Karaoke of Dreams: A multi-modal neural-network generated music experience

Derek Kwan[1], Sophia Sun[2], and Sofy Yuditskaya[3] ⋆

[1] Stony Brook University
[2] University of California, San Diego
[3] New York University

Abstract. Karaoke of Dreams (KoD) is a deep learning karaoke environment that generates songs and video based on user input of song titles. KoD is a multi-dimensional machine learning matrix operating in the aural dimension on harmony, melody, lyrics and style. In the musical dimension it plays 5-track pop songs generated by a Generative Adversarial Network (GAN) rendered through a Pure Data patch. In the visual dimension it operates on images tied to words using attention GANs. In the linguistic dimension it is generated by GPT-2 fine-tuned on pop song lyrics. In the spatial dimension it is a 4-fold rotational tesseract, outlining 3D space. The human element is the participant's voice and performance of the generated concert experience. KoD envelops the singer like an AI womb, a safe, warm, pulsating center allowing the singer to pour their heart out. [4]

Keywords: generative karaoke, neural network, machine creativity, interactive concert

1 Introduction

In the world of technology we see Machine Learning (ML) as the new catchall solution to social and technical problems. From directing our financial markets to doing our political organizing, from what music we consume to the healthcare we receive, so much of the fabric of our society is mediated by ML algorithms, and their influence is only likely to increase. What does it mean to forgo control of our environments to algorithms? How do we coexist, and better yet, collaborate with these impenetrable systems?

With Karaoke of Dreams (KoD), we translated these questions into a tender and intimate musical experience. KoD is an ML karaoke that generates songs and video based on user input of song titles. Karaoke, literally "empty orchestra" in Japanese, refers to singing with a sing-along machine which provides the singers with prerecorded accompaniment and vocal. We chose the karaoke as a medium because (1) karaoke is a performance of intimacy, often done in front of friends, to

⋆ The authors contributed equally to this project.
[4] Demo: https://vimeo.com/405964174

songs that are collectively known (Mitsui & Hosokawa, 1998); (2) it is an activity enabled by technology, where the machine assists and supports the human, often an amateur singer, to perform; (3) as a form of entertainment, it exists virtually everywhere in the world, and most people know how to interact with it. (Mitsui & Hosokawa, 1998) In some contemporary cultures, the karaoke box even signifies an escape from corporate and domestic norms, providing an emotional safe space while creating a sense of belonging. (Ma, 1994)

KoD creates a space where the song lyrics, music, harmonies, and stage lighting are all generated and leave nothing for the human to do but perform in a totalizing environment created by ML. With KoD, we challenged our participants by adding machines and neural networks to the mix and make them a proactive creator of the music being performed. How did we build rapport with the machine?

2 Related Works

Artists have long worked with algorithms, sometimes as medium, sometimes as subject, sometimes both. In the US perhaps the most famous algorithmic artist working before 1960 is Sol Lewitt, whose artworks are sets of instructions intended to create drawings or structures, as such he is still an artist producing new works today, despite having died in 2007. (Susan Cross (Editor), 2009) Prof. Max Bense was an early algorithmic artist using computers in the late 1950's in Stuttgart. He argued in this televised conversation with mystic artist Joseph Beuys that algorithmic art is devoid of emotional appeal and therefore cannot be abused for political oppression. (Medien-Archiv, 1970) Post-1970 artists like Frieder Nake became heavily invested in systems design that produce massive image sets. Frieder Nake and others such as Harold Cohen are invested in systems design more than achieving a specific output from those systems, tying AI art back to conceptual art. (Offert, 2019)

Since 2015, deep neural networks (NNs) have shown a different potential in computational creativity. Powered by neural networks, computers have composed music that sounds "more Bach than Bach" (Hadjeres, Pachet, & Nielsen, 2016), written screenplays (Sharp, 2016), and powered a mirror which reflects the world back in the Cubist painting style (Karras, Laine, & Aila, 2019)(Kogan, 2017). The terminology of describing the process of deep generative NNs producing samples as "dreaming" originated from Google Research's DeepDream project (Szegedy et al., 2015), where a network trained on pictures of objects and animals enhances patterns it finds in user-uploaded images. The processed result has a dream-like hallucinogenic appearance, and the authors used biology-inspired words such as *algorithmic pareidolia* to explain its function. Later artworks with similar aesthetics, especially the ones using Generative Adversarial Networks (GANs), have kept the terminology.

