PetCoCre: Zero-Shot Cross-Species Musical Co-Creation for Wellness Therapy

Zihao Wang[1,2] , **Le Ma**[1] , **Yuhang Jin**[1] , **Yongsheng Feng**[4] , **Xin Pan**[5] , **Shulei Ji**[1,3] and **Kejun Zhang**[1,3*]
[1]Zhejiang University
[2]Carnegie Mellon University
[3]Innovation Center of Yangtze River Delta, Zhejiang University
[4]Central Conservatory of Music, China
[5]Independent Researcher
{carlwang, maller, 3210103422}@zju.edu.cn, fengyongsheng@mail.ccom.edu.cn,
darthpanath@gmail.com, {shuleiji, zhangkejun}@zju.edu.cn

Abstract

This paper explores AI-mediated human-pet musical co-creation from an interdisciplinary perspective, leveraging recent advancements in animal-assisted therapy. These advancements have shown significant psychosocial benefits, especially in reducing anxiety and enhancing social engagement. Building on these findings, this study innovatively employs pet vocal timbres as 'digital avatars' to enhance emotional investment during the music creation process. We propose PetCoCre, a novel system that applies pet vocal timbres in three distinct character paradigms within AI music creation: (1) PetRhythm: using pet voices as rhythmic percussion through beat synchronization. (2) PetMelody: enabling pet voices to act as melodic instruments via pitch-shifting alignment. (3) PetVocalia: utilizing pet vocal timbres as the target timbre for SVC (Singing Voice Conversion), where the converted singing voice replaces the original singer's voice, thus preserving the original semantic content. Beyond these character paradigms, our technical innovation lies in proposing SaMoye, the first open-source, high-quality zero-shot SVC model that effectively overcomes existing methods' zero-shot limitations by employing mixed speaker embeddings for timbre enhancement and leveraging a large-scale singing voice dataset. In our experiments, we collected dog and cat vocalization data from pet stores and conducted experiments with 30 participants. Results demonstrate that the human-pet co-creation mode led to significant enhancements in pleasure and creative satisfaction compared to solo AI music generation, along with a significant reduction in participants' anxiety levels. Through collaborative art creation, this research pioneers new paradigms for animal-assisted therapeutic interventions and expands the boundaries of AI-assisted creative collaboration.

1 Introduction

The increasing prevalence of pet ownership has transformed the perception of animals from mere companions to integral members of the family unit. This shift has sparked interest in exploring innovative therapeutic modalities that leverage the unique bond between humans and their pets.

Recent advances in animal-assisted therapy (AAT) have demonstrated significant psychosocial benefits, particularly in reducing anxiety and improving social engagement in ASD (Autism Spectrum Disorder) populations [Sissons *et al.*, 2022]. However, traditional AAT modalities predominantly focus on tactile interactions (e.g., petting), while auditory co-creation remains underexplored.

Current music therapy paradigms primarily focus on therapists playing music based on user feedback [Hunter *et al.*, 2011], AI algorithms generating music according to user emotions [Wang *et al.*, 2024c], and involving users in the music creation process through gamification [Hahn *et al.*, 2006]. However, the role of pets has not yet been incorporated into music therapy.

This paper introduces a pioneering approach to animal-assisted therapy by proposing a novel framework for musical co-creation between humans and their pets. By integrating the vocal characteristics of pets into the music creation process, we aim to enhance emotional engagement and therapeutic outcomes, thus providing a fresh perspective on AAT. To achieve this, we have developed a system, PetCoCre, that incorporates three distinct functionalities: PetRhythm, PetMelody, and PetVocalia.

Beyond these three character paradigms, our technical innovation lies in proposing SaMoye, the first open-source, high-quality zero-shot SVC model. Existing open-source SVC methods struggle with zero-shot conversion due to incomplete feature disentanglement or reliance on speaker look-up tables. SaMoye disentangles singing voice features into content, timbre, and pitch. Furthermore, we enhance timbre features by unfreezing the speaker encoder, and mixing the speaker embedding with those of the top-3 most similar speakers. We also collected and established a large-scale singing voice dataset (comprising 1,815 hours of pure singing voice and 6,367 speakers) to ensure zero-shot SVC perfor-

[*] Corresponding author.

Figure 1. The PetCoCre system transforms any input song by integrating pet-derived timbres into the resulting audio, where these timbres can function as a rhythmic instrument, a melodic instrument, or the target timbre for singing voice conversion (SVC).

mance. We conduct objective and subjective experiments to find that SaMoye outperforms other models in zero-shot SVC tasks under extreme conditions like converting singing to animals' timbre.

Furthermore, we conducted a series of experiments to validate the efficacy of this co-creation paradigm. By engaging participants in collaborative music-making with pets, we aimed to assess the impact of this interactive experience on emotional well-being, particularly in alleviating anxiety and enhancing pleasure. The results show that participants' creative satisfaction and pleasure were significantly enhanced, and their anxiety levels were markedly reduced, underscoring the potential of this interdisciplinary approach to address pressing social issues.

Open-source links: Code[1] and Model Checkpoint[2].

2 Background

2.1 Current Status and Potential Value of Animal-Assisted Therapy

Pandey et al. conducted a systematic study of the qualitative and quantitative evidence of AAT's role in enhancing patients' well-being, finding that it significantly reduces anxiety and improves overall quality of life [Pandey *et al.*, 2024]. Flynn et al. reviewed how AAT can improve engagement in behavioral and mental health services, highlighting its ability to attract patients and increase treatment adherence [Flynn *et al.*, 2022]. Sissons et al. provided a systematic review of AAT for improving social functioning in children with autism, noting that interactions with animals, particularly horses, can enhance social skills and reduce anxiety [Sissons *et al.*, 2022]. Dimolareva and Dunn (2021) conducted a meta-analysis showing that AAT can improve social skills and communication in children with ASD (Autism Spectrum Disorder), though the effectiveness varies by intervention type [Dimolareva and Dunn, 2021].

