ACappellaSet: A Multilingual A Cappella Dataset for Source Separation and AI-assisted Rehearsal Tools

Anonymous Author(s)
Affiliation
Address
email

Abstract

A cappella music presents unique challenges for source separation due to its diverse vocal styles and the coexistence of harmonic and percussive voices. Current a cappella datasets are limited in size and diversity, hindering the development of robust source separation models. In this paper, we present *ACappellaSet*, a collection of 55 professionally recorded a cappella songs performed by three professional groups. In addition, we present experimental results showing that fine-tuning HTDemucs on ACappellaSet substantially improves vocal percussion (VP) separation, raising VP SDR from 5.22 dB to 7.62 dB, and enabling scalable multi-stem modeling. 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 [12]. Unlike large choral ensembles, a cappella groups typically feature one singer per part and perform diverse vocal styles [5]. A distinctive feature is vocal percussion (VP), or beatboxing, which provides a rhythmic backbone by "imitating existing drum sounds" but is rarely notated [5], 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 [11, 17, 16] contain only 20–40 songs totaling 100–200 minutes of audio, which is insufficient for training robust models. To address this limitation, we present *ACappellaSet*, a repository of 55 a cappella songs performed by three professional groups. We present statistics and potential applications of the golden dataset. We also present a comparison between our dataset and other existing music separation datasets.

Based on our dataset, we conducted experiments with HTDemucs [14], demonstrating that fine-tuning on *ACappellaSet* significantly improves vocal percussion separation performance and extends effectively to multi-stem configurations. Inspired by Sarkar et al.'s work with synthetic data for chamber ensemble separation [15], we also explore AI-driven dataset expansion through voice cloning and synthesis, particularly for underrepresented vocal parts. Finally, we discuss potential applications and the importance of a cappella source separation.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

2 ACappellaSet: A Multilingual A Cappella Dataset

2.1 Dataset Statistics and Applications

ACappellaSet contains 55 professionally recorded a cappella songs performed by three vocal groups, each captured in isolated stems corresponding to soprano (S), alto (A), tenor (T), bass (B), and vocal percussion (VP), with occasional mezzo-soprano (M) or baritone (Bar) parts.

The collection totals **2 hours 37 minutes 30 seconds** on 55 unique songs (Figure 1a). Recordings were made with either a **single microphone in studio** or a **multi-microphone setup** (up to five microphones), enabling clean isolation of individual voice parts when multiple microphones were used. All files were **exported as stereo WAV files** at **44.1 kHz, 24-bit** resolution. For ensemble synchrony, click tracks were used in most sessions, though some rough recordings preserve natural timing variation. This design allows for both clean source separation benchmarks and more realistic "in-the-wild" conditions.

Each song is available in up to 3 production stages. The **rough recordings** are minimally edited and captured in a casual setting, preserving natural imperfections. The **dry mixes** undergo post-processing steps such as tuning, equalization, and balancing to create a more refined version while maintaining clarity. Finally, the **wet mixes** represent polished studio-style productions, enhanced with effects, reverb, and spatial panning.

Together, these versions produce **110 files** totaling approximately **5 hours and 11 minutes** of audio.

The dataset includes 11 original arrangements, 37 covers, and 7 medleys, performed in 6 languages (Mandarin, English, Hakka Chinese, Taiwanese, Korean, and non-lyrical vocalizations; Figure 1b). Ensemble configurations are dominated by SATB+VP but also include others such as ATBarB and SMAB (Figure 1c). Figure 2 illustrates the distribution of different recording stages across the corpus.

Potential Applications. Beyond source separation, *ACappellaSet* enables studies in **voice-part classification**, **automatic arrangement analysis**, and **cross-lingual style modeling**. The inclusion of rough, dry, and wet versions allows investigation of how production quality influences learning-based models. By combining controlled isolation with real ensemble variability, the dataset provides a practical foundation for AI-driven a cappella research and rehearsal applications.

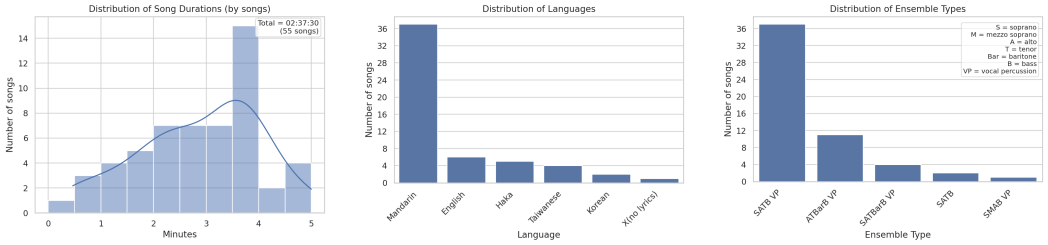
2.2 Comparison with Existing Datasets

Compared with widely used music separation datasets such as **MUSDB18** [13] and **MedleyDB** [2], which primarily feature instrumental mixtures, *ACappellaSet* is designed specifically for purely vocal ensembles. It also differs from existing corpora in several aspects:

