VIDEO TO AUDIO GENERATION

Riccardo F. Gramaccioni, Christian Marinoni, Danilo Comminiello, Aurelio Uncini

Dipartimento di Ingegneria dell'Informazione, Elettronica e Telecomunicazioni (DIET)
"Sapienza" Università di Roma
Via Eudossiana 18, 00184 Roma

Parole chiave: video-to-audio, audio-video synchronization, multimodal audio synthesis

Introduzione

La generazione automatica di audio sincronizzato con contenuti visivi rappresenta una sfida aperta nel campo dell'intelligenza artificiale multimodale, con applicazioni in ambito cinematografico, videoludico e della realtà virtuale. L'obiettivo è realizzare un sistema capace di generare effetti sonori coerenti e sincronizzati a partire da sequenze video, riducendo il carico manuale del sound design. Tuttavia, i modelli esistenti non offrono un controllo esplicito sulla sincronizzazione temporale o sulla coerenza semantica tra suono e immagine. In questo contesto, viene proposto FolAI [1], un nuovo framework per la sintesi di effetti sonori sincronizzati, basato su modelli di diffusione controllabili. Il metodo introduce un'architettura a due stadi che separa il controllo temporale da quello semantico, offrendo al tempo stesso generazione realistica e flessibilità creativa.

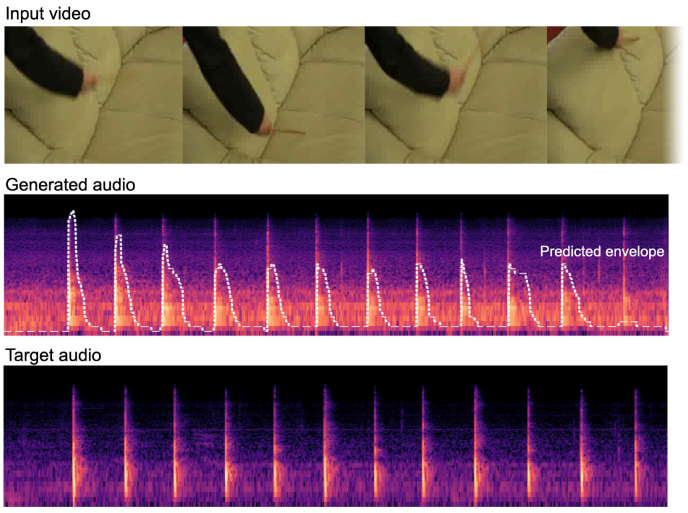

Figura 1: Esempio di video in input in cui una bacchetta batte su un oggetto, di audio generato corrispondente ai colpi della bacchetta sull'oggetto temporalmente allineato al target.

FolAI prevede una prima fase in cui viene stimato un inviluppo che identifica la distribuzione temporale degli eventi audio attesi. Successivamente, il processo generativo viene guidato da embedding audio forniti dall'utente per condizionare semanticamente il suono generato. In questo modo, il sistema può produrre suoni coerenti con il contesto visivo ma anche personalizzabili a seconda dell'intenzione artistica. Il modello sfrutta come generatore Stable Audio Open, adattato per ricevere sia la traccia video che l'inviluppo temporale, integrando inoltre la possibilità di controllare semanticamente l'audio tramite tecniche ispirate ai modelli di diffusione condizionata (es. ControlNet [2]). Le metriche di valutazione sperimentale

mostrano che FolAI raggiunge prestazioni superiori rispetto ai metodi esistenti, mantenendo al contempo un alto livello di controllo creativo.

FolAI rappresenta un significativo passo avanti nella generazione automatica di effetti sonori video-guidati, grazie alla sua capacità di disaccoppiare la componente temporale da quella semantica e alla possibilità di integrare facilmente il contributo di professionisti del suono nel processo generativo.

Bibliografia

[1] R. Fosco Gramaccioni et al., “Stable-V2A: Synthesis of Synchronized Sound Effects with Temporal and Semantic Controls”, arXiv:2412.15023, 2025.
[2] H. Chen et al., “ControlNet: Adding Conditional Control to Text-to-Image Diffusion Models”, CVPR 2023.