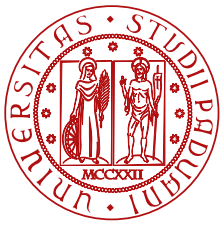


Università degli Studi di Padova

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA INFORMATICA

Analisi e valutazione di modelli diffusion basati su spettrogrammi per la generazione di musica ed effetti sonori

Relatore
Prof. Rodà Antonio

Laureando
Bulmaga Daniela

ANNO ACCADEMICO 2023-2024

Data di laurea 15/07/2024

Sommario

Negli ultimi due anni i modelli di diffusione sono stati presi in esame per la loro capacità di modellare distribuzioni di dati complesse, mostrando risultati promettenti in compiti di *inpainting*, generazione di immagini, *super-resolution*, *denoising* di immagini e *instance segmentation*. Oltre a questo i modelli di diffusione si sono dimostrati efficaci anche nella sintesi vocale, nel miglioramento dell'audio e nella generazione di audio.

L'obiettivo di questa tesi sarà, appunto, la valutazione di quest'ultimo compito. Lo studio analizza e confronta diversi modelli di diffusione basati su spettrogrammi per la sintesi di audio di alta qualità, concentrandosi sulla generazione di musica ed effetti sonori.

Per rispondere alla domanda delle ricerca si usa un'indagine a base di sondaggio, in cui ai partecipanti viene presentata una serie di campioni audio generati che comprendono brani musicali ed effetti sonori e viene chiesto loro di valutare vari attributi come la chiarezza del suono, la fedeltà al prompt testuale e il realismo.

Sebbene l'architettura dei modelli abbia influenzato i risultati, è emerso che la qualità e la diversità dei dataset utilizzati per l'addestramento giocano un ruolo cruciale nel determinare l'eccellenza dei nuovi campioni audio generati. In altre parole, un dataset più ricco e diversificato consente ai modelli di apprendere una gamma più ampia di caratteristiche audio, migliorando così la loro capacità di produrre suoni realistici e di alta qualità.

Indice

1 Introduzione **4**

2 Introduzione all'IA Generativa **6**

2.1 Concetti base 6

2.2 Autoencoder 8

2.2.1 Variational Autoencoder 9

2.3 Diffusion Models 10

2.3.1 Denoising Diffusion Probabilistic Model 10

2.3.2 Latent Diffusion Models 13

2.3.3 Architettura 14

2.4 Spettrogramma 15

2.4.1 Scala Mel 16

2.4.2 Spettrogrammi Mel e Musica 16

2.5 Modelli 17

2.5.1 Auffusion 19

2.5.2 Tango 21

2.5.3 Mustango 22

2.5.4 AudioLDM 2 24

3 Metodologia **26**

3.1 Materiali 26

3.2 Raccolta delle risposte 27

4 Analisi dei dati **29**

4.1 Effetti sonori 29

4.2 Musica 32
4.3 Analisi demografica 35

5 Conclusioni **36**

A Informazioni sul sondaggio **44**

Elenco delle figure

2.1 Evoluzione dei *dataset* negli anni [Villalobos and Ho, 2022] 7
2.2 Esempio di un Autoencoder [Bank et al., 2021] 8
2.3 Progressione del rumore usando una varianza lineare (sopra) e una varianza coseno (sotto) [Nichol and Dhariwal, 2021] 12
2.4 Esempio di un *Latent Diffusion Model* [Rombach et al., 2022] 14
2.5 Esempio di una U-Net [Ronneberger et al., 2015] 15
2.6 Spettrogrammi Mel di due passaggi della stessa melodia. 17
2.7 Architettura del modello Auffusion [Xue et al., 2024] 19
2.8 Architettura del modello Tango [Ghosal et al., 2023] 21
2.9 Architettura del modello Mustango [Melechovsky et al., 2024] . . . 22
2.10 Architettura del modello AudioLDM 2 [Liu et al., 2024] 24

4.1 Media delle valutazioni per gli effetti sonori 31
4.2 Media delle valutazioni per la musica 33
4.3 Distribuzione dei generi del gruppo campione 35
4.4 Valutazioni dei due sondaggi 35

A.1 Pagina informativa del sondaggio 46
A.2 Continuazione della pagina informativa 47
A.3 Prova del suono eseguita dai partecipanti prima di proseguire con la valutazione dei campioni audio generati 48
A.4 Esempio di un clip audio e le domande poste ai partecipanti 49
A.5 Domande demografiche poste ai partecipanti 50

Capitolo 1

Introduzione

Definire cosa sia l'intelligenza artificiale (AI) non è mai stato semplice, poiché richiederebbe prima di tutto definire il concetto di "intelligenza" e comprendere in che modo quella delle macchine si differenzi da quella considerata intrinseca agli esseri umani.

Nell'ambito ingegneristico l'intelligenza artificiale è la disciplina il cui scopo è riprodurre o emulare alcune funzioni dell'intelligenza umana, tra cui l'apprendimento, la risoluzione dei problemi, la percezione e la comprensione del linguaggio naturale. Dotata di queste abilità, una macchina può addirittura simulare ciò che comunemente viene definita "creatività".

È proprio questa capacità che ha catturato l'interesse del pubblico nel novembre del 2022 con l'introduzione di ChatGPT, portando sotto i riflettori anche altri modelli generativi come DALL-E, Stable Diffusion e Midjourney, che sono in grado di generare immagini di qualunque tipo in pochi minuti a partire da un *input* testuale di un utente. Questi modelli sono in grado generare dati, immagini, testo, suoni o altri tipi di informazioni che fino a ora si pensava potessero essere replicati solo dagli esseri umani.

Spesso quando si parla di AI generativa si parla di sintesi di immagini, mentre la generazione di audio è rimasta relativamente poco conosciuta al pubblico; tuttavia, il 2023 ha segnato un momento significativo per il campo della generazione del suono, che comprende la creazione di contenuti audio sintetici, che vanno dalla voce umana ai brani musicali. Questo progresso è stato evidenziato dal lancio di

diversi importanti modelli per la generazione di audio, come UniAudio, MusicGen e MusicLM.

Questo lavoro si propone di esaminare quattro modelli di intelligenza artificiale che utilizzano il processo di diffusione per la generazione di audio. Attraverso un'indagine basata su un sondaggio, verranno valutate le loro capacità nel creare audio nuovo e realistico. Tale valutazione fornirà indicazioni sull'efficacia e sulle potenziali applicazioni di questi modelli in diversi ambiti.

La presente tesi è strutturata nel presente modo:

- Nel capitolo 2 verrà introdotto il processo di diffusione su cui si basa la generazione di spettrogrammi Mel, essenziali per la sintesi del suono. Inoltre, verranno presentati i modelli scelti per questa tesi, le loro architetture e i *dataset* utilizzati.
- Nel capitolo 3 sarà descritta la metodologia utilizzata per condurre l'indagine.
- Nel capitolo 4 saranno esposti i risultati delle indagini e le ipotesi che potrebbero spiegare tali risultati.
- Nel capitolo 5 saranno tratte le conclusioni relative alla ricerca effettuata.

Capitolo 2

Introduzione all'IA Generativa

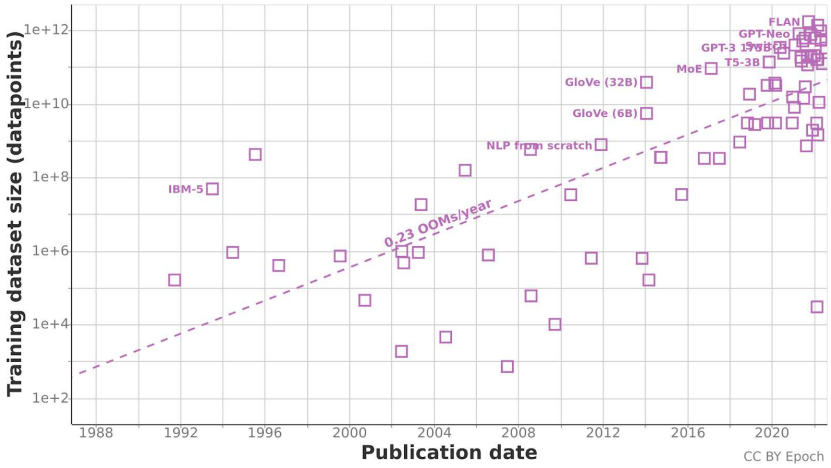
Questo capitolo illustra alcune architetture comunemente impiegate per la generazione di dati, ponendo particolare attenzione alle strutture e ai concetti utilizzati nei quattro modelli selezionati. Nella sezione 2.1 viene introdotta l'intelligenza artificiale generativa, nella sezione 2.2 vengono introdotti gli *autoencoder*, nella sezione 2.3 viene esaminato il processo di diffusione, nella sezione 2.4viene spiegato il ruolo degli spettrogrammi Mel e la loro applicazione nell'analisi musicale. Infine, nella sezione 2.5 vengono descritti i modelli analizzati in questa tesi.

2.1 Concetti base

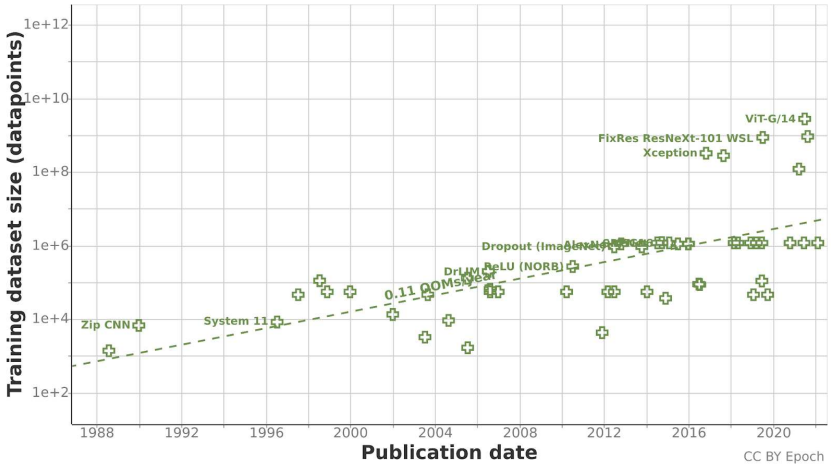
I modelli di apprendimento automatico si suddividono in due categorie principali: discriminativi e generativi. I modelli discriminativi si concentrano nel prevedere la probabilità di un evento in base a un contesto, rendendoli ideali per compiti di classificazione in cui i dati vengono etichettati. I modelli generativi, invece, mirano a comprendere le distribuzioni dei dati per generare nuovi contenuti simili. Questi modelli sono utilizzati principalmente per creare testi, immagini, musica, audio e video.

Questo è stato possibile in gran parte grazie alla convergenza di due fattori cruciali: L'aumento delle capacità di calcolo, reso possibile da innovazioni hardware come i *neural processing unit*, e l'uso di *dataset* di larga scala appositamente concepiti per compiti generativi che hanno migliorato la capacità dei modelli di

apprendere da una vasta gamma di dati di alta qualità, aumentando così la loro capacità di generare contenuti realistici [Villalobos and Ho, 2022].



(a) Dimensione di *dataset* linguistici in funzione del tempo



(b) Dimensione di *dataset* contenenti immagini e video in funzione del tempo

Figura 2.1: Evoluzione dei *dataset* negli anni [Villalobos and Ho, 2022]

Nonostante la preoccupazione che l'intelligenza artificiale possa sostituire gli artisti persiste, molti la vedono come uno strumento che potrebbe rendere l'espressione artistica più accessibile e meno laboriosa. Mentre gli investimenti privati complessivi nell'IA sono diminuiti nel 2022, i finanziamenti per l'IA generativa sono aumentati notevolmente. Nel 2023, il settore ha attirato 25,2 miliardi di dollari,

quasi nove volte l'investimento del 2022 e circa 30 volte quello del 2019. Inoltre, l'IA generativa ha rappresentato oltre un quarto di tutti gli investimenti privati legati all'IA nel 2023 [Maslej et al., 2024].

2.2 Autoencoder

L'*autoencoder* è una rete neurale progettata per comprimere i dati in ingresso, riducendoli alle loro caratteristiche essenziali, e poi ricostruire l'*input* originale a partire da questa rappresentazione compressa. Gli *autoencoder* utilizzano l'apprendimento auto-supervisionato, il che significa che non necessitano di set di dati etichettati. Tuttavia, come nei modelli supervisionati, dispongono di una *ground truth*, rappresentata dal proprio *input*, con cui confrontano le loro prestazioni [Bank et al., 2021].

