CURATING AN A CAPPELLA DATASET FOR SOURCE SEPARATION

Ting-Yu Pan*
University of Michigan
pting@umich.edu

Kexin Phyllis Ju*
University of Michigan
kexinju@umich.edu

Hao-Wen Dong
University of Michigan
hwdong@umich.edu

Late-Breaking / Demo Session Extended Abstract, ISMIR 2025 Conference

ABSTRACT

A cappella music presents unique challenges for source separation due to its diverse vocal styles and the presence of vocal percussion. Current a cappella datasets are limited in size and diversity, hindering the development of robust source separation models. To address this, we curated a dataset of 55 studio-quality a cappella songs performed by 3 professional groups. Then, we introduce a two-step a cappella source separation pipeline and present preliminary results on vocal percussion separation. Finally, we discuss future work on AI-driven dataset augmentation and supporting tools for asynchronous a cappella rehearsals.

1. INTRODUCTION

A cappella is a music genre performed solely by the human voice and body [1]. Unlike large choral ensembles, a cappella groups typically feature one singer per part and perform diverse vocal styles [2]. A distinctive feature is vocal percussion (VP), or beatboxing, which provides a rhythmic backbone by "imitating existing drum sounds" but is rarely notated [2], limiting the effectiveness of score-informed separation methods. These characteristics position a cappella between traditional choirs and pop bands, demanding source separation techniques capable of handling both intricate harmonies and percussive vocal effects.

A significant challenge in developing source separation models for a cappella music is the scarcity of large, high-quality datasets with isolated vocal stems. Existing a cappella datasets [3–5] contain only 20–40 songs totaling 100–200 minutes of audio, which is insufficient for training robust models. To address this limitation, we compiled a "golden dataset": a repository of 55 a cappella songs performed by 3 professional groups.

Based on the dataset, we introduce a two-step source separation pipeline for a cappella music and present preliminary results on VP separation, which demonstrate fine-tuning on our curated dataset yields a $\approx$ 10% relative gain in VP separation SDR over pretrained models(see section 4). Next, inspired by Sarkar et al.'s work with synthetic data for chamber ensemble separation [6], we propose an AI-driven approach to expand the "golden dataset" through voice cloning and synthesis, supplemented with symbolic a cappella data in MIDI and MusicXML formats. We believe that generative AI can play a role in curating audio datasets, particularly in resource-limited domains such as the underexplored area of a cappella. Finally, we discuss potential applications of the dataset and source separation model for a cappella.

2. INITIAL "GOLDEN DATASET"

At the core of our dataset is the "Golden Dataset," a collection of 55 studio-quality a cappella songs performed by 3 distinct a cappella groups. These recordings offer a diverse representation of styles, such as pop music, jazz, R&B, and medley. Each track was professionally recorded in studios, with isolated stems for individual parts (e.g., soprano, alto, tenor, bass (SATB) and VP). The recordings cover languages including English, Mandarin, Korean, and Hakka Chinese. This dataset serves as the foundation for building an a cappella source separation model.

3. A TWO-STEP SOURCE SEPARATION PIPELINE FOR A CAPPELLA

Modern a cappella blends percussive and pitched voices, so we first evaluated a state-of-the-art choral model [7] on high-quality covers (e.g., Pentatonix [8]). Informal listening tests showed that VP was not isolated but remained across the SATB outputs. To address this, we propose:

Step 1: Vocal Percussion Extraction (Ongoing) We configure Demucs [9] to output four stems and treat its "drums" channel as VP. Human listening and preliminary SDR gains (see §4) confirm this reliably captures VP.

Step 2: Vocal Harmony Separation (Future Work) Subtracting the extracted VP yields a "VP-less" mix like a traditional choral track. We will apply UMSS [7] (or a similar model) to decompose this residual into SATB parts.

This sequence leverages Demucs's strength on rhythmic sources and a choral model's strength on pitched voices, ensuring each stage uses the most suitable tool.

