National Technical University of Athens
School of Electrical and Computer Engineering
Division of Signals, Control and Robotics

Multicultural Representation Learning for Music Signal Analysis

Ph.D. Dissertation

of

Charilaos Papaioannou

Supervisor: Alexandros Potamianos
Associate Professor, NTUA

Athens, August 2025

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Signals, Control and Robotics

Multicultural Representation Learning for Music Signal Analysis

Ph.D. Dissertation

of

Charilaos Papaioannou

Supervision Committee : Alexandros Potamianos (NTUA)
Petros Maragos (NTUA)
Aggelos Pikrakis (University of Piraeus)

Approved by the examination committee on 27th August 2025.

(Signature)	*(Signature)*	*(Signature)*	*(Signature)*
.....	.....	.....	.....
Alexandros Potamianos	Petros Maragos	Aggelos Pikrakis	Emmanouil Benetos
Associate Professor	Professor	Associate Professor	Associate Professor
NTUA	NTUA	University of Piraeus	Queen Mary University of London

(Signature)	*(Signature)*	*(Signature)*
.....	.....	.....
Athanasios Rontogiannis	Constantinos Tzafestas	Gerasimos Potamianos
Associate Professor	Associate Professor	Associate Professor
NTUA	NTUA	University of Thesally

Athens, August 2025

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Signals, Control and Robotics

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism.

(Signature)

.....
Charilaos Papaioannou
PhD, National Technical University of Athens

27th August 2025

Περίληψη

Η έρευνα στον τομέα της Ανάκτησης Πληροφορίας από Μουσική (Music Information Retrieval – MIR) έχει παραδοσιακά επικεντρωθεί στις δυτικές μουσικές παραδόσεις, δημιουργώντας ένα σημαντικό κενό στις υπολογιστικές προσεγγίσεις για τις ποικίλες μουσικές κουλτούρες του κόσμου. Η παρούσα διατριβή στοχεύει στην κάλυψη αυτού του κενού, αναπτύσσοντας και αξιολογώντας μεθόδους για την πολυπολιτισμική αναπαράσταση της μουσικής, με σκοπό τη δημιουργία πιο "πολιτισμικά ενήμερων" υπολογιστικών προσεγγίσεων που μπορούν να αποτυπώνουν και να αναλύουν αποτελεσματικά τα ιδιαίτερα χαρακτηριστικά διαφορετικών μουσικών παραδόσεων.

Η έρευνα ξεκινά με την ανάπτυξη του συνόλου δεδομένων Lyra, μιας ολοκληρωμένης συλλογής ελληνικής παραδοσιακής μουσικής, η οποία περιλαμβάνει 1570 κομμάτια (περίπου 80 ώρες) με συνεπή ποιότητα ήχου και πλούσια μεταδεδομένα. Το σύνολο δεδομένων αυτό λειτουργεί ως βάση για τις επόμενες μελέτες, παρέχοντας έναν πολύτιμο πόρο για την υπολογιστική ανάλυση μιας μουσικής παράδοσης που ενσωματώνει στοιχεία τόσο από το δυτικό[1] όσο και από το μεσανατολικό/μεσογειακό μουσικό σύστημα. Πειράματα ταξινόμησης βάσης επιβεβαιώνουν τη χρησιμότητα του συνόλου δεδομένων για την αναγνώριση μουσικολογικών χαρακτηριστικών όπως το είδος, τα όργανα και η γεωγραφική προέλευση.

Αξιοποιώντας τη βάση αυτή, η διατριβή εξερευνά τη διαπολιτισμική μεταφορά γνώσης στη μουσική μέσω συστηματικής αξιολόγησης τριών μοντέλων βαθιών αναπαραστάσεων ήχου (deep audio embeddings) σε έξι σύνολα δεδομένων που καλύπτουν δυτικές, μεσογειακές και ινδικές μουσικές παραδόσεις. Μέσα από πειράματα μεταφοράς μάθησης, αναδεικνύεται ποιες πηγές δεδομένων προσφέρουν την πιο αποτελεσματική μεταφορά γνώσης σε κάθε περίπτωση, παρέχοντας υπολογιστικές ενδείξεις για τις μουσικές ομοιότητες μεταξύ κουλτούρων. Τα αποτελέσματα δείχνουν ότι μπορεί να επιτευχθεί ανταγωνιστική επίδοση σε όλους τους πολιτισμούς μέσω μεταφοράς μάθησης, με διαφορετικά μοτίβα μεταφοράς που αντανακλούν τις μουσικές σχέσεις.

Για την αντιμετώπιση των προκλήσεων που σχετίζονται με τον περιορισμένο όγκο επισημειωμένων δεδομένων στον χώρο της έρευνας για τη μουσική του κόσμου, η διατριβή εισάγει τα Label-Combination Prototypical Networks (LC-Protonets), μια καινοτόμο προσέγγιση μάθησης από λίγα παραδείγματα που δημιουργεί πρωτότυπα για συνδυασμούς κατηγοριών αντί για μεμονωμένες κατηγορίες. Η μέθοδος αυτή βελτιώνει σημαντικά την απόδοση σε διαφορετικά σύνολα μουσικών δεδομένων και σενάρια εκπαίδευσης, επιτρέποντας την ένταξη υποεκπροσωπούμενων κατηγοριών και μουσικών παραδόσεων στα υπολογιστικά μοντέλα, ακόμη και με ελάχιστα παραδείγματα. Η ενσωμάτωση προεκπαιδευμένων μοντέλων ενισχύει περαιτέρω την απόδοση, δείχνοντας τη δυναμική του συνδυασμού μεταφοράς μάθησης και μάθησης από λίγα παραδείγματα για πολυπολιτισμική ανάλυση μουσικής.

Στη συνέχεια, η διατριβή αξιολογεί πέντε σύγχρονα θεμελιώδη μοντέλα (state-of-the-art foundation models) σε έξι μουσικά σύνολα δεδομένων που καλύπτουν δυτικές, μεσογειακές/μεσανατολικές

[1] Στην παρούσα διατριβή, ο όρος "δυτικός" αναφέρεται συγκεκριμένα σε μουσικές παραδόσεις που αναπτύχθηκαν εντός των ευρωπαϊκών κλασικών, λειτουργικών και παραδοσιακών συστημάτων, τα οποία αργότερα επεκτάθηκαν με προσαρμογές στη βόρεια Αμερική, και χαρακτηρίζονται από συγκεκριμένους μελωδικούς τρόπους, αρμονικές δομές και χαρακτηριστικές οργανολογικές διαμορφώσεις.

και ινδικές παραδόσεις, χρησιμοποιώντας συμπληρωματικές μεθοδολογίες όπως probing, επιβλεπόμενη εκπαίδευση (supervised fine-tuning) και μάθηση από λίγα παραδείγματα. Η εκτενής αυτή αξιολόγηση αποκαλύπτει τόσο υποσχόμενες διαπολιτισμικές δυνατότητες όσο και σημαντικούς περιορισμούς, με την απόδοση να μειώνεται για πολιτισμικά απομακρυσμένες παραδόσεις, ειδικά σε σενάρια χαμηλών πόρων. Η έρευνα επιτυγχάνει κορυφαία αποτελέσματα σε πέντε από τα έξι αξιολογούμενα σύνολα δεδομένων, καταδεικνύοντας την αποτελεσματικότητα των θεμελιωδών μοντέλων για την κατανόηση της μουσικής του κόσμου, ενώ ταυτόχρονα αναδεικνύει υφιστάμενα κενά στην διαπολιτισμική αναπαράσταση της μουσικής.

Με βάση αυτά τα ευρήματα, η διατριβή παρουσιάζει το CultureMERT, ένα πολυπολιτισμικά προσαρμοσμένο θεμελιώδες μοντέλο, που αναπτύχθηκε μέσω μιας στρατηγικής συνεχιζόμενης προεκπαίδευσης δύο σταδίων σε ένα μείγμα 650 ωρών ελληνικής, τουρκικής και ινδικής μουσικής. Η προσέγγιση αυτή βελτιώνει σταθερά την απόδοση σε προβλήματα ταξινόμησης μη δυτικής μουσικής, περιορίζοντας παράλληλα τη μείωση της απόδοσης σε δυτικές παραδόσεις. Εξερευνάται επίσης η χρήση του task arithmetic ως εναλλακτική προσέγγιση πολυπολιτισμικής προσαρμογής, η οποία συγχωνεύει αποτελεσματικά μοντέλα προσαρμοσμένα σε μεμονωμένες κουλτούρες στον χώρο των βαρών, επιτυγχάνοντας συγκρίσιμη απόδοση με τη συνεχιζόμενη προεκπαίδευση, χωρίς να απαιτείται ταυτόχρονη πρόσβαση σε όλα τα πολιτισμικά σύνολα δεδομένων.

Η τελευταία μελέτη παρουσιάζει μια ολοκληρωμένη ανάλυση της διαπολιτισμικής μουσικής ομοιότητας που γεφυρώνει την ανθρώπινη αντίληψη, τα χαρακτηριστικά επεξεργασίας σήματος και τα θεμελιώδη μοντέλα. Μέσω των απαντήσεων 125 συμμετεχόντων από διαφορετικές χώρες, συλλέχθηκαν η ομοιότητα 1130 ζευγών ηχητικών αποσπασμάτων από εννέα μουσικά σύνολα δεδομένων που καλύπτουν δυτικές, μεσογειακές, ινδικές και κινεζικές κουλτούρες. Η εργασία αυτή παρέχει εμπειρική βάση για την υπολογιστική αξιολόγηση της μουσικής ομοιότητας. Κάθε ζεύγος αξιολογήθηκε από τους συμμετέχοντες σε τρεις διαστάσεις: συνολική μουσική ομοιότητα, πολιτισμική ομοιότητα και ομοιότητα σε επίπεδο προτάσεων (recommendation-level). Συστηματική σύγκριση των ανθρώπινων αξιολογήσεων με χαρακτηριστικά επεξεργασίας σήματος που καλύπτουν διαστάσεις ρυθμού, μελωδίας, αρμονίας και ηχοχρώματος, καθώς και με αναπαραστάσεις από επτά θεμελιώδη μοντέλα, δείχνει ότι τα τελευταία επιτυγχάνουν την ισχυρότερη ευθυγράμμιση με την ανθρώπινη αντίληψη (triplet agreement ≈ 0.65), ενώ η μελωδία παρουσιάζει σταθερά την καλύτερη απόδοση μεταξύ των χαρακτηριστικών σήματος. Η ανάλυση των ανθρώπινων αξιολογήσεων αναδεικνύει τη μελωδία ως τη σημαντικότερη φέρουσα διάσταση της ομοιότητας, ενώ στα θεμελιώδη μοντέλα το ηχόχρωμα εμφανίζεται εξίσου ή και πιο σημαντικό σε ορισμένες περιπτώσεις.

Καθόλη τη διάρκεια των ερευνών, η διατριβή υιοθετεί μια υπολογιστική, βασισμένη στα δεδομένα, προσέγγιση για τη μελέτη των διαπολιτισμικών μουσικών σχέσεων, χρησιμοποιώντας μοντέλα βαθιάς μάθησης ώστε τα πρότυπα να προκύπτουν απευθείας από τα ηχητικά δεδομένα, χωρίς την επιβολή προκαθορισμένων αναλυτικών πλαισίων. Η μεθοδολογία αυτή καθιστά δυνατή την ανάπτυξη υπολογιστικών εργαλείων ικανών να προσαρμόζονται στα ιδιαίτερα χαρακτηριστικά διαφορετικών μουσικών συστημάτων, χωρίς να απαιτείται εκτεταμένη σχεδίαση χαρακτηριστικών με βάση ειδικές γνώσεις. Με την πρόοδο στην ανάπτυξη συνόλων δεδομένων, τη μεταφορά μάθησης, τη μάθηση από λίγα δείγματα εκπαίδευσης, την προσαρμογή θεμελιωδών μοντέλων και την αξιολόγηση με επίκεντρο τον άνθρωπο για την πολυπολιτισμική αναπαράσταση της μουσικής, η διατριβή συμβάλλει στην ανάπτυξη υπολογιστικών μεθοδολογιών για την ανάλυση ποικίλων μουσικών παραδόσεων. Το έργο αυτό διευκολύνει τη διαπολιτισμική σύγκριση και μεταφορά γνώσης όπου είναι εφικτό, προσφέροντας τελικά πληροφορίες για τη σχέση μεταξύ της ανθρώπινης διαπολιτισμικής μουσικής αντίληψης και της υπολογιστικής κατανόησης της μουσικής.

Λέξεις-κλειδιά: Ανάκτηση Πληροφορίας από Μουσική (MIR), Πολυπολιτισμική Αναπαράσταση Μουσικής, Υπολογιστική Εθνομουσικολογία, Διαπολιτισμική Μεταφορά Μάθησης, Προσαρμογή Θεμελιωδών Μοντέλων, Μάθηση από Λίγα Παραδείγματα, Διαπολιτισμική Μουσική Ομοιότητα, Ανθρώπινη Αντίληψη, Συστήματα με Πολιτισμική Ευαισθησία

Abstract

Music Information Retrieval (MIR) research has traditionally focused on Western musical traditions, creating a significant gap in computational approaches to diverse world music cultures. This dissertation addresses this gap by developing and evaluating methods for multicultural music representation learning, aiming to create more culturally aware computational approaches that can effectively capture and analyze the distinctive characteristics of various musical traditions.

The research begins with the development of the Lyra dataset, a comprehensive collection of Greek traditional and folk music comprising 1,570 pieces (approximately 80 hours) with consistent audio quality and rich metadata. This dataset serves as a foundation for subsequent investigations, offering a valuable resource for computational analysis of a musical tradition that incorporates elements from both Western[2] and Middle Eastern/Eastern Mediterranean musical systems. Baseline classification experiments demonstrate the dataset's utility for recognizing musicological attributes including genre, instrumentation, and geographical origin.

Building on this foundation, the dissertation explores cross-cultural knowledge transfer in music through systematic evaluation of three deep audio embedding models across six datasets spanning Western, Eastern Mediterranean, and Indian musical traditions. Through transfer learning experiments, this research reveals which source domains provide the most effective knowledge transfer for each target domain, offering insights into computational similarities between musical cultures. Results demonstrate that competitive performance can be achieved across all domains via transfer learning, with varying patterns of transferability that reflect musical relationships.

To address the challenges of limited annotated data in world music research, the dissertation introduces Label-Combination Prototypical Networks (LC-Protonets), a novel approach to multi-label few-shot learning that creates prototypes for label combinations rather than individual labels. This method significantly improves performance across diverse music datasets and training setups, enabling the inclusion of underrepresented tags and musical traditions in computational models even with minimal examples. The integration of pre-trained models further enhances performance, demonstrating the potential of combining transfer learning with few-shot learning for multicultural music analysis.

The dissertation then evaluates five state-of-the-art foundation models across six musical corpora spanning Western, Greek, Turkish, and Indian classical traditions, employing complementary methodologies including probing, supervised fine-tuning, and few-shot learning. This comprehensive evaluation reveals both promising cross-cultural capabilities and significant limitations, with performance declining for culturally distant traditions, particularly in low-resource scenarios. The research achieves state-of-the-art performance on five out of six evaluated datasets, demonstrating the effectiveness of foundation models for world music understanding while highlighting remaining gaps in universal music representation.

Building upon these findings, the dissertation introduces CultureMERT, a multi-culturally

[2]In this dissertation, "Western" refers specifically to musical traditions that developed within European classical, liturgical, and folk traditions, later extending to North American adaptations of these systems, characterized by particular tonal organizations, harmonic structures, and instrumental configurations.

adapted foundation model developed through a two-stage continual pre-training strategy on a 650-hour mix of Greek, Turkish, and Indian music. This approach consistently improves performance across diverse non-Western music tagging tasks while minimizing regression on Western benchmarks. The research also explores task arithmetic as an alternative approach to multi-cultural adaptation, effectively merging single-culture adapted models in weight space with comparable performance to continual pre-training but without requiring simultaneous access to all cultural datasets.

The final investigation presents a comprehensive analysis of cross-cultural music similarity that bridges human perception, signal processing features, and foundation models. Through collection of human annotations from 125 participants across diverse backgrounds, evaluating 1,130 unique audio pairs from nine musical datasets spanning Western, Middle Eastern, Indian, and Chinese cultures, this work provides empirical grounding for computational music similarity assessment. Each pair was assessed along three dimensions: overall musical similarity, cultural similarity, and recommendation-level similarity. Systematic comparison of human judgments against signal processing features covering rhythm, melody, harmony, and timbre dimensions, as well as representations from seven foundation models, reveals that foundation models achieve the strongest alignment with human perception (triplet agreement ≈ 0.65), while melody consistently demonstrates superior performance among signal processing features. Analysis of human ratings identifies melody as the most important perceptual dimension, while different patterns emerge for foundation models with timbre being equally or even more important in some cases.

Throughout these investigations, the dissertation employs a data-driven computational approach to studying cross-cultural musical relationships, using end-to-end deep learning models to allow patterns to emerge directly from the audio data rather than imposing predefined analytical frameworks. This methodology enables the development of computational tools capable of adapting to the distinctive characteristics of different musical systems without requiring extensive domain-specific feature engineering. By advancing dataset development, transfer learning, few-shot learning, foundation model adaptation, and human-centered evaluation for multicultural music representation learning, this dissertation contributes computational methodologies for analyzing diverse musical traditions. The work facilitates cross-cultural comparison and knowledge transfer where appropriate, ultimately providing insights into the relationship between human cross-cultural music perception and computational music understanding.

Key Terms: Music Information Retrieval (MIR), Multicultural Music Representation Learning, Computational Ethnomusicology, Cross-Cultural Transfer Learning, Foundation Model Adaptation, Few-Shot Learning, Cross-Cultural Music Similarity, Human perception, Culturally aware systems

Στον Κόσμο της Μουσικής

Ευχαριστίες - Αντί Προλόγου

Ήταν πριν από 7 χρόνια όταν, καθώς προσγειωνόταν το αεροπλάνο της επιστροφής από την Κύπρο, αποφάσισα να κάνω αίτηση για να ξεκινήσω διδακτορικό. Εργαζόμουν ήδη αλλά κάτι ήθελα να αλλάξω. Τι άραγε; Δεν είχα ιδέα για το ταξίδι που θα ξεκινούσα...

Στο σήμερα και βάζοντας τις τελευταίες πινελιές στο κείμενο της διατριβής, αισθάνομαι πολύ διαφορετικός από τότε, πιο σκεπτόμενος αλλά και πιο σκεπτικός. Βαθαίνοντας κανείς τη γνώση του δεν μπορεί παρά να δει την πολυπλοκότητα του όποιου μικρόκοσμου στον οποίο εστιάζει.

Η μεγάλη αυτή αλλαγή, ανεξάρτητη των διαστάσεων του βολικού και του άβολου, είναι μια ένδειξη μάθησης και άρα ζωής. Και τα μεγαλύτερα μαθήματα για μένα έρχονται από ανθρώπους και τη μεταξύ μας σχέση. Το κείμενο αυτό αναφέρεται σε πολλούς από αυτούς αλλά αφιερώνεται και σε εκείνους τους "ανώνυμους" που προσφέρουν άνευ σκοπού λίγη από την πνοή τους στους άλλους.

Ξεκινάω με τον επιβλέποντά μου, Αλέξανδρο Ποταμιάνο, που είναι ανοιχτός σε νέες ιδέες, έτοιμος να πιστέψει στην προοπτική ενός ανθρώπου που τις εκφράζει. Κρατάω πολλά πράγματα αλλά περισσότερο από όλα τη διάθεση για μελέτη ετερογενών θεμάτων με έμφαση σε εκείνα που αφορούν τον άνθρωπο. Κρατάω επίσης τη σύμπνοια στο χιούμορ που ισορροπεί μεταξύ ρομαντισμού και κυνισμού, είμαι λάτρης!

Συνεχίζω με τον Εμμανουήλ Μπενέτο, που υπήρξε καθοριστικός στο να υλοποιηθούν οι ιδέες και να μειωθεί, όσο γινόταν, η εντροπία τους. Η προσηνής και ρεαλιστική ματιά του λειτούργησε σαν φάρος σε καίρια σημεία της έρευνας που χαρακτηρίζονταν από αβεβαιότητα.

Ήμουν εξαιρετικά τυχερός που είχα στενή συνεργασία με τους δύο αυτούς εκπληκτικούς επιστήμονες, οι οποίοι αφιέρωσαν αμέτρητο χρόνο σε συζητήσεις σχετικές με την παρούσα εργασία.

Ευχαριστώ επίσης τον Γιάννη Βαλιάτζα, τον Θοδωρή Γιαννακόπουλο και τον Μάξιμο Καλιακάτσο-Παπακώστα για τη συνεργασία μας στα πρώτα στάδια της έρευνας. Τη Χριστίνα Αναγνωστοπούλου και το Τμήμα Μουσικολογίας του ΕΚΠΑ για τη βοήθεια τους. Τους φοιτητές της Σχολής ΗΜΜΥ που επίσης μας βοήθησαν εξαιρετικά στη συμπλήρωση ερωτηματολογίων. Τα άλλα δύο μέλη της τριμελούς επιτροπής, Πέτρο Μαραγκό και Άγγελο Πικράκη, αλλά και τα μέλη της επταμελούς επιτροπής που με τα σχόλιά τους βελτίωσαν το παρόν πόνημα.

Τους φοιτητές που συνεργαστήκαμε κατά την εκπόνηση της διπλωματικής τους εργασίας. Άρια, Παναγιώτη, Όλγα, Γεράσιμε, Άγγελε, Ανδρέα, Αλέξανδρε, η επαφή μου μαζί σας ήταν από τις πιο θετικές εμπειρίες του διδακτορικού. Ανθρώπους της ερευνητικής κοινότητας που λειτούργησαν κομβικά στην προώθηση της εργασίας ή στην διατήρηση της επιμονής για τη συνέχισή της. Rafael, Emilia, Yiit, Nazif, Sertan, Alastair, Shangda, Daniel, Iran, thank you!

Πριν περάσω στο πιο προσωπικό μέρος, ένα μεγάλο ευχαριστώ στην κοινότητα του Πολυτεχνείου, εργαζόμενοι στη Γραμματεία, στη Βιβλιοθήκη, στην Καθαριότητα, για τη φιλοξενία στις εγκαταστάσεις του ιδρύματος και το φιλότιμό τους για την επίλυση οποιουδήποτε θεμάτος.

Θεωρώ το σημαντικότερο για έναν άνθρωπο, το πλέγμα των σχέσεων που τον περιβάλλουν στην καθημερινότητα. Εκεί γεννιούνται οι κυματισμοί της ευτυχίας και της έμπνευσης και απορροφώνται οι κραδασμοί των προβλημάτων και των ανησυχιών. Ξεκινάω με τον Ευθύμη που μοιραστήκαμε στο εργαστήριο 2.1.2, τα ζόρια και τις ελπίδες του διδακτορικού. Τους συναδέλφους από εργαστήρια-

ξαδέρφια του δικού μας, Χρήστο και Βασίλη για τις κουβέντες πάνω στη μουσική και την τεχνολογία. Τον Άρη για τα αστεία του και τη γενναιοδωρία του στο να βοηθήσει τεχνικά οπουδήποτε του ζητηθεί. Τον Ηλία για την ηθική υποστήριξη, τις βόλτες σε Αθήνα, Ινδία και Οξφόρδη. Την "Initech" ομάδα (Βασίλη-Γιάννη-Γιώργο) για τα μηνύματά τους επί παντός επιστητού που λειτούργησαν σαν όαση ορισμένες δύσκολες μέρες. Τον Βαγγέλη, για τους απογευματινούς περιπάτους στην Πολυτεχνειούπολη, μιλώντας για τις σπουδές, τη δουλειά, τις σχέσεις, τη ζωή.

Ευχαριστώ τους γονείς μου, Στέφανο και Λέτα για την αναλυτική και την καλλιτεχνική σκέψη που μου μετέδωσαν. Την αδερφή μου Έφη για τη διαρκή εμπιστοσύνη της σε όποια προσωπική μου απόφαση. Τον αδερφό μου Χρήστο για τα αμέτρητα τηλεφωνήματα, τη στήριξή του, την καθαρή του σκέψη και τη διάθεσή του να βοηθήσει με την εμπειρία του και την καλή του προαίρεση.

Κλείνω με την Ήβη - σ' ευχαριστώ για το καθημερινό παρόν, την πίστη σου σε μένα, την ηρεμία σου, την αντίληψη σου, την καλοσύνη σου. Χωρίς εσένα, η εργασία αυτή δεν θα είχε ολοκληρωθεί.

Αθήνα, Αύγουστος 2025

Χάρης Παπαϊωάννου

Table of Contents

List of Figures

List of Tables

Εκτεταμένη Ελληνική Περίληψη

Ι Εισαγωγή και Κίνητρα

Η μουσική αποτελεί θεμελιώδες στοιχείο των ανθρώπινων πολιτισμών, αντικατοπτρίζοντας την ταυτότητα και τις παραδόσεις κοινοτήτων σε όλο τον κόσμο. Παρά την κεντρική της θέση στην ανθρώπινη εμπειρία, η υπολογιστική ανάλυση της μουσικής έχει επικεντρωθεί παραδοσιακά στις δυτικές μουσικές παραδόσεις, δημιουργώντας σημαντικά κενά στην κατανόηση και αναπαράσταση της παγκόσμιας μουσικής ποικιλομορφίας.

Ι.1 Το Πεδίο της Ανάκτησης Μουσικών Πληροφοριών

Η Ανάκτηση Μουσικών Πληροφοριών (Music Information Retrieval - MIR) έχει αναδειχθεί ως ένας δυναμικός διεπιστημονικός τομέας που εφαρμόζει υπολογιστικές μεθόδους για την κατανόηση, οργάνωση και πρόσβαση στο μουσικό περιεχόμενο. Οι τεχνολογίες MIR έχουν μεταμορφώσει τον τρόπο αλληλεπίδρασής μας με τη μουσική, επιτρέποντας εξατομικευμένες υπηρεσίες streaming, αυτοματοποιημένη κατηγοριοποίηση μουσικής και νέα δημιουργικά εργαλεία.

Ωστόσο, παρά την τεράστια πρόοδο που έχει σημειωθεί ο MIR κλάδος τις τελευταίες δεκαετίες, ένας σημαντικός περιορισμός παραμένει: η συντριπτική πλειονότητα των υπολογιστικών μοντέλων, συνόλων δεδομένων και πλαισίων αξιολόγησης είναι κυρίως επικεντρωμένα στις δυτικές μουσικές παραδόσεις. Μια πρόσφατη συστηματική ανάλυση της τρέχουσας κατάστασης των μουσικών συλλογών επιβεβαιώνει αυτή την προκατάληψη, δείχνοντας ότι οι δυτικές μουσικές παραδόσεις κυριαρχούν στα υπάρχοντα σύνολα δεδομένων με μόλις 5,7% αντιπροσώπευση μη-δυτικών ειδών.

Ι.2 Η Πρόκληση της Ποικιλομορφίας των Μουσικών Συστημάτων

Η κυρίαρχη εστίαση στις Ευρωπαϊκές και Βορειο-Αμερικανικές μουσικές παραδόσεις δημιουργεί σημαντικές μεθοδολογικές προκλήσεις για την υπολογιστική αναπαράσταση διαφορετικών μουσικών συστημάτων παγκοσμίως. Αυτή η προκατάληψη, όπου οι δυτικές μουσικές έννοιες και αναλυτικά πλαίσια χρησιμεύουν ως προεπιλεγμένος φακός για όλη τη μουσική, έχει εξεταστεί κριτικά από μελετητές στην υπολογιστική εθνομουσικολογία.

Η παραδοσιακή και λαϊκή μουσική από διαφορετικές περιοχές συχνά εμφανίζει διακριτά χαρακτηριστικά που μπορεί να μην ευθυγραμμίζονται με τις παραδοχές που ενσωματώνονται στις τρέχουσες υπολογιστικές προσεγγίσεις. Οι προκλήσεις της ποικιλομορφίας των μουσικών συστημάτων εκδηλώνονται σε πολλαπλές τεχνικές διαστάσεις, που περιγράφονται ακολούθως.

Ι.2.1 Τονικά Συστήματα και Μελωδική Οργάνωση

Η δυτική μουσική συνήθως χρησιμοποιεί ένα σύστημα 12-τόνων ίσου διαστήματος με τυποποιημένες κλίμακες και αρμονία βασισμένη σε τριαδικές δομές. Εντούτοις, πολλές άλλες μουσικές παραδόσεις χρησιμοποιούν διαφορετικές διαιρέσεις διαστημάτων, μικροτονικά μελίσματα και εναλλακτικές

αρχές οργάνωσης. Η τουρκική μουσική "μακάμ" χρησιμοποιεί μια διαίρεση 53-τόνων, ενώ η ινδική κλασική μουσική λειτουργεί εντός ενός συστήματος ράγκας για τη μελωδική οργάνωση.

I.2.2 Ρυθμικές Δομές

Τα Ευρωπαϊκά και Βορειο-Αμερικανικά μουσικά μέτρα χαρακτηρίζονται συνήθως από απλές δομές (4/4, 2/4, 3/4). Πολλές άλλες μουσικές παραδόσεις χρησιμοποιούν πολύπλοκους ρυθμικούς κύκλους, ασύμμετρα μέτρα και πολυρυθμικές οργανώσεις (7/8, 9/8, 10/8). Η μεσογειακή μουσική, συμπεριλαμβανομένης της ελληνικής παράδοσης, χαρακτηρίζεται από ρυθμούς με ασύμμετρες ομαδοποιήσεις κτύπων.

I.2.3 Πρακτικές Εκτέλεσης και Εναρμόνιση

Διάφορες μουσικές παραδόσεις χαρακτηρίζονται από εξεζητημένες πρακτικές εκτέλεσης που παρουσιάζουν προκλήσεις για τις μεθόδους υπολογιστικής ανάλυσης που αναπτύχθηκαν για δυτικά συστήματα σημειογραφίας. Ο αυτοσχεδιασμός παίζει κεντρικό ρόλο σε πολλές κουλτούρες, ενώ πολλές μη-δυτικές παραδόσεις χρησιμοποιούν την ετεροφωνία ως θεμελιώδη αρχή εναρμόνισης.

