iː goʊ weɪ

JONATHAN REUS, Intelligent Instruments Lab, University of Iceland (IS); EMUTE Lab, University of Sussex (UK)

Additional Key Words and Phrases: AI-mediated voice performance,Neural audio synthesis, Live coding, Vocal archives and dataset practice, machine learning instruments

1 Program Notes

iː goʊ weɪ (IPA for "I go away") is a semi-improvised solo performance for voice and real-time neural voice conversion, working in the tradition of Dadaist phonetic poetry and extended vocal practice. The performance departs from Kurt Schwitters' *Ursonate* - one of the most celebrated sound poems of the twentieth century - expanding its text through fragments of language assembled in dialogue with a predictive text model. As the performance progresses, the performer's voice is augmented and transformed through a system of four parallel voice conversion models, beginning with a clone of his own voice before moving through the voice of Jaap Blonk, a definitive interpreter of the Ursonate, and onward into choral, animal, and affectively raw vocal territories.

The work is not primarily concerned with transformation as spectacle. What it seeks is a narrow corridor - a state in which vocal coherence is stretched but not broken, where the visible body remains perceptually necessary yet the sounds it produces exceed what any single body could plausibly generate. Within that corridor, the singular voice becomes elastic: neither fully possessed nor fully external, but pursued. The performance moves between the primal, the beautiful, and the categorically uncertain - treating rupture, drift, and absurdity as expressive resources rather than failures.

In a media environment saturated with disembodied and algorithmically generated voices, Hugo Ball's provocation feels newly urgent: we can no longer write poetry using these languages, we must find new languages. *iː goʊ weɪ* is one attempt at that. The performer's augmented voice moves onward, dissolving and coalescing into multi-human, polyphonic, choral, schizophonic and alien forms. The interest of this performance is to celebrate the unraveling of voice as a marker of singular identity.

2 Technical Description

iː goʊ weɪ is a live, real-time neural voice conversion performance implemented in SuperCollider. The performer's microphone signal is continuously routed through four parallel custom-trained voice conversion models built on the RAVE (Realtime Audio Variational AutoEncoder) and BRAVE architectures [1, 2]. Each model maps the incoming audio into a compressed latent representation and reconstructs it in the timbral domain of a distinct target dataset, running at low latency sufficient to remain responsive to breath, articulation, and dynamic intensity.

Four active models may be pulled from a repository of models in real-time. Within the repository, different models correspond to distinct vocal territories: for example, a Self model trained on the performer's own voice; a Sound Poet model trained on recordings of Jaap Blonk; a Tutti model trained on choral material from collaborations with the University of Twente student choir; and more extreme territories such a model trained on non-human vocal and animal sounds, and a deliberately underfit model trained on vocal expressions of distress drawn from social media. The underfitting of this last model is a deliberate aesthetic and ethical decision: the model cannot faithfully reproduce its source material, instead producing a field of reconstruction noise and artefact from which voice-like sounds must be drawn. This obfuscates the identity of individual contributors while treating the model's incompleteness as a compositional resource.

The four currently active models are mapped onto the corners of a two-dimensional morphing grid, controllable via a MIDI surface or a wireless inertial sensor worn on the performer's arm. Moving within this grid enables continuous timbral interpolation between vocal identities or abrupt transitions between distinct regions of vocal character. In addition to inter-model morphing, the performer can directly manipulate the latent representation of a single active model, exaggerating instability and inhabiting areas of the vocal territory which would be otherwise inaccessible due to the timbral limitations of his own voice. Models may also be hot-swapped during performance - one of the four active models replaced in real time without interrupting signal flow - allowing new vocal territories to be entered while gestural continuity is maintained.

Author's Contact Information: Jonathan Reus, Intelligent Instruments Lab, University of Iceland (IS); EMUTE Lab, University of Sussex (UK).

