UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA DELL'INFORMAZIONE

"Standard MPAI basato su AI per la conservazione di nastri magnetici audio"

Relatore: Prof. Sergio Canazza

Laureando: Filippo Zago

Correlatori: Dott. Matteo Spanio
Dott. Alessandro Russo

ANNO ACCADEMICO 2023 - 2024

Data di laurea: 14 / 03 / 2024

Indice

1 Introduzione

Questo elaborato si propone di contestuallizzare e riportare l'esperienza svolta nel tentativo di sviluppare un modello di *machine learning* per il riconoscimento della direzione di riproduzione di un nastro magnetico.
L'esperienza si inserisce in un progetto di ampio respiro del Centro di Sonologia Computazionale (CSC) del Dipartimento di Ingegneria dell'Informazione dell'Università di Padova: la realizzazione di precisi riferimenti software per la realizzazione dello standard internazionale MPAI/IEEE CAE-ARP, per la preservazione di registrazioni audio - uno standard di cui il CSC stesso è stato uno dei principali contibutori. L'esperienza è stata impostata come indagine esplorativa per il successivo sviluppo di una componente per il riconoscimento della direzione di riproduzione di un nastro magnetico.
La linea d'indagine proposta dal CSC verte sulla scelta delle *features* per la caratterizzazione dei segnali audio e per la definizione della capacità discriminatoria del modello. L'indagine è stata condotta attraverso del codice Python in Jupyter Notebook utilizzando alcune librerie dedicate all'analisi audio e all'allenamento di modelli di *machine learning*. Viste le limitate capacità di calcolo disponibili si sono selezionate solo un paio di *features* spettrali: RMS e MFCC. Le conclusioni dell'indagine spiegano se queste siano adatte per l'allenamento di tale modello.
Precedono il report sull'esperienza svolta un'ampia sezione compilativa sullo standard di MPAI, sulle tecnologie adoperate per la realizzazione dello standard e un breve approfondimento sulla fisica della lettura di nastri magnetici; per informare il lettore riguardo a questi temi prima della lettura del report.

2 Lo standard MPAI/IEEE *Context-based Audio Enhancement* (CAE)

Gli archivi audio in Europa sono caratteristicamente di ridotte dimensioni e personali. Molti di essi non sono ancora digitalizzati, per cui questi documenti sonori analogici, registrati anche più di 50 anni fa, rischiano di diventare inaccessibili.
Negli ultimi anni si è assistito a un notevole interesse in favore della digitalizzazione dei processi aziendali, ma anche di documenti audio/video storici, si veda ad esempio il "Piano Nazionale di Ripresa e Resilienza" (PNRR) in Italia, un programma che mira a utilizzare i fondi *NextGenerationEU* per reintegrare e proteggere il patrimonio di documenti audio/video storici.
Purtroppo, i piani di digitalizzazione sono spesso progettati e condotti da professionisti che non hanno esperienza nella conservazione del patrimonio culturale musicale. Inoltre, l'assenza di standard internazionali ampiamente accettati nel campo dell'archiviazione audio rappresenta una sfida significativa per l'interoperabilità tra gli archivi.
Sebbene nel corso degli anni siano state sviluppate diverse linee guida, è persistita la mancanza di standard internazionali approvati con precisi riferimenti software. Questa situazione crea il rischio di produrre oggetti digitali incompatibili tra loro, rendendo praticamente inutile il lavoro di digitalizzazione e impedendo agli studiosi e al pubblico di accedere ai documenti.

Lo standard MPAI/IEEE-CAE ARP (Cap. 3) colma questa lacuna nel campo dell'archiviazione audio, fornendo precisi riferimenti software al lavoro di conservazione. Le specifiche tecniche CAE-ARP adottano una metodologia di conservazione che si basa su un approccio filologico al documento audio e incorpora strumenti informatici all'avanguardia basati sull'intelligenza artificiale. Sfruttando questo approccio, CAE-ARP mira a garantire che gli oggetti digitali prodotti siano compatibili tra loro, promuovendo così l'interoperabilità tra gli archivi.

Il Centro di Sonologia Computazionale (CSC) del Dipartimento di Ingegneria dell'Informazione dell'Università di Padova è uno dei principali contributori al CAE-ARP. Il CSC è un laboratorio multidisciplinare attivamente impegnato nella produzione musicale, nella ricerca audio, nella didattica e nella divulgazione. Un'importante area di ricerca del CSC è la conservazione e la valorizzazione del patrimonio culturale, in particolare dei documenti audio vocali e musicali. Audio Innova (http://audioinnova.com), uno spin-off del CSC, è un

membro fondatore e un importante contributore di MPAI-CAE (Cap.2). Lo standard internazionale MPAI-CAE è stato approvato nel maggio 2022 e successivamente adottato dall'IEEE Standards Association con il numero 3302-2022 nel dicembre dello stesso anno.

2.1 MPAI risponde alla necessità di standardizzare prodotti basati su AI con *AI Framework* (AIF)

Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) è un'associazione internazionale con la missione di sviluppare degli standard di codifica dei dati, basati sull'applicazione di intelligenza artificiale. Ricerche hanno dimostrato che la codifica dei dati con le tecnologie basate sull'intelligenza artificiale è più efficiente rispetto alle tecnologie esistenti.

La maggior parte degli standard applicativi MPAI sono compositi, nel senso che un'applicazione completa che realizzi uno dei casi d'uso (Cap.2.2) è tipicamente implementata con un insieme di elementi di elaborazione di base chiamati *AI Moduls* (AIM), collegati per formare flussi di lavoro (*AI Workflow*, AIW) e eseguiti in un contesto (*AI Framework*, AIF). Lo standard fondamentale di MPAI *AI Framework* (MPAI-AIF) specifica l'architettura, le interfacce, i protocolli e le *Application Programming Interfaces* (API). MPAI sta facendo il possibile per identificare moduli di elaborazione che siano riutilizzabili e aggiornabili senza necessariamente cambiare la logica interna, in modo tale da assicurare l'interoperabilità tra le diverse implementazioni.