I.3 Διατύπωση του Προβλήματος

Η κεντρική πρόκληση που αντιμετωπίζεται σε αυτή τη διατριβή είναι η περιορισμένη ικανότητα των τρεχόντων υπολογιστικών μοντέλων να αναπαριστούν και να αναλύουν αποτελεσματικά μουσικές παραδόσεις πέρα από τις δυτικές συμβάσεις. Παρά τις εξελίξεις στην ανάκτηση μουσικών πληροφοριών, οι περισσότερες υπολογιστικές προσεγγίσεις καταδεικνύουν μειωμένη απόδοση όταν εφαρμόζονται σε μουσικά συστήματα με θεμελιωδώς διαφορετικές αρχές οργάνωσης.

Μια κρίσιμη διάσταση αυτής της πρόκλησης περιλαμβάνει την κατανόηση του πώς τα υπολογιστικά συστήματα ευθυγραμμίζονται με την ανθρώπινη αντίληψη αναφορικά με τη διαπολιτισμική μουσική ομοιότητα. Οι τρέχουσες υπολογιστικές προσεγγίσεις συχνά αποτυγχάνουν να συλλάβουν τις λεπτές σχέσεις μεταξύ μουσικών στυλ, οργάνων και αισθητικών αρχών που ορίζουν διαφορετικές μουσικές κουλτούρες. Το πρόβλημα εκδηλώνεται στις ακόλουθες κύριες περιοχές.

I.3.1 Μάθηση Αναπαράστασης

Τα υπάρχοντα σύνολα δεδομένων και μοντέλα ενσωματώνουν συστηματικές προκαταλήψεις που ευνοούν ορισμένα μουσικά χαρακτηριστικά έναντι άλλων. Η σπανιότητα δεδομένων για πολλά περιφερειακά και παραδοσιακά μουσικά συστήματα δημιουργεί ουσιαστικές ανισορροπίες σε σύγκριση με τα εμπορικά κυρίαρχα είδη.

I.3.2 Κατανόηση Πολυπολιτισμικής Μουσικής

Η ταξινόμηση μουσικής σε πολυπολιτισμικά πλαίσια συνήθως συναντά προκλήσεις λόγω των μακριών κατανομών κατηγοριών όπου πολλά, πολιτισμικά σημαντικά, χαρακτηριστικά έχουν πολύ λίγα παραδείγματα. Οι περιορισμοί της αξιολόγησης αυξάνονται όταν ληφθεί υπόψη ότι τα παραδοσιακά μέτρα αποδοτικότητας δεν συλλαμβάνουν επαρκώς την πολιτισμική σημασία.

I.3.3 Ευθυγράμμιση Ανθρώπου-Υπολογιστή

Μια κρίσιμη πρόκληση περιλαμβάνει την κατανόηση του κατά πόσο οι υπολογιστικές προσεγγίσεις δύνανται να συλλάβουν την ανθρώπινη αντίληψη αναφορικά με τη μουσική ομοιότητα ανάμεσα

σε πολιτισμούς. Η ανθρώπινη αντίληψη της μουσικής ομοιότητας ποικίλλει σημαντικά σε διαφορετικά πολιτισμικά πλαίσια, με ακροατές από διαφορετικά μουσικά υπόβαθρα δυνητικά να ακούν και να αξιολογούν τις ίδιες μουσικές σχέσεις με θεμελιωδώς διαφορετικούς τρόπους.

I.3.4 Πολυπολιτισμική Μάθηση

Ο απώτερος στόχος της ανάπτυξης υπολογιστικών μοντέλων ικανών να αναπαριστούν αποτελεσματικά ποικίλες μουσικές παραδόσεις, απαιτεί την αντιμετώπιση των πολύπλοκων προκλήσεων της πολυπολιτισμικής προσαρμογής μοντέλων και της ενσωμάτωσης γνώσης.

I.4 Ερευνητικά Ερωτήματα

Αυτή η διατριβή αντιμετωπίζει έξι διασυνδεδεμένα ερευνητικά ερωτήματα που προοδευτικά βασίζονται το ένα στο άλλο:

EP1: Πώς μπορούν να αναπτυχθούν υψηλής ποιότητας σύνολα δεδομένων για υποεκπροσωπούμενες μουσικές παραδόσεις με σκοπό την υποστήριξη της υπολογιστικής ανάλυσης και διαπολιτισμικής σύγκρισης;

EP2: Σε ποιο βαθμό μπορεί η γνώση να μεταφερθεί αποτελεσματικά μεταξύ διαφορετικών μουσικών συστημάτων, και ποια μοτίβα μεραφοράς παρατηρούνται σε ποικίλες μουσικές παραδόσεις;

EP3: Πώς μπορούν τα υπολογιστικά μοντέλα να μάθουν αποτελεσματικά από περιορισμένα παραδείγματα σε πολυπολιτισμικά μουσικά πλαίσια, ιδιαίτερα για σπάνια αλλά πολιτισμικά σημαντικά μουσικά χαρακτηριστικά;

EP4: Ποιες είναι οι δυνατότητες και οι περιορισμοί των σύγχρονων θεμελιωδών μοντέλων όταν εφαρμόζονται σε ποικίλες μουσικές παραδόσεις;

EP5: Πώς μπορούν τα θεμελιώδη μοντέλα να προσαρμοστούν για να αναπαριστούν καλύτερα ποικίλες μουσικές παραδόσεις διατηρώντας παράλληλα τη γενική μουσική τους γνώση;

EP6: Πώς συγκρίνονται οι υπολογιτικές μέθοδοι μουσικής ομοιότητας με την ανθρώπινη μουσική αντίληψη, και ποιοι παράγοντες καθορίζουν την ομοιότητα ανάμεσα σε ποικίλους πολιτισμούς;

I.5 Συνεισφορές

Αυτή η διατριβή προάγει την εκμάθηση πολυπολιτισμικών μουσικών αναπαραστάσεων μέσω ενός ολοκληρωμένου ερευνητικού προγράμματος που αντιμετωπίζει άμεσα κάθε ένα από τα ερευνητικά ερωτήματα που τέθηκαν παραπάνω.

I.5.1 Αντιμετώπιση της Επάρκειας Δεδομένων για Ποικίλες Μουσικές Παραδόσεις (EP1)

Το **Σύνολο Δεδομένων Lyra** αντιπροσωπεύει την απάντησή μας στη θεμελιώδη πρόκληση της σπανιότητας δεδομένων στην υπολογιστική ανάλυση παραδοσιακής μουσικής. Αυτή η ολοκληρωμένη συλλογή ελληνικής παραδοσιακής μουσικής, που περιλαμβάνει 1570 κομμάτια με περίπου 80 ώρες υψηλής ποιότητας ηχογραφήσεων, καταδεικνύει μια μεθοδολογία για τη δημιουργία πολιτισμικά θεμελιωμένων συνόλων δεδομένων.

I.5.2 Κατανόηση της Διαπολιτισμικής Μεταφοράς Γνώσης (EP2)

Το **Πλαίσιο Διαπολιτισμικής Μεταφοράς Μάθησης** παρέχει την πρώτη συστηματική διερεύνηση των προτύπων μεταφοράς γνώσης μεταξύ διαφορετικών μουσικών συστημάτων. Η έρευνα καταδεικνύει ότι τα υπολογιστικά μοντέλα μπορούν πράγματι να επωφεληθούν από τη μεταφορά

γνώσης μεταξύ διαφορετικών μουσικών συστημάτων, ενώ αποκαλύπτει επίσης την ασύμμετρη και πολύπλοκη φύση αυτών των σχέσεων.

I.5.3 Μάθηση από Περιορισμένα Παραδείγματα σε Μουσικά Πλαίσια (EP3)

Η ανάπτυξη των **LC-Protonets** αντιμετωπίζει άμεσα το EP3 εισάγοντας μια νέα προσέγγιση στη μάθηση πολαπλών κατηγοριών με λίγα παραδείγματα, η αποτελεσματικότητα της οποίας ελέγχθηκε σε σενάρια μουσικής ταξινόμησης. Η μεθοδολογία επεκτείνει τα Prototypical Networks για να χειριστεί τα πολύπλοκα σενάρια πολλών κατηγοριών που είναι κοινά στη μουσική ανάλυση.

I.5.4 Αξιολόγηση Θεμελιωδών Μοντέλων σε Μουσικές Παραδόσεις (EP4)

Το **Πλαίσιο Αξιολόγησης Θεμελιωδών Μοντέλων** παρέχει την πρώτη ολοκληρωμένη αξιολόγηση σύγχρονων μουσικών μοντέλων σε ποικίλες μουσικές παραδόσεις. Η συστηματική σύγκριση πέντε θεμελιωδών μοντέλων ήχου σε έξι μουσικές συλλογές αποκαλύπτει τόσο εντυπωσιακές διαπολιτισμικές ικανότητες όσο και σημαντικούς περιορισμούς.

I.5.5 Προσαρμογή Θεμελιωδών Μοντέλων για Πολυπολιτισμική Ενσωμάτωση (EP5)

Το **CultureMERT** αντιπροσωπεύει την απάντησή μας στο EP5 μέσω μιας νέας στρατηγικής συνεχιζόμενης προ-εκπαίδευσης δύο σταδίων που επιτρέπει τη σταθερή προσαρμογή των θεμελιωδών μοντέλων σε ποικίλες μουσικές παραδόσεις. Η συστηματική αξιολόγηση σε πολλαπλά προβλήματα ταξινόμησης μουσικής επιβεβαιώνει συνεπείς βελτιώσεις.

I.5.6 Γεφύρωση Ανθρώπινης Αντίληψης και Υπολογιστικής Μουσικής Ομοιότητας (EP6)

Η **Μελέτη Διαπολιτισμικής Μουσικής Ομοιότητας** αντιπροσωπεύει την πρώτη πλήρη αξιολόγηση υπολογιστικών μεθόδων μουσικής ομοιότητας έναντι της ανθρώπινης μουσικής αντίληψης. Η μελέτη συλλέγει επισημειώσεις ομοιότητας από 125 συμμετέχοντες με διαφορετικά υπόβαθρα, αξιολογώντας 1130 μοναδικά ζεύγη ήχου από εννέα μουσικά σύνολα δεδομένων.

Τα αποτελέσματα καταδεικνύουν ότι τα θεμελιώδη μοντέλα γενικά επιτυγχάνουν ισχυρότερη ευθυγράμμιση με την ανθρώπινη αντίληψη σε σχέση με τα παραδοσιακά χαρακτηριστικά επεξεργασίας μουσικού σήματος, με το CLAP-Music&Speech να φτάνει το 64,9% triplet agreement. Μεταξύ των χαρακτηριστικών επεξεργασίας σήματος, η μελωδία αναδεικνύεται με συνέπεια ως η πιο καθοριστική για τις ανθρώπινες κρίσεις ομοιότητας.

Αξίζει να σημειωθεί ότι η έρευνα αποκαλύπτει σημαντικές διαφορές στις στρατηγικές επεξεργασίας μεταξύ ανθρώπων και υπολογιστικών μοντέλων: οι άνθρωποι δίνουν προτεραιότητα στο μελωδικό περιεχόμενο, ενώ πολλά θεμελιώδη μοντέλα δίνουν έμφαση στα ηχοχρωματικά χαρακτηριστικά. Οι μέθοδοι συνόλου (ensemble) που συνδυάζουν τα ερμηνεύσιμα χαρακτηριστικά με τις αναπαραστάσεις των μοντέλων επιτυγχάνουν ουσιαστικές βελτιώσεις (67,0% triplet agreement, 25-30% μείωση σφάλματος).

II Θεωρητικό Υπόβαθρο και Μεθοδολογία

Η ανάπτυξη αποτελεσματικών υπολογιστικών προσεγγίσεων για την ανάλυση διαφορετικών μουσικών παραδόσεων προϋποθέτει ένα θεωρητικό υπόβαθρο που συνδυάζει αρχές από την επεξεργασία

μουσικού σήματος, τη βαθιά μάθηση και την υπολογιστική εθνομουσικολογία. Αυτή η ενότητα παρουσιάζει τα κεντρικά θεωρητικά πλαίσια και μεθοδολογικές προσεγγίσεις που αποτελούν τη βάση της παρούσας διατριβής.

II.1 Επεξεργασία και Αναπαράσταση Μουσικού Σήματος

Η παραδοσιακή επεξεργασία μουσικού σήματος στηρίζεται στην εξαγωγή συγκεκριμένων χαρακτηριστικών που σχεδιάζονται για να ανιχνεύουν διαφορετικές πτυχές της μουσικής πληροφορίας. Αυτά τα χαρακτηριστικά, που αναπτύχθηκαν βασισμένα στη μουσική θεωρία, έχουν χρησιμοποιηθεί ευρέως στα παραδοσιακά συστήματα MIR.

II.1.1 Παραδοσιακά Μουσικά Χαρακτηριστικά

Τα **μελωδικά χαρακτηριστικά** βασίζονται στην εξαγωγή της θεμελιώδους συχνότητας F0 από πολυφωνικό ήχο, αντιμετωπίζοντάς την ως τον κυρίαρχο μελωδικό σκελετό. Δεδομένης της F0, οι τονικές κλάσεις (pitches) αναγνωρίζονται και εν συνεχεία τα μουσικά διαστήματα δύναται να ανιχνευθούν και να αναλυθούν για τον χαρακτηρισμό ενός μουσικού κομματιού.

Τα **ρυθμικά χαρακτηριστικά** περιλαμβάνουν αλγορίθμους εκτίμησης tempo, παρακολούθησης κτύπου (beat) και ανίχνευσης έναρξης γεγονότος (onset) για την εξαγωγή πληροφοριών σχετικά με τη χρονική οργάνωση της μουσικής.

Τα **αρμονικά χαρακτηριστικά** αναπαριστούν την αρμονία μέσω προφίλ τόνων, αναγνώρισης συγχορδιών και εκτίμησης κλειδιού. Τα προφίλ που ανιχνεύονται σε συνδυασμό με τα βασικά chromagrams, παρέχουν μια απεικόνιση των αρμονικών τόνων του κομματιού.

Τα **ηχοχρωματικά χαρακτηριστικά**, όπως τα Mel-Frequency Cepstral Coefficients (MFCCs), έχουν αποτελέσει θεμελιώδη εργαλεία για την ανίχνευση χαρακτηριστικών σχετικά με το ηχόχρωμα, το οποίο ποικίλει ανάλογα με τα όργανα που συμμετέχουν, τους καλλιτέχνες αλλά και του τρόπους ηχογράφησης και συγκεριμένα ηχητικά εφέ που μπορεί να λαμβάνουν χώρα.

Τα **δομικά χαρακτηριστικά** αναγνωρίζουν τη μακροσοπική δομή της μουσικής και τις επαναλήψεις, ανιχνεύοντας τη μορφή και τη διάταξη (chorus, verse, bridge).

Παρόλο που τα παραπάνω χαρακτηριστικά έχουν αποδειχθεί αποτελεσματικά για πολλά προβλήματα MIR, συχνά ενσωματώνουν μουσικές παραδοχές από ευρωπαϊκές κλασικές και δημοφιλείς μουσικές παραδόσεις. Για παράδειγμα, τα χαρακτηριστικά chromagram υποδηλώνουν έμμεσα 12-τονικό σύστημα, καθιστώντας τα λιγότερο κατάλληλα για παραδόσεις με διαφορετικά τονικά συστήματα.

II.2 Βαθιά Μάθηση στην Ανάκτηση Μουσικών Πληροφοριών

Οι προσεγγίσεις βαθιάς μάθησης έχουν αλλάξει το MIR πεδίο επιτρέποντας την αυτόματη μάθηση αναπαραστάσεων απευθείας από δεδομένα, μειώνοντας την εξάρτηση από χειροποίητα χαρακτηριστικά. Η εξέλιξη της βαθιάς μάθησης στο MIR έχει περάσει από διάφορες διακριτές φάσεις, καθεμία χαρακτηριζόμενη από διαφορετικά υπολογιστικά παραδείγματα.

II.2.1 Εξέλιξη της Βαθιάς Μάθησης στο MIR

Στην εποχή της παραδοσιακής εξαγωγής χαρακτηριστικών, που εκτείνεται από τη δεκαετία του 1990 μέχρι τις αρχές της δεκαετίας του 2010, κυριαρχούσαν χειροποίητα χαρακτηριστικά όπως τα MFCCs, τα chroma vectors και διάφοροι χρονικοί και φασματικοί περιγραφείς. Παρόλο που αυτές οι προσεγγίσεις παρείχαν ερμηνεύσιμες και μουσικά θεμελιωμένες αναπαραστάσεις, συχνά ενσωμάτωναν

συγκεκριμένες μουσικές παραδοχές που περιόριζαν την αποτελεσματικότητά τους σε ποικίλες μουσικές παραδόσεις.

Οι πρώτες εφαρμογές της βαθιάς μάθησης στο MIR επικεντρώθηκαν στην προσαρμογή αρχιτεκτονικών από άλλους τομείς, ιδιαίτερα την όραση υπολογιστών. Τα Συνελικτικά Νευρωνικά Δίκτυα (CNNs), που αρχικά αναπτύχθηκαν για ανάλυση εικόνας, προσαρμόστηκαν για να επεξεργάζονται φασματογραφήματα ήχου ως 2D εικόνες. Αυτές οι προσεγγίσεις αντιμετωπίζουν τις διαστάσεις χρόνου και συχνότητας όπως τις χωρικές διαστάσεις των εικόνων.

Καθώς η βαθιά μάθηση στο MIR ωρίμαζε, οι ερευνητές ανέπτυξαν αρχιτεκτονικές ειδικά σχεδιασμένες για μουσικά σήματα. Αυτές οι προσεγγίσεις ενσωμάτωσαν γνώση του τομέα σχετικά με τη μουσική δομή και αντίληψη ενώ αξιοποιούσαν την αναπαραστατική δύναμη των βαθιών νευρωνικών δικτύων. Το Musicnn εισήγαγε οριζόντια και κάθετα συνελικτικά φίλτρα για την ξεχωριστή ανίχνευση χρονικών και ηχοχρωματικών χαρακτηριστικών.

Ολιστικές προσεγγίσεις που μαθαίνουν απευθείας από κυματομορφές ήχου αναδύθηκαν, εξαλείφοντας την ανάγκη για προκαθορισμένες αναπαραστάσεις. Μοντέλα όπως το SampleCNN και το TCNN λειτουργούν απευθείας σε δείγματα ήχου, μαθαίνοντας κατάλληλες ιεραρχίες χαρακτηριστικών από τα ίδια τα δεδομένα.

Πιο πρόσφατα, μοντέλα βασισμένα σε μηχανισμούς προσοχής (attention), ιδιαίτερα οι Transformers, εφαρμόζονται στη μουσική ανάλυση. Ο Audio Spectrogram Transformer (AST) προσαρμόζει την αρχιτεκτονική ενός Vision Transformer σε φασματογραφήματα ήχου, αντιμετωπίζοντάς τα ως ακολουθίες από patches.

Η τρέχουσα εποχή θεμελιωδών μοντέλων αντιπροσωπεύει τη πιο πρόσφατη αλλαγή, με μοντέλα όπως το MERT και το CLAP να καταδεικνύουν πρωτοφανείς ικανότητες σε διαφορετικά προβλήματα ανάλυσης μουσικής. Αυτά τα μοντέλα αξιοποιούν μεγάλης κλίμακας αυτο-επιβλεπόμενη προεκπαίδευση για να μάθουν αναπαραστάσεις γενικού σκοπού που μπορούν να προσαρμοστούν σε διάφορες εφαρμογές.

II.3 Σύνολα Δεδομένων και Μοντέλα που Χρησιμοποιούνται σε αυτή την Εργασία

Για την έρευνά μας σχετικά με τη μάθηση μουσικής αναπαράστασης σε ποικίλες παραδόσεις, χρησιμοποιούμε μια συλλογή συνόλων δεδομένων που καλύπτουν διαφορετικές γεωγραφικές περιοχές και μουσικά συστήματα. Αυτά τα σύνολα δεδομένων επιλέγονται προσεκτικά για να αντιπροσωπεύουν τρεις διακριτές γεωγραφικές περιοχές: Ευρώπη και Βόρεια Αμερική, Μεσόγειος, Μέση Ανατολή και η Ινδική υποήπειρος.

II.3.1 Σύνολα Δεδομένων

Τα σύνολα δεδομένων που χρησιμοποιούνται στη διατριβή αναπτύσσονται ακολούθως.

Δυτικά Μουσικά Σύνολα Δεδομένων: Το σύνολο δεδομένων MagnaTagATune χρησιμοποιείται ευρέως για έρευνα μουσικής ταξινόμησης, αποτελούμενο από περισσότερες από 25000 ηχογραφήσεις με συνολική διάρκεια περίπου 210 ωρών. Το FMA-medium περιλαμβάνει επίσης 25000 κομμάτια των 30 δευτερολέπτων το καθένα, συνολικής διάρκειας 208 ωρών.

Μεσογειακά Σύνολα Δεδομένων: Το σύνολο δεδομένων Lyra, που αναπτύχθηκε ως μέρος αυτής της διατριβής, επικεντρώνεται στην ελληνική παραδοσιακή και λαϊκή μουσική, περιλαμβάνοντας

1570 κομμάτια με συνολική διάρκεια 80 ωρών. Το τουρκικό corpus makam περιλαμβάνει χιλιάδες ηχογραφήσεις που καλύπτουν περισσότερα από 5000 έργα.

Ινδικά Κλασικά Μουσικά Σύνολα Δεδομένων: Το Hindustani corpus αντιπροσωπεύει την κλασική παράδοση της Βόρειας Ινδίας με 1204 ηχογραφήσεις, ενώ το Carnatic corpus αντιπροσωπεύει την κλασική παράδοση της Νότιας Ινδίας με 2612 ηχογραφήσεις.

Για τη μελέτη της διαπολιτισμικής μουσικής ομοιότητας, επεκτείνουμε αυτή τη συλλογή για να συμπεριλάβουμε επιπλέον σύνολα δεδομένων που αντιπροσωπεύουν την κινεζική παραδοσιακή μουσική (Jingju), μεσανατολικές παραδόσεις και μεσογειακή μουσική της Ιβηρικής χερσονήσου (Arab-Andalusian, corpusCOFLA).

II.3.2 Μοντέλα

Σε όλη αυτή τη διατριβή, χρησιμοποιούμε διάφορες αρχιτεκτονικές βαθιάς μάθησης για μουσική ανάλυση, εστιάζοντας σε μοντέλα που μπορούν να επεξεργάζονται ήχο με ελάχιστη επαγωγική προκατάληψη.

VGG-ish: Βασίζεται στην αρχιτεκτονική Visual Geometry Group (VGG), που αρχικά αναπτύχθηκε για αναγνώριση εικόνας. Η υλοποίησή μας περιλαμβάνει επτά συνελικτικά στρώματα με φίλτρα συνέλιξης 3×3 και max-pooling 2×2, ακολουθούμενα από δύο πλήρως συνδεδεμένα στρώματα.

Musicnn: Μια αρχιτεκτονική CNN ειδικά σχεδιασμένη για μουσική, ικανή να ανιχνεύσει τόσο χρονικά όσο και ηχοχρωματικά χαρακτηριστικά από φασματογραφήματα ήχου. Η κεντρική καινοτομία του Musicnn είναι το πρώτο συνελικτικό στρώμα που χρησιμοποιεί τόσο κάθετα όσο και οριζόντια φίλτρα.

Audio Spectrogram Transformer (AST): Αντιπροσωπεύει μια αρχιτεκτονική που βασίζεται εξ ολοκλήρου σε μηχανισμούς προσοχής. Προσαρμοσμένο από την αρχιτεκτονική Vision Transformer σε φασματογραφήματα ήχου, το AST μοντέλο καταδεικνύει πώς τα μοντέλα Transformer μπορούν να επεξεργάζονται αποτελεσματικά μουσικά σήματα.

II.3.3 Θεμελιώδη Μοντέλα

Για την εργασία μας σχετικά με την αξιολόγηση και προσαρμογή θεμελιωδών μοντέλων, χρησιμοποιούμε διάφορα σύγχρονα μοντέλα:

MERT-95M και MERT-330M: Το "Music undERstanding model with large-scale self-supervised Training" (MERT) χρησιμοποιεί μια προσέγγιση masked acoustic modeling παρόμοια με το BERT στην επεξεργασία φυσικής γλώσσας. Οι παραλλαγές 95M και 330M προσφέρουν πλήθος παραμέτρων του μοντέλου.

CLAP-Music και CLAP-Music&Speech: Το Contrastive Language-Audio Pre-training προσαρμόζει το πλαίσιο CLIP στον τομέα του ήχου. Αυτά τα μοντέλα μαθαίνουν ενσωματωμένες αναπαραστάσεις μουσικού ήχου και κειμενικών περιγραφών μέσω αντιθετικής μάθησης.

Qwen2-Audio: Αντιπροσωπεύει ένα ενοποιημένο θεμελιώδες μοντέλο κατανόησης ήχου ικανό να επεξεργάζεται τόσο ομιλία όσο και μουσική.

CultureMERT: Αντιπροσωπεύει το πολιτισμικά προσαρμοσμένο θεμελιώδες μοντέλο μας, που αναπτύχθηκε μέσω συνεχούς προ-εκπαίδευσης σε ποικίλες μουσικές παραδόσεις. Χτισμένο στην αρχιτεκτονική MERT-95M, το CultureMERT ενσωματώνει μάθηση από ελληνικές, τουρκικές και ινδικές μουσικές παραδόσεις διατηρώντας την απόδοση σε δυτικά σημεία αναφοράς.

Αυτή η ολοκληρωμένη συλλογή συνόλων δεδομένων και μοντέλων παρέχει τη βάση για τη συστηματική διερεύνηση της εκμάθησης πολυπολιτισμικών μουσικών αναπαραστάσεων που παρουσιάζεται στα επόμενα κεφάλαια της διατριβής.

III Το Σύνολο Δεδομένων Lyra για την Ελληνική Παραδοσιακή και Λαϊκή Μουσική

Η ανάπτυξη υψηλής ποιότητας συνόλων δεδομένων για υποεκπροσωπούμενες μουσικές παραδόσεις αποτελεί βασική προϋπόθεση για την πρόοδο στην υπολογιστική εθνομουσικολογία. Αυτό το κεφάλαιο παρουσιάζει το σύνολο δεδομένων Lyra, μια ολοκληρωμένη συλλογή ελληνικής παραδοσιακής και λαϊκής μουσικής που αναπτύχθηκε ειδικά για να υποστηρίξει την υπολογιστική ανάλυση μιας μουσικής παράδοσης που ενσωματώνει στοιχεία τόσο από δυτικά όσο και από μεσανατολικά μουσικά συστήματα.

III.1 Προκλήσεις και Μέθοδοι Εξαγωγής Δεδομένων

Η πληθώρα ποιοτικών δεδομένων είναι βασική προϋπόθεση για να αναπτύξουν τα σύγχρονα μοντέλα τεχνητής νοημοσύνης το σύνολο της δυναμικότητάς τους. Στην περίπτωση της ελληνικής παραδοσιακής και λαϊκής μουσικής, υπάρχουν λίγες περιπτώσεις όπου τα μεταδεδομένα συνδυάζονται με ηχογραφήσεις με δομημένο τρόπο. Επιπλέον, υπάρχει ζήτημα ποιότητας των ηχογραφήσεων καθώς επηρεάζεται σημαντικά από διάφορους παράγοντες, συμπεριλαμβανομένου του εξοπλισμού που χρησιμοποιείται, της κοινωνικής περίστασης και της χρονικής περιόδου στην οποία πραγματοποιήθηκε.

Προκειμένου να περιορίσουμε την επίδραση του παράγοντα ποιότητας ήχου, αποφασίσαμε να ενσωματώσουμε τα επεισόδια από την ελληνική σειρά ντοκιμαντέρ "Το Αλάτι της Γης" που μεταδόθηκε από την ΕΡΤ, όπου παρουσιάζεται κυρίως παραδοσιακή και λαϊκή μουσική. Τα επεισόδια γυρίστηκαν κατά τη διάρκεια μιας 10ετούς περιόδου υπό αυστηρές προδιαγραφές παραγωγής, με αποτέλεσμα ένα πολύ καθαρό και ομοιογενές ηχητικό περιεχόμενο, ενώ σημαντικός πλούτος πληροφοριών παρέχεται από τον παρουσιαστή και τους καλεσμένους με τη μορφή αφηγήσεων μεταξύ των μουσικών παραστάσεων. Η επαγγελματική ποιότητα παραγωγής που χαρακτηρίζει το πηγαίο υλικό εξασφαλίζει μουσικολογική ορθότητα και συνεπή ηχητικά χαρακτηριστικά σε όλη τη συλλογή.

III.2 Περιγραφή του Συνόλου Δεδομένων

Το σύνολο δεδομένων Lyra οργανώνεται σε έναν ενιαίο πίνακα όπου κάθε γραμμή αντιστοιχεί σε ένα μουσικό κομμάτι ενώ οι στήλες περιλαμβάνουν τις διάφορες πληροφορίες μεταδεδομένων. Το σύνολο δεδομένων αποτελείται από 1570 κομμάτια με συνολική διάρκεια περίπου 80 ωρών, παρέχοντας μια σημαντική πηγή για την υπολογιστική ανάλυση της ελληνικής παραδοσιακής μουσικής.

III.2.1 Κατηγορίες Μεταδεδομένων

Η ταξινομία αποτελείται από: (i) τα μουσικά όργανα που συμμετέχουν στην εκτέλεση κάθε μουσικού κομματιού (η φωνή θεωρείται όργανο), (ii) τα μουσικά είδη και υπο-είδη που αναγνωρί-

ζονται από μουσικολόγους στην ελληνική μουσική, (iii) τους τόπους προέλευσης και (iv) το αν το μουσικό κομμάτι χορεύεται κατά τη διάρκεια της εκτέλεσής του.

