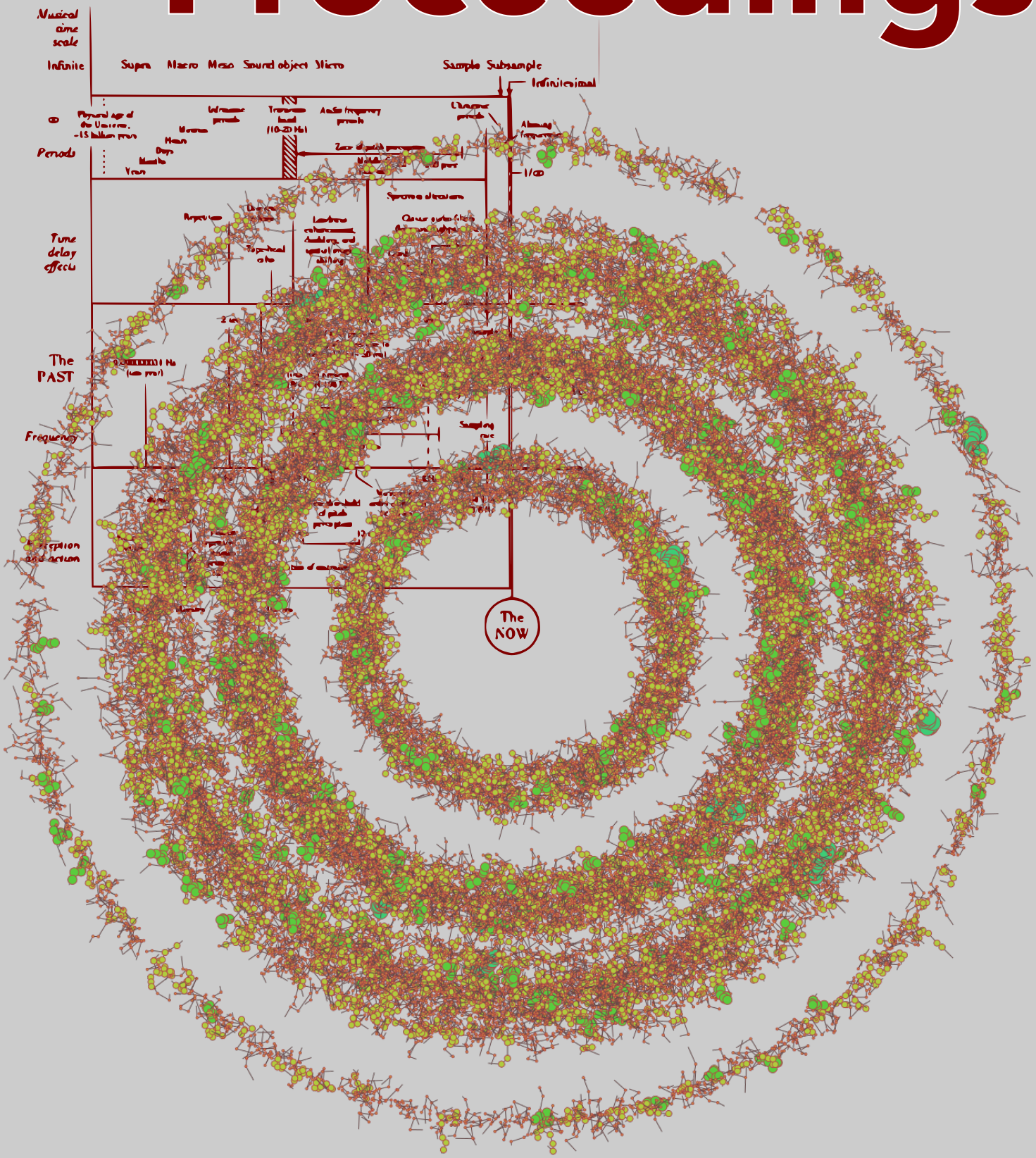
XXIV CIM

Colloquio di Informatica Musicale/Colloquium on Music Informatics

Memorie proiettive/Projecting Memories

Torino, 30/09 - 02/10 2024

Atti
Proceedings



XXIV CIM – Colloquio di Informatica Musicale
24th CIM – Colloquium on Music Informatics

Atti della Conferenza
Conference Proceedings

Torino, 30 Settembre – 2 Ottobre 2024
Torino, September 30th – October 2nd 2024

Davide Andrea Mauro, Simone Spagnol and Andrea Valle, a cura di/*eds.*

AA.VV.
Memorie proiettive/Projecting Memories

XXIV Colloquio di Informatica Musicale
24th Colloquium on Music Informatics

A cura di/eds.:
Davide Andrea Mauro, Simone Spagnol, Andrea Valle

ISBN: 9788890341373
ISSN: 2611-7355
Per gentile collaborazione di Davide Rocchesso e Stefano Delle Monache
Courtesy of Davide Rocchesso and Stefano Delle Monache

Pubblicato da/publisher:
DADI - Dipartimento Arti e Design Industriale, Università IUAV di Venezia

In co-edizione con/co-published with:
CIRMA – Università degli Studi di Torino
AIMI – Associazione Informatica Musicale Italiana
`https://aimi-musica.org/`

BibTeX:

```
@proceedings{24CIMConf,
Editor = {Davide Andrea Mauro, Simone Spagnol and Andrea Valle},
Organization = {AIMI - Associazione Informatica Musicale Italiana},
Publisher = {DADI - Dip. Arti e Design Industriale. Università IUAV di Venezia},
Title = {Memorie proiettive/Projecting Memories. Atti del {XXIV} Colloquio di Informatica Musicale/
Proceedings of the {XXIV} Colloquium on Music Informatics},
Year = {2024}}
```


Cover drawing: algorithmic radial arborescences generated with the Shoebot Python Library (`http://shoebot.net/`) by Andrea Valle
Typeset with ConTEXt (`http://wiki.contextgarden.net/`) by Andrea Valle

With the support of:

UNIVERSITÀ DI TORINO

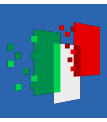
CONSERVATORIO STATALE DI MUSICA GIUSEPPE VERDI TORINO

TO LISTEN TO

SMET Scuola di Musica Elettronica del Conservatorio di Torino

ART IN MED
L'Arte nella divulgazione delle Scienze Mediche

Ministero dell'Università e della Ricerca



UNIVERSITÀ DI TORINO

General chairs
Stefano Bassanese, Conservatorio di Torino
Andrea Valle, Università di Torino

Scientific chairs
Davide Andrea Mauro, Paderborn University
Simone Spagnol, Università IUAV di Venezia
Andrea Valle, Università di Torino