Sebbene le AIM debbano esporre interfacce standard per poter operare in un MPAI *AI Framework*, le loro prestazioni possono variare a seconda delle tecnologie utilizzate per implementarle. MPAI ritiene che gli sviluppatori in concorrenza tra loro che si sforzano di fornire AIM proprietarie, interoperabili e più performanti, promuoveranno mercati orizzontali di soluzioni AI che si basano sull'innovazione dell'AI e la promuovono ulteriormente.

2.2 Scopo dello standard MPAI-CAE e casi d'uso

La scopo finale dello standard è il miglioramento dell'esperienza dell'utente per le applicazioni legate all'audio, tra le quali intrattenimento, comunicazione, teleconferenza,

videogiochi, post-produzione, restauro, ecc., in una varietà di contesti, in casa, in auto, in movimento, nello studio, ecc., utilizzando le informazioni contestuali per agire sul contenuto audio in ingresso e potenzialmente fornire un output elaborato tramite un protocollo appropriato.

Nella prima versione dello standard [2] sono stati identificati quattro casi d'uso:

- *Emotion Enhanced Speech* (EES), consente all'utente di convertire un singolo segmento di parlato privo di emozioni in un segmento con un'emozione specifica espressa come *tag* appartenente a un elenco standard di emozioni o derivata dall'estrazione di caratteristiche da un modello di enunciato.
- *Audio Recording Preservation* (ARP), consente all'utente di creare una copia dell'audio digitalizzato di un nastro magnetico a bobina aperta per la conservazione a lungo termine e una copia ad accesso ripristinato per la corretta riproduzione.
- *Speech Restoration System* (SSR) consente all'utente di ripristinare un segmento danneggiato di un segmento audio contenente il parlato di un singolo oratore, fornendo il testo del segmento danneggiato e un modello di rete neurale ottenuto da parti non danneggiate del segmento di parlato in ingresso.
- *Enhanced Audioconference Experience* (EAE), consente all'utente di migliorare l'esperienza uditiva durante un'audioconferenza: estraendo i segnali vocali dai singoli altoparlanti; riducendo il rumore di fondo e il riverbero da un array di microfoni; estraendo gli attributi spaziali dei partecipanti alla conferenza per consentire la rappresentazione spaziale dei segnali vocali al ricevitore.

La versione più recente dello standard [3] riporta inoltre il seguente sottosistema *Human* CAV *Interaction* (HCI) pertinente però allo standard MPAI, *Connected Autonomous Vehicle* (CAV), il quale sfrutta AIM dello standard CAE.
Un CAV è un sistema in grado di eseguire un comando di spostamento basato su 1) l'analisi e l'interpretazione dei dati rilevati da una serie di sensori di bordo che esplorano l'ambiente e 2) le informazioni trasmesse da altre fonti nel raggio d'azione, ad esempio altri CAV, semafori e unità stradali.
Il sottosistema di interazione tra persona e CAV (Human CAV Interaction) è in grado di separare le diverse fonti vocali interne o esterne al veicolo per poter partecipare alla conversazione, ad esempio per rispondere a domande specifiche o ricevere indicazioni sulla prossima destinazione.

3 Tecnologia CAE-ARP

La conservazione di risorse audio registrate su una varietà di supporti (vinili, nastri, cassette, ecc.) è un'attività importante per una varietà di domini applicativi, in particolare per il patrimonio culturale, perché la conservazione richiede più di un trasferimento di informazioni audio dal dominio analogico a quello digitale. Per esempio:

- è necessario recuperare e conservare le informazioni sul contesto, ovviamente, ma non esclusivamente, audio;
- la registrazione di un evento acustico non può essere un'operazione neutra perché la qualità timbrica e il valore plastico del suono registrato, che sono di grande importanza a esempio nella musica contemporanea, sono influenzati dal posizionamento dei microfoni utilizzati durante la registrazione;
- l'elaborazione viene effettuata dal *Tonmeister*, cioè da colui che ha una conoscenza teorica e pratica approfondita di tutti gli aspetti della registrazione sonora. Tuttavia, a differenza di un tecnico del suono, il *Tonmeister* deve avere anche una profonda formazione musicale: le competenze musicologiche e storico-critiche sono essenziali per l'identificazione e la corretta catalogazione delle informazioni contenute nei documenti audio;
- i supporti sonori, essendo costituiti da materiali di base instabili, sono maggiormente soggetti a danni causati da una manipolazione inadeguata. La commistione tra una formazione tecnico-scientifica e una conoscenza storico-filologica diventa essenziale per le operazioni di ri-registrazione a scopo di conservazione, andando oltre la semplice conversione analogico/digitale.

Nel caso dei nastri magnetici, il supporto può contenere informazioni importanti: il nastro può includere giunzioni multiple; può essere annotato (dal compositore o dai tecnici) e/o presentare diversi tipi di irregolarità (ad esempio, corruzione del supporto, nastro di diverso colore o composizione chimica).

Il caso d'uso CAE-ARP si occupa della creazione di una copia digitale dell'audio digitalizzato di nastri magnetici a bobina aperta per la conservazione a lungo termine (*Preservation Master File*) e di una copia di accesso (*Access Copy File*), ripristinata se necessario, per la corretta riproduzione della registrazione digitalizzata.

Questo approccio filologico è stato affinato con l'esperienza del CSC e è andato formalizzandosi in una metodologia precisa. Lo schema relativo si trova in Fig.1.

3.1 Metodologia di preservazione dell'audio

3.1.1 Preparazione del supporto

In questa prima fase, il documento audio viene fotografato insieme alla sua scatola, al fine di documentarne lo stato di conservazione e tenere traccia di annotazioni/osservazioni sulla scatola, che spesso danno informazioni utili sul modo corretto di riprodurre un nastro, come la velocità di registrazione e la configurazione dei canali.
L'ispezione visiva del nastro aiuta a diagnosticare problematiche o sindromi chimico-fisiche che possono colpirlo. In alcuni casi, potrebbe essere necessaria un'analisi chimica approfondita con specifica strumentazione di laboratorio. L'ottimizzazione del supporto comprende il trattamento termico che può essere applicato per risolvere la sindrome *Sticky-Shed Soft Binder* (una condizione creata dal deterioramento dei leganti adesivi del nastro, che trattengono il rivestimento magnetizzabile in ossido ferrico al suo supporto in plastica, dovuto all'assorbimento di umidità nell'aria); la pulizia della superficie del nastro per rimuovere muffe, polvere e altre particelle che possono presenti su di esso; il restauro manuale e la riparazione di vecchie giunzioni; l'aggiunta del nastro leader, ecc.