Traditional AAT modalities predominantly focus on tactile interactions (e.g., petting), while auditory co-creation remains underexplored despite music therapy's proven efficacy in neurorehabilitation. De Witte et al. demonstrated through a systematic review and meta-analysis that music therapy is highly effective in reducing stress and improving mental health [De Witte *et al.*, 2022]. Barnett and Vasiu reviewed the neural mechanisms behind the therapeutic effects of creative arts, showing how they can enhance psychological and physical health [Barnett and Vasiu, 2024].

2.2 Existing SVC Models and Their Limitations

SVC (Singing Voice Conversion) aims to convert the timbre in a given song to the reference audio without disrupting the original content. This technique has wide applications such as virtual singers, music production and other artistic domains which currently experience considerable growth thanks to the advances in AI such as AI-based music and art [Wang *et al.*, 2022; Wang *et al.*, 2024b; Wang *et al.*, 2024a; Mao *et al.*, 2023].

SVC methods predominantly rely on the recognition-synthesis framework, which involves recognizing the content of the singing voice and then synthesizing it in the target voice. [Nercessian, 2021] use a pretrained LSTM to extract the speaker embedding and concatenate it with the original speaker's phoneme and loudness embedding in the decoder to generate the converted audio. PitchNet [Deng *et al.*, 2020] introduces a singer prediction network and pitch regression network to control the timbre and pitch stability. The embeddings from the two networks are fed into the decoder with the output from the encoder to generate the audio. [Luo *et al.*, 2020] use separate encoders to extract singer and techniques embedding for singer and techniques classification tasks respectively and are concatenated before feeding into the decoder and refinement network to generate the converted audio. [Polyak *et al.*, 2020] takes as input the speech features by a pre-trained automatic speech recognition (ASR) model Wav2Letter, the F0 feature by Crepe, and the loudness feature from the power spectrum. The speaker embedding from the target singer is included in the generator for converted audio generation. [Li *et al.*, 2021] introduce F0 features, PPGs features as content features, and the Mel-spectrogram as the timbre features, which is enhanced through singer classification and reconstruction. FastSVC [Liu *et al.*, 2021b] leverages sine-excitation signals and loudness features and uses a Conformer model to extract content features. These studies introduce HiFi-GAN and BigVGAN as the vocoder to generate the converted audio. VITS refers to Variational Inference with Adversarial Learning for End-to-End Text-to-Speech.

3 Method

3.1 PetCoCre System Overview

This paper aims to explore the cognitive processes underlying the enhancement of user emotional engagement through the use of ”pet timbre” as a digital avatar, utilizing theories of social presence and examining the psychological mechanisms involved in anthropomorphic projection. We introduce PetCoCre, a human-pet musical art co-creation system, which defines three paradigms for incorporating pet sounds

[1] https://github.com/CarlWangChina/SaMoye-SVC

[2] https://huggingface.co/karl-wang/SaMoyeSVC/tree/main

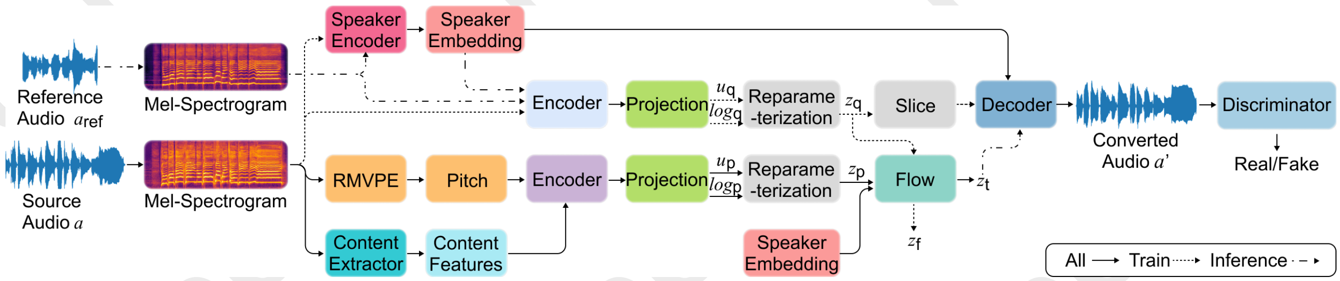


Figure 2. The overall framework of the PetVocalia (SaMoye model). u_q and log_q are of the posterior distribution, and u_p and log_p are of the prior distribution. z_q, z_p, z_t, z_t are sampled from their corresponding latent space, among which z_t is from the forward-process of the Flow model and z_f is from the inverse process.

into AI-driven music creation: (1) PetRhythm: Rhythmic percussion through beat synchronization, where pet sounds are used as percussive instruments; (2) PetMelody: Melodic instruments via pitch-shifting alignment, employing pet sounds as primary melodic instruments; and (3) PetVocalia: Vocal synthesis combining pet timbre with human semantic content, which involves extracting pet timbres and performing sing voice conversion.

3.2 PetRhythm

PetRhythm involves the use of pet sounds as rhythmic percussion instruments, synchronized with beats to create a unique musical texture, thereby enhancing the rhythmic complexity and richness of the music. We employ an real-time beat recognition algorithm to align the pet vocalizations with the musical beats, effectively using them as rhythmic percussion instruments. To enhance the richness and naturalness of this alignment, we introduce a stochastic model that simulates the natural fluctuations in biological rhythms by incorporating randomness in the intervals between beats and vocalizations.

3.3 PetMelody

PetMelody transforms pet sounds into melodic instruments by aligning their pitches, allowing pets' voices to serve as primary musical elements, thus integrating pet vocalizations seamlessly into the melodic structure of compositions.