La struttura di un *autoencoder* comprende due parti principali:

- L'*encoder*, che comprime i dati di *input* in una rappresentazione a bassa dimensionalità chiamata spazio latente, catturando le caratteristiche più rilevanti dei dati. Questa operazione viene anche chiamata *"bottleneck"*.
- Il *decoder*, che ricostruisce i dati di *input* dalla rappresentazione dello spazio latente. L'obiettivo del *decoder* è generare un *output* che sia più simile possibile all'*input* originale

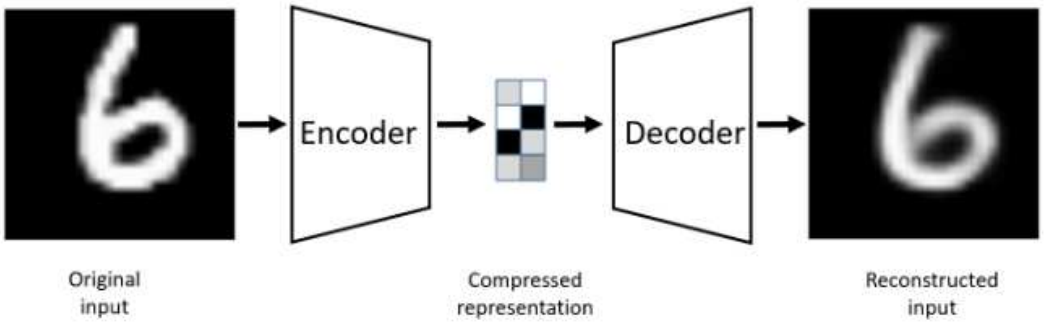


Figura 2.2: Esempio di un Autoencoder [Bank et al., 2021]

Questo processo di compressione ed estrazione delle caratteristiche riduce il volume dei dati senza perdere informazioni importanti semplificando e accelerando l'addestramento del modello.

Dalla loro introduzione sono emerse diverse tipologie di *autoencoder*. Queste varianti mirano principalmente a risolvere i punti deboli e soddisfare particolari bisogni. Alcuni esempi notevoli sono i *Denoising Autoencoder* (DAE), i *Sparse Autoencoder* (SAE) e più recentemente i *Variational Autoencoder* (VAE). Ognuno di questi diversi modelli ha i propri punti di forza.

I *Denoising Autoencoder* ricostruiscono dati puliti a partire da un *input* rumoroso. Il loro funzionamento prevede la corruzione casuale dei dati in ingresso con rumore, seguita dall'addestramento dell'*autoencoder* per ricostruire i dati originali privi di rumore. Questo tipo di *autoencoder* è stato principalmente applicato alle immagini, con l'obiettivo di migliorarne la qualità e facilitarne l'analisi successiva [Vincent et al., 2008].

Per gli *Sparse Autoencoder* viene imposto il vincolo della sparsità, ossia durante l'addestramento rimangono attivi solo una parte dei nodi. In questo modo, l'*autoencoder* impara le caratteristiche fondamentali dei dati in ingresso, evitando informazioni ridondanti. Questo tipo di *autoencoder* offre ottime prestazioni nei compiti di classificazione e nel riconoscimento di pattern in dati ad alta dimensionalità [Makhzani and Frey, 2014].

2.2.1 Variational Autoencoder

I VAE (*Variational Autoencoders*) [Kingma and Welling, 2022] sono modelli di intelligenza artificiale generativa che utilizzano concetti di inferenza bayesiana per apprendere e generare dati. A differenza di altri tipi di *autoencoder* che operano in uno spazio latente discreto, i VAE lavorano in uno spazio latente continuo. Questo approccio consente loro di apprendere gli attributi latenti dei dati come una distribuzione di probabilità, anziché come un modello deterministico.

In altre parole, i VAE adottano un approccio stocastico, che permette di campionare da questa distribuzione per generare dati nuovi. Questa caratteristica

rende i VAE particolarmente potenti per la generazione di dati nuovi e variabili, ampliando notevolmente le loro applicazioni in vari campi dell'intelligenza artificiale. [Bergmann and Stryker, 2023]

Come per gli *autoencoder* tradizionali, l'*encoder* e il *decoder* degli VAE possono essere divisi e utilizzati separatamente.

2.3 Diffusion Models

I modelli di diffusione hanno le loro radici nella termodinamica del non equilibrio e nei metodi Monte Carlo [Sohl-Dickstein et al., 2015]. Introdotti per la prima volta da Sohl-Dickstein et al., si sono gradualmente affermati come una valida alternativa ai modelli generativi convenzionali, come i GAN e i VAE. In effetti, i modelli di diffusione costituiscono la base del popolare modello *open-source* di generazione di immagini, Stable Diffusion. Le loro capacità generative possono essere impiegate per sintetizzare una vasta gamma di dati, tra cui immagini, testi, audio e video.

Il successo dei modelli di diffusione può essere attribuito alla loro flessibilità e trattabilità, che permettono di apprendere da vasti insiemi di dati complessi. Inoltre, essi offrono vantaggi aggiuntivi in termini di scalabilità e parallelizzazione.

Dalla loro introduzione, sono stati sviluppati diversi modelli di diffusione, tra cui i *Denoising Diffusion Probabilistic Models* (DDPM) [Ho et al., 2020] e i *Noise-Conditioned Score Networks* (NCSN). Questi modelli condividono formulazioni simili, basate su principi fondamentali comuni. Tuttavia, in questo lavoro, mi concentrerò sui DDPM e su una delle loro varianti, il *Latent Diffusion Model* (LDM) [Rombach et al., 2022].

2.3.1 Denoising Diffusion Probabilistic Model

Il DDPM è costruito a partire da una gerarchia di *denoising autoencoder* e si compone da due parti principali: *forward process* e *reverse process*.

Forward Process

Durante il *forward process*, noto anche come processo di diffusione, si utilizza una catena di Markov per aggiungere rumore gaussiano all'immagine di *input* secondo un programma di varianza β, che regola la quantità di rumore aggiunto a ogni passo. In questo modo, ogni passo dipende solo dallo stato precedente [Weng, 2021].

L'immagine iniziale, indicata come x_0, viene campionata dalla distribuzione dei dati $q(x_0)$. L'immagine rumorosa finale, nota come x_T, si ottiene dopo l'applicazione sequenziale di T passi di rumore a partire da x_0. Ciascun passo di aggiunta del rumore, da x_1 a x_T, rappresenta una variabile latente con la stessa dimensionalità di x_0. Man mano che il numero di passi T aumenta, l'immagine si trasforma gradualmente in puro rumore e alla fine viene completamente distrutta. Se il programma di varianza $\beta_1 < \beta_2 < ... < \beta_T$, allora per un T infinitamente grande x_T diventa una Gaussiana così che $x_T \sim N(0, I)$. Questo permette di generare campioni semplicemente attingendo da una distribuzione gaussiana.

$$q(x_1,\ ...\ ,\ x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{2.1}$$

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{1-\beta_t}\ x_{t-1},\ \beta_t I) \tag{2.2}$$

Il programma di varianza scelto riveste un ruolo significativo nel determinare la qualità e l'efficacia del processo di avanzamento. Nichol et al. hanno introdotto una varianza che usa la funzione coseno progettata in modo che il livello di rumore nella parte centrale del processo aumenti in modo lineare, mentre l'aumento sia smorzato vicino ai punti estremi $t = 0$ e $t = T$ per evitare repentini cambiamenti nell'intensità dell'aggiunta del rumore. In contrasto, il programma lineare proposto da Ho et al. aumenta i livelli di rumore molto più rapidamente, potenzialmente causando una perdita di informazioni. Il programma coseno presenta il vantaggio di ridurre il numero di passaggi di campionamento richiesti rispetto al programma lineare.

Figura 2.3: Progressione del rumore usando una varianza lineare (sopra) e una varianza coseno (sotto) [Nichol and Dhariwal, 2021]

Reverse Process

Il *reverse process* ha un ruolo esattamente opposto a quello del *forward process*. Esso mira a rimuovere il rumore aggiunto durante il processo di diffusione e a ricostruire il campione di dati in *input*. Idealmente, il processo inverso definisce una catena di Markov variazionale che costruisce x_T, $x_T - 1$, ..., x_1, x_0 in modo che siano campionati da $q(x_t - 1|x_t)$. Tuttavia, non è possibile stimare la distribuzione posteriore $q(x_t - 1|x_t)$ poiché richiederebbe l'accesso alla distribuzione completa dei dati. È possibile, invece, apprendere una distribuzione approssimativa p_θ, per eseguire il processo inverso, tale che:

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t),\ \Sigma_\theta(x_t, t)) \tag{2.3}$$

Quindi definiamo:

$$p\theta(x_0,\ \dots\ ,\ x_T) = p(x_T) \prod_{t=1}^{T} p\theta(x_{t-1}|x_t) \tag{2.4}$$

dove la media $\mu_\theta(x_t, t)$ e la varianza $\Sigma_\theta(x_t, t)$ possono essere calcolate durante l'addestramento.

I *diffusion models* sono particolarmente impegnativi dal punto di vista computazionale a causa delle ripetute valutazioni delle funzioni e dei calcoli del gradiente richiesti sia durante l'addestramento sia durante l'inferenza. Questi processi avvengono nello spazio ad alta dimensionalità delle immagini RGB, il che aumenta ulteriormente la complessità computazionale. Di conseguenza, l'addestramento e l'inferenza con i modelli di diffusione richiedono sostanziali risorse computazionali,

che a loro volta comportano un notevole consumo energetico nei centri elaborazione dati. [Luccioni et al., 2024]

Questo consumo energetico non è associato solo all'elettricità utilizzata per alimentare l'hardware di calcolo, ma anche al funzionamento continuo dei robusti sistemi di condizionamento dell'aria necessari per raffreddare i processori. Questo crea un'impronta ambientale considerevole e contribuisce ai costi operativi.

Alla luce di queste sfide, è fondamentale ridurre le richieste computazionali dei modelli di diffusione senza comprometterne le prestazioni. Il raggiungimento di questo obiettivo non solo renderebbe questi modelli più accessibili a un'ampia gamma di utenti e organizzazioni, soprattutto a quelle con risorse computazionali limitate, ma contribuirebbe anche a mitigare l'impatto ambientale.

2.3.2 Latent Diffusion Models

I modelli di diffusione latente (LDM), come suggerisce il nome, sono modelli di diffusione che, a differenza di quelli tradizionali, operano in uno spazio latente piuttosto che direttamente nello spazio dei pixel. Questo approccio è stato introdotto per la prima volta da Rombach et al. con l'obiettivo di ridurre la complessità computazionale e consentire l'addestramento di modelli probabilistici di diffusione (DPM) su sistemi con risorse computazionali limitate, pur mantenendo alta la qualità e la flessibilità.

L'idea chiave degli LDM prevede un processo in due fasi. Il primo passo è l'addestramento di un *autoencoder* per creare uno spazio di rappresentazione a dimensionalità ridotta che catturi fedelmente le caratteristiche essenziali dei dati. Questo spazio a bassa dimensionalità è equivalente allo spazio dei dati originali, ma con una complessità ridotta. Una volta addestrato l'*autoencoder*, il modello di diffusione viene addestrato su questo spazio latente. Durante questa fase, il rumore viene applicato alle codifiche delle immagini originali, che ora hanno una dimensione molto più piccola rispetto alle loro controparti a grandezza naturale.

Il campionamento in questo modello prevede l'estrazione del rumore gaussiano nello spazio del VAE in più passi seguito dalla decodifica del risultato dallo spazio latente allo spazio immagine. Questo processo ricostruisce l'immagine originale dalla rappresentazione latente [Rombach et al., 2022].

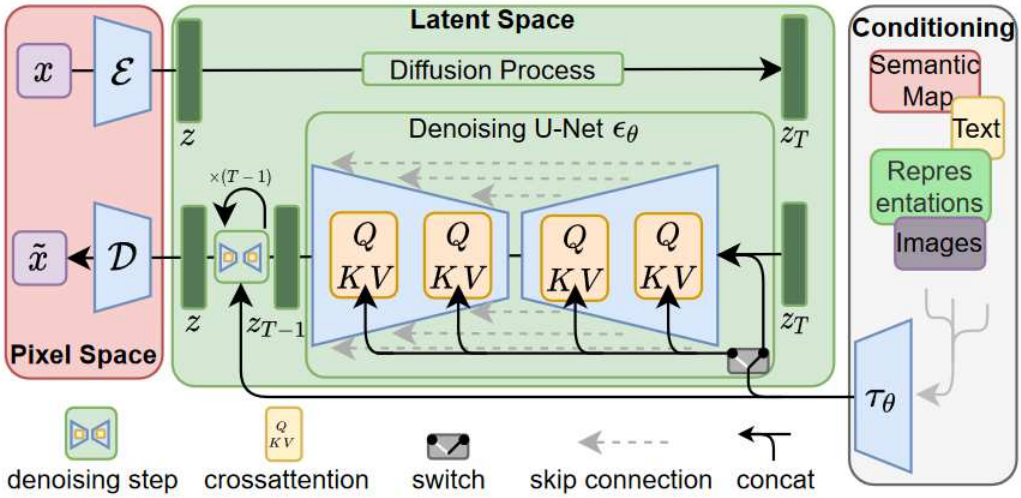


Figura 2.4: Esempio di un *Latent Diffusion Model* [Rombach et al., 2022]

Uno dei vantaggi più importanti di questo approccio è la riusabilità della fase di *autoencoder* universale. Poiché l'*autoencoder* deve essere addestrato una sola volta, può essere riutilizzato per più addestramenti DPM. Questo alleggerisce il processo e apre la possibilità di esplorare compiti completamente diversi senza dover riaddestrare l'*autoencoder* da zero ogni volta. Ciò consente non solo di risparmiare risorse computazionali, ma anche di accelerare la sperimentazione e l'applicazione dei modelli di diffusione in vari domini.

