RICONOSCIMENTO IN TEMPO REALE DI TECNICHE ESPRESSIVE PER CHITARRA SU EMBEDDED COMPUTERS

Domenico Stefani
Creative, Intelligent & Multisensory Interactions Laboratory
Università di Trento
domenico.stefani@unitn.it

ABSTRACT

Negli ultimi decenni dello scorso secolo, le innovazioni nel campo dell'elettronica analogica hanno portato alla creazione di una serie di strumenti-controller basati sulla chitarra che potevano produrre un'ampia gamma di suoni controllando direttamente dei sintetizzatori audio. Nonostante i vari miglioramenti portati a questi sistemi negli anni, essi sono sempre stati limitati al tracciamento della frequenza e intensità nel tempo delle note suonate, mancando di considerare le sottili tecniche espressive generalmente usate dai chitarristi per mutare il suono dello strumento. In questo documento presento la mia ricerca, che si propone di utilizzare le tecniche più moderne del deep learning per permettere il riconoscimento in tempo reale della tecnica espressiva usata su di una chitarra. Particolare attenzione verrà dedicata a proporre implementazioni su dispositivi embedded e a considerare le limitazioni di questi, di maniera da poter permettere la creazione di nuovi strumenti intelligenti dove l'analisi del segnale e sintesi di nuovi suoni venga svolta in maniera auto-contenuta. Assieme ad una descrizione della ricerca e della metodologia utilizzata, vengono presentati i risultati ottenuti fino ad ora nel deep learning per classificazione in tempo reale di audio, classificazione di tecniche espressive e onset detection. Infine, vengono presentate le linee di ricerca che verranno seguite nel futuro prossimo.

1. INTRODUZIONE

L'evoluzione della tecnologia audio degli ultimi cinquanta anni ha reso possibile l'uso della chitarra per produrre un'ampia gamma di nuovi suoni, generati controllando direttamente sintetizzatori audio. Questa famiglia di strumenti-controller ha potuto beneficiare da diverse innovazioni elettroniche, che hanno portato ad un miglioramento dei processi di tracciamento dell'*intonazione* ed evoluzione temporale delle note suonate dal chitarrista. Questo ha poi permesso di controllare dei sintetizzatori audio usando la chitarra al posto di una tastiera. Tuttavia, codificare la performance di un chitarrista tramite questi unici due parametri risulta limitante, in quanto manca di considerare le differenze timbriche introdotte tramite l'uso di varie tecniche espressive, proprie delle performance di ogni chitarrista. Alcuni esempi sono l'uso di diverse posizioni del plettro, gli armonici, mutare le corde o il vibrato. La mancanza di riconoscimento automatico di queste tecniche porta ad avere un controllo ridotto sulla generazione del suono sintetico, oltre che ad un'interazione atipica tra il chitarrista e strumento, dove l'uso di ogni tecnica espressiva non ha effetto sul suono e viene generalmente evitato. Questo porta gli utenti di questo tipo di controller a dover ricorrere interfacce fisiche convenzionali di sintetizzatore per modificare il suono sintetico, staccando le mani dallo strumento invece di utilizzare le tecniche espressive che sono abituati ad utilizzare.

Questo articolo presenta la mia ricerca, sullo sviluppo di modelli di deep learning che possano riconoscere in tempo-reale le tecniche espressive usate da un chitarrista, così che questa informazione possa essere riutilizzata per controllare algoritmi di sintesi sonora o diversi sistemi per performance audio-video o effettistica da palco come sistemi di illuminazione. I requisiti principali di questi sistemi sono l'accuratezza del riconoscimento, la ridotta latenza nell'esecuzione di questi algoritmi in tempo reale. In aggiunta, la mia ricerca si concentra principalmente sull'esecuzione di questi sistemi in dispositivi embedded come piccoli single-board computers che possano essere integrati direttamente all'interno di strumenti musicali, risultando in nuovi strumenti musicali con intelligenza "incorporata", come la famiglia degli smart musical instruments [1].

2. LAVORI CORRELATI

Questa ricerca si svolge nel campo del Music Information Retrieval (MIR), campo che si concentra sull'estrazione di informazioni dalla musica. Alcuni esempi di sfide che vengono affrontate nel campo del MIR includono la trascrizione automatica di performance musicali [2], beat-detection [3], onset detection [4], e classificazione del genere musicale [5].

