International Journal of Web of Multidisciplinary Studies
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
website: *https://ijwos.com*
Volume 03, Issue 11,November 2026
OPEN ACCESS
E-ISSN : 3049-2424
DOI: https://doi.org/10.71366/ijwos

Listening to Uncertainty: A Fuzzy Logic Approach to Visual-to-Audio Generation

MAKIREDDI DURGA[1], MAMIDI REVATHI KALYANI[2], MANABARTHI YAMUNA[3], UPPE NANAJI[4]
[1],[2],[3] Master of Computer Applications Final Year Students, Department of CSE, Avanthi Institute of Engineering & Technology, Anakapalli, Andhra Pradesh, India
[4] Associate Professor, Department of CSE, Avanthi Institute of Engineering & Technology, Anakapalli, Andhra Pradesh, India

Article Info

Article History:
Published: 04 Nov 2026

Publication Issue:
Volume 03, Issue 11
November 2026

Page Number:
17-22

Corresponding Author:
MAKIREDDI DURGA

Abstract:
Cross-modal synthesis from silent video to audio is challenging because visual events do not determine a unique sound and the relationship between visual and acoustic representations is uncertain. This paper presents a Neuro-Fuzzy Visual-to-Audio (NF-V2A) framework that combines a pre-trained 3D Convolutional Neural Network (3D-CNN) for spatial-temporal visual feature extraction, an adaptive Takagi-Sugeno-Kang (TSK) Fuzzy Inference System (FIS) for uncertainty-aware mapping, and a generative adversarial network (GAN) decoder for Mel-spectrogram synthesis. A Genetic Algorithm (GA) optimizes fuzzy membership parameters and rule weights using a fitness objective based on spectrogram Mean Squared Error (MSE) and Fréchet Audio Distance (FAD). Experiments are reported on VGG-Sound and a curated AudioSet subset. On the VGG-Sound test set, the proposed model reports FAD of 3.04, Inception Score of 7.41, KL divergence of 1.12, and spectrogram MSE of 0.043. Compared with the Diff-V2A baseline, the reported FAD reduction is approximately 14.3%, while spectrogram MSE is reduced by approximately 15.6%. The results suggest that combining fuzzy inference and evolutionary optimization can improve the reported perceptual and reconstruction measures for visual-to-audio generation.

Keywords: Visual-to-audio generation, neuro-fuzzy system, fuzzy inference system, genetic algorithm, multimodal learning, Mel-spectrogram, cross-modal synthesis.

1. Introduction

The integration of visual and auditory information is central to human perception. When a person observes an event such as a glass breaking or a dog barking, visual context can suggest the sound likely to accompany that event. Visual-to-Audio (V2A) generation seeks to reproduce this capability by synthesizing audio that is contextually and temporally aligned with a silent video stream. Potential applications include automated Foley generation for film production, immersive virtual-reality environments, and assistive technologies for people with visual impairments.

The mapping from video to audio is inherently ambiguous. A visible action may produce different sounds depending on material, force, distance, and environmental acoustics. Visual information is spatially dense and often changes more slowly, whereas audio has high temporal variability. Consequently, strict deterministic visual-to-audio mappings may be sensitive to occlusion, noise, and imperfect temporal alignment.

This paper presents the Neuro-Fuzzy Visual-to-Audio (NF-V2A) model. It combines deep spatial-temporal feature extraction with a TSK fuzzy inference mechanism that maps visual features into an acoustic latent space. A Genetic Algorithm optimizes the fuzzy membership functions and rule parameters, while a GAN-based decoder generates the Mel-spectrogram representation of the output audio.

A. Research Contributions

Copyright © 2026 The Author(s) : This is an open access article under the CC BY license (http://creativecommons.org/licenses/by-nc/4.0/)

The principal contributions described in this work are: (1) a hybrid architecture that integrates a 3D-CNN with an adaptive FIS for cross-modal generation; (2) uncertainty management through fuzzy membership functions and rule-based inference; (3) evolutionary optimization of fuzzy membership parameters and rule weights; and (4) quantitative evaluation using FAD, Inception Score (IS), KL divergence, and spectrogram MSE on the reported benchmark datasets.

2. RELATED WORK

A. Indirect Visual-to-Audio Mapping

Early V2A systems frequently used an intermediate text representation. A visual model first described the scene or action, after which a text-to-speech or sound-generation component produced audio from the description. CNN and recurrent neural network architectures, including LSTM networks, were commonly used for visual feature extraction and sequence modeling. Such pipelines can lose non-verbal details and may introduce latency or alignment artifacts because the audio is generated from an intermediate semantic description.

