Algorithm Based GEN Video to Partial GEN Audio Synthesis System

Xiaosha Li
Berklee College of Music
xli5@berklee.edu

Neil Leonard
Berklee College of Music
nleonard@berklee.edu

ABSTRACT

This paper presents a novel Gen-to-Gen synthesis system that converts generative video signals into multichannel audio through six interconnected modules. The low-level architecture processes video pixel streams to generate partial audio synthesis, while higher-level modules map RGB channels from five video sources to control a 32-channel amplifier with user-defined sound shapes. The system features MIDI and computer keyboard integration for real-time interaction, employing parametric controls including contrast, spatial transformations, and spectral mapping. Experimental results demonstrate the system's capability to produce complex spatial audio textures through nonlinear video-to-audio conversion, offering new possibilities for audiovisual performance systems.

I. INTRODUCTION

Real-time audiovisual synthesis is expanding artistic and technological possibilities, particularly in interactive performance and multimedia art. This paper introduces a novel system that bridges generative video and audio synthesis through algorithmic processing, converting visual patterns into multichannel audio in real time. Implemented in *Max/MSP* [1], the system features a modular video-to-audio synthesis architecture. The system employs a two-tiered approach. At the lower level, it extracts pixel data from a 256x256 video frame, segments it into 32 normalized values, and cycles these at 512 Hz to generate audio signals. At a higher level, five video playbacks, with RGB channel mapping, control a 32-channel audio amplifier, enabling spatial audio synthesis through 15 predefined sound shapes. Designed for interactivity, the system integrates MIDI and computer keyboard controls for real-time adjustments of contrast, brightness, and speed. Its modular design in *Max/MSP* [1] ensures flexibility for diverse audiovisual applications.

II. SYSTEM ARCHITECTURE

The proposed system implements two hierarchical synthesis layers through six modular components. The lower layer converts real-time video pixel data into audio signals via buffer-based analysis, while the upper layer employs RGB channel decomposition for spatial sound shaping.

II-A Core Synthesis Pipeline

$$A_{out}[n] = \sum_{k=1}^{32} \left(\frac{1}{512} \sum_{i=16(k-1)}^{16k} V_{frame}[i] \right) \cdot W_k[n] \quad (1)$$

where V_{frame} represents normalized pixel values and W_k window functions for partial extraction.

III. MODULE SPECIFICATIONS

III-A Video Generation Modules

Module 1 implements a generative video algorithm that dynamically creates visual patterns based on pixel position and parametric controls. The module generates intricate shapes and textures by applying mathematical transformations to pixel coordinates, resulting in a continuously evolving visual output. The following parameters are available for real-time control:

- **Contrast Level** (-25.0 to 25.0): Adjusts the contrast between black and white pixels, enhancing or reducing the visual intensity of the generated patterns.
- **X/Y Zoom** (0.1-4.0x): Controls the scaling of the visual output along the X and Y axes, allowing for dynamic resizing of the generated shapes.
- **Brightness** (0.0 to 2.0): Modifies the overall luminance of the video output, enabling fine-tuning of the visual appearance.

Module 2 generates rotating square geometries with the following controls:

- **Alignment**: Adjusts the alignment of the square geometries, ranging from 0.0 to 1.0.
- **Rotation**: Controls the rotation angle of the squares, ranging from 0° to 360°.
- **Speed Synchronization**: Synchronizes the rotation speed with the master clock for consistent timing across modules.

III-B Video Playback and Control Module

Module 3 implements video playback control and background noise generation, serving as the central hub for video source management and synchronization. The module features:

- Dual video stream processing with crossfade control
- Background noise generation with spectral characteristics matching the visual texture
- MIDI input routing for real-time parameter control

- Global speed synchronization across all video sources

III-C Audio Synthesis Modules

Module 4 is the core of the system, responsible for converting video pixel data into partial audio synthesis. The module processes the combined video signal through the following steps:

1) **Frame Segmentation**: The 256x256 video frame is divided into 32 vertical segments, each representing a partial.
2) **Normalization and Buffering**: Segments are normalized (0–1) and stored in a 512-sample buffer, cycled based on global speed.
3) **Waveform Generation**: Pixel values define partial amplitudes, forming a custom oscillator waveform looped in the buffer.
4) **Frequency Control**: The vertical zoom (Zoom Y) adjusts the base frequency, shaping the spectral characteristics.