Nonostante questo, la ricerca in questo campo si è sempre principalmente concentrata nell'ideazione di metodi offline (quindi non-in-tempo-reale) che vendono applicati all'analisi di grandi basi di dati, dove la principale metrica di valutazione è l'accuratezza del sistema creato, mentre generalmente non viene considerato il tempo di esecuzione. Sebbene questo permetta di sperimentare potenti metodi e algoritmi che non debbano necessariamente essere veloci o completare l'esecuzione in ridotti tempi prestabiliti, l'utilizzo di sistemi prettamente offline può essere applicato

solo a contesti offline come l'analisi di basi di dati, mentre ogni tipo di sistema sviluppato per performance musicali deve necessariamente essere eseguito in tempo reale.

Il campo del MIR comprende anche una serie di metodi per il riconoscimento delle tecniche espressive, in particolare della chitarra, proposti negli anni [6], [7], [8], [9]. Tuttavia, in maniera simile alla gran parte dei sistemi di MIR esplorati in ricerca, anche i vari approcci sopracitati sono basati sull'esecuzione offline, quindi inadatti all'utilizzo in tempo reale, dove il riconoscimento della tecnica espressiva deve essere ottenuto in pochi millisecondi dall'inizio (anche detto *onset*) della nota relativa [10]. In particolare, Moore [11] descrive come $30ms$ sia il massimo intervallo fra due toni complessi, oltre il quale il sistema uditivo umano possa percepire i due come distinti. Questo offre un massimo limite empirico alla latenza di un sistema controller-sintetizzatore.

Con alcune eccezioni limitate (es. Reboursiere et al. [12]), il riconoscimento in tempo reale delle tecniche espressive per chitarra resta un campo largamente inesplorato, dove i recenti avanzamenti nel campo del deep learning possono dare risultati utili. In aggiunta, le piattaforme embedded per sistemi audio rese disponibili negli ultimi anni ([13], [14]) hanno aperto la possibilità di ideare strumenti intelligenti [1] che, tra le varie caratteristiche, sono definiti dalla presenza di *embedded intelligence*. Questa si riferisce alla capacità di processare l'audio dello strumento all'interno dello strumento stesso.

3. PROGRESSO E RISULTATI

Il problema affrontato in questa ricerca, ovvero il riconoscimento in tempo reale e su dispositivi embedded delle tecniche espressive per chitarra, offre una serie di sfide. Prima di tutto, in maniera simile agli approcci offline, il riconoscimento della tecnica deve essere accurato. Tuttavia, il riconoscimento deve anche avvenire con un ritardo quanto più minimo da quando il musicista suona una nota con lo strumento, di modo che sia possibile riutilizzare questa informazione per sintetizzare suoni che vengano percepiti come contemporanei alla nota originale. Questo limita la gamma di algoritmi che possono essere utilizzati ai più veloci, e richiede di trovare un compromesso con l'accuratezza del riconoscimento. Infine, eseguire simili sistemi di riconoscimento su di dispositivi embedded pone ulteriori limiti sulla quantità di operazioni che possono essere eseguite, rispetto a più potenti personal computers.

3.1 Primi studi e risultati

La parte iniziale di questa ricerca è stata concentrata sulla raccolta di dati audio per le tecniche espressive più comuni e lo sviluppo di una pipeline di classificazione a più stadi.

La pipeline, o catena di classificazione è composta da un onset detector, una serie di estrattori di proprietà timbriche del segnale (features) e infine una rete neurale che assolve il compito di classificatore, processando le features estratte dal segnale e calcolando a quale tecnica espressiva corrisponde la nota suonata. L'estrazione separata delle features richiede una ridotta potenza computazionale, contrariamente alle pratiche adottate nell'offline end-to-end deep learning, e permette di ottenere un sistema eseguibile in tempo reale. Questo passo si rivela fondamentale per un sistema che deve essere eseguito in tempo reale con limiti imposti dalla piattaforma hardware designata. Le features utilizzate includono Mel-frequency cepstral coefficients (MFCC), Bark-frequency cepstral coefficients (BFCC) e Real Cepstrum. Figura 1 mostra lo schema a blocchi del sistema di classificazione e alcune possibili applicazioni.

