Composing for AI Voice Model Choir

Nick Collins
Durham University Music Department
nick.collins@durham.ac.uk

ABSTRACT

The rapid growth of generative AI capability has not left singing voice synthesis and voice transformation untouched. This paper details compositional experiments in the use of AI voice conversion models outside of their normal intended purpose. Deepfake vocal models emulating well known celebrity singers have been used for singing voice substitution in popular song to make new songs with the vocal timbre of already established voices, or for new forms of mash-up. In the current paper, however, more experimental vocal improvisation, or non-vocal sounds, are used as the guide track for singing voice substitution, often multi-tracked across multiple voice models, leading to some interesting experimental AI choral music. The current technology for singing voice substitution is reviewed, and more radical vocal model experiments detailed. A suite of six studies, the 'Music for Celebrity AI Voice Model Choir' are detailed, and ethical issues discussed.

1. INTRODUCTION

Generative AI models offer powerful new facilities to computer music composition, at the same time as they raise new ethical issues. Singing voice analysis and synthesis has been one of the hardest challenges in computer music historically [1, 2], but has seen rapid advances with the growth of new deep learning techniques [3]. Aside from full synthesis of the singing voice, the substitution of vocal timbre has seen great success, often called voice conversion. Rather convincing singing voice deepfakes grafting the timbre of established singers into new songs, or within new forms of mash-up, have been released outside of academic labs [4,5] due to the availability of such open source software as RVC (Retrieval Voice Conversion) and so-vits-svc (SoftVC VITS Singing Voice Conversion) [6]. Such software was widely available by 2023, coming to public consciousness through novel deepfake Eminem tracks, or an invented late 90s Oasis album from 'AIsis' with a Liam Gallagher vocal model to give it an authentic Mancunian twang. [1] A high profile instance was within the war of words between Drake and Kendrick Lamar in 2024, when Drake goaded Lamar in *Taylor Made Freestyle* by rapping through the voices of other rap notaries including the deceased Tupac Shakur [7]. Such voices as Freddie Mercury, Kurt Cobain and Michael Jackson have been much utilised online, though the technology also supports a much wider array of singer models from amateurs to celebrities and politicians [4].

There are precedents in computer music composition manipulating the voice. Charles Dodge's *Any Resemblance is Purely Coincidental* (1978) used a linear predictive coding model to manipulate a 1907 Caruso recording; Philippe Manoury's opera *K...* (2001) was an opportunity for IRCAM to develop effective synthesized simulations of a choir [8]. The compositional experimentation here can be contextualised alongside the work of glitch musicians who have sought to find new ways of using and manipulating digital audio through the intentional mis-use and novel application of music software [9].

The paper proceeds to outline the two primary voice conversion softwares which were explored, before detailing specific compositional experimentation within the six studies of the work *Music for Celebrity AI Voice Model Choir* (2024). [2] . Ethical issues arising in such work is confronted.

2. CURRENT DEEP LEARNING VOICE SUBSTITUTION SOFTWARE; RVC AND SOVITS

For the explorations detailed in this paper, so-vits-svc [3] was primarily used; RVC [4] was also compared. RVC often gave more accurate output in terms of fidelity of pitch to the source, though the greater number of pitch tracking errors in so-vits-svc was more creatively stimulating when voices were multi-tracked. Both softwares often crashed; so-vits-svc was more reliable run from the command line on a Mac, though the gui was more stable on PC. Voice models were sourced from Hugging Face creator QuickWick and included both famous deceased singers, and current voices, in various popular music styles. [5]

Both so-vits-svc and RVC rest on a front end pitch detector. Whilst RVC was quite stable, the default option of the crepe [10] pitch detector in so-vits-svc could be improved in pitch tracking by using the 'dio' option [11]. Figure 1 demonstrates this with a plot over a 30 second excerpt from a vocal a cappella, the lead voice from Mirroman's cover of Fever available in Mike Senior's multitracks for mixing practice. [6] Pitch tracking for the plot was carried out with the librosa [12] python library's probabilistic YIN