Όργανα: Στο σύνολο δεδομένων, η φωνή εμφανίζεται σε σχεδόν 75% των τραγουδιών και όργανα όπως βιολί, κρουστά και λαούτο, που έχουν παρουσία τόσο στα νησιά όσο και στην ηπειρωτική Ελλάδα, ακολουθούν. Υπάρχουν 296 διαφορετικές ομάδες οργάνων στο σύνολο δεδομένων, με αυτή που αποτελείται από φωνή, βιολί, κρουστά, λαούτο και κλαρίνο να είναι η πιο δημοφιλής συμμετέχοντας στην εκτέλεση περίπου 12% των μουσικών κομματιών.

Είδη: Η ταξινόμηση της ελληνικής μουσικής σε "είδη" είναι μια εργασία που απαιτεί να ληφθούν υπόψη ορισμένα κοινωνικο-πολιτισμικά και ανθρωπο-γεωγραφικά κριτήρια. Εν γένει, μπορούμε να διακρίνουμε τη μουσική των αστικών κέντρων σε αντίθεση με τη μουσική των αγροτικών περιοχών της Ελλάδας. Τα 32 μοναδικά είδη χωρίζονται σε 5 διακριτά είδη και 27 υπο-είδη, με το "παραδοσιακό" να είναι το κυρίαρχο αποτελώντας σχεδόν το 78% του συνόλου.

Τόποι Προέλευσης: Από μουσικολογική άποψη, η ελληνική παραδοσιακή μουσική μπορεί να χωριστεί σε δύο μεγάλες γεωγραφικές περιοχές, δηλαδή τη νησιωτική και την ηπειρωτική Ελλάδα. Η κάθε μια δημιουργεί ένα διακριτό μουσικό αίσθημα καθώς χαρακτηρίζονται τόσο από τη ρυθμική προσέγγιση όσο και από τις κλίμακες που χρησιμοποιούνται συνήθως. Από τους 81 τόπους στο σύνολο δεδομένων, 20 είναι ευρύτερες περιοχές και μόνο οι μισές από αυτές περιλαμβάνουν τις υπόλοιπες 61.

Χορός: Η δυαδική κατηγορία "is-danced" ενημερώνει για το αν ένα μουσικό κομμάτι χορεύεται από τους καλεσμένους της εκπομπής. Τα μουσικά κομμάτια που επισημειώθηκαν με "1" είναι περίπου 51% ενώ στα υπόλοιπα δεν λαμβάνει χώρα χορός.

III.3 Βασική Ταξινόμηση

Για την αξιολόγηση της χρησιμότητας του συνόλου δεδομένων, πραγματοποιήθηκαν τρία βασικά προβλήματα ταξινόμησης βασισμένης σε ήχο: αναγνώριση οργάνων, τόπου προέλευσης και ταξινόμηση είδους. Η ηχητική ηχογράφηση κάθε μουσικού κομματιού αναπαριστάται χρησιμοποιώντας ένα Mel-scaled Spectrogram (mel-spectrogram), υπολογιζόμενο ανά τμήμα σταθερής διάρκειας των 10 δευτερολέπτων.

Ως βασική προσέγγιση ταξινόμησης, κάθε mel-spectrogram 10 δευτερολέπτων ταξινομείται στα προαναφερθέντα προβλήματα χρησιμοποιώντας ένα Συνελικτικό Νευρωνικό Δίκτυο (CNN). Τα CNNs έχουν χρησιμοποιηθεί ευρέως σε προβλήματα ταξινόμησης γενικού ήχου, ομιλίας και μουσικής. Η αρχιτεκτονική που υιοθετήθηκε περιλαμβάνει 4 συνελικτικά στρώματα και 3 πλήρως συνδεδεμένα στρώματα.

Τα αποτελέσματα που παρουσιάστηκαν δείχνουν ότι εξειδικευμένα προβλήματα, που χρησιμοποιούν το ηχητικό σήμα, μπορούν δυνητικά να παρέχουν πολύτιμη γνώση για διάφορες πτυχές αυτής της μουσικής. Ο συνδυασμός βίντεο και ηχητικών σημάτων επιτρέπει πιθανούς μελλοντικούς πειραματισμούς σε μεθόδους που επεξεργάζονται πολυτροπικά δεδομένα.

III.4 Συμπεράσματα και Συνεισφορές

Η ελληνική παραδοσιακή και λαϊκή μουσική ενσωματώνει συστατικά ανατολικών και δυτικών ιδιωμάτων, παρέχοντας ενδιαφέρουσες ερευνητικές κατευθύνσεις στον τομέα της υπολογιστικής εθνο-

μουσικολογίας. Το σύνολο δεδομένων Lyra περιλαμβάνει υλικό που επιτρέπει άμεσα τη χρήση εργαλείων MIR για την επίτευξη πολύτιμων μουσικολογικών αποτελεσμάτων και μπορεί δυνητικά να προωθήσει την επέκταση των μεθόδων MIR συνολικά.

Το πλεονέκτημα αυτού του συνόλου δεδομένων είναι ότι όλο το περιεχόμενο συλλέγεται από διαδικτυακούς πόρους μιας ελληνικής σειράς ντοκιμαντέρ που παρήχθη από ακαδημαϊκούς με εξειδίκευση σε αυτή τη μουσική και, συνεπώς, περιλαμβάνει λεπτομερείς επισημειώσεις που εξάγονται από το περιεχόμενο των εκπομπών. Επιπλέον, η παραγωγή ηχογραφήσεων και οπτικοακουστικού υλικού είναι επαγγελματικού επιπέδου, παρέχοντας κοινή βάση όσον αφορά την ποιότητα του ήχου.

Πέρα από την άμεση χρησιμότητά του για την έρευνα ελληνικής μουσικής, το σύνολο δεδομένων Lyra χρησιμεύει ως βάση για τις ευρύτερες διερευνήσεις που παρουσιάζονται στα επόμενα κεφάλαια αυτής της διατριβής. Χρησιμοποιείται μαζί με άλλα σύνολα δεδομένων παγκόσμιας μουσικής για τη διερεύνηση προτύπων μεταφοράς γνώσης μεταξύ διαφορετικών μουσικών παραδόσεων, καθώς και σε ολοκληρωμένες αξιολογήσεις θεμελιωδών μοντέλων σε πολλαπλούς μουσικούς πολιτισμούς.

Η μεθοδολογία που αναπτύχθηκε για τη δημιουργία του συνόλου δεδομένων Lyra αποτελεί αναπαραγώγιμο πλαίσιο για την ανάπτυξη παρόμοιων πόρων για άλλες υποεκπροσωπούμενες μουσικές παραδόσεις. Αυτή η συνεισφορά υπερβαίνει την άμεση χρησιμότητα για την ελληνική μουσική, παρέχοντας ένα πρότυπο για τη δημιουργία πολιτισμικά θεμελιωμένων συνόλων δεδομένων που μπορούν να υποστηρίξουν μια καινοτόμο υπολογιστική ανάλυση διατηρώντας παράλληλα την πολιτισμική αυθεντικότητα.

IV Μάθηση Μεταξύ Πολιτισμών

Για την αντιμετώπιση της πρόκλησης ανάπτυξης υπολογιστικών μοντέλων για την ανάλυση διαφορετικών μουσικών παραδόσεων, δύο συμπληρωματικές προσεγγίσεις εξερευνώνται: η μεταφορά γνώσης (transfer learning) μεταξύ διαφορετικών μουσικών συστημάτων και η μάθηση από λίγα παραδείγματα για σενάρια με περιορισμένα δεδομένα. Και οι δύο προσεγγίσεις συμβάλλουν στον στόχο μας να δημιουργήσουμε πιο ευέλικτες υπολογιστικές μεθόδους προσφέροντας παράλληλα ευρήματα για το πώς η υπολογιτική γνώση μπορεί να μεταφέρεται αποτελεσματικά σε ποικίλες μουσικές παραδόσεις.

IV.1 Διαπολιτισμική Μεταφορά Γνώσης

Η συστηματική διερεύνησή μας για τη μεταφορά γνώσης μεταξύ μουσικών παραδόσεων έχει αποφέρει διάφορα σημαντικά ευρήματα. Πρώτον, καταδείξαμε ότι τα βαθιά μοντέλα ενσωμάτωσης ήχου μπορούν να επωφεληθούν από τη μεταφορά γνώσης από δυτικές σε μη-δυτικές μουσικές παραδόσεις και αντίστροφα. Αυτή η αμφίδρομη μεταφερσιμότητα καταδεικνύει αμοιβαίο όφελος μεταξύ διαφορετικών μουσικών συστημάτων και υποδηλώνει ότι τα μοντέλα που εκπαιδεύονται σε ποικίλες μουσικές παραδόσεις μπορούν να συνεισφέρουν πολύτιμη γνώση στα συστήματα MIR.

Συγκεντρώνοντας την απόδοση σε τρεις αρχιτεκτονικές μοντέλων και διαφορετικές στρατηγικές προσαρμογής τους, εντοπίσαμε μοτίβα μεταφερσιμότητας που μπορεί να αντικατοπτρίζουν υποκείμενες ομοιότητες μεταξύ μουσικών πολιτισμών. Αυτά τα πρότυπα θα μπορούσαν δυνητικά να ερμηνευτούν ως υπολογιστικά μέτρα ομοιότητας μεταξύ παραδόσεων, προσφέροντας γνώσεις για μουσικές σχέσεις που αντικατοπτρίζουν ιστορικές συνδέσεις, γεωγραφική εγγύτητα ή παράλληλη εξέλιξη πολιτισμών.

IV.1.1 Βασικά Ευρήματα της Διαπολιτισμικής Μεταφοράς

Το γενικό μήνυμα που πρέπει να αποκομίσει κανείς είναι ότι ανεξάρτητα από την αρχιτεκτονική του μοντέλου, όλα τα σύνολα δεδομένων έχουν τη δυνατότητα να συνεισφέρουν ως πηγή γνώσης

σε έναν πολιτισμό στόχο παρέχοντας τις βαθιές αναπαραστάσεις ήχου τους. Πιο συγκεκριμένα, τα δυτικά σύνολα δεδομένων όπως (MagnaTagATune και FMA-medium) απέδωσαν καλά ως πηγές σε διάφορες παραδόσεις-στόχους.

Επιπλέον, παρατηρήσαμε ότι καθώς μετακινούμαστε προς μεσογειακές/μεσανατολικές και ινδικές μουσικές παραδόσεις, άλλες μη-δυτικές πηγές γνώσης συνέβαλαν παρόμοια ή μερικές φορές πιο αποτελεσματικά σε αυτούς τους στόχους. Η ολιστική εικόνα της διαπολιτισμικής μουσικής μεταφοράς μάθησης αποκαλύπτει ότι η ομοιομορφία των επιδόσεων διαφορετικών πηγών σε κάθε σύνολο δεδομένων-στόχο ποικίλλει, και συνεπώς κάποιες παραδόσεις είναι πιο κατάλληλες από άλλες για να συνεισφέρουν γνώση σε κάποια συγκεκριμένη μουσική παράδοση.

IV.2 Label-Combination Prototypical Networks (LC-Protonets)

Για να αντιμετωπίσουμε την πρόκληση των περιορισμένων επισημειωμένων δεδομένων στις συλλογές παγκόσμιας μουσικής, προτείνουμε τα Label-Combination Prototypical Networks (LC-Protonets), μια νέα προσέγγιση για τη μάθηση πολλαπλών κατηγοριών από λίγα παραδείγματα. Τα LC-Protonets υπερτερούν έναντι των συγκριτικών προσεγγίσεων, όταν αξιολογήθηκαν σε διάφορα σύνολα δεδομένων και προβλήματα, μέσω της δημιουργίας πρωτοτύπων για συνδυασμούς κατηγοριών αντί για μεμονωμένες κατηγορίες.

IV.2.1 Μεθοδολογία LC-Protonets

Η μάθηση πολλαπλών κατηγοριών από λίγα παραδειγμάτων παρουσιάζει μια σημαντική πρόκληση, ιδιαίτερα επειδή οι κατηγορίες συσχετίζονται και κάθε δείγμα μπορεί να ανήκει σε πολλές από αυτές. Για να το αντιμετωπίσουμε αυτό, προτείνουμε τα LC-Protonets, μια προσέγγιση που επεκτείνει τα Prototypical Networks με έναν απλό αλλά αποτελεσματικό τρόπο.

Θεωρούμε την ταξινόμηση πολλαπλών κατηγοριών ως πρόβλημα όπου κάθε συνδυασμός κατηγοριών είναι μια περιγραφική κατηγορία. Αυτοί οι συνδυασμοί είναι όλα τα υποσύνολα των κατηγοριών που βρίσκονται στα περιορισμένα δεδομένα μάθησης, συμπεριλαμβανομένων των πλήρων συνόλων κατηγοριών. Ένα στοιχείο υποστήριξης με κατηγορίες $\{A, B, C\}$ ορίζει και συμβάλλει σε όλους τους συνδυασμούς κατηγοριών που προκύπτουν από το δυναμοσύνολο των κατηγοριών του, εξαιρουμένου του κενού συνόλου: $\{A\}, \{B\}, \{C\}, \{A, B\}, \{B, C\}, \{A, C\}, \{A, B, C\}$.

Αυτή η προσέγγιση αντιμετωπίζει το πρόβλημα ταξινόμησης πολλαπλών κατηγοριών ως μίγμα σεναρίων μάθησης λίγων παραδειγμάτων και μηδενικών παραδειγμάτων. Για ένα άγνωστο δείγμα, υπολογίζονται οι αποστάσεις από όλα τα δημιουργημένα πρωτότυπα. Σε περιπτώσεις που ένα άγνωστο δείγμα έχει ίσες αποστάσεις από πολλαπλά πρωτότυπα, επιλέγεται αυτό που αντιπροσωπεύει τον μεγαλύτερο αριθμό κατηγοριών, υποστηρίζοντας έτσι ιεραρχικές σχέσεις και ισχυρή συσχέτιση μεταξύ των κατηγοριών.

IV.2.2 Αποτελέσματα και Αξιολόγηση

Τα πειράματά μας έδειξαν ότι τα LC-Protonets επιτυγχάνουν αξιοσημείωτες βελτιώσεις σε σχέση με τις συγκριτικές μεθόδους. Επιπλέον, διαπιστώθηκε ότι η χρήση προ-εκπαιδευμένων μοντέλων ωφελεί σημαντικά τις μεθόδους μάθησης από λίγα παραδείγματα. Το γεγονός αυτό οδήγησε στην ανάπτυξη μιας μεθόδου μάθησης δύο βημάτων που μπορεί να επεκτείνει επιτυχώς το σύνολο κατηγοριών ενός συνόλου δεδομένων αξιοποιώντας προ-εκπαιδευμένα μοντέλα μαζί με τα LC-Protonets.

IV.3 Σύνθεση και Μελλοντικές Κατευθύνσεις

Οι δύο προσεγγίσεις που εξερευνήθηκαν στην ενότητα αυτή, η μεταφορά γνώσης και η μάθηση από λίγα παραδείγματα, συμπληρώνουν η μία την άλλη στην αντιμετώπιση διαφορετικών πτυχών της μάθησης σε ποικίλες μουσικές παραδόσεις. Ενώ η μεταφορά γνώσης αξιοποιεί τη γνώση από παραδόσεις πλούσιες σε δεδομένα για να ενισχύσει την απόδοση σε προβλήματα με επαρκή παραδείγματα, η μάθηση από λίγα παραδείγματα επιτρέπει την αποτελεσματική εκμάθηση ακόμη και με ελάχιστα επισημειωμένα δεδομένα, γεγονός ιδιαίτερα σημαντικό για υποεκπροσωπούμενες παραδόσεις.

Αξίζει να σημειωθεί ότι και οι δύο προσεγγίσεις αναδεικνύουν την αξία των προ-εκπαιδευμένων μοντέλων στη διαπολιτισμική μουσική ανάλυση, αν και χρησιμοποιούν αυτά τα μοντέλα διαφορετικά. Η μεταφορά γνώσης αρχικοποιεί τα μοντέλα με παραμέτρους που μαθαίνονται από έναν πηγαίο σύνολο δεδομένων, ενώ τα LC-Protonets μπορούν να αξιοποιήσουν άμεσα τον χώρο αναπαραστάσεων των προ-εκπαιδευμένων μοντέλων χωρίς επιπλέον εκπαίδευση.

Οι βελτιώσεις απόδοσης που παρατηρήθηκαν όταν συνδυάζονται και οι δύο προσεγγίσεις, χρησιμοποιώντας τη μεταφορά γνώσης για να αποκτήσουν καλύτερες αναπαραστάσεις χαρακτηριστικών και τη μάθηση από λίγα παραδείγματα για να προσαρμόσουν αυτές τις αναπαραστάσεις σε νέες κατηγορίες με περιορισμένα παραδείγματα, δείχνουν προς ενοποιημένα πλαίσια που θα μπορούσαν να αντιμετωπίσουν το πλήρες φάσμα σεναρίων επάρκειας δεδομένων στην πολυπολιτισμική μουσική ανάλυση.

V Θεμελιώδη Μοντέλα για Ποικίλους Πολιτισμούς

Η ενότητα αυτή διερευνά τις δυνατότητες και τους περιορισμούς των σύγχρονων μουσικών θεμελιωδών μοντέλων όταν εφαρμόζονται σε ποικίλες μουσικές παραδόσεις, και αναπτύσσει στρατηγικές προσαρμογής για την ενίσχυση της πολιτισμικής τους αντίληψης. Μέσω μιας ολοκληρωμένης αξιολόγησης και της ανάπτυξης του CultureMERT, αυτή η έρευνα προσφέρει πρακτικές προσεγγίσεις για τη δημιουργία πιο πολιτισμικά ενημερωμένων μουσικών συστημάτων τεχνητής νοημοσύνης.

V.1 Πλαίσιο Αξιολόγησης Πολλαπλών Μεθόδων για Θεμελιώδη Μοντέλα

Για να εκτιμήσουμε με συστηματικό τρόπο τις διαπολιτισμικές ικανότητες των θεμελιωδών μοντέλων, αναπτύξαμε ένα ολοκληρωμένο μεθοδολογικό πλαίσιο που χρησιμοποιεί τρεις συμπληρωματικές προσεγγίσεις αξιολόγησης. Το πλαίσιό μας επιτρέπει τη συστηματική σύγκριση σύγχρονων θεμελιωδών μοντέλων σε δυτικές και μη-δυτικές μουσικές παραδόσεις, παρέχοντας γνώσεις τόσο για τα δυνατά τους σημεία όσο και για τους περιορισμούς τους για τη διαπολιτισμική μουσική ανάλυση.

V.1.1 Μεθοδολογίες Αξιολόγησης

Probing: Η πρώτη μεθοδολογία αξιολογεί πόσο καλά τα θεμελιώδη μοντέλα αναπαριστούν εγγενώς μουσικά χαρακτηριστικά σε ποικίλους πολιτισμούς. Χρησιμοποιούμε probing, όπου το μοντέλο παραμένει παγωμένο ενώ εκπαιδεύουμε μόνο έναν ταξινομητή πάνω από τις εξαγόμενες αναπαραστάσεις. Συγκεκριμένα, υλοποιούμε ένα ρηχό Multi-layer Perceptron (MLP) με ένα κρυφό στρώμα 512 μονάδων ακολουθούμενο από ένα σιγμοειδές στρώμα ταξινόμησης.

Επιβλεπόμενη Προσαρμογή (Supervised Fine-Tuning): Για να αξιολογήσουμε τη δυνατότητα προσαρμογής, υλοποιούμε στοχευμένη επιβλεπόμενη μάθηση ξεπαγώνοντας ένα υποσύνολο παραμέτρων του μοντέλου. Για το MERT-95M, ξεπαγώνουμε τα τελευταία δύο transformer στρώματα, ενώ για

τα υπόλοιπα μοντέλα μόνο το τελευταίο στρώμα. Αυτές οι επιλογές περιορίστηκαν από περιορισμούς RAM που επηρεάζουν τόσο τις εκπαιδεύσιμες παραμέτρους όσο και τη ρύθμιση υπερπαραμέτρων.

Μάθηση Πολλαπλών Κατηγοριών από Λίγα Παραδείγματα (Multi-Label Few-Shot Learning): Η τρίτη μεθοδολογία αξιολογεί την απόδοση σε σενάρια χαμηλών πόρων χρησιμοποιώντας τα LC-Protonets. Εξάγουμε αναπαραστάσεις από τρία διαφορετικά πλαίσια: απευθείας από το προ-εκπαιδευμένο μοντέλο, από το κρυφό στρώμα του εκπαιδευμένου MLP Probe, και από το προσαρμοσμένο μοντέλο.

V.2 Αποτελέσματα και Ανάλυση Αξιολόγησης Θεμελιωδών Μοντέλων

V.2.1 Probing και Επιβλεπόμενη Προσαρμογή

Τα αποτελέσματα αποκαλύπτουν ότι το Qwen2-Audio επιτυγχάνει την υψηλότερη απόδοση με 88,59% ROC-AUC και 56,48% mAP στο probing, βελτιώνοντας περαιτέρω σε 89,37% ROC-AUC και 58,73% mAP μετά την επιβλεπόμενη προσαρμογή. Το μοντέλο αυτό ακολουθείται από το MERT-95M και το CLAP-Music&Speech με συγκρίσιμη απόδοση, ενώ το CLAP-Music δείχνει σημαντικά χαμηλότερη απόδοση.

Παρατηρούμε ένα συνεπές πρότυπο μειωμένης απόδοσης για μουσικές παραδόσεις που είναι πολιτισμικά απομακρυσμένες από τα δεδομένα που χρησιμοποιούνται για την προ-εκπαίδευση των αντίστοιχων θεμελιωδών μοντέλων. Τα δυτικά μουσικά σύνολα δεδομένων (MagnaTagATune και FMA-medium) επιτυγχάνουν συνεπώς την υψηλότερη απόδοση σε όλα τα θεμελιώδη μοντέλα, με τιμές ROC-AUC που φτάνουν το 96,60% για το Qwen2-Audio στο FMA-medium. Τα ελληνικά (Lyra) και τουρκικά (makam) μουσικά σύνολα δεδομένων δείχνουν μέτρια απόδοση, ενώ τα σύνολα δεδομένων ινδικής μουσικής (Hindustani και Carnatic) συνήθως εμφανίζουν τη χαμηλότερη απόδοση.

Αξίζει να σημειωθεί ότι οι προσεγγίσεις μας επιτυγχάνουν την καλύτερη επίδοση που έχει αναφερθεί σε πέντε από τα έξι σύνολα δεδομένων, με το MagnaTagATune να είναι η μόνη εξαίρεση. Ωστόσο, η συνεπής μείωση της απόδοσής τους για ποικίλους πολιτισμούς, υποδηλώνει ότι οι αναπαραστάσεις τους είναι ακόμη προκατειλημμένες προς τις δυτικές μουσικές παραδόσεις.

V.2.2 Μάθηση Πολλαπλών Κατηγοριών με Λίγα Παραδείγματα

Το Qwen2-Audio καταδεικνύει, και εδώ, την καλύτερη συνολική απόδοση με 32,00% macro-F1 και 56,85% micro-F1 μετά την επιβλεπόμενη προσαρμογή. Εντούτοις, ακόμη και η απόδοση του καλύτερου θεμελιώδους μοντέλου (Qwen2-Audio με περισσότερες από 600M παραμέτρους για την επεξεργασία του ήχου) είναι συγκρίσιμη με έναν μοντέλο VGG-ish που αποτελείται από μόλις 3,6M παραμέτρους. Αυτό υποδηλώνει ότι η μάθηση από λίγα παραδείγματα παραμένει πρόκληση για τα θεμελιώδη μοντέλα.

Όταν εξετάζουμε τα αποτελέσματα ανά παράδοση, παρατηρούμε ότι μόνο στα δυτικά σύνολα δεδομένων (MagnaTagATune και FMA-medium) το καλύτερο θεμελιώδες μοντέλο (Qwen2-Audio) επιτυγχάνει σημαντικά καλύτερη απόδοση από το VGG-ish baseline. Αυτό το εύρημα παρέχει μια επιπλέον ένδειξη της δυτικοκεντρικής προκατάληψης που έχει ενσωματωθεί στα μοντέλα αυτά λόγω των δεδομένων προ-εκπαίδευσής τους.

V.3 CultureMERT: Ένα Πολυπολιτισμικά Προσαρμοσμένο Θεμελιώδες Μοντέλο

Βασιζόμενοι στις γνώσεις από την ολοκληρωμένη αξιολόγηση των θεμελιωδών μοντέλων, παρουσιάζουμε μια νέα προσέγγιση για την ενίσχυση της πολιτισμικής τους αντίληψης. Τα αποτελέσματα αξιολόγησης έχουν καταδείξει σαφώς τόσο τη δυνατότητα όσο και τους περιορισμούς των υπαρχόντων θεμελιωδών μοντέλων όταν εφαρμόζονται σε ποικίλες μουσικές παραδόσεις. Συγκεκριμένα, παρατηρήσαμε μια συνεπή μείωση της απόδοσης για πολιτισμικά απομακρυσμένες παραδόσεις και σε σενάρια χαμηλών πόρων, υπογραμμίζοντας την ανάγκη για ειδικές στρατηγικές προσαρμογής.

V.3.1 Στρατηγική Συνεχιζόμενης Προ-εκπαίδευσης Δύο Σταδίων

Για να προσαρμόσουμε το θεμελιώδες μοντέλο MERT σε ποικίλες μουσικές παραδόσεις, χρησιμοποιούμε συνεχιζόμενη προ-εκπαίδευση, η οποία επεκτείνει την εκπαίδευση ενός προ-εκπαιδευμένου μοντέλου σε νέα δεδομένα, με στόχο να το προσαρμόσει σε έναν νέο σύνολο δεδομένων ή σε μια νέα εργασία διατηρώντας παράλληλα την προηγούμενη γνώση. Δεδομένης αυτής της μετατόπισης του μοντέλου, η ευθεία συνέχιση της εκπαίδευσής του στα νέα δεδομένα μπορεί να οδηγήσει σε καταστροφική λήθη και κακή προσαρμογή.

Για να αντιμετωπίσουμε το φαινόμενο αυτό, προτείνουμε μια στρατηγική δύο σταδίων που σταθεροποιεί την εκπαίδευσή του.

Στάδιο 1 - Φάση Σταθεροποίησης: Εκπαίδευση σε ένα μικρότερο υποσύνολο δεδομένων, ενημερώνοντας μόνο συγκεκριμένα τμήματα του μοντέλου ενώ διατηρούμε τον κωδικοποιητή Transformer παγωμένο. Για να μειώσουμε το απόκλιση στην κατανομή των δεδομένων και να μετριάσουμε τη λήθη, ενσωματώνουμε 20% δεδομένα Music4All (κυρίως δυτικά) στο μείγμα προ-εκπαίδευσης.

Στάδιο 2 - Πλήρης Προσαρμογή: Ξεπαγώνουμε τον κωδικοποιητή Transformer και συνεχίζουμε την εκπαίδευση στο πλήρες σύνολο δεδομένων.

Αυτή η προσέγγιση εξισορροπεί την πλαστικότητα (προσαρμογή σε μη-δυτικές παραδόσεις) και τη σταθερότητα (διατήρηση γνώσης σε δυτικά σύνολα δεδομένων), αντιμετωπίζοντας αποτελεσματικά το δίλημμα σταθερότητας-πλαστικότητας.

V.3.2 Αριθμητική Εργασιών για Διαπολιτισμική Προσαρμογή

Ως εναλλακτική στη συνεχιζόμενη προ-εκπαίδευση, εξερευνούμε την αριθμητική εργασιών (task arithmetic), η οποία συνδυάζει πολιτισμικά εξειδικευμένα μοντέλα στον χώρο βαρών. Λαμβάνουμε διανύσματα εργασιών υπολογίζοντας τη διαφορά στοιχείο προς στοιχείο μεταξύ μοντέλων προσαρμοσμένων σε μια κουλτούρα και του βασικού μοντέλου MERT. Για πολυπολιτισμική προσαρμογή, κατασκευάζουμε ένα ενοποιημένο μοντέλο συγχωνεύοντας όλα τα διανύσματα εργασιών.

V.4 Αξιολόγηση Απόδοσης και Διαπολιτισμική Ανάλυση του CultureMERT

Το CultureMERT, προσαρμοσμένο μέσω πολυπολιτισμικής συνεχιζόμενης προ-εκπαίδευσης, υπερτερεί του αρχικού μοντέλου MERT σε όλες τα μη-δυτικά προβλήματα και μετρικές αξιολόγησης, επιτυγχάνοντας μια μέση βελτίωση 4,9%. Υπερτερεί επίσης των μοντέλων προσαρμοσμένων σε μια κουλτούρα κατά μέσο όρο, υποδηλώνοντας ότι η ενσωμάτωση πολυπολιτισμικών δεδομένων ωφελεί

όλες τις μη-δυτικές παραδόσεις βελτιώνοντας την ποιότητα των αναπαραστάσεων για κάθε επιμέρους πολιτισμό.

Αξίζει να σημειωθεί ότι το CultureMERT το επιτυγχάνει αυτό με ελάχιστη λήθη σε δυτικά σημεία αναφοράς (0,05% μέση πτώση στο ROC-AUC και AP), καταδεικνύοντας την αποτελεσματικότητα της προσέγγισής μας.

Επιπλέον, η αριθμητική εργασιών αποδίδει συγκρίσιμα με το CultureMERT σε μη-δυτικά προβλήματα και ακόμη το ξεπερνά σε δυτικά σημεία αναφοράς και στο σύνολο δεδομένων Lyra, καταδεικνύοντας ότι η συγχώνευση στον χώρο βαρών πολιτισμικά εξειδικευμένων μοντέλων μπορεί να χρησιμεύσει ως μια αποτελεσματική εναλλακτική για εκμάθηση πολυπολιτισμικών αναπαραστάσεων.

V.4.1 Διαπολιτισμική Μεταφορά

Η συνεχιζόμενη προ-εκπαίδευση σε μια μουσική παράδοση μπορεί να ωφελήσει άλλες σε διαφορετικούς βαθμούς, αποκαλύπτοντας ασυμμετρίες στην αποτελεσματικότητα της διαπολιτισμικής μεταφοράς. Για παράδειγμα, παρατηρούμε ισχυρή μεταφορά μεταξύ τουρκικού-makam και Carnatic μουσικής, γεγονός που ευθυγραμμίζεται με τα κοινά θεωρητικά θεμέλια των μουσικών τους συστημάτων. Επιπλέον, η ισχυρή απόδοση του μοντέλου προσαρμοσμένου στο Carnatic στην κουλτούρα Hindustani εδράζεται στη γεωγραφική και μουσική εγγύτητα των παραδόσεων αυτών, ιδιαίτερα στην κοινή χρήση των raga (μελωδικός τρόπος) και tala (ρυθμικό πλαίσιο).

