Oltre il DSP: Tecniche, Sfide e Applicazioni dell'AI Generativa per l'Audio

DIET Dipartimento di Ingegneria dell'Informazione, Elettronica e Telecomunicazioni

Intelligent Signal Processing and MultiMedia lab

Giuliano Galadini,
giuliano.galadini@uniroma1.it

Intelligent Signal Processing and MultiMedia Lab

- **Deep Learning per l'audio**: Modelli generativi, analisi semantica, sintesi neurale
- **Multimodalità**: Integrazione di **audio, video, testo** per applicazioni innovative.
- **Representation Learning: rappresentazioni astratte e compatte** che catturano le **caratteristiche essenziali** del dato
- **Hypercomplex Neural Network:** modellare **relazioni complesse e multidimensionali** tra i dati

Intelligent Signal Processing
and MultiMedia lab

Tecniche Generative – Passato e Presente

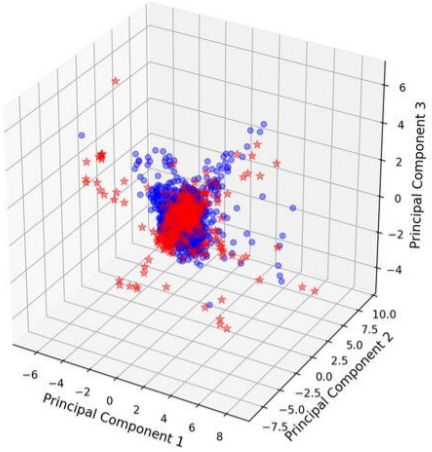
- L’idea di usare **processi stocastici** per esplorare nuove frontiere artistiche non è nuova: risale almeno al **XX secolo**.

- Oggi, una **tecnica generativa** in ambito audio/musicale è un metodo che usa algoritmi (spesso basati su IA) per creare suoni o musica **nuovi e originali** a partire da dati esistenti o regole

IA Generativa – Variational Autoencoder (VAE)

Per generare nuovi audio, il VAE:

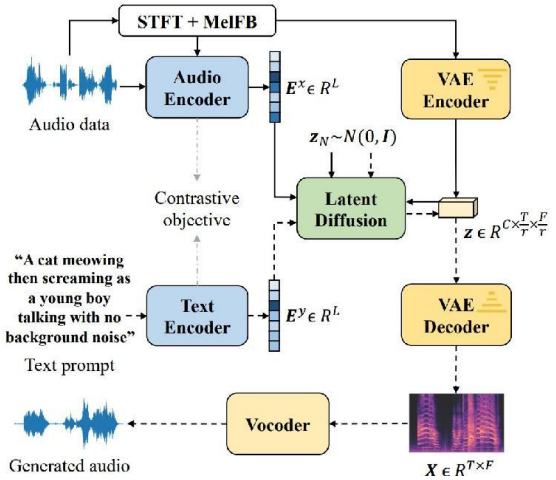
- Viene allenato su un **numero enorme di audio**
- Impara le **caratteristiche astratte** di ogni audio e le ordina (spazio latente)
- Di conseguenza, impara cosa c'è nello **spazio «vuoto»** tra gli audio che ha visto



IA Generativa – Diffusion Models

Un modello a diffusione impara a generare il dato richiesto a partire da **rumore casuale**:

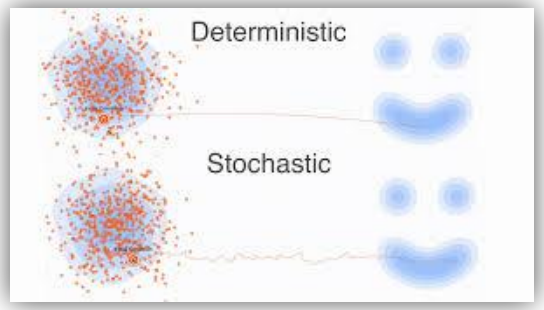
- Se il dato è troppo complesso *(es. audio raw a 44.1 kHz)* si preferisce **lavorare sulla sua astrazione** (generazione nello spazio latente)
- È un processo **iterativo** che richiede un numero elevato di step
- La qualità della generazione è notevolmente **più alta** rispetto al VAE