2.3.3 Architettura

La maggior parte dei modelli di diffusione attuali si basa sull'architettura U-Net [Ronneberger et al., 2015] con livelli di *self-attention* incorporati in quanto si è dimostrato efficace nel *denoising* di immagini. La U-Net è una rete neurale convoluzionale progettata per fornire segmentazioni precise anche con un numero limitato di immagini di addestramento.

L'architettura è composta da due parti principali: un percorso contrattivo, simile a una tipica rete convoluzionale, e un percorso espansivo, che esegue *up-convolutions* per aumentare la risoluzione spaziale e ridurre il numero di canali delle caratteristiche.

Per trasferire informazioni di basso livello dal percorso contrattivo agli strati corrispondenti del percorso espansivo vengono utilizzate connessioni di salto. Questo aiuta a preservare i dettagli importanti in tutta la rete.

Il percorso espansivo è più o meno simmetrico al percorso contrattivo, dando vita a un'architettura a U. La figura 2.5 illustra un esempio di questa architettura a U.

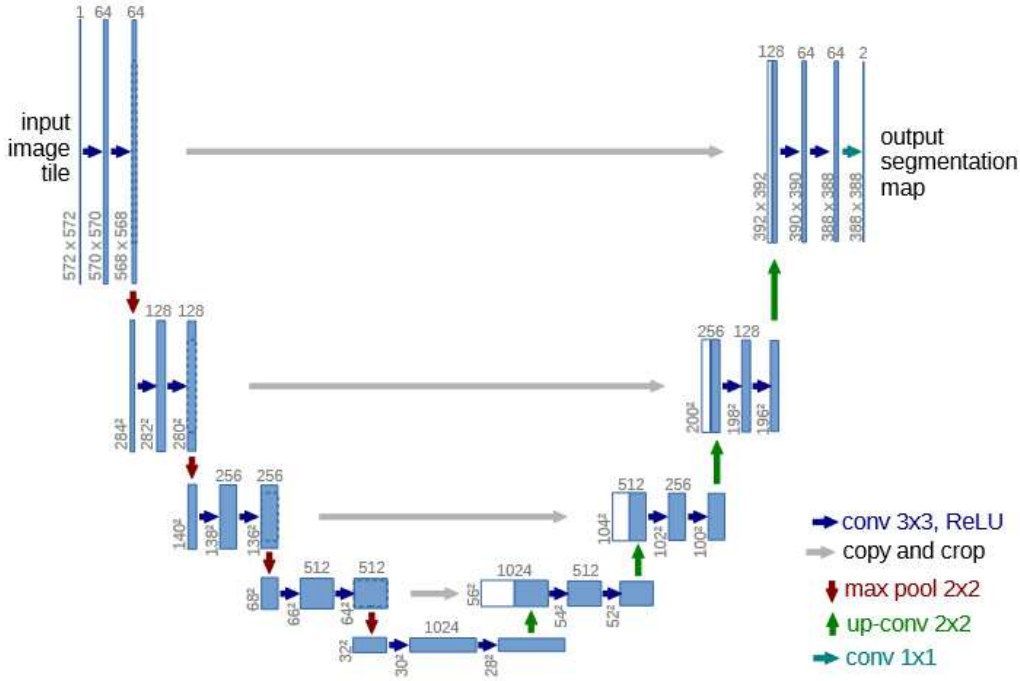


Figura 2.5: Esempio di una U-Net [Ronneberger et al., 2015]

2.4 Spettrogramma

Lo spettrogramma è una rappresentazione grafica dello spettro delle frequenze in funzione del tempo. Nell'asse delle ascisse è rappresentato il tempo, nell'asse delle ordinate la frequenza del suono in Hz, e il colore indica l'intensità del suono. Lo spettrogramma viene creato suddividendo il segnale audio in brevi segmenti, generalmente di pochi millisecondi, e calcolando la trasformata discreta di Fourier per ciascun segmento, al fine di ottenere il relativo spettro di frequenza. Gli

spettri risultanti vengono poi sovrapposti lungo l'asse del tempo per creare lo spettrogramma completo.

La capacità degli spettrogrammi di identificare visivamente note musicali, accordi e altri elementi musicali [Alm and Walker, 2002] li rendono molto utili per compiti di classificazione del genere musicale [Costa et al., 2011], mentre la rappresentazione grafica del suono può essere usata per determinare le proprietà acustiche dei suoni del parlato e per capire come vengono prodotti e percepiti i diversi elementi fonetici [Zue and Cole, 1979]).

2.4.1 Scala Mel

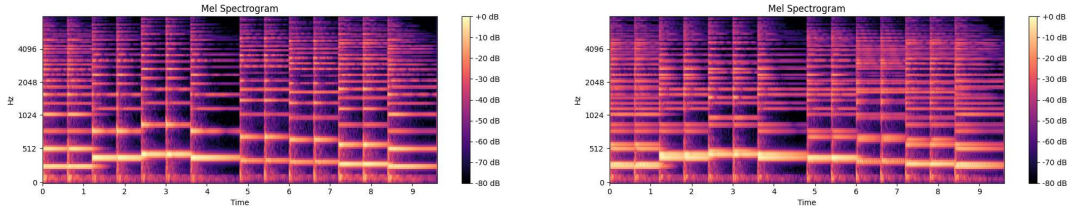
Una variante dello spettrogramma è lo spettrogramma Mel, che utilizza la scala Mel per misurare la frequenza. La scala Mel è stata sviluppata sperimentalmente per riflettere la sensibilità del sistema uditivo umano, che percepisce i suoni solo entro un determinato intervallo di frequenze. Questo rende lo spettrogramma Mel più adatto a catturare le caratteristiche del segnale audio più significative per la percezione umana, risultando particolarmente utile in compiti come il riconoscimento del parlato e l'identificazione del parlante [Stevens et al., 1937].

2.4.2 Spettrogrammi Mel e Musica

Nella tradizione della musica occidentale i suoni così chiamati consonanti vengono descritti come armoniosi, gradevoli e stabili; mentre i suoni dissonanti vengono descritti come spiacevoli e poco armoniosi.

La consonanza e la dissonanza possono essere identificate visibilmente su uno spettrogramma Mel analizzando le distribuzioni delle frequenze. Gli intervalli consonanti mostrano strutture armoniche, chiare e stabili, mentre gli intervalli dissonanti mostrano schemi più irregolari e meno stabili che sembrano lasciare un'ombra. I modelli di intelligenza artificiale possono apprendere questi schemi visivi allenandosi su ampie serie di dati musicali, consentendo loro di riconoscere e replicare le relazioni armoniche che producono suoni piacevoli e composizioni esteticamente più gradevoli.

Nella figura 2.6 sono rappresentati due spettrogrammi Mel dello stesso passaggio di 'Twinkle twinkle little star'; il primo passaggio è in ottava, quindi appare consonante, mentre il secondo è in seconda, apparendo dissonante.



(a) (b)

Figura 2.6: Spettrogrammi Mel di due passaggi della stessa melodia.

2.5 Modelli

In questa sezione verranno illustrati i quattro modelli *text-to-audio* scelti per questa tesi, verranno esposte le loro architetture e i *dataset* utilizzati per il loro addestramento, per offrire una comprensione completa di ciò che rende ciascuno di essi efficace.

Quando si tratta di sviluppare un potente modello *text-to-audio* (TTA), ci sono due obiettivi principali da raggiungere: innanzitutto, essere in grado di generare una vasta gamma di segnali audio ad alta definizione, come la voce umana, la musica e gli effetti sonori, creando suoni il più realistici possibile. In secondo luogo, assicurarsi che questi segnali audio seguano fedelmente il prompt testuale fornito.

Questi modelli sono stati selezionati perché hanno in comune l'architettura generale e il processo di generazione dell'audio, ma si differenziano per i set di dati e i *text encoders* che utilizzano. Questa combinazione di somiglianze e differenze offre un ricco terreno di confronto e mette in evidenza i punti di forza e di debolezza unici di ciascun approccio.

La generazione dell'audio tramite il processo di diffusione *fine-tuned* sugli spettrogrammi Mel comporta la conversione dei dati tra quattro *feature spaces*: audio, spettrogramma, pixel e spazio latente. Durante la fase di *training*, l'audio viene

trasformato in uno spettrogramma Mel e normalizzato per essere convertito in un'immagine RGB, che viene poi compressa nello spazio latente dal *variational autoencoder* (VAE). Il processo di diffusione opera in questo spazio latente, aggiungendo rumore gaussiano alle immagini degli spettrogrammi. Il modello viene poi addestrato per invertire questa trasformazione, recuperando la distribuzione originale dei dati a partire dal rumore.

Nella fase di inferenza, il processo viene invertito. Si inizia con il codificatore di prompt testuali, che codifica il prompt per la generazione del audio. Questa rappresentazione testuale viene poi utilizzata per costruire una rappresentazione latente dell'audio tramite la diffusione inversa. Il decodificatore del VAE costruisce quindi uno spettrogramma Mel a partire dalla rappresentazione latente. Infine, questo spettrogramma Mel viene inviato a un *vocoder* per generare l'audio finale. In questo modo, il modello impara a generare audio realistico e di alta qualità dalle descrizioni testuali.

I quattro modelli utilizzati in questo elaborato sono: Auffusion [Xue et al., 2024], Tango [Ghosal et al., 2023], Mustango [Melechovsky et al., 2024] e AudioLDM 2 [Liu et al., 2024].

2.5.1 Auffusion

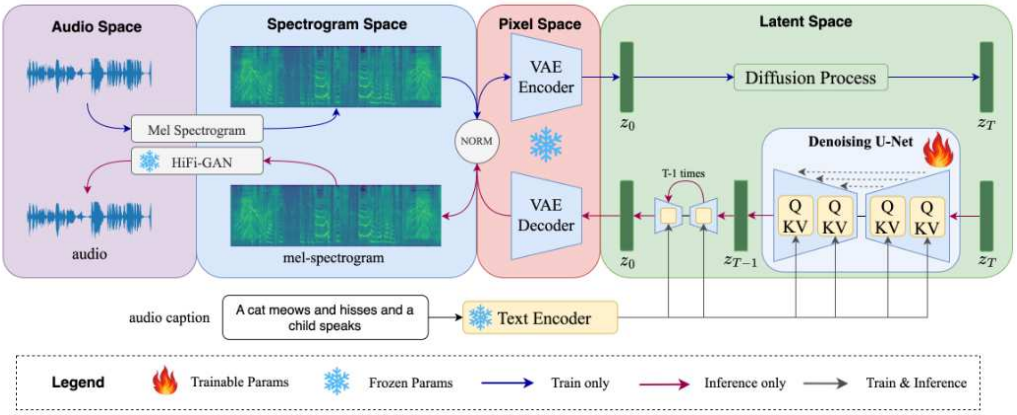


Figura 2.7: Architettura del modello Auffusion [Xue et al., 2024]

Affusion è stato introdotto nel gennaio del 2024 ed è quindi il modello più recente che verrà discusso. Auffusion impiega un processo semplice ed elegante per la generazione dell'audio, senza allontanarsi molto dalla tecnica sopra citata. Esso utilizza il *pretrained* Stable Diffusion v1.5, compresa la sua VAE, che è stata mantenuta congelata, e U-Net che è stata *fine-tuned* sul *dataset* audio, e utilizza il *vocoder* HiFi-Gan [Kong et al., 2020] per sintetizzare l'audio dallo spettrogramma Mel generato.

Auffusion integra i *text encoders Contrastive Language-Audio Pretraining* (CLAP) [Elizalde et al., 2022] e FLAN-T5 [Chung et al., 2022] per estrarre le informazioni chiave dalle descrizioni testuali e abbinarle con precisione all'audio desiderato. Secondo Elizalde et al. CLAP può categorizzare gli input in categorie non conosciute durante l'addestramento, eliminando così la necessità di addestrare e forzare un insieme predefinito di categorie e consentendo una previsione flessibile delle classi al momento dell'inferenza.

Per l'addestramento sono stati utilizzati diversi set di dati, tra cui AudioCaps (AC) [Kim et al., 2019], WavCaps [Mei et al., 2023], MACS [Morato and Mesaros, 2021], Clotho [Drossos et al., 2019], ESC50 [Piczak,], UrbanSound [Salamon et al., 2014], Music Instruments Dataset, e GTZAN [Tzanetakis and Cook, 2002]. I campioni audio del *training* set sono stati limitati a soli 10 secondi per motivi di

conformità; ciò si traduce in un set di dati contenenti campioni audio della durata totale di circa 2000 ore per l'addestramento del modello.

Le fonti principali dei campioni audio sono WavCaps, un *dataset* di audio estratti da varie fonti online ed etichettati con descrizioni rese più conformi utilizzando ChatGPT, e AudioCaps, un sottoinsieme del *dataset* AudioSet [Gemmeke et al., 2017] con etichette ad hoc contenente circa 46K clip audio.

AudioSet è un *dataset* di audio che consiste da oltre 2 milioni di clip video etichettati di 10 secondi. Circa la metà di tutti i campioni presenti nel *dataset* AudioSet sono etichettati come "musica"; tuttavia, questi clip sono stati raccolti da YouTube, quindi, molti di essi sono di scarsa qualità e contengono più fonti sonore. Per annotare questi dati viene utilizzata una raccolta di 632 classi, il che significa che lo stesso clip potrebbe essere annotato con etichette diverse.