4. PRELIMINARY RESULTS

Our experiments evaluate a two-step source separation pipeline, focusing on the first step—VP extraction. All models are Demucs variants. We benchmark pretrained

[*] These authors contributed equally to this work.

Attribution: T. Pan, K. Ju, and H. Dong, "Curating an A Cappella Dataset for Source Separation", in *Extended Abstracts for the Late-Breaking Demo Session of the 26th Int. Society for Music Information Retrieval Conf.,* Daejeon, South Korea, 2025.

models to establish a baseline and then fine-tune the best performer on our data to gauge benefit of specialization.

4.1 Experimental Setup

While our full "golden" dataset comprises 55 a cappella tracks, only 17 have completed preprocessing. For these preliminary experiments, we split them as follows:

Training (10 tracks): Studio recordings from Group A for fine-tuning.

Validation (2 tracks): Additional Group A recordings.

Test Dataset 1 (In-Distribution, 2 tracks): Held-out Group A studio recordings.

Test Dataset 2 (Out-of-Distribution): This dataset contains three studio tracks from Groups B and C.

4.2 Analysis of Results

We first tested two pretrained Demucs models: Model 1, the official `htdemucs` model, and Model 2, a version fine-tuned on a drum dataset [1]. Table 1 shows that Model 1 outperforms Model 2 on both test sets, confirming that VP behaves differently from conventional drum sources. We then fine-tuned `htdemucs` to output two stems (VP vs. Other) with the configurations in Table 3. As shown in Table 2, VP SDR on Test 1 rises from 3.09 dB to 3.3–3.4 dB after fine-tuning ($\approx 10\%$ relative gain), demonstrating that even limited, targeted training yields a notable improvement in VP separation, while the minimal differences between configurations likely reflect our small dataset size.

Model	**Test Dataset 1**			**Test Dataset 2**		
	VP	Other	All	VP	Other	All
Pretrained Model 1	3.09	9.28	6.18	3.03	20.43	11.73
Pretrained Model 2	1.07	7.88	4.47	2.71	17.00	9.86

Table 1: SDR (dB) of pretrained models.

Model	**Test Dataset 1**			**Test Dataset 2**		
	VP	Other	All	VP	Other	All
Pretrained Model 1	3.09	9.28	6.18	3.03	20.43	11.73
Finetuned Model A	3.43	9.25	6.34	3.07	20.48	11.78
Finetuned Model B	3.37	9.29	6.33	3.04	20.73	11.89
Finetuned Model C	3.40	9.24	6.32	3.07	20.38	11.72
Finetuned Model D	3.38	9.25	6.31	3.04	20.32	11.68

Table 2: SDR (dB) before and after fine-tuning.

Parameter	**Finetuned Models**			
	A	**B**	**C**	**D**
Epoch	10	20	10	20
Weights*	1.0, 1.0, 1.0, 1.0	1.0, 1.0, 1.0, 1.0	1.0, 0.0, 0.0, 1.0	1.0, 0.0, 0.0, 1.0

* Weights correspond to the stems [VP, dummy1, dummy2, other]

Table 3: Configuration of Finetuned Models

5. DATASET AUGMENTATION

To expand the dataset of 55 songs, we plan to augment the recordings in the following ways.

Pitch shifting: To enhance model robustness against minor tuning variations, we will generate versions of each track in the "golden dataset" shifted by –1 and +1 semitone. This augmentation triples the number of available audio tracks for each song.

Voice cloning: Using voice cloning techniques[2], we will convert each audio track (including the pitch-shifted variants) into a different timbre. This process doubles the number of tracks and enables the creation of both *all-AI mixes* (where all parts are cloned) and *hybrid mixes* (where AI-cloned and original human voices are combined).

Voice synthesis: We experimented with AI singing voice synthesizers (e.g., VOCALOID6[3] and Synthesizer V Studio 2 Pro[4]). These tools can transform a MIDI file with annotated syllables into an AI singing voice that reproduces both melody and lyrics. We input MIDI files of each voice part to generate a cappella recordings. The resulting quality was satisfactory, with Synthesizer V Studio 2 Pro producing more lifelike results for English a cappella songs. AI voice synthesis potentially allows us to augment our dataset by adding additional songs.