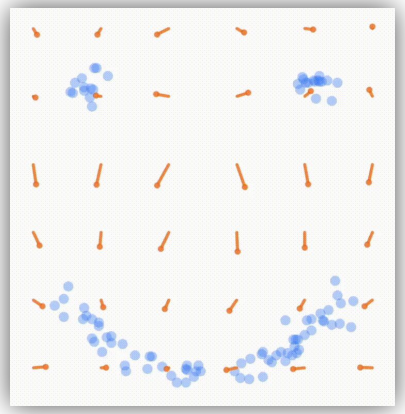


IA Generativa – Flow Matching (FM)

Flow Matching impara a generare il dato richiesto **a partire da rumore**:

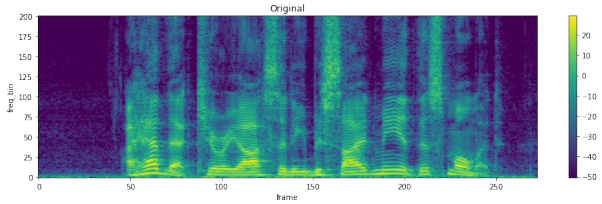
- Non lavora passo-passo, ma impara direttamente il **percorso ottimale** per trasformare il rumore nel dato desiderato
- È un processo **deterministico** che richiede **pochi step** (o addirittura uno solo)
- La qualità della generazione è **paragonabile alla diffusione**, ma con maggiore efficienza e **velocità**

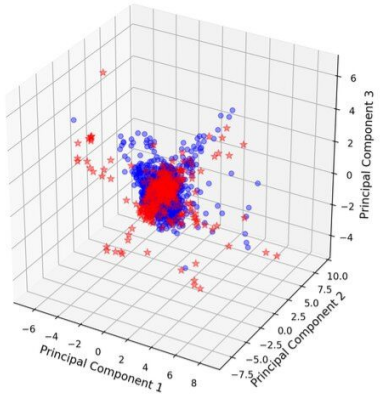




Come si rappresenta l'audio?

- **Raw Audio:** audio «grezzo», rappresenta l'ampiezza del suono nel tempo
- **Spettrogramma (Mel/STFT):** tempo (asse X) vs. frequenza (asse Y), con intensità (colori).
- **Rappresentazione Latente:** codice astratto (vettore numerico) che descrive le macro-caratteristiche del suono





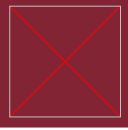
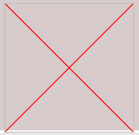
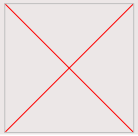
Confronto tra le rappresentazioni

Rappresentazione	Vantaggi	Svantaggi	Uso tipico
Raw	Tutti i dettagli (fase, timbro, dinamica). **Nessuna perdita di informazione**	**Alta dimensionalità** (es. 44.100 campioni/sec). Difficile da elaborare per l'IA.	Registrazione, editing diretto, sintesi con modelli raw
Spettrogramma	**Comprime l'informazione,** cattura pattern temporali e frequenziali	**Perde la fase** (difficile ricostruire l'onda originale), sensibile a rumore	Classificazione, sintesi vocale e sonora, feature extraction
Latente	**Bassa dimensionalità** (es. 512 valori), **efficiente** per l'IA, cattura **semantica** (es. "chitarra acustica" vs. "violino").	**Perdita di interpretabilità**, dipende dal modello usato per generarla.	Generazione audio, retrieval semantico e multimodalità