We first utilize the SheetSage algorithm to extract the main melody in the form of a MIDI file from a chosen audio song. Subsequently, the median F0 of each pet sound is transposed to match the corresponding MIDI note pitches found in the extracted melody. This requires converting the pet's natural vocal frequency into the nearest equivalent MIDI note, thereby assigning it a specific pitch within the scale. For example, if a dog's bark corresponds to a fundamental frequency of 440 Hz, this would be matched to MIDI note A4, which also has a frequency of 440 Hz. Finally, we map out the duration and beat placement for every pet sound based on its corresponding MIDI note and integrate them seamlessly into the melody, thereby making the pet sounds an integral part of the composition.

3.4 PetVocalia

In PetVocalia, the challenge of converting pet vocal timbres, which are generally 'unseen' data by conventional Singing Voice Conversion (SVC) models, necessitates a zero-shot SVC approach capable of timbre conversion from short reference audio without fine-tuning. However, existing open-source SVC models typically perform poorly on such zero-shot tasks, often requiring minutes to hours of singing data for fine-tuning. Given the absence of open-source zero-shot SVC models that met our performance requirements, we developed SaMoye model by modifying, optimizing, and scaling up the training of the Whisper-VITS-SVC[3] architecture.

Overall Architecture

The overall architecture is illustrated in Figure 2. Considering the importance of pitch in singing, we use RMVPE to extract pitch features. We use GE2E as the speaker encoder to extract speaker embedding from the audio as the timbre features. We use a Flow model to obtain the prior distribution from these features, while another posterior encoder derives the posterior distribution from the original waveform and speaker embedding. SaMoye is trained on audio reconstruction by aligning the posterior distribution with the prior distribution through minimizing their KL divergence. We also introduce a multi-scale discriminator for adversarial learning. In the inference stage, we extract content features and pitch features from the original audio, and timbre features from the reference audio. These are used to obtain their prior distribution via the Flow model, which then generates the converted results. For timbre features, instead of freezing the speaker encoder during training, we train it jointly to enhance the timbre information in the speaker embedding. For content features, we fuse the features from existing ASR models, HubertSoft and Whisper.

For content features, we fuse the features from existing ASR models, HubertSoft and Whisper. We use a Flow model to obtain the prior distribution from these features, while another posterior encoder derives the posterior distribution from the original waveform and speaker embedding. SaMoye is trained on audio reconstruction by aligning the posterior distribution with the prior distribution through minimizing their KL divergence. We also introduce a multi-scale discriminator for adversarial learning. In the inference stage, we extract content features and pitch features from the original audio, and timbre features from the reference audio. These are used to obtain their prior distribution via the Flow model, which then generates the converted results. Instead of freezing the

[3] https://github.com/PlayVoice/whisper-vits-svc

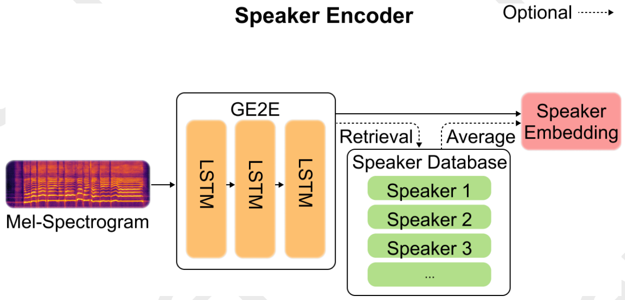


Figure 3. The process of extracting the speaker embedding with Retrieval Inference Strategy.

speaker encoder during training, we train it jointly to enhance the timbre information in the speaker embedding.

Timbre Leakage Problem

Some early studies introduced speaker look-up tables for timbre features, making it difficult to extend the table for unseen speakers. More recent studies dropped speaker look-up tables and trained a speaker encoder to extract speaker embeddings from a given audio. However, this led to the problem of *timbre leakage* in content features. Content features are extracted by pre-trained automatic speech recognition (ASR) models such as Hubert, Whisper or ContentVec. Although these ASR models preserve semantic information while reducing timbre information in content features, timbre leakage when faced with unseen data can result in the converted audio being more similar to the original than the reference audio.

Adversarial Learning

We introduce an adversarial training strategy to train the generator to synthesize converted audio and a discriminator to distinguish between generated and real audio. For the generator, the Mel-spectrogram and speaker embedding pass through the posterior encoder to get the posterior latent variables z_q. We apply the BigVGan decoder to generate the audio from the latent variables. Meanwhile, the pitch and content features pass another encoder to get the latent variables z_p. We use the speaker embedding as the condition for a Flow model to generate the prior latent variables z_t to be aligned with the posterior latent space. We compute the Kullback–Leibler divergence (KL) between the mean μ and variance σ^2 of z_t and z_q, as well as z_f and z_p, and sum them up as $\mathcal{L}_{kl}$. We also sum the *L1* and *L2* loss for the waveform as $\mathcal{L}_{wav}$ and for the mel-spectrogram as $\mathcal{L}_{mel}$ for audio reconstruction. We also use $\mathcal{L}_{stft}$ following [Takaki *et al.*, 2019].

$$\mathcal{L}_{dis} = E\left[D(a)^2 + (1 - D(a'))^2\right] \tag{1}$$

$$\mathcal{L}_{adv} = E\left[(1 - D(a'))^2\right] \tag{2}$$

For the discriminator, we use the multi-scale and multi-period discriminator in our study, which takes the generated and real audio as input to compute the *L1* loss between their feature maps in the discriminator as $\mathcal{L}_{fmap}$. The loss for the discriminator is shown in (1) and the adversarial loss for the generator is described in (2), where D is the discriminator, and a and a' are real and generated audio, respectively.

