SepACap: Source Separation for A Cappella Music

Luca A. Lanzendörfer* **Constantin Pinkl*** **Florian Grötschla** **Roger Wattenhofer**

ETH Zurich

{lanzendoerfer, cpinkl, fgroetschla, wattenhofer}@ethz.ch

Abstract

In this work, we study the task of multi-singer separation in a cappella music, where the number of active singers varies across mixtures. To address this, we use a power set-based data augmentation strategy that expands limited multi-singer datasets into exponentially more training samples. To separate the singers we introduce SepACap, an adaptation of SepReformer, a state-of-the-art speaker separation model architecture. We adapt the model with periodic activations and a composite loss function that remains effective when stems are silent, enabling robust detection and separation. Experiments on the JaCappella dataset demonstrate that our approach achieves state-of-the-art performance in both full-ensemble and subset singer separation scenarios, outperforming spectrogram-based baselines while generalizing to realistic mixtures with varying numbers of singers.

1 Introduction

The field of music source separation has recently seen rapid progress [5, 17, 12, 22], primarily focusing on separating different instruments from each other by learning spectral masks that isolate the desired source from the mixture. In contrast, separating multiple singers in purely vocal (a cappella) recordings remains underexplored. A cappella ensembles range from small groups (duets and quartets) to chamber choirs, often organized by part (soprano, alto, tenor, and bass), as well as subset variants [15, 2, 8]. These recordings exhibit dense harmonic overlap, frequent unison/octave doubling within a part, voice crossing, vibrato and portamento, tightly aligned consonant onsets, and breath and sibilant noise. These properties reduce timbral diversity and make source separation more challenging than instrument-wise separation [4]. Contemporary a cappella may also include vocal percussion or beatboxing, further increasing spectral overlap despite all sources being human voices [6, 23].

Historically, research on music source separation has been broadly split into two categories: frequency-based masking and direct waveform modeling. In the time-frequency domain, mask-based approaches predict source-specific masks applied to the mixture, where early CNN and RNN models have been surpassed by Transformer architectures that better capture spectral structure. Band-Split RoFormer [22] interleaves time and frequency Transformers with a subband projection, winning the Sound Demixing Challenge 2023 [7], while Mel-Band RoFormer [22] replaces heuristic bands with mel-scale projections to yield overlapping subbands aligned with perception, improving vocal separation and melody transcription. TF-Locoformer [18] further shows that local convolution plus global Transformer modeling in the time-frequency domain outperforms dual-path RNNs. In contrast, waveform-domain methods avoid masks and reconstruct sources directly: Conv-TasNet [11] learns a latent time-domain encoder and decoder, Demucs [5] and hybrid-Demucs [17] adopt a U-net waveform autoencoder to separate instrument stems. Compared to instrument separation, the field of singer separation has received comparably less attention despite applications in transcription,

*Equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Music.

remixing, and choir analysis. Furthermore, multi-singer datasets are limited: JaCappella [14] contains six vocal stems, containing a total of 34 minutes gathered from 35 songs across diverse genres, while choral and ensemble resources are typically even more limited, such as the Dagstuhl ChoirSet [16] containing choral singing with 7 minutes from 3 songs, and ESMUC Choir [3] containing 31 minutes of music from 3 songs.

In this work, we present SepACap, an adaptation of the recently proposed SepReformer [20], a state-of-the-art speaker separation model, for the a cappella setting where the number of active singers may vary across mixtures. We adapt the model by (i) introducing periodic activations [25] as we find them to perform better compared to the default ReLU activations; (ii) replacing the default training loss with a silence-aware composite objective that remains well-defined when stems are absent, combining waveform, multi-scale mel, and multi-resolution spectral losses; and (iii) coupling the model with a power set-based data augmentation scheme that creates mixtures for all subsets of stems, enabling joint separation and detection of active singers. We evaluate SepACap on JaCappella [14] in two scenarios: all-stems and subset-stems, and report both separation quality and silent-stem suppression quality, along with detection metrics. Across these settings, SepACap outperforms strong spectrogram-masking and waveform baselines, achieving state-of-the-art results in full-ensemble separation while markedly reducing bleed-through and correctly outputting silence for inactive stems.

Our contributions can be summarized as follows:

- We propose SepACap, an a cappella source separation model achieving state-of-the-art performance on multi-singer separation while operating in the waveform domain.
- To increase available training data, and to enable model generalizability to subset singer separation, we use a power set-based data augmentation method that transforms a standard multi-singer dataset into exponentially more training samples, including cases with absent singers, enabling more robust separation performance.
- We extend the loss function to allow for stable training with empty stems, ensuring that models operating on the waveform learn to handle silent signals in the data.