3.1.2 Trasferimento del segnale

Il trasferimento del segnale audio da analogico a digitale comporta un processo completo. Prima della digitalizzazione vera e propria è fondamentale analizzare i formati di registrazione, parametri e la configurazione di riproduzione, nonché l'apparecchiatura di bonifica e di monitoraggio. Il monitoraggio dell'intero processo di acquisizione aiuta a prevenire errori, come l'errata interpretazione della configurazione del canale o della velocità di registrazione. Seguendo linee guida stabilite, il segnale audio viene digitalizzato con una profondità di 24 bit per campione audio e una frequenza di campionamento di 96 kHz.
Durante il processo di digitalizzazione, il nastro viene filmato con una videocamera puntata sulla testina di riproduzione del registratore a bobina aperta. La documentazione video tiene annotazioni, segni e altre irregolarità che possono essere presenti sulla superficie del

nastro. Inoltre, problemi audio nel file digitalizzato, come ad esempio cali di volume, che possono essere legati a disallineamenti del nastro sulla testina di riproduzione e sulla documentazione video in considerazione.
Dopo il trasferimento del segnale, la sorgente originale viene archiviata.

3.1.3 Elaborazione e archiviazione dei dati

L'ultima fase della metodologia di conservazione riguarda l'elaborazione dei dati e la creazione di una copia di conservazione e di una copia di accesso.
La copia di conservazione comprende il file audio digitale di alta qualità, con l'audio memorizzato di solito con una profondità di 24 bit e una frequenza di campionamento di 96 kHz, senza alcun restauro o filtro applicato. Per le registrazioni multicanale, viene fornito un file audio separato per ogni canale. Vengono eseguite acquisizioni multiple, quando i nastri sono stati registrati a velocità diverse, e generati file audio separati. Oltre ai file audio digitali, la copia di conservazione comprende documentazione fotografica e video, checksum e immagini scannerizzate della documentazione che eventualmente accompagnava l'articolo originale.
La raccolta dei metadati riveste un ruolo centrale in questo processo. I dati relativi al documento originale, come la marca, il diametro della bobina, la configurazione del canale, velocità di registrazione, ecc. sono memorizzati in un database dedicato e riassunti in un file .pdf, anch'esso incluso nella copia di conservazione. La copia di accesso è generalmente in un formato compresso (ad esempio, MPEG AAC).

3.2 Modello di riferimento CAE-ARP e funzioni dei relativi moduli AI

L'ARP AIW e i diversi componenti sono mostrati in Fig.2. L'architettura tecnica dello standard comprende cinque AIM che si occupano e elaborano diversi ingressi digitali: *Audio Analyser, Video Analyser, Tape Irregularities Classifier, Tape Audio Restoration, Packager*.

3.2.1 *Audio Analyser* e *Video Analyser*

Il segnale audio analogico viene estratto dal nastro a bobina aperta durante il processo di digitalizzazione. La metodologia di conservazione adottata prevede che il nastro venga

filmato durante l'intera digitalizzazione per tenere traccia delle annotazioni, dei segni e di altre irregolarità che possono essere presenti sulla sua superficie. Queste informazioni vengono memorizzate nel *Preservation Audio-Visual File* per la conservazione. La telecamera è puntata sulla testina di riproduzione del registratore a bobina aperta, e la registrazione video include una traccia audio di bassa qualità estratta direttamente da una delle uscite del registratore (ad esempio l'uscita per le cuffie) per consentire la corretta sincronizzazione del video con il segnale audio di alta qualità memorizzato nel *Preservation Audio File*.

Il *Video Analyser* individua le irregolarità sul nastro attraverso algoritmi di *computer vision*, fornendo immagini di irregolarità e assegna un ID univoco a ciascuna irregolarità.
Il *Video Analyser* riceve anche un *Irregularities File* contenente le irregolarità rilevate dall'*Audio Analyser*, l'offset tra il file audio di conservazione e la traccia audio del *Preservation Audio-Visual File* . L'*Audio Analyser* estrae i frammenti e li classifica utilizzando algoritmi di apprendimento automatico (un *random forest classifier*).
L'identificazione delle irregolarità avviene sulla base del riconoscimento delle curve di equalizzazione e delle velocità di registrazione. I moduli *Audio/Video Analyser* inviano i blocchi audio analizzati, le immagini di irregolarità e i file di irregolarità corrispondenti al *Tape Irregularity Classifier* per la classificazione.

3.2.2 Tape Irregularity Classifier

Le informazioni elaborate da *Audio Analysers* e *Video Analyser*, come l'*Irregularities File* e le immagini e i file audio corrispondenti, vengono ricevute in ingresso dal modulo *Tape Irregularity Classifier*. Questo AIM permette di classificare le immagini di irregolarità e di selezionare quelle rilevanti tramite una *Deep Neural Network*. Il file di irregolarità finale viene inviato ai moduli *Tape Audio Restoration* e *Packager;* quest'ultimo riceve anche le immagini di irregolarità corrispondenti.

3.2.3 Tape Audio Restoration

I nastri a bobina aperta possono essere registrati a diverse velocità. Le macchine professionali da studio di solito adottano velocità più elevate, mentre quelle portatili sono caratterizzate da valori di velocità più bassi per aumentare la lunghezza di registrazione dei nastri più piccoli. Un altro parametro critico nel restauro e nella conservazione dei nastri è l'equalizzazione

adottata nella registrazione originale. La curva di equalizzazione viene utilizzata durante la registrazione come pre-enfasi per estendere la gamma dinamica e migliorare il rapporto segnale/rumore (SNR) del segnale registrato. Durante la riproduzione, viene applicata la curva di post-enfasi inversa per ripristinare la risposta in frequenza originale.