NIME '26, June 23–26, 2026, London, UK

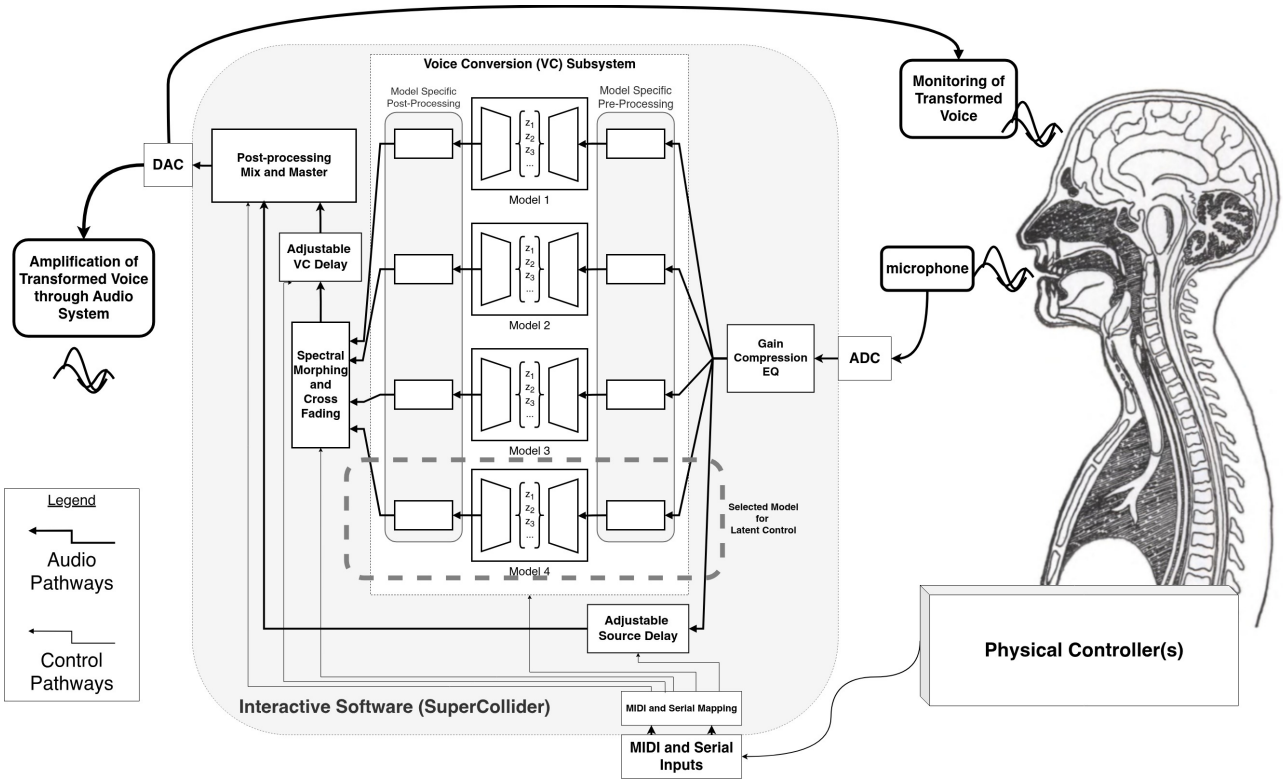

Fig. 1. System Diagram

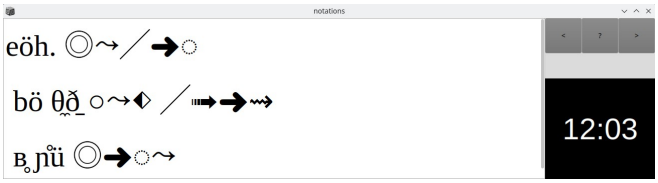

Fig. 2. Score System UniVoNo

The system supports three distinct performance configurations. In the first, the transformed voice replaces the acoustic voice entirely, decoupling audible output from visible effort. In the second, raw and transformed voices are blended simultaneously, placing the two in dialogue. In the third, an adjustable delay places the transformed voice in antiphonal relation to the acoustic voice — echoing, answering, or gradually converging with it - so that the performer appears to duet with a version of themselves. This antiphonal configuration is particularly central to the vocal territories that resemble traditional voice conversion effects, where the performer can conduct a real-time dialogue with a transformed vocal identity carrying a distinct performer’s articulatory habits and sonic character.

