

Token Telephone

HUGO FLORES GARCÍA, Northwestern University, USA

STEPHAN MOORE, Northwestern University, USA

Additional Key Words and Phrases: interactive sound installation, neural networks, generative ai, spatial sound

ACM Reference Format:

Hugo Flores García and Stephan Moore. 2024. Token Telephone. 1, 1 (September 2024), 4 pages.

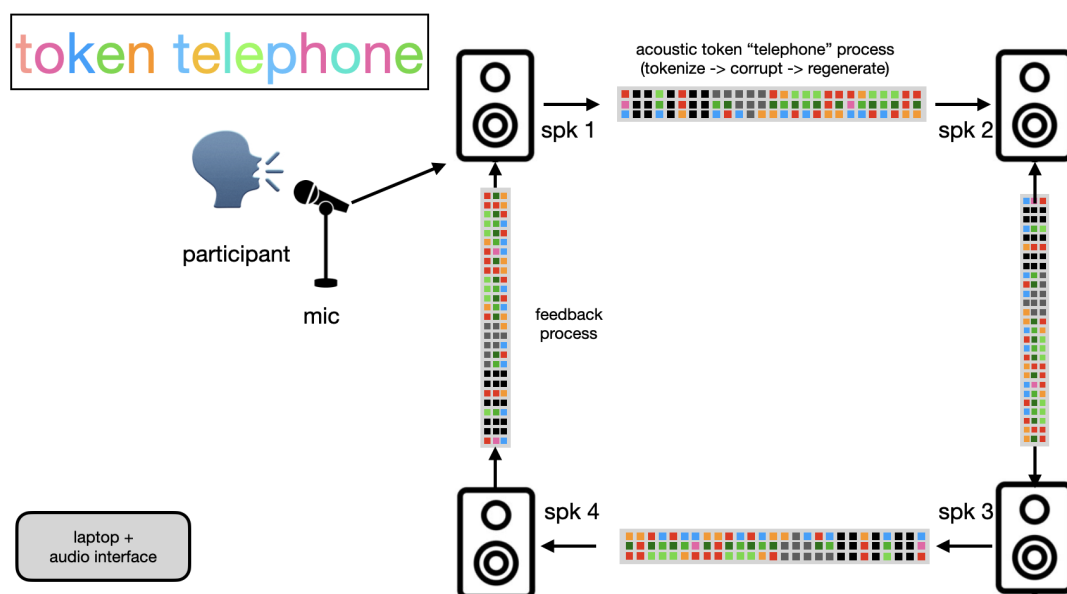


Fig. 1. Token Telephone is a co-creative AI sound installation where participants interact with a chain of generative AI models, initiating a generative game of telephone. The installation space is circled by four neural networks, each represented by a loudspeaker. Participants make sounds into a microphone at the entrance of the installation space. Their sounds are iteratively transformed by each neural network in a feedback loop, deviating further from the original with every pass. This iterative process reveals patterns between the input and the training data of the networks, morphing human utterances into new and unexpected sound textures.

Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). Copyright remains with the author(s).

1 PROGRAM NOTES

Token Telephone is a co-creative AI sound installation. Participants enter a space equipped with a microphone and a quartet of generative sound neural networks, each represented by a loudspeaker. Upon vocalizing into the microphone, the participants' utterance is transformed into neural acoustic tokens and played back, initiating a game of telephone between the neural networks. Each network encodes, processes and reconstructs the sound, distorting the original utterance into new textures guided by the network's training data. The newly reconfigured sound is then passed to the next network/loudspeaker in a clockwise direction, and the process repeats. The sound produced by the fourth network is passed back to the first network in the cycle, creating a feedback loop wherein the original utterance incrementally loses all of its original characteristics and disintegrates into textures that reflect the inherent biases of the generative models in play. In time, the resonant properties of the processes are revealed in front of the participant. Inspired by the popular children's game of telephone, Token Telephone illuminates the gradual formation of hallucinations through the iterative processing and re-processing of audio, reflecting the biases introduced by the model's understanding of sound objects, as well as the data that was provided to it.

2 MEDIA LINKS

To listen to a stereo demonstration of token telephone, visit the following YouTube link:

- video: <https://youtu.be/vEaYoEgtSUo>

3 PROJECT DESCRIPTION

3.1 Motivation

Telephone is a popular children's game in which children try to communicate a message through a noisy information channel. Humans are lossy information machines, and they do not store the utterance they hear as a raw audio signal in their brains but rather a compressed representation that contains a mix of semantic (what was said) and acoustic (what that sounded like) information.

When we repeat a spoken utterance from memory, we are forced to rebuild an acoustic signal from the lossy representation stored in our memory. This means that we may hallucinate words that were not there, transforming the meaning of the original message we meant to pass along.

Unlike the traditional game, this installation employs neural networks instead of humans to encode, distort, and regenerate sound. It illustrates the fascinating and often unpredictable ways AI interprets and manipulates patterns in the input data to generate new sounds.

Audiences engaging with Token Telephone will be able to hear, iteratively and in real time, the formation of hallucinatory audio information. The sounds are themselves compelling as they imitate and amplify the rhythms and nuances of our vocalizations. But beyond the aesthetic interest of its output, this installation provides a rare opportunity to hear generative neural network at work. As AI moves into our daily lives, there may be some value in understanding how the biases inherent in the data sets we use for training these systems can influence their output.