A volte, data la difficoltà di trovare macchine analogiche perfettamente funzionanti con la corretta configurazione di velocità e le impostazioni di equalizzazione richieste dallo specifico documento audio in questione, gli operatori devono eseguire la digitalizzazione adattandosi ai dispositivi disponibili durante il processo. In altri casi, l'operatore potrebbe non rilevare le variazioni di velocità avvenute durante la registrazione originale. Queste situazioni comuni possono introdurre errori che non possono essere annullati.

Il modulo *Tape Audio Restoration* corregge gli errori legati alle variazioni di velocità e quelli legati all'applicazione di curve di equalizzazione errate. Al termine del processo, il modulo fornisce i file audio ripristinati e un elenco di modifiche al *Packager*.

3.2.4 *Packager*

Questo AIM raccoglie il *Preservation Audio File*, i file audio restaurati, l'elenco delle modifiche, l'*Irregularities File*, le immagini delle irregolarità e i dati del *Preservation Audio-Visual File* e produce *Preservation Master File* e *Access Copy File*.

3.3 Attuale implementazione

La stretta collaborazione di Audio Innova con il CSC dell'Università di Padova ha permesso all'azienda di sfruttare al meglio la ricerca sulla conservazione dei documenti audio e di implementare, testare e commercializzare la tecnologia ARP. Questo ha fornito un'opzione per archivi grandi e piccoli che non sempre è disponibile quando si lavora con istituzioni meno agili, come le università. Grazie a questa sinergia e allo sviluppo dello standard CAE-ARP, sono state digitalizzate e opportunamente conservate diverse migliaia di registrazioni audio provenienti da archivi internazionali, tra i quali la Fondazione Archivio Luigi Nono di Venezia, il Centro Studi Luciano Berio e la Paul Sacher Stiftung di Basilea (computer music e musica elettroacustica), Fondazione Giorgio Cini Onlus di Venezia, Istituto Superiore di Studi

Musicali di Reggio Emilia e Castelnuovo ne' Monti (registrazioni vocali ed etnomusicologia), Archivio Storico del Maggio Fiorentino, Archivio Storico Fondazione Arena di Verona e Archivio Storico del Teatro Regio di Parma (opera e musica classica).

4 Approfondimento sulla lettura inversa di nastri magnetici

Su un nastro magnetico, le informazioni acustiche sono organizzate in "tracce". Esistono diverse configurazioni possibili delle tracce (Fig.3). Ad eccezione dei nastri magnetici solo su un lato, in cui tutte le tracce devono essere lette nella stessa direzione, la maggior parte dei nastri presenta due lati. L'apparecchiatura di riproduzione deve essere compatibile con le tracce presenti sul nastro.
Quando si lavora con un segnale digitale, sono possibili diverse manipolazioni del segnale, sia che il segnale sia stato digitalizzato (trasferimento analogico-digitale) sia che sia stato estratto da un supporto digitale e trasferito a un dispositivo digitale di elaborazione (trasferimento digitale-digitale). Le manipolazioni più semplici includono la normalizzazione, l'equalizzazione, il *trimming* o l'inversione di un'intera traccia. Dalla possibilità di invertire una traccia digitale in questo modo deriva il fatto che si potrebbero estrarre tutte le tracce su un nastro a due facciate in un unico passaggio, indipendentemente dalla direzione prevista, poiché possiamo successivamente invertire il segnale digitalmente.
Questa è un'opzione interessante, perché il tempo totale di digitalizzazione per ogni nastro si riduce del 50%. Questo comporta non solo un importante risparmio di tempo, che si traduce poi in un risparmio anche economico, ma anche una minor usura del supporto, il quale necessariamente si indebolisce con ogni riproduzione.

4.1 Fisica della lettura di nastri magnetici

I nastri magnetici possono essere registrati e riprodotti per mezzo di un registratore a bobine (vedi schema semplificato in Fig.4). Durante la riproduzione, il nastro si muove da sinistra a destra, dalla bobina di origine a quella di ricezione. L'avanzamento del nastro avviene attraverso il capstan, un cilindro rotante azionato da un motore elettrico, contro il quale il nastro viene fatto aderire per mezzo di uno speciale rullo. La velocità angolare del capstan è controllata in modo da garantire che il nastro scorra con una velocità il più possibile costante. Per leggere il segnale magnetico registrato, il nastro viene fatto passare davanti alla testina di riproduzione sulla quale il nastro induce un campo magnetico variabile proporzionale al segnale registrato. Ciò determina la generazione di una forza elettromotrice indotta alle estremità dell'avvolgimento della testina. Seguendo lo schema, questo segnale passa attraverso varie trasformazioni che hanno lo scopo di compensare in parte le perdite di

ampiezza alle varie frequenze che si verificano nel processo di riproduzione e registrazione. La complessità del sistema di registrazione magnetica appena descritto ha alcune conseguenze: l'attrito del nastro e la dinamica del sistema di trazione, costituito da bobine, motore e rulli, fanno sì che la velocità del nastro e la sua posizione relativa rispetto alla testina di riproduzione siano soggette a continue piccole variazioni difficili da controllare. Il risultato è che un segnale letto dallo stesso nastro magnetico sarà leggermente diverso a ogni lettura. In altre parole, esiste una certa "tolleranza" nella riproduzione analogica: solo dopo una certa soglia si potrebbe ipotizzare che l'apparecchiatura sia difettosa. Inoltre, l'intero sistema elettromeccanico presenta alcune caratteristiche non lineari e un comportamento dinamico che avvalorerebbero l'ipotesi che la lettura "inversa" dei nastri introduca differenze misurabili rispetto all'alternativa standard.

4.2 Discussione sulla validità della strategia di lettura inversa

La letteratura sul tema è scarna perché questo campo è ancora in fase di esplorazione ma alcuni risultati [6] rivelano che la strategia di leggere i nastri al contrario introduca alterazioni misurabili nel segnale. Queste alterazioni sono maggiori della variabilità insita nel processo di riproduzione con i registratori analogici a bobina.
È importante però far notare che l'introduzione di queste alterazioni non implica che la lettura dei nastri al contrario (nel contesto dei progetti di digitalizzazione) sia sbagliata o da evitare. Indica solamente che i due approcci di lettura (ortodossa e inversa) non danno gli stessi risultati a livello di segnale, ma non si hanno informazioni sicure sulla percezione soggettiva delle alterazioni, ovvero se un ascoltatore umano sia in grado di notare la differenza tra le riprese. Ciò significa che la decisione se sia conveniente e sicuro risparmiare tempo leggendo i nastri al contrario deve essere valutata caso per caso.