Music chairs
Andrea Agostini, Conservatorio di Torino
Stefano Bassanese, Conservatorio di Torino
Daniele Ghisi, Conservatorio di Torino

Program committee
Andrea Agostini, Conservatorio di Torino
Giacomo Albert, Università di Torino
Carlo Barbagallo, Conservatorio di Torino
Stefano Bassanese, Conservatorio di Torino
Francesco Bianchi, Conservatorio di Torino
Leonardo Gabrielli, Università Politecnica delle Marche
Daniele Ghisi, Conservatorio di Torino
Francesco Giomi, Tempo Reale
Luca Guidarini, Ensemble Collettivo 21
Davide Andrea Mauro, Paderborn University
Simone Spagnol, Università IUAV di Venezia
Anna Terzaroli, Conservatorio di Adria
Luca Turchet, Università di Trento
Andrea Valle, Università di Torino
Paolo Zavagna, Conservatorio di Venezia

Scientific Reviewers
Giacomo Albert, Università di Torino
Adriano Baratè, Università di Milano
Alessio Brutti, FBK - Fondazione Bruno Kessler
Luca Comanducci, Politecnico di Milano
Stefano D'Angelo, dangelo.audio
Stefano Delle Monache, IRCAM
Michele Ducceschi, Università di Bologna
Davide Fantini, Università di Milano
Vanessa Faschi, Università di Milano
Federico Fontana, Università di Udine
Leonardo Gabrielli, Università Politecnica delle Marche
Michele Geronazzo, Università di Padova
Riccardo Giampiccolo, Politecnico di Milano
Luca Andrea Ludovico, Università di Milano
Nicola Orio, Università di Padova
Giorgio Presti, Università di Milano
Niccolò Pretto, Libera Università di Bolzano
Alessandro Giuseppe Privitera, Università di Udine
Davide Rocchesso, Università di Milano
Antonio Rodà, Università di Padova
Cristina Rottondi, Politecnico di Torino
Sylviane Sapir, Conservatorio di Milano
Stefania Serafin, Aalborg University
Giuseppe Silvi, Conservatorio di Bari
Federico Simonetta, Gran Sasso Science Institute
Domenico Stefani, Università di Trento
Anna Terzaroli, Conservatorio di Adria
Marco Tiraboschi, Università di Milano
Gualtiero Volpe, Università di Genova
Stefano Zambon, MIND Music Labs
Victor Zappi, Northeastern University
Laura Zattra, IRCAM / Conservatorio di Rovigo e Vicenza
Paolo Zavagna, Conservatorio di Venezia

Music Reviewers
Alessandro Anatrini, Hochschule für Musik und Theater Hamburg
Nicoletta Andreuccetti, Independent
Massimo Avantaggiato, Conservatorio di Milano
Lorenzo Ballerini, Conservatorio di Trapani
Carlo Barbagallo, Conservatorio di Torino
Angelo Benedetti, Conservatorio di Perugia
Nicola Bernardini, Conservatorio di Roma
Francesco Bianchi, Conservatorio di Torino
Nicola Buso, Conservatorio di Trieste
Francesco Canavese, Tempo Reale
Diego Capoccitti, Independent
Nicola Cappelletti, Conservatorio di Perugia
Maura Capuzzo, Conservatorio di Venezia
Nicola Casetta, Conservatorio di Pesaro
Daniela Cattivelli, Conservatorio di Avellino
Luigi Ceccarelli, Edison Studio
Antonino Chiaramonte, Conservatorio di Frosinone
Walter Cianciusi, Conservatorio di Sassari
Fabio Cifariello Ciardi, Conservatorio di Perugia
Pasquale Citera, Conservatorio di Bologna
Giorgio Colombo Taccani, Conservatorio di Torino
Diego Dall'Osto, Conservatorio di Vicenza
Riccardo Dapelo, Conservatorio di Piacenza
Maria Cristina De Amicis, Conservatorio di L'Aquila
Fabio De Sanctis De Benedictis, ISSM "P. Mascagni" di Livorno
Stefano Delle Monache, IRCAM
Agostino Di Scipio, Conservatorio di L'Aquila
Roberto Doati, Conservatorio di Genova
Giulia Francavilla, Independent
Nicola Frattegiani, Conservatorio di Frosinone
Francesco Giomi, Tempo Reale
Luca Guidarini, Università di Pavia
Giorgio Klauer, Conservatorio di Trieste
Sergio Lanza, Conservatorio di Alessandria
Silvia Lanzalone, CRM, Roma
Ilaria Lemmo, Conservatorio di Torino
Giulia Lorusso, Independent
Marco Marinoni, Conservatorio di Latina
Matteo Marson, Conservatorio di Torino
Damiano Meacci, Conservatorio di Firenze
Alessandro Olla, TiConZero
Federico Ortica, Independent
Claudio Panariello, KTH, Stockholm
Mattia Parisse, Independent
Pietro Polotti, Conservatorio di Trieste
Alessandro Ratoci, Conservatorio di Mantova
Luca Richelli, Conservatorio di Vicenza
Silvia Rosani, Goldsmiths University of London
Dario Sanfilippo, Independent
Andrea Sarto, Independent
Francesco Scagliola, Conservatorio di Bari
Julian Scordato, Conservatorio di Padova
Carlo Siega, 'A. Bruckner' Privatuniversität, Linz/Conservatorio di Trieste
Giuseppe Silvi, Conservatorio di Bari
Anna Terzaroli, Conservatorio di Adria
Matteo Tundo, Independent
Gianluca Verlingieri, Conservatorio di Cuneo
Giovanni Verrando, Scuola Civica "C. Abbado" di Milano
Roberto Zanata, Conservatorio di Cesena
Paolo Zavagna, Conservatorio di Venezia

Table of Contents

UNA PROSPETTIVA CRITICA SUI DATASET PER LA SINTESI TEXT-TO-AUDIO

Riccardo Ancona
Conservatorio Licinio Refice
anconariccardo@gmail.com

ABSTRACT

La sintesi text-to-audio (TTA) promette generazione sonora mediata esclusivamente da linguaggio naturale. Come tutte le tecniche di deep learning generativo, essa deriva la sua memoria del mondo e i suoi confini semantici dal dataset con cui viene addestrata. Seguendo una tradizione umanistica dell'analisi critica degli algoritmi, questo articolo prende in esame i dataset impiegati da diversi modelli di sintesi text-to-audio e il modo in cui essi sono concepiti e manipolati dai ricercatori. I dataset non sono strutture informazionali inerti, bensì oggetti socio-tecnici che implicano specifiche simbolizzazioni mediate da prospettive culturali, politiche e industriali. Uno studio comparato di labeled dataset, caption-based dataset e dataset aumentati algoritmicamente rivela i limiti tecnici ed etici legati a un approccio quantitativo alla raccolta dati. A fronte di un paradigma di valutazione dei dataset di matrice puramente computazionale, si propone una metodologia di analisi qualitativa volta a interpretare la capacità di connotazione dei modelli e suggerire pratiche alternative nella collezione delle informazioni.