[1] https://www.youtube.com/watch?v=whB21dr2Hlc

[2] https://sicklincoln.bandcamp.com/album/music-for-celebrity-ai-voice-model-choir

[3] https://github.com/voicepaw/so-vits-svc-fork

[4] https://github.com/IAHispano/Applio

[5] https://huggingface.co/QuickWick/Music-AI-Voices

[6] https://cambridge-mt.com/ms/mtk

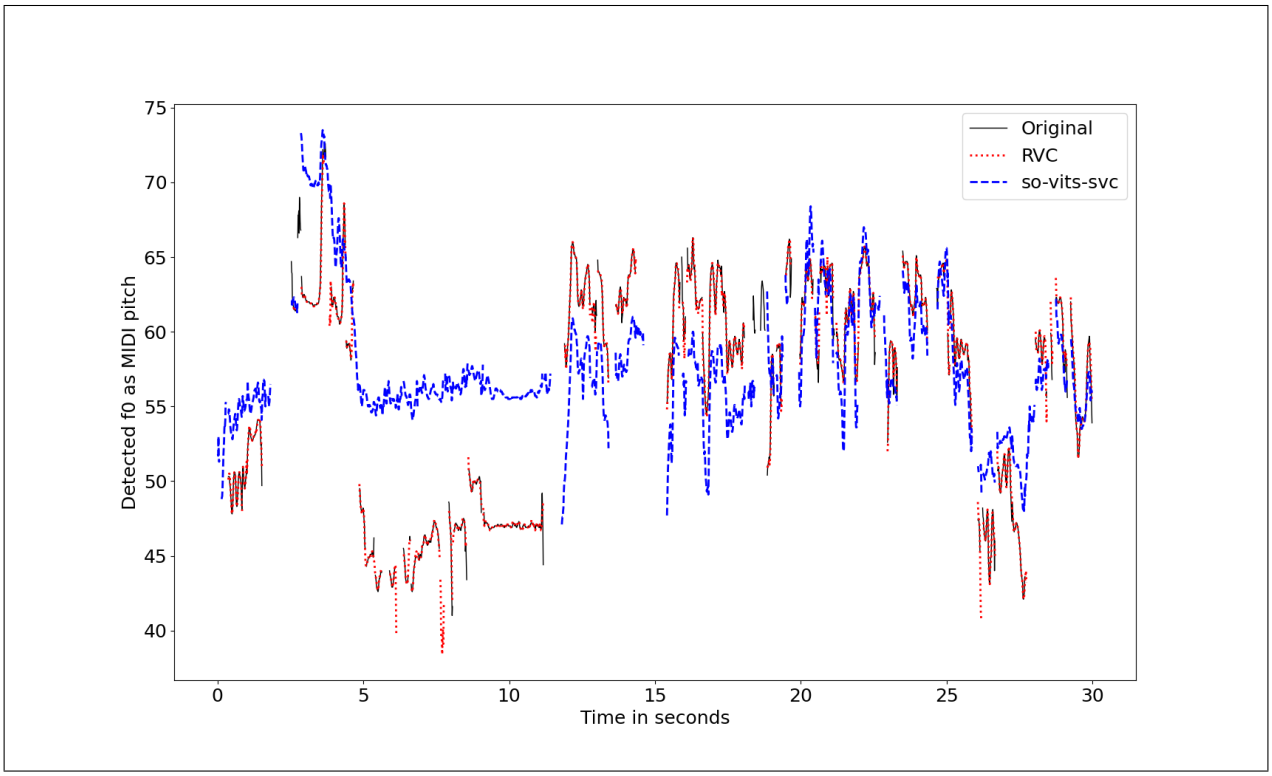

Figure 1. Pitch detection comparison of an original audio source, RVC, and so-vits-svc.

pitch detector.[7] so-vits-svc tracks pitch poorly when using crepe, and there is an interaction with the model such that different output pitch trails result for different models, giving a very weird atonal chorusing effect. Rhythm is still absolutely precise (inhumanly so).