5 Rapporto su un'esperienza di sviluppo di un modello di *machine learning* per il riconoscimento di tracce audio lette inversamente

Se si considera di utilizzare la strategia di lettura inversa delle tracce, è necessario che l'AIM *Audio Analyser* sia in grado di distinguere tra segmenti audio letti direttamente o inversamente. Quei segmenti che sono riconosciuti come letti inversamente saranno digitalmente ribaltati in modo da ristabilirne l'orientamento. Successivamente si costruisce un segnale audio, completo delle sue parti lette direttamente e di quelle lette inversamente e ribaltate, il quale andrà a costituire l'*Audio File* da passare in ingresso all'AIM *Tape Irregularity Classifier*.

L'esperienza svolta si prefiggeva l'obbiettivo di scegliere e allenare un modello di *machine learning* come classificatore binario per svolgere questa funzione, e non tanto con l'intenzione di realizzare un componente utilizzabile quanto piuttosto di produrre qualche spunto o indicazione per il futuro sviluppo dell'effettivo componente.
Sono state svolte due indagini, una prima indagine esplorativa e una seconda più formale, entrambe utilizzando il linguaggio *Python* all'interno di *Jupyter Notebook*.
Riguardo agli strumenti utilizzati: i modelli di classificazione sottoposti a analisi sono modelli standard messi a disposizione dalla libreria *scikit-learn*; per l'elaborazione dei segnali audio sono utilizzate le librerie *Librosa* e *SoundFile*.
Seguono un resoconto dell'esperienza e una breve discussione dei risultati ottenuti.

5.1 Indagine esplorativa su un ridotto dataset di prova

In una prima fase è stato fornito un esiguo dataset di prova per familiarizzare con gli strumenti da utilizzare. Questo dataset si compone di sei segmenti audio (in formato *.wav*), di cui una metà sono stati ottenuti riproducendo un nastro secondo il normale orientamento e selezionando tre segmenti, l'altra metà riproducendo il nastro al contrario e selezionando segmenti corrispondenti. Un esempio delle forme d'onda di una coppia di segmenti letti correttamente e inversamente è riportata in Fig.5: si nota un'evidente simmetria tra i due segnali. Caratteristiche comuni di questo dataset sono: frequenza di campionamento a 96 kHz, canale singolo e profondità a 24 bit.

Prima di procedere all'allenamento di un modello, i dati devono essere processati in modo da presentarsi in una forma adatta al modello. Per ogni entrata del dataset serve definire una coppia (*feature*,*label*) dove *feature* è un oggetto che descrive l'entrata mentre *label* è l'etichetta su cui si basa la classificazione. In questo caso la classificazione è binaria e si sono usati i valori 0 e 1 come *labels* rispettivamente per file 'letto direttamente' e 'letto inversamente'. Le *features* che si è deciso di usare sono ricavate dalle sequenze temporali di ambiezze che descivono le forme d'onda dei segnali audio attraverso predefinite funzioni di *Librosa* sono *MFCC* e *RMS* (vedi sottopunto successivo per maggiori dettagli). Inizialmente si volevano utilizzare altre *features* che però generano strutture dati più corpose e data la scarsa capacità di calcolo dell'elaboratore utilizzato si è optato per queste.

Un altro necessario intervento in questa fase di pre-elaborazione dei dati riguarda la durata dei file audio. I segmenti audio forniti sono di durata troppo estesa per l'allenamento di un modello: non ci si aspetta di avere sempre a che fare con file di diversi minuti di lunghezza e inoltre le *features* sono molto più caratterizzanti per sezioni di minore estensione. Si procede quindi prima di estrarre le *features* a una riconfigurazione del dataset operando una segmentazione di ogni file audio in porzioni di qualche secondo. La scelta di una precisa durata per queste porzioni deve essere motivata dal contesto audio. Poiché il contenuto vocale è un forte indicatore per l'orientamento di un segnale (ci appare evidente ascoltandole registrazioni lette al contrario) si può ipotizzare di utilizzare come durata di ogni porzione un tempo che permetta la produzione di una frase. In prima analisi si è scelto di segmentare i file in porzioni di due secondi per ottenere un dataset quanto più corposo possibile, in questo caso si sono ottenuti poco più di ottanta segmenti; il dataset risulta comunque esiguo rispetto alle dimensioni di norma utilizzate per allenare un modello.

Dopo aver eseguito le operazioni di segmentazione, estrazione delle *features* e assegnazione del corretto *label*, si procede a suddividere il dataset nelle porzioni utili all'allenamento (*train set*) e al test (*test set*) del modello – inutile dividere ulteriormente il dataset in una porzione di validazione (*validation set*) per la sua dimensione già estremamente ridotta.
Si può quindi allenare un modello sul *train set* e valutarne la precisione di classificazione sulle sue prestazioni nell'attribuire un *label* alle entrate del *test set*.
Comparando le precisioni sul *test set* su una variegata selezione di modelli presi da *scikit-learn* (KNeighborsClassifier, SVC, DecisionTreeClassifier, RandomForestClassifier, AdaBoostClassifier, GaussianNB, QuadraticDiscriminantAnalysis), appaiono maggiormente

adatti allo scopo algoritmi basati su alberi - RandomForest e DecisionTree – e l'algoritmo KNeighbors.
Si decide di impostare la successiva analisi sull'ottimizzazione dei modelli RandomForestClassifier e KNeighborsClassifier.