Sebbene sia impossibile stabilire l'esatto rapporto tra i campioni audio musicali di alta qualità e quelli contenenti musica di bassa qualità o effetti sonori, è molto probabile che il numero di ore di campioni musicali puliti sia molto limitato.

2.5.2 Tango

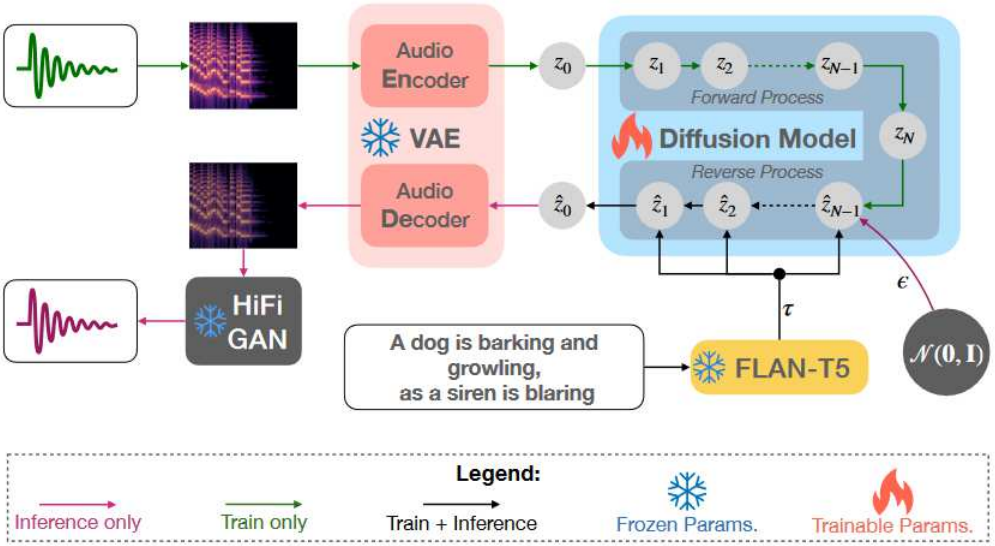


Figura 2.8: Architettura del modello Tango [Ghosal et al., 2023]

Tango è stato presentato per la prima volta nell'aprile del 2023. Sebbene condivida alcune somiglianze con gli altri modelli discussi, Tango ha le proprie caratteristiche che lo rendono unico.

Il suo modello di diffusione si basa sull'architettura U-Net di Stable Diffusion, incorpora il VAE di Liu et al. e utilizza HiFi-GAN come *vocoder* per sintetizzare l'audio dagli spettrogrammi Mel generati. A differenza di Auffusion, Tango utilizza esclusivamente il modello linguistico FLAN-T5 come codificatore di testo.

Per l'addestramento del LDM di Tango è stato utilizzato il *dataset* AudioCaps.

Per affrontare le sfide poste da questo set di dati relativamente piccolo, Tango utilizza tecniche di *data augmentation* per migliorare le sue prestazioni. Questo processo prevede la creazione di coppie testo-audio supplementari attraverso la fusione di clip audio esistenti e la unione delle loro etichette. Durante questa operazione, si tiene conto della percezione uditiva umana, prestando particolare attenzione al livello di pressione sonora garantendo che i campioni con livelli di pressione sonora più elevati non offuschino quelli con livelli più bassi. Questo attento bilanciamento aiuta a mantenere l'integrità e la chiarezza del nuovo set

di dati, generando segnali audio le cui relazioni sono coerenti con le descrizioni in linguaggio naturale.

2.5.3 Mustango

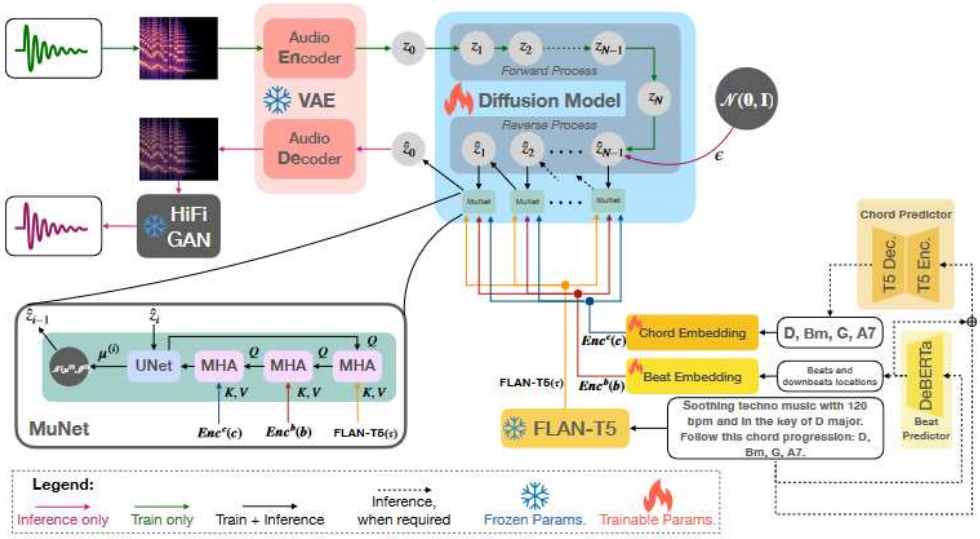


Figura 2.9: Architettura del modello Mustango [Melechovsky et al., 2024]

Mustango è stato sviluppato in parte dallo stesso team di ricerca che ha creato Tango e rilasciato nel novembre 2023. Questo modello è stato progettato specificamente per la generazione di musica, il che è evidente sia nella sua architettura che nel set di dati utilizzato per l'addestramento.

Pur impiegando un processo di generazione audio simile a quello di Auffusion e Tango, Mustango introduce modifiche significative alla sua architettura, rendendolo particolarmente adatto per la generazione di musica. Melechovsky et al. propongono una nuova variante di UNet: la *Music-Domain-Knowledge informed UNet (MuNet)*. La MuNet include due livelli aggiuntivi di *cross-attention* che migliorano la capacità del modello di identificare le informazioni rilevanti nei dati in ingresso, fornendo una comprensione più profonda del contesto musicale.

Questi vengono integrati da due codificatori creati ad hoc per estrarre informazioni cruciali sul ritmo e sull'accordo della musica dal prompt in ingresso. Oltre ai

due codificatori mustango usa anche FLAN-T5 per estrare le informazioni pertinente dai prompt in ingresso e DeBERTa [He et al., 2023] per estrare le informazioni riguardante il metro e i battiti per minuto.

Questa integrazione consente di catturare adeguatamente le caratteristiche musicali e di preservare diverse proprietà musicali fondamentali.

A causa della limitata disponibilità di set di dati di musica e descrizioni testuali accoppiati, è stato creato un nuovo set di dati chiamato MusicBench, che ha ampliato il set di dati MusicCaps [Agostinelli et al., 2023]. Il *dataset* MusicCaps contiene 5.479 campioni audio provenienti dal *dataset* AudioSet, ciascuno abbinato a etichette scritte da un esperto umano. Questi campioni contengono clip musicali di alta qualità in un'ampia gamma di generi.

Melechovsky et al. hanno scelto di non utilizzare il metodo per *data augmentation* applicato in Tango in quanto ciò poteva potenzialmente corrompere la struttura della musica, che è fortemente basata su tempo, tonalità e progressione degli accordi. Invece, sono state modificate la tonalità, il tempo e il volume dei campioni audio per creare 37.000 nuove variazioni. Le etichetti corrispondenti sono state poi riformulate con ChatGPT per riflettere questi cambiamenti. Questo approccio espande efficacemente il set di dati, mantenendo l'integrità della struttura intrinseca della musica.

2.5.4 AudioLDM 2

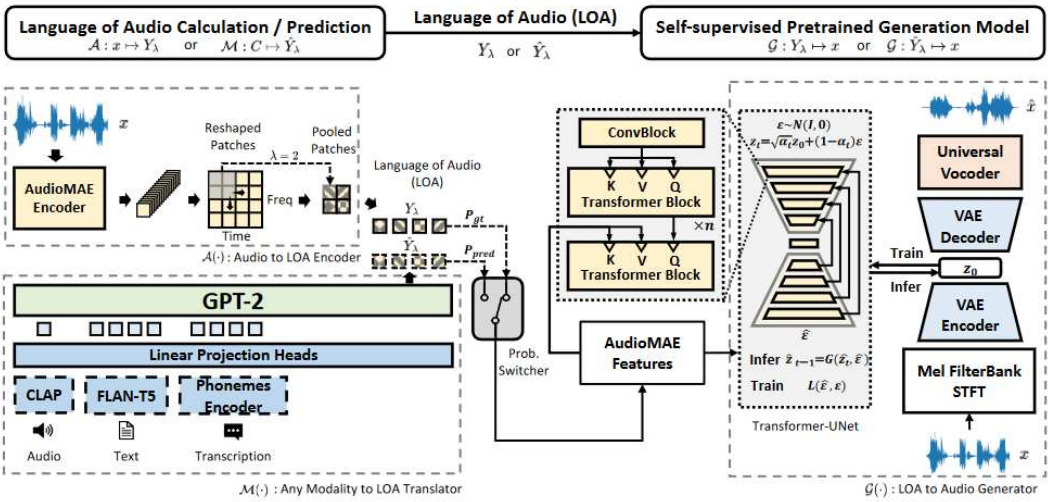


Figura 2.10: Architettura del modello AudioLDM 2 [Liu et al., 2024]

AudioLDM 2 è stato presentato per la prima volta nell'agosto 2023 da Liu et al. Questo modello è caratterizzato da un complesso processo di codifica che lo rende particolarmente abile nel generare un'ampia gamma di audio, tra cui la voce umana, effetti sonori e musica.

Ciò che distingue AudioLDM 2 è il suo approccio unico alla codifica delle informazioni semantiche, chiamato *Language of Audio (LOA)*. Il LOA è progettato per catturare sia i dettagli acustici tecnici sia le informazioni semantiche più ampie. Per ottenere questo risultato, il modello utilizza le caratteristiche estratte da un AudioMAE, un *autoencoder* audio, che apprende rappresentazioni da dati audio non etichettati.

AudioMAE inizia con la codifica di settori di spettrogrammi, utilizzando un elevato rapporto di copertura in modo che solo i token non coperti vengano elaborati dal codificatore. Durante il pre-addestramento auto-supervisionato gli spettrogrammi Mel sono coperti in modo casuale. Il decodificatore impara quindi a ricostruire questi settori coperti. Per completare il processo, il decodificatore riordina e decodifica il contesto codificato per ricostruire lo spettrogramma originale.

Per la generazione di audio e musica, le caratteristiche vengono estratte dal testo richiesto con la combinazione dei codificatore di testo CLAP e FLAN-T5.

I *dataset* utilizzati per l'addestramento di questo modello includono AudioSet, WavCaps, AudioCaps (AC), VGGSound [Chen et al., 2020], Free Music Archive (FMA) [Defferrard et al., 2017], Million Song Dataset (MSD) [Bertin-Mahieux et al., 2011], LJSpeech (LJS) [Ito and Johnson, 2017] e GigaSpeech [Chen et al., 2021].

FMA è un grande set di dati musicali non etichettati, contenente 100.000 brani musicali. Del Million Song Dataset è stato utilizzato solo il sottoinsieme etichettato, che contiene circa 510.000 brani musicali. LJSpeech è un set di dati contenente audio vocale con 13.000 brevi clip audio e trascrizioni dettagliate. GigaSpeech è un corpus di riconoscimento vocale inglese su larga scala con circa 10.000 ore di audio etichettati con trascrizioni.

	text encoders	*datasets*
$Auffusion$	Flan-T5 e Clap	AC, WavCaps, MACS, Clotho + altri
$Tango$	Flan-T5	AC
$Mustango$	Flan-T5, DeBERTa + due altri	MusicCaps
$AudioLDM\ 2$	Flan-T5, Clap e un Phoneme Encoder	AudioSet, WavCaps, AC, VGGSound, FMA, MSD, LJSpeech, e GigaSpeech

Tabella 2.1: Riassunto modelli

Capitolo 3

Metodologia

L'obiettivo di questo studio è valutare la qualità di campioni audio generati usando l'intelligenza artificiale attraverso il raccoglimento e l'analisi delle valutazioni di alcuni partecipanti su diverse metriche. Questa sezione illustra la metodologia utilizzata per condurre l'indagine e raccogliere i dati, fornendo una chiara comprensione delle procedure seguite. Nella sezione 3.1 verranno spiegati i strumenti utilizzati per la generazione dei campioni audio e del sondaggio; nella sezione 3.2 verrà spiegata la modalità di raccolta delle risposte.

3.1 Materiali

Lo studio ha utilizzato una raccolta di campioni audio generati con quattro diversi modelli basati sulla diffusione: Auffusion, Tango, Mustango e AudioLDM 2.

L'attenzione è stata principalmente rivolta all'efficacia dei modelli nel generare effetti sonori e musica. L'efficacia nel generare campioni di voce non è stata considerata, poiché non tutti i modelli scelti sono stati progettati per questo scopo.