Generative models have also been studied in the domain of music creation. One approach is to generate music symbolically in the form of a piano roll or MIDI file (Huang et al., 2019)(Dong, Hsiao, Yang, & Yang, 2018)(Genchel, Pati,

& Lerch, 2019), continuing a long history of algorithmic music composition from the ancient Greeks up through John Cage (Maurer, 1999). Recently there has been success in generating raw audio directly by leveraging the expressiveness of deep neural networks. (Dhariwal et al., 2020)(Engel et al., 2019). In application, these systems has resulted in applications to help people learn improvisation (Johnson, Keller, & Weintraut, 2017), create whole albums (Yacht, 2019), and make creativity-augmenting DAW plugins (Roberts et al., 2019).

3 KoD Content Generation

The generation of the karaoke songs consists of three parts: lyrics, music, and the music video. The generation pipeline is modular, where both the music and the video condition on the generated lyrics. For each generation module, we used a fine-tuned open-source neural network model. The rationale behind the design is (1) for the ability to replace modules as better models become available as research progresses, and (2) to be considerate in the computing power we use for environmental and accessibility reasons. All the processing, including fine-tuning models and generating the 10 songs in our song bank, was done on one Nvidia 1080 Ti Graphic Processing Unit, totalling to around 6 hours. [5]

The installation was created for and by our community; we requested submissions of original song titles from the showcase participants. We asked them to devise titles that speak to them and used all the submissions. We then generated the lyrics, accompanying music, and music videos to these song titles.

3.1 Lyrics

OpenAI's GPT-2 (Radford et al., 2019) is a large-scale transformer-based language model pretrained on a large corpora scraped from the internet. GPT-2 has provided state-of-the-art results for many text-generation tasks, from powering chatbots (Budzianowski & Vulic, 2019) to writing patent claims (Lee & Hsiang, 2019). OpenAI, noting the model's power, delayed the release of its largest model (1.5B) to avoid malicious use. In our project we used the `SMALL-124M` model for fine-tuning, as we found it to be sufficiently expressive. Our parameters for "sufficient expressivity" were that the resultant lyrics were able to capture the diverse vocabulary of song lyrics and had a structure approximating verse-chorus form.

The model is fine-tuned with the `lyricsfreak` dataset (Mohdazfar, 2018), consisting of lyrics to 57,650 pop songs crowdsourced on `lyricsfreak.com`, with 500 epochs of training. To generate new lyrics, the model is prompted with the song title and performs a greedy search of 100 tokens. The generated text manifests lyrical properties such as short sentences, repetitive song structures, and poetic language pertaining to sentiments common in popular music.

[5] Code for content generation is open-sourced under the MIT licence at `https://github.com/suisuis/karaoke`

```
But there's nothing I can do
      To change you.
    I can't stop you.
 You and I are the world,
    We're the world,
   The world is ours.

                Your body,
                Your mind.

Innocent man, Innocent man
Innocent man, Innocent man
```

Fig. 1: Generated lyrics excerpt for song titled *Dust of Stars, Surf the Universe.*

3.2 Music

We use a pre-trained MuseGAN model (Dong et al., 2018) to generate the accompanying music for the karaoke. MuseGAN is a convolutional generative adversarial network (GAN) trained on a subset of the Lakh piano roll dataset (Raffel, 2016), consisting of 21,425 multi-track MIDI files. A GAN is a deep neural network that captures the statistical distributions of its training data which is able to sample from random noise as input. In some sense, the network *hallucinates* the result from that sample of noise; we find it to be most befitting to the *karaoke of dreams* concept. MuseGAN generates 4-bar-long 5-instrument piano rolls; the five instruments are bass, drums, guitar, strings, and piano.