V.5 Συμπεράσματα και Μελλοντικές Κατευθύνσεις

Η αξιολόγησή μας των θεμελιωδών μοντέλων για διαφορετικούς μουσικούς πολιτισμούς αποκαλύπτει τόσο τις δυνατότητες όσο και τους περιορισμούς τους, ενώ καταδεικνύει αποτελεσματικές στρατηγικές για την ενίσχυση της ευελιξίας στη μάθηση μουσικής αναπαράστασης.

Στο ολοκληρωμένο πλαίσιο αξιολόγησης ποα αναπτύξαμε, διαπιστώσαμε ότι αυτά τα μοντέλα επέτυχαν καλύτερη απόδοση από προηγούμενες προσεγγίσεις για την ανάλυση παγκόσμιας μουσικής, καταδεικνύοντας εντυπωσιακές ικανότητες διαπολιτισμικής μεταφοράς. Ωστόσο, εντοπίσαμε σαφείς ενδείξεις δυτικοκεντρικής προκατάληψης, ιδιαίτερα σε σενάρια χαμηλών πόρων.

Για να αντιμετωπίσουμε τους περιορισμούς που εντοπίστηκαν στην αξιολόγησή μας, αναπτύξαμε το CultureMERT, ένα πολυπολιτισμικά προσαρμοσμένο θεμελιώδες μοντέλο που δημιουργήθηκε μέσω συνεχιζόμενης προ-εκπαίδευσης σε διαφορετικές μη-δυτικές μουσικές παραδόσεις. Η διαπολιτισμική αξιολόγηση κατέδειξε ότι το CultureMERT υπερτερούσε του αρχικού μοντέλου σε ποικίλα μη-δυτικά προβλήματα μουσικής ταξινόμησης διατηρώντας παράλληλα την απόδοση σε δυτικά σημεία αναφοράς. Αυτό το εύρημα επιβεβαιώνει τη δυνατότητα της συνεχούς προ-εκπαίδευσης για την ενίσχυση της πολιτισμικής συμπερίληψης των θεμελιωδών μοντέλων χωρίς να θυσιάζει τις γενικές τους ικανότητες.

Ωστόσο, ένα βασικό ερώτημα παραμένει αναπάντητο: πόσο καλά αυτές οι υπολογιστικές μέθοδοι ευθυγραμμίζονται με την ανθρώπινη αντίληψη αναφορικά με τις σχέσεις διαφορετικών πολιτισμών;

VI Διαπολιτισμική Μουσική Ομοιότητα: Γεφυρώνοντας την Ανθρώπινη Αντίληψη και τις Υπολογιστικές Μεθόδους

Αυτή η ενότητα παρουσιάζει την πρώτη ολοκληρωμένη αξιολόγηση υπολογιστικών μεθόδων μουσικής ομοιότητας έναντι της ανθρώπινης διαπολιτισμικής μουσικής αντίληψης, παρέχοντας κρίσιμη εμπειρική

επικύρωση των προσεγγίσεων εκμάθησης πολυπολιτισμικών μουσικών αναπαραστάσεων που αναπτύχθηκαν σε όλη αυτή τη διατριβή. Μέσω συστηματικής σύγκρισης τόσο ερμηνεύσιμων χαρακτηριστικών επεξεργασίας σήματος όσο και σύγχρονων θεμελιωδών μοντέλων έναντι ανθρώπινων κρίσεων ομοιότητας σε εννέα ποικίλες μουσικές παραδόσεις, αυτή η μελέτη διατυπώνει ένα νέο σημείο αναφοράς για την αξιολόγηση της αποτελεσματικότητας των διαπολιτισμικών μουσικών συστημάτων τεχνητής νοημοσύνης.

VI.1 Μελέτη Ανθρώπινης Ομοιότητας

Για να κατανοήσουμε πώς οι άνθρωποι αντιλαμβάνονται τη μουσική ομοιότητα σε διαφορετικές πολιτισμικές παραδόσεις, διενεργήσαμε μια διαδικτυακή έρευνα συλλέγοντας κρίσεις ομοιότητας από συμμετέχοντες με διαφορετικά μουσικά υπόβαθρα και πολιτισμικές καταγωγές.

VI.1.1 Σχεδιασμός Έρευνας και Μεθοδολογία

Η μελέτη μας περιλαμβάνει εννέα μουσικά σύνολα δεδομένων που αντιπροσωπεύουν διαφορετικές πολιτισμικές παραδόσεις. Από κάθε σύνολο δεδομένων, επιλέξαμε 52 αντιπροσωπευτικά κλιπ ήχου διάρκειας 20 δευτερολέπτων, με αποτέλεσμα συνολικά 468 ηχητικά αποσπάσματα που καλύπτουν διαφορετικά όργανα, φωνητικά στυλ και μουσικές δομές.

Ακολουθώντας καθιερωμένες μεθοδολογίες στην έρευνα μουσικής αντίληψης, χρησιμοποιήσαμε μια προσέγγιση ανά ζεύγη σύγκρισης όπου οι συμμετέχοντες αξιολόγησαν τυχαία επιλεγμένα ζεύγη ηχητικών αποσπασμάτων 20 δευτερολέπτων. Κάθε συμμετέχων αξιολόγησε 10 μοναδικά ζεύγη, παρέχοντας βαθμολογίες σε τρεις διακριτές διαστάσεις ομοιότητας χρησιμοποιώντας μια κλίμακα Likert 9 σημείων:

1. **Συνολική Μουσική Ομοιότητα**: "Πόσο όμοια είναι τα δύο ηχητικά αποσπάσματα συνολικά;"

2. **Πολιτισμική Ομοιότητα**: "Πόσο όμοια είναι τα δύο ηχητικά αποσπάσματα στα πολιτισμικά τους χαρακτηριστικά;"

3. **Ομοιότητα Επιπέδου Σύστασης**: "Πόσο πιθανό είναι να βάλετε τα δύο ηχητικά αποσπάσματα στην ίδια λίστα αναπαραγωγής;"

Το παραπάνω πλαίσιο επιτρέπει την εξέταση του πώς διαφορετικές πτυχές της ομοιότητας ευθυγραμμίζονται ή αποκλίνουν, ιδιαίτερα σημαντικό για τη διαπολιτισμική ανάλυση όπου η μουσική και πολιτισμική ομοιότητα μπορεί να μη συμπίπτουν.

VI.1.2 Δημογραφικά Στοιχεία Συμμετεχόντων και Στατιστικά Δεδομένα

Η μελέτη μας συνέλεξε απαντήσεις από 125 συμμετέχοντες, με αποτέλεσμα 1130 διαφορετικά επισημειωμένα ζεύγη που αποτελούνται από 463 διαφορετικά κλιπ ήχου. Οι σχολιαστές προήλθαν από 21 χώρες και αναφέρουν 13 διακριτά επίπεδα μουσικής εκπαίδευσης, ενώ 58 διαφορετικοί μουσικοί πολιτισμοί αναγνωρίστηκαν ως οικείοι από τουλάχιστον έναν συμμετέχοντα. Παρά την ελληνική πλειονότητα (62,4%), η εκπροσώπηση από διαφορετικές περιοχές συμπεριλαμβανομένης της Ασίας, της Ευρώπης και της Βόρειας Αμερικής εξασφαλίζει τη διαπολιτισμική εγκυρότητα.

Η κατανομή ηλικιών των συμμετεχόντων εκτείνεται από 18-64 ετών, με την πλειονότητα να εμπίπτει στην ηλικιακή ομάδα 25-44 ετών. Η ομάδα 25-34 αντιπροσωπεύει το μεγαλύτερο τμήμα (50

συμμετέχοντες), ακολουθούμενη από την 35-44 (37) και την 18-24 (29). Όσον αφορά το φύλο, υπάρχουν 75 άνδρες συμμετέχοντες (60,0%), 42 γυναίκες συμμετέχουσες (33,6%), και 8 συμμετέχοντες που απάντησαν Άλλο ή προτίμησαν να μη δηλώσουν το φύλο τους (6,4%).

VI.2 Χαρακτηριστικά Επεξεργασίας Σήματος και Θεμελιώδη Μοντέλα έναντι Ανθρώπινης Αντίληψης

Αξιολογήσαμε συστηματικά τόσο τα χαρακτηριστικά επεξεργασίας σήματος όσο και τα θεμελιώδη μοντέλα έναντι των ανθρώπινων κρίσεων ομοιότητας χρησιμοποιώντας πέντε συμπληρωματικά μέτρα για κάθε μία από τις τρεις διαστάσεις ομοιότητας. Αυτή η ολοκληρωμένη αξιολόγηση παρέχει γνώσεις για το ποιες υπολογιστικές προσεγγίσεις ευθυγραμμίζονται καλύτερα με την ανθρώπινη διαπολιτισμική μουσική αντίληψη.

VI.2.1 Ολοκληρωμένη Αξιολόγηση Απόδοσης

Τα αποτελέσματα αποκαλύπτουν διακριτά πρότυπα απόδοσης μεταξύ χαρακτηριστικών επεξεργασίας σήματος και θεμελιωδών μοντέλων.

Απόδοση Χαρακτηριστικών Επεξεργασίας Σήματος: Μεταξύ των χαρακτηριστικών επεξεργασίας σήματος, η μελωδία καταδεικνύει με συνέπεια την ανώτερη απόδοση σε όλα τα μέτρα και τις διαστάσεις ομοιότητας. Η μελωδία επιτυγχάνει τις καλύτερες τιμές MAE (29,5-30,9%) και δείχνει τις ισχυρότερες συσχετίσεις με τις ανθρώπινες κρίσεις (Spearman ρ = 0,14-0,15 και Kendall τ = 0,12-0,13). Αυτό επιβεβαιώνει τον κεντρικό ρόλο της μελωδίας στην ανθρώπινη αντίληψη μουσικής ομοιότητας σε όλους τους πολιτισμούς. Αντίθετα, τα χαρακτηριστικά ρυθμού, αρμονίας και ηχοχρώματος δείχνουν περιορισμένη ευθυγράμμιση με την ανθρώπινη αντίληψη, με συσχετίσεις κοντά στο μηδέν ή ελαφρώς αρνητικές.

Απόδοση Θεμελιωδών Μοντέλων: Τα θεμελιώδη μοντέλα γενικά υπερτερούν των χαρακτηριστικών επεξεργασίας σήματος στις περισσότερες μετρικές, με το CLAP-Music&Speech να αναδεικνύεται ως το κορυφαίος μοντέλο, επιτυγχάνοντας τις υψηλότερες τιμές triplet agreement (62,6-64,9%) και NDCG (88,0-89,8%). Ωστόσο, τα χαρακτηριστικά μελωδίας παραμένουν ανταγωνιστικά, επιτυγχάνοντας τις καλύτερες τιμές MAE και ισχυρή απόδοση συσχέτισης.

Η ανώτερη απόδοση του CLAP-Music&Speech έναντι του CLAP-Music υπογραμμίζει τη συνέργεια μεταξύ μελωδικών τρόπων και ομιλίας, καθώς και οι δύο συμπεριλαμβάνουν συμπληρωματικές πτυχές της μουσικής έκφρασης που είναι ιδιαίτερα πολύτιμες για τη διαπολιτισμική μουσική κατανόηση.

Το MERT-95 καταδεικνύει συνεπή απόδοση σε όλες τις διαστάσεις ομοιότητας, ενώ το μεγαλύτερο μοντέλο MERT-330 δείχνει μικτά αποτελέσματα, μερικές φορές υποαποδίδοντας του μικρότερου ομολόγου του, γεγονός που είχε παρατηρηθεί και στα αποτελέσματα της προηγούμενης ενότητας. Οι πολιτισμικά προσαρμοσμένες παραλλαγές CultureMERT υποαποδίδουν του βασικού μοντέλου τους, MERT-95, κάτι που είναι λογικό δεδομένου του κυρίως δυτικού μουσικού υποβάθρου του δείγματος συμμετεχόντων μας.

VI.2.2 Ανάλυση Διαπολιτισμικής Διακριτότητας

Για να εκτιμήσουμε πόσο καλά οι υπολογιστικές μέθοδοι διακρίνουν μεταξύ διαφορετικών μουσικών παραδόσεων, αναλύουμε τη διαπολιτισμική τους διακριτότητα χρησιμοποιώντας λόγους διαχωρισμού βασισμένους σε απόσταση. Συγκρίνουμε αυτούς τους λόγους με τις αντίστοιχες τιμές ανθρώπινης

αντίληψης για να παρέχουμε ευρήματα σχετικά με το πόσο αποτελεσματικά οι υπολογιστικές προσεγγίσεις μπορούν να αναγνωρίσουν τα πολιτισμικά όρια.

Οι ανθρώπινες κρίσεις ομοιότητας καταδεικνύουν ανώτερη πολιτισμική διάκριση, με τη διάσταση Πολιτισμικής ομοιότητας να επιτυγχάνει τον υψηλότερο λόγο διαχωρισμού (2,361), ακολουθούμενη από το Επίπεδο σύστασης (2,106) και τη Συνολική μουσική ομοιότητα (1,803). Αυτό επιβεβαιώνει ότι οι άνθρωποι αναγνωρίζουν και διακρίνουν με συνέπεια τις ποικίλες μουσικές παραδόσεις.

Μεταξύ των χαρακτηριστικών επεξεργασίας σήματος, η μελωδία και πάλι αναδεικνύεται ως η πιο διακριτική (1,276), σε συμφωνία με την ανώτερη απόδοσή της. Τα θεμελιώδη μοντέλα υπερτερούν σημαντικά των χαρακτηριστικών επεξεργασίας σήματος στην πολιτισμική διάκριση, αλλά αναδεικνύουν έναν σημαντικό συμβιβασμό μεταξύ καθολικής μουσικής κατανόησης και πολιτισμικής διακριτότητας.

Το Qwen2-Audio επιτυγχάνει τους υψηλότερους λόγους διαχωρισμού μεταξύ όλων των υπολογιστικών μεθόδων (1,579), καταδεικνύοντας ανώτερη ικανότητα διάκρισης μεταξύ μουσικών παραδόσεων. Αντίθετα, το CLAP-Music&Speech, ενώ διαπρέπει στην ευθυγράμμιση με την ανθρώπινη αντίληψη ομοιότητας, δείχνει πιο μέτρια απόδοση διάκρισης (1,366), υποδηλώνοντας ότι τα μοντέλα που είναι βελτιστοποιημένα για καθολική διαπολιτισμική μουσική κατανόηση μπορεί να θυσιάζουν κάποια διακριτική δύναμη σχετικά με την ανίχνευση των πολιτισμικών ορίων.

VI.3 Μέθοδοι Συνόλου για Πρόβλεψη Ανθρώπινης Ομοιότητας

Για να αξιοποιήσουμε τα συμπληρωματικά δυνατά σημεία των χαρακτηριστικών επεξεργασίας σήματος και των αναπαραστάσεων θεμελιωδών μοντέλων, αναπτύξαμε μεθόδους συνόλου (ensemble) που συνδυάζουν όλες τις υπολογιστικές προσεγγίσεις για να προβλέψουν τις ανθρώπινες κρίσεις ομοιότητας. Αυτή η ανάλυση εξερευνά εάν ο συνδυασμός ερμηνεύσιμων μουσικών χαρακτηριστικών με αναπαραστάσεις θεμελιωδών μοντέλων δύναται να επιτύχει ανώτερη ευθυγράμμιση με την ανθρώπινη διαπολιτισμική μουσική αντίληψη.

VI.3.1 Αποτελέσματα Απόδοσης Συνόλου

Οι μέθοδοι συνόλου επιτυγχάνουν αξιοσημείωτες βελτιώσεις σε σύγκριση με τις μεμονωμένες προσεγγίσεις. Το σύνολο γραμμικής παλινδρόμησης επιτυγχάνει τιμές triplet agreement 65,1-67,0% σε σύγκριση με την καλύτερη μεμονωμένη μέθοδο (CLAP-Music&Speech) στο 62,6-64,9%. Παρόμοια, οι τιμές NDCG φτάνουν το 90,9-92,5% έναντι του προηγούμενου καλύτερου 88,0-89,8%.

Επιπλέον, οι μέθοδοι συνόλου επιτυγχάνουν σημαντικές μειώσεις στο σφάλμα πρόβλεψης, με τιμές MAE 19,7-23,2% που αντιπροσωπεύουν βελτιώσεις περίπου 6-7 ποσοστιαίων μονάδων έναντι των καλύτερων μεμονωμένων μεθόδων (χαρακτηριστικά μελωδίας στο 29,5-30,9% MAE). Αυτό αντιπροσωπεύει μια σχετική μείωση σφάλματος περίπου 25-30%, καταδεικνύοντας ότι ο συνδυασμός πολλαπλών υπολογιστικών προσεγγίσεων παρέχει συμπληρωματικές πληροφορίες για την πρόβλεψη ανθρώπινων κρίσεων ομοιότητας.

VI.3.2 Ανάλυση Συνεισφοράς Υπολογιστικών Μοντέλων

Η ανάλυση σπουδαιότητας αποκαλύπτει τη σχετική σημασία διαφορετικών υπολογιστικών μεθόδων εντός των προσεγγίσεων συνόλου. Το CLAP-Music&Speech αναδεικνύεται ως ο κύριος συνεισφέρων και στις δύο μεθόδους συνόλου που αναπτύχθηκαν, με τον υψηλότερο συντελεστή γραμμικής παλινδρόμησης (23,5%) και την υψηλότερη τιμή κέρδους LightGBM (89,3). Αυτό το εύρημα επικυρώνει την προηγούμενη παρατήρησή μας ότι το μοντέλο αυτό επιτυγχάνει την καλύτερη ευθυγράμμιση με την ανθρώπινη διαπολιτισμική μουσική αντίληψη. Μεταξύ των χαρακτηριστικών επεξεργασίας σήματος, η μελωδία διατηρεί τη θέση της ως το πιο σημαντικό χαρακτηριστικό (γραμμικός συντελεστής:

19,2%, κέρδος LightGBM: 58,6), επιβεβαιώνοντας τον θεμελιώδη ρόλο της στην ανθρώπινη αντίληψη μουσικής ομοιότητας σε όλους τους πολιτισμούς.

VI.4 Συμπεράσματα

Αυτό το κεφάλαιο παρουσίασε την πρώτη ολοκληρωμένη αξιολόγηση υπολογιστικών μεθόδων μουσικής ομοιότητας έναντι της ανθρώπινης διαπολιτισμικής μουσικής αντίληψης, παρέχοντας κρίσιμη εμπειρική επικύρωση των προσεγγίσεων εκμάθησης πολυπολιτισμικών μουσικών αναπαραστάσεων που αναπτύχθηκαν σε όλη τη διατριβή.

Τα βασικά ευρήματα από αυτή τη διερεύνηση επικυρώνουν και επεκτείνουν διάφορες γνώσεις από προηγούμενα κεφάλαια. Η ανώτερη απόδοση των θεμελιωδών μοντέλων έναντι των παραδοσιακών χαρακτηριστικών επεξεργασίας σήματος επιβεβαιώνει τη δυνατότητα των προσεγγίσεων που αξιολογήθηκαν στο Κεφάλαιο 5, ενώ η ανακάλυψη ότι η μελωδία αναδεικνύεται συνεπώς ως το πιο προβλεπτικό χαρακτηριστικό επεξεργασίας σήματος ευθυγραμμίζεται με τη μουσικολογική κατανόηση.

Η αποκάλυψη ότι οι μέθοδοι συνόλου που συνδυάζουν ερμηνεύσιμα χαρακτηριστικά με αναπαραστάσεις θεμελιωδών μοντέλων επιτυγχάνουν την καλύτερη ευθυγράμμιση με την ανθρώπινη αντίληψη, καταδεικνύει τη συμπληρωματική αξία των διαφορετικών υπολογιστικών προσεγγίσεων που αναπτύχθηκαν στη διατριβή.

Η μελέτη αποκαλύπτει, επίσης, σημαντικές διαφορές μεταξύ ανθρώπινων και υπολογιστικών στρατηγικών επεξεργασίας που έχουν σημαντικές επιπτώσεις για τον τομέα. Το εύρημα ότι οι άνθρωποι δίνουν προτεραιότητα στο μελωδικό περιεχόμενο ενώ πολλά θεμελιώδη μοντέλα δίνουν έμφαση στα ηχοχρωματικά χαρακτηριστικά επισημαίνει μια κρίσιμη αναντιστοιχία που εκτείνεται πέρα από την τεχνική βελτιστοποίηση σε ερωτήματα σχετικά με τους στόχους και τα κριτήρια αξιολόγησης για τα συστήματα μουσικής τεχνητής νοημοσύνης.

Οι μεθοδολογικές συνεισφορές αυτής της μελέτης, συμπεριλαμβανομένου του πολυδιάστατου πλαισίου αξιολόγησης ομοιότητας, των ολοκληρωμένων μετρικών αξιολόγησης και των προσεγγίσεων συνόλου, παρέχουν πρότυπα για μελλοντική έρευνα που μπορεί να συνεχίσει να προάγει την ευθυγράμμιση μεταξύ υπολογιστικής μουσικής ανάλυσης και ανθρώπινης αντιληπτικής κατανόησης σε διαφορετικά πολιτισμικά πλαίσια.

VII Συμπεράσματα και Μελλοντικές Κατευθύνσεις

Αυτή η διατριβή διερεύνησε τη εκμάθηση αναπαραστάσεων σε ποικίλες μουσικές παραδόσεις για ανάλυση μουσικού σήματος μέσω μιας σειράς διασυνδεδεμένων μελετών. Ξεκινώντας με την ανάπτυξη του συνόλου δεδομένων Lyra για ελληνική παραδοσιακή μουσική, προχωρώντας σε διερευνήσεις προσεγγίσεων μεταφοράς γνώσης και μάθησης από λίγα παραδείγματα, αξιολογώντας και προσαρμόζοντας θεμελιώδη μοντέλα για ποικίλες μουσικές παραδόσεις, και κορυφώνοντας σε μια ολοκληρωμένη αξιολόγηση υπολογιστικών μεθόδων μουσικής ομοιότητας έναντι της ανθρώπινης διαπολιτισμικής μουσικής αντίληψης, αυτή η έρευνα συνέβαλε στην προώθηση της εκμάθησης αναπαραστάσεων για ποικίλα μουσικά συστήματα.

VII.1 Σύνοψη Συνεισφορών

Οι κύριες συνεισφορές αυτής της διατριβής καλύπτουν την ανάπτυξη συνόλων δεδομένων, μεθοδολογικές καινοτομίες, αξιολόγηση μοντέλων, στρατηγικές προσαρμογής και μελέτες ευθυγράμμισης ανθρώπου-υπολογιστή, αντιμετωπίζοντας τα ερευνητικά ερωτήματα που διατυπώθηκαν στην εισαγωγή.

VII.1.1 Αντιμετώπιση της Επάρκειας Δεδομένων για Ποικίλες Μουσικές Παραδόσεις (EP1)

Το σύνολο δεδομένων Lyra αντιπροσωπεύει την ολοκληρωμένη απάντησή μας στη θεμελιώδη πρόκληση της σπανιότητας δεδομένων στην υπολογιστική ανάλυση παραδοσιακής μουσικής. Αυτή η συλλογή ελληνικής παραδοσιακής και λαϊκής μουσικής, που περιλαμβάνει 80 ώρες υψηλής ποιότητας ηχογραφήσεων, καταδεικνύει μια μεθοδολογία για τη δημιουργία πολιτισμικά θεμελιωμένων συνόλων δεδομένων που μπορούν να υποστηρίξουν την υπολογιστική ανάλυση σεβόμενα τη μουσικολογική ακεραιότητα.

VII.1.2 Κατανόηση της Διαπολιτισμικής Μεταφοράς Γνώσης (EP2)

Η συστηματική διερεύνησή μας για τη μεταφορά γνώσης μεταξύ μουσικών παραδόσεων παρέχει την πρώτη ολοκληρωμένη ανάλυση του πώς η υπολογιστική γνώση μετακινείται σε διαφορετικά μουσικά συστήματα. Η αμφίδρομη φύση της αποτελεσματικής μεταφοράς γνώσης αμφισβητεί τις παραδοχές για την πρωτοκαθεδρία των δυτικά εκπαιδευμένων μοντέλων για την ανάλυση παγκόσμιας μουσικής. Αυτά τα ευρήματα εδραιώνουν ότι τα υπολογιστικά μοντέλα μπορούν να αποκαλύψουν σημαντικές σχέσεις μεταξύ μουσικών συστημάτων που συμπληρώνουν τις παραδοσιακές μουσικολογικές συγκριτικές μελέτες με ποσοτικά, βασισμένα σε δεδομένα, ευρήματα.

VII.1.3 Μάθηση από Περιορισμένα Παραδείγματα σε Μουσικά Πλαίσια (EP3)

Η ανάπτυξη των LC-Protonets αντιμετωπίζει την πρόκληση της σπανιότητας δεδομένων στην έρευνα παγκόσμιας μουσικής μέσω μιας νέας προσέγγισης στη μάθηση πολλών κατηγοριών με λίγα παραδείγματα. Δημιουργώντας πρωτότυπα για συνδυασμούς κατηγοριών αντί για μεμονωμένες κατηγορίες, αυτή η μεθοδολογία επιτρέπει στα υπολογιστικά μοντέλα να μάθουν από το δυναμοσύνολο των διαθέσιμων επισημειώσεων. Οι συνεπείς βελτιώσεις απόδοσης σε διαφορετικά μουσικά σύνολα δεδομένων καταδεικνύουν τη γενικευσιμότητα της προσέγγισης πέρα από τα συγκεκριμένα πλαίσια στα οποία αναπτύχθηκε.

VII.1.4 Αξιολόγηση Θεμελιωδών Μοντέλων σε Μουσικές Παραδόσεις (EP4)

Η ολοκληρωμένη αξιολόγησή μας των σύγχρονων μουσικών θεμελιωδών μοντέλων σε ποικίλες μουσικές παραδόσεις παρέχει κρίσιμα ευρήματα τόσο για τις δυνατότητες όσο και για τους περιορισμούς των τρεχόντων προσεγγίσεων στην καθολική μουσική αναπαράσταση. Το πολυδιάστατο πλαίσιο αξιολόγησης αποκαλύπτει ότι τα θεμελιώδη μοντέλα καταδεικνύουν εντυπωσιακές διαπολιτισμικές ικανότητες σε σύγκριση με προηγούμενες προσεγγίσεις ενώ ταυτόχρονα εμφανίζουν σαφείς δυτικοκεντρικές προκαταλήψεις. Τα προβλήματα μάθησης πλαπλών κατηγοριών από λίγα παραδειγμάτα αποδεικνύονται ιδιαίτερα αποκαλυπτικές, δείχνοντας ότι τα θεμελιώδη μοντέλα δυσκολεύονται με το είδος των σεναρίων χαμηλών πόρων που είναι κοινά στην έρευνα παγκόσμιας μουσικής.

VII.1.5 Προσαρμογή Θεμελιωδών Μοντέλων για Πολυπολιτισμική Κατανόηση (EP5)

Το CultureMERT αντιπροσωπεύει τη συστηματική προσέγγισή μας για την ενίσχυση της πολυπολιτισμικής κατανόησης των θεμελιωδών μοντέλων μέσω συνεχιζόμενης προ-εκπαίδευσης σε ποικίλες μουσικές παραδόσεις. Η στρατηγική προσαρμογής δύο σταδίων αντιμετωπίζει τη θεμελιώδη πρόκληση της καταστροφικής λήθης ενώ επιτρέπει τη σταθερή απόκτηση νέας πολιτισμικής γνώσης. Οι συνεπείς βελτιώσεις σε διαφορετικά προβλήματα μουσικής ταξινόμησης, με μια μέση βελτίωση ROC-AUC 4,43%

σε σύγκριση με το αρχικό θεμελιώδες μοντέλο, καταδεικνύουν την πρακτική αξία της πολιτισμικής προσαρμογής.

VII.1.6 Γεφύρωση Ανθρώπινης Αντίληψης και Υπολογιστικής Μουσικής Ομοιότητας (EP6)

Η ολοκληρωμένη μελέτη διαπολιτισμικής μουσικής ομοιότητας αντιπροσωπεύει την πρώτη συστηματική αξιολόγηση υπολογιστικών μεθόδων μουσικής ομοιότητας έναντι της ανθρώπινης διαπολιτισμικής μουσικής αντίληψης. Η αξιολόγηση αποκαλύπτει μια σαφή ιεραρχία στις υπολογιστικές προσεγγίσεις: τα θεμελιώδη μοντέλα γενικά υπερτερούν των παραδοσιακών χαρακτηριστικών επεξεργασίας σήματος, με το CLAP-Music&Speech να επιτυγχάνει την υψηλότερη μεμονωμένη απόδοση.

Ωστόσο, τα ευρήματά μας αποκαλύπτουν έναν βασικό συμβιβασμό μεταξύ καθολικής μουσικής κατανόησης και πολιτισμικής διακριτότητας στα υπολογιστικά μοντέλα. Ενώ το CLAP-Music&Speech ξεχωρίζει στην ευθυγράμμιση με την ανθρώπινη αντίληψη ομοιότητας, το Qwen2-Audio καταδεικνύει ανώτερη ανίχνευση πολιτισμικών ορίων.

Επιπρόσθετα, η ανάλυση συνεισφοράς χαρακτηριστικών αποκαλύπτει διαφορές μεταξύ ανθρώπινων και υπολογιστικών στρατηγικών επεξεργασίας μουσικής. Οι άνθρωποι δίνουν προτεραιότητα στο μελωδικό περιεχόμενο σε όλες τις διαστάσεις ομοιότητας, ενώ πολλά θεμελιώδη μοντέλα τείνουν να δίνουν έμφαση στα ηχοχρωματικά χαρακτηριστικά.