1. CONTESTO

Il rapido avanzare delle tecnologie di generazione di immagini da prompt testuali (text-to-image) e dei modelli linguistici di grandi dimensioni (LLM) ha catalizzato le ricerche nel campo della sintesi da testo a suono. Gli investimenti industriali per il perfezionamento di tali tecniche di deep learning hanno permesso a numerosi ricercatori, afferenti principalmente ad aziende tecnologiche, ma anche a centri di ricerca universitari, a tentare di sviluppare algoritmi e piattaforme che permettano di generare suono mediante prompt scritti in linguaggio naturale. Se già dal 2019 erano stati compiuti alcuni esperimenti di sintesi del suono mediante *generative adversarial networks* (GAN [1, 2, 3]), il successo di nuovi paradigmi di architettura per le reti neurali, quali i *modelli di diffusione* nella generazione text-to-image [4] e i *modelli autoregressivi* nei LLM [5], ha portato allo sviluppo di numerose implementazioni adattate al contesto della sintesi text-to-audio. Dalla fine del 2022 nuovi modelli TTA si susseguono a distanza di pochi mesi o giorni, proponendo perfezionamenti tecnici e differenti architetture, che puntualmente superano i risultati nei test benchmark dei modelli precedenti. Oggetto di analisi per questo studio sono stati sia modelli di diffusione, quali Diffsound [6], Noise2Music [7], AudioLDM [8], Moûsai [9], Tango [10], sia modelli autoregressivi, come Audiogen [11], MusicLM [12] e MusicGen [13]. Mediante una lettura comparata degli articoli di ricerca che presentano questi modelli – intesi, secondo gli insegnamenti di Bruno Latour [14], come veri e propri documenti antropologici della ricerca ingegneristica – si osservano comportamenti, presupposizioni e scelte di design caratteristiche del paradigma statistico-inferenziale dominante [1] . Nell'ambito di questo articolo si è voluto studiare nello specifico un aspetto cruciale, eppure ancora sottostimato, quello dei dataset impiegati per l'addestramento dei modelli. Si è osservato che gli ingegneri autori di tali ricerche sembrano essere quasi esclusivamente interessati all'architettura delle reti neurali, esaminando molto più sinteticamente (o sorvolando completamente) la questione dei dati impiegati [2] . I dataset sono manufatti tecnici ma anche storici e sociali, il cui contenuto non può mai essere neutrale; al contrario, informa sulla visione del mondo dei suoi creatori [3] . L'orizzonte conoscitivo dei modelli di deep learning, la loro finestra sul mondo reale, si limita a ciò che i dataset contengono e descrivono. Nel caso della sintesi TTA, dove si tenta di relazionare il dominio acustico con quello testuale, sono di estrema rilevanza non soltanto i dati in sé, ma anche il modo in cui sono interconnessi tra loro. Ci si è proposti dunque di analizzare i dataset impiegati per la realizzazione di modelli TTA, il modo in cui essi sono interpretati e manipolati, nonché l'effetto che producono sui risultati sonori, a partire da una già nutrita letteratura critica impegnata nello studio dei dataset [20, 18, 21]. Melanie Feinberg [22] ha suggerito che una lettura lenta e diretta del contenuto e della struttura dei dataset, contrapposta a tecniche analitiche automatizzate, possa far emergere criticità altrimenti nascoste; Lindsay Poirier [19] ha identificato tre modalità analitiche dei dataset, denotazione, connotazione e decostruzione; il *ground truth tracing* (GTT), proposto da Edward B Kang [23], è un procedimento analitico di tipo umanistico per rapportare il contenuto di un dataset alla supposta caratteristica oggettiva a cui si riferisce, in-

[1] Una prospettiva etnografica sugli algoritmi è stata sviluppata in [15] e in [16]. Si confronti anche con l'attenta lettura sociologica di [17].

[2] Come osservato da Sambasivan et al [18]: "Paradoxically, data is the most under-valued and de-glamorised aspect of AI."

[3] Poirier [19] suggerisce che i dataset non vadano intesi come "merely instrumental artifacts tarnished by politics, but as always already iterating cultural artifacts privileging certain symbolic orders over others ".

dividuandone il grado di dissonanza ontologica. Un'interpretazione dei dataset per la sintesi text-to-audio informata da questi principi metodologici ha lo scopo di rilevare eventuali criticità e limiti insiti nei dati raccolti e nella loro capacità rappresentativa.

2. DATASET

Gli ingegneri occupati nella realizzazione di algoritmi di deep learning che analizzano testi o immagini (o una combinazione dei due) possono contare su collezioni di dati pressoché illimitate, estratte dalla rete attraverso specifici algoritmi di *web crawling*, che ogni mese setacciano internet e raccolgono trilioni di byte [24]. Di difficile reperibilità sono invece gli accoppiamenti tra suono e descrizioni testuali. Causa di tale scarsità è la struttura del web contemporaneo che incoraggia maggiormente la comunicazione visiva rispetto a quella uditiva, nonché la natura stessa delle informazioni sensoriali, essendo la vista figurativa per definizione, mentre l'udito presenta un quadro epistemologico più sfumato e imprendibile [4]. È ben più semplice osservare un'immagine e descriverne il contenuto testualmente, che non fare la stessa cosa col suono, i cui contorni restano ambigui. Contribuisce probabilmente anche un'educazione orientata a fornire una terminologia adeguata per gli stimoli visivi, ma spesso carente per ciò che concerne le informazioni uditive. Per sopperire a tali difficoltà, gli ingegneri che si occupano di sintesi TTA devono spesso ricorrere a dati scarsamente connotativi [5], oppure a tecniche algoritmiche per l'aumento artificiale dei dati [27]. Entrambi i compromessi sollevano dubbi sulla qualità dei dataset impiegati.