The final loss for the generator is shown below, where we set α to 1.0, β to 0.2, and γ to 9.0 during training.

$$\begin{aligned}\mathcal{L}_{gen} &= \mathcal{L}_{wav} + \mathcal{L}_{mel} + \beta * \mathcal{L}_{kl} \\ &+ \mathcal{L}_{adv} + \mathcal{L}_{fmap} + \gamma * \mathcal{L}_{stft}\end{aligned} \tag{3}$$

Retrieval Inference Strategy

We are inspired by retrieval-based voice conversion to introduce another inference strategy shown in Figure 3. In the inference stage, SaMoye retrieves the top-3 most similar speaker embeddings from a speaker embedding database and averages them with the reference speaker embedding. We build the speaker embedding database from the dataset to compute the cosine similarity between the reference audio and the embeddings in the database. Then, in the inference stage, we can retrieve the top three similar speaker embeddings and average them with the reference speaker embedding to serve as the timbre features. This method may reduce the timbre similarity of the converted audio but can improve zero-shot performance by exploiting seen speaker embeddings to fill the gap introduced by unseen reference audio, which can otherwise result in adverse outcomes such as mute voice.

Expanding Training Data

The primary limitation for zero-shot SVC task performance is the training data. SVC training data can be classified as parallel (multiple singers performing the same song) or non-parallel. Due to the limited availability of parallel data, existing models are typically trained on non-parallel data. The number of speakers in the base model significantly impacts zero-shot SVC performance, as zero-shot SVC requires models to see as much data as possible to enhance their generalization capability. Therefore, in addition to pet sounds, we have collected a substantial amount of human voice data to support subsequent training.

By collecting online audio and open-source datasets, we established a large-scale non-parallel SVC open-source singing and speech dataset comprising 1,815 hours and 6,367 speakers. We manually checked this dataset to ensure its quality. Specifics of these datasets are detailed in Table 1.

For data processing, all audio sampling rates are unified to 32kHz. When converting waveforms to Mel-spectrograms with 80 filters, we use a filter length of 1024, a hop length of 320, and a Hanning window of length 1024. This dataset includes 36.5 hours of online music, from which we separate pure human vocals using Demucs [Défossez *et al.*, 2019]. Subsequently, the obtained data are manually checked to remove any errors in recognition results.

Due to the inclusion of multiple singing datasets, there are balancing considerations regarding aspects such as language and style (e.g., differences in timbre and tone across various languages). As this is not the primary research focus of this paper, interested readers can refer to existing studies that have conducted relevant comparisons, such as [Creel *et al.*, 2023].

Datasets	**Speakers**	**Duration (hours)**
JSUT-Song [Sonobe *et al.*, 2017]	1	0.41
PJS [Koguchi and Takamichi, 2020]	1	0.60
KiSing [Shi *et al.*, 2022]	1	0.88
Jvs Music [Tamaru *et al.*, 2020]	100	4.00
CSD [Choi *et al.*, 2020]	1	4.86
Opencpop [Huang *et al.*, 2021]	1	5.20
DSD100 [Liutkus *et al.*, 2017]	100	6.99
Popcs [Liu *et al.*, 2021a]	117	5.89
KSS [Park, 2018]	1	12.85
M4Singer [Zhang *et al.*, 2022]	20	29.77
OpenSinger [Huang *et al.*, 2021]	66	50.00
VCTK [Valentini-Botinhao and others, 2017]	109	44.00
Aishell-3 [Shi *et al.*, 2020]	218	85.00
DAMP VPB [Smule, 2017]	5428	1529.00
Online Music	203	36.5
Total	6367	1815.95

Table 1. The statistics of the Collected Human Vocal Datasets.

4 Experiment Setup

4.1 Singing Voice Conversion Experiment Metrics and Participants

Since pet timbre conversion is inherently a zero-shot SVC task, and this particular task represents an extreme testing scenario that may result in performance degradation, we need to determine which model is best suited to serve as the base model for the PetVocalia functionality. We use several metrics to evaluate the quality of the converted audio and how similar the timbre is to the target pet timbre:

- Timbre Similarity: This metric is based on a percentage score, where 1 indicates identical timbre and 0 indicates completely different timbre. In this study, the metric is evaluated against the target pet timbre.
- Mean Opinion Score on Quality(MOS): MOS is a widely-used audio or video quality evaluation standard based on expert evaluation. The score of MOS is from 1 to 5, where a higher score means higher quality.
- Perceptual Evaluation of Speech Quality (PESQ): PESQ computes multiple perspectives like temporal alignment and perceptual filtering. PESQ score is from -0.5 to 4.5, where a higher score stands for better perceptual quality.
- Short-Time Objective Intelligibility (STOI): STOI represents how well the audio can be comprehended, and the STOI score is from 0 to 1. The higher STOI score means that the audio is easier to understand.
- Non-Intrusive Speech Quality Assessment (NISQA) [4]: NISQA is a non-intrusive metrics based on a pre-trained deep learning model. Given audio, the NISQA predicts MOS for audio quality, NOI for noise degree, DIS for audio coherence, COL for timbre quality, and Loud for loudness. All these predicted scores are the higher the better.

We selected five song clips, each 20 seconds in length, and 11 speakers (6 humans and 5 pets) to evaluate the model's performance. These five songs cover the full range from bass to treble. The human timbres included 4 male and 2 female voices, while the 5 pet timbres consisted of 3 cat and 2 dog sounds.

For the subjective evaluation, we recruited 19 professional musicians. These participants (9 female, 10 male; aged 21-37) are distinct from those involved in the subsequent human-pet co-creation experiment. They listened to the original music and target timbres, and then evaluated the converted song clips. We have documented their areas of musical specialization (e.g., electronic music production, musicology, music education). An anonymized list of these participants, with names redacted for privacy reasons, can be provided in the appendix.