B. Soft Computing and Feature Extraction

Soft computing methods, including fuzzy logic and probabilistic models, are designed to represent imprecision and partial truth. Fuzzy inference can model overlapping categories and uncertain visual evidence, such as partial occlusion, blur, or ambiguous object motion. The general conversion can be represented as O = F(I, C), where O is the generated audio, I is the visual input, and C denotes environmental context.

C. Direct Generative and Foundation Models

Recent approaches map visual latent representations directly into acoustic latent spaces. Contrastive models such as CLIP and CLAP can support cross-modal semantic representations, while GAN and diffusion-based decoders can synthesize audio conditioned on those representations. These models can improve perceptual fidelity, but large generative architectures may require substantial computational resources.

D. Research Gap

The source work identifies a need to balance uncertainty-aware mapping, audio quality, and computational practicality. NF-V2A addresses this gap by inserting a fuzzy inference layer between visual feature extraction and audio generation, and by using a Genetic Algorithm to tune the fuzzy system rather than relying solely on manual rule design.

3. PROPOSED NF-V2A FRAMEWORK

The NF-V2A framework contains four sequential modules: (1) visual feature extraction, (2) adaptive FIS, (3) audio generation network, and (4) Genetic Algorithm optimization. The system receives a silent video segment $V = \{v_1, v_2, \ldots, vT\}$ and produces a corresponding audio representation $A = \{a_1, a_2, \ldots, aT\}$. The goal is to learn a nonlinear mapping from the sequence of visual observations to a temporally coherent audio signal.

$A = F(V; \theta), \quad V = \{v_1, v_2, \ldots, vT\}$

Fig. 1 illustrates the model architecture. The 3D-CNN extracts spatial-temporal visual features. The fuzzy system converts these deterministic features into an uncertainty-aware acoustic latent sequence. The GAN decoder synthesizes a Mel-spectrogram, which is converted into a waveform by Griffin-Lim reconstruction or a pre-trained vocoder such as HiFi-GAN. The GA optimizes the fuzzy parameters using the generation objective.

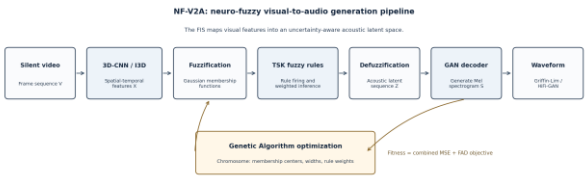


Fig. 1. Architecture of the proposed Neuro-Fuzzy Visual-to-Audio (NF-V2A) framework.

4. METHODOLOGY

A. Visual Feature Extraction

An input video segment is represented by a sequence of T frames, $V = \{v_1, v_2, \ldots, vT\}$. To capture both spatial appearance and temporal motion, the framework uses a pre-trained 3D Convolutional Neural Network. The source specifies initialization using pre-trained I3D weights. The extracted feature matrix is represented as:

$X = F_{3D}\text{-CNN}(V; \theta v), \quad X \in R^{TxD}$

Here, D is the feature dimension and $x_t \in R^D$ denotes the feature vector at time step t. The experimental configuration uses D = 256 for the FIS input.

B. Adaptive Fuzzy Inference System

The FIS maps the visual feature vector x_t into an intermediate acoustic latent representation z_t. A Takagi-Sugeno-Kang (TSK) fuzzy model is used because it provides a computationally efficient rule-based mapping with continuous outputs.

1) Fuzzification

Each scalar feature $x_{t,i}$ is converted into a linguistic membership degree using a Gaussian membership function. For the j-th fuzzy set associated with input dimension i, the membership is:

$\mu_{i,j}(x_{t,i}) = \exp(-(x_{t,i} - c_{i,j})^2 / (2\sigma_{i,j}^2))$

The parameter $c_{i,j}$ is the center and $\sigma_{i,j}$ is the width (standard deviation) of the Gaussian membership function. These parameters control the overlap and sensitivity of the fuzzy sets.

2) Fuzzy Rule Base and Inference

The fuzzy rule base consists of R IF–THEN rules. A generic TSK rule Rk is defined as:

Rk: IF $x_{t,1}$ is $A_{k,1}$ AND … AND $x_{t,D}$ is $A_{k,D}$, THEN $y_k = w_{k,0} + \Sigma_{i=1}^{D} w_{k,i}\, x_{t,i}$

Using the algebraic product T-norm, the firing strength of rule k is calculated as:

$\omega_k = \Pi_{i=1}^{D} \mu_{k,i}(x_{t,i})$

3) Defuzzification

The crisp acoustic latent vector is obtained through a normalized weighted average of the rule outputs:

$z_t = (\Sigma_{k=1}^{R} \omega_k\, y_k) / (\Sigma_{k=1}^{R} \omega_k)$

The resulting latent sequence $Z = \{z_1, z_2, \ldots, zT\}$ retains a continuous representation suitable for the downstream audio decoder.