2 Methodology

Model. To separate arbitrary singers in an a cappella setting, we introduce SepACap, an adaptation of SepReformer [20], a recently proposed state-of-the-art speaker separation model. We found that updating the existing ReLU activation functions to the SNAKE activation [25] markedly increased model performance. Additionally, we changed the training objective from SNR-based training to a composite time and time-frequency loss. Many separation models from the speech domain use a variant of signal-to-noise ratio (SNR) or signal-to-distortion ratio (SDR) as their loss during training [20, 21, 24], such as the scale invariant SDR (SI-SDR) [10] defined as

$$\text{SI-SDR} := 10\log_{10}\frac{\|e_{target}\|^2}{\|e_{\text{res}}\|^2} = 10\log_{10}\frac{\left\|\frac{\hat{s}^T s}{\|s\|}s\right\|^2}{\left\|\frac{\hat{s}^T s}{\|s\|}s-\hat{s}\right\|^2}, \tag{1}$$

where s is the target signal and $\hat{s}$ is the predicted signal. The problem when using SI-SDR as a loss function is that, for stems without a signal, SI-SDR provides no informative gradient, making it ineffective as a training objective in such cases. Therefore, we utilize a different class of losses that provide a similar information level as the SI-SDR loss, but do not rely on a signal always being present. Therefore, we propose to use a combination of three different losses. We use an L1 loss on the waveform as well as a multi-scale Mel loss, which measures the L1 distance between log-mel spectrograms. We additionally use a spectral loss, which combines L1 losses on magnitude and log-magnitude STFT features to capture spectral consistency across resolutions. This loss combination has been effectively applied in the audio compression-reconstruction domain [9]. We find that this loss combination works well and demonstrate its successful transfer to the task of a cappella separation.

Data Augmentation. Since our model is designed to handle arbitrary subsets of sources, we construct training data by generating mixtures corresponding to the power set of available stems. This yields an exponential increase in the number of possible separation targets, covering both full mixtures and cases where some stems are absent. To make this method computationally feasible, we further

segment the audio into short fixed-length snippets, which both increases the number of training samples and reduces memory requirements during training.

3 Experiments

Setup. We use the JaCappella dataset [14] to train and evaluate SepACap and baseline approaches. The dataset provides Japanese a cappella music and their 6 corresponding stems (Alto, Bass, Lead Vocal, Soprano, Tenor, and Vocal Percussion). Using the data augmentation strategy discussed in Section 2, we increase the dataset size to 105k samples up from the original 35 music clips. In terms of duration, the augmentation strategy increases the dataset duration from 0.57 hours to 145h hours.

We train SepACap on 4-second snippets of the augmented JaCappella dataset. The model is designed to separate up to six stems directly in the waveform domain. For training, we adopt the composite loss setup described in Section 2, which combines time-domain and spectral reconstruction objectives. The spectral loss operates at three STFT window lengths (512, 1024, and 2048), and the mel loss spans seven bins (5, 10, 20, 40, 80, 160, and 320) with window lengths of 32, 64, 128, 256, 512, 1024, and 2048. We empirically found that 0.3, 0.7, and 1.0 as the weights for the spectral loss, the mel loss, and the waveform loss, respectively, work well. Additionally, as a baseline comparison we train a Mel-Band RoFormer model [22] on the same dataset.[2]

We evaluate our model in two settings. First, we assess model performance when all stems are present. In this setting we can directly compare against prior work DPTNet [1], X-UMX [19], and MRDLA [13], by reporting the SI-SDR-improvement (SDRi), defined as

$$\text{SI-SDR-improvement} = \text{SI-SDR}_{\text{Pred}} - \text{SI-SDR}_{\text{Mixture}} \tag{2}$$

and average the results across stems. Second, we evaluate the subset condition, where only a subset of stems is present in the mixture. In this setting, the model must both separate active sources and correctly output silence for absent sources. To capture this dual objective, we use two complementary metrics. When a reference signal is present, we report SDRi, which quantifies separation quality relative to the input mixture. When the reference stem is silent, SI-SDR is not meaningful; instead, we evaluate the model's ability to suppress spurious output by measuring the root-mean-square energy relative to full scale (RMS-DBFS), defined as

$$\text{RMS-DBFS} = 20 \cdot \log_{10} \sqrt{\frac{1}{T} \sum_{t} x_t^2 + \varepsilon}, \tag{3}$$

where x is the signal, T is the length of the signal, and ε is a small constant. This silence metric directly reflects the residual energy of the predicted signal and therefore serves as an indicator of how well the model avoids false positives in stems that should be silent.