The system runs on a locally hosted machine. Audio input is captured via standard interface; model inference is handled using the NN.ar SuperCollider UGen [3], enabling synchronous and asynchronous operation of torchscript-encoded RAVE models at reduced latency. A software GUI provides a live visualization of morph position, model selection, and signal routing during performance [4].

3 Technical Notes

This performance requires roughly 20 minutes of set-up/sound check time assuming the sound system, microphone stand, lighting, table/stool and optional camera feed are already available.

- Best presented in a theatrical setting with dim lights and spot lighting on the performer, the presentation should focus on the performing body of the performer as a solitary presence. Ideally the audience should be arranged close enough to the performer to see his facial and bodily movements, if this is not possible, a live camera feed of the performer’s mouth projected on to a screen is preferrable.

- Requires a small table, preferrably with seating on a small stool, where the performer can set up a laptop + physical controller + audio interface (MOTU M4).
- Audio setup is simple: 2x mono balanced jack (stereo) from the MOTU M4 audio interface to the house PA.
- Optionally, but strongly preferred for larger venues where the performer is not intimately visible, is to have a live projected video feed close-up of the performer's mouth. Camera and video projection must be provided by venue.
- The performance itself lasts approx. 15 minutes and can be modified for shorter or longer times.

4 Media Links

- Project documentation: https://www.researchcatalogue.net/view/3866796/3866858

5 Ethical Standards

iː goʊ weɪ is a live solo performance using real-time neural voice conversion models trained on multiple vocal datasets assembled through distinct relational contexts. These include: (1) a dataset of the performer's own voice; (2) a dataset created in close artistic collaboration with vocal improviser Jaap Blonk, who contributed recordings and consented to their use in model training and performance; and (3) choral recordings developed in collaboration with the MUSILON student choir at the University of Twente, who provided full consent, alongside augmentation with publicly available research choir datasets. One model ("A Cry") was trained on publicly accessible social media recordings documenting acute geopolitical events [4]. These recordings were processed to prevent identification of individual speakers, and the model was intentionally underfit to prevent accurate reproduction or simulation of identifiable voices. The performance does not involve recording or collecting audience data, and no biometric identification or commercial voice cloning is conducted. This work does not involve research with animals.

Acknowledgments

This research has been supported by The Leverhulme Trust as part of the University of Sussex Interdisciplinary Doctoral Scholarship Programme *From Sensation and Perception to Awareness*. Additional financial and in-kind support for the development of *i Go Way* was provided by the Intelligent Instruments Lab, University of Iceland, supported by the European Research Council (ERC) under the Horizon 2020 research and innovation programme (Grant Agreement No. 101001848). The author declares that the research and artistic development of *iː goʊ weɪ* conducted in the absence of commercial or financial relationships that could be construed as a potential conflict of interest.

References

[1] Antoine Caillon and Philippe Esling. 2021. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. arXiv:2111.05011 (Dec. 2021). https://doi.org/10.48550/arXiv.2111.05011 arXiv:2111.05011 [cs, eess].

[2] Franco Caspe, Andrew McPherson, and Mark Sandler. 2025. Waveform Autoencoding at the Edge of Perceivable Latency. (2025).

[3] Gianluca Elia. 2023. NN.ar: nn_tilde adaptation for SuperCollider. https://github.com/elgiano/nn.ar

[4] Jonathan Reus. 2026. The Data-driven Voice-Body in Performance: AI Voices as Materials, Mediators, and Gifts. *Frontiers in Computer Science* 8 (Jan. 2026). https://doi.org/10.3389/fcomp.2026.1686763