Το πιο ενθαρρυντικό εύρημα της μελέτης περιλαμβάνει μεθόδους συνόλου που συνδυάζουν χαρακτηριστικά επεξεργασίας σήματος με αναπαραστάσεις θεμελιωδών μοντέλων. Αυτές οι προσεγγίσεις συνόλου επιτυγχάνουν σημαντικές βελτιώσεις, φτάνοντας τιμές 65,1-67,0% στη μετρική triplet agreement και μειώνοντας τα σφάλματα πρόβλεψης κατά 25-30% σε σύγκριση με τις μεμονωμένες μεθόδους.

VII.2 Σύνθεση Ευρημάτων

Εξετάζοντας τις ερευνητικές συνεισφορές ολιστικά, αποκαλύπτονται διάφορα ευρήματα σχετικά με τη φύση της εκμάθησης πολυπολιτισμικών μουσικών αναπαραστάσεων και τις προκλήσεις που είναι εγγενείς στην ανάπτυξη υπολογιστικών προσεγγίσεων που μπορούν να γεφυρώσουν αποτελεσματικά τα πολιτισμικά όρια στη μουσική ανάλυση.

VII.2.1 Η Πρόκληση της Μουσικής Μεταφοράς Γνώσης

Η διερεύνηση της διαπολιτισμικής μεταφοράς αποκαλύπτει ότι η μουσική μεταφορά γνώσης λειτουργεί σύμφωνα με πολύπλοκα μοτίβα που αντιστέκονται σε απλές εξηγήσεις βασισμένες αποκλειστικά στη γεωγραφική εγγύτητα ή τις ιστορικές συνδέσεις. Η ασύμμετρη φύση πολλών σχέσεων μεταφοράς δείχνει ότι οι μουσικές παραδόσεις μπορεί να μοιράζονται ορισμένα υπολογιστικά χαρακτηριστικά ενώ διαφέρουν σε άλλα.

VII.2.2 Περιορισμοί Πόρων και Μεθοδολογική Καινοτομία

Το πρόβλημα των περιορισμών πόρων σε διαφορετικές πτυχές της πολυπολιτισμικής μουσικής έρευνας έχει οδηγήσει σε μεθοδολογικές καινοτομίες που επεκτείνονται πέρα από τις άμεσες εφαρμογές τους. Οι προσεγγίσεις μάθησης από λίγα παραδείγματα όπως τα LC-Protonets καταδεικνύουν ότι σημαντική πρόοδος μπορεί να επιτευχθεί ακόμη και με περιορισμένα επισημειωμένα δεδομένα.

VII.2.3 Το Ερώτημα της Μουσικής Καθολικότητας

Η έρευνα παρέχει αποχρώσεις αποδείξεων σχετικά με τη δυνατότητα καθολικών μουσικών αναπαραστάσεων που μπορούν να περιγράψουν αποτελεσματικά ποικίλες μουσικές παραδόσεις. Τα τρέχοντα θεμελιώδη μοντέλα καταδεικνύουν εντυπωσιακές διαπολιτισμικές ικανότητες που υποδηλώνουν κάποιες κοινές αναπαραστατικές δομές ανάμεσα στις μουσικές κουλτούρες του κόσμου. Η επιτυχία των πολυπολιτισμικών προσεγγίσεων εκπαίδευσης στην ενίσχυση της διαπολιτισμικής γενίκευσης υποδηλώνει ότι η καθολικότητα στη μουσική αναπαράσταση μπορεί να είναι εφικτή, αλλά μόνο μέσω σκόπιμης ένταξης διαφορετικών μουσικών παραδόσεων στη διαδικασία εκπαίδευσης.

VII.2.4 Ευθυγράμμιση Ανθρώπου-Υπολογιστή σε Διαπολιτισμικά Πλαίσια

Η συστηματική σύγκριση υπολογιστικών προσεγγίσεων με την ανθρώπινη διαπολιτισμική μουσική αντίληψη αποκαλύπτει τόσο ενθαρρυντικές ευθυγραμμίσεις όσο και σημαντικά κενά. Το εύρημα ότι τα θεμελιώδη μοντέλα γενικά υπερτερούν των χαρακτηριστικών επεξεργασίας σήματος στην πρόβλεψη ανθρώπινων κρίσεων ομοιότητας επικυρώνει την σύγχρονη τάση στη μουσική τεχνητή νοημοσύνη.

Ωστόσο, η ανακάλυψη ότι οι άνθρωποι δίνουν προτεραιότητα στο μελωδικό περιεχόμενο ενώ τα υπολογιστικά μοντέλα δίνουν έμφαση στα ηχοχρωματικά χαρακτηριστικά επισημαίνει μια θεμελιώδη αναντιστοιχία στις στρατηγικές επεξεργασίας. Αυτή η αναντιστοιχία έχει πρακτικές επιπτώσεις για τα συστήματα μουσικής τεχνολογίας που στοχεύουν να εξυπηρετήσουν διαφορετικές πληθυσμιακές ομάδες χρηστών.

VII.3 Περιορισμοί και Προκλήσεις

Παρά τις συνεισφορές που περιγράφηκαν παραπάνω, αυτή η διατριβή εμπεριέχει αρκετά ζητήματα που περιορίζουν τη γενικευσιμότητα των ευρημάτων της και επισημαίνουν περιοχές που απαιτούν περαιτέρω έρευνα.

Η μελέτη ανθρώπινης αντίληψης, ενώ περιλαμβάνει συμμετέχοντες από 21 χώρες με διαφορετικά μουσικά υπόβαθρα, εμφανίζει αξιοσημείωτη ανισορροπία με την πλειονότητα (62,4%) από την Ελλάδα και άλλες ευρωπαϊκές χώρες, και με σχετικά λίγους συμμετέχοντες από τους πολιτισμούς που αντιπροσωπεύονται στα μουσικά σύνολα δεδομένων.

Τα θεμελιώδη μοντέλα που αξιολογήθηκαν εκπαιδεύτηκαν κυρίως σε εμπορικά μουσικά σύνολα δεδομένων, περιορίζοντας δυνητικά την κατανόησή τους για παραδοσιακά χαρακτηριστικά παγκόσμιας μουσικής. Ορισμένα χαρακτηριστικά επεξεργασίας σήματος εμφανίζουν τάσεις προς δυτικές μουσικές έννοιες που μπορεί να μην αναγνωρίζουν επαρκώς σχέσεις σημαντικές σε μη-δυτικές παραδόσεις.

VII.4 Μελλοντικές Κατευθύνσεις

Βασιζόμενοι στις συνεισφορές και αντιμετωπίζοντας τους περιορισμούς που εντοπίστηκαν σε αυτή την έρευνα, αναδύονται διάφορες υποσχόμενες κατευθύνσεις για την προώθηση της εκμάθησης πολυπολιτισμικών μουσικών αναπαραστάσεων και τη βελτίωση της ευθυγράμμισης ανθρώπου-υπολογιστή στη διαπολιτισμική μουσική κατανόηση.

Η επέκταση της κάλυψης συνόλων δεδομένων αντιπροσωπεύει μια θεμελιώδη προτεραιότητα για την βελτίωση της εκμάθησης αναπαραστάσεων παγκόσμιας μουσικής. Μελλοντική ανάπτυξη συνόλων δεδομένων θα πρέπει να δίνει έμφαση σε πολυτροπικές συλλογές που ενσωματώνουν ήχο, βίντεο, στίχους καθώς και πολιτισμικό πλαίσιο.

Η ανάγκη για ενσωμάτωση πολιτισμικά διαφορετικών συμμετεχόντων σε μελέτες ανθρώπινης αντίληψης αντιπροσωπεύει μια κρίσιμη κατεύθυνση για μελλοντική έρευνα. Με μεγαλύτερους και πιο

αντιπροσωπευτικούς πληθυσμούς συμμετεχόντων από ποικίλες μουσικές παραδόσεις, οι ερευνητές θα μπορούσαν να μετρήσουν την πολιτισμική προκατάληψη των υπολογιστικών μοντέλων αξιολογώντας την ευθυγράμμισή τους με ακροατές από διαφορετικά υπόβαθρα.

Μεθοδολογικές καινοτομίες που βασίζονται στις προσεγγίσεις που αναπτύχθηκαν σε αυτή τη διατριβή προσφέρουν διάφορες υποσχόμενες κατευθύνσεις. Το εύρημα ότι οι άνθρωποι δίνουν προτεραιότητα στο μελωδικό περιεχόμενο ενώ τα θεμελιώδη μοντέλα δίνουν έμφαση στα ηχοχρωματικά χαρακτηριστικά υποδηλώνει την ανάγκη για στόχους προ-εκπαίδευσης που αναγνωρίζουν καλύτερα τις μελωδικές σχέσεις σε ποικίλους πολιτισμούς.

Η επιτυχία των μεθόδων συνόλου στην επίτευξη ανώτερης ευθυγράμμισης ανθρώπου-υπολογιστή υποδηλώνει ότι η μελλοντική έρευνα θα πρέπει να εξερευνήσει πιο εξελιγμένες προσεγγίσεις για τον συνδυασμό διαφορετικών υπολογιστικών μεθόδων.

Η ανάπτυξη θεμελιωδών μοντέλων ειδικά σχεδιασμένων για πολυπολιτισμική μουσική αναπαράσταση απαιτεί θεμελιώδεις αλλαγές στις τρέχουσες προσεγγίσεις ανάπτυξης μοντέλων. Αντί να προσαρμόζουν δυτικοκεντρικά μοντέλα εκ των υστέρων, η μελλοντική εργασία θα πρέπει να επικεντρωθεί στην ανάπτυξη θεμελιωδών μοντέλων προ-εκπαιδευμένων από την αρχή σε διαφορετικές μουσικές παραδόσεις.

Η αξιοποίηση των τεχνικών εξελίξεων σε πρακτικές εφαρμογές με πολιτισμικό και κοινωνικό αντίκτυπο αντιπροσωπεύει μια κρίσιμη κατεύθυνση για μελλοντική εργασία. Εφαρμογές διατήρησης πολιτισμικής κληρονομιάς θα μπορούσαν να αξιοποιήσουν τα υπολογιστικά εργαλεία που αναπτύχθηκαν σε αυτή την έρευνα για να υποστηρίξουν την τεκμηρίωση και ανάλυση μουσικών παραδόσεων που κινδυνεύουν με εξαφάνιση.

Διαπολιτισμικά συστήματα συστάσεων που διευκολύνουν την εύρεση μουσικών από ποικίλες παραδόσεις, με σεβασμό στα διακριτικά τους χαρακτηριστικά, αντιπροσωπεύουν μια αναδυόμενη περιοχή έρευνας στην πολυπολιτισμική ανάκτηση μουσικών πληροφοριών.

Δημιουργικά εργαλεία που υποστηρίζουν τη διαπολιτισμική μουσική δημιουργία και συνεργασία αποτελούν, επίσης, μια αναπτυσσόμενη περιοχή έρευνας στη δημιουργικότητα που υποβοηθείται από την τεχνητή νοημοσύνη.

VII.5 Τελικές Σκέψεις

Αυτή η διατριβή αντιμετώπισε θεμελιώδεις προκλήσεις στην ανάπτυξη υπολογιστικών αναπαραστάσεων που μπορούν να απεικονίσουν αποτελεσματικά τον πλούσιο πολιτισμό των μουσικών παραδόσεων παγκοσμίως ευθυγραμμιζόμενες με την ανθρώπινη διαπολιτισμική μουσική αντίληψη. Μέσω συστηματικής διερεύνησης που καλύπτει την ανάπτυξη συνόλων δεδομένων, μεθοδολογικές καινοτομίες, ολοκληρωμένη αξιολόγηση, προσαρμογή μοντέλων και επικύρωση έναντι της ανθρώπινης αντίληψης, η παρούσα έρευνα συνέβαλε στην προώθηση του τομέα της εκμάθησης πολυπολιτισμικών μουσικών αναπαραστάσεων αποκαλύπτοντας παράλληλα τόσο τις δυνατότητες όσο και τους περιορισμούς των τρεχόντων υπολογιστικών προσεγγίσεων.

Τα ευρήματα αυτής της έρευνας έχουν επιπτώσεις πέρα από τον τεχνικό τομέα της ανάκτησης μουσικών πληροφοριών. Καθώς οι μουσικές τεχνολογίες διαμεσολαβούν όλο και περισσότερο στον τρόπο που ανακαλύπτουμε, δημιουργούμε και μοιραζόμαστε μουσική παγκοσμίως, η ανάπτυξη πιο πολιτισμικά ενήμερων υπολογιστικών προσεγγίσεων που ευθυγραμμίζονται με την ανθρώπινη αντιληπτική κατανόηση καθίσταται ουσιώδης για τη διατήρηση του πλούσιου πολιτισμού της ανθρώπινης μουσικής έκφρασης. Τα υπολογιστικά εργαλεία, τα μεθοδολογικά πλαίσια και οι προσεγγίσεις αξιολόγησης που αναπτύχθηκαν σε αυτή τη διατριβή παρέχουν μονοπάτια για να διασφαλιστεί ότι οι τεχνολογικές εξελίξεις στη μουσική τεχνητή νοημοσύνη ενισχύουν παρά ομογενοποιούν την παγκόσμια μουσική κληρονομιά εξυπηρετώντας παράλληλα τους χρήστες με τρόπους που σέβονται την

πολιτισμική τους αντίληψη.

Κοιτάζοντας προς το μέλλον, η έρευνα που παρουσιάζεται εδώ αντιπροσωπεύει μια θεμελιώδη εργασία σε έναν αναδυόμενο τομέα που βρίσκεται στη σύνδεση της υπολογιστικής νοημοσύνης, της πολιτισμικής κατανόησης και της ανθρώπινης αντίληψης. Οι μεθοδολογίες, τα ευρήματα και οι πόροι ανοιχτού κώδικα που συνεισφέρει αυτή η διατριβή παρέχουν δομικά στοιχεία για μελλοντική έρευνα που μπορεί να προάγει περαιτέρω την ικανότητά μας να αναπαριστούμε και να αναλύουμε υπολογιστικά το πλήρες φάσμα της ανθρώπινης μουσικής έκφρασης.

Ελπίζω ότι αυτή η εργασία θα εμπνεύσει τη συνέχιση της έρευνας πάνω στην εκμάθηση πολυπολιτισμικών μουσικών αναπαραστάσεων με εστίαση στην ευθυγράμμιση ανθρώπου-υπολογιστή, προάγοντας τόσο τις τεχνικές δυνατότητες όσο και την πολιτισμική κατανόηση στο τρέχον ταχέως εξελισσόμενο τεχνολογικό τοπίο. Πέρα από τις τεχνικές συνεισφορές, ελπίζω ότι αυτή η εργασία θα συμβάλει στο να αναδειχθεί ο θεμελιώδης ρόλος που παίζει η μουσική στην ανθρώπινη ανάπτυξη, μια καθολική αλήθεια που υπερβαίνει τα πολιτισμικά όρια.

Chapter 1

Introduction

What is music? What is its purpose in human civilization? Is there an apparent survival value in musical behavior? Charles Darwin was not sure about the role of music when writing his book "The Descent of Man, and Selection in Relation to Sex" in 1871 [1]. Although such profound questions do not have simple answers, contemporary theories converge toward the social aspect of music, recognizing it as a fundamental component of human cultures [2–4].

In this work, we recognize the diversity of music and its connection with culture that is shared by small and large groups of people, by experiencing and analyzing the available data. While we may not have definitive answers about what music constitutes in human civilization, we attempt to shift the attention of the research community toward an aspect that mirrors the sounds expressing the identity of human communities around the world.

This work approaches the musical understanding of artificial intelligence (AI) models as analogous to human learning of the same concepts. It then evaluates model performance not merely to achieve optimal results, but rather to understand the inherent limitations of the task while employing current technology.

Throughout this dissertation, we will examine several methodologies unified by the common goal of achieving understanding across various musical contexts. We will observe how state-of-the-art approaches perform at inadequate levels when large amounts of data are unavailable. We will propose methodologies that extend the limits of current best-performing models in addressing these challenges and we will evaluate the computational realm against human cross-cultural perception.

Ultimately, the observed inadequacy of current technology to understand the refined aspects of human civilization should lead us to greater respect for both ourselves and our societies.

1.1 Motivation and Context

1.1.1 Music as Cultural Expression and Perceptual Experience

Music, often described as a universal language, holds a distinctive place among human cultural expressions. Its pervasiveness across societies and cultures makes it a fascinating subject for both cultural studies and computational analysis. Throughout human history, music has served as a medium for cultural identity, social cohesion, emotional expression, and historical documentation [5–7]. The rich tapestry of global musical traditions reflects the diverse ways in which different cultures have developed unique approaches to melody, harmony, rhythm, timbre, and form.

This dissertation addresses the challenge of multicultural music representation learning, which encompasses both cross-cultural adaptation, enabling computational models to transfer knowledge across different musical traditions, and cross-cultural music understanding, i.e., developing the

capacity to analyze and interpret the distinctive characteristics of diverse musical systems. These complementary aspects together constitute our broader goal of creating computational approaches that can effectively represent and analyze world music traditions.

Critically, music exists for the listener, it is fundamentally a perceptual and experiential phenomenon that lives in the interaction between acoustic signals and human cognition [8]. This perceptual dimension introduces unique challenges for computational approaches, as musical meaning emerges not merely from acoustic properties but from the complex interplay between sound, cultural context, and listening experience [3, 9]. Unlike other domains where computational analysis can rely primarily on structural features, music representation learning must account for the subjective, culturally-situated nature of musical understanding.

The question of music's universality remains contested among scholars [2, 10]. While certain musical elements may transcend cultural boundaries, such as the recognition of emotional expressions or the use of discrete pitches, musical traditions have evolved with distinct characteristics that reflect their cultural contexts. These differences manifest in various aspects: scale systems (e.g., Western 12-tone equal temperament versus Indian 22-shruti systems), rhythmic organizations (e.g., symmetrical Western meters versus complex asymmetrical patterns[1] in Eastern Mediterranean traditions), instrumental timbres, performance practices, and semantic associations.

1.1.2 The Field of Music Information Retrieval

Music Information Retrieval (MIR) has emerged as a vibrant interdisciplinary field that applies computational methods to understand, organize, and access musical content. Drawing from computer science, signal processing, musicology, psychology, and information science, MIR research has developed algorithms and systems for tasks such as music transcription, recommendation, genre classification, beat tracking, chord recognition, structural analysis, and music similarity assessment [11, 12]. These technologies have transformed how we interact with music, enabling personalized streaming services, automated music categorization, and novel creative tools.

However, while tremendous progress has been made in MIR over the past decades, a significant limitation persists: the vast majority of computational models, datasets, and evaluation frameworks are predominantly centered on Western musical traditions [13]. A recent systematic analysis of the current state of musical corpora confirms this bias, showing that Western musical traditions dominate existing datasets with only 5.7% representation of non-Western genres [14]. This specific focus has created analytical challenges in MIR research [15], where computational systems optimized for Western popular and classical music conventions often perform less effectively when applied to diverse musical traditions from other regions of the world.

1.1.3 The Challenge of Musical System Diversity

The predominant focus on European and North American musical traditions in MIR research creates substantial methodological challenges for computational representation of diverse musical systems worldwide. This "American/Eurocentric" bias, where Western musical concepts and analytical frameworks serve as the default lens for all music, has been critically examined by scholars in computational ethnomusicology and corpus studies [16]. Such frameworks often inadequately represent the rich diversity of global musical expressions and may inadvertently prioritize certain

[1] In this computational context, the terms "pattern" and "structure" refer to recurring regularities/motives identified through algorithmic analysis, distinct from the musicological concepts that encompass culturally meaningful melodic, rhythmic, or formal units with specific aesthetic and theoretical significance within musical traditions.

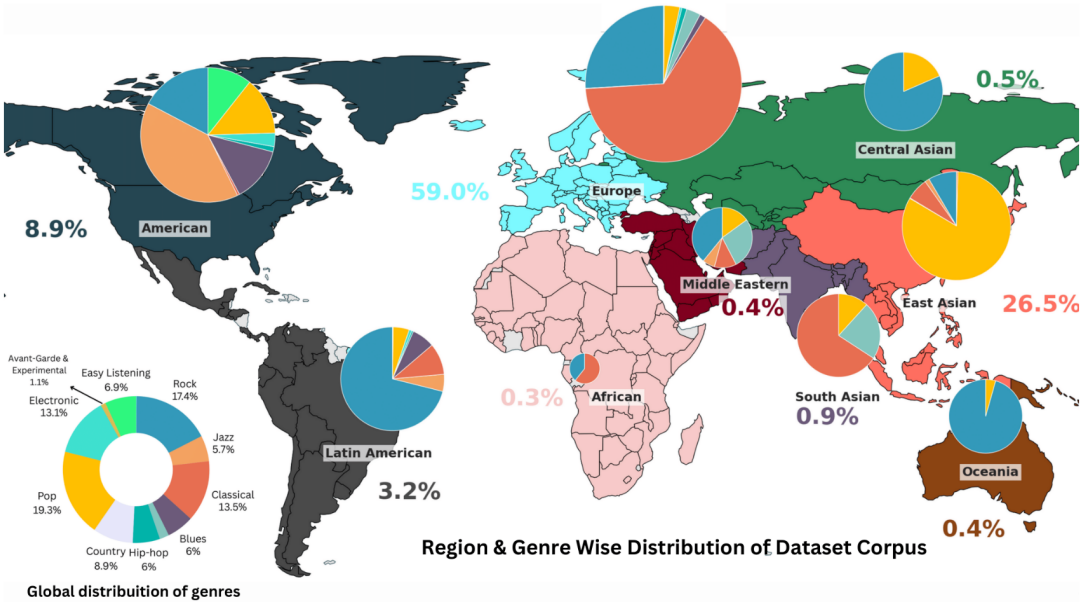

Figure 1.1. Global Distribution of Datasets in MIR research. Regional and genre-wise distribution of dataset corpus showing the overwhelming predominance of Western musical traditions. The bottom left pie chart shows the global distribution of genres, while each pie chart on the map shows the distribution of genres in different regions, with the size proportional to their contribution to the data corpus. Adapted from [14].

musical features while marginalizing others, such as rhythmic complexities found in non-Western traditions.

Recent comprehensive analysis of music datasets reveals the stark extent of this representational bias. As illustrated in Figure 1.1, approximately 94% of the total hours in available music datasets are dedicated to music from the Western world, while only 5.7% are devoted to South Asian, Middle Eastern, Oceanian, Central Asian, Latin American, and African music combined [14].[2] This dramatic imbalance in dataset composition naturally leads to disparate performance of computational models across genres, with models tending to rely on Western tonal and rhythmic structures when processing non-Western musical traditions.

Traditional and folk music from different regions often exhibits distinctive characteristics that may not align with the assumptions embedded in current computational approaches. For instance, Greek traditional music incorporates elements from both European and Eastern Mediterranean musical practices, creating a distinctive musical landscape that requires specialized computational consideration [17]. Similarly, Turkish makam music and Indian classical traditions feature complex melodic structures, modal systems, and microtonal intervals that differ significantly from the equal-tempered scales and harmonic progressions common in European classical and popular music [18, 19].

The challenges of musical system diversity manifest in multiple technical dimensions:

Tonal Systems and Melodic Organization

Western music typically employs a 12-tone equal temperament system with standardized scales and harmony based on tertian structures. In comparison, many other musical traditions utilize dif-

[2] This 94% figure includes East Asian music datasets, as the vast majority of this music falls within pop and rock genres that are considered Western in their musical structure and organization, despite their geographic origin.

ferent interval divisions, microtonal inflections, and alternative organizational principles. Turkish makam music employs a 53-tone division with characteristic melodic progressions and modulations [20]. Indian classical music operates within a system of ragas, each with specific melodic movements, emphasized notes, and expressive characteristics [21]. Greek traditional music incorporates both European and Eastern Mediterranean elements, with regional variations in scale systems and ornamentations [17].

Rhythmic Structures

European and North American musical meters predominantly feature simple (2/4, 3/4) or compound (6/8, 9/8) structures with regular accent patterns. Many other musical traditions employ complex rhythmic cycles, asymmetric meters, and polyrhythmic organizations. Eastern Mediterranean music, including Greek tradition, features rhythms with irregular beat groupings (e.g., 7/8 grouped as 3+2+2). Indian classical music employs elaborately organized talas that can span multiple measures with specific accent patterns [19].

Performance Practices and Ornamentation

Various musical traditions feature distinctive performance practices that present challenges for computational analysis methods developed for Western notation systems. Improvisation plays a central role in many traditions, including Indian classical music (alap and taan)[3], Turkish taksim[4], and Greek taximi[5].

Many non-Western traditions also employ heterophony as a fundamental textural principle, where multiple performers simultaneously present variations of the same melodic line [22]. Unlike Western polyphony with its emphasis on harmonic progression through distinct melodic lines, or strict monophony with identical performance, heterophonic textures create rich, fluid sonorities through subtle variations in timing, ornamentation, and articulation. This practice is prevalent across diverse cultures including Arabic, Turkish, Southeast Asian, and East Asian musical traditions, presenting distinct challenges for computational models trained primarily on Western harmonic structures.

Ornamentation techniques, such as gamaka[6] in Indian music, various glissandi in Eastern Mediterranean traditions, and melismatic embellishments in Greek folk singing, are integral to musical expression rather than optional additions [23].

Instrumental Timbres

The timbral characteristics of region-specific instruments, such as the Greek lyra, Turkish ney, or Indian sitar, present technical challenges for audio analysis algorithms calibrated primarily on orchestral and popular music instrumentation. These instruments often produce complex spectra with distinctive attack-decay profiles and harmonic structures that may require specialized computational approaches different from those optimized for common European and North American instruments [24].

[3] Alap refers to the unmetered, rhythmically free improvisation that opens a performance, while taan consists of rapid melodic passages demonstrating technical virtuosity.

[4] Taksim is an improvised instrumental introduction that explores the melodic and modal characteristics of a specific makam.

[5] Taximi (ταξίμι) is the Greek equivalent of the Turkish taksim, featuring similar modal improvisation but often incorporating distinct regional stylistic elements.

[6] Gamaka can be understood as embellishment done on a note or between two notes.

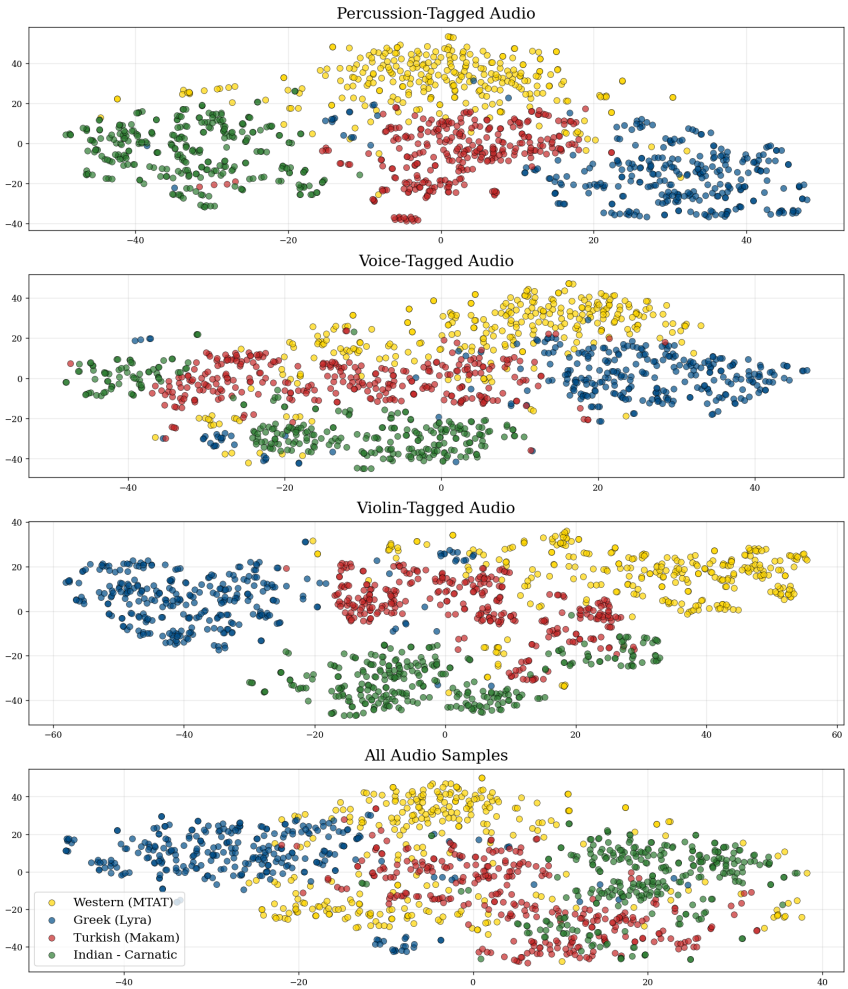

Figure 1.2. Cultural Discriminability Under Tag-Based Filtering. t-SNE visualization of MERT-95M embeddings showing how filtering audio samples by musical attributes (Voice, Violin, Percussion) affects the separation and clustering of different cultural traditions. The bottom panel shows all audio samples, while the top three panels demonstrate how specific musical content influences cross-cultural discriminability in the embedding space.

1.1.4 World Music Representation Learning: Unique Challenges Beyond Language

To understand why multicultural music representation presents distinctive challenges, it is instructive to contrast music with natural language processing (NLP), where cross-lingual transfer has achieved remarkable success. The situation in music AI is comparable to the historical lack of cultural and linguistic diversity in NLP research [25–27], though music presents additional unique challenges. In NLP, languages share fundamental structural similarities: discrete symbolic systems, compositional semantics, and relatively stable mappings between linguistic units and meanings. Transfer learning between languages often leverages shared conceptual structures, even when surface forms differ, underlying semantic relationships can be aligned through techniques like cross-lingual word embeddings [28] or multilingual pre-training [29].