4.2 Human-Pet Co-Creation Experiment Metrics and Participants

Our experiment adopts a dual-group control design involving 30 healthy adult participants. This design enables comparative analysis of response differences between groups under various conditions. Each participant was instructed to engage in AI music creation under different paradigms. These sessions could not be conducted consecutively on the same day, in order to avoid carryover effects between the three paradigms that might result from continuous sessions. Before and after each creation session, participants were required to complete web-based questionnaires to record relevant metrics. Three key psychological scales were selected to assess subjective experiences:

- **Creative Pleasure:** Measured using the PANAS-X (Positive and Negative Affect Schedule - Expanded Form), particularly its Positive Affect subscale, containing 10 semantic differential items (e.g., excited-calm, inspired-dull).
- **Anxiety Reduction:** Evaluated through the state subscale S-AI of STAI (State-Trait Anxiety Inventory), adapted for music creation contexts with 12 retained items. It assesses anxiety in various populations, including medical and psychiatric patients, occupational groups, and evaluates psychotherapy and medication effectiveness.
- **Pet Anthropomorphism Perception:** Developed from IDAQ scale, this 5-dimensional assessment system (emotional resonance, intentional attribution, mental simulation, role projection, social interaction) uses 7-point Likert scales.

These scales were chosen for their cross-cultural adaptability and established reliability in assessing emotional states and mental health.

5 Experiment Results & Analysis

5.1 Technical Performance

Baseline Comparison SVC Models

To evaluate the SaMoye model in PetVocalia, we compare it against several baseline SVC models trained on our datasets:

[4] https://github.com/gabrielmittag/NISQA

Metric	Human as Target Voice					Pets as Target Voice				
	PESQ↑	STOI↑	NISQA↑	Timbre↑ Similarity	MOS↑	PESQ↑	STOI↑	NISQA↑	Timbre↑ Similarity	MOS↑
SoVITS-Flow[5]	2.486	0.572	2.893	0.585	3.61	2.413	0.545	2.812	0.423	3.499
GPT-Sovits[6]	2.31	0.546	2.434	0.573	3.592	2.253	0.526	2.314	0.454	3.405
RVC [Kamble *et al.*, 2023]	2.252	0.622	3.093	0.56	3.822	2.208	0.592	2.855	0.46	3.447
whisper-vits-SVC [Ning *et al.*, 2023]	2.643	0.651	2.978	0.574	3.945	2.592	0.643	2.916	0.476	3.767
DiffSVC [Liu *et al.*,]	2.917	0.673	3.201	**0.598**	4.231	2.821	0.628	3.002	0.487	3.956
CoMoSVC [Lu *et al.*, 2024]	**2.948**	**0.686**	**3.212**	0.585	**4.277**	2.867	0.631	2.971	0.478	3.903
SaMoye[7]	2.890	0.675	3.187	0.580	4.167	**2.877**	**0.667**	**3.133**	**0.556**	**4.131**

Table 2. Experimental results of the benchmark test of the Zero-shot SVC models in PetVocalia.

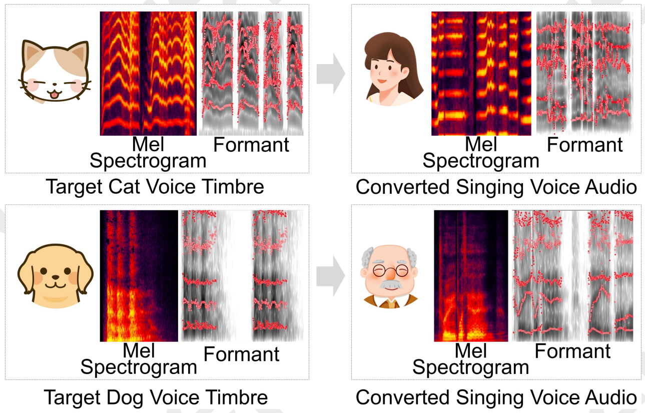


Figure 4. Specific case analysis of audio converted using the PetVocalia (SaMoye model), where the pet timbre is the target timbre, shows that the mel-spectrogram and formant of the converted audio closely resemble those of the animal timbre.

- SoVITS-Flow[8]: A Flow model-based SoVITS-SVC, using *Contentvec* for representations.
- GPT-Sovits[9]: Originally a TTS model, adapted for SVC by replacing phoneme with F0 embeddings and training on our datasets.
- RVC [Kamble *et al.*, 2023][10]: Retrieval-based Voice Conversion (RVC) is a popular open-source, VITS-based SVC technique using retrieval of similar acoustic features.
- whisper-vits-SVC [Ning *et al.*, 2023][11]: Combines OpenAI's Whisper for speaker-independent content feature extraction and VITS for singing voice synthesis.
- DiffSVC [Liu *et al.*,]: A Denoising Diffusion Probabilistic Model (DDPM)-based SVC system, using Phonetic PosteriorGrams (PPGs) as content features and recovering target mel-spectrograms via a diffusion process.
- CoMoSVC [Lu *et al.*, 2024]: A Consistency Model-based SVC method via distillation from a pre-trained diffusion model (teacher) for few-step inference, significantly improving speed while maintaining performance.

Speaker Encoder	Pets as Target Voice				
	PESQ↑	STOI↑	NISQA↑	TBS↑	MOS↑
Fixed	2.105	0.614	2.796	0.405	3.350
Unfrozen	**2.877**	**0.667**	**3.133**	**0.556**	**4.131**

Table 3. The results of fixed and unfrozen speaker encoder in the PetVocalia (SaMoye model). TBS is the abbreviation for Timbre Similarity. All subjective metrics exhibit statistically significant differences with $p < 0.05$.

Zero-Shot SVC Benchmark Testing

As shown in Table 2, we found that compared to using human singers as the target timbre, using animals as the target timbre leads to a decline in model performance. This is because animal timbres are unseen data for all SVC models, representing an extreme condition for zero-shot singing voice conversion. Furthermore, the results show that SaMoye demonstrates significant performance improvements on both Human and Pet timbres compared to whisper-vits-SVC. Although its performance on Human timbres still lags behind CoMoSVC and DiffSVC, SaMoye's zero-shot performance on Pet timbres surpasses other models, thus making it more suitable as the base model for the PetVocalia functionality. This further explains that SaMoye's retrieval strategy during inference can fill information gaps in speaker embeddings, especially for animal cases where the timbre information gap is larger.