- **Style and structure:** *ACappellaSet* captures contemporary small-group a cappella pop performances, where singers frequently **switch lead melodies** across parts (S, A, T, B). This arrangement style is common in modern a cappella but presents greater separation difficulty due to overlapping timbres and dynamic role changes. In contrast, the **JaCappella Corpus** [11] features Japanese children's songs arranged for six parts (lead, S, A, T, B, and VP), where a fixed *lead vocal* stem consistently carries the melody.
- **Expanded stem configuration:** *ACappellaSet* includes isolated **vocal percussion (VP)** tracks, enabling full SATB+VP modeling. Choral datasets such as CSD [4] and Cantoria [3] instead feature large ensembles with multiple singers per part and no percussive voices.
- **Multilingual and production-aware:** *ACappellaSet* spans six languages and three production stages, facilitating cross-lingual and production-conditioned learning.

3 Experiments and Results

We evaluate **Hybrid Transformer Demucs (HTDemucs)** [14, 8], a state-of-the-art time-domain source separation model, on our curated **golden dataset** (§2), focusing on the challenge of isolating **vocal percussion (VP)** from other voice parts.



(a) Song durations (b) Languages (c) Ensemble types

Figure 1: Dataset statistics: (a) song durations, (b) language distribution, and (c) ensemble types.

Table 1: SDR (dB) for pretrained and fine-tuned HTDemucs models. Fine-tuning substantially improves VP separation.

Model	VP	Other	All
Pretrained (official)	5.22	10.66	7.94
Pretrained (drum)	3.66	9.24	6.45
Fine-tuned (ours)	**7.62**	**11.63**	**9.62**

3.1 Experimental Setup and Results

We used 49 songs (103 files) from the golden dataset, split into 40 for training, 4 for validation, and 5 for testing (Appendix B, Table 4). Each song may have up to three production versions (rough, dry, wet), totaling 4:46:55 of audio. Evaluation used the SDR metric from the MDX 2021 definition (NSDR) [19, 18].

We compared two pretrained **HTDemucs** variants—(i) the official general-purpose `htdemucs` and (ii) a drum-fine-tuned version—to test whether VP behaves more like percussion or requires targeted adaptation. HTDemucs, originally trained for four stems (drums, bass, vocals, other), was reconfigured for two stems: VP (mapped from drums) and Other (all remaining voices). We tuned learning rate and stem-weight configurations; full settings are listed in Appendix B, Table 5.

Table 1 shows that the official model outperformed the drum-fine-tuned variant, confirming that VP is acoustically distinct from conventional drums. Our fine-tuned model further improved VP SDR from 5.22 dB to 7.62 dB (+46% relative), while also raising scores for Other and overall—without degrading harmonic separation.

The test split (five songs, 14 files) includes the only two **Korean** tracks in the dataset, enabling **zero-shot cross-lingual** evaluation. Because all test songs include rough, dry, and wet versions, we also analyzed robustness to production conditions. Table 2 shows that **wet mixes degraded performance**, with VP SDR dropping to 4.84 dB versus 8.73 dB for dry mixes, consistent with prior findings that reverberation smears temporal and spectral cues [10, 6, 7]. Dry and rough versions yielded stronger results, indicating that minimally processed audio benefits separation.

3.2 Multi-Stem and Two-Stage Evaluation

To further assess the scalability of our approach, we conducted additional experiments on multi-stem separation using the same pretrained family of **HTDemucs** models. Specifically, we compared:

- **(i)** one-stage separation using a 6-stem model fine-tuned on *ACappellaSet* to predict {VP, Bass (B), Soprano (S), Alto (A), Tenor (T), Baritone (Bar)} simultaneously, and
- **(ii)** a **two-stage pipeline** where a 2-stem model first isolates VP, followed by a 5-stem model that separates the remaining harmonic voices.

The two-stage configuration yielded clear gains across most stems, with VP improving from -0.6 dB to 6.8 dB and Bass from 1.7 dB to 7.1 dB. Isolating VP first reduces spectral overlap and rhythmic interference in harmonic voices, leading to cleaner downstream separation. Interestingly, the 5-stem model—trained purely on vocal mixtures—generalized better than the 6-stem variant that retained

Table 2: SDR (dB) by recording condition for the fine-tuned HTDemucs. Wet versions consistently degrade quality.

Condition	VP	Other	All
Dry	8.73	12.13	10.43
Rough	6.01	14.56	10.28
Wet	4.84	10.08	7.46

Table 3: Comparison of one-stage vs. two-stage a cappella separation (NSDR [dB]).