C. Audio Generation Network

The defuzzified latent representation Z is supplied to a GAN-based decoder. The generator produces a Mel-spectrogram S with frequency and time dimensions:

$S = G\theta gen(Z), \quad S \in R^{fxT'}$

Here, F represents the number of frequency bins and T′ is the number of audio time frames. The generated Mel-spectrogram is converted into a waveform using the Griffin-Lim algorithm or a pre-trained vocoder, such as HiFi-GAN.

D. Genetic Algorithm Optimization

Manually selecting fuzzy membership centers, widths, and rule weights is difficult for high-dimensional cross-modal data. The framework therefore uses a Genetic Algorithm. Each chromosome C encodes the membership centers c, widths σ, and linear rule weights w. The fitness function minimizes a combined objective based on spectrogram reconstruction error and FAD:

$\text{Fitness}(C) = 1 / [\alpha \cdot \text{MSE}(S, S^*) + \beta \cdot \text{FAD}(S, S^*) + \varepsilon]$

Here, S is the generated spectrogram, S* is the ground-truth spectrogram, α and β are weighting coefficients, and ε is a small constant. The GA applies selection, uniform crossover, and Gaussian mutation over successive generations

until convergence. The stated configuration uses a population of 100 chromosomes, 50 generations, crossover rate 0.8, and mutation rate 0.05.

Fig. 2 summarizes the training and evolutionary optimization workflow. The GA evaluates candidate fuzzy parameters using the generated audio objective and returns improved membership parameters and rule weights to the FIS.

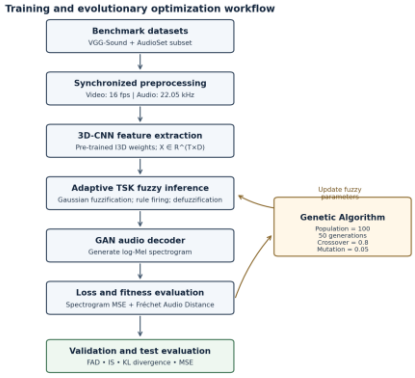


Fig. 2. Training and Genetic Algorithm optimization workflow.

E. Inference Workflow

During inference, the trained visual feature extractor, optimized FIS, and audio generator process a silent video clip to produce a waveform. The video is sampled at the stated rate of 16 frames per second. The visual features are mapped into the acoustic latent space, decoded as a Mel-spectrogram, and converted into audio. The GA is used during optimization; the optimized FIS parameters are held fixed during inference. Fig. 3 illustrates this sequence.

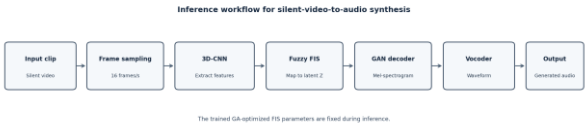


Fig. 3. Inference workflow for silent-video-to-audio synthesis.

5. EXPERIMENTAL SETUP

A. Datasets and Preprocessing

The model is reported to be trained and evaluated using two audio-visual datasets. VGG-Sound contains more than 200,000 video clips across 300 audio classes. A curated AudioSet subset focuses on classes with strong visual-audio correspondence, including playing musical instruments, dog barking, and breaking glass; the subset contains approximately 50,000 clips of 10 seconds each. Video is sampled at 16 frames per second, and audio is resampled to 22,050 Hz. The data is divided into 70% training, 15% validation, and 15% testing sets.

B. Implementation Configuration

The 3D-CNN is initialized with pre-trained I3D weights. The FIS uses D = 256 input dimensions and R = 50 initial fuzzy rules. The GA uses a population size of 100 chromosomes over 50 generations, with crossover rate 0.8 and mutation rate 0.05. The framework is trained with the Adam optimizer using an initial learning rate of 1×10^{-4} on two NVIDIA A100 GPUs.

C. Evaluation Metrics

Four metrics are reported to evaluate generated-audio quality and cross-modal alignment:

1) Fréchet Audio Distance (FAD): Measures the distance between real and generated audio distributions using VGGish embeddings; lower values indicate closer distributions.

2) Inception Score (IS): Measures the clarity and diversity of generated audio; higher values indicate better performance.

3) Kullback-Leibler (KL) divergence: Quantifies the divergence between visual and audio representations; lower values indicate better alignment.

4) Spectrogram Mean Squared Error (MSE): Measures reconstruction error between predicted and ground-truth log-Mel spectrograms; lower values indicate better reconstruction.