Evaluation. In Table 1 we observe the performance of the different methods. We find that SepACap outperforms previous approaches in 5 out of 6 stems (Bass, Lead Vocal, Soprano, Tenor, and Vocal Percussion) in SDRi even though only a fraction of the samples seen during training contain all stems simultaneously. Furthermore, the Mel-Band RoFormer seems to significantly underperform at this task, which suggests that the time-frequency domain masking struggles to separate multiple sources contained in similar frequency bands. The reported values for X-UMX, DPTNet, and MRDLA are taken from JaCappella [14].

For the subset objective, the results in Table 2 highlight a clear trade-off between the two models. SepACap generalizes especially well to this setting, as it produces fewer instances of bleed-through when stems are absent and can more effectively suppress inactive sources. However, this comes at the cost of introducing more audible artifacts in the reconstructed signals. In contrast, the Mel-Band RoFormer yields cleaner outputs with fewer artifacts, but it frequently fails to fully suppress silent stems, leading to noticeable bleed-through between sources. This difference is consistent with the underlying model designs: the Mel-Band RoFormer operates by masking unwanted frequency components, which prevents artifact creation but makes complete suppression of inactive signals difficult, whereas SepACap generates waveforms directly and is therefore more prone to artifact introduction.

[2]We train with `https://github.com/KimberleyJensen/Mel-Band-Roformer-Vocal-Model`

Table 1: Performance of models when all stems are present, measured on the test split of the JaCappella dataset. The metric is per-stem SDRi (higher is better) for each model. Best performance in bold, and second best performance is underlined.

Method	Alto	Bass	Lead Vocal	Soprano	Tenor	Vocal Perc.
X-UMX [19]	13.5	9.1	7.5	10.7	10.2	21.0
DPTNet [1]	11.9	<u>19.7</u>	<u>8.9</u>	8.5	<u>14.9</u>	21.9
MRDLA [13]	**14.7**	10.2	8.7	<u>11.8</u>	11.3	<u>22.1</u>
Mel-Band RoFormer [22]	6.3	17.8	0.7	4.5	10.3	19.3
SepACap (Ours)	<u>14.6</u>	**23.2**	**13.0**	**13.1**	**17.0**	**22.5**

Table 2: Subset condition performance of DPTNet (publicly-available checkpoint), our trained Mel-Band RoFormer (MBR), and our proposed model SepACap, on the test split of the augmented JaCappella dataset. The per-stem SDRi is only reported when a reference signal is present, and RMS silence scores evaluate suppression quality for silent stems. Unsurprisingly, we find that DPTNet underperforms on the subset-stem task as it was only trained on full mixes. SepACap also significantly outperforms the Mel-Band RoFormer as it does not rely on frequency-based masking.

	SI-SDRi (dB)↑			RMS (dBFS)↓		
Stem	DPTNet	MBR	SepACap (ours)	DPTNet	MBR	SepACap (ours)
Alto	-17.2	3.9	**11.6**	-19.6	-59.1	**-92.7**
Bass	-30.8	15.5	**20.4**	-33.7	-70.8	**-95.1**
Lead Vocal	-44.0	1.6	**9.1**	-41.5	-63.6	**-91.9**
Soprano	-46.9	1.6	**11.1**	-44.7	-55.5	**-85.6**
Tenor	-25.9	7.6	**13.0**	-27.2	-75.3	**-95.7**
Vocal Perc.	-32.4	18.3	**18.4**	-33.6	-73.1	**-95.3**

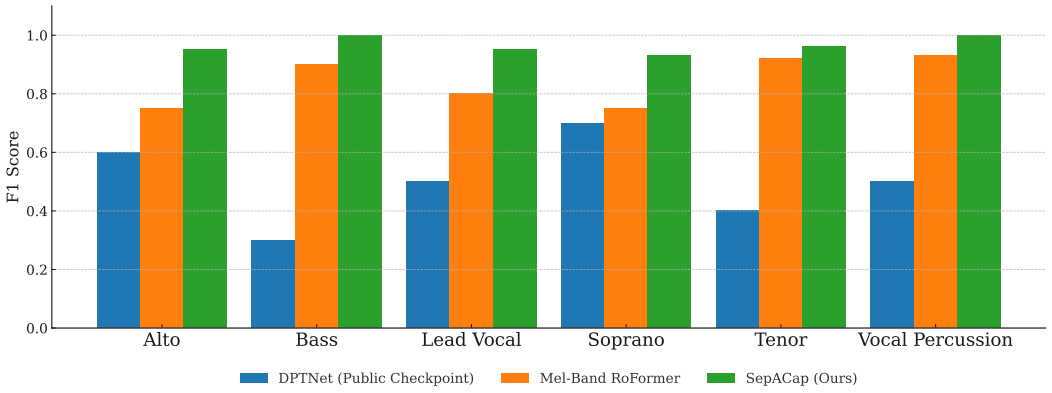

Figure 1: We test the ability of models to predict whether a stem is present in a mix. We report F1 scores for stem detection on the test split of the augmented JaCappella dataset. The results show per-stem detection performance for DPTNet, Mel-Band RoFormer, and SepACap, where higher is better in the interval. Our proposed model SepACap achieves the best overall performance in stem detection.