To further push creative options, using polyphonic audio as a hidden guide track was explored. Precise and inhumanly fast rhythms with vocal timbre could be created by passing rhythmic polyphonic material through a vocal conversion; noise music (such as Merzbow [13]) made a very nice source to push vocal conversion models in unexpected ways, activating vocal noises such as sibilance. In previous work by David Kant, transcription errors in computer music analysis have been shown to be productive in creating new pieces which rest on the misreading of the source [14]. For the present project, pitch tracking inaccuracy was a particularly strong compositional stimulant.

3. MUSIC FOR CELEBRITY AI VOICE MODEL CHOIR

A collection of six studies is available for free download from `https://sicklincoln.bandcamp.com/album/music-for-celebrity-ai-voice-model-choir`. Table 1 lists the six movements of the collection, and in particular the source materials driving the process.

For each study, a particular set of sources was used to drive a collection of AI voice models (4-8) within so-vits-svc. Despite, or because, of its deficiencies, the default crepe pitch tracker was used; when the multiple renders (one per voice model) were combined, the pitch errors within individual models led to very interesting harmonies, whilst the rhythmic flow was accurately preserved and inhumanly tight. Few additional effects were used so as to focus on the vocal timbre, just varying levels of reverb, panning to give a stereo spread to help discern the sense of a choir, and occasional use of more filtering and distortion (particularly in the sixth study). To create variety, source-driven lines were often deployed in canon across voices, including tempo canon in the style of Nancarrow with convergence points [15]; time stretching was sometimes used to stretch towards a particular convergence point after staggered entries.

The inspiration for the vocal improvisation underlying the first and last studies were the vocal utterances of Trevor Wishart [16, 17]; the composer's vocals were always to be guide tracks, the original never heard but instead mediated by more famous singers (who may themselves have never thought of more extended vocal play).

The second study combined the effect of passing a full polyphonic recording through vocal models, and a recording of the Wilhelm scream.[8] The full recording as source had much faster choppier material than could be accomplished with a normal singing voice; the models act as creative vocoders, giving a singer's timbre to the percussive rhythms. The Wilhelm scream is less recognisable than the polyphonic electronic track in transformation.

The source for the third study is a tenor horn which happened to be around the family home, and which the performer was extremely unfamiliar with and had never recorded before (they just about managed to get out some sustained

[7] `https://librosa.org/doc/0.10.2/generated/librosa.pyin.html#librosa.pyin`

[8] Actually the recording session where it was created, available at `https://archive.org/details/SSE_Library_VOICES` as 'VOXScrm Man eaten by alligator; screams [Wilhelm CS USC00:3911]'

Track name	Duration	Source	Comments
Study for Celebrity Voices	03:45	Vocal improvisation	Unison and canonic effects across a quintet of voice models, two male and three female.
Study 2: Raw and Wilhelm	02:41	An existing noisy electronica track/ the famous Wilhelm scream movie sound effect	Sextet, three male, three female
Study 3: Tenor Horn	02:47	A tenor horn, played badly	Quartet (two male, two female), but tenor horn source is also allowed through sometimes, providing a strange accompaniment
Study 4: On a Microtonal Guide Track	03:25	Microtonal vocal track, itself the result of trying to sing along to a computer generated microtonal guide	Quartet (two male, two female)
Study 5: I'm not a Celebrity	01:11	Speech	The spoken voice is used as a singing voice cue; nonet (six male, three female)
Study 6	06:35	Vocal improvisation	Nonet of voices, six male, three female

Table 1. The six studies within the collection.

louder notes). This beginner playing is in keeping with the ethos of the Portsmouth Sinfonia, whose members, even if they had musical training, took on instruments with which they had little chance of success, but whose attempts made for interesting cacophony [18].

For the fourth study, a SuperCollider program was written to generate a microtonal guide track which a singer had to quickly follow; this hidden score was generated on the fly and could not be anticipated. The engagement with microtones was to further push the pitch detection within so-vits-svc, and again to explore the sort of vocals that the voice model's famous popular musicians would not typically create.