6. RESULTS AND DISCUSSION

The proposed NF-V2A model is compared with a deterministic CNN-RNN model, a V2A-GAN model, and a diffusion-based Diff-V2A model. Table I reproduces the quantitative results reported for the VGG-Sound test set. A downward arrow indicates that lower values are better, while an upward arrow indicates that higher values are better.

TABLE I
QUANTITATIVE COMPARISON ON THE VGG-SOUND TEST SET

Model architecture	**FAD ↓**	**IS ↑**	**KL ↓**	**Spectrogram MSE ↓**
CNN-RNN (Baseline)	5.82	4.12	2.45	0.084
V2A-GAN	4.31	5.85	1.89	0.062
Diff-V2A (Diffusion)	3.55	6.72	1.45	0.051
NF-V2A (Proposed)	3.04	7.41	1.12	0.043

A. Comparative Performance

The NF-V2A model reports the lowest FAD (3.04), KL divergence (1.12), and spectrogram MSE (0.043), together with the highest IS (7.41) among the four listed configurations. Relative to Diff-V2A, the FAD decreases from 3.55 to 3.04, corresponding to approximately 14.3%. Spectrogram MSE decreases from 0.051 to 0.043, corresponding to approximately 15.7% (reported in the source as 15.6%). These results are consistent with the claim that the FIS provides a more tolerant mapping for ambiguous visual events.

B. Ablation Study

The ablation experiment evaluates the effect of removing the fuzzy FIS or the GA optimization. In the first variant, the FIS is replaced with a standard linear dense layer. In the second variant, the FIS is retained but its parameters are optimized using backpropagation instead of the GA. The complete architecture includes both components.

TABLE II
ABLATION STUDY OF FUZZY AND GA COMPONENTS

Architecture configuration	**FAD ↓**	**KL divergence ↓**
NF-V2A (without fuzzy FIS)	4.56	2.01
NF-V2A (without GA optimization)	3.89	1.54
NF-V2A (full model)	3.04	1.12

The ablation values show that removing the fuzzy FIS produces the largest degradation among the listed variants. Removing GA optimization also worsens both FAD and KL divergence compared with the full model. The results support the combined use of fuzzy inference and evolutionary parameter optimization in the reported configuration.

7. CONCLUSION AND FUTURE WORK

This paper presents NF-V2A, a neuro-fuzzy framework for generating audio from silent video. A pre-trained 3D-CNN extracts spatial-temporal features, a TSK FIS maps visual features to an uncertainty-aware acoustic latent space, and

a GAN decoder generates a Mel-spectrogram for waveform reconstruction. A Genetic Algorithm optimizes the fuzzy membership functions and rule weights using an objective based on MSE and FAD. The reported VGG-Sound evaluation shows that NF-V2A achieves FAD of 3.04, IS of 7.41, KL divergence of 1.12, and spectrogram MSE of 0.043, outperforming the listed CNN-RNN, V2A-GAN, and Diff-V2A baselines on these measures.

Future work identified in the source includes multimodal IoT integration, real-time edge deployment through model quantization and pruning, and zero-shot cross-modal transfer to visual objects and actions not seen during training.

References

[1] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "VGGSound: A Large-scale Audio-Visual Dataset," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 721–725.

[2] J. F. Gemmeke et al., "Audio Set: An Ontology and Human-Labeled Dataset for Audio Events," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 776–780.

[3] J.-S. R. Jang, "ANFIS: Adaptive-Network-Based Fuzzy Inference System," IEEE Transactions on Systems, Man, and Cybernetics, vol. 23, no. 3, pp. 665–685, 1993.

[4] S. Luo et al., "Diff-Foley: Synchronized Video-to-Audio Synthesis with Latent Diffusion Models," in Advances in Neural Information Processing Systems (NeurIPS), 2023.

[5] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually Indicated Sounds," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2405–2413.

[6] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6299–6308.

[7] K. Kilgour et al., "Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms," in Proc. Interspeech, 2019, pp. 2350–2354.

[8] F. Herrera, M. Lozano, and J. L. Verdegay, "Tuning Fuzzy Logic Controllers by Genetic Algorithms," International Journal of Approximate Reasoning, vol. 12, nos. 3–4, pp. 299–315, 1995.

[9] C. Gan, D. Huang, P. Chen, J. B. Tenenbaum, and A. Torralba, "Foley Music: Learning to Generate Music from Videos," in Proc. European Conference on Computer Vision (ECCV), 2020, pp. 758–775.

[10] D. W. Griffin and J. S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 2, pp. 236–243, 1984.