2.1 Labeled dataset

Fino a pochi anni fa l'approccio prevalente nel campo della descrizione testuale del suono non contemplava l'utilizzo di intere frasi, bensì di etichette (in inglese *label* o *tag*). La ricerca nel campo del *music information retrieval* prevede l'assegnazione di etichette a grandi corpora di suoni, al fine di addestrare algoritmi di classificazione in grado di distinguere identità sonore, eventi, parole, stati d'animo dei parlanti, generi musicali e stili [28, 29, 30]. La classificazione del suono per etichette è un processo fruttuoso anche per la gestione e navigazione di grandi database, come quelli che raccolgono i suoni stock forniti in licenza per applicazioni pubblicitarie e multimediali. Librerie di suoni come Audiostock [31] ed Epidemic Sound [32], entrambe tutt'ora usate come dataset per l'addestramento di modelli TTA, raccolgono decine di migliaia di file audio classificati secondo categorie funzionali, suddivise in luoghi, tipi d'uso, forme e strumenti musicali, tipologie d'immaginario. La destinazione economica di queste librerie ha un influsso determinante sui loro principi tassonomici, che dedicano ampio spazio a tipologie sonore di frequente utilizzo, come i suoni per videogiochi, i sottofondi per i video d'intrattenimento e i loghi sonori. La tassonomia di Audiostock dedica un'intera sezione a suoni di armi e combattimenti, suddivisi in ventisei tag, mentre la categoria dedicata alle emozioni consta soltanto di sei descrittori: feel down, fun, funny, sad, scary/fear, surprise. Per quanto questa categorizzazione del suono possa assolvere efficacemente alle funzioni per cui è stata progettata, risulta evidente che ereditare una tale struttura del mondo acustico in un modello TTA sia tutt'altro che auspicabile. Per tentare di descrivere il reale con delle etichette, è necessaria una chiara tassonomia, derivata da un'ontologia del suono che non è mai neutrale, ma sempre frutto di scelte ingegneristiche o industriali precise, le quali a loro volta riflettono prospettive culturali implicite ma non irrilevanti. Nel tentativo di costruire un dataset di tag fondativo e quanto più neutrale, Google ha sviluppato Audioset [33], un corpus di 1789621 segmenti audio di dieci secondi, estrapolati da video Youtube ed etichettati da ascoltatori umani seguendo una tassonomia predefinita di 527 parole chiave. Il gruppo di ricerca ha impiegato tecniche statistiche per inferire l'ontologia sonora di Audioset, in primo luogo collezionando gli eponimi del suono più ricorrenti sul web, dunque eliminando le ridondanze e costruendo una gerarchia tassonomica, ulteriormente filtrata verificando che ciascuna tag fosse statisticamente rappresentativa e intelligibile dagli operatori umani. La tassonomia ottenuta non è pertanto un ordinamento ontologico universale, bensì il riflesso statistico di ciò che in un dato momento temporale risulta essere quantitativamente più rappresentato nel discorso digitale in lingua inglese. Ciononostante Audioset rappresenta al momento il dataset più ricco e significativo per numerose applicazioni che riguardano la classificazione automatica e il riconoscimento dei suoni [6]. Ai fini dell'addestramento di un modello TTA, l'uso di etichette appare invece limitante. Lo scopo dichiarato dagli sviluppatori dei modelli più recenti è l'interazione per mezzo di frasi strutturate scritte in linguaggio naturale, che definiscono *caption*, come "qualcuno suona virtuosamente arpeggi discendenti su un arciliuto, all'improvviso passa un treno e migliaia di tortore volano via". Ammesso che il modello sia effettivamente in grado di interpretare il linguaggio naturale, ridurre questa frase alle tag Audioset "plucked string instrument, train, flapping wings" non sortirebbe lo stesso risultato, perché una tale semplificazione perde tutte le specificità della connotazione testuale: il grado di competenza dello strumentista, il tipo di figura melodica prodotta, il timbro specifico dello strumento impiegato, la consequenzialità temporale degli eventi, il numero e la specie di uccelli – mancano, in sostanza, tutte le relazioni logiche, causali, temporali, contestuali, come di una lingua fatta solo di sostantivi e aggettivi, ma priva di avverbi e complementi. Inoltre un labeled dataset può constare di qualche centinaia o migliaia di tag, un numero insufficiente a etichettare la totalità del mondo acustico; incrementi di scala nel numero di tag sono fortemente impratici. Per questo è necessario che nell'addestramento di modelli TTA si impieghino dataset di caption.

[4] È ciò che sostiene Salomé Voegelin in [25]: "Listening's focus on the dynamic nature of things renders the perceptual object unstable, fluid and ephemeral."

[5] In [26] vengono esplicitamente definiti "weakly-associated".

[6] Alternative, seppur in scala ridotta, sono i dataset FSD-50K [34] e MagnaTagATune [35].

2.2 Caption-based dataset

Il più ampio dataset open source contente coppie di caption testuali e suoni è opera di LAION, un gruppo no-profit intento a fornire dataset accessibili e trasparenti [7] . LAION-Audio-630K [37], formato da 633526 caption per un totale di oltre 4000 ore di audio, è un aggregato di alcune delle più grandi librerie sonore gratuite, quali Freesound [38], BBC Sound Effects [39] e i già citati Audiostock ed Epidemic Sound. La maggioranza delle caption, circa mezzo milione, provengono da Freesound, un database collaborativo attivo dal 2005, nel quale chiunque può condividere i propri suoni con licenza Creative Commons. Al suo interno si trovano field recordings, frammenti vocali, musica, suoni oggettuali, effetti sonori, timbri elettronici. Si tratta di un importantissimo manufatto di condivisione acustica che meriterebbe una ricerca antropologica approfondita. Bisogna domandarsi, tuttavia, se i dati in esso contenuti siano opportuni allo scopo di addestrare algoritmi di connotazione testuale del suono. Sfogliando il dataset, ci si può imbattere in caption contenenti link promozionali o precisazioni sui diritti d'autore ("For commercial purposes, please visit http:[ecc.]"), dati di natura diaristica o personale ("This one was interesting to create"), informazioni tecniche eccessivamente accurate ("Edited with TDR Kotelnikov and Waves Loudness Meter"), contestualizzazioni troppo solerti ("Date/time: June 21th 2022, 7:34 PM. Weather: 20c, clouds 67%, wind E - NE 2 bft with gusts to 13km/h, 1014 hPa, humidity 50 - 60%"), commenti agli ascoltatori ("Hey guys new track", "I would normally explain how I recorded it, but I want to see if anyone can guess how I did it"). Questo genere di annotazioni abbondano nel dataset e sono usate per l'addestramento di modelli text-to-audio [8] , introducendo un gran numero di informazioni scarsamente significative in quanto poco o per nulla connotative del contenuto sonoro associato. Unico sottoinsieme di caption in LAION-Audio-630K raccolte con una metodologia formalizzata è la libreria BBC Sound Effects, che tuttavia ne rappresenta soltanto una porzione estremamente esigua e dunque poco rappresentativa dal punto di vista statistico [9] . Un'alternativa a LAION è Clotho [40], un dataset ottenuto selezionando 4981 suoni da Freesound e chiedendo a esseri umani di descriverli con annotazioni testuali non vincolate da specifiche prescrizioni. Le persone preposte a tale compito, tutte parlanti la lingua inglese, sono state ingaggiate mediante la piattaforma di crowdsourcing Amazon Mechanical Turk [10] . Il dataset consta di 24905 caption, un numero considerato generalmente insufficiente per la sintesi text-to-audio – difatti, nessuno dei modelli studiati lo usa in maniera esclusiva [11] . In mancanza di dataset gratuiti sufficientemente estesi, per il suo modello TTA incentrato sulla generazione di musica strumentale Meta ha acquisito licenze da Shutterstock Music [43] e Pond5 [44], due servizi di musica stock generalmente impiegati per l'intrattenimento multimediale. Osservando i dati forniti dal paper [13], risulta che il dataset sia costituito prevalentemente da suoni etichettati come "electronic/edm" e "corporate", il che riflette la destinazione per cui sono stati originariamente realizzati. Un quadro complessivo dei dataset a disposizione per l'addestramento di modelli TTA mostra dunque un'evidente carenza di dati disponibili e una scarsa diversità al di fuori di categorie sonore fortemente funzionalizzate. La quantità di tempo e risorse necessarie per annotare con pertinenza terminologica centinaia di migliaia o milioni di file audio eccede le possibilità dei gruppi di ricerca coinvolti, costretti pertanto a fare uso di dataset di bassa qualità o a dover trovare metodologie algoritmiche non supervisionate per estrarre moli di dati che non si ha il tempo di far ascoltare a orecchie umane.

