

Improving Choral Music Separation through Expressive Synthesized Data from Sampled Instruments

Ke Chen¹, Hao-Wen Dong¹, Yi Luo², Julian McAuley¹, Taylor Berg-Kirkpatrick¹, Miller Puckette¹, Shlomo Dubnov¹

¹University of California San Diego

²Tencent AI Lab

What We Do

Choral music separation refers to the task of extracting tracks of voice parts (e.g., soprano, alto, tenor, and bass) from mixed audio. We make three contributions on the choral music separation task:

- An automated pipeline for synthesizing choral music data from **sampled instrument plugins** within controllable options for instrument expressiveness.
- A **8.2-hour-long** choral music dataset from the JSB Chorales Dataset.
- Multiple separation models of **4-part choral music separation** from different backbones.

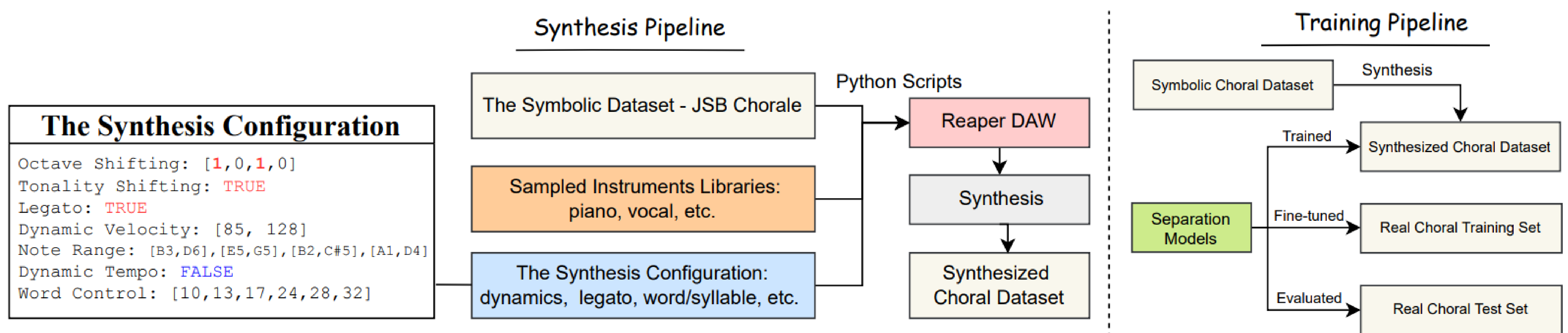
Dataset Specification

Our provided dataset is synthesized from there instrument plugins: piano, vocal, and strings. The sampled instruments we use are listed below:

Name	Type	Pitch range			
		Soprano	Alto	Tenor	Bass
Standard MIDI		A0–C8	A0–C8	A0–C8	A0–C8
Noire [17]	Piano	A0–C8	A0–C8	A0–C8	A0–C8
Grandeur [18]	Piano	A0–C8	A0–C8	A0–C8	A0–C8
Voices Of Rapture [19]	Vocal	B3–D6	E3–G5	B2–C#5	A1–D4
Dominus Choir [20]	Vocal	G3–A5	G3–A5	E2–G4	E2–G4

String datasets are sampled from *Intimate Strings Bundle*, which is detail provided in our paper.