Per l'addestramento e il testing del classificatore si è resa necessaria la registrazione di un dataset di tracce audio, registrate direttamente dai pickup interni di cinque diverse chitarre, grazie ad altrettanti chitarristi con esperienza. È stato scelto di registrare una serie esaustiva di suoni (oltre 20 mila note in totale) con dodici diverse tecniche espressive suonate sulla chitarra acustica. Inoltre, le tecniche prese in considerazione sono in parte tecniche "convenzionali' suonate sulle corde (es. palm-mute, armonici, diverse posizioni del plettro), e in parte tecniche percussive che comportano l'uso del corpo della chitarra acustica come strumento percussivo.

Una demo preliminare del sistema è stata presentata in [15], mentre il primo classificatore è stato presentato in [16]. Data la complessità introdotta dai requisiti di bassa latenza ed esecuzione embedded, il problema è stato ridotto alla classificazione di singole note (monofonia).

Il sistema creato è in grado di distinguere tra tecniche percussive e convenzionali (classificazione binaria) con un' accuratezza del 99.2%, mentre ha raggiunto una simile accuratezza di 99.1% nel distinguere quattro diverse tecniche percussive e una macro-classe comprendente tutte le tecniche convenzionali in analisi. Al contrario, il problema di classificazione completo, quindi la categorizzazione tra dodici tecniche, si è rivelato considerevolmente più complesso a causa delle più sottili differenze timbriche nell'attacco delle tecniche suonate sulle corde. I risultati di questo primo sistema con il problema completo hanno infatti raggiunto solo un'accuratezza del 56.5%.

Tutti i risultati di classificazione sono stati ottenuti con una latenza media di $30.7\ ms$, vicina ai minimi requisiti discussi da Moore [11], ma migliorabile in un'iterazione successiva del sistema. La latenza media totale è il risultato dalla somma delle latenze di onset detection, allineamento della finestra di estrazione, computazione delle features, e tempo di esecuzione del classificatore, che corrispondono rispettivamente a $19.0\ ms$, $7.77\ ms$, $0.78\ ms$, e $3.15\ ms$.

Il valore aggiunto del lavoro presentato consiste nell'implementazione dell'intero sistema in un Raspberry PI 4: un single-board computer che per l'occasione è stato fornito del sistema operativo open-source per processing di audio in tempo reale, Elk OS [17]. Figura 2 mostra i dispositivi che compongono il sistema. Il segnale audio in input viene campionato a 48 kHz e raggruppato in buffer di 64 campioni per una ridotta latenza.

3.2 Raffinamento della pipeline di classificazione

Nonostante alcune delle limitazioni e i modesti risultati presentati in [16], la pipeline di classificazione ha dimostrato di essere funzionale, eseguibile in tempo reale e ra-

