Yang, "A no-reference speech quality assessment method based on neural network with densely connected convolutional architecture," in *INTERSPEECH 2023, Dublin, Ireland*, 2023.

[17] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brebisson, Y. Bengio, and A. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," 2019.

[18] V. V. Krishnan, N. Alben, A. A. Nair, and N. Condit-Schultz, "Sanidha: A studio quality multi-modal dataset for carnatic music, San Francisco, United States," in *Proc. of the 25th Int. Society for Music Information Retrieval Conf.*, 2024. [Online]. Available: http://arxiv.org/abs/2501.06959