Symbolic Data: To further enrich our dataset, we will include a symbolic dataset comprising a cappella arrangements in MIDI and MusicXML formats. These arrangements are sourced from MuseScore and include the songs present in both the "golden dataset" and the AI voice synthesis dataset.

6. DISCUSSION & FUTURE WORK

Our project is based on the belief that generative AI can play a role in augmenting audio datasets [6] for resource-limited domains such as a cappella. By leveraging voice cloning and synthesis technologies, we can generate new audio samples from a small set of data in a cost-effective manner.

A key contribution of our dataset is its potential to support model training and evaluation for various tasks, such as a cappella source separation. Because our dataset includes recordings that differ in timbre, pitch, and origin (AI-generated vs. human-performed), it can also be used for voice style conversion and singer identification [10]. Furthermore, the symbolic dataset paired with singing audio can be utilized as a baseline for evaluating the performance of generative audio models.

Our preliminary experiments indicate a promising direction for source separation of a cappella music. Fine-tuning on our initial dataset yielded a $\approx 10\%$ relative gain in VP separation SDR over pretrained models.

In the future work, we will augment the dataset by leveraging AI and explore practical applications of the source separation model, such as the development of an asynchronous a cappella rehearsal system. A future design could allow singers to rehearse asynchronously by listening to or mixing subsets of separated voice parts. This could support a more flexible and collaborative rehearsal when group members are geographically dispersed.

[1] `https://github.com/facebookresearch/demucs`

[2] `https://studio.moises.ai/voice-studio/`

[3] `https://www.vocaloid.com/`

[4] `https://dreamtonics.com/synthesizerv/`

7. ACKNOWLEDGEMENT

We would like to thank the three a cappella groups—Annappella from the University of Michigan, Real A Cappella Group from Cornell University, and Seagull-K from Taiwan—for generously providing high-quality recordings. We also extend our gratitude to Zexu Du and Rae Chen for their assistance in obtaining recordings.

8. REFERENCES

[1] N. L. Norden, "A new theory of untempered music: A few important features with special reference to "a cappella" music," *The Musical Quarterly*, vol. 22, no. 2, p. 217–233, 1936.

[2] J. S. Duchan, "Collegiate a cappella: Emulation and originality," *American Music*, vol. 25, no. 4, p. 477–506, 2007.

[3] T. Nakamura, S. Takamichi, N. Tanji, S. Fukayama, and H. Saruwatari, "JaCappella Corpus: A Japanese a Cappella Vocal Ensemble Corpus," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[4] R. Schramm, A. McLeod, M. Steedman, and E. Benetos, "Multi-pitch detection and voice assignment for A cappella recordings of multiple singers," in *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, S. J. Cunningham, Z. Duan, X. Hu, and D. Turnbull, Eds., 2017, pp. 552–559. [Online]. Available: https://ismir2017.smcnus.org/wp-content/uploads/2017/10/26_Paper.pdf

[5] R. Schramm and E. Benetos, "Automatic transcription of a cappella recordings from multiple singers," 2017.

[6] S. Sarkar, L. Thorpe, E. Benetos, and M. Sandler, "Leveraging synthetic data for improving chamber ensemble separation," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA: IEEE, Oct. 2023, p. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10248118/

[7] G. Richard, P. Chouteau, and B. Torres, "A fully differentiable model for unsupervised singing voice separation," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 946–950.

[8] Pentatonix, "Pentatonix official website," https://www.home.ptxofficial.com/, accessed: 2025-08-08.

[9] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," 2022. [Online]. Available: https://arxiv.org/abs/2211.08553

[10] Z. Deng and R. Zhou, "Vocal92: Audio dataset with a cappella solo singing and speech," *IEEE Access*, vol. 11, pp. 140 958–140 966, 2023.