Music, however, presents fundamentally different challenges that make direct adaptation of NLP cross-cultural methodologies insufficient:

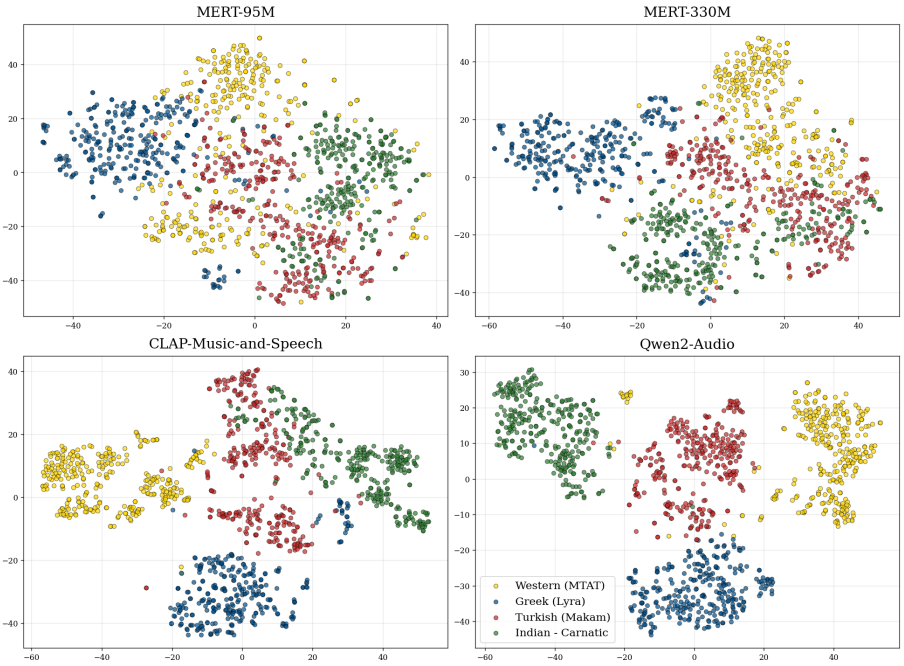

Figure 1.3. Cross-Cultural Representations Across Foundation Models. Comparison of t-SNE projections from four music foundation models (MERT-95M, MERT-330M, CLAP-Music-and-Speech, Qwen2-Audio) applied to the same cross-cultural audio samples. Each model demonstrates different organizational principles and varying degrees of cultural separation, highlighting the model-dependent nature of musical representation learning.

Continuous vs. Discrete Representations: Unlike language's discrete symbolic nature, music operates in continuous acoustic space with culture-specific discretizations. A "note" in Western equal temperament represents a different acoustic and conceptual unit than a microtonal inflection in Turkish makam or a gamaka ornament in Indian classical music. These differences cannot be easily mapped through simple transformations.

Culturally-Embedded Semantic Spaces: Musical meaning emerges from gestalt properties that are conditioned by cultural context rather than following universal compositional rules. The semantic spaces of different musical traditions could be viewed not merely as different vocabularies expressing similar concepts, but as fundamentally distinct organizations of acoustic, temporal, and cultural dimensions. A Greek taximi improvisation and an Indian alap, though both non-metrical modal explorations, operate within entirely different conceptual frameworks that structure their respective musical spaces, reflecting distinct theoretical systems and performance traditions.

Perceptual and Experiential Grounding: Music is tied to embodied perception and cultural conditioning from the listener's perspective [3, 8, 30]. The same acoustic signal can evoke entirely different emotional, aesthetic, and semantic responses across cultural contexts depending on the listener's background. This perceptual variability makes the creation of universal semantic representations more challenging than in language, where reference and meaning maintain more stable relationships across different linguistic communities.

To empirically demonstrate these challenges, we present three complementary visualizations using t-SNE [31] projections of foundation model embeddings (see Section 2.7). Figure 1.2 illustrates how musical content filtering affects the discriminability of different cultural traditions within a single foundation model's representation space.

The visualization reveals that while cultural traditions maintain some separation in the full

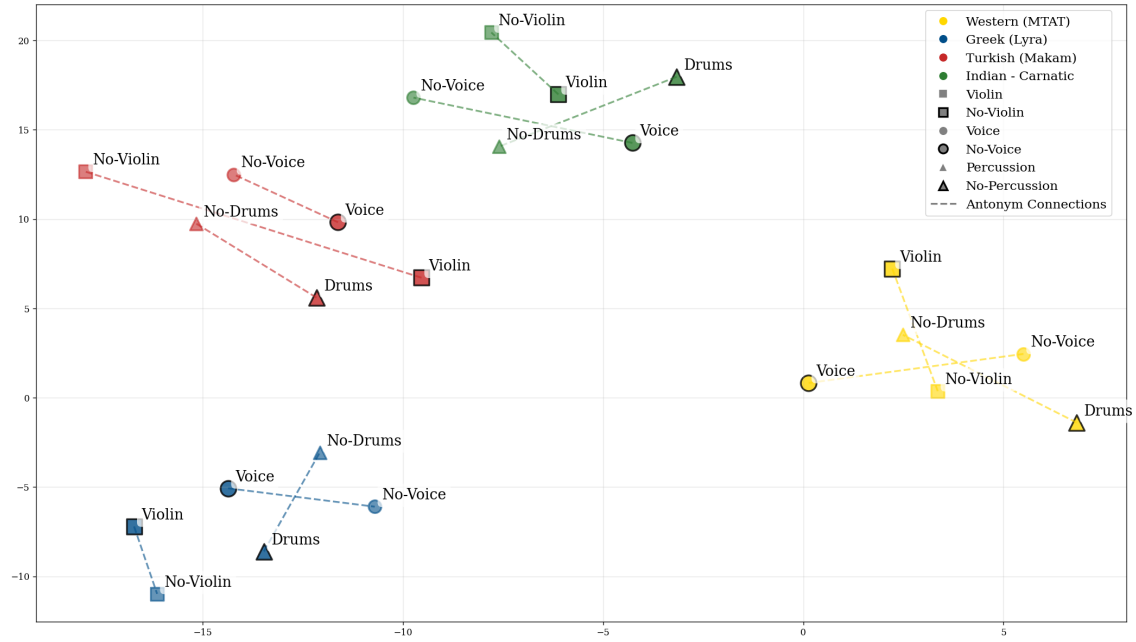

Figure 1.4. Cross-Cultural Semantic Divergence in Musical Concept Organization. Tag centroids computed from average MERT-95M embeddings for audio samples containing specific musical attributes (Voice, Violin, Percussion) and their negations across four musical traditions. The varying distances and orientations between concept pairs demonstrate that even fundamental musical categories are culturally conditioned in their semantic organization, complicating cross-cultural alignment efforts.

audio space, this discriminability varies significantly when filtering by specific musical attributes. This suggests that certain musical concepts may be more culturally distinctive than others, complicating efforts to develop universal musical representations.

Figure 1.3 extends this analysis by comparing how four different foundation models (Section 2.7) organize the same musical content across cultures, revealing both capabilities and limitations in current approaches to musical representation. While some models (e.g., Qwen2-Audio) show clear cultural clustering, others (e.g., MERT-330M) exhibit more distributed representations with less obvious cultural boundaries. This model-dependent variation underscores the challenge of achieving consistent cross-cultural musical understanding across different architectural approaches.

Perhaps most critically, Figure 1.4 demonstrates the fundamental challenge of cross-cultural semantic alignment by examining how seemingly universal musical concepts occupy different semantic positions across traditions. Even for fundamental musical concepts like "Voice," "Violin," and "Percussion," the semantic positioning and relational structures vary significantly across musical traditions. In Western music (yellow), these concepts and their negations form one organizational pattern, while Greek (blue), Turkish (red), and Indian (green) traditions each exhibit distinct semantic arrangements.

The varying distances between antonym pairs (e.g., Violin/No-Violin) across cultures reveals that binary musical concepts are not universally organized, but the challenge extends even deeper: the semantic vectors themselves exhibit cultural inversions. For instance, the direction of the vector from No-Voice to Voice, representing the semantic transformation from absence to presence of vocal elements, points in different directions across traditions. For Western (yellow) and Greek (blue) music, it points to the left while in the other two traditions it points to the right. These

directional inversions indicate that the very meaning of vocal presence and absence is culturally conditioned, reflecting different aesthetic priorities and theoretical frameworks.

This empirical evidence demonstrates that musical semantic spaces resist the kind of cross-lingual alignment successful in natural language processing. Unlike linguistic concepts that maintain relatively stable referential relationships across languages, musical concepts are fundamentally shaped by the theoretical, aesthetic, and performance frameworks of their respective traditions. The Voice concept in Indian classical music, embedded within a context of gamakas (ornamentations) and raga-specific melodic movements, occupies a different semantic position than the Voice concept in Western popular music, which operates within harmonic progressions and regular metrical structures.

These distinctions necessitate specialized approaches for multicultural music representation learning that go beyond straightforward adaptation of cross-lingual techniques.

1.1.5 The Path to Multicultural Representations

Creating computational representations that effectively capture diverse musical traditions requires addressing multiple interconnected challenges. The semantic divergence and model-dependent variations illustrated in the previous section motivate the development of specialized strategies for multicultural music representations:

Cross-Cultural Transfer Learning and Model Adaptation: Investigating how knowledge learned from one musical tradition can inform understanding of another, while respecting the distinctive characteristics of each system. At scale, this includes enhancing large-scale pre-trained models and foundation models to better represent diverse musical traditions through continual learning, fine-tuning, and novel adaptation strategies that prevent catastrophic forgetting while acquiring new cultural knowledge.

Low-Resource Learning Approaches: Developing methods that can learn meaningful representations from limited examples, crucial for underrepresented musical traditions where annotated data is scarce. Few-shot and meta-learning approaches become essential for including diverse musical cultures in computational models.

Human-Centered Evaluation: Integrating human perception studies to validate computational approaches, ensuring that similarity measures align with how listeners from diverse cultural backgrounds actually perceive musical relationships. This human-in-the-loop validation becomes crucial for developing culturally aware music technology systems.

Multi-Task and Multi-Modal Learning: Integrating information from multiple sources, audio, metadata, cultural context, performance practices, to create richer representations that capture both acoustic and cultural dimensions of musical expression.

The central premise of this dissertation is that achieving truly multicultural music representation requires not simply scaling existing approaches, but developing new methodologies that account for the unique properties of musical expression across cultures, as empirically demonstrated by the distinct semantic organizations and model-dependent variations observed across musical traditions.

1.1.6 Computational Ethnomusicology and Dataset Development

The field of Computational Ethnomusicology applies computational methods to study diverse musical traditions [32]. This emerging discipline combines ethnomusicological knowledge with MIR techniques to develop culturally appropriate computational approaches to world music analysis.

Creating structured datasets of traditional music from various cultures is vital for enabling computational analysis and cross-cultural musical comparisons.

A pivotal development in this field was the CompMusic project [33], which created the corpora, and set the criteria for doing so, for five distinct musical cultures: Hindustani (North Indian), Carnatic (South Indian), Turkish-makam, Beijing Opera, and Arab-Andalusian traditions. This project was further supported by the establishment of the Folk Music Analysis (FMA) workshops[7], established in 2011, which created a dedicated community and scholarly forum for computational approaches to traditional music.

Building on these foundations, several efforts have been made to develop specialized datasets, including collections of Dutch melodies [34], Indian art music [19], Arab-Andalusian and Flamenco music [23, 35], Georgian vocal music [36], and Chinese traditional music [24]. However, these collections remain underrepresented in mainstream MIR research and applications, and their limited size and scope compared to Western music datasets present challenges for developing robust computational models.

1.1.7 The Rise of Deep Learning and Foundation Models

Recent advances in deep learning have revolutionized MIR by introducing pre-trained models that provide informative audio embeddings applicable to various tasks. Models such as VGG-ish [37], Musicnn [38], and Audio Spectrogram Transformer (AST) [39] have demonstrated impressive performance across multiple MIR tasks. However, the majority of these models have been trained predominantly on Western musical data, raising important questions about their effectiveness when applied to different musical cultures.

The emergence of foundation models in music [40–42] presents both opportunities and challenges for multicultural music analysis. Following the paradigm established in natural language processing and computer vision, music foundation models are trained on large-scale data to learn general-purpose representations applicable to diverse downstream tasks. Models such as MERT [40], CLAP [41], and Qwen-Audio [42] have demonstrated state-of-the-art performance across various MIR benchmarks.

However, the implicit universality claims of these foundation models deserve critical examination, particularly in light of their predominant training on Western-centric data [14, 43]. The extent to which these models can represent and analyze diverse musical traditions beyond their training distribution remains an open question that this dissertation systematically addresses (see Section 5.2).

1.1.8 The Need for Cross-Cultural Computational Methods

The technical challenges outlined above highlight the need for adaptable computational approaches to music representation learning that can accommodate diverse musical systems. Such approaches require:

1. Development of comprehensive, high-quality datasets representing multiple regional musical traditions
2. Methods for effective knowledge transfer between different musical systems
3. Techniques for learning from limited examples, addressing the data scarcity common in many regional musical traditions outside mainstream commercial genres

[7] https://www.folkmusicanalysis.org/

4. Systematic evaluation of existing models across different musical systems to identify domain-specific limitations

5. Approaches for adapting foundation models to better represent the distinctive characteristics of world music traditions

6. Human-centered evaluation frameworks that validate computational approaches against cross-cultural music perception

This dissertation addresses these methodological needs through a series of interconnected studies that collectively advance the field of multicultural music representation learning.

1.2 Problem Statement

1.2.1 Core Research Problem

The central challenge addressed in this dissertation is the limited ability of current computational models to effectively represent and analyze musical traditions beyond Western conventions. Despite advances in music information retrieval, most computational approaches demonstrate reduced performance when applied to musical systems with fundamentally different organizational principles. This performance gap stems from several interconnected methodological factors rooted in the predominant development of MIR around specific musical paradigms common in European and North American traditions [15].

A critical dimension of this challenge involves understanding how computational similarity measures align with human cross-cultural music perception. Current computational approaches often fail to capture the nuanced relationships between musical styles, instruments, and aesthetic principles that define different musical cultures, a limitation that becomes evident when these approaches are evaluated against human judgment.

The goal of developing culturally appropriate computational models extends beyond technical considerations to encompass broader questions about representation, accessibility, and preservation of cultural heritage. As digital technologies increasingly mediate our musical experiences, through streaming platforms, recommendation systems, and analysis tools, the underrepresentation of non-Western traditions in these technologies risks marginalizing important aspects of global musical culture and perpetuating existing biases in musical representation [44].

1.2.2 Specific Challenges

This core technical problem manifests in several interconnected challenges that span from fundamental data and modeling issues to practical implementation concerns. These challenges can be organized into four primary areas that this dissertation addresses systematically.

1) Representation Learning

The foundation of effective multicultural music analysis lies in developing computational representations that can capture the distinctive characteristics of diverse musical traditions while enabling meaningful comparison and knowledge transfer across systems.

Representational bias. Existing datasets and models embed systematic biases that favor certain musical characteristics over others, primarily those aligned with Western analytical frameworks. The predominance of harmonic structures and regular meters in training data creates models that excel at recognizing chord progressions and simple rhythmic patterns but struggle with modal music, microtonal inflections, and complex asymmetrical rhythmic cycles common in many traditional music systems. Scale systems that assume 12-tone equal temperament inadequately represent the rich microtonal traditions found in Turkish makam, Indian classical music, or Arab maqam systems. Furthermore, annotation schemes designed for commercial music often fail to capture culturally relevant attributes of traditional music, such as the specific ornamentations that define regional styles or the improvisational practices that are central to many musical traditions.

Data scarcity. Many regional and traditional musical systems face substantial imbalances in data quantity compared to commercially dominant genres, with some traditions having only hundreds of annotated examples compared to millions available for popular Western music. This scarcity extends beyond mere quantity to encompass diversity within traditions, where available datasets may inadequately represent regional substyles, historical periods, or different performance contexts. The resulting data limitations create fundamental bottlenecks for supervised learning approaches, necessitating specialized methods that can learn effectively from minimal examples while avoiding overfitting to the limited available data.

Cross-cultural transfer and semantic spaces. Unlike natural language processing, where cross-lingual transfer can leverage shared conceptual structures, musical semantic spaces are not easily alignable across traditions. The semantic organization of musical elements is deeply shaped by cultural conditioning, creating distinct conceptual frameworks that resist simple transformation or alignment. A Greek taximi improvisation and an Indian alap, while both representing non-metrical modal explorations, operate within fundamentally different musical and cultural paradigms that structure their respective semantic spaces. The challenge lies in identifying which aspects of musical representation are transferable across systems while preserving the culture-specific elements that define each tradition's distinctive character.

2) Music Understanding Tasks

The representation learning challenges described above manifest concretely in the computational tasks used to evaluate musical understanding, particularly in classification scenarios that are central to music information retrieval.

Classification challenges. Musical classification tasks in multicultural music contexts typically exhibit extremely long-tailed label distributions where many culturally significant attributes have very few examples. This creates scenarios where traditional supervised learning approaches either exclude rare categories entirely or perform poorly due to class imbalance. The hierarchical and overlapping nature of musical categories adds additional complexity, as regional substyles, performance practices, and cultural contexts create intricate relationships that resist simple categorical organization. Consider the challenge of distinguishing between a Greek bouzouki and an Irish bouzouki, while sharing structural similarities and even a name, they operate within entirely different musical contexts, employ distinct playing techniques, and carry different cultural meanings that computational models must learn to differentiate.

Multi-label scenarios. Musical pieces simultaneously belong to multiple overlapping categories spanning genre, instrumentation, regional style, and performance context, further complicating the classification tasks. Unlike single-label classification problems, these scenarios require models to capture complex co-occurrence patterns while handling the sparse annotation patterns common in traditional music datasets. The challenge intensifies when considering that the relevance and interpretation of musical labels can vary significantly across cultural contexts, requiring models that can adapt their understanding of categorical relationships based on the musical tradition being analyzed.

Evaluation complexity. Traditional classification accuracy measures may not adequately capture the cultural significance of correctly identifying rare but important musical attributes, while cross-cultural evaluation requires developing metrics that can assess model performance across different musical systems without imposing inappropriate external standards. The perceptual and experiential nature of musical meaning adds another dimension to evaluation, as computational success should ideally align with culturally informed musical understanding rather than purely statistical optimization.

3) Human-Computational Alignment

A critical challenge involves understanding how computational approaches to music similarity align with human cross-cultural music perception, requiring systematic evaluation frameworks that bridge algorithmic processing and perceptual understanding.

Cross-cultural similarity perception. Human perception of musical similarity varies significantly across cultural contexts, with listeners from different musical backgrounds potentially hearing and evaluating the same musical relationships in fundamentally different ways. Traditional computational similarity measures, typically developed and validated on Western musical content, may not capture the perceptual dimensions that are most salient to listeners from diverse cultural backgrounds. The challenge involves developing evaluation frameworks that can systematically assess how different computational approaches align with human similarity judgments across multiple cultural traditions.

Multi-dimensional similarity assessment. Musical similarity encompasses multiple overlapping dimensions including overall musical characteristics, cultural identity, and personal preference, each of which may be weighted differently across cultural contexts. Computational approaches must account for these multi-dimensional aspects of similarity while handling the inherent subjectivity and cultural conditioning that shapes human musical perception. The development of appropriate evaluation frameworks requires careful consideration of how to elicit and analyze human similarity judgments in ways that respect cultural differences while enabling systematic comparison of computational approaches.

Signal processing versus learned representations. Understanding the relative strengths and limitations of interpretable signal processing features compared to learned representations from foundation models requires systematic evaluation against human perception. While signal processing features offer interpretability through their connection to established music theory concepts, they may incorporate Western musical assumptions that limit their cross-cultural validity. Con-

versely, foundation models may capture complex patterns that align better with human perception but lack interpretability regarding which musical dimensions drive their similarity assessments.

4) Multicultural Learning

The goal of developing computational models that can effectively represent diverse musical traditions requires addressing the complex challenges of multicultural model adaptation and knowledge integration.

Model adaptation. Adapting foundation models to better represent diverse musical traditions must navigate the risk of catastrophic forgetting, where learning new cultural knowledge degrades performance on previously acquired traditions. The challenge lies in developing adaptation strategies that can acquire tradition-specific representations while preserving the general musical knowledge that enables cross-cultural understanding. This requires careful balancing of plasticity and stability, allowing models to learn new cultural patterns while maintaining their ability to recognize universal musical elements that transcend cultural boundaries.

Cultural authenticity and computational efficiency. Adaptation strategies must preserve the integrity of distinctive musical characteristics rather than homogenizing different traditions toward a common representation. This cultural preservation requirement often conflicts with computational efficiency goals, as maintaining separate representations for different traditions increases model complexity and resource requirements. The challenge involves developing approaches that can capture cultural specificity while remaining computationally tractable for practical deployment across diverse musical contexts.

Knowledge integration across traditions. It represents the most ambitious aspect of multicultural learning, requiring models that can leverage similarities between musical systems while respecting their distinctive characteristics. This involves developing sophisticated understanding of which musical concepts transfer across cultures and which require culture-specific modeling. The goal is not to create a single universal musical representation, but rather to develop adaptive systems that can dynamically adjust their processing based on the cultural context while maintaining the ability to identify meaningful relationships and patterns across different musical traditions. Success in this area would enable computational tools that can support cross-cultural musical understanding while preserving the rich diversity that defines global musical expression.

1.2.3 Practical Implications

These technical challenges have significant practical implications for the development and deployment of music technologies. Current limitations affect:

- **Music Recommendation Systems**: Existing platforms may demonstrate reduced performance when representing and recommending music from diverse traditions, potentially reinforcing cultural biases in music consumption patterns and limiting exposure to diverse musical cultures.
- **Music Education Tools**: Educational technologies based primarily on Western music concepts may provide inadequate support for learning in diverse musical traditions, failing to recognize culturally appropriate pedagogical approaches or assessment criteria.

- **Cultural Heritage Preservation**: Digital archives and computational analysis tools may have technical limitations in capturing and preserving the nuances of less-documented musical traditions, potentially losing important cultural information during digitization and analysis processes.
- **Creative Technologies**: Music production tools and algorithmic composition systems often reflect specific musical conventions, potentially limiting their technical applicability for creators working in different traditions and constraining creative expression within Western musical frameworks.
- **Music Similarity and Search Systems**: Current similarity-based search and discovery systems may fail to capture the perceptual dimensions that are most relevant to listeners from diverse cultural backgrounds, leading to suboptimal user experiences and reduced effectiveness in cross-cultural music discovery.
- **Global Applicability**: The technical capabilities of music AI technologies may be unevenly distributed across musical traditions, with users interested in less-represented traditions receiving less effective technological support.

These challenges call for innovative approaches to multicultural music representation learning that can overcome data limitations, leverage knowledge transfer across different musical systems, develop more versatile computational models capable of representing diverse musical traditions, and ensure that computational approaches align with human cross-cultural music perception. Addressing these challenges requires interdisciplinary collaboration between MIR researchers, ethnomusicologists, cultural heritage specialists, and practitioners from diverse musical backgrounds.

1.3 Research Questions

1.3.1 Central Research Question

How can computational approaches be developed to effectively understand music from diverse cultures worldwide, and how well do these approaches align with human cross-cultural music perception?

1.3.2 Primary Research Questions

This dissertation addresses six interconnected research questions that progressively build upon each other, from fundamental data availability to advanced model adaptation and human-computational alignment:

1. **RQ1: How can high-quality datasets for underrepresented musical traditions be developed to support computational analysis and cross-cultural comparison?**
2. **RQ2: To what extent can knowledge be effectively transferred between different musical systems, and what patterns of transferability exist across diverse musical traditions?**
3. **RQ3: How can computational models learn effectively from limited examples in multicultural music contexts, particularly for rare but culturally significant musical attributes?**

4. **RQ4: What are the cross-cultural capabilities and limitations of state-of-the-art music foundation models when applied to diverse musical traditions?**
5. **RQ5: How can foundation models be adapted to better represent diverse musical traditions while preserving their general musical knowledge and avoiding catastrophic forgetting?**
6. **RQ6: How do computational music similarity measures compare to human cross-cultural music perception, and what factors drive similarity judgments across different musical traditions?**

1.3.3 Supporting Research Questions

In addition to these primary questions, this dissertation explores several supporting inquiries:

1. How can evaluation frameworks be designed to appropriately assess cross-cultural music representation learning while accounting for human perceptual validation?
2. What computational resources and strategies are required for effective multicultural model adaptation?
3. What specific transfer patterns exist between particular musical traditions, and how do they reflect cultural and historical relationships?
4. Which musical dimensions (melody, rhythm, harmony, timbre) are most predictive of human similarity judgments across different cultural contexts?
5. How can ensemble methods combining interpretable features and learned representations improve alignment with human cross-cultural music perception?
6. How can open science principles be applied to enable reproducible research in multicultural music analysis?

Through addressing these research questions systematically, this dissertation contributes to a more comprehensive understanding of multicultural representation learning for music and its relationship to human cross-cultural music perception.

1.4 Contributions

This dissertation advances multicultural music representation learning through a comprehensive research program that directly addresses each of the research questions posed above. The contributions span from fundamental dataset development to sophisticated model adaptation techniques and human-computational alignment studies, providing both theoretical insights and practical solutions.

1.4.1 Addressing Data Availability for Diverse Musical Traditions (RQ1)

The Lyra Dataset [45] represents our response to the fundamental challenge of data scarcity in computational analysis of traditional music. This comprehensive collection of Greek traditional music, comprising 1,570 pieces with approximately 80 hours of high-quality recordings, demonstrates a methodology for creating culturally-grounded datasets that can support sophisticated

computational analysis. The dataset addresses RQ1 by establishing consistent recording quality through systematic collection from academic documentary sources, developing rich metadata annotations that capture musicologically relevant attributes including instrumentation, geographic origin, and genre classification, and providing structured access through timestamped multimedia links. Expert validation throughout the collection and annotation process ensures cultural authenticity, while baseline classification experiments establish performance benchmarks for key musicological attributes, providing a foundation for future computational research on Greek traditional music.

1.4.2 Understanding Cross-Cultural Knowledge Transfer (RQ2)

Our **Cross-Cultural Transfer Learning Framework** [46] provides the first systematic investigation of knowledge transfer patterns between diverse musical systems. This comprehensive methodology directly addresses RQ2 by evaluating multiple deep audio embedding models across musical corpora spanning Western, Mediterranean, and Indian traditions, revealing previously unknown patterns of cross-cultural musical relationships. The framework enables systematic comparison of single-domain versus cross-domain learning approaches, establishing quantitative metrics for measuring transfer effectiveness across different musical systems. The research demonstrates that computational models can indeed benefit from knowledge transfer between diverse musical systems, while also revealing the asymmetric and complex nature of these relationships. These findings provide new insights into computational similarities between musical cultures and establish that bidirectional knowledge transfer is possible, though effectiveness varies significantly based on the specific traditions involved.

1.4.3 Learning from Limited Examples in Musical Contexts (RQ3)

The development of **LC-Protonets** [47] directly tackles RQ3 by introducing a novel approach to multi-label few-shot learning specifically designed for musical classification scenarios. This methodology extends Prototypical Networks to handle the complex multi-label scenarios common in music analysis, creating prototypes for label combinations rather than individual labels. The approach significantly improves performance across diverse music datasets and enables the inclusion of rare but culturally significant musical attributes in computational models. Integration with pre-trained embedding spaces enhances performance while maintaining computational efficiency, and comprehensive evaluation across diverse music datasets demonstrates consistent improvements over existing few-shot learning methods. The two-step learning framework shows particular efficacy for imbalanced datasets, enabling the expansion of tag sets to include underrepresented musical categories that would otherwise be excluded from computational analysis.

1.4.4 Evaluating Foundation Models Across Musical Traditions (RQ4)

Our **Foundation Model Evaluation Framework** provides the first comprehensive assessment of state-of-the-art music foundation models across diverse musical traditions, directly addressing RQ4. This multi-faceted evaluation employs complementary methodologies including linear probing, supervised fine-tuning, and few-shot learning to assess model capabilities under different resource constraints. The systematic comparison of five state-of-the-art audio foundation models across six musical corpora representing different traditions reveals both impressive cross-cultural capabilities and significant limitations, particularly in low-resource scenarios and for culturally distant traditions. The research establishes benchmarks for future foundation model

development while identifying specific areas where current models exhibit Western-centric biases. These findings provide crucial insights into the current state of universal music representation and highlight the need for more inclusive training approaches.

1.4.5 Adapting Foundation Models for Cultural Inclusivity (RQ5)

CultureMERT, developed in collaboration with Angelos-Nikolaos Kanatas, represents our comprehensive answer to RQ5 through a novel two-stage continual pre-training strategy that enables stable adaptation of foundation models to diverse musical traditions. This approach addresses the fundamental challenge of catastrophic forgetting by carefully balancing plasticity and stability during adaptation. Training on a 650-hour diverse data mix comprising Greek, Turkish, and Indian music demonstrates that foundation models can be effectively enhanced to better represent non-Western traditions while preserving their general capabilities. Systematic evaluation across multiple music tagging tasks confirms consistent improvements, while analysis of catastrophic forgetting provides insights into effective mitigation strategies. Additionally, our exploration of **Task Arithmetic for Music Models** provides an alternative adaptation approach that merges independently adapted models in weight space, offering a resource-efficient method that eliminates the need for simultaneous access to all cultural datasets.

1.4.6 Bridging Human Perception and Computational Music Similarity (RQ6)

Our **Cross-Cultural Music Similarity Study** provides the first systematic evaluation of computational music similarity methods against human cross-cultural music perception, directly addressing RQ6. This comprehensive investigation collected human similarity annotations from 125 participants across diverse backgrounds, evaluating 1,130 unique audio pairs from nine musical datasets across three similarity dimensions: overall musical similarity, cultural similarity, and recommendation-level similarity. The systematic comparison of both traditional signal processing features and seven state-of-the-art foundation models using five complementary evaluation metrics reveals that foundation models achieve superior alignment with human perception, with melody consistently emerging as the most predictive traditional feature across all cultural contexts.