I campioni audio sono stati creati utilizzando sei prompt - tre per effetti sonori e tre per composizioni musicali – facendo attenzione che ogni prompt producesse suoni distinti. Le richieste specifiche e le domande del sondaggio sono riportate in dettaglio nell'appendice A. In totale, sono stati prodotti ventiquattro campioni audio, ciascuno della durata di dieci secondi.

Per generare i campioni audio, sono state utilizzate due diverse strategie: l'uso delle API fornite dai creatori dei modelli sulla piattaforma Hugging Face Spaces per Tango, Mustango e AudioLDM, oppure il download del modello preaddestrato ed eseguito localmente per Auffusion.

Lo strumento di indagine è stato un questionario online creato con Google Forms, contenente domande di tipo qualitativo, concepito per valutare vari aspetti dei campioni audio. Questo approccio ha garantito un feedback completo sulla qualità dell'audio generato.

3.2 Raccolta delle risposte

Per raccogliere le valutazioni dei partecipanti sui campioni audio generati è stato utilizzato un approccio basato sul sondaggio. I partecipanti sono stati selezionati attraverso un campionamento di convenienza quindi facevano parte di un pubblico generico, senza che fosse richiesta o misurata una precedente esperienza nella creazione di musica.

Il sondaggio è stato progettato in due versioni identiche per quanto riguarda le domande, ma con i campioni audio presentati in un ordine diverso. Questo accorgimento è stato adottato per garantire che l'ordine di presentazione degli audio non influenzasse i risultati, permettendo così una valutazione più imparziale e accurata delle risposte dei partecipanti. I sondaggi sono stati proposti a demografiche simili e hanno avuto un numero simile di risposte, 20 per un sondaggio e 18 per l'altro, per un totale di 38 partecipanti.

Il sondaggio è stato redatto sia in italiano che in inglese al fine di raggiungere un gruppo demografico più ampio. Ogni partecipante poteva scegliere la lingua di sua preferenza.

I sei prompt si dividono in due categorie: effetti sonori e musica. Di questi, un solo prompt richiede la presenza di voci umane nello sfondo.

I partecipanti hanno completato il sondaggio online durante un periodo di un mese. Inizialmente, hanno ascoltato un clip audio di prova per assicurarsi che il loro metodo di ascolto preferito funzionasse correttamente. Poi, sono stati istruiti ad ascoltare ogni campione audio e a valutarlo in base alla qualità dell'audio, all'aderenza al prompt e al realismo utilizzando una scala Likert a cinque punti. Per

quest'ultima metrica, è stato specificato che per 'realistico' si intende la probabilità di confondere il campione audio generato con uno 'naturale', non creato da un modello di intelligenza artificiale.

L'ultima sezione del sondaggio comprendeva domande demografiche. Ai partecipanti è stato chiesto se avessero problemi di udito, in modo che le loro risposte potessero essere considerate in modo appropriato, e se si considerassero fan della musica alternativa.

Capitolo 4

Analisi dei dati

In questo capitolo verranno presentati i risultati del sondaggio e discusse alcune ipotesi che potrebbero spiegare tali risultati. Nella sezione 4.1 verranno esaminati i risultati pertinenti ai primi tre prompt per la generazione di effetti sonori; nella sezione 4.2 verranno esaminati i risultati pertinenti ai seguenti tre prompt per la generazione di musica; nella sezione 4.3 verranno delucidati alcuni dettagli sul gruppo campione scelto.

4.1 Effetti sonori

La tabella 4.1 mostra le medie delle valutazioni per diverse metriche relative tre prompt per la generazione di effetti sonori.

Dall'analisi risulta evidente che Mustango ha ricevuto le valutazioni più basse in tutte le categorie. Per la pertinenza al prompt, il 59% delle valutazioni ha ottenuto il punteggio minimo di 1, mentre per il realismo questa percentuale è stata del 63%. La metrica in cui Mustango ha ottenuto i risultati relativamente migliori è stata la qualità dell'audio; tuttavia, anche in questo caso, solo il 26% delle valutazioni erano pari o superiori a 3, suggerendo una generale insoddisfazione da parte dei partecipanti. Questi risultati indicano chiaramente che ci sono significativi margini di miglioramento per Mustango in tutte le aree analizzate.

Rispetto a Mustango, Tango ha ricevuto valutazioni discrete per la qualità dell'audio, con il 45% delle valutazioni superiori a 3. Questo suggerisce che i

partecipanti hanno trovato i suoni prodotti da Tango accettabili in termini di purezza e chiarezza. Tuttavia, Tango ha ottenuto valutazioni relativamente scarse per la pertinenza al prompt e per il realismo, con il 57% e il 64% rispettivamente inferiori o uguali a 3. Ciò indica che i campioni audio di Tango sono stati facilmente riconosciuti come generati da un modello artificiale.

	Tango			*Mustango*		
	Pertinenza	*Qualità*	*Realismo*	*Pertinenza*	*Qualità*	*Realismo*
Prompt 1	2,74 $\pm 1,11$	3,42 $\pm 1,00$	3,03 $\pm 1,10$	1,84 $\pm 1,03$	2,08 $\pm 0,97$	1,58 $\pm 0,83$
Prompt 2	4,34 $\pm 0,81$	3,39 $\pm 1,13$	3,63 $\pm 1,15$	1,71 $\pm 0,80$	2,16 $\pm 1,15$	1,82 $\pm 0,95$
Prompt 3	2,11 $\pm 0,98$	3,03 $\pm 0,97$	3,06 $\pm 1,11$	1,64 $\pm 1,02$	1,96 $\pm 0,91$	1,56 $\pm 0,89$

	Auffusion			*AudioLDM2*		
	Pertinenza	*Qualità*	*Realismo*	*Pertinenza*	*Qualità*	*Realismo*
Prompt 1	2,74 $\pm 1,11$	3,42 $\pm 1,00$	3,03 $\pm 1,10$	1,84 $\pm 1,03$	2,08 $\pm 0,97$	1,58 $\pm 0,83$
Prompt 2	4,34 $\pm 0,81$	3,39 $\pm 1,13$	3,63 $\pm 1,15$	1,71 $\pm 0,80$	2,16 $\pm 1,15$	1,82 $\pm 0,95$
Prompt 3	2,11 $\pm 0,98$	3,03 $\pm 0,97$	3,06 $\pm 1,11$	1,64 $\pm 1,02$	1,96 $\pm 0,91$	1,56 $\pm 0,89$

Tabella 4.1: Media delle valutazioni dei campioni audio contenenti effetti sonori

Auffusion e AudioLDM 2 hanno ottenuto valutazioni complessivamente simili, ma con alcune differenze degne di nota. Auffusion ha eccelso nelle valutazioni di pertinenza al prompt, dove quasi la metà delle valutazioni hanno ricevuto il punteggio massimo, indicando che i partecipanti ritenevano i campioni audio di Auffusion molto fedeli alle indicazioni fornite. In particolare, Auffusion ha ottenuto valutazioni ottime nella categoria del realismo, con il 61% delle valutazioni superiori a 3.

D'altro canto, AudioLDM 2 ha ottenuto valutazioni più alte nelle categorie di qualità dell'audio con il 69% delle valutazioni superiori a 3, rispetto al 54% di Auffusion. Questo suggerisce che i campioni di AudioLDM 2 sono stati percepiti come più naturali e meglio realizzati dal punto di vista tecnico. Inoltre, il 58% dei partecipanti hanno assegnato un punteggio superiore a 3 nella categoria del realismo per AudioLDM 2.

Le deviazioni standard variano da un minimo di 0,8 nella categoria concordanza al prompt per Mustango a un massimo di 1,27 nella categoria di realismo per Auffusion. In generale, i modelli, tranne Mustango, mostrano deviazioni standard più elevate nelle categorie di realismo, indicando un'incoerenza nelle valutazioni. Questa inconsistenza può essere attribuita alla vaghezza della metrica del realismo e al fatto che ogni persona ha una concezione diversa di cosa significhi per un suono essere realistico.

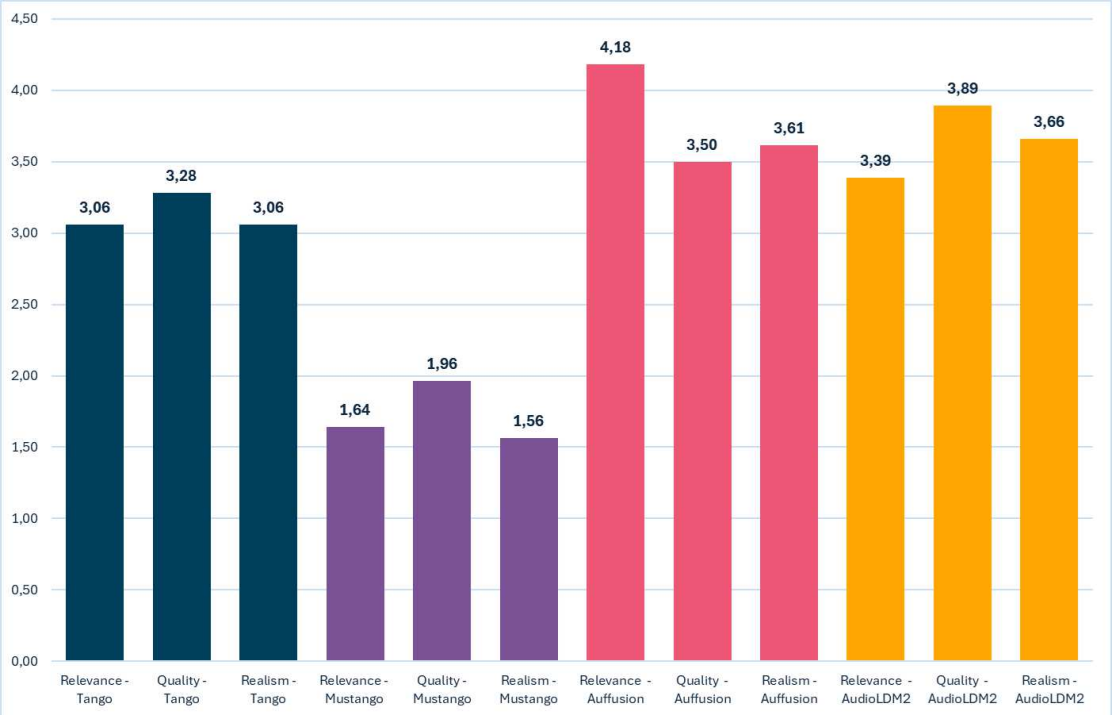


Figura 4.1: Media delle valutazioni per gli effetti sonori

La scarsa prestazione di Mustango può essere attribuita al fatto che il *dataset* utilizzato per l'addestramento di questo modello consiste esclusivamente di campioni musicali. Questo limite nel *dataset* ha probabilmente influenzato la capacità di Mustango di generare effetti sonori di alta qualità. Senza una varietà sufficiente di campioni, il modello non ha avuto l'opportunità di apprendere le caratteristiche diverse e complesse necessarie per produrre suoni realistici e pertinenti a prompt non musicali. Pertanto, la mancanza di diversità nel *dataset* di training è una spie-

gazione chiave per i risultati deludenti ottenuti da Mustango nelle varie metriche valutate.

Auffusion e AudioLDM2, invece, hanno utilizzato *dataset* molto variegati, che contenevano un gran numero di effetti sonori. Questa diversità ha reso l'apprendimento e la generazione di esempi simili molto più efficace e favorevole. Di conseguenza, questi modelli sono stati in grado di produrre suoni più realistici e pertinenti rispetto a Mustango, che aveva un *dataset* limitato solo a campioni musicali.

4.2 Musica

La tabella 4.2 mostra le medie delle valutazioni per le diverse metriche relative agli ultimi tre prompt, che riguardano la generazione di musica. Tango ha ottenuto i punteggi più bassi in tutte le categorie, con il 53% delle valutazioni per la pertinenza al prompt, il 67% per la qualità dell'audio e 50% per il realismo minori di 3.

	Tango			*Mustango*		
	Pertinenza	*Qualità*	*Realismo*	*Pertinenza*	*Qualità*	*Realismo*
Prompt 1	3,42 ±1,18	2,47 ±1,03	3,13 ±1,02	3,87 ±1,07	3,24 ±1,10	3,42 ±1,27
Prompt 2	2,32 ±1,14	2,00 ±1,04	2,29 ±1,27	3,45 ±1,13	2,89 ±1,29	2,76 ±1,28
Prompt 3	2,08 ±1,02	2,24 ±1,17	2,26 ±1,08	3,34 ±0,97	2,61 ±1,28	2,81 ±1,34

	Auffusion			*AudioLDM2*		
	Pertinenza	*Qualità*	*Realismo*	*Pertinenza*	*Qualità*	*Realismo*
Prompt 1	3,84 ±0,95	2,39 ±1,03	3,03 ±1,22	2,76 ±1,21	3,47 ±1,13	3,37 ±1,28
Prompt 2	3,50 ±1,20	2,97 ±1,44	2,71 ±1,45	4,55 ±0,86	4,13 ±1,04	4,24 ±1,02
Prompt 3	3,66 ±1,21	2,37 ±1,24	2,82 ±1,50	4,13 ±0,91	4,11 ±0,83	4,05 ±1,01

Tabella 4.2: Media delle valutazioni dei campioni audio contenenti musica

Queste tre categorie presentano medie di valutazione molto simili, intorno a 2,5, risultando comunque nettamente superiori ai punteggi di Mustango nella generazione di effetti sonori. Le scarse prestazioni di Tango possono essere attribuite

al *dataset* molto ridotto, composto principalmente da effetti sonori, utilizzato per il suo addestramento.