Ablation Study: Evaluation of Unfreezing the Speaker Encoder

We also evaluate the effect of unfreezing the speaker encoder during training. The results, as shown in Table 3, indicate that unfreezing the speaker encoder during training enhances performance across all metrics.

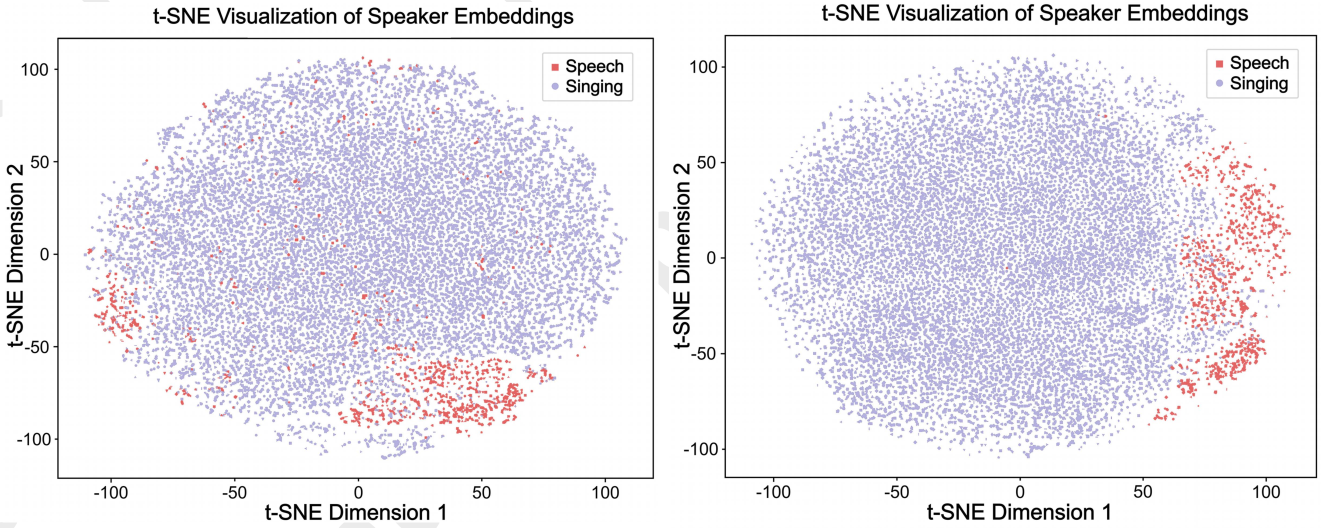
In addition, we use t-SNE to reduce the speaker embedding to two dimensions for visualization, as shown in Figure 5. We selected 328 speakers from speech datasets including Aishell, KSS, and VCTK (denoted in red), and 6,299 speakers from our singing datasets (denoted in purple). The figure illustrates that unfreezing the speaker encoder during training allows for a clearer distinction between speaker embeddings from speech and singing. This helps to enhance the timbre features of singing voices and improve the performance of zero-shot SVC tasks. This is because the wider pitch range in singing leads to more complex timbres compared to speech timbres, which typically have a very narrow pitch range. If not differentiated, using speech timbres for singing can lead to adverse consequences, such as intermittent sound issues in high and low pitch regions.

[8] https://github.com/svc-develop-team/so-vits-svc

[9] https://github.com/RVC-Boss/GPT-SoVITS

[10] https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI

[11] https://github.com/PlayVoice/whisper-vits-svc



(a) Fixed speaker encoder.

(b) Unfrozen speaker encoder.

Figure 5. The t-SNE visualization of the speaker embedding from fixed (left) and unfrozen (right) speaker encoders in the PetVocalia (SaMoye model). Each point represents a speaker.

Metric	PANAS-X	STAI	IDAQ
Main Effect (Co-creation vs. Solo)			
Co-creation Group	**4.32** (0.68)	**3.85** (0.72)	**4.90** (0.65)
Solo Creation Group	3.15 (0.72)	2.93 (0.81)	3.24 (0.59)
Paradigm Comparison (Within Co-Creation Group)			
PetRhythm	**4.51** (0.61)	**3.92** (0.65)	4.03 (0.72)
PetMelody	4.23 (0.57)	3.65 (0.71)	4.35 (0.68)
PetVocalia	4.18 (0.63)	3.57 (0.69)	**5.37** (0.61)

Table 4. Experimental Results of Human-Pet Co-Creation. The meanings corresponding to each metric: PANAS-X (Pleasure), STAI (Anxiety), IDAQ (Anthropomorphism). Anxiety reduction scores represent change values (higher=better). Bold values denote optimal paradigm performance. Standard deviations in parentheses. All $p<0.05$.

5.2 Human-Pet Co-Creation Experiment Results

The experiment is implemented through the PetCoCre system, investigating a novel form of animal-assisted therapy involving collaborative music creation between humans and pets.

Main Effect (Co-creation vs. Solo)

In the first experiment, we aim to investigate whether higher satisfaction is exhibited when collaborating with pets in artistic creation compared to independent AI music composition. To validate this hypothesis, the independent creation mode of the control group (Solo Creation Group) allows participants to freely select digital instruments as rhythm/lead instruments and extract personal vocal timbre as lead vocals, without utilizing pet sounds as core elements. In contrast, the co-creation conditions of the experimental group (Co-Creation Group) require participants to use their pets' vocal characteristics (or a standard sample library) for AI music generation.

Experimental results, as shown in Table 4, reveal that artistic co-creation with pets significantly enhances users' creative pleasure and anxiety reduction effects compared to solo AI music creation. These findings provide empirical support for the digital transformation in animal-assisted therapy.