The fifth study allowed the intonation of spoken voice to drive singing voice models, deliberately clashing speech and song in a mismatch of models (so-vits-svc derives from the speech literature but is specialised for singing voice conversion). The lyrics are "Help I'm trapped inside this computer; help an AI stole my voice. Please let me out of this celebrity choir, I didn't ask to join and I'm not a celebrity" four times repeated, each time more and more rhythmically misaligned between voices. This track was concealed later in the collection as an ambiguous plea for help from a victim of vocal theft, with the tension that only celebrity voices might be true targets of mass appropriation.

Ideas for a follow-up second suite of studies have already been explored, with draft vocal transformations including:

- *Recursive application*: study 6 becomes the guide basis to commence a new round. The pitch detection finds the predominant f0, so loses much of the polyphony, which is built up again through multitracking over many models (and repeated).
- *Balloon sounds to voice*: following the work of Judy Dunaway [9]
- *Anti-celebrity*: A celebrity vocal track acts as a hidden guide for re-rendering with non-celebrity voices
- *Military industrial complex*: rocket launches and tank tracks drive vocal substitution for the choral parts
- *Requiem for a requiem*: a classical Requiem for large chorus becomes the driving signal
- *Vocal Pedagogy*: A singing voice data set [10] is used as the hidden signal

4. ETHICS

Rather than the problem of insufficient rendering capacity for polyphonic voices in a synthesizer, vocal conversion software brings a new tension to the term 'voice stealing'. Though there are great possibilities for compositional innovation, generative AI tools come with ethical issues, particularly when the data sets on which they are trained involve copyrighted audio, and the permission of the original musicians involved in the creation of that audio is not resolved [19, 20]. This is particularly acute for deceased musicians, but may also arise if a musician was recorded some time before the current age of AI, and could not have anticipated new uses to which their work might be put. However, the ethics of singing voice substitution is actually already well anticipated by the rise of sampling in the 1980s [21], the plunderphonics movement [22], 2000s mash-ups [23], and concatenative synthesis [24]. There is only an order of difference in the quality of voice models used and the facility to manipulate raw audio with recent deep learning advances. Copyright qualms have not slowed down engagement with voice conversion: it is trivial to find innumerable RVC and so-vits voice models.

Indeed, we might explore the long view where the majority of human voices are eventually recorded and modelled as a matter of course, just by interacting within society with computer agents so equipped. With a huge population of human voices, surely plentiful overlaps of accent, register and timbre must happen. There are already professional singers that it would be difficult to tell apart for non-specialists, and vocal imitation by humans can be convincing even without voice conversion technology to assist (including in imitations initiated because a professional has chosen not to release their music online and soundalikes step into the gap [25]). Perhaps it is unlikely that people will retain any sense of copyright of their own vocal timbre.

[9] See for instance `https://composersrecordingsinc.bandcamp.com/album/judy-dunaway-balloon-music`

[10] for example `https://zenodo.org/records/1442513`

If this seems like a distopian justification, there are benefits to voice substitution technology. You can sing with deceased family and friends, improve amateur voices, access a choir without leaving your house, collaborate with any professional singer even though they wouldn't normally work with you...

5. CONCLUSION

Singing voice synthesis and substitution is a growth area for computer music composition, as the technology continues to improve.

The ethics of using models trained on commercial recordings, often without the knowledge of the estate of a dead singer, is a source of legal uncertainty for future creation, but in practice has not stopped a tide of hobbyist investigation. Fair dealing defences are only easier to stranger experimental composition unlikely to attract as many views on social media as another Eminem/Kurt Cobain mash-up.

This author has made their music available for free download to avoid any sense of profiteering from those whose voices have been explored for rendering.

6. REFERENCES

[1] J. Bonada and X. Serra, "Synthesis of the singing voice by performance sampling and spectral models," *IEEE signal processing magazine*, vol. 24, no. 2, pp. 67–79, 2007.

[2] P. R. Cook, "Singing voice synthesis: History, current work, and future directions," *Computer Music Journal*, vol. 20, no. 3, pp. 38–46, 1996.

[3] Y.-P. Cho, F.-R. Yang, Y.-C. Chang, C.-T. Cheng, X.-H. Wang, and Y.-W. Liu, "A survey on recent deep learning-driven singing voice synthesis systems," in *2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 2021, pp. 319–323.