Despite SepACap's stronger overall performance in the subset setting, it does not always maintain consistent performance across all stems. In particular, both SepACap and Mel-Band RoFormer struggle on the Alto stem compared to the all-stems setting. The lower quantitative results observed in this evaluation can often be attributed to failures in detecting a present stem and instead predicting silence, as illustrated in Table 2. Because the SI-SDR metric assigns large negative values in such cases, these errors disproportionately reduce the average scores.

Conclusion. We introduced SepACap, a source separation model trained for a cappella mixtures. Evaluated on JaCappella, SepACap achieves state-of-the-art performance when all stems are present and substantially improves subset separation by suppressing inactive stems and reducing bleed-through compared to baseline approaches.

References

[1] Jingjing Chen, Qirong Mao, and Dong Liu. Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation. *arXiv preprint arXiv:2007.13975*, 2020.

[2] Ke Chen et al. Improving choral music separation through expressive synthesized data from sampled instruments. *arXiv preprint arXiv:2209.02871*, 2022.

[3] Helena Cuesta. Data-driven pitch content description of choral singing recordings. 2022.

[4] Helena Cuesta et al. A framework for multi-f0 modeling in satb choir recordings. *arXiv preprint arXiv:1904.05086*, 2019.

[5] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Demucs: Deep extractor for music sources with extra unlabeled data remixed. *arXiv preprint arXiv:1909.01174*, 2019.

[6] Alejandro Delgado et al. A new dataset for amateur vocal percussion analysis. In *Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound*, pages 17–23, 2019.

[7] Giorgio Fabbro et al. The sound demixing challenge 2023 music demixing track. *arXiv preprint arXiv:2308.06979*, 2023.

[8] Chang-Bin Jeon et al. Medleyvox: An evaluation dataset for multiple singing voices separation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[9] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36:27980–27993, 2023.

[10] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE, 2019.

[11] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27 (8):1256–1266, 2019.

[12] Yi Luo and Jianwei Yu. Music source separation with band-split rnn. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1893–1901, 2023.

[13] Tomohiko Nakamura, Shihori Kozuka, and Hiroshi Saruwatari. Time-domain audio source separation with neural networks based on multiresolution analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1687–1701, 2021.

[14] Tomohiko Nakamura, Shinnosuke Takamichi, Naoko Tanji, Satoru Fukayama, and Hiroshi Saruwatari. jacappella corpus: A japanese a cappella vocal ensemble corpus. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[15] Darius Petermann et al. Deep learning based source separation applied to choir ensembles. *arXiv preprint arXiv:2008.07645*, 2020.

[16] Sebastian Rosenzweig, Helena Cuesta, Christof Weiß, Frank Scherbaum, Emilia Gómez, and Meinard Müller. Dagstuhl choirset: A multitrack dataset for mir research on choral singing. *Transactions of the International Society for Music Information Retrieval*, 3(1), 2020.

[17] Simon Rouard et al. Hybrid transformers for music source separation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[18] Kohei Saijo, Gordon Wichern, François G. Germain, Zexu Pan, and Jonathan Le Roux. Tf-locoformer: Transformer with local modeling by convolution for speech separation and enhancement. In *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 205–209, 2024. doi: 10.1109/IWAENC61483.2024.10694313.

[19] Ryosuke Sawata, Stefan Uhlich, Shusuke Takahashi, and Yuki Mitsufuji. All for one and one for all: Improving music separation by bridging networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 51–55. IEEE, 2021.

[20] Ui-Hyeop Shin, Sangyoun Lee, Taehan Kim, and Hyung-Min Park. Separate and Reconstruct: Asymmetric Encoder-Decoder for Speech Separation, January 2025. URL `http://arxiv.org/abs/2406.05983`. arXiv:2406.05983 [eess].

[21] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation, 2021. URL `https://arxiv.org/abs/2010.13154`.

[22] Ju-Chiang Wang, Wei-Tsung Lu, and Minz Won. Mel-band roformer for music source separation. *arXiv preprint arXiv:2310.01809*, 2023.

[23] Sangeon Yong et al. A phoneme-informed neural network model for note-level singing transcription. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[24] Shengkui Zhao and Bin Ma. Mossformer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[25] Liu Ziyin, Tilman Hartwig, and Masahito Ueda. Neural networks fail to learn periodic functions and how to fix it. *Advances in Neural Information Processing Systems*, 33:1583–1594, 2020.