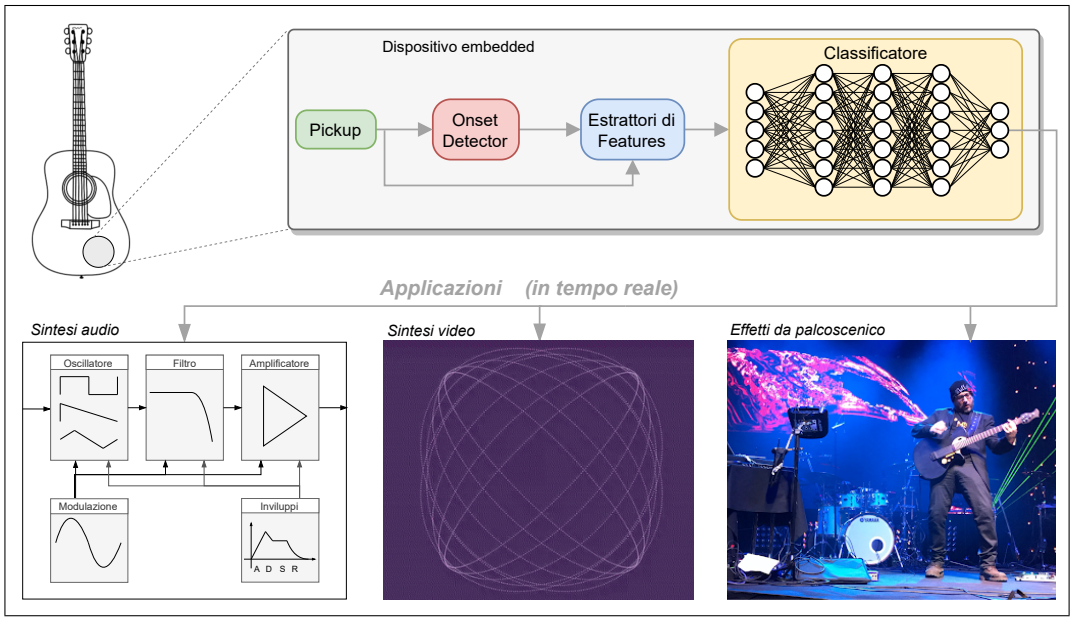

Figure 1. *Sistema di classificazione della tecnica espressiva in tempo reale e possibili applicazioni. Il segnale audio registrato dai pickup dello strumento viene analizzato da un onset detector, che attiva il calcolo di una serie di features timbriche quando viene identificata una nota nel segnale. In seguito, le features calcolate vengono utilizzate da una rete neurale profonda per categorizzare la tecnica espressiva usata. L'immagine mostra alcune possibili applicazioni in tempo reale che non si limitano alla sintesi audio, ma includono la sintesi video e il controllo di effettistica da palco come l'illuminazione.*

Figure 2. *Prototipo del classificatore di tecniche espressive. Il segnale audio dei pickup della chitarra viene passato ad un Raspberry PI 4 tramite un interfaccia audio (Elk PI Hat). Il risultato della classificazione viene comunicato tramite le cuffie con dei semplici toni (sinusoidali) con una diversa frequenza in base alla tecnica espressiva usata.*

gionevolmente accurata con il problema più semplice di classificazione percussiva, anche a fronte di una regolazione poco fine dei parametri di tutta la catena di classificazione.

Il lavoro di ricerca è quindi continuato con il raffinamento dei vari stadi della classificazione, a partire dal riconoscimento dell'inizio di una nota in tempo reale (Onset Detection). In [18] abbiamo presentato un approccio basato su algoritmi genetici per ottimizzare i parametri iniziali di vari onset parametrici, di maniera da raggiungere il compromesso ottimale tra l'accuratezza dell'individuazione dell'onset e la sua latenza. Questo ha permesso contemporaneamente di ridurre gli errori di individuazione (falsi positivi e negativi) e di ridurre la latenza di individuazione, ma soprattutto la sua variabilità. Ridurre la variabilità della latenza di individuazione dell'onset ha permesso di apprezzare un miglioramento sostanziale della performance del classificatore neurale, derivata da un allineamento più accurato della finestra di estrazione delle features, e di un aumento della qualità di queste ultime.

In secondo luogo, l'estrazione delle features è stata abbinata ad un algoritmo di selezione automatica di features. Questa ha permesso di mantenere solo i coefficienti più importanti e ridurre notevolmente il rumore delle features passate al classificatore neurale, che a sua volta ha permesso di ridurre le dimensioni della rete neurale, rimuovendo una grande parte di neuroni che altrimenti avrebbero dovuto essere dedicati dalla rete alla selezione delle features.

Infine, un lavoro più fine di progettazione e raffinamento del classificatore neurale è attualmente in corso, con soddisfacenti risultati preliminari superiori al 90% di accuratezza con una media di 20 millisecondi di latenza tra onset e risultati (10 millisecondi inferiore al lavoro presentato in precedenza).

La necessità di ottimizzare la latenza della classificazione ha portato anche alla comparazione di quattro diversi sistemi di esecuzione di modelli di Deep Learning su dispositivi con risorse computazionali limitate [19].

4. CONCLUSIONI E LAVORI FUTURI

Questo documento ha riportato il progresso della mia ricerca diretta al riconoscimento automatico in tempo reale delle tecniche espressive per chitarra su dispositivi embed-

ded. I risultati ottenuti fino ad ora hanno aperto la possibilità per alcune linee di ricerca per il futuro che verranno ora discusse brevemente.

Il lavoro attuale sul miglioramento della classificazione monofonica delle tecniche espressive per chitarra sarà completato con una serie di test tecnici e percettivi per verificare l'accuratezza del sistema e l'impercettibilità del ritardo tra le note suonate dal musicista e i suoni sintetici generati secondo la tecnica classificata.

In seguito, questa ricerca mirerà a coprire le limitazioni del sistema attuale, in particolare la classificazione monofonica. L'estensione alla classificazione polifonica della tecnica espressiva sarà affrontata tramite l'uso di un pickup esafonico, che permette di avere un segnale audio separato per ogni corda, su cui poter effettuare la classificazione.