3.2 Interaction and Underlying Process

Refer to Figure 1 for a diagram of the installation layout. The installation is quadraphonic, with four speakers placed in a ring around the room, with a microphone at the entrance of the installation space. Each speaker in the room embodies

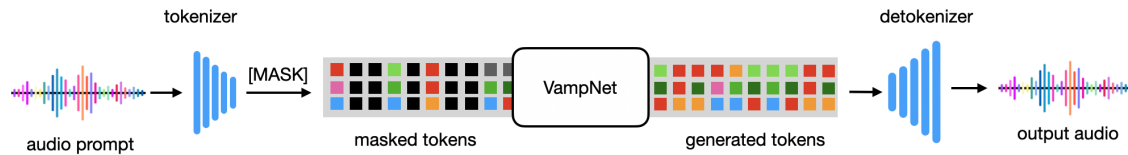


Fig. 2. Illustration of the tokenization -> corruption (masking) -> generation process occurring at each speaker in the installation.

a co-creative generative sound agent capable of receiving a sound, encoding it, and reconstructing it from its lossy neural acoustic token representation.

Upon entering the installation space, the participant is greeted with a microphone to utter any sound they like. Shortly after, the participant's utterance starts playing in a loop on the speaker nearest to the them, which sets off the game of telephone.

When the player's utterance enters the telephone chain, it is encoded into a sequence of neural acoustic tokens [1, 2]. These tokens are a compressed encoding of the audio signal, and they are used by generative sound model (VampNet) to reconstruct the encoded sounds. These generative acoustic tokens are organized hierarchically, where "coarse" tokens loosely encode higher-level information about a signal like its rhythmic structure, while "fine" tokens may represent high frequency details and other characteristics that further define a sound event.

While the tokens are being passed around the chain, a percentage of these tokens are "corrupted" (i.e. masked), meaning that the generative model will have to infer and fill in the missing spots, resorting to the model of its training data to reconstruct the missing tokens. As the uttered sound undergoes many passes through the token telephone loop, the sound retains some of its original characteristics and rhythm, but the sound identities are incrementally transformed from human speech to the sounds and patterns present in model's training data.

This resulting audio transformation makes the installation feel like a voice-controlled interface for musical expression, where the sonic gestures made by the participant are preserved by the generative algorithm, but the actual contents and perceptual identity of the sound are given a special timbre, texture and color, originating from the neural network's own distribution of sounds.

The generative model underlying in this process is VampNet [1], a generative model capable of generating variations on an input signal through this tokenize → corrupt → generate process.

Participants can listen to their utterance undergo this process through models trained on different sound libraries, resulting in different "flavors" of generated sound (e.g. operatic textures, choral textures, natural sounds).

4 PERFORMANCE NOTES

4.1 Space Requirements and Suitable Venues

This piece can be installed in any small room or medium hallway (e.g., minimum 14 x 14 ft), as long as it comfortably fits a 4-speaker array, a mic + mic stand positioned outside of the ring of speakers, a small table to house the computer and audio interface (outside the speakers or between two speakers), and a minimum of 4-5 people walking around the area enclosed by the speakers.

We imagine that this installation would be suitable for exhibition in a number of situations, and therefore in several of the mentioned venues. It may be most appropriate in the Academy Gallery, but it might receive more interaction if it were placed in a corner of one of the open spaces in the main performance venue. The HKU building's corridors could

also be used if they are sufficiently wide. We can imagine the outdoor venue and the train station being possibilities, as long as concerns about security and shelter from weather could be addressed.

4.2 Network Requirements

Because the installation requires a stable internet connection to a remote server for neural network processing, we require that the location of the installation have good network coverage and an available internet connection capable of uploading/downloading 10-second audio files reliably (e.g. at least 50Mbps of download and 30Mbps of upload speed).

4.3 Equipment

The installation requires the following equipment. **Equipment to be provided by the conference is highlighted in yellow.**

- **1x dynamic vocal microphone (e.g. Shure SM58)**
- **1x microphone stand.**
- **1x XLR cable (25ft, or long enough to reach from the mic to the equipment table comfortably in the installation space).**
- **4x Speakers, large enough for the installation space, with appropriate cabling for each speaker back to the equipment table. If speakers are not active, amplifiers for the speakers are required as well.**
- **1x small table for housing the computer, electronics.**
- **a reliable internet connection for processing incoming sound.**
- 1x 4+ channel audio interface (e.g. Scarlett 4i4, PreSonus Studio1824c)
- 1x Mac OSX computer capable of running Max/MSP 8.6 or later.

5 ETHICAL STATEMENT

AI models are ubiquitous in different aspects of our modern society. Often, AI models are obscurely present in the digital services we use daily (e.g. web search rankings, music recommendations), and their biases can influence our decision processes without us even knowing it. One of the aims of this installation is to illustrate the strengths and flaws of the pattern recognition and signal synthesis capabilities of generative AI systems, as being familiar with the properties and processes behind these AI models is becoming more and more important. Additionally, generative models have gathered much controversy as large for-profit AI companies will train their generative music models on large collections of copyrighted music without an artist's consent, and provide techno-solutionist arguments along the lines of "generative AI will democratize music". We believe that this is an ill-formed goal, as music is not a homogeneous blob, but an umbrella term encompassing countless evolving communities of artistic practice, each with a unique set of styles, techniques, and aesthetic values [3].

REFERENCES

- [1] Hugo Flores García, Prem Seetharaman, Rithesh Kumar, and Bryan Pardo. 2023. VampNet: Music Generation via Masked Acoustic Token Modeling. *ISMIR (2023)*.
- [2] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-Fidelity Audio Compression with Improved RVQGAN. *arXiv:cs.SD/2306.06546*
- [3] Andrew McPherson, Fabio Morreale, and Jacob Harrison. 2019. Musical instruments for novices: Comparing NIME, HCI and crowdfunding approaches. *New directions in music and human-computer interaction (2019)*, 179–212.