5.1.1 Significato delle *features* estratte: RMS e MFCC

Un file audio può essere rappresentato come una serie temporale in cui l'asse dipendente è l'ampiezza della forma d'onda audio. La forma d'onda del file audio è l'unica informazione di cui disponiamo per creare le caratteristiche con cui addestrare il nostro modello. Tuttavia, la forma d'onda non contiene abbastanza informazioni discriminanti, motivo per cui da questa si estraggono delle *features* maggiormente caratterizzanti. Lo sviluppo di un modello che si basa su *features* comporta vantaggi che non si limitano a: modelli più accurati e generalizzabili, approfondimento del comportamento decisionale del modello, maggiore flessibilità nella scelta dei modelli, *training* più rapido.

Il R*oot Mean Square* (RMS), o valore efficace, è definito per una sequenza di valori semplicemente come la radice quadrata della media dei quadrati dei valori. Per un segnale audio la in questione è quella delle ampiezze del segnale. Questo indice lega il concetto di volume mediamente percepito con le metriche più facilmente quantificabili che si utilizzano per misurare l'ampiezza del sengale audio.

MFCC è l'acronimo di Mel-Frequency Cepstral Coefficients. Questa tipo di analisi viene spesso utilizzata per la descrizione e il confronto del timbro. Questi coefficienti imbrigliano sinteticamente il contenuto informativo dello spettro del segnale.
I valori MFCC sono ricavati a partire dallo spettro in scala Mel. La scala Mel mette in relazione la frequenza percepita di un tono puro con la sua frequenza effettiva misurata. Gli esseri umani sono molto più abili nel discernere le piccole variazioni di intonazione alle basse frequenze rispetto alle alte frequenze. L'incorporazione di questa scala fa sì che le nostre caratteristiche corrispondano maggiormente a ciò che gli esseri umani sentono.
Il calcolo dei coefficienti dalla forma d'onda richiede diverse trasformazioni. Si passa alla rappresentazione in frequenza attraverso la *Short-Time Fourier Transform* (STFT) che taglierà la nostra forma d'onda audio in brevi segmenti eventualmente sovrapposti, di uguale lunghezza e ne prenderà la trasformata di Fourier per ogni segmento individualmente. Ogni

spettro viene portato in scala Mel e successivamente dato in ingresso alla *Discrete Cosine Transform* (DCT). Ciò significa che la forma degli spettri in scala Mel viene confrontata con una serie di forme d'onda cosinusoidali (forme cosiniche diverse create da frequenze diverse). Ogni valore MFCC rappresenta quanto lo spettro di ogni segmento sia simile a una di queste forme cosinusoidali.
In parole semplici, i coefficienti MFCC ci danno un'idea del cambiamento di tonalità di un segnale audio.

Per ogni entrata del dataset vengono estratti 40 valori MFCC (valore consigliato in *Librosa*) e un valore RMS, allineati in un unico vettore di *features* poi associato al *label* corripondente al file di origine.

5.1.2 Concetti base degli algoritmi *Random Forest* e *K-Nearest Neighbour*

L'algoritmo *Random Forest* crea molteplici istanze di *Decision Tree* che vengono uniti per ottenere una previsione più accurata.
La logica alla base del modello *Random Forest* è che più modelli non correlati (i singoli alberi) funzionano molto meglio come gruppo che da soli. Quando si utilizza *Random Forest* per la classificazione, ogni albero fornisce una classificazione o un "voto". La foresta sceglie la classificazione con la maggioranza dei "voti". Quando si utilizza *Random Forest* per la regressione, la foresta sceglie la media dei risultati di tutti gli alberi.
È fondamentale che la correlazione tra i singoli modelli, cioè tra i *Decision Tree* che compongono il modello più ampio di Random Forest, è bassa (o nulla). Anche se i singoli alberi possono produrre errori, la maggior parte del gruppo sarà corretta, portando così il risultato complessivo nella giusta direzione.

L'algoritmo *K-Nearest Neigbours* si basa sull'idea che punti di dati simili tendono ad avere etichette o valori simili.
Durante la fase di addestramento, l'algoritmo memorizza l'intero set di dati di addestramento e li fissa in punti in uno spazio di dimensioni pari al numero di caratteristiche. Al momento di produrre delle previsioni, viene calcolata la distanza tra il punto che identifica un dato nuovo in ingresso e tutti i punti noti, utilizzando una metrica di distanza scelta, come la distanza euclidea. L'algoritmo identifica i K punti più prossimi al nuovo punto in ingresso in base alle loro distanze.

Nel caso della classificazione, l'algoritmo assegna l'etichetta di classe più comune tra i K vicini come etichetta prevista per il punto di ingresso. Nel caso della regressione, calcola la media o la media ponderata dei valori target dei K vicini per prevedere il valore del punto dati in ingresso.
Le sue prestazioni possono essere influenzate dalla scelta di K e della metrica di distanza, per cui è necessaria un'attenta regolazione dei parametri per ottenere risultati ottimali.

5.2 Indagine su dataset basato su *GTZAN* dataset

Il nuovo dataset fornito è basato su una famosa raccolta di file audio utilizzata per allenare classificatori al riconoscimento dei generi musicali. La struttura è del tutto simile al dataset di prova. Contiene più di 18 ore di registrazione tra file audio originali e file audio ribaltati digitalmente. Inoltre è già diviso in *train set* e *test set*. Ogni file audio è in formato *.wav*, campionato a 22050 Hz, profondità a 16 bit, durata di 30 secondi.

La pre-elaborazione è analoga a quella dell'indagine precedente: i file vengono segmentati in porzioni della durata di cinque secondi (producendo un dataset con oltre 8˙400 elementi), un valore *RMS* e quaranta *MFCC* vengono estratti per ogni porzione e uniti in un vettore di *features*, che viene associato al proprio *label*.
Il *train set* viene diviso ulteriormente in *train set* effettivo e *validation set* con rapporto 80/20.