Le valutazioni di Mustango per la generazione di musica sono state positive, superando quelle di Auffusion nella qualità dell'audio e avvicinandosi molto ai risultati di Auffusion nelle altre due categorie. Il 60% delle valutazioni per la pertinenza al prompt ha ottenuto un punteggio superiore a 3, indicando che, nonostante le sue scarse prestazioni nella generazione di effetti sonori, Mustango ha dimostrato una certa competenza nella generazione musicale.

Auffusion continua ad ottenere risultati buoni nella categoria della concordanza con il prompt, con il 57% delle valutazioni superiori a 3. Tuttavia, Auffusion ha ottenuto risultati mediocri nella qualità dell'audio e realismo, con il 63% e 54% rispettivamente delle valutazioni inferiori a 3.

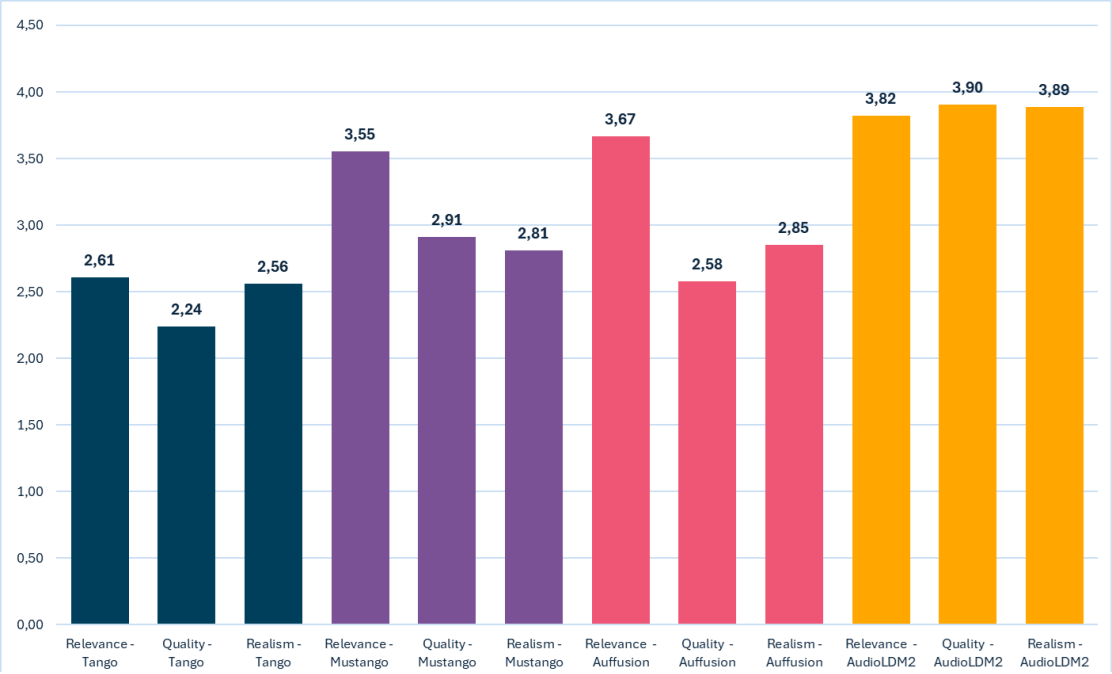


Figura 4.2: Media delle valutazioni per la musica

AudioLDM2 si è rivelato il miglior modello per la generazione di musica, ottenendo risultati eccellenti, soprattutto nella categoria del realismo, dove il 40% delle valutazioni hanno ricevuto il voto massimo. Inoltre, il 73% delle valutazioni per la qualità dell'audio hanno ricevuto un voto superiore a 3, una performan-

ce significativamente migliore rispetto agli altri modelli. Questa superiorità di AudioLDM2 suggerisce una capacità più avanzata di creare musica che sembra naturale e convincente.

Le valutazioni dei campioni musicali presentano in media una deviazione standard leggermente più elevata rispetto alle valutazioni degli effetti sonori per tutti i modelli. Questa discrepanza nelle valutazioni può essere attribuita alle differenze di competenza e conoscenza tra i partecipanti. Coloro che possiedono una formazione musicale o una conoscenza approfondita dei vari generi musicali potrebbero valutare gli stessi campioni in modo diverso rispetto a chi non ha alcuna formazione, aumentando così la deviazione standard delle valutazioni.

Va notato che il sondaggio non ha verificato il livello di familiarità dei partecipanti con i generi musicali utilizzati nei prompt. Un'altra spiegazione per la maggiore variazione nelle valutazioni potrebbe essere la variabilità delle preferenze musicali dei partecipanti.

È importante notare che i prompt utilizzati non sono stati ottimizzati per Mustango, poiché non contengono informazioni sul ritmo e sull'accordatura che, se presenti, possono essere estratte dai codificatori appositamente creati. Di conseguenza, l'architettura di Mustango non è stata sfruttata al meglio delle sue capacità. La ragione di questa scelta risiede nel mantenimento dell'uniformità, poiché gli altri modelli non dispongono di una struttura progettata per l'estrazione di tali informazioni, il che li avrebbe posti in una situazione di svantaggio.

I modelli Auffusion e Tango si differenziano principalmente per l'uso di codificatori di testo: Auffusion utilizza sia CLAP che Flan-T5, mentre Tango impiega solo Flan-T5. Inoltre, i due modelli sono stati addestrati su dataset diversi: Auffusion ha utilizzato un assortimento di dataset, impiegando quindi una quantità di dati molto maggiore, mentre Tango ha fatto uso di un dataset più ridotto, ricorrendo a tecniche di *data augmentation* per ampliarlo.

Le valutazioni superiori di Auffusion nelle categorie di qualità del suono e realismo dimostrano che un dataset più diversificato favorisce la generazione di contenuti nuovi di alta qualità. Le valutazioni inferiori di Tango nella categoria

di pertinenza al prompt indicano che l'uso del solo FLAN-T5 come codificatore di testo potrebbe non essere sufficiente per garantire prestazioni ottimali.

4.3 Analisi demografica

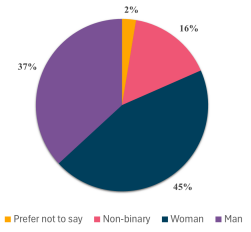


Figura 4.3: Distribuzione dei generi del gruppo campione

Il sondaggio ha ricevuto risposte da un gruppo di partecipanti abbastanza diverso da garantire una distribuzione equa per quanto riguarda il genere: il 37% degli intervistati erano uomini, il 45% donne e il 16% persone non binarie. Questa distribuzione equilibrata ha permesso di ottenere un campione rappresentativo e diversificato, assicurando che le opinioni e i feedback riflettessero una varietà di prospettive.

Come descritto nel capitolo 3, l'indagine è stata eseguita usando due sondaggi con differenze minime nell'ordine dei campioni audio. È importante notare che uno dei sondaggi ha ricevuto sistematicamente valutazioni più alti, suggerendo un pregiudizio sistematico piuttosto che un evento casuale. La questione, tuttavia, va oltre la portata di questo lavoro.

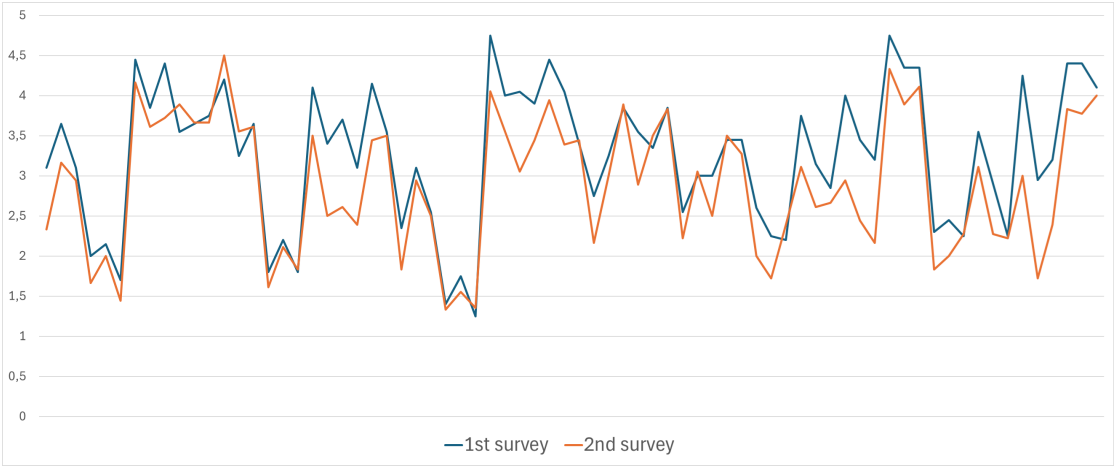


Figura 4.4: Valutazioni dei due sondaggi

Capitolo 5

Conclusioni

L'obiettivo primario di questa tesi è stato quello di analizzare le capacità dei modelli di intelligenza artificiale basati sul processo di diffusione nel sintetizzare campioni audio nuovi e creativi. Un obiettivo secondario è stato quello di valutare l'importanza dell'architettura del modello e dei *dataset* utilizzati per l'addestramento nello sviluppo di un modello efficace. Esplorando questi aspetti, la ricerca intende contribuire al crescente campo della sintesi dell'audio usando l'intelligenza artificiale, offrendo una nuova prospettiva su come le diverse architetture e risorse influiscono sulla qualità e sulla creatività dei campioni audio generati.

Nella sezione iniziale della tesi sono state esaminate le varie componenti di un modello basato sulla diffusione, fornendo una panoramica del processo di diffusione stesso e i principi matematici coinvolti nel loro funzionamento. Sono stati, inoltre, illustrati i vantaggi dell'addestramento del modello sugli spettrogrammi, discutendo di come questo approccio aumenti la capacità del modello di catturare e replicare gli intricati dettagli dei segnali audio.

Successivamente, è stata fornita un'analisi dettagliata dei quattro modelli selezionati per questo studio. Ogni modello è stato valutato per evidenziarne le somiglianze e le differenze, offrendo una prospettiva sulla loro progettazione e implementazione. L'analisi ha incluso una discussione delle scelte strutturali fatte per ciascun modello, come le configurazioni e i parametri specifici utilizzati. Inoltre, abbiamo esaminato i set di dati utilizzati per l'addestramento, valutando l'impatto di diversi tipi e dimensioni di set di dati sulle prestazioni e sulla creatività dei

campioni audio generati dall'intelligenza artificiale.

Per determinare quale dei modelli selezionati fosse il migliore nella generazione di audio, è stata condotta un'indagine basata su un sondaggio. Ai partecipanti è stato chiesto di ascoltare vari campioni audio generati dall'IA e di valutarli in base a tre metriche soggettive.

I risultati hanno indicato che il modello AudioLDM 2 ha fornito le migliori prestazioni nella generazione di effetti sonori e musica. Questo modello è stato seguito da vicino da Auffusion per la generazione di effetti sonori e da Mustango per la musica. Al contrario, Tango ha dimostrato prestazioni mediocri nella generazione di effetti sonori e scarse nella generazione di musica. La disparità di risultati tra Auffusion e Tango, nonostante le architetture simili, suggerisce che l'uso di Flan-T5 come codificatore da solo non è sufficiente per ottenere risultati soddisfacenti.

Nel complesso, le prestazioni superiori di AudioLDM 2 possono essere attribuite alla sua complessa struttura interna e all'addestramento su un set di dati molto ampio e diversificato, comprendente vari tipi di suoni. Questa combinazione di elementi consente ad AudioLDM2 di produrre campioni audio di qualità superiore e più creativi.

Le sfide principali incontrate in questo studio sono state il gruppo campione piccolo e il numero limitato di campioni audio generati utilizzati per la valutazione. Queste limitazioni possono aver prodotto risultati poco affidabili. Per superare queste limitazioni, le ricerche future potrebbero trarre vantaggio dell'uso di un campione più ampio e dall'inclusione di partecipanti con una conoscenza pregressa della musica e del suono. Ciò non solo aumenterebbe l'affidabilità dei risultati, ma fornirebbe anche valutazioni più approfondite sull'audio generato dall'intelligenza artificiale.

Inoltre, si potrebbe eseguire uno studio simile per esaminare gli effetti della dimensione e della diversità dei *dataset* sulle prestazioni del modello. Ciò potrebbe comportare l'addestramento di un singolo modello su insiemi di dati di dimensioni e livelli di diversità vari per osservare come questi fattori influenzino la qualità dell'audio generato. Inoltre, si potrebbero addestrare modelli diversi sugli stessi insiemi di dati per confrontare le loro prestazioni. Questo approccio consentirebbe

di chiarire le intricate relazioni tra le caratteristiche dei set di dati, l'architettura del modello e la qualità dell'audio risultante, fornendo così una visione più chiara dei fattori determinanti per il successo della sintesi audio usando l'intelligenza artificiale.