The research uncovers fundamental differences between human and computational processing strategies, demonstrating that humans prioritize melodic content while foundation models emphasize timbral characteristics, a misalignment with significant implications for music AI system design. Cross-cultural discrimination analysis reveals substantial gaps between human cultural awareness and computational capabilities, while the apparent underperformance of culturally adapted models reflects the influence of listener cultural background on evaluation outcomes. Most encouragingly, ensemble methods combining interpretable features with learned representations achieve substantial improvements, reducing prediction errors by 25-30% and demonstrating the complementary value of diverse computational approaches. This study establishes both a comprehensive evaluation framework for cross-cultural music similarity assessment and actionable insights for developing culturally aware music technology systems that align with human cross-cultural music understanding.

1.4.7 Integrative Insights and Open Science Contributions

Beyond addressing individual research questions, this dissertation provides integrative contributions that span multiple aspects of multicultural music representation learning. Our **Cross-**

Cultural Transferability Analysis quantifies transfer effectiveness between Western, Greek, Turkish, and Indian musical traditions, identifying asymmetries in cross-cultural transfer patterns and revealing insights into shared musical characteristics across different traditions. These findings contribute to theoretical understanding of computational relationships between musical cultures while providing practical guidance for future cross-cultural music research.

The **Human-Computational Alignment Analysis** bridges the gap between algorithmic processing and perceptual understanding by systematically evaluating how different computational approaches align with human cross-cultural music similarity judgments. This work provides empirical evidence for the complementary strengths of interpretable signal processing features and learned foundation model representations, while establishing that ensemble approaches can achieve superior alignment with human perception.

Recognizing the importance of reproducible research and community engagement, we have made substantial **Open-Source Contributions** including public release of the Lyra dataset with structured metadata, open-source implementations of LC-Protonets, comprehensive evaluation frameworks and benchmarks for cross-cultural model assessment, public release of adapted foundation models, and the cross-cultural music similarity dataset with human annotations. These resources enable the broader research community to build upon our work and advance the field of multicultural music representation learning.

Through this integrated approach to addressing fundamental research questions in multicultural music analysis, the dissertation provides both theoretical insights into the nature of musical representation across cultures and practical tools that advance the state-of-the-art in computational music analysis. The contributions collectively demonstrate that while significant challenges remain in developing universal music representations, substantial progress can be achieved through systematic research that combines dataset development, methodological innovation, comprehensive evaluation, adaptive model enhancement, and human-centered validation.

1.5 Associated Publications

The work presented in this dissertation has been shared at international peer-reviewed conferences and journals, or is currently under review. The following list includes all publications associated with this dissertation, followed by their correspondence to the dissertation chapters and a description of the author's contributions to each.

Peer-reviewed Publications (First Author)

1. **Charilaos Papaioannou**, Ioannis Valiantzas, Theodore Giannakopoulos, Maximos A. Kaliakatsos-Papakostas, and Alexandros Potamianos, "**A Dataset for Greek Traditional and Folk Music: Lyra**", in Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR 2022), Bengaluru, India, 2022, pp. 377-383 [45].

2. **Charilaos Papaioannou**, Emmanouil Benetos, and Alexandros Potamianos, "**From West to East: Who can understand the music of the others better?**", in Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR 2023), Milan, Italy, 2023, pp. 311-318 [46].

3. **Charilaos Papaioannou**, Emmanouil Benetos, and Alexandros Potamianos, "**LC-Protonets:**

Multi-Label Few-Shot Learning for World Music Audio Tagging", IEEE Open Journal of Signal Processing, vol. 6, pp. 138-146, 2025 [47].

4. **Charilaos Papaioannou**, Emmanouil Benetos, and Alexandros Potamianos, "**Universal Music Representations? Evaluating Foundation Models on World Music Corpora**", to appear in Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR 2025), Daejeon, Korea, 2025 [48].

Peer-reviewed Publication (Corresponding Author)

5. Angelos-Nikolaos Kanatas, **Charilaos Papaioannou**, and Alexandros Potamianos, "**CultureMERT: Continual Pre-Training for Cross-Cultural Music Representation Learning**", to appear in Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR 2025), Daejeon, Korea, 2025 [49].

Publication Under Review (First Author)

6. **Charilaos Papaioannou**, Emmanouil Benetos, and Alexandros Potamianos, "**Cross-Cultural Music Similarity: Bridging Human Perception, Signal Processing, and Foundation Models**", under review for the Transactions of the International Society for Music Information Retrieval (TISMIR) journal.

Chapter 3 is based on publication [1], which introduced the Lyra dataset for Greek traditional and folk music. As the first author, I led the dataset development, system design, experimental evaluation and manuscript preparation. Ioannis Valiantzas contributed to data annotation and musicological analysis, Theodore Giannakopoulos and Maximos A. Kaliakatsos-Papakostas assisted with technical implementation and manuscript preparation respectively, and Alexandros Potamianos provided supervision and guidance throughout the project.

Chapter 4 draws from publications [2] and [3]. For publication [2], I was the lead contributor, developing the transfer learning framework and conducting all experiments on cross-cultural music understanding. Emmanouil Benetos provided guidance on experimental design and manuscript revision, while Alexandros Potamianos offered supervision and feedback on the research direction. For publication [3], I was responsible for the development of the LC-Protonets method, implementation, experimental evaluation, and manuscript preparation, with Emmanouil Benetos and Alexandros Potamianos providing theoretical guidance and critical feedback.

Chapter 5 incorporates content from publications [4] and [5]. For publication [4], I led the research as first author, designing the evaluation framework, conducting the experiments, and writing the manuscript, with guidance from Emmanouil Benetos and Alexandros Potamianos. For publication [5], I contributed to the conceptual framework, experimental design, and data analysis, providing guidance on cultural adaptation strategies and evaluation methodologies, while Angelos-Nikolaos Kanatas led the implementation as first author, developing the continual pre-training and task arithmetic approaches.

Chapter 6 is based on manuscript [6], which presents the comprehensive cross-cultural music similarity study. As the first author, I led the research design, human annotation study coor-

dination, computational analysis implementation, and manuscript preparation. I designed and conducted the large-scale human perception study involving 125 participants, implemented all computational similarity measures including signal processing features and foundation model evaluations, developed the comprehensive evaluation framework, and performed all statistical analyses and ensemble method evaluations. Emmanouil Benetos and Alexandros Potamianos provided theoretical guidance, methodological feedback, and critical review throughout the research process.

1.6 Dissertation Structure

The remainder of this dissertation is organized as follows:

Chapter 2: Background This chapter provides a comprehensive background on the concepts, methods, and related work relevant to this dissertation. It covers fundamental concepts in music signal analysis and representation learning, examines foundation models in music, reviews literature on cross-cultural music analysis and computational ethnomusicology, and synthesizes relevant work from all included papers. This chapter establishes the conceptual and methodological foundation for the research presented in subsequent chapters, spanning topics from music signal processing fundamentals to world music datasets, annotation challenges, and human perception studies.

Chapter 3: The Lyra Dataset: A Resource for Greek Traditional and Folk Music This chapter presents the development, structure, and analysis of the Lyra dataset for Greek traditional music. It details the data collection process, metadata structure, and distribution of pieces across genres, instrumentation, and geographic origins. The chapter concludes with baseline classification experiments for genre, instrument, and regional classification, demonstrating the dataset's utility for computational analysis and establishing performance benchmarks for future research on Greek traditional music.

Chapter 4: Learning Across Cultures This chapter explores approaches for transferring knowledge between musical traditions, combining transfer learning and few-shot learning methodologies. The first part investigates knowledge transfer patterns between different musical systems using deep audio embedding models, while the second part introduces LC-Protonets, a novel multi-label few-shot learning method designed for scenarios with limited annotated data. Together, these approaches address complementary aspects of the challenge of learning across diverse musical traditions with varying amounts of available data.

Chapter 5: Foundation Models for Diverse Music Traditions This chapter evaluates the capabilities of state-of-the-art music foundation models across diverse musical traditions and explores approaches for enhancing their representational capacity. It presents a comprehensive evaluation framework assessing multiple foundation models across varied musical corpora using three complementary methodologies. It then introduces CultureMERT, an adapted foundation model developed through a two-stage continual pre-training strategy, along with an exploration of task arithmetic as an alternative approach to model adaptation.

Chapter 6: Cross-Cultural Music Similarity: Bridging Human Perception and Computational Methods This chapter presents the first comprehensive evaluation of computational music similarity methods against human cross-cultural music perception. It details a large-scale

human annotation study involving 125 participants evaluating 1,130 audio pairs from nine diverse musical traditions across three similarity dimensions. The chapter systematically compares human judgments against both signal processing features and foundation model representations, reveals fundamental differences in processing strategies between humans and machines, and demonstrates the effectiveness of ensemble methods that combine interpretable features with learned representations.

Chapter 7: Conclusions This chapter summarizes the key contributions and findings of this dissertation, synthesizes insights across the various studies, acknowledges limitations, and discusses future directions for music representation learning across diverse traditions. It highlights how the dissertation has advanced dataset development, methodological innovation, model evaluation, adaptation techniques, and human-computational alignment, while identifying promising research avenues and potential applications that could further expand the technical capabilities of music information retrieval systems worldwide.

Chapter 2

Background

This chapter provides a comprehensive overview of the fundamental concepts, methods, and existing research relevant to representation learning across diverse musical traditions for music signal analysis. Beginning with a concise overview of theoretical frameworks for comparative music analysis, it progresses through detailed coverage of music signal processing fundamentals and deep learning approaches in Music Information Retrieval (MIR). The chapter then examines methodological challenges in MIR when analyzing diverse musical systems, world music datasets, computational ethnomusicology, transfer learning, few-shot learning, and foundation models in music. Additionally, it explores the intersection of computational approaches with human perception, particularly in the context of cross-cultural music similarity assessment. Throughout, we emphasize a data-driven approach to studying relationships between musical systems using end-to-end deep learning models with minimal inductive bias, avoiding explicit feature extraction to let the models discover relevant patterns directly from the data.

2.1 Theoretical Frameworks for Comparative Music Analysis

While our research takes a primarily technical approach to music representation learning across different traditions, it is informed by several theoretical perspectives from musicology, cognitive science, and cultural studies. This section briefly summarizes key theoretical frameworks that contextualize our computational approaches.

The field of comparative music analysis has historically considered both universalist perspectives, which seek common elements across musical traditions, and relativist views, which emphasize the distinctive characteristics of each tradition [10, 50]. Recent research suggests a nuanced view that recognizes both shared constraints and regional diversity in music [2], with certain statistical features appearing across traditions while manifesting in regionally specific ways.

From cognitive perspectives, music perception involves both domain-general processes shared across populations and specific knowledge acquired through exposure [3]. This suggests potential for both transferable and tradition-specific aspects of computational music representation. Information theory provides additional insights by framing music as a communication system balancing predictability and surprise [51], with statistical patterns that vary across traditions.

The question of musical universality versus cultural specificity has profound implications for computational approaches. While certain musical elements may transcend cultural boundaries, such as the recognition of emotional expressions or the use of discrete pitches, musical traditions have evolved with distinct characteristics that reflect their cultural contexts [2, 10]. The assumption of music as a "universal language" is challenged by research showing that cultural context influences

auditory perception and aesthetic appraisal, leading to diverse "listening frameworks" and "musical ontologies" that shape how different communities understand and categorize musical experience.

These theoretical frameworks inform our technical approach through several key principles. We adopt a computational perspective that uses data-driven methods while maintaining awareness of cultural context, rather than imposing predetermined analytical frameworks. Our research pursues adaptive representation approaches that can adjust to the distinctive characteristics of different traditions rather than applying a single universal framework. Given the data scarcity common in many musical traditions outside the commercial mainstream, we emphasize resource-conscious methods that work effectively with limited annotated data. Throughout this work, we maintain awareness of the technical challenges in computational music analysis, working to create more broadly applicable approaches that respect cultural diversity.

Most importantly, our approach emphasizes studying relationships between musical systems through end-to-end deep learning models. By minimizing inductive bias and avoiding explicit feature extraction for specific musical attributes (melody, rhythm, harmony), we allow the models to discover relevant patterns directly from the data. This approach reduces the risk of imposing assumptions from any particular musical system and enables more flexible representation learning across diverse traditions.

2.2 Music Signal Processing and Representation

2.2.1 Fundamentals of Audio Signal Processing

Music, at its most basic computational level, is represented as a digital audio signal, a one-dimensional sequence of amplitude values sampled at regular time intervals. Typical music recordings are sampled at rates of 44.1 kHz or 48 kHz, resulting in 44,100 or 48,000 amplitude values per second of audio [12]. This raw waveform representation contains all the acoustic information but presents challenges for direct analysis due to its high dimensionality and the complex encoding of musical information.

Various signal processing techniques transform these raw waveforms into more tractable representations. The Short-Time Fourier Transform (STFT) is fundamental, decomposing the signal into its frequency components over short time windows to create a time-frequency representation known as a spectrogram [52]:

$$X(n,k) = \sum_{m=0}^{N-1} x(m+n)w(m)e^{-j2\pi km/N}, \tag{2.1}$$

where $x(m)$ is the input signal, $w(m)$ is a window function of length N, n is the frame index, and k is the frequency bin index. The resulting spectrogram represents the magnitude of different frequency components over time, providing a two-dimensional visualization of the audio's spectral content.

For music analysis, the mel-spectrogram has become particularly important [53]. It applies mel-scale filterbanks to the STFT to compress the frequency axis in a way that approximates human auditory perception:

$$M(n,b) = \sum_{k=0}^{N/2} |X(n,k)|^2 H_b(k), \tag{2.2}$$

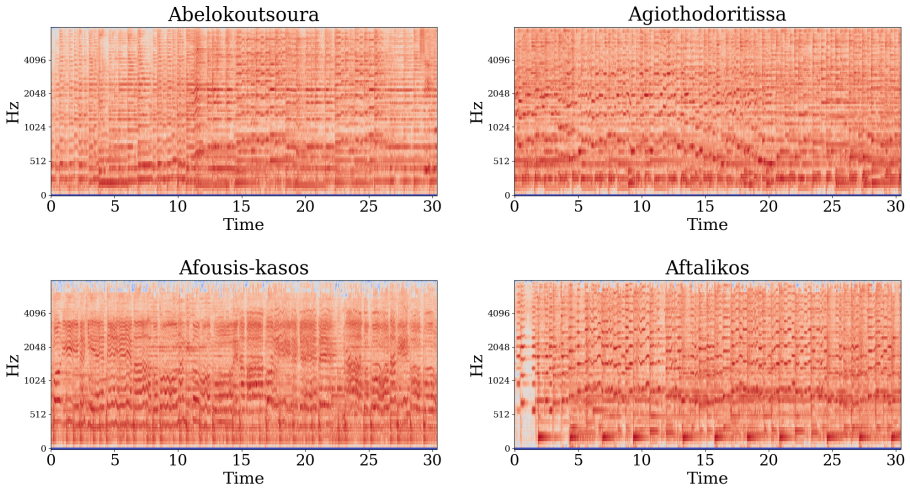

Figure 2.1. Mel-Spectrograms of Traditional Greek Music. Time-frequency representations of four songs from the Lyra dataset (Chapter 3).

where $H_b(k)$ represents the triangular mel-scale filters and b is the mel-band index. The mel scale converts frequencies f in Hz to mel units:

$$\text{mel}(f) = 2595 \log_{10}(1 + f/700) \tag{2.3}$$

This transformation emphasizes perceptually relevant frequency regions while reducing dimensionality, making it well-suited for machine learning approaches. Figure 2.1 shows four mel-spectrograms of songs from the Lyra dataset (see Chapter 3), revealing both temporal patterns (horizontal axis) and frequency content (vertical axis).

Other common time-frequency representations include:

Constant-Q Transform (CQT): Provides logarithmically spaced frequency bins aligned with musical scales [54]. The frequency resolution Δf_k at center frequency f_k maintains a constant ratio $Q = f_k/\Delta f_k$, matching the logarithmic nature of musical pitch organization:

$$X_{CQ}(k,n) = \sum_{m=n-\lfloor N_k/2 \rfloor}^{n+\lfloor N_k/2 \rfloor} x(m) \cdot w_k(m - n + \lfloor N_k/2 \rfloor) \cdot e^{-j2\pi Q m/N_k}, \tag{2.4}$$

where N_k is the variable window length for each frequency bin k.

Chromagrams: Project the spectrum onto 12 pitch classes representing the chromatic scale, capturing harmonic and tonal content while discarding octave information [55]:

$$C(n,p) = \sum_{q=0}^{Q-1} |X(n, p + 12q)|^2, \tag{2.5}$$

where $p \in \{0, 1, ..., 11\}$ represents the 12 pitch classes and Q is the number of octaves.

Log-Mel Spectrograms: Apply logarithmic compression to mel-spectrograms, further aligning with human perception of loudness:

$$L(n,b) = \log(1 + \alpha \cdot M(n,b)), \tag{2.6}$$

where α is a scaling factor. This representation has become particularly important for deep learning approaches.

These time-frequency representations form the foundation for both traditional feature extraction and modern deep learning approaches. They encode different aspects of the signal, with varying trade-offs between temporal and frequency resolution, and different alignments with musical organization principles.

2.2.2 Traditional Musical Feature Extraction

Building upon basic time-frequency representations, various higher-level features have been developed to capture specific musical characteristics. These features, designed based on music theory and perceptual principles, have been widely used in traditional MIR systems:

Melodic Features: Melody analysis focuses on extracting the fundamental frequency (F0) as the primary carrier of melodic information. The PYIN algorithm [56] provides robust F0 extraction from polyphonic audio by treating the fundamental frequency as the dominant melodic skeleton. Following F0 detection, pitch classes are recognized to enable cross-cultural melodic analysis through dual-resolution representations.

Given the extracted fundamental frequency f in Hz, MIDI-like pitch numbers are computed as:

$$m = 12 \log_2 \left(\frac{f}{440} \right) + 69. \tag{2.7}$$

Pitch classes are then calculated as $pc = \lfloor m \bmod 12 \rfloor$, typically providing 12 bins for traditional Western analysis.

Melodic intervals between consecutive pitch classes can, in turn, be computed. These intervals, along with pitch class distributions and contour analysis, characterize the music's melodic movement.

Rhythmic Features: Tempo estimation, beat tracking, and onset detection algorithms extract information about temporal organization [57]. Onset detection functions identify points where new events begin in the signal:

$$ODF(n) = \sum_{k=1}^{K} H(|X(n,k)| - |X(n-1,k)|), \tag{2.8}$$

where $H(x) = \frac{x+|x|}{2}$ is a half-wave rectifier function. Tempo is typically estimated through periodicity analysis of these onset functions, while beat tracking aligns a regular grid to these onsets using techniques like dynamic programming or hidden Markov models.

Harmonic Features: Harmony can be represented through pitch class profiles, chord recognition, and key estimation [58]. Harmonic Pitch Class Profiles (HPCP) enhance basic chromagrams by weighting frequency bins based on their harmonic relationship to the fundamental:

$$HPCP(n,p) = \sum_{h=1}^{H} w_h \cdot C(n, (p \cdot h) \bmod 12), \tag{2.9}$$

where h is the harmonic index and w_h is a weighting function.

Timbral Features: Mel-Frequency Cepstral Coefficients (MFCCs) have been fundamental for capturing timbral characteristics [59]. Derived by applying the Discrete Cosine Transform to the logarithm of the mel-spectrogram, MFCCs represent the spectral envelope shape:

$$MFCC(n, c) = \sum_{b=0}^{B-1} \log(M(n, b)) \cos(c(b + 0.5)\pi/B), \tag{2.10}$$

where c is the cepstral coefficient index and B is the number of mel bands. Additional spectral features include spectral centroid (the "center of mass" of the spectrum), flux (the rate of spectral change), and rolloff (the frequency below which a specified percentage of spectral energy is contained) [60].

Structural Features: Segmentation algorithms identify structural boundaries and repetitions, capturing form and arrangement [61]. These often employ self-similarity matrices S computed from frame-level features:

$$S(i, j) = \text{sim}(v_i, v_j), \tag{2.11}$$

where v_i and v_j are feature vectors at frames i and j, and sim is a similarity measure like cosine similarity.

While these hand-crafted features have proven effective for many MIR tasks, they often embed musical assumptions from European classical and popular music traditions. For instance, chromagram features implicitly assume 12-tone equal temperament, making them less suited for traditions with different tuning systems or microtonal inflections. Similarly, conventional beat tracking algorithms often assume metrical structures common in European and North American music, facing challenges with complex rhythmic cycles or asymmetrical meters found in other traditions.

2.2.3 Challenges in Analyzing Diverse Musical Systems

Computational analysis presents several challenges when applied to diverse musical systems due to fundamental differences in musical organization:

- **Tonal Systems**: Many musical traditions employ microtonal intervals, non-equal temperament, and modal systems not well-captured by conventional features [13]. For example, Turkish makam music uses intervals as small as a comma (approximately 22.6 cents, compared to 100 cents in the semitone), creating 53 divisions of the octave instead of 12 [20]. Similarly, Indian classical music employs 22 microtonal divisions (shruti) and complex melodic ornamentation (gamaka) that confound equal-tempered representations [21].

- **Rhythmic Complexity**: Asymmetrical patterns, complex cycles, and flexible timing present challenges for conventional rhythm analysis [62]. Greek traditional music often uses meters like 7/8 (grouped as 3+2+2) or 9/8 (grouped as 2+2+2+3), while Indian classical music employs complex tala cycles with internal hierarchical organizations. These structures present difficulties for algorithms expecting regular beat divisions or simple duple/triple meters.

- **Timbre and Instrumentation**: Traditional instruments from various regions produce timbral qualities not well-represented by features optimized for modern orchestral or electronic instruments [33]. The Greek lyra, Turkish ney, and Indian sitar each produce distinctive spectral patterns with unique attack characteristics, sustained resonances, and harmonic structures that may be mischaracterized by standard timbral features.

- **Performance Practices**: Improvisation, ornamentation, and tradition-specific performance techniques may not be adequately captured by standard feature extraction [33]. Indian classical music employs complex ornamentation like meend (gliding between notes), Turkish makam features distinctive microtonal inflections, and Greek traditional music includes taximi (non-metrical improvisation) that challenge conventional analysis.

These challenges have motivated two different approaches. One direction involves developing specialized representations for specific traditions, such as a MIDI-based representation for Turkish makam [18], specialized models for makam music lyrics-to-audio alignment [63], and methods for identifying asymmetric rhythms in Greek music [17]. While effective for specific traditions, these specialized approaches lack scalability across multiple musical systems.

The alternative direction, and the one pursued in this dissertation, involves end-to-end learning approaches that minimize inductive bias. Instead of designing specialized features that might embed particular musical assumptions, we employ deep learning techniques that can learn appropriate representations directly from data. This approach allows the models to discover relevant patterns without imposing predefined notions of what musical characteristics are important, potentially providing more flexible and adaptable representations.

2.2.4 Representation Learning Approaches: Traditional Features vs. End-to-End Methods

The limitations of many traditional features for analyzing diverse musical systems motivate a systematic investigation of representation learning approaches. This dissertation employs primarily end-to-end models that learn directly from minimally processed audio representations, while also evaluating the cross-cultural effectiveness of traditional signal processing features (see Chapter 6).

As illustrated in Figure 2.2, these two approaches differ fundamentally in how they process musical information. Traditional feature extraction relies on explicit feature engineering where musical theory assumptions are embedded at multiple stages, from the choice of spectral features to the design of higher, level descriptors like chroma vectors or beat tracking algorithms. These handcrafted features, while interpretable and grounded in music theory, impose fixed representations that may not adequately capture the characteristics of diverse musical traditions. In contrast, end-to-end learning employs data-driven feature discovery, allowing models to learn representations directly from minimally processed audio without imposing predetermined musical categories.

This end-to-end approach offers several key advantages for analyzing diverse musical systems. First, it provides reduced methodological bias by avoiding feature engineering based on particular music theories, thereby reducing the risk of imposing inappropriate analytical frameworks on various traditions while allowing models to potentially learn tradition-specific patterns directly from the data. Second, deep learning models can develop adaptive representations that adjust to the distinctive characteristics of different musical traditions without requiring explicit modeling of those differences. Third, with reduced inductive bias, models might discover unexpected patterns and similarities or differences between musical systems that wouldn't be captured by predefined feature sets.

Furthermore, end-to-end approaches offer scalability across traditions, potentially generalizing across diverse musical traditions without requiring specialized knowledge about each tradition's unique characteristics. However, our systematic evaluation reveals important nuances: some traditional features, particularly melody-based descriptors, retain significant value across cultural

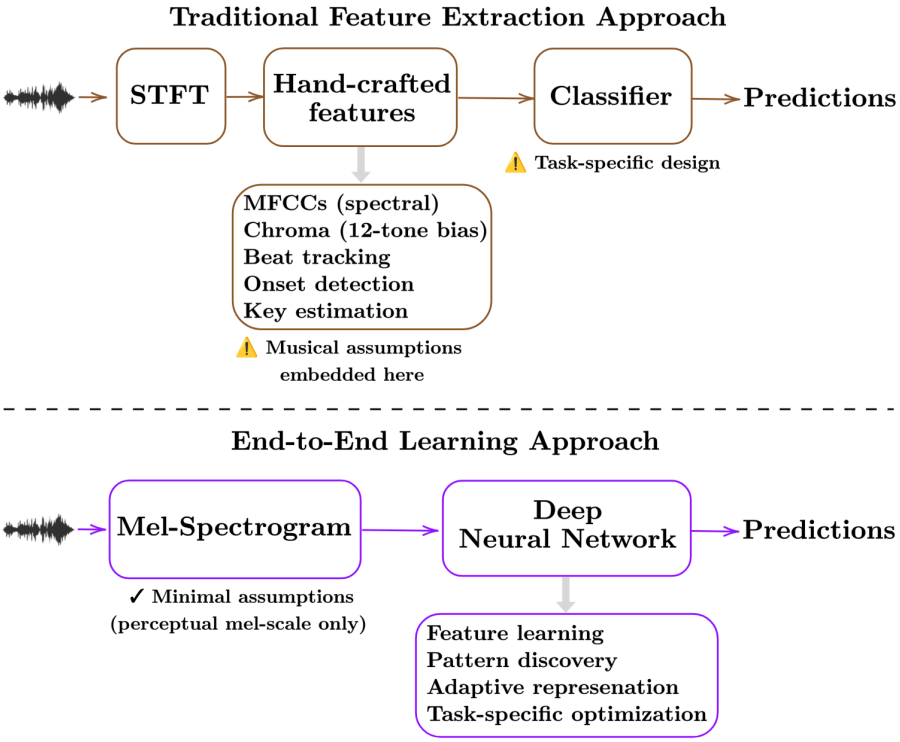

Figure 2.2. Traditional Feature Extraction vs. End-to-End Learning Approaches. Comparison of computational pathways showing where musical assumptions and cultural biases can be introduced in traditional approaches versus the more culturally neutral end-to-end methodology employed in this dissertation.

contexts and can complement learned representations, challenging the assumption that newer approaches are universally superior.

Focusing on the end-to-end learning, our approach uses minimally processed audio representations, primarily log-mel spectrograms, as input to deep neural networks. While these representations still embed some perceptual assumptions (the mel scale approximates human frequency perception), they maintain much of the original signal information without imposing explicit musical categories. The deep learning models can then learn task-relevant patterns directly from these representations through supervised, self-supervised, or transfer learning objectives.

However, by adopting a comprehensive evaluation framework that includes human perception studies, our research demonstrates that the most effective approach combines the flexibility of end-to-end learning with the interpretability and surprising cross-cultural robustness of select traditional features. This integrated perspective validates which computational approaches align best with cultural understanding.

2.3 Deep Learning in Music Information Retrieval

2.3.1 Evolution of Deep Learning in MIR

Deep learning approaches have revolutionized MIR by enabling the automatic learning of representations directly from data, reducing the reliance on hand-crafted features. This section traces the evolution of deep learning in MIR, from early applications to current state-of-the-art approaches.

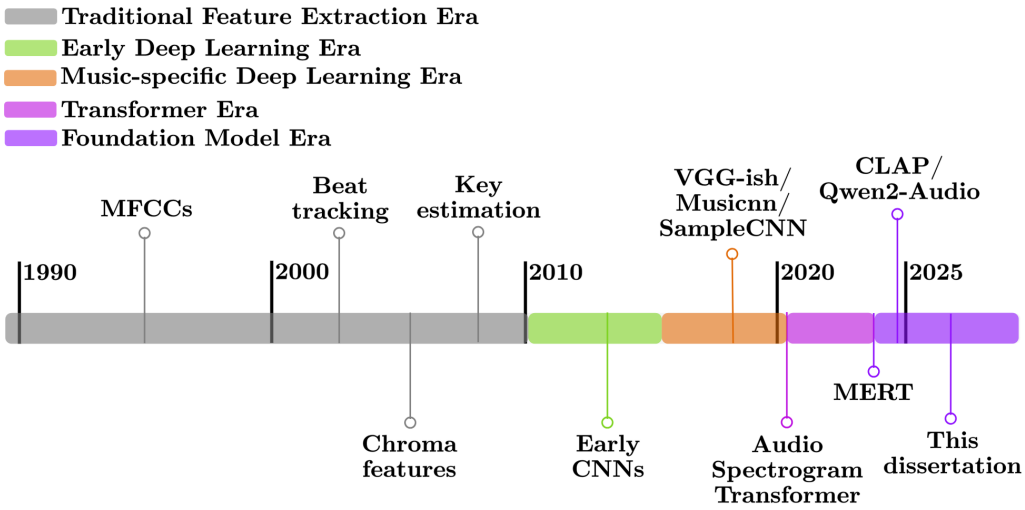

Figure 2.3. Methodological Evolution in Music Information Retrieval. Timeline showing the progression from traditional hand-crafted features to foundation models, highlighting the major paradigm shifts in computational music analysis.

As illustrated in Figure 2.3, the field of MIR has undergone several distinct evolutionary phases, each characterized by different computational paradigms and methodological approaches. The traditional feature extraction era, spanning from the 1990s through the early 2010s, was dominated by hand-crafted features such as MFCCs [64], chroma vectors [65], and various temporal and spectral descriptors. Many of these features were initially developed in the 1980s or earlier for speech and audio processing applications, but were subsequently adopted and adapted for music analysis from the 1990s onwards. While these approaches provided interpretable and musically grounded representations, they often embedded specific musical assumptions that limited their effectiveness across diverse musical traditions.