To make the accompaniment for a song, each 4-bar phrase (we refer to them as *chunks*) is matched to a line in the lyrics. The chunks are then transposed to the same key and put together to form a complete piece. The MIDI file is sent to a Pd patch, see section 4.1. The specific melodies that maps to the lyric lines are left for the karaoke singer to come up as they hear the accompaniment.

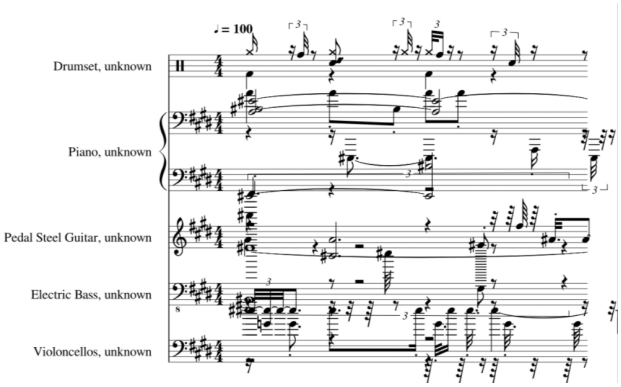

Fig. 2: A generated chunk.

3.3 Video

We looked to recent works in text-conditioned image generation for generating the karaoke's videos. Much like (Frosst & Kereliuk, 2019), we made use of At-

tnGAN (Xu et al., 2018), generating images that condition on the lyrics of a song one line at a time. We then interpolated between the images using Adobe's morphing algorithm to create a smoother transition. We used the model trained on the MSCOCO dataset (Lin et al., 2014) which given specific descriptions of common objects will generate realistic images. In our case, the images are often abstract yet, to various extents, resemble what is written in the lyrics to the human eye. Figure 3 shows three examples with the strongest correlation from word to image.

(a) *Desert de Bebe* (b) *My body is a battleground* (c) *seagulls over home*

Fig. 3: AttnGAN generated images and their corresponding lyrics.

4 System Overview

After the machine has generated the karaoke's content, the question that follows is how we should materialize and showcase our work. In this section we introduce the system setup of the art installation, illustrated in figure 4.

4.1 MIDI Playback

The MIDI files generated by MuseGAN were loaded into the Pd programming environment using the external library Cyclone's [seq] object. (Czaja et al., 2020) [seq]'s output is interpreted within the Pd patch by frequency-modulation synths with runtime-randomized parameters for the pitch material and subtractive-synthesis drum synths for percussion.

4.2 Raspberry Pi

The principal hardware used for the installation was a Raspberry Pi 3B+. This single-board computer hosted an ad-hoc wireless network with iPad and visuals-running laptop as clients. On the Pi, Pd and Node.js communicated via Open Sound Control (OSC) with Pd serving as the source of timing information.

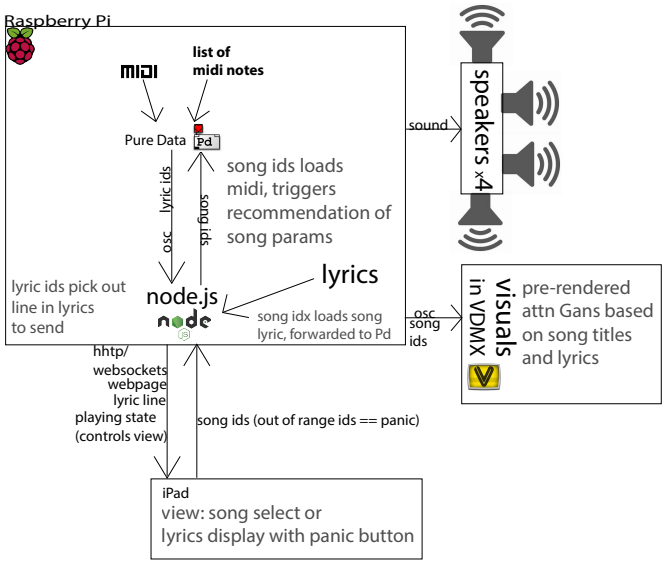

Fig. 4: I/O diagram for the KoD's system setup

Node.js hosted a website that served as the iPad's graphical interface and bidirectional communication between the two was facilitated by WebSockets (Node.js's socket.io library). Node.js served as the communications conduit between the iPad and Pd with the iPad sending song choice by index and Pd sending timed lyrics lines by index. Pd also sent OSC messages containing song index information to the laptop through their WiFi connection. [6]

4.3 Hardware

The Pi output Pd's audio via 4 mini speakers. The microphone was a 2BOOM Wireless Karaoke Microphone with Speaker, used for comedic effect and vocals processing. An iPad Mini served as the visual interface for the installation.