IA Generativa vs. DSP: Imparare ciò che non sappiamo modellare

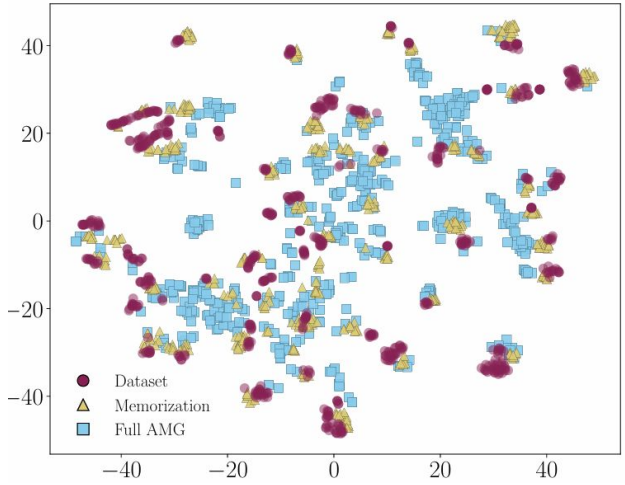
- Il DSP classico si basa su **funzioni matematiche esplicite** che **sappiamo definire manualmente**
- L'IA generativa non si limita a regole predefinite: **impara** le **funzioni complesse** direttamente dai dati
- L'IA generativa non sostituisce il DSP, ma lo **estende**, permettendo di fare cose che prima erano **impossibili da definire**.

Opportunità e Sfide nel Mondo dell'Audio

	**Chi possiede la musica generata da un'IA?**
	**Tecnologie che permettono di clonare voci umane con precisione**
	**I modelli possono 'memorizzare' dati di training e riprodurli quasi identici**
	I modelli riflettono i bias presenti nei dati di addestramento

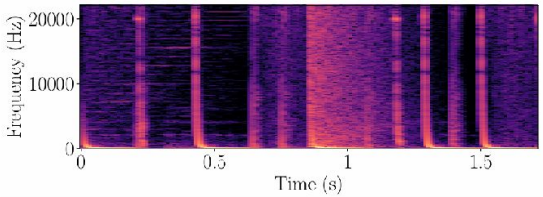
Un Caso di Studio: la Memorizzazione nei Modelli Generativi

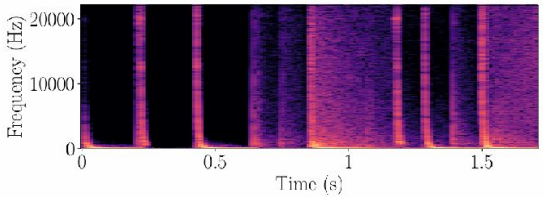
- Se un modello generativo crea una canzone **identica a una esistente**, chi ne detiene i diritti? L'artista originale o chi ha addestrato il modello?
- **Bisogna mitigare questo rischio** senza sacrificare la qualità della generazione.

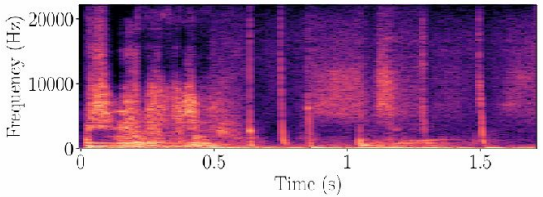


Allontanarsi dalle regioni «memorizzate»

- Il nostro approccio considera **tutti i vicini semantici** nel dataset
- Usiamo una **misura geometrica** per assicurarci che la generazione sia **lontana dall'intero gruppo di esempi simili**







Source Separation

- È una tecnica che **scompone un segnale audio** nei suoi **componenti originali**
- LARS 2.0 è specificatamente pensato per separare loop percussivi nelle singole tracce

Neural Virtual Analog

- I **sintetizzatori virtual analog** emulano il suono dei sintetizzatori analogici classici via software
- Il **Neural Virtual Analog** impara da dati audio reali di sintetizzatori analogici e genera suoni indistinguibili da quelli originali.

Neural Audio Synthesis

- **RAVE** (Realtime Audio Variational autoEncoder) è un **modello di sintesi audio basato su VAE** che permette di **generare suoni in tempo reale**
- Divide il segnale audio in **bande di frequenza** (PQMF) per migliorare la qualità e ridurre la latenza

Conclusioni e Domande

Grazie!

Giuliano Galadini,
giuliano.galadini@uniroma1.it

Intelligent Signal Processing
and MultiMedia lab