Inoltre, sperimenterò le possibilità offerte dai più recenti dispositivi di accelerazione hardware come Tensor Processing Unit (TPU), le quali potrebbero permettere di eseguire il modello di classificazione più velocemente, oppure di avere modelli più complessi senza introdurre latenza eccessiva.

In maniera simile, sarà possibile utilizzare la tecnica del transfer-learning di modo da adattare un modello generico di classificazione della tecnica ad una chitarra specifica.

5. RIFERIMENTI

[1] L. Turchet, "Smart Musical Instruments: vision, design principles, and future directions," *IEEE Access*, vol. 7, pp. 8944–8963, 2019.

[2] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 927–939, 2016.

[3] S. Böck and M. Schedl, "Enhanced beat tracking with context-aware neural networks," in *Proc. 14th Int. Conf. on Digital Audio Effects (DAFx-11)*, pp. 135–139, 2011.

[4] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[5] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[6] T. H. Özaslan, E. Guaus, E. Palacios, and J. L. Arcos, "Attack based articulation analysis of nylon string guitar," in *Proc. 7th Int. Symposium on Computer Music Modeling and Retrieval (CMMR)*, 2010.

[7] C. Kehling, J. Abeßer, C. Dittmar, and G. Schuller, "Automatic tablature transcription of electric guitar recordings by estimation of score- and instrument-related parameters," *Proc. 17th Int. Conf. on Digital Audio Effects (DAFx 2014)*, pp. 1–8, 2014.

[8] Y. P. Chen, L. Su, and Y. H. Yang, "Electric guitar playing technique detection in real-world recordings based on F0 sequence pattern recognition," *Proc. 16th Int. Society for Music Information Retrieval Conference (ISMIR)*, pp. 708–714, 2015.

[9] J. Abeßer, H. Lukashevich, and G. Schuller, "Feature-based extraction of plucking and expression styles of the electric bass guitar," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2290–2293, 2010.

[10] A. McPherson, R. Jack, and G. Moro, "Action-sound latency: Are our tools fast enough?," in *Proc. Int. Conference on New Interfaces for Musical Expression*, pp. 20–25, 2016.

[11] B. C. J. Moore, *An introduction to the psychology of hearing*. Brill, 2012.

[12] L. Reboursiere, O. Lähdeoja, R. Chessini, T. Drugman, S. Dupont, C. Picard, and N. Riche, "Guitar as controller," in *Journal of The Acoustical Society of America*, vol. 80, pp. 41–54, 01 2011.

[13] L. Vignati, S. Zambon, and L. Turchet, "A comparison of real-time linux-based architectures for embedded musical applications," *Journal of the Audio Engineering Society*, vol. 70, no. 1/2, pp. 83–93, 2022.

[14] E. Meneses, J. Wang, S. Freire, and M. Wanderley, "A comparison of open-source linux frameworks for an augmented musical instrument implementation," in *Proc. Int. Conf. on New Interfaces for Musical Expression*, pp. 222–227, June 2019.

[15] D. Stefani and L. Turchet, "Demo of the timbreid-vst plugin for embedded real-time classification of individual musical instruments timbres," in *Proc. 27th Conf. of Open Innovations Association (FRUCT)*, vol. 2, pp. 412–413, 2020.

[16] D. Stefani and L. Turchet, "On the Challenges of Embedded Real-Time Music Information Retrieval," in *Proceedings of the 25-th Int. Conf. on Digital Audio Effects (DAFx20in22)*, vol. 3, pp. 177–184, Sept. 2022.

[17] L. Turchet and C. Fischione, "Elk Audio OS: an open source operating system for the Internet of Musical Things," *ACM Transactions on the Internet of Things*, vol. 2, no. 2, pp. 1–18, 2021.

[18] D. Stefani and L. Turchet, "Bio-Inspired Optimization of Parametric Onset Detectors," in *Proc. 24th Int. Conf. on Digital Audio Effects (DAFx20in21)*, vol. 2, pp. 268–275, Sept. 2021.

[19] D. Stefani, S. Peroni, and L. Turchet, "A Comparison of Deep Learning Inference Engines for Embedded Real-Time Audio Classification," in *Proceedings of the 25-th Int. Conf. on Digital Audio Effects (DAFx20in22)*, vol. 3, pp. 256–263, Sept. 2022.