2.3 Dataset aumentati

Una tecnica di generazione artificiale dei dataset consiste nell'annotazione automatica delle caption mediante algoritmi di classificazione. VGG-Sound [45] è un dataset audiovisivo costituito da 200000 caption prodotte artificialmente da un algoritmo di riconoscimento visivo. Esso analizza il contenuto di clip YouTube raccolte mediante web scraping e verifica la presenza di entità nelle immagini, restringendo il campo di etichette possibili a oggetti e azioni concrete aventi un chiaro corrispettivo visivo. Si tratta di una forma di classificazione intermodale in grado di denotare eventi ed enti predefiniti, ma non di connotarli linguisticamente: le annotazioni ottenute somigliano più a delle tag che a delle caption in linguaggio naturale. Inoltre, la mera presenza visiva di un oggetto non garantisce che esso stia producendo suono durante la clip, né esclude che nella clip siano presenti suoni di altri oggetti o processi non riscontrati dall'immagine. Per sopperire a tale problematica, un algoritmo di classificazione del suono tenta di escludere i falsi positivi. Al netto di quanto osservato, sembrerebbe che chiedere a un algoritmo sordo ma con un'ottima vista di riconoscere i suoni non sia la migliore soluzione. Una metodologia alternativa consiste nel trasformare un labeled dataset come Audioset in un caption-based dataset attraverso l'utilizzo di un algoritmo di processamento testuale, come il modello linguistico T5 [46]. Il modello TTA AudioLDM ha sfruttato questa procedura *key-to-text* per trasformare Audioset in un dataset di caption [8]. Non è chiaro tuttavia in che modo un algoritmo di processamento del testo possa giungere a una maggiore connotazione del suono, non avendo esso alcuna esperienza del contenuto audio relazionato. Un tale procedimento sarebbe forse utile a sintetizzare un testo lungo in uno più breve e puntuale, ma non può di certo aggiungere informazione che non esiste nelle tag fornite, tutt'al più può diluirla in un maggior numero di parole. L'esempio mostrato sulla pagina Github dell'algoritmo di estensione testuale indica come tag "India, Capital, New Delhi" e dà in output il testo "The capital of India is New Delhi" [47]. Un test

[7] Si noti che LAION è stato altrove criticato per la metodologia di raccolta dati, ad es. in [36].

[8] Lo impiega ad esempio AudioLDM [8] come dataset primario.

[9] La libreria BBC contiene caption del tipo: "Church Atmosphere – Church congregation assembling before service with bells in background".

[10] Data la paga media infima [41], questo tipo di crowdsourcing è stato talora considerato una forma di colonialismo mediante intelligenze artificiali [42].

[11] Lo impiegano Audiogen [11] e AudioLDM [8] in aggiunta ad altri dataset.