Pd sent OSC messages to a second computer installed a distance away running the AttnGANs. Each time a new song was selected, its associated video is pulled up by VDMX and projected onto the KoD. VDMX allows for projection mapping and real-time pixel manipulation. To further tighten the relationship between our visual and aural experiences, we had the overall color schema of each video determined by the overall pitch produced by the singer(s) in KoD using VDMX's built-in Audio Analysis tool.

The AttnGAN acted as contextual lights for the performers in KoD. From the inside of the projection, visuals do not resolve themselves into identifiable

[6] The Pd and JavaScript code for KoD can be found at `https://github.com/derekxkwan/karaokeofdreams-engine`.

images but do provide a responsive light environment to indicate that a song has begun. From the outside, viewers can see a video illustrating the song going on within and perhaps be drawn in by the projection. The projection acted as atmospheric lighting for the performers within. Inside KoD, we placed red 5-volt lasers at the intersection of each tesseract node. These lasers pointed at cube prisms on motors, further expanding and diffracting the light through hypercubes.

4.4 Interaction Design

KoD is set up like a shade structure in a sculpture park in a desert environment, inviting visitors to walk inside for shade if for no other reason. It is covered in mosquito netting which serves simultaneously as a projection surface, shade cloth, and a means of keeping bugs out of the karaoke. Visitors can step inside through any one of the eight openings of the rotated hypercube. Once inside, there is a miniature version of the structure they are inside of suspended in the center. This small structure houses the screen, speakers, and microphone karaoke users are all so familiar with. The lasers twinkle gently above reflecting on the fabric within, and they have but to choose a song from the glowing list and start singing.

Traditional karaokes are designed to be forgiving and supportive of the amateur singer, with their simplified MIDI versions of popular tunes playing in highly compressed audio backing the singing. Our karaoke was no different and despite the generated MIDI files, effects, and lyrics, the resulting format was familiar for the singer.

Some testimonials of participants include:

Arseniy Klempner: Usually in karaoke the rhythm of the lyrics are fixed to the song, but in this case it was on the participant to spontaneously interpret how the words would flow best. This felt much more creative and engaging than trying to match the rhythm of an existing song, as in traditional karaoke, where the machine's role is more of a lenient taskmaster; like a production-line for reproducing the same song, albeit with slight variations introduced by the singer. The KoD did not challenge me to reproduce, but to interpret and create.

At first, the music felt amusing and strange. It was obvious that the notes and lyrics were not written by a human. But as I spent more time there...I adjusted to the strangeness; it faded to the background, and the more human-like lyrics and melodies, the ones that felt charged with emotion and creative cohesiveness, stood out amidst the mechanical nature of the songs.

Jonas Johansen: It did feel silly, unexpected - fun, also a bit of uneasiness since you don't know what comes next and maybe you're not feeling the song but perhaps it'll get good again? Sometimes u get stuck in a loop but remain curious. It's more fun to sing a song that u don't know - i think - since there's no comparison of anyone who have sung it before.

Phil Stearns: It was so freeing to know for sure that no interpretation could possibly be construed as wrong. I humbly offer that the project opened the doors to a whole new kind of authenticity within that sacred discipline of Karaoke.

Fig. 5: Participants singing in the karaoke.

5 Future Work

The human machine reckoning is upon us! KoD is a playful locus for participants to perform this reckoning.

Since the completion of this work, there have been advances in state-of-the art models for all our generation domains - a brand new and more powerful GPT-3 (Brown et al., 2020), for example, and OpenAI's Jukebox (Dhariwal et al., 2020) that can generate audio directly from genre, artist, and lyrical primers. In future renditions of the karaoke, we hope to incorporate these advances. Future iterations of KoD could include wiring up the lasers and light fixtures to generative machine learning input and live vocal effects.