Method	VP	B	S	A	T	Bar
(i) one-stage (6s)	−0.6	1.7	2.3	2.1	4.5	0.6
(ii) two-stage (2s→5s)	**6.8**	**7.1**	2.9	2.7	5.0	0.7

unused instrumental heads, suggesting that domain-specific fine-tuning benefits from simplified architectures.

4 Future Work

Exploring Generative AI-powered Data Augmentation To expand the dataset, we plan to augment the recordings in the following ways.

- **Pitch shifting**: 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 [1], 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 and Synthesizer V Studio 2 Pro) that transform a MIDI file with annotated syllables into an AI singing voice. 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 allows us to augment our dataset by adding additional songs.
- **Symbolic Data**: 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.

ACAMate: Towards AI-assisted A Capella Rehearsal Tools Curating a dataset for a cappella source separation is valuable because of its potential to support rehearsals, particularly for novice singers in collegiate a cappella groups. Collegiate a cappella groups are typically small and lack a conductor [5], and may even be non-auditioned, with members who have little singing experience [9]. As a result, these groups often struggle to access professional guidance. During rehearsals, singers must monitor their own voices and identify mistakes in real time. A source separation–based rehearsal tool could allow them to review separated voice parts after group singing, making it easier to detect and learn from errors. Such a tool could also analyze and provide feedback on separated voice parts, helping singers understand concrete next steps for improvement.

The lack of professional guidance is even more challenging in individual practice. A cappella arrangements are rhythmically and harmonically interdependent, and singers rely on vocal cues from others to stay synchronized. Without those cues, novices may struggle to grasp how their part fits within the ensemble. Source separation could address this by enabling singers to practice with either group or reference recordings in which voice parts are isolated (see figure 3A). They could then create customized practice materials by adjusting or remixing individual parts.

Therefore, a cappella source separation forms the foundation for rehearsal-support systems. Given a mixed recording, such a system could first separate the parts and then allow users to manipulate them (e.g., by muting or unmuting a line) and receive feedback through analysis of the separated voices.

References

[1] Moises voice studio. https://studio.moises.ai/voice-studio/.

[2] Rachel Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Bello. Medleydb: A multitrack dataset for annotation-intensive mir research.

[3] Helena Cuesta and Emilia Gómez. Cantoría dataset, January 2022.

[4] Helena Cuesta, Emilia Gómez, Agustín Martorell, and Felipe Loáiciga. Choral singing dataset, June 2018.

[5] Joshua S. Duchan. Collegiate a cappella: Emulation and originality. *American Music*, 25(4):477–506, 2007.

[6] Alexandre Défossez, Simon Rouard, Robin Wightman, Sam Green, and E. Gordon. Hybrid transformers for music source separation. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, pages 1362–1369, 2021.

[7] Sebastian Ewert, Felix Weninger, Florian Eyben, and Björn Schuller. Speech dereverberation using ideal ratio masks in the stft domain. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5372–5376. IEEE, 2014.

[8] Facebook AI Research. Demucs: Music source separation. https://github.com/facebookresearch/demucs. Accessed: 2025-08-29.

[9] Marshall Haning. "everyone has a voice": Informal learning in student-led collegiate a cappella ensembles. page 61–76, 2019.

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

[11] 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, 2023.

[12] N. Lindsay Norden. A new theory of untempered music: A few important features with special reference to "a cappella" music. *The Musical Quarterly*, 22(2):217–233, 1936.

[13] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. Musdb18 - a corpus for music separation, December 2017.

[14] Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation, 2022.

[15] Saurjya Sarkar, Louise Thorpe, Emmanouil Benetos, and Mark Sandler. Leveraging synthetic data for improving chamber ensemble separation. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, page 1–5, New Paltz, NY, USA, October 2023. IEEE.

[16] Rodrigo Schramm and Emmanouil Benetos. Automatic transcription of a cappella recordings from multiple singers. 2017.

[17] Rodrigo Schramm, Andrew McLeod, Mark Steedman, and Emmanouil Benetos. Multi-pitch detection and voice assignment for A cappella recordings of multiple singers. In Sally Jo Cunningham, Zhiyao Duan, Xiao Hu, and Douglas Turnbull, editors, *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, pages 552–559, 2017.

[18] Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito. The 2018 signal separation evaluation campaign. In *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 293–305, 2018.

[19] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. In *Proceedings of the International Conference on Independent Component Analysis and Signal Separation (ICA)*, pages 94–99, 2006.