più approfondito usando come input delle tag arbitrarie di Audioset ne mette in evidenza i limiti: le etichette "cash register, sobbing, boiling" danno come risultato "A man is sobbing and boiling cash at a cash register", con "harp, whistling, jingle" si ottiene "A man is playing a jingle and whistling on the harp", con "bassline, didgeridoo, singing bowl, tender music" si ha "A young boy is playing the bassline of a didgeridoo while a young girl is singing". In luce di quanto ottenuto dai test empirici, questo tipo di tecnica sembra essere più un modo di inquinare i dati che di aumentarli. Risulta evidente che impiegare il testo o l'immagine per espandere i dataset sia poco fruttuoso, poiché si perde la relazione diretta con l'informazione acustica. Una soluzione più efficace consiste nell'utilizzo di metodologie di codifica che proiettino sia il testo sia il suono in un unico spazio latente, vale a dire in una rappresentazione numerica condivisa. Una volta definite le proprietà dello spazio latente multimodale sulla base di un addestramento preliminare – a sua volta compiuto con informazioni provenienti da un caption-based dataset, ma di dimensioni inferiori – è possibile usare questo modello (detto modello di *pre-training*) per generare automaticamente caption avendo in input un qualsiasi suono. Questo tipo di procedura è stata inventata per la relazione tra immagini e linguaggio naturale ed è una forma di apprendimento contrastivo auto-supervisionato [48]. MuLAN [26] e CLAP [49] sono due implementazioni per il suono, rispettivamente impiegate per addestrare i modelli text-to-audio MusicLM [12] e AudioLDM [8]. Entrambi gli algoritmi di pre-training dapprima comprimono le informazioni usando un encoder per i testi (BERT o suoi derivati, [50]) e uno per i suoni (variabile a seconda delle diverse implementazioni), ottenendo i cosiddetti *embedding*, cioè dei vettori multidimensionali contenuti nello spazio latente, dunque impiegano l'apprendimento contrastivo per trovare delle relazioni trai due insiemi numerici. Una volta addestrati, è sufficiente fornire l'embedding di un nuovo suono per ottenere l'inferenza di un possibile embedding testuale; quest'ultimo non prende mai la forma di un testo intelligibile da un essere umano, ma viene immediatamente immagazzinato per essere usato come dato per l'addestramento del modello text-to-audio. Tale procedura riduce la problematica delle caption a una questione totalmente numerica, sostituendo le annotazioni testuali con rappresentazioni latenti difficili da interpretare da un essere umano, eliminando così la possibilità di verificare fenomenologicamente i risultati ottenuti durante la fase di pre-training. I dataset impiegati non sono più studiabili con un metodo qualitativo, ma dipendono da altri algoritmi di basso livello, a loro volta addestrati usando altri dataset, la cui provenienza non è sempre specificata in maniera chiara. I ricercatori che hanno sviluppato MuLAN dichiarano di aver usato 44 milioni di caption scarsamente connotate, raccolte in maniera automatizzata da video musicali trovati su internet. Dal momento che non esistono dataset pubblici così ampi, è verosimile che siano state usate come caption le descrizioni di video YouTube; le quali, notoriamente, non sono sempre connotative del contenuto sonoro, poiché spesso rimandano a link esterni, promozioni, descrizioni non pertinenti direttamente al suono e riflessioni a latere. Il presupposto logico implicito è che anche se le annotazioni potrebbero contenere molte informazioni non rilevanti o errate, il mero fatto di estrarre una tale mole di dati tenderà a diluire statisticamente gli errori e a espandere il campo semantico rappresentato: una soluzione puramente quantitativa a un problema sia di natura quantitativa sia qualitativa. Un numero così grande di video raccolti sembra anche aprire all'eventualità che alcuni dei materiali presenti potrebbero essere coperti da copyright; tuttavia un algoritmo come MuLAN dissolve ogni traccia del suo dataset di origine, perché lo riduce a vettori matematici usati in fasi intermedie della produzione di un modello TTA. Non c'è modo di stabilire a posteriori con certezza assoluta che un dato video musicale sia stato usato per l'addestramento, in quanto esso ha soltanto la funzione di informare il meccanismo di generazione automatica delle caption, ma non viene usato direttamente come esempio per l'apprendimento della generazione sonora. Eppure le nozioni linguistiche che verranno usate per descrivere il dataset di suoni sono ereditate dalle informazioni originariamente impiegate per l'addestramento del modello di pre-training: non sopravvivono i dati specifici, ma se ne deriva una rappresentazione numerico-semantica del mondo. Usando questi modelli di pre-training, si può addestrare un modello TTA senza fare ricorso a dataset di caption, sfruttando invece grandi raccolte di musica e suoni non annotate. Qualunque file audio diventa materiale potenziale per l'addestramento, senza alcun vincolo di annotazione, una notevole conquista tecnica non priva di criticità. MusicLM [12] segue questa metodologia e dichiara di aver usato cinque milioni di file musicali, senza specificarne la provenienza [12] . L'opacità di questa procedura fa in modo che non sia possibile stabilire a posteriori la presenza di una traccia musicale nel dataset, eccetto in casi di memorizzazione da parte dell'algoritmo, che gli autori del paper hanno misurato essere statisticamente rarissima [13] . La segretezza industriale circa i dataset impiegati, verosimilmente imposta per evitare problematiche di copyright e per non informare la concorrenza, provoca anche l'impossibilità di compiere uno studio approfondito della natura dei dataset, valutando qualitativamente la varietà dei contenuti, il grado di connotazione delle caption, nonché la presenza di eventuali bias.

Table 1 mostra i dataset impiegati dai modelli TTA presi in esame; sono stati esclusi dallo studio i modelli a uso commerciale che non forniscono alcuna informazione sulla provenienza dei dati, come Suno [51] e Udio [52].

3. CONNOTAZIONE

Se la frequente omissione di informazioni sulla provenienza dei dati e l'artificialità delle caption generate dagli algo-

[12] Ho chiesto chiarimenti sull'origine dei dati agli autori dell'articolo, ma non ho ricevuto risposta.

[13] Ho compiuto dei test nella versione distribuita da Google di MusicLM, MusicFX, dove ho riscontrato che la piattaforma analizza i prompt e li blocca se individua stringhe di caratteri corrispondenti a nomi di musicisti noti. Il prompt "Meshuggah", nome di un gruppo djent metal svedese, nonché sequenza di caratteri altrimenti poco ricorrente, non viene bloccato e produce generazioni aventi sonorità estremamente simili alla musica del gruppo. Questo non dimostra con certezza assoluta che la loro musica sia stata impiegata nell'addestramento, poiché non si ottengono corrispondenze 1:1.

Modello	Dataset
Diffsound	AudioCaps, AudioSet
AudioGen	AudioCaps, AudioSet, Clotho, VGG-Sound, FSD-50k e altre librerie di sound effects
MusicLM	Embedding testuali generati da MuLan usando più di 280000 ore di musica
AudioLDM	CLAP addestrato con LAION-Audio-630K, AudioSet aumentato da T5, AudioCaps, Clotho
Noise2Music	Caption generate da un LLM e assegnate automaticamente a suoni mediante MuLan
Moûsai	7000 "Spotify top recommendations" con caption estratte automaticamente dai metadati YouTube
MusicGen	Shutterstock, Pond5
Tango	AudioCaps

Table 1. Dataset impiegati dai modelli text-to-audio.