Due to network issues at the site, we were not able to generate contents on request. To make the experience more engaging, a pipeline can be built so that the song title, lyrics, music, and visuals can be generated in real-time immediately upon participant input. In this way, the machine and the human may dream with one another, weaving together an intimate and ephemeral musical experience.

Acknowledgements

KoD was built as part of the authors' residency at `Brahman.ai`. We thank Gene Kogan and Freeman Murray for organizing the residency and providing resources. We thank the local community for participating in and contributing to our artwork. And finally, we thank M. Schedel for valuable discussions and suggestions.

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *CoRR*, *abs/2005.14165*.

Budzianowski, P., & Vulic, I. (2019). Hello, it's GPT-2 - how can I help you? towards the use of pretrained language models for task-oriented dialogue systems. *CoRR*, *abs/1907.05774*.

Czaja, K., Steiner, H.-C., Kraan, F. J., Porres, A., Kwan, D., & Barber, M. (2020). *pd-cyclone.* `https://github.com/porres/pd-cyclone`. GitHub.

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., & Sutskever, I. (2020). Jukebox: A generative model for music. *CoRR*, *abs/2005.00341*.

Dong, H.-W., Hsiao, W.-Y., Yang, L.-C., & Yang, Y.-H. (2018). Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment..

Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., & Roberts, A. (2019). Gansynth: Adversarial neural audio synthesis..

Frosst, N., & Kereliuk, J. (2019). Text conditional lyric video generation. In (chap. Machine Learning for Creativity and Design Workshop).

Genchel, B., Pati, A., & Lerch, A. (2019). Explicitly conditioned melody generation: A case study with interdependent rnns. In (Vol. abs/1907.05208).

Hadjeres, G., Pachet, F., & Nielsen, F. (2016). Deepbach: a steerable model for bach chorales generation. *arXiv/1612.01010*.

Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., ... Eck, D. (2019). Music transformer: Generating music with long-term structure..

Johnson, D. D., Keller, R. M., & Weintraut, N. (2017). Learning to create jazz melodies using a product of experts. In A. K. Goel, A. Jordanous, & A. Pease (Eds.), *Proceedings of the eighth international conference on computational creativity, atlanta, georgia, usa, june 19-23, 2017* (pp. 151–158). Association for Computational Creativity (ACC).

Karras, T., Laine, S., & Aila, T. (2019, Jun). A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kogan, G. (2017). *Cubist mirror.* Retrieved from `https://vimeo.com/167910860`

Lee, J., & Hsiang, J. (2019). Patent claim generation by fine-tuning openai GPT-2. *CoRR*, *abs/1907.02052*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In (p. 740–755). Springer.

Ma, R. (1994). Karaoke and interpersonal communication in east asia..

Maurer, J. (1999). A brief history of algorithmic composition. *Stanford CCRMA*.

Medien-Archiv, J. B. (1970). *Provokation - lebenselement der gesellschaft.* Retrieved from `https://archive.org/details/PodiumsdiskussionJosephBeyusArnoldGehlenAntiKunst`

Mitsui, T., & Hosokawa, S. (1998). *Karaoke around the world: Global technology, local singing.* (Vol. 21). London and New York: Routledge.

Mohdazfar. (2018). *Song lyrics dataset.* Retrieved from `https://www.kaggle.com/mousehead/songlyrics`

Offert, F. (2019). The past, present, and future of ai art. *The Gradient*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Raffel, C. (2016). Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching. *PhD Thesis*.

Roberts, A., Engel, J., Mann, Y., Gillick, J., Kayacik, C., Nørly, S., ... Eck, D. (2019). Magenta studio: Augmenting creativity with deep learning in ableton live. In *Proceedings of the international workshop on musical metacreation (mume).*

Sharp, O. (2016). *Sunspring.* Therefore Films.

Susan Cross (Editor), D. M. E. (2009). *Sol lewitt: 100 views.* Yale University Press.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi: 10.1109/cvpr.2015.7298594

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *2018 ieee/cvf conference on computer vision and pattern recognition* (p. 1316-1324).

Yacht. (2019). *Chain tripping.* album.