Bibliografia

[Agostinelli et al., 2023] Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., and Frank, C. (2023). Musiclm: Generating music from text.

[Alm and Walker, 2002] Alm, J. and Walker, J. (2002). Time-frequency analysis of musical instruments. *Society for Industrial and Applied Mathematics*, 44:457–476.

[Bank et al., 2021] Bank, D., Koenigstein, N., and Giryes, R. (2021). Autoencoders.

[Bergmann and Stryker, 2023] Bergmann, D. and Stryker, C. (2023). Cos'è un autoencoder? URL: `https://www.ibm.com/it-it/topics/autoencoder`, Accessed: 2024-06-27.

[Bertin-Mahieux et al., 2011] Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*.

[Chen et al., 2021] Chen, G., Chai, S., Wang, G., Du, J., Zhang, W.-Q., Weng, C., Su, D., Povey, D., Trmal, J., Zhang, J., Jin, M., Khudanpur, S., Watanabe, S., Zhao, S., Zou, W., Li, X., Yao, X., Wang, Y., Wang, Y., You, Z., and Yan, Z. (2021). Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio.

[Chen et al., 2020] Chen, H., Xie, W., Vedaldi, A., and Zisserman, A. (2020). Vggsound: A large-scale audio-visual dataset.

[Chung et al., 2022] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. (2022). Scaling instruction-finetuned language models.

[Costa et al., 2011] Costa, Y., Soares de Oliveira, L., Koericb, A., and Gouyon, F. (2011). Music genre recognition using spectrograms. pages 1 – 4.

[Defferrard et al., 2017] Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X. (2017). Fma: A dataset for music analysis.

[Drossos et al., 2019] Drossos, K., Lipping, S., and Virtanen, T. (2019). Clotho: An audio captioning dataset.

[Elizalde et al., 2022] Elizalde, B., Deshmukh, S., Ismail, M. A., and Wang, H. (2022). Clap: Learning audio concepts from natural language supervision.

[Gemmeke et al., 2017] Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA.

[Ghosal et al., 2023] Ghosal, D., Majumder, N., Mehrish, A., and Poria, S. (2023). Text-to-audio generation using instruction-tuned llm and latent diffusion model.

[He et al., 2023] He, P., Gao, J., and Chen, W. (2023). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

[Ho et al., 2020] Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models.

[Ito and Johnson, 2017] Ito, K. and Johnson, L. (2017). The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

[Kim et al., 2019] Kim, C. D., Kim, B., Lee, H., and Kim, G. (2019). Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*.

[Kingma and Welling, 2022] Kingma, D. P. and Welling, M. (2022). Auto-encoding variational bayes.

[Kong et al., 2020] Kong, J., Kim, J., and Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis.

[Liu et al., 2024] Liu, H., Yuan, Y., Liu, X., Mei, X., Kong, Q., Tian, Q., Wang, Y., Wang, W., Wang, Y., and Plumbley, M. D. (2024). Audioldm 2: Learning holistic audio generation with self-supervised pretraining.

[Luccioni et al., 2024] Luccioni, S., Jernite, Y., and Strubell, E. (2024). Power hungry processing: Watts driving the cost of ai deployment? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24. ACM.

[Makhzani and Frey, 2014] Makhzani, A. and Frey, B. (2014). k-sparse autoencoders.

[Maslej et al., 2024] Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., and Clark, J. (2024). Artificial intelligence index report 2024. pages 242–245.

[Mei et al., 2023] Mei, X., Meng, C., Liu, H., Kong, Q., Ko, T., Zhao, C., Plumbley, M. D., Zou, Y., and Wang, W. (2023). Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv preprint arXiv:2303.17395*.

[Melechovsky et al., 2024] Melechovsky, J., Guo, Z., Ghosal, D., Majumder, N., Herremans, D., and Poria, S. (2024). Mustango: Toward controllable text-to-music generation.

[Morato and Mesaros, 2021] Morato, I. M. and Mesaros, A. (2021).

[Nichol and Dhariwal, 2021] Nichol, A. and Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. *CoRR*, abs/2102.09672.

[Piczak,] Piczak, K. J. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.

[Rombach et al., 2022] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation.

[Salamon et al., 2014] Salamon, J., Jacoby, C., and Bello, J. P. (2014). A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, page 1041–1044, New York, NY, USA. Association for Computing Machinery.

[Sohl-Dickstein et al., 2015] Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics.

[Stevens et al., 1937] Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190.

[Tzanetakis and Cook, 2002] Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302.

[Villalobos and Ho, 2022] Villalobos, P. and Ho, A. (2022). Trends in training dataset sizes. URL: `https://epochai.org/blog/trends-in-training-dataset-sizes`, Accessed: 2024-06-27.

[Vincent et al., 2008] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. pages 1096–1103.

[Weng, 2021] Weng, L. (2021). What are diffusion models? *lilianweng.github.io*.

[Xue et al., 2024] Xue, J., Deng, Y., Gao, Y., and Li, Y. (2024). Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation.

[Zue and Cole, 1979] Zue, V. and Cole, R. (1979). Experiments on spectrogram reading. volume 4, pages 116 – 119.

Appendice A

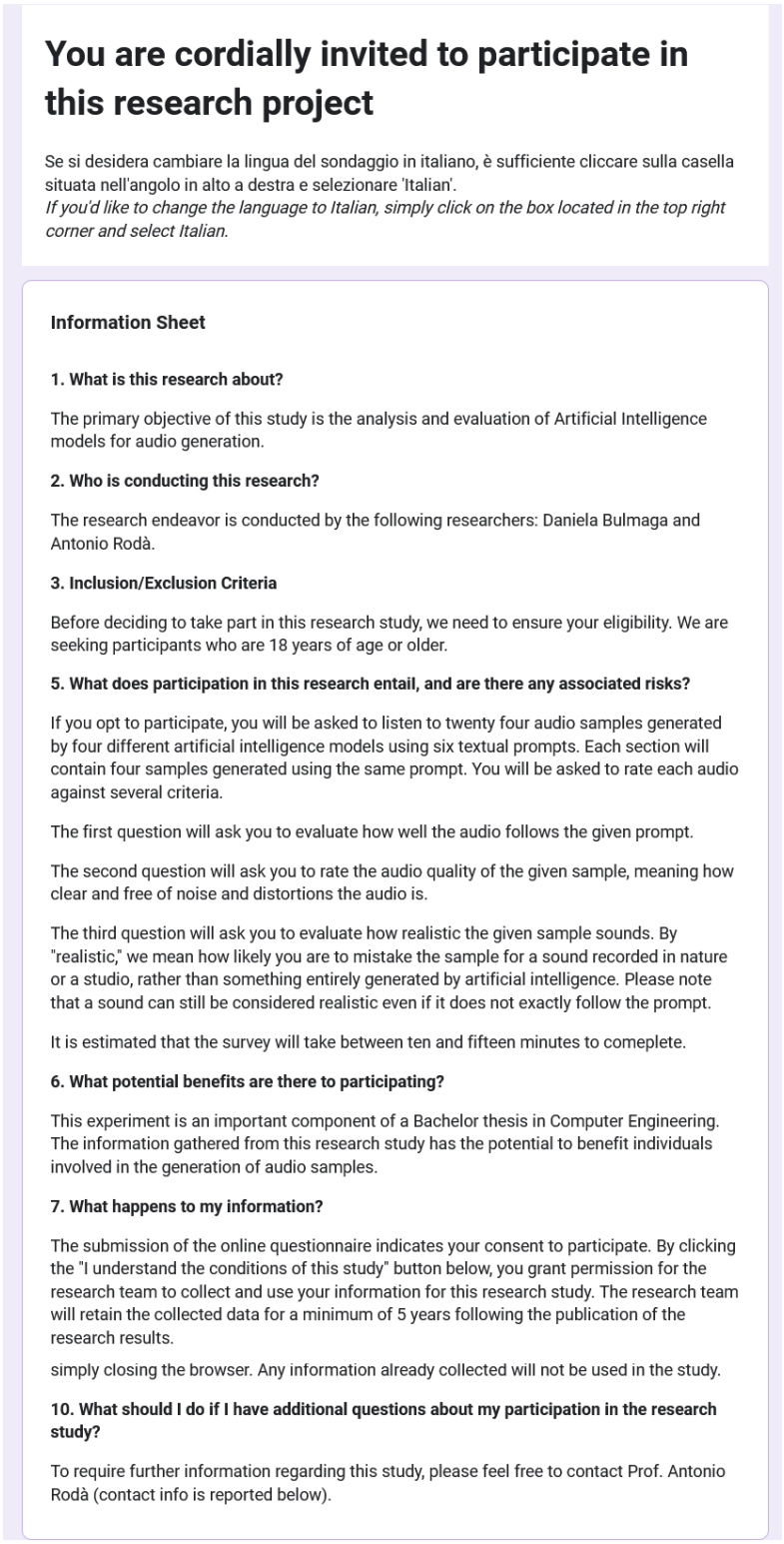
Informazioni sul sondaggio

	Italiano	*Inglese*
Prompt 1	Metallo che sbatte contro metallo in una grande stanza	Metal crashing against metal in a big room
Prompt 2	Cibo che frigge in una cucina piena e gente che parla	Food frying in a full kitchen and people talking
Prompt 3	Il ruggito di un leone e poi uccelli che volano via	The roar of a lion and then birds flying away
Prompt 4	Un'introduzione pianistica delicata e serena per una sonata	A delicate and serene piano introduction for a sonata
Prompt 5	Un potente riff di basso con la batteria in sottofondo	A powerful bass riff with drums in the background
Prompt 6	Disco pop anni '80, melodico, orecchiabile, ritornello con tonalità in minore	Driving 1980s Disco pop, melodic, catchy, chorus in minor

Tabella A.1: Prompt utilizzati per il sondaggio

	Italiano	*Inglese*
Domanda 1	Quanto è pertinente il secondo audio al prompt?	How relevant is the first audio to the prompt?
Domanda 2	Valutare il secondo audio in termini di qualità complessiva (l'audio è chiaro e nitido, senza rumori o distorsioni)	Evaluate the first audio in terms of overall audio qualit (the audio is clear and crisp, without noise or sound distortion)
Domanda 3	Quanto realistico (non generato artificialmente) vi sembra il secondo audio?	How realistic (non artificially generated) does the first audio sound?

Tabella A.2: Domande poste per ogni campione audio



You are cordially invited to participate in this research project

Se si desidera cambiare la lingua del sondaggio in italiano, è sufficiente cliccare sulla casella situata nell'angolo in alto a destra e selezionare 'Italian'.
If you'd like to change the language to Italian, simply click on the box located in the top right corner and select Italian.

Information Sheet

1. What is this research about?

The primary objective of this study is the analysis and evaluation of Artificial Intelligence models for audio generation.

2. Who is conducting this research?

The research endeavor is conducted by the following researchers: Daniela Bulmaga and Antonio Rodà.

3. Inclusion/Exclusion Criteria

Before deciding to take part in this research study, we need to ensure your eligibility. We are seeking participants who are 18 years of age or older.

5. What does participation in this research entail, and are there any associated risks?

If you opt to participate, you will be asked to listen to twenty four audio samples generated by four different artificial intelligence models using six textual prompts. Each section will contain four samples generated using the same prompt. You will be asked to rate each audio against several criteria.

The first question will ask you to evaluate how well the audio follows the given prompt.

The second question will ask you to rate the audio quality of the given sample, meaning how clear and free of noise and distortions the audio is.

The third question will ask you to evaluate how realistic the given sample sounds. By "realistic," we mean how likely you are to mistake the sample for a sound recorded in nature or a studio, rather than something entirely generated by artificial intelligence. Please note that a sound can still be considered realistic even if it does not exactly follow the prompt.

It is estimated that the survey will take between ten and fifteen minutes to comeplete.

6. What potential benefits are there to participating?

This experiment is an important component of a Bachelor thesis in Computer Engineering. The information gathered from this research study has the potential to benefit individuals involved in the generation of audio samples.

7. What happens to my information?

The submission of the online questionnaire indicates your consent to participate. By clicking the "I understand the conditions of this study" button below, you grant permission for the research team to collect and use your information for this research study. The research team will retain the collected data for a minimum of 5 years following the publication of the research results.

simply closing the browser. Any information already collected will not be used in the study.

10. What should I do if I have additional questions about my participation in the research study?

To require further information regarding this study, please feel free to contact Prof. Antonio Rodà (contact info is reported below).