Paradigms Comparison (Within Co-Creation Group)

Furthermore, within the Co-Creation Group, we also systematically compared the differences among three human-pet music co-creation paradigms: PetRhythm (rhythm paradigm), PetMelody (melody paradigm), and PetVocalia (vocal paradigm). Each co-creation paradigm fundamentally incorporates pet sounds to enhance emotional engagement and explore their impact on the creative process.

As shown in Table 4, among the three paradigms, PetRhythm exhibited optimal performance in positive affect arousal and anxiety reduction, potentially due to rhythm synchronization-induced motor empathy effects, whereas PetVocalia achieved higher scores in anthropomorphism perception.

6 Conclusion & Discussion

This study introduces the PetCoCre system and the SaMoye model, which effectively integrates pet vocalizations into music creation by addressing the limitations of Zero-shot SVC open-source models. This process, utilizing pet vocalizations as their digital avatars, enhances emotional bonds and elevates the creative experience. Furthermore, this work pioneers a new paradigm for animal-assisted therapy, demonstrating significant therapeutic benefits in alleviating anxiety and enhancing pleasure. Moreover, this innovative approach highlights the potential of artistic co-creation in mental health interventions.

Ethical Statement

This study primarily focuses on human experiences and agency, emphasizing human emotional changes and anxiety alleviation during creation, while neglecting the active participation and experiences of animals. The current perspective only treats animals as passive participants, limiting the breadth of the overall definition of human-pet co-creation. Future research should aim to develop a more equitable co-creation system that allows animals to actively engage and express their emotions, thereby improving animal welfare and enhancing mutual understanding in human-pet interactions.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.62272409).

Contribution Statement

Authors Zihao Wang and Le Ma contributed equally to this work.

References

[Barnett and Vasiu, 2024] Kelly Sarah Barnett and Fabian Vasiu. How the arts heal: a review of the neural mechanisms behind the therapeutic effects of creative arts on mental and physical health. *Frontiers in behavioral neuroscience*, 18:1422361, 2024.

[Choi *et al.*, 2020] Soonbeom Choi, Wonil Kim, Saebyul Park, Sangeon Yong, and Juhan Nam. Children's song dataset for singing voice research. In *International Society for Music Information Retrieval Conference (ISMIR)*, volume 4, 2020.

[Creel *et al.*, 2023] Sarah C Creel, Michael Obiri-Yeboah, and Sharon Rose. Language-to-music transfer effects depend on the tone language: Akan vs. east asian tone languages. *Memory & Cognition*, 51(7):1624–1639, 2023.

[De Witte *et al.*, 2022] Martina De Witte, Ana da Silva Pinho, Geert-Jan Stams, Xavier Moonen, Arjan ER Bos, and Susan Van Hooren. Music therapy for stress reduction: a systematic review and meta-analysis. *Health psychology review*, 16(1):134–159, 2022.

[Défossez *et al.*, 2019] 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.

[Deng *et al.*, 2020] Chengqi Deng, Chengzhu Yu, Heng Lu, Chao Weng, and Dong Yu. Pitchnet: Unsupervised singing voice conversion with pitch adversarial network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7749–7753. IEEE, 2020.

[Dimolareva and Dunn, 2021] Mirena Dimolareva and Thomas J Dunn. Animal-assisted interventions for school-aged children with autism spectrum disorder: A meta-analysis. *Journal of autism and developmental disorders*, 51(7):2436–2449, 2021.

[Flynn *et al.*, 2022] Erin Flynn, Alexandra G Zoller, Jaci Gandenberger, and Kevin N Morris. Improving engagement in behavioral and mental health services through animal-assisted interventions: A scoping review. *Psychiatric Services*, 73(2):188–195, 2022.

[Hahn *et al.*, 2006] Mariah S Hahn, James B Kobler, Barry C Starcher, Steven M Zeitels, and Robert Langer. Quantitative and comparative studies of the vocal fold extracellular matrix i: elastic fibers and hyaluronic acid. *Annals of Otology, Rhinology & Laryngology*, 115(2):156–164, 2006.

[Huang *et al.*, 2021] Rongjie Huang, Feiyang Chen, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3945–3954, 2021.

[Hunter *et al.*, 2011] Patrick G. Hunter, E. Glenn Schellenberg, and Andrew T. Griffith. Misery loves company: Mood-congruent emotional responding to music. *Emotion*, 11(5):1068 – 1072, 2011.

[Kamble *et al.*, 2023] Anand Kamble, Aniket Tathe, Suyash Kumbharkar, Atharva Bhandare, and Anirban C. Mitra. Custom data augmentation for low resource asr using bark and retrieval-based voice conversion, 2023.

[Koguchi and Takamichi, 2020] Junya Koguchi and Shinnosuke Takamichi. Pjs: phoneme-balanced japanese singing voice corpus, 2020.

[Li *et al.*, 2021] Zhonghao Li, Benlai Tang, Xiang Yin, Yuan Wan, Ling Xu, Chen Shen, and Zejun Ma. Ppg-based singing voice conversion with adversarial representation learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7073–7077. IEEE, 2021.

[Liu *et al.*,] Songxiang Liu, Yuewen Cao, Dan Su, and Helen Meng. DiffSVC: A diffusion probabilistic model for singing voice conversion. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 741–748. IEEE.

[Liu *et al.*, 2021a] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, Peng Liu, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. *arXiv preprint arXiv:2105.02446*, 2, 2021.

[Liu *et al.*, 2021b] Songxiang Liu, Yuewen Cao, Na Hu, Dan Su, and Helen Meng. Fastsvc: Fast cross-domain singing voice conversion with feature-wise linear modulation. In *2021 ieee international conference on multimedia and expo (icme)*, pages 1–6. IEEE, 2021.