ritmi di pre-training rendono talora impossibile una valutazione qualitativa diretta del contenuto dei dataset, l'unico metodo per indagarne le proprietà è osservarne i risultati a posteriori, studiando le generazioni sonore ottenute dai modelli TTA addestrati. Come valutare i risultati di un modello di sintesi text-to-audio? La questione è tutt'altro che banale e richiede una chiara definizione di quali sono gli obiettivi perseguiti nell'addestramento del modello. Comparando le diverse implementazioni prese in esame, sembrerebbe che uno degli scopi prevalenti sia la generazione di suono per mezzo del linguaggio ordinario, senza necessità di ulteriori mediazioni né competenze a carico dell'utente, cercando quindi di operare una traduzione diretta dal pensiero verbale al suono. Si presume quindi che un modello propriamente programmato e addestrato sia capace, per il solo mezzo dell'inferenza statistica, di esibire un comportamento di tipo interpretativo nei confronti del prompt scritto dall'utente e che di conseguenza produca un risultato sonoro quanto più affine alla richiesta. I paper di alcuni modelli forniscono esempi di caption impiegate per l'addestramento, supponendo che tali modelli ne abbiano assimilato le informazioni e siano in grado di rappresentarle efficacemente. Una caption come "enchanting jazz song with a memorable saxophone solo and a solo singer" (menzionata in [12]), aggettivando con "enchanting" e "memorable" sembra intendere che lo scopo del modello TTA non sia soltanto di tipo denotativo, ma anche connotativo. Per stabilire l'efficacia della generazione è necessario dunque valutare quanto essa sia propriamente connotativa del prompt dato. Il che, naturalmente, è tutt'altro che semplice da stabilire in termini obiettivi, perché non sempre esiste una *ground truth* a cui fare riferimento: se domando "musica ruvida ma spiritosa per saltare la corda", è evidente che sto chiedendo all'algoritmo qualcosa di puramente qualitativo, per il quale io stesso non ho che un'idea approssimativa. Il linguaggio del prompt in questo caso è un'allusione, non ha valore puramente indicale. A meno che non si forniscano nel prompt indicazioni quantitative inequivocabili, come "un'onda quadra a 726.3 Hz", le richieste espresse in linguaggio naturale chiamano in causa referenti dalla natura ambigua e polivalente; a tali richieste, l'algoritmo tenterà di generare un contenuto coerente sulla base della mera rappresentazione statistica dei dati forniti, ragion per cui la capacità connotativa di un modello TTA, al netto delle diverse architetture e implementazioni, dipende dal modo in cui è connotante il dataset di addestramento. Se però il dataset è ignoto oppure è a sua volta il risultato di un processo computazionale, non rimane che compiere il percorso inverso, valutando quanto il risultato sonoro risulti essere connotativo del prompt fornito. Nei modelli di deep learning generativo la forma del prompt ha un impatto sostanziale sui risultati possibili, al punto che esiste una disciplina preposta allo studio delle formalizzazioni linguistiche, il *prompt engineering* [53]. Da studi relativi ai modelli di diffusione text-to-image risulta – e lo si può verificare empiricamente anche in tutti i modelli text-to-audio disponibili – che questi algoritmi non mostrano di saper interpretare né la grammatica né la sintassi del prompt, ignorando o confondendo negazioni, relazioni e consequenzialità [54]. Prendendo atto di tale limite intrinseco, forse valicabile con miglioramenti tecnici delle architetture dei modelli o con un mero incremento di scala dei dati impiegati per l'addestramento, si può comunque tentare di testare l'efficacia connotativa dei risultati sonori specificando nei prompt soggetti, eventi e aggettivazioni. Una valutazione empirica approfondita dell'efficacia connotativa dei vari modelli a seconda del tipo di prompt esula dalle possibilità e dagli scopi di questo articolo, tuttavia è opportuno affermare che la valutazione delle generazioni text-to-audio intesa in questo senso richiede una capacità interpretativa di tipo fenomenologico e qualitativo, per sua natura soggettiva e personale. Quando si domanda "un suono vetroso", si richiama una categoria esperienziale derivante da una complessa stratificazione di memoria e percezione, ottenuta da una pluralità di esperienze private, soggettive e talvolta contraddittorie tra di loro, che insieme formano qualia uditivi di complessa definizione [14] . Tali identità sonore divergono tra gli individui e non è detto che una loro media statistica sia sufficiente a coglierne la complessità. La valutazione dell'efficacia connotativa di un modello dovrebbe perciò basarsi sulla negoziazione tra la percezione soggettiva, le proprie aspettative e le possibilità tecniche del modello, mettendo al centro l'esperienza fenomenologica di chi compie l'analisi. Confrontando i risultati ottenuti con uno studio sia qualitativo sia statistico dei dataset, è possibile trovare delle correlazioni tra risultati inaspettati e aporie nei dati, verificando l'efficacia del modello a seconda dello scopo prefissato e individuando soluzioni specifiche per la redazione di un dataset maggiormente rappresentativo. Se la connotatività di un modello dipende in larga parte dal suo dataset, le osservazioni raccolte durante l'analisi fenomenologica dei risultati dovrebbero infor-

[14] Ho affrontato questo argomento in [55].

mare la creazione di nuovi dataset, innescando un circuito di feedback tra design algoritmico e percezione. Quanto osservato sinora nella ricerca sulla sintesi text-to-audio sembra tenere molto poco in considerazione questo genere di metodologia analitica qualitativa, affidando la valutazione dell'efficacia dei modelli a tecniche quantitative, come l'uso di *benchmark test*. Alcuni modelli, come MusicLM e AudioLDM, hanno compiuto test con esseri umani, ma sempre di natura quantitativa, domandando agli ascoltatori quale suono preferissero tra due opzioni e comparando il numero di preferenze, oppure chiedendo di indicare il grado di "qualità" di un suono su un scala numerica data [15]. Tra le metriche di benchmark più impiegate per la sintesi text-to-audio si menzionano la Fréchet Audio Distance [56] e la divergenza di Kullback-Leibler [57]. Se queste metriche possono fornire risultati utili in termini statistici per il miglioramento delle architetture, sembra tuttavia paradossale delegare il compito della valutazione delle capacità connotative e interpretative del linguaggio naturale di un algoritmo a un altro algoritmo. Misurare l'efficacia di un modello con un benchmark permette agli ingegneri di stabilire delle metriche quantitative da superare, deviando l'obiettivo della ricerca dal miglioramento effettivo dei risultati al semplice superamento di soglie numeriche precedentemente raggiunte da altri algoritmi. Seguendo tale criterio si rischia di addestrare modelli che soddisfano le richieste degli algoritmi di benchmark, ma non quelle degli utenti. L'idea che l'efficacia nella connotazione linguistica di un modello possa essere misurata quantitativamente presenta limiti evidenti se si considera la natura incommensurabile del linguaggio. Soltanto una valutazione di tipo esperienziale può fornire un'indicazione significativa delle caratteristiche qualitative dei risultati sonori; se la denotazione è tutto sommato computabile come rappresentazione media di un referente, la connotazione è per sua natura allusiva, ragion per cui solo un'entità in grado di cogliere le allusioni può esaminarne la qualità.

4. CONCLUSIONI

Uno studio comparato dei dataset e dei modelli di sintesi text-to-audio ha rilevato una scarsa curatela dei dati, derivante da un'obiettiva insufficienza di informazioni significative a disposizione e dall'uso di tecniche di aumento artificiale dei dati di dubbia efficacia. Si sono riscontrate possibili problematiche etiche legate alla raccolta dei dati, il cui processo è spesso opaco, nonché l'impiego di sistemi di valutazione dell'efficacia dei modelli di tipo puramente quantitativo, i cui risultati non rappresentano necessariamente le loro capacità di connotazione. In luce di quanto osservato, si è proposto un possibile cambio di paradigma nella raccolta dei dati e nella valutazione dei modelli, orientato all'analisi qualitativa e fenomenologica delle informazioni. Si suggerisce inoltre l'inclusione di musicisti, musicologi e semiologi all'interno dei gruppi di ricerca, al fine di orientare il design dei modelli e i processi decisionali a partire da punti di vista e competenze differenziate.