Figura A.1: Pagina informativa del sondaggio

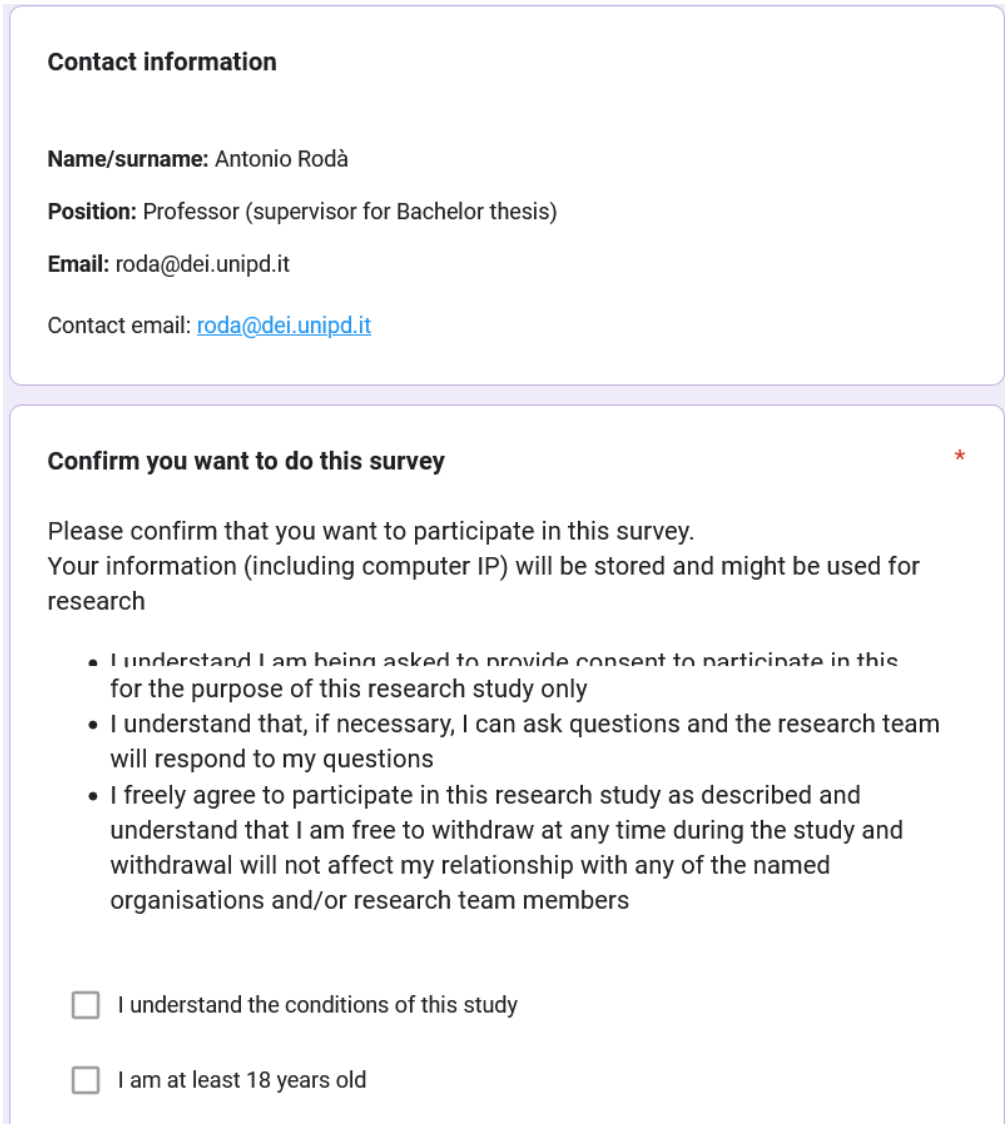


Figura A.2: Continuazione della pagina informativa

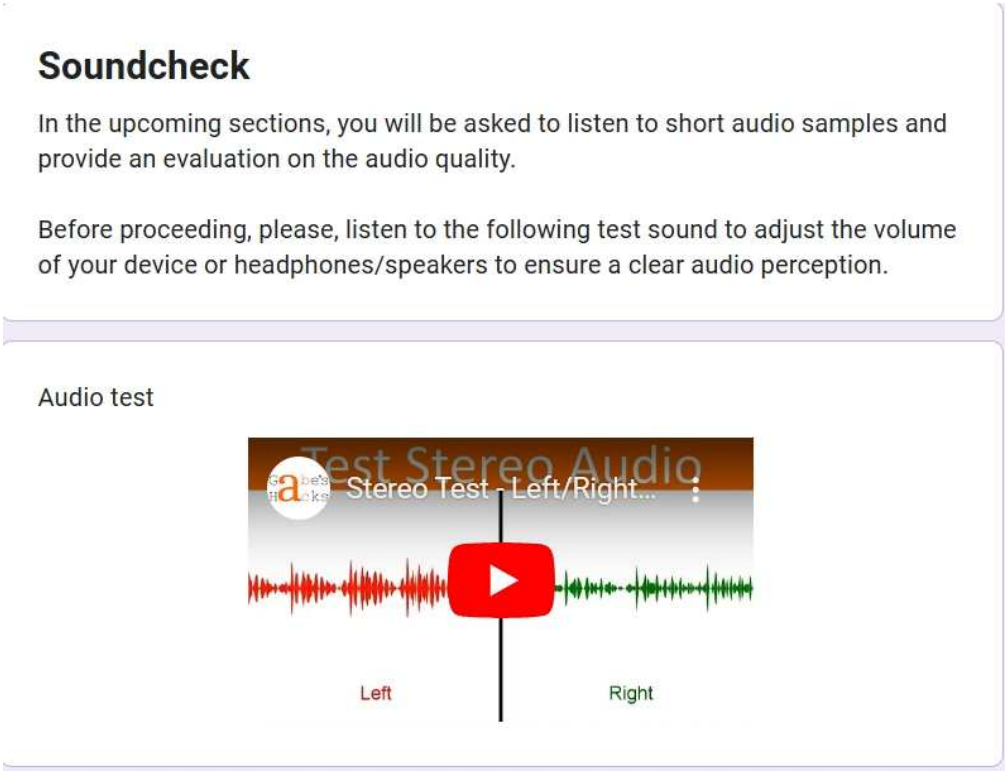


Figura A.3: Prova del suono eseguita dai partecipanti prima di proseguire con la valutazione dei campioni audio generati

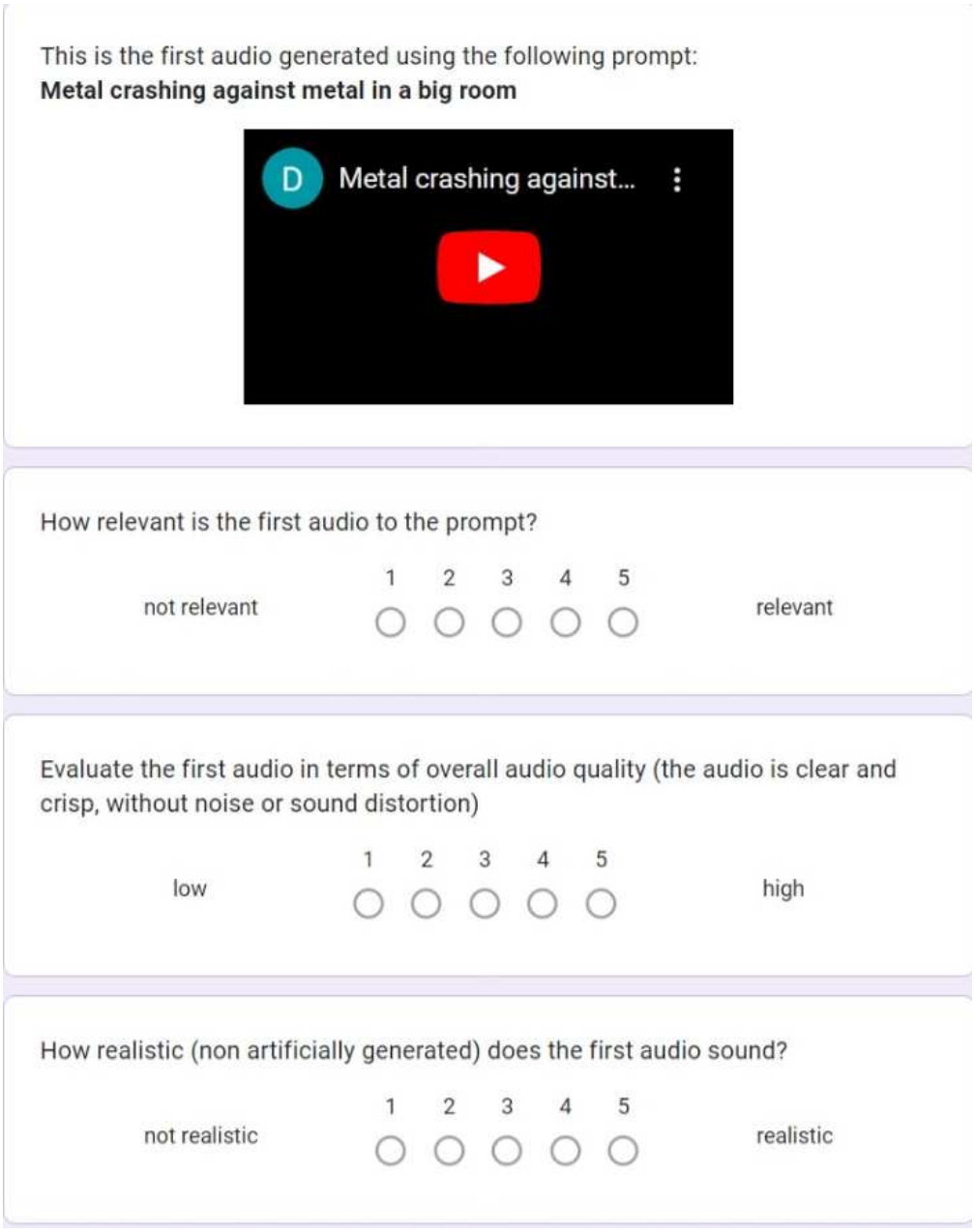


Figura A.4: Esempio di un clip audio e le domande poste ai partecipanti

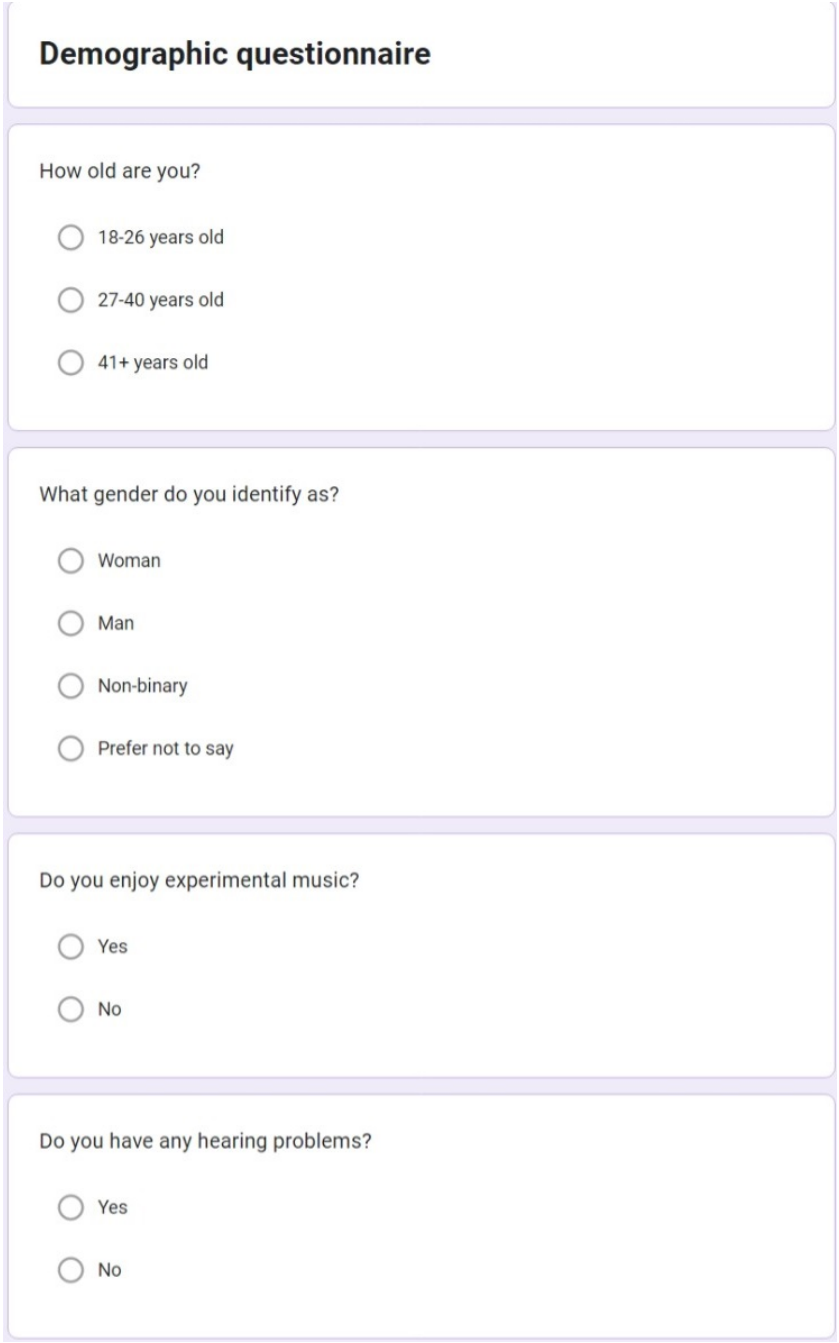


Figura A.5: Domande demografiche poste ai partecipanti