[Liutkus *et al.*, 2017] Antoine Liutkus, Fabian-Robert Stöter, Zafar Rafii, Daichi Kitamura, Bertrand Rivet, Nobutaka Ito, Nobutaka Ono, and Julie Fontecave. The 2016 signal separation evaluation campaign. In Petr Tichavský, Massoud Babaie-Zadeh, Olivier J.J. Michel, and Nadège Thirion-Moreau, editors, *Latent Variable Analysis and Signal Separation - 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25-28, 2015, Proceedings*, pages 323–332, Cham, 2017. Springer International Publishing.

[Lu *et al.*, 2024] Yiwen Lu, Zhen Ye, Wei Xue, Xu Tan, Qifeng Liu, and Yike Guo. CoMoSVC: Consistency

Model-based Singing Voice Conversion. *arXiv preprint arXiv:2401.01792*, 2024.

[Luo *et al.*, 2020] Yin-Jyun Luo, Chin-Cheng Hsu, Kat Agres, and Dorien Herremans. Singing voice conversion with disentangled representations of singer and vocal technique using variational autoencoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3277–3281. IEEE, 2020.

[Mao *et al.*, 2023] Weihang Mao, Bo Han, and Zihao Wang. Sketchffusion: Sketch-guided image editing with diffusion model. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 790–794, 2023.

[Nercessian, 2021] Shahan Nercessian. End-to-end zero-shot voice conversion using a ddsp vocoder. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5. IEEE, 2021.

[Ning *et al.*, 2023] Ziqian Ning, Yuepeng Jiang, Zhichao Wang, Bin Zhang, and Lei Xie. VITS-BASED SINGING VOICE CONVERSION LEVERAGING WHISPER AND MULTI-SCALE F0 MODELING. In *Proc. Singing Voice Conversion Challenge (SVCC)*, Oct 2023. arXiv:2310.02802.

[Pandey *et al.*, 2024] Ramendra Pati Pandey, Riya Mukherjee, Chung-Ming Chang, et al. The role of animal-assisted therapy in enhancing patients' well-being: Systematic study of the qualitative and quantitative evidence. *Jmirx med*, 5(1):e51787, 2024.

[Park, 2018] K Park. Kss dataset: Korean single speaker speech dataset. https://www.kaggle.com/datasets/bryanpark/korean-single-speaker-speech-dataset, 2018.

[Polyak *et al.*, 2020] Adam Polyak, Lior Wolf, Yossi Adi, and Yaniv Taigman. Unsupervised cross-domain singing voice conversion. *arXiv preprint arXiv:2008.02830*, 2020.

[Shi *et al.*, 2020] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567*, 2020.

[Shi *et al.*, 2022] Jiatong Shi, Shuai Guo, Tao Qian, Tomoki Hayashi, Yuning Wu, Fangzheng Xu, Xuankai Chang, Huazhe Li, Peter Wu, Shinji Watanabe, and Qin Jin. Muskits: an End-to-end Music Processing Toolkit for Singing Voice Synthesis. In *Proc. Interspeech 2022*, pages 4277–4281, 2022.

[Sissons *et al.*, 2022] Jon H Sissons, Elise Blakemore, Hannah Shafi, Naomi Skotny, and Donna M Lloyd. Calm with horses? a systematic review of animal-assisted interventions for improving social functioning in children with autism. *Autism*, 26(6):1320–1340, 2022.

[Smule, 2017] I Smule. Damp-vpb: Digital archive of mobile performances-smule vocal performances balanced, 2017.

[Sonobe *et al.*, 2017] Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. JSUT corpus: free large-scale japanese speech corpus for end-to-end speech synthesis. *CoRR*, abs/1711.00354, 2017.

[Takaki *et al.*, 2019] Shinji Takaki, Toru Nakashika, Xin Wang, and Junichi Yamagishi. Stft spectral loss for training a neural speech waveform model. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7065–7069. IEEE, 2019.

[Tamaru *et al.*, 2020] Hiroki Tamaru, Shinnosuke Takamichi, Naoko Tanji, and Hiroshi Saruwatari. Jvs-music: Japanese multispeaker singing-voice corpus, 2020.

[Valentini-Botinhao and others, 2017] Cassia Valentini-Botinhao et al. Noisy speech database for training speech enhancement algorithms and tts models. *University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR)*, 2017.

[Wang *et al.*, 2022] Zihao Wang, Kejun Zhang, Yuxing Wang, Chen Zhang, Qihao Liang, Pengfei Yu, Yongsheng Feng, Wenbo Liu, Yikai Wang, Yuntao Bao, et al. Songdriver: Real-time music accompaniment generation without logical latency nor exposure bias. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1057–1067, 2022.

[Wang *et al.*, 2024a] Zihao Wang, Shuyu Li, Tao Zhang, Qi Wang, Pengfei Yu, Jinyang Luo, Yan Liu, Ming Xi, and Kejun Zhang. Muchin: A chinese colloquial description benchmark for evaluating language models in the field of music. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 7771–7779, 8 2024.

[Wang *et al.*, 2024b] Zihao Wang, Haoxuan Liu, Jiaxing Yu, Tao Zhang, Yan Liu, and Kejun Zhang. Mudit & musit: Alignment with colloquial expression in description-to-song generation. *arXiv preprint arXiv:2407.03188*, 2024.

[Wang *et al.*, 2024c] Zihao Wang, Le Ma, Chen Zhang, Bo Han, Yunfei Xu, Yikai Wang, Xinyi Chen, Haorong Hong, Wenbo Liu, Xinda Wu, and Kejun Zhang. Remast: Real-time emotion-based music arrangement with soft transition. *IEEE Transactions on Affective Computing*, pages 1–15, 2024.

[Zhang *et al.*, 2022] Lichao Zhang, Ruiqi Li, Shoutong Wang, Liqun Deng, Jinglin Liu, Yi Ren, Jinzheng He, Rongjie Huang, Jieming Zhu, Xiao Chen, et al. M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. *Advances in Neural Information Processing Systems*, 35:6914–6926, 2022.