[4] M. Feffer, Z. C. Lipton, and C. Donahue, "DeepDrake ft. BTS-GAN and TayloRVC: An Exploratory Analysis of Musical Deepfakes and Hosting Platforms." in *HCMIR@ ISMIR*, 2023.

[5] Y. Zang, Y. Zhang, M. Heydari, and Z. Duan, "Singfake: Singing voice deepfake detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 156–12 160.

[6] J.-E. Choi, K. Schäfer, and S. Zmudzinski, "Introduction to Audio Deepfake Generation: Academic Insights for Non-Experts," in *Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation*, 2024, pp. 3–12.

[7] M. Conteh, "EVERYTHING THAT'S HAPPENED IN THE DRAKE-KENDRICK BEEF: A chronological rundown of one of rap's most messy tussles," January 2025. [Online]. Available: https://www.rollingstone.com/music/music-news/drake-kendrick-lamar-beef-explained-1235015540/

[8] M. Ramstrum, "Philippe Manoury's Opera K...," in *Analytical Methods of Electroacoustic Music*, M. Simoni, Ed. Routledge, 2005, pp. 239–274.

[9] K. Cascone, "The Aesthetics of Failure:"Post-Digital" Tendencies in Contemporary Computer Music," *Computer Music Journal*, vol. 24, no. 4, pp. 12–18, 2000.

[10] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 161–165.

[11] M. Morise, H. Kawahara, and H. Katayose, "Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," in *Audio Engineering Society Conference: 35th International Conference: Audio for Games*. Audio Engineering Society, 2009.

[12] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python." in *SciPy*, 2015, pp. 18–24.

[13] P. Hegarty, "Noise threshold: Merzbow and the end of natural sound," *Organised Sound*, vol. 6, no. 3, pp. 193–200, 2001.

[14] D. Kant, "The Happy Valley Band: Creative (Mis) Transcription," *Leonardo Music Journal*, vol. 26, pp. 76–78, 2016.

[15] K. Gann, *The Music of Conlon Nancarrow*. Cambridge University Press, 1995.

[16] D. Witts, "Trevor Wishart and'Vox'," *The Musical Times*, vol. 129, no. 1747, pp. 452–454, 1988.

[17] Y. Vassilandonakis and T. Wishart, "An interview with trevor wishart," *Computer Music Journal*, vol. 33, no. 2, pp. 8–23, 2009.

[18] C. Sun, "Brian Eno, Non-Musicianship, and the Experimental Tradition," in *Brian Eno: Oblique Music*, S. Albiez and D. Pattie, Eds. Bloomsbury Publishing USA, 2016, pp. 29–48.

[19] B. L. Sturm, M. Iglesias, O. Ben-Tal, M. Miron, and E. Gómez, "Artificial intelligence and music: open questions of copyright law and engineering praxis," in *Arts*, vol. 8, no. 3. MDPI, 2019, p. 115.

[20] Y. Zhang, Y. Zang, J. Shi, R. Yamamoto, T. Toda, and Z. Duan, "Svdd 2024: The inaugural singing voice deepfake detection challenge," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 782–787.

[21] P. D. Miller, *Sound unbound: sampling digital music and culture*. Mit Press, 2008.

[22] K. Holm-Hudson, "Quotation and context: sampling and John Oswald's Plunderphonics," *Leonardo Music Journal*, pp. 17–25, 1997.

[23] V. Golosker, "The transformative tribute: How mash-up music constitutes fair use of copyrights," *Hastings Comm. & Ent. LJ*, vol. 34, no. 3, pp. 380–401, 2011.

[24] B. L. Sturm, "Concatenative sound synthesis and intellectual property: An analysis of the legal issues surrounding the synthesis of novel sounds from copyright-protected work," *Journal of New Music Research*, vol. 35, no. 1, pp. 23–33, 2006.

[25] S. Parler, "Garth Brooks Soundalikes, YouTube Misinformation, and Authenticity Politics," *The Journal of Musicology*, vol. 41, no. 4, pp. 494–529, 2024.