[15] I test soggettivi quantitativi potrebbero costituire una metrica significativa per l'analisi dei modelli se la loro valutazione si basasse su un protocollo formalizzato, che dovrebbe tener conto delle teorie cognitive contemporanee e delle perturbazioni indotte nelle misurazioni dal contesto e dal formato dei test.

5. REFERENCES

[1] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brebisson, Y. Bengio, and A. Courville, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," Dec. 2019. arXiv preprint.

[2] J. Nistal, S. Lattner, and G. Richard, "DrumGAN: Synthesis of Drum Sounds With Timbral Feature Conditioning Using Generative Adversarial Networks," June 2022. arXiv preprint.

[3] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," Oct. 2020. arXiv preprint.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," Apr. 2022. arXiv preprint.

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," July 2020. arXiv preprint.

[6] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete Diffusion Model for Text-to-sound Generation," 2022. arXiv preprint.

[7] Q. Huang, D. S. Park, T. Wang, T. I. Denk, A. Ly, N. Chen, Z. Zhang, Z. Zhang, J. Yu, C. Frank, J. Engel, Q. V. Le, W. Chan, Z. Chen, and W. Han, "Noise2Music: Text-conditioned Music Generation with Diffusion Models," Mar. 2023. arXiv preprint.

[8] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models," Sept. 2023. arXiv preprint.

[9] F. Schneider, O. Kamal, Z. Jin, and B. Schölkopf, "Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion," Oct. 2023. arXiv preprint.

[10] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-Audio Generation using Instruction-Tuned LLM and Latent Diffusion Model," May 2023. arXiv preprint.

[11] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "AudioGen: Textually Guided Audio Generation," Mar. 2023. arXiv preprint.

[12] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank, "MusicLM: Generating Music From Text," Jan. 2023. arXiv preprint.

[13] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and Controllable Music Generation," Jan. 2024. arXiv preprint.

[14] B. Latour, *Laboratory life: the construction of scientific facts*. Princeton, N.J: Princeton University Press, 1986.

[15] N. Seaver, "What Should an Anthropology of Algorithms Do?," *Cultural Anthropology*, vol. 33, pp. 375–385, Aug. 2018.

[16] A. Christin, "The ethnographer and the algorithm: beyond the black box," *Theory and Society*, vol. 49, pp. 897–918, Oct. 2020.

[17] M. Airoldi, *Machine Habitus. Sociologia degli algoritmi*. LUISS, 2024.

[18] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, ""Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, (Yokohama Japan), pp. 1–15, ACM, May 2021.

[19] L. Poirier, "Reading datasets: Strategies for interpreting the politics of data signification," *Big Data & Society*, vol. 8, July 2021.

[20] F. Jaton, *The constitution of algorithms: ground-truthing, programming, formulating*. Inside technology, Cambridge, Massachusetts: The MIT Press, 2020.

[21] W. Orr and K. Crawford, "The social construction of datasets: On the practices, processes and challenges of dataset creation for machine learning," Nov. 2023. knowingmachines.org.

[22] M. Feinberg, "Reading databases: slow information interactions beyond the retrieval paradigm," *Journal of Documentation*, vol. 73, pp. 336–356, Mar. 2017.

[23] E. B. Kang, "Ground truth tracings (GTT): On the epistemic limits of machine learning," *Big Data & Society*, vol. 10, p. 205395172211461, Jan. 2023.

[24] "Common Crawl." https://commoncrawl.org/.

[25] S. Voegelin, *Listening to Noise and Silence*. Continuum, 2010.

[26] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. W. Ellis, "MuLan: A Joint Embedding of Music Audio and Natural Language," Aug. 2022. arXiv preprint.

[27] K. Drossos, S. Adavanne, and T. Virtanen, "Automated Audio Captioning with Recurrent Neural Networks," Oct. 2017. arXiv preprint.

[28] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," June 2016. arXiv preprint.

[29] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," Oct. 2017. arXiv preprint.

[30] S. Vishnupriya and K. Meenakshi, "Automatic Music Genre Classification using Convolution Neural Network," in *2018 International Conference on Computer Communication and Informatics (ICCCI)*, (Coimbatore), pp. 1–4, IEEE, Jan. 2018.

[31] "Audiostock." https://audiostock.net/.

[32] "Epidemic Sound." https://www.epidemicsound.com/.

[33] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (New Orleans, LA), pp. 776–780, IEEE, Mar. 2017.

[34] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An Open Dataset of Human-Labeled Sound Events," 2020. arXiv preprint.

[35] E. Law, K. West, M. I. Mandel, M. Bay, and J. Stephen Downie, "Evaluation of Algorithms Using Games: The Case of Music Tagging," *10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, Oct. 2009.

[36] C. Buschek and J. Thorp, "Models All the Way Down," 2024. https://knowingmachines.org/models-all-the-way.

[37] "Laion-audio-630k," 2022. https://github.com/LAION-AI/audio-dataset.

[38] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proceedings of the 21st ACM international conference on Multimedia*, (Barcelona Spain), pp. 411–412, ACM, Oct. 2013.

[39] "Bbc sound effects." https://sound-effects.bbcrewind.co.uk/.

[40] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: an Audio Captioning Dataset," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Barcelona, Spain), pp. 736–740, IEEE, May 2020.

[41] K. Hara, A. Adams, K. Milland, S. Savage, C. Callison-Burch, and J. Bigham, "A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk," Dec. 2017. arXiv preprint.

[42] K. Hao and A. P. Hernández, "How the AI industry profits from catastrophe," *MIT Technology Review*, Apr. 2022.

[43] "Shutterstock music." https://www.shutterstock.com/music.

[44] "Pond5." www.pond5.com.

[45] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "VGGSound: A Large-scale Audio-Visual Dataset," Sept. 2020. arXiv preprint.

[46] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," 2019. arXiv preprint.

[47] G. Bhatia, "keytotext." https://github.com/gagan3012/keytotext.

[48] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," Feb. 2021. arXiv preprint.

[49] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP: Learning Audio Concepts From Natural Language Supervision," June 2022. arXiv preprint.

[50] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 2019. arXiv preprint.

[51] "Suno." https://suno.com/.

[52] "Udio." https://www.udio.com/.

[53] M. Diab, J. Herrera, M. Sleep, B. Chernow, and C. Mao, "Stable Diffusion Prompt Book," tech. rep., OpenArt, Oct. 2022.

[54] H.-H. Wu, O. Nieto, J. P. Bello, and J. Salamon, "Audio-text models do not yet leverage natural language," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Rhodes Island, Greece), IEEE, 2023.

[55] R. Ancona, "Material identities in corpus-based algorithmic improvisation," Master's thesis, Institute of Sonology, 2023.

[56] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms," Jan. 2019. arXiv preprint.

[57] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, pp. 79–86, Mar. 1951.