Control Parameters:

- **Volume**: Adjusts overall output.
- **Brightness-to-Amplitude**: Links video brightness to audio intensity.
- **Zoom Y to Frequency**: Modulates base frequency via vertical zoom.
- **Global Speed**: Controls buffer cycling speed, affecting tempo.

Module 5 enhances audio synthesis by mapping RGB channels from five video sources to 32 output channels, creating dynamic spatial sound.

1) **Video-Controlled Audio Synthesis**: Five video sources correspond to five stored sound patterns, influencing the generated audio.
2) **RGB Processing**: Each video's RGB channels are processed independently to control audio movement across 32 channels.
3) **Channel Mapping**: The audio amplitude for each channel is shaped by the weighted sum of its RGB values:

$$C_{ch}[n] = \frac{R[n] + G[n] + B[n]}{3} \cdot M_{shape}[k] \qquad (2)$$

where:

- $R[n], G[n], B[n]$ represent the normalized RGB values of the pixel output for the n^{th} channel.
- $M_{shape}[k]$ represents one of 15 predefined spectral envelopes, which are used to shape the final audio output.
- $C_{ch}[n]$ is the amplitude of the n^{th} audio channel.

4) **Spatial Audio Control**: The RGB values dynamically shape spatial distribution, generating immersive soundscapes responsive to visual content.

Control Parameters:

- **Volume**: Adjusts overall output.
- **Modulation**: Alters amplitude and frequency for dynamic textures.
- **Global Speed**: Modifies video playback speed, influencing tempo and rhythm.

III-D RGB Channel Downscaling

Module 5.5 implements a downscaling process to map the RGB channels of the video signals to 32 audio output channels. The downscaling algorithm operates as follows:

$$A_{ch}[n] = \frac{1}{10} \sum_{i=10(k-1)}^{10k} (R[i] + G[i] + B[i]) \qquad (3)$$

where:

- $R[i], G[i], B[i]$ represent the RGB channel values of the video signal.
- $A_{ch}[n]$ is the amplitude of the n^{th} audio channel.
- The downscaling factor is 10:1, reducing 320 values to 32 control points.

III-E Global Playback Control

Module 6 provides centralized control over the entire system's video playback and rendering. It combines all video signals through a series of operations, including blending, masking, and spatial transformations. The module features the following control parameters:

- **Fullscreen On/Off**: Toggles fullscreen rendering mode for video output.
- **Visible On/Off**: Controls the visibility of individual video layers.
- **Blend Mode**: Selects between additive, subtractive, and multiplicative blending.
- **Global Opacity**: Adjusts the transparency of combined video signals (0-100%).

IV. CONTROL IMPLEMENTATION

The system integrates MIDI and computer keyboards for real-time interaction and parameter control, enabling dynamic adjustments during performances. The **MIDI keyboard** modulates key parameters, including contrast, brightness, speed, modulation depth, volume, and noise selection. It also supports zoom adjustments and MIDI-triggered events for audiovisual synchronization. The **computer keyboard** toggles module states, playback, and geometric transformations. It enables switching video-to-audio mappings and controlling rotation and movement in various modules. This dual-control scheme balances real-time modulation via MIDI with discrete toggling via the computer keyboard, offering an intuitive and flexible interface for live performance.

V. CONCLUSIONS

This work demonstrates effective video-to-audio synthesis through parametric mapping of visual features to spectral components. Future work will explore machine learning-based mapping strategies and expanded gesture control.

VI. REFERENCES

[1] Cycling '74. *Max/MSP*. [Online]. Available: https://cycling74.com/.