A More Dataset Statistics

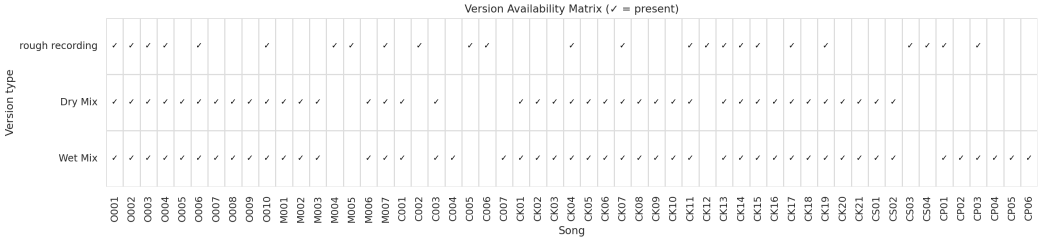


Figure 2: Version availability matrix showing rough, dry, and wet versions for each song.

B Implementation Details

B.1 Dataset Split

Table 4: Dataset split by songs. Each song may have up to three production versions (rough/dry/wet), yielding 103 files in total.

Split	Song Count	Duration
Train	40	1:53:49
Valid	4	0:11:18
Test	5	0:11:54
Total	49	2:17:01

B.2 Fine-tuning Configurations

Table 5: Fine-tuning configurations of htdemucs. Best-performing run (e76885f2) achieved VP SDR of 7.62 dB.

ID	Epochs	LR	Weights	VP SDR	Other SDR	All SDR
959da2b8	20	3e-4	[1.2, 0.5, 0.5, 1.0]	7.21	11.27	9.24
e76885f2*	20	3e-4	[**1.5**, 0.3, 0.3, 1.0]	**7.62**	**11.63**	**9.62**
704097c5	20	3e-4	[1.0, 0.7, 0.7, 1.0]	7.26	11.39	9.32
86f19ed3	20	3e-4	[1.3, 0.4, 0.4, 1.0]	7.20	11.51	9.36
4dea81d0	20	3e-4	[1.0, 1.0, 1.0, 1.0]	7.19	11.28	9.23
cb1aef50	10	2e-5	[1.0, 1.0, 1.0, 1.0]	6.55	10.68	8.61
f3fab4cf	10	1e-4	[1.0, 1.0, 1.0, 1.0]	4.43	9.89	7.16
a935fdf0	10	3e-5	[1.0, 1.0, 1.0, 1.0]	6.25	10.61	8.43
8038f405	10	4e-2	[1.0, 1.0, 1.0, 1.0]	0.28	4.15	2.21

* Best performing configuration.
Weights correspond to the stems [VP, dummy1, dummy2, other].

B.3 Additional 4-Stem (SATB) Evaluation

For completeness, we also fine-tuned an **HTDemucs-4s** model to separate the four classical voice parts—Soprano (S), Alto (A), Tenor (T), and Bass (B)—on *ACappellaSet*. Table 6 reports mean NSDR (dB) for each stem. The model achieved solid separation quality across all parts, with higher scores for lower voices, consistent with prior observations that lower-frequency stems (e.g., Bass) exhibit less spectral overlap than higher-pitched ones.

These results (mean NSDR $= 6.72$ dB) demonstrate that the same dataset supports both multi-stem harmonic separation and VP-inclusive configurations, highlighting its flexibility for different a cappella modeling setups.

Table 6: NSDR (dB) for fine-tuned 4-stem (SATB) HTDemucs model.

Stem	S	A	T	B
NSDR (dB)	4.05	4.41	7.91	10.50

C Proposed Workflow of ACAMate

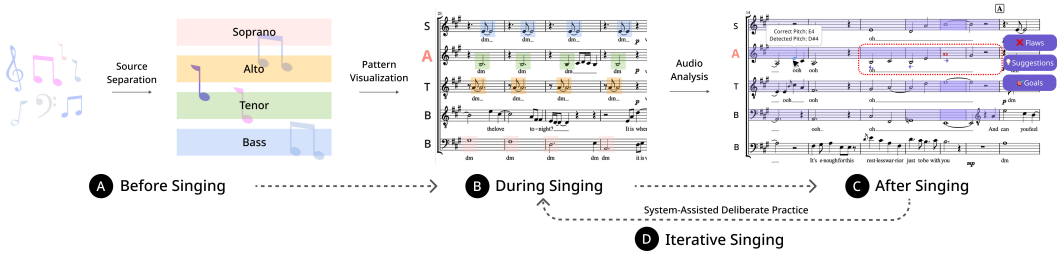


Figure 3: Workflow of ACAMate. Users iteratively practice during an individual A cappella rehearsal. Before singing (A), ACAMate separates the group recording into different voice parts, allowing users to create flexible and authentic mixes. During singing (B), ACAMate highlights musical patterns to help users perceive relationships between their part and others. After singing (C), ACAMate delivers intuitive feedback on pitch, rhythm, and dynamics. To facilitate iterative practice (D), ACAMate analyzes weak segments, offers suggestions, and sets practice goals for users.