Data Synthesis Pipeline and Model Training Pipeline



Experiments on Different Separation Backbones

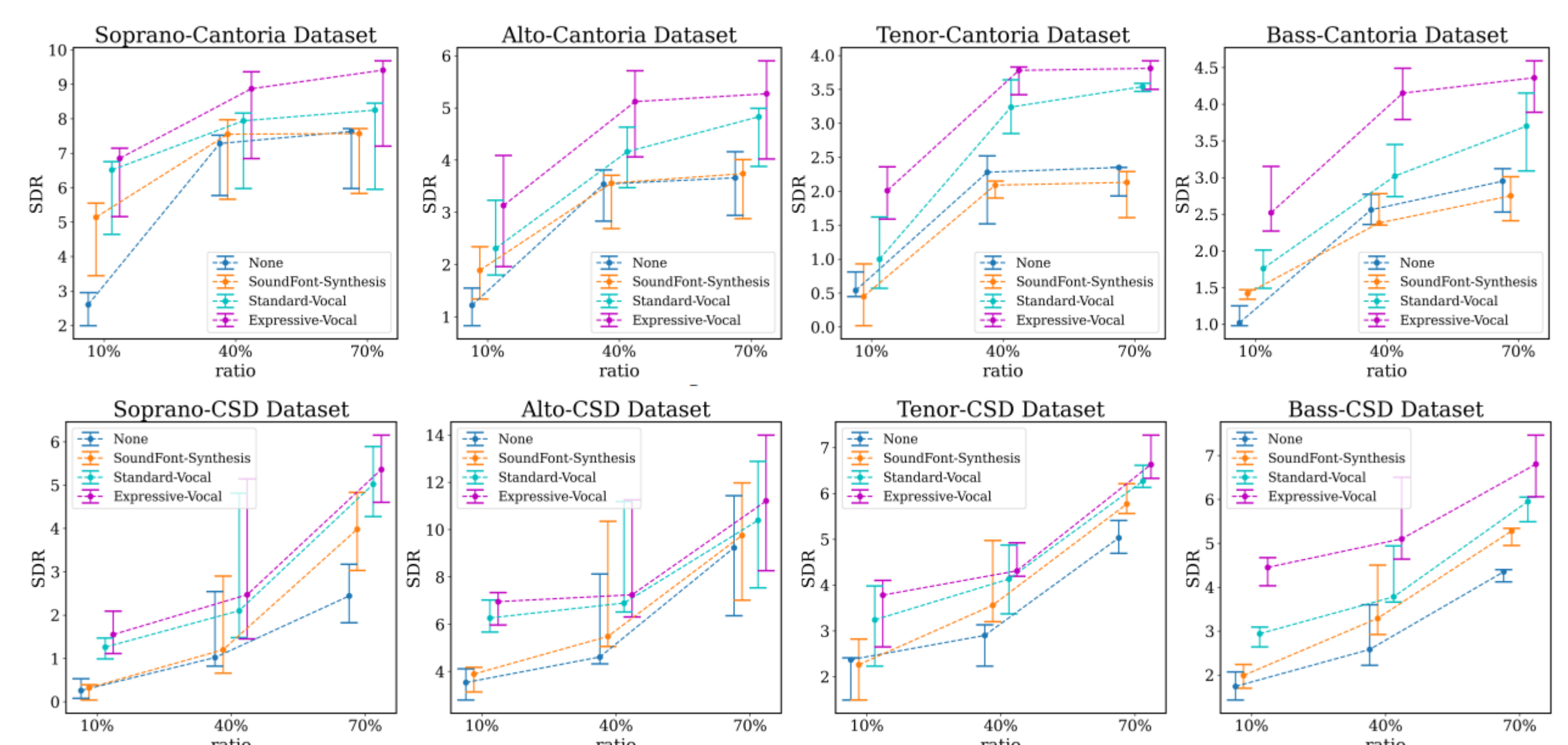
We apply four backbones to train the choral music separation models on our proposed dataset:

Standard Dataset	Model	Median Source-Distortion Ratio (dB)				
		Soprano	Alto	Tenor	Bass	Avg.
Piano	Spec-U-Net [7]	9.78	9.46	10.35	10.60	10.05
Piano	Res-U-Net [8]	8.53	9.01	9.97	12.23	9.94
Piano	Wave-U-Net [9]	6.95	5.36	7.21	9.82	7.34
Piano	Conv-TasNet [10]	7.04	6.98	7.29	7.82	7.28
Vocal	Spec-U-Net [7]	10.45	10.19	12.25	9.53	10.61
Vocal	Res-U-Net [8]	9.35	10.87	10.20	10.77	10.30
Vocal	Wave-U-Net [9]	2.65	3.08	3.06	3.90	3.17
Vocal	Conv-TasNet [10]	6.60	6.12	6.41	6.58	6.43

Among four backbones, Spec-U-Net and Res-U-Net, as frequency-domain models, perform better than Wave-U-Net and Conv-TasNet as time-domain models.

Experiments on Finetuning Experiments

We evaluated the separation models pretrained by our dataset on the real choral music separation datasets, to determine if they get improvements:



In both Cantoria Dataset and Choral Singing Dataset, our pretrained models achieve significant better results in the source-to-distortion ratio performance.

Code and Demo