Tentando di allenare *RandomForestClassifier* e *KneighborsClassifier* con iper-parametri di default sul nuovo *train set* si ottengono risultati pessimi (precisione < 10%). Si procede quindi a un *tuning* degli iper-pararametri per ottimizzare i modelli.
A questo scopo si utilizza la funzione *gridsearchCV* di *scikit-learn*. Questa permette di definire un dizionario popolato di liste di valori da provare per gli iper-parametri di un modello. Al momento di un *fit*, questa griglia tenta tutte le combinazioni di valori forniti per iper-parametro e aggiorna un campo *best_params_* con la combinazione che porta alla precisione maggiore sul *validation set*.
Operare un *gridsearchCV* con un dizionario corposo aumenta esponenzialmente il numero di *fit* del modello da calcolare. Il tempo che è stato richiesto per svolgere gridsearchCV per entrambi i modelli su un ampio dizionario di valori è stato superiore alle 5 ore per un elaboratore con RAM di 4 Gb.

I migliori parametri trovati per ciascun modello sono stati utilizzati per riallenare ciascuno dei due modelli. Come ci si aspetta il risultato sul *validation set* aumenta drasticamente, anche se purtroppo non oltre il 60% di precisione.

Per verificare che questi risultati sono veritieri e non dipendano dagli specifici set di dati utilizzati, si ripete l'allenamento dei due modelli, questa volta su un *train set* composto dall'unione dei precedenti *train set* e *validation set*. Operando il *fit* dei modelli sul *train set*, si ottengono valori di precisione del 16.67% per *KNeighborsClassifier* e 9.87% per *RandomForestClassifier*, rivelando l'incostanza e l'incapacità dei due modelli anche dopo un primo tentativo di ottimizzazione degli iper-parametri.

5.3 Discussione dei risultati

Ulteriori prove nella ricerca di parametri portano sempre a risultati insoddisfacenti, rivelando come non sia in questo caso tanto rilevante la scelta del modello e dei suoi parametri quanto la scelta delle features che descrivono le caratteristiche di un segmento audio. Un singolo valore RMS e quaranta MFCC non sono decisamente sufficienti a definire caratteristiche contraddistintive dell'orientamento di lettura del nastro.
Ulteriori esperienze di sviluppo di un modello per il riconoscimento di traccie lette inversamente possono scartare questa scelta di *features*.

6 Bibliografia

[1] M. Bosi, S. Canazza, A. Russo, N. Pretto e L. Chiariglione, *An MPAI/IEEE International Standard for Audio: Overview of CAE Audio Recording Preservation (ARP) Technology*, AES Conference Paper, 2023.
http://www.aes.org/e-lib/browse.cfm?elib=22136

[2] MPAI, *Introduction to MPAI-CAE*, 2022, N549.
https://mpai.community/wp-content/uploads/2022/02/N549-Introduction-to-MPAI-CAE.docx

[3] MPAI, *Technical Specification MPAI Contextbased Audio Enhancement (MPAI-CAE)*, Version V2, 2024.
https://mpai.community/wp-content/uploads/2024/02/Technical-Specification-Context-based-Audio-Enhancement-MPAI-CAE-V2.pdf

[4] MPAI, *MPAI-CAE Use Cases and Functional Requirements*, 2021, N151.
https://mpai.community/wp-content/uploads/2021/03/N151-MPAI-CAE-Use-Cases-and-Functional-Requirements.docx

[5] F. Bressan, V. Burini, E. Micheloni, A. Rodà, R. L. Hess e S. Canazza, *Reading Tapes Backwards: A Legitimate Approach to Saving Time and Money in Digitization Projects?*, Appl. Sci. 2021, 11, 7092.
doi: https://doi.org/10.3390/app11157092

[6] F. Bressan, R. L. Hess, *Non-standard track configuration in historical audio recordings: Technical and philological consequences for preservation*. Fontes Artis Music. 2020, 67, 229–252.
https://muse.jhu.edu/article/768869

7 Appendici

A Figure di Riferimento

Fig.1: metodologia adottata per CAE-ARP (tratto da [1], p.3)

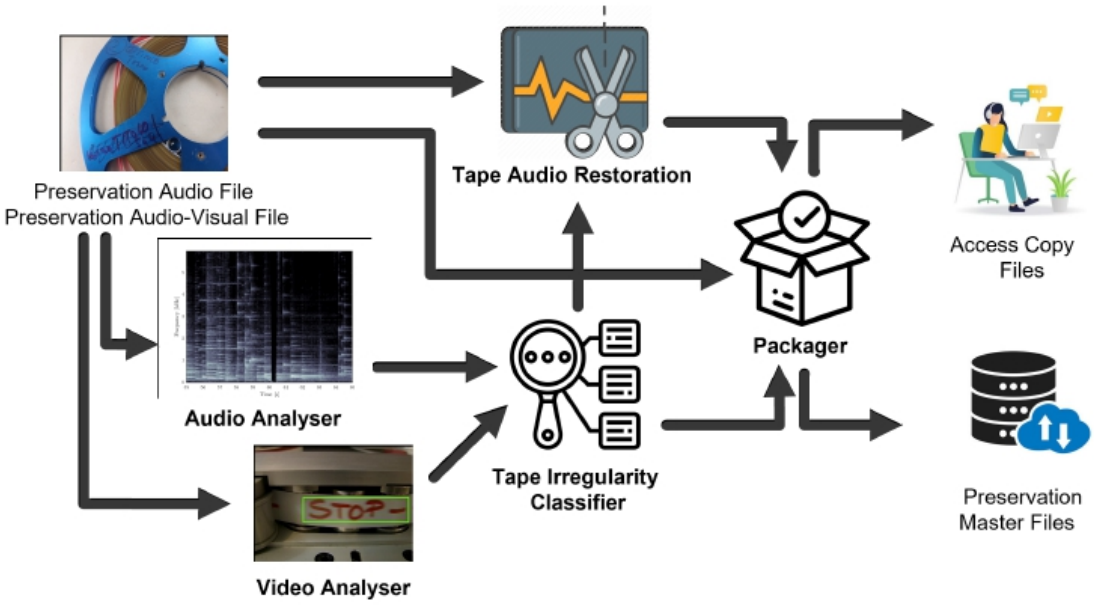

Fig.2: architettura di riferimento AIW CAE-ARP (tratto da [3], p.14)

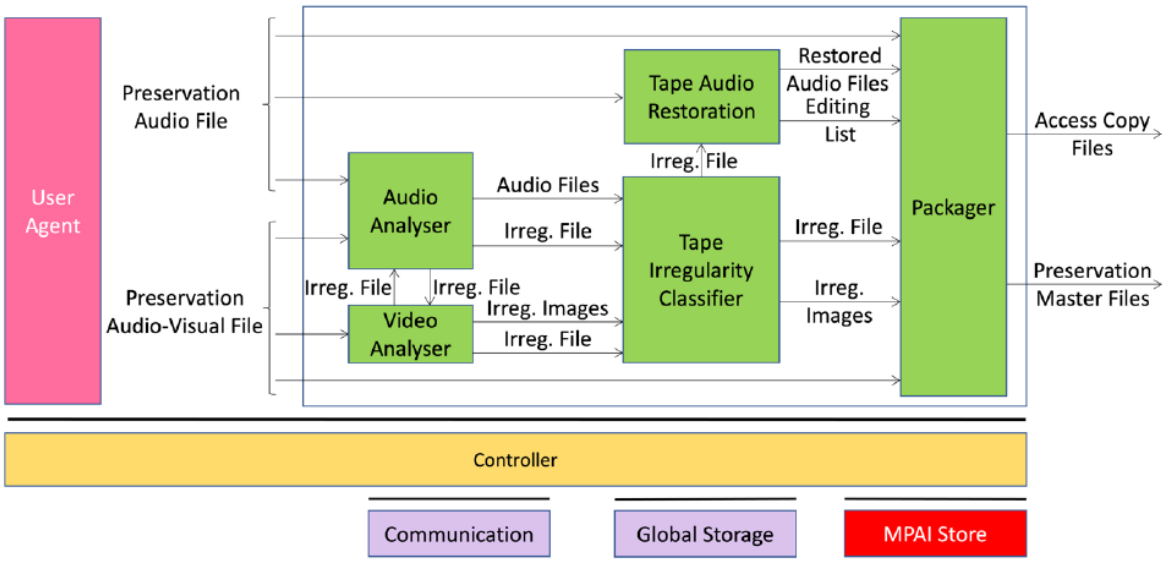

Fig.3: configurazioni standard tracce su nastro magnetico (tratto da [6], p.2)

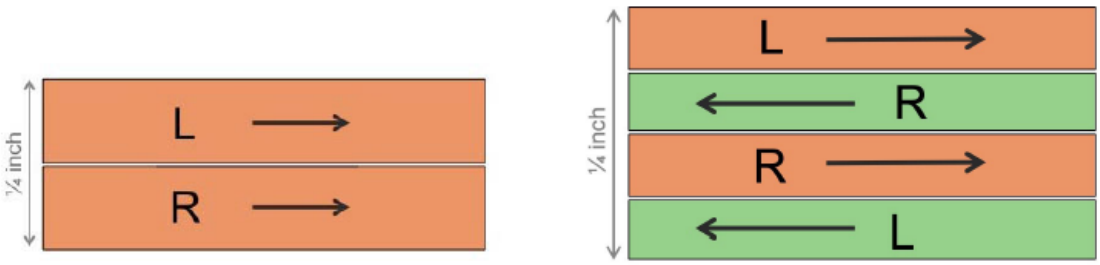

Fig.4: schema a blocchi di un registratore per bobine aperte (adattato da [6], p.4)

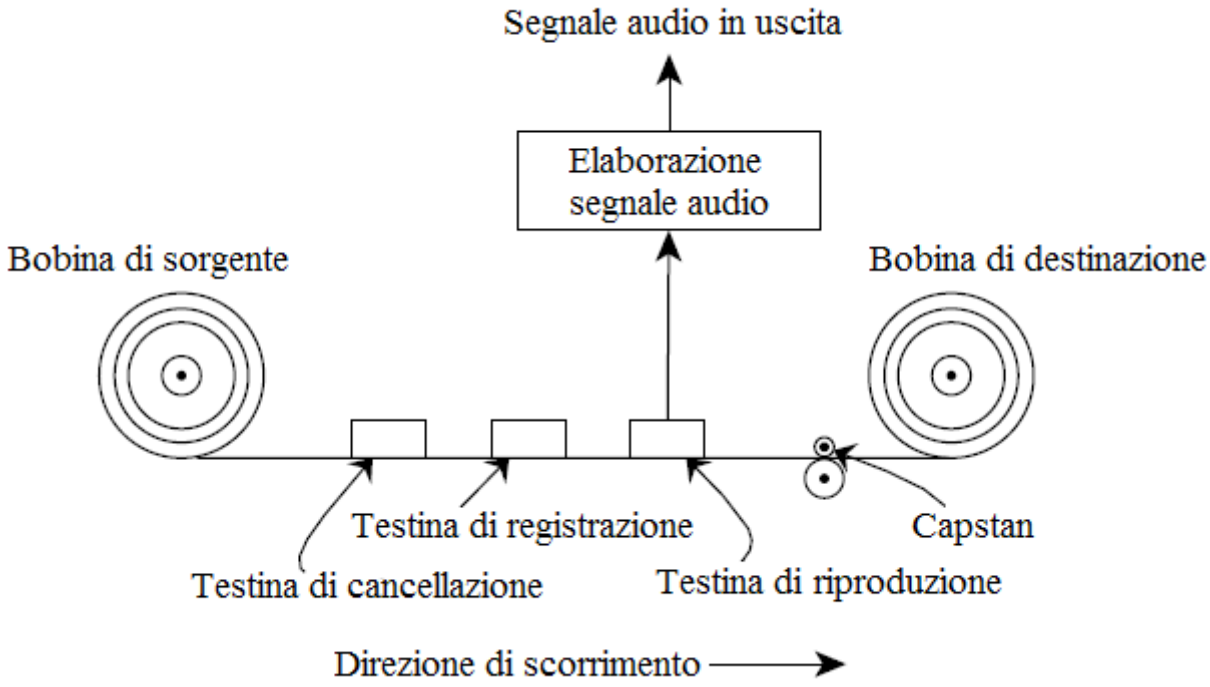

Fig.5: esempio di una coppia di forme d'onda di segmenti audio mono provenienti da una registrazione riprodotta con orientamento normale (in alto) e orientamento inverso (in basso); figura realizzata caricando le clip in *Audacity*.