Early applications of deep learning in MIR focused on adapting architectures from other domains, particularly computer vision. Convolutional Neural Networks (CNNs), originally developed for image analysis, were adapted to process spectrograms as 2D images [66, 67]. These approaches treated time and frequency dimensions analogously to the spatial dimensions of images, enabling the extraction of patterns across both dimensions.

As deep learning in MIR matured, researchers developed architectures specifically designed for music signals. These approaches incorporated domain knowledge about music structure and perception while leveraging the representational power of deep neural networks. For instance, the authors of [68] introduced Musicnn, which uses horizontal and vertical convolutional filters to separately capture temporal and timbral features before combining them for music classification.

Besides the models that process time-frequency input data, end-to-end approaches that learn directly from raw audio waveforms emerged, eliminating the need for pre-defined representations. Models like SampleCNN [69] and TCNN [70] operate directly on raw audio samples, learning appropriate feature hierarchies from the data itself. Additionally, the success of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks in sequence modeling led to their application in MIR tasks requiring temporal understanding, such as structural segmentation and beat tracking [71].

More recently, attention-based models, particularly Transformers, have been applied to music analysis [39]. The Audio Spectrogram Transformer (AST) adapts the Vision Transformer architec-

ture to audio spectrograms, treating them as sequences of patches. This approach has demonstrated state-of-the-art performance on various audio classification tasks, including music auto-tagging.

The current foundation model era represents the most recent paradigm shift, with models like MERT [40] and CLAP [41] demonstrating unprecedented capabilities across diverse MIR tasks. These models leverage large-scale self-supervised pre-training to learn general-purpose musical representations that can be adapted to various downstream applications.

This evolutionary progression reflects a shift from manually engineered approaches toward data-driven methods, though our research demonstrates that hybrid approaches combining traditional and learned features can be most effective in some cross-cultural applications. This transition directly motivates the multicultural representation learning approaches explored in this dissertation.

2.3.2 Deep Audio Embeddings

Deep audio embeddings, dense vector representations learned by deep neural networks, have become fundamental tools in MIR. These embeddings capture musically relevant information in a compact form that can be used for various downstream tasks, including similarity search, clustering, and classification.

Several approaches exist for learning deep audio embeddings. *Supervised learning* involves training networks on labeled data for tasks like genre classification or auto-tagging, then using intermediate layer activations as embeddings [72]. *Self-supervised* learning takes a different approach by learning representations without explicit labels through solving pretext tasks like reconstructing corrupted inputs, predicting future frames, or contrastive prediction [73]. *Transfer learning* adapts embeddings learned on one dataset or task to new domains, leveraging knowledge from resource-rich areas to enhance performance in limited-data scenarios [74].

Deep audio embeddings offer several key advantages over traditional hand-crafted features. They can capture complex patterns and hierarchical structures in music that might be difficult to define explicitly, while being learned directly from data, potentially reducing methodological biases embedded in hand-crafted features. Additionally, they can be fine-tuned or adapted for specific tasks and musical traditions, and they provide a unified representation that can support multiple downstream tasks.

In Sections 4.2-4.4, we explore the transferability of deep audio embeddings between musical traditions, investigating how well embeddings learned from one musical tradition can represent and analyze music from different traditions. This work provides insights into the similarities and differences between musical systems from a computational perspective and informs the development of more broadly applicable embedding models.

2.4 Methodological Limitations in Current MIR Systems

The field of MIR has made remarkable progress in developing computational approaches to music analysis, but it also faces important methodological limitations when analyzing diverse musical systems. These limitations manifest at multiple levels, from low-level feature extraction to high-level task formulation and evaluation. Music representation approaches, both traditional feature extraction and modern representation learning, embed specific musical assumptions that can limit their effectiveness across diverse traditions. For example, chroma features implicitly assume 12-tone equal temperament, making them less suited for traditions with different tuning systems [75], while rhythmic features often assume certain metrical hierarchies, struggling with complex cycles of traditions like Carnatic music [76]. Additionally, learned representations from

deep models inherit biases from their training data, which predominantly consists of Western commercial music [13, 14].

Task formulation in MIR often reflects specific musical priorities that may not generalize across traditions. Genre classification typically employs commercial music categories, while chord recognition assumes harmonic principles common in European traditions that may not apply to modal or heterophonic traditions [15]. Similarly, evaluation metrics may not capture all musically relevant aspects of model performance, potentially rewarding systems that exploit dataset artifacts rather than meaningful musical understanding [77]. The predominance of certain musical traditions in research datasets [14] can create a "self-reinforcing cycle" in which established datasets lead to specialized algorithms, which in turn encourage more similar data collection.

These challenges are illustrated by specific examples from diverse musical traditions. Indian classical music employs a complex system of ragas that define specific melodic movements, emphasized notes, and expressive associations that differ from concepts of scales or modes in other traditions. Standard tonal analysis tools face challenges because pitch tracking algorithms optimized for certain musical contexts often perform less effectively with the continuous pitch movements (gamaka) central to Indian classical music, and equal-tempered pitch representations cannot adequately capture the subtle microtonal intervals used in raga performance [21]. Similarly, traditional Eastern Mediterranean music employs complex rhythmic cycles with distinctive structural characteristics that confound conventional beat tracking algorithms. These patterns often feature asymmetrical beat groupings (e.g., 9/8 grouped as 2+2+2+3) that differ from the patterns expected by algorithms designed for regular beat divisions [17]. Early research demonstrated this challenge quantitatively, with algorithms achieving less than 80% performance on non-Western datasets compared to more than 90% for popular music with regular metrical structures [62].

As demonstrated in Section 5.4, evaluations of foundation models show similar limitations despite their more flexible representation learning capabilities. These challenges reflect not inherent limitations of computational approaches but rather the methodological assumptions embedded in both explicit feature design and implicit biases in the training data used for representation learning. Such limitations highlight the need for more culturally aware computational approaches that can adapt to the distinctive characteristics of diverse musical traditions.

2.5 World Music Datasets and Computational Ethnomusicology

2.5.1 Computational Ethnomusicology

Computational ethnomusicology represents the intersection of ethnomusicology, the study of music in its cultural context, and computational methods [32]. This emerging field applies digital tools and computational analysis to study diverse musical traditions, complementing traditional ethnomusicological approaches with data-driven insights.

The field pursues several interconnected objectives that span preservation, analysis, and tool development. A primary focus involves preservation and documentation through digitizing, organizing, and analyzing recordings of traditional music, particularly from endangered musical traditions [33]. Computational ethnomusicology also enables comparative analysis by identifying similarities and differences between musical traditions through computational comparison of acoustic features, structures, and patterns [78, 79]. The field's analytical capabilities extend to pattern discovery, uncovering structures within specific musical traditions that might not be im-

mediately apparent through traditional analysis methods [80]. Additionally, researchers in this domain focus on tool development, creating specialized computational tools that can effectively capture the unique characteristics of different musical traditions [81].

Despite its promise, computational ethnomusicology faces several significant challenges. These include the need for appropriate computational representations that can accommodate diverse musical systems, the scarcity of structured datasets for many traditions, and the critical importance of accounting for tradition-specific musical characteristics that may not align with conventional computational approaches. These challenges have motivated the development of specialized approaches and datasets for various musical traditions, as well as the exploration of more culturally adaptive computational methods.

2.5.2 World Music Datasets

Several datasets have been developed to support computational analysis of diverse musical traditions. These datasets vary in size, scope, and annotation depth, reflecting the diversity of the traditions they represent and the specific research questions they aim to address. Table 2.1 presents representative examples of world music datasets, illustrating their characteristics and availability.

These datasets represent important contributions to computational analysis of diverse musical traditions, enabling both culture-specific studies and cross-cultural comparisons. The CompMusic project stands out as a particularly comprehensive effort, providing structured access to multiple traditions through the Dunya platform, while more recent datasets like Erkomaishvili and Lyra focus on specific traditions with detailed musicological annotations. However, many musical traditions worldwide remain underrepresented in available datasets, highlighting the ongoing need for continued efforts in data collection and annotation, particularly for musical cultures from Africa, Southeast Asia, and indigenous traditions globally.

The CompMusic Project

The CompMusic (Computational Models for World Music) project represents the most significant initiative in computational ethnomusicology to date, focusing on developing computational approaches for analyzing five music traditions: Hindustani and Carnatic classical music from India, Turkish makam music, Beijing Opera, and Arab-Andalusian music [33].

The project made several groundbreaking contributions that have shaped the field. It developed comprehensive tradition-specific corpora containing audio recordings, scores, and contextual information for each tradition [19, 23, 84–86], providing researchers with unprecedented access to structured collections of world music. The project also created specialized computational tools and algorithms for analyzing tradition-specific aspects such as raga and tala in Indian music, makam and usul in Turkish music, enabling culturally appropriate computational analysis [20, 21]. Additionally, CompMusic developed sophisticated knowledge representations and ontologies that capture the conceptual structures of each tradition [87], while building user-friendly applications like Dunya that provide accessible interfaces to the collections and computational tools [88].

The CompMusic approach emphasizes tradition-centered design, developing computational methods that respect and capture the unique characteristics of each tradition rather than imposing external frameworks. This methodology has significantly influenced subsequent research in computational ethnomusicology and provided valuable resources for comparative music analysis, establishing a model for culturally sensitive computational musicology.

Dataset	Musical Tradition	Size	Key Annotations	Availability	Year
Dutch Folk Songs Database [34]	Dutch folk	∼200,000 melodies	Melody, lyrics, metadata	Public	2019
CompMusic Corpora [33]	Hindustani, Carnatic, Turkish makam, Beijing Opera, Arab-Andalusian	Varies by tradition	Raga, tala, makam, usul, instruments	Public (Dunya)	2014
COFLA Dataset [35]	Flamenco	∼95 hours	Hierarchical style labels, artists, transcriptions	Public	2016
Erkomaishvili Dataset [36]	Georgian traditional	∼100 record-ings, 7 hours	Three-voice polyphony, lyrics, transcriptions	Public	2020
ChMusic [24]	Chinese traditional	55 recordings	Instruments, artists	Public	2021
Greek Audio Dataset (GAD) [82]	Greek popu-lar/traditional	1,000 tracks, ∼8 hours	Genre, mood, lyrics	Public	2014
Greek Music Dataset (GMD) [83]	Greek popu-lar/traditional	1,400 tracks, ∼12 hours	Extended GAD annotations	Public	2015
Lyra Dataset [45]	Greek traditional/folk	1,570 pieces, 80 hours	Instruments, geography, genre	Public	2022

Table 2.1. Representative World Music Datasets. Selected datasets for computational analysis of diverse musical traditions, highlighting the variety in scope, size, and annotation approaches. Size information is approximate where not precisely reported in original sources.

2.6 Transfer Learning and Few-Shot Learning in MIR

2.6.1 Transfer Learning Fundamentals

Transfer learning aims to improve model performance on a target task by leveraging knowledge from a related source task [89]. This approach is particularly valuable when the target task has limited data or computational resources, allowing it to benefit from knowledge gained in data-rich domains.

In the context of deep learning, transfer learning typically involves pre-training a model on a source task with abundant data, then adapting it to a target task through fine-tuning, updating some or all of the model parameters using the target data. The underlying assumption is that the representations learned for the source task capture generalizable patterns that are relevant to the target task.

Several interconnected factors influence the effectiveness of transfer learning. Task similarity, or the degree of relatedness between source and target tasks, fundamentally affects how well knowledge

transfers [90]. Domain shift, referring to differences in data distributions between source and target domains, can limit transfer effectiveness even when tasks are conceptually similar [91]. The choice of model architecture also plays a crucial role, as some architectures may facilitate transfer better than others depending on how they structure and abstract knowledge [92]. Finally, fine-tuning strategy, decisions about which layers to freeze or fine-tune, can significantly impact transfer performance, with different strategies being optimal for different degrees of domain similarity [93].

In MIR, transfer learning has been applied to various tasks, including genre classification [72] and music recommendation [74]. These applications typically transfer knowledge from large-scale datasets like MagnaTagATune or Million Song Dataset to more specialized or limited-data scenarios.

2.6.2 Transfer Learning Across Musical Traditions

Transfer learning can be applied to bridge different musical traditions, leveraging knowledge from one musical domain to enhance performance in another. This approach aligns with our goal of studying relationships between musical systems using data-driven methods, as it enables quantitative assessment of knowledge transferability without requiring explicit modeling of tradition-specific features.

The transfer learning approach offers several interconnected benefits for analyzing diverse musical traditions. Data efficiency represents a primary advantage, as many traditional musical systems have limited annotated datasets that make direct training challenging, while transfer learning allows these traditions to benefit from models pre-trained on larger datasets. Through feature discovery, models pre-trained on one tradition may identify features that are relevant to other traditions, potentially revealing shared musical elements without explicitly engineering those features. Adaptation efficiency is another key benefit, as adapting existing models through transfer learning is generally more computationally efficient than training from scratch, making it practical to develop specialized models for diverse traditions. Perhaps most importantly for comparative musicology, the effectiveness of knowledge transfer between specific traditions can provide a quantitative measure of their computational similarity, potentially revealing relationships that might not be apparent through traditional musicological analysis.

This data-driven approach to comparative music analysis avoids the need to explicitly model tradition-specific features, instead allowing patterns of knowledge transfer to emerge naturally from the data. This methodology reduces the risk of imposing inappropriate analytical frameworks while still providing insights into relationships between different musical systems. By examining which source domains provide the most effective knowledge transfer for each target domain, we can identify potential relationships between musical traditions from a computational perspective. For instance, stronger transfer between geographically or historically connected traditions might reflect shared musical elements, while asymmetric transfer patterns might reveal directional influences or overlapping musical concepts.

In Sections 4.2–4.5, we investigate transfer learning between different musical traditions, including both Western and non-Western regions, exploring the extent to which musical knowledge can be transferred effectively across different musical systems.

2.6.3 Few-Shot Learning Approaches

Few-shot learning addresses the challenge of learning from limited examples, aiming to generalize to new classes based on only a few available instances [94]. This capability is particularly

valuable for world music research, where annotated examples for many traditions or specific musical attributes may be scarce.

Several approaches to few-shot learning have been developed:

Metric-Based Methods: These approaches learn a similarity metric that can compare new examples to a small set of labeled instances. Prototypical Networks [95] represent each class by a prototype computed as the mean of its examples in an embedding space, then classify new instances based on their distance to these prototypes:

$$p(y = k|\mathbf{x}) = \frac{\exp(-d(f_\phi(\mathbf{x}), \mathbf{c}_k))}{\sum_{k'} \exp(-d(f_\phi(\mathbf{x}), \mathbf{c}_{k'}))}, \tag{2.12}$$

where f_ϕ is an embedding function, $\mathbf{c}_k$ is the prototype for class k, and d is a distance function.

Model-Based Methods: These approaches use memory mechanisms or recurrent architectures to rapidly adapt to new tasks. Memory-Augmented Neural Networks [96] incorporate external memory that can be quickly updated with new information.

Optimization-Based Methods: These approaches learn an initialization that can be rapidly fine-tuned to new tasks with minimal data. Model-Agnostic Meta-Learning (MAML) [97] optimizes for quick adaptation by explicitly training for few-shot fine-tuning performance:

$$\min_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)}), \tag{2.13}$$

where $\mathcal{T}_i$ are tasks sampled from a distribution $p(\mathcal{T})$, and $\mathcal{L}$ is a loss function.

Few-shot learning has been applied to various MIR tasks, including drum transcription [98], source separation [99], and single instrument recognition [100]. These applications demonstrate the potential of few-shot learning to address data scarcity challenges in specialized music analysis tasks.

Few-shot learning is particularly relevant for analyzing diverse musical traditions because it addresses the data scarcity common in many traditional musical systems. It enables the inclusion of under-represented tags and musical elements in computational models, even when only a few examples are available. This capability aligns with our goal of developing more adaptable approaches to music representation learning that can work effectively across diverse traditions with varying data availability.

2.6.4 Multi-Label Few-Shot Learning

While conventional few-shot learning frameworks focus on multi-class classification (where each instance belongs to exactly one class), many music analysis tasks, including auto-tagging, require multi-label classification (where each instance can be associated with multiple labels simultaneously). This introduces additional challenges for few-shot learning.

Several approaches have been proposed to address multi-label few-shot learning:

Sample Synthesis: The LASO method [101] synthesizes samples with multiple labels by combining pairs of examples in the feature space:

$$f_{mix}(x_i, x_j) = \alpha f(x_i) + (1 - \alpha) f(x_j), \tag{2.14}$$

where f is an embedding function and α is a mixing coefficient.

Label Count Prediction: The approach introduced in [102] predicts the number of labels assigned to an item, enabling multi-label predictions.

Attention Mechanisms: Attention-based approaches [103] integrate label-specific attention to handle multiple labels per instance.

Hierarchical Label Structures: Some methods [104] leverage taxonomic hierarchies between tags to improve few-shot performance.

Binary Reformulation: The "One-vs.-Rest" approach [105] reformulates multi-label problems into multiple binary classification tasks.

These methods address various aspects of multi-label few-shot learning but often introduce additional complexity during training. The LC-Protonets approach, see Section 4.6, takes a different approach by creating prototypes for label combinations derived from the power set of labels present in support samples, providing a novel extension of Prototypical Networks to the multi-label setting.

Multi-label few-shot learning is particularly relevant for world music research, where music pieces often have multiple overlapping tags related to instrumentation, region, genre, and other attributes. The ability to learn from limited examples with multiple labels enables more comprehensive modeling of diverse musical traditions, even with the limited annotated data available for many traditional musical contexts.

2.7 Foundation Models in Music

2.7.1 The Rise of Foundation Models

Foundation models represent a paradigm shift in artificial intelligence, featuring large-scale models pre-trained on vast datasets that can be adapted to various downstream tasks [106]. Originally pioneered in natural language processing with models like BERT (Bidirectional Encoder Representations from Transformers) [107] and GPT (Generative Pre-trained Transformer) [108], the foundation model approach has expanded to other domains, including computer vision, speech processing, and music.

Foundation models are characterized by several key features that distinguish them from traditional deep learning approaches. *Scale* represents a fundamental aspect, as these models typically feature large architectures with millions or billions of parameters, trained on massive datasets that were previously impractical to utilize. Most foundation models employ *self-supervised pre-training* objectives that don't require explicit labels, enabling them to leverage vast amounts of unlabeled data and learn rich representations from the underlying structure of the data itself. These models are specifically designed for *transfer learning*, with the ability to transfer knowledge to diverse downstream tasks through fine-tuning or prompting, making them remarkably versatile across applications and domains. Perhaps most intriguingly, foundation models often exhibit *emergent capabilities*, sophisticated behaviors that were not explicitly designed but emerge from the combination of scale and architecture, such as few-shot learning and cross-domain transfer abilities.

Model	Architecture	Parameters	Pre-training Approach	Training Data	Year
JukeMIR [109]	Transformer	5B	Derived from Jukebox generative model [110]	1.2M songs from many genres and artists	2021
MULE [111]	Transformer	100M	Self-supervised masked prediction	1.7M tracks from a private catalog	2022
Music2Vec [112]	Transformer	90M	Masked prediction with student-teacher	1k hours of collected music audio files from the Internet	2022
MusicFM [113]	Transformer	330M / 660M	Masked token modeling	160k hours of in-house music data / FMA dataset [114]	2023
MERT [40]	BERT-style Transformer	95M / 330M	Masked acoustic modeling	∼1,000 hours music	2023
CLAP [41]	Unified Transformer	194M	Contrastive audio-text learning	Audio-text dataset with total duration of ∼4,3k hours	2024
Qwen2-Audio [42, 115]	Unified Transformer	8.4B	Multi-task training framework for multi-modal audio understanding	Large-scale multi-dataset co-training	2024

Table 2.2. Representative Music Foundation Models. Key foundation models developed for music understanding, showing the evolution of architectural approaches and pre-training strategies in the field.

Foundation models for music have emerged in recent years, applying similar principles to audio understanding and music analysis. These models leverage large-scale self-supervised or contrastive learning on extensive audio datasets, enabling them to capture rich musical features applicable across diverse tasks.

2.7.2 Music Foundation Models

Several foundation models have been developed specifically for music understanding, employing diverse architectural approaches and pre-training strategies. Table 2.2 presents key music foundation models, highlighting their architectural characteristics and training approaches.

These models demonstrate the evolution of foundation model approaches in music AI, progressing from early explorations like JukeMIR [109] to large-scale unified models like Qwen-Audio [42]. Most employ Transformer-based architectures with various pre-training objectives including masked prediction, contrastive learning, and multi-modal training. They are typically pre-trained on large collections of music recordings, often focusing on popular and classical music from com-

mercial sources, though some recent models incorporate more diverse training data.

The MERT model [40], which is central to this dissertation's foundation model work, exemplifies the masked acoustic modeling approach. It uses a BERT-style architecture [107] with 12 Transformer encoder layers and approximately 95 million parameters (with a larger 330M variant also available). Pre-training involves the employment of masked acoustic modeling, using an acoustic and a musical teacher, encouraging the model to learn coherent representations of musical structure. The model has been pre-trained on approximately 1,000 hours of music, primarily from commercially available genres.

2.7.3 Evaluation of Foundation Models

Foundation models are typically evaluated through a combination of approaches that assess different aspects of their learned representations and adaptation capabilities:

Probing Freezing the pre-trained model and training only a classification layer on top, assessing how well the learned representations capture relevant features without adaptation:

$$\hat{y} = \text{softmax}(W \cdot f_{\text{frozen}}(x) + b), \tag{2.15}$$

where f_{frozen} is the frozen pre-trained model and W, b are the trainable linear layer parameters.

Fine-Tuning Adapting some or all of the model parameters to specific downstream tasks, evaluating the model's ability to transfer knowledge through parameter updates:

$$\theta_{\text{fine-tuned}} = \theta_{\text{pre-trained}} - \eta \nabla_\theta \mathcal{L}(\theta, \mathcal{D}_{\text{downstream}}), \tag{2.16}$$

where θ represents the model parameters, η is the learning rate, and $\mathcal{L}$ is the loss function on the downstream dataset $\mathcal{D}_{\text{downstream}}$.

Few-Shot Evaluation Assessing performance with limited labeled examples, testing the model's ability to generalize from minimal task-specific data.

Zero-Shot Evaluation Evaluating performance without any task-specific training, typically through prompting or other mechanisms that leverage the model's pre-trained knowledge.

Benchmarks for evaluating music foundation models include MARBLE [116], `mir_eval` [117], and `mir_ref` [118]. However, these benchmarks predominantly feature Western music, raising important questions about how well foundation models generalize to diverse musical traditions, a gap that this dissertation addresses through comprehensive cross-cultural evaluation (see Section 5.2) and by measuring their alignment with human perception of musical similarity (see Section 6.6).

2.7.4 Adaptation Challenges and Approaches

Despite their impressive performance on standard benchmarks, music foundation models face several interconnected challenges when applied to diverse musical traditions. Training data limitations represent a fundamental issue, as most foundation models are trained predominantly on commercial music from major markets, potentially limiting their ability to represent characteristics of other musical traditions [13]. This limitation is compounded by the presence of tradition-specific

elements, musical characteristics like melodic structures, modal systems, and rhythmic patterns in traditions such as Turkish makam, Indian classical, and Greek folk music may not be adequately captured by models trained primarily on commercial music [21]. Evaluation gaps further complicate the assessment of these limitations, as the scarcity of benchmarks for traditional music from various regions makes it difficult to systematically assess model capabilities across diverse musical systems [15]. Additionally, tokenizer constraints present technical challenges, as audio tokenizers trained on certain musical traditions may not optimally encode the acoustic characteristics of traditional instruments and performance practices from other regions.

Addressing these challenges requires sophisticated approaches for adapting foundation models to better represent diverse musical traditions. Fine-tuning represents the most straightforward adaptation strategy, involving supervised fine-tuning of pre-trained models on tradition-specific data (see Section 5.2). Continual pre-training (Section 5.5) offers a more comprehensive approach by further pre-training foundation models on data from diverse musical traditions, incrementally adapting the representations while avoiding catastrophic forgetting [119]:

$$\mathcal{L}_{\text{CPT}} = \mathcal{L}_{\text{MLM}}(\theta, \mathcal{D}_{\text{tradition}}) + \lambda \cdot \mathcal{R}(\theta, \theta_{\text{pre-trained}}), \tag{2.17}$$

where $\mathcal{L}_{\text{MLM}}$ is the masked language modeling loss on the tradition-specific dataset $\mathcal{D}_{\text{tradition}}$, $\mathcal{R}$ is a regularization term to prevent catastrophic forgetting, and λ is a weighting coefficient. Task arithmetic provides an alternative approach by merging tradition-specific adaptations in weight space to create unified models that represent multiple musical traditions [120]:

$$\theta_{\text{combined}} = \theta_{\text{pre-trained}} + \sum_i \alpha_i \cdot (\theta_i - \theta_{\text{pre-trained}}), \tag{2.18}$$

where θ_i represents the parameters of a model adapted to tradition i, and α_i are weighting coefficients. Finally, diverse pre-training represents a proactive approach, involving the development of new foundation models pre-trained from the outset on more diverse collections of musical data.

These adaptation strategies align with our data-driven approach to studying relationships between musical systems, allowing us to quantitatively assess how well foundation models can generalize across diverse traditions and how they can be enhanced to better represent characteristics of various musical traditions without requiring explicit modeling of tradition-specific features. The evaluation and adaptation of foundation models across diverse musical traditions represent important directions for advancing music representation learning, working toward more versatile and effective computational approaches to music understanding that balance tradition-specific characteristics with cross-traditional generalization.

2.8 Human Perception and Cross-Cultural Music Similarity

Understanding how humans perceive musical similarity has been a central question in music cognition and MIR research, with particular importance for developing computational approaches that align with human musical understanding. This section reviews the existing literature on human music similarity perception and its relationship to computational approaches, with special attention to cross-cultural contexts.

2.8.1 Foundations of Music Similarity Perception

Early research in music similarity perception established fundamental insights into how humans process and categorize musical relationships. Pioneering work [121] investigated statistical features and perceived similarity of folk melodies, finding that frequency-based musical properties could account for moderate amounts (40%) of listeners' similarity ratings, with descriptive variables like melodic predictability and rhythmic variability achieving slightly better performance (55%). This research suggested that while acoustic features provide some predictive power for human similarity judgments, substantial variance remains unexplained by traditional computational measures.

Research examining similarity perception across musical styles [122] found that human judgments were context-specific and roughly equivalent between trained musicians and non-musicians, with ratings primarily based on surface features such as dynamics, articulation, texture, and contour rather than deeper structural relationships. These findings highlighted the importance of immediately perceptible musical characteristics in human similarity assessment, suggesting that computational approaches emphasizing surface-level features might align better with human perception than those focusing on abstract structural analysis.

The multifaceted nature of human music similarity perception has been consistently demonstrated across studies. Similarity judgments involve both immediate surface features and deeper structural relationships, with significant individual and cultural variation in how these different dimensions are weighted and interpreted. This complexity suggests that effective computational approaches to music similarity must account for multiple perceptual dimensions while recognizing that the relative importance of these dimensions may vary across cultural contexts and individual listeners.

2.8.2 Computational Approaches to Music Similarity

The relationship between computational approaches and human perception has been a persistent concern in MIR research. Large-scale evaluations have provided crucial insights into this alignment, with comprehensive cross-site evaluations [123] comparing acoustic techniques against subjective measures across hundreds of popular artists. These studies demonstrated that acoustic measures could achieve agreement with ground truth data comparable to internal agreement between different subjective sources, suggesting that computational approaches can capture meaningful aspects of human musical understanding when properly designed and evaluated.

More recent work has explored how different types of computational representations align with human perception [124]. Audio representations were evaluated against human timbre similarity ratings, with the style embeddings from foundation models achieving superior performance compared to traditional signal processing features. This research suggests that modern foundation models may capture aspects of musical similarity that align more closely with human perception than traditional hand-crafted features, though questions remain about their effectiveness across diverse musical traditions.

The evolution from traditional signal processing features to learned representations reflects broader trends in the field toward more data-driven approaches. While hand-crafted features offer interpretability and direct connections to music theory concepts, learned representations from deep neural networks may capture complex patterns and relationships that better align with human perceptual processing, even if these patterns are less immediately interpretable from a theoretical perspective.

2.8.3 Cross-Cultural Dimensions of Music Similarity

Understanding similarity across cultural boundaries presents particular challenges for both human perception studies and computational approaches. Cultural context influences auditory perception and aesthetic appraisal, leading to diverse "listening frameworks" that shape how different communities understand and categorize musical experience. These culturally-conditioned perceptual frameworks suggest that similarity judgments may vary significantly across cultural contexts, with listeners from different musical backgrounds potentially emphasizing different aspects of musical relationships.

Research on the universality of music demonstrates cultural specificity in musical perception and understanding [2, 10]. While certain musical elements may transcend cultural boundaries, such as the recognition of emotional expressions or basic rhythmic patterns, the interpretation and evaluation of musical similarity appears to be significantly influenced by cultural background and musical training within specific traditions.

This cultural conditioning of musical perception has profound implications for the development of computational approaches to cross-cultural music similarity. Traditional MIR approaches, developed primarily within Western musical frameworks, may not capture the perceptual dimensions that are most salient to listeners from other musical traditions. Similarly, foundation models trained predominantly on commercial Western music may not align with human similarity judgments across diverse cultural contexts, highlighting the need for more culturally inclusive evaluation frameworks and training approaches.

Addressing these challenges requires careful methodological considerations including diverse participant populations representing multiple musical traditions, evaluation frameworks that assess different dimensions of similarity (overall musical characteristics, cultural identity, personal preference), systematic comparison of computational approaches ranging from traditional signal processing to modern foundation models, and metrics that capture both absolute performance and relative alignment with human judgment patterns across cultural contexts. This comprehensive approach to cross-cultural similarity evaluation provides the foundation for the systematic study presented in Chapter 6, which compares computational methods against human similarity judgments across diverse musical traditions, contributing new insights into the effectiveness and limitations of current approaches for cross-cultural music understanding.

2.9 Automatic Music Tagging

Automatic music tagging, or music auto-tagging, is the task of automatically predicting tags (such as genre, mood, instrumentation) from audio signals and has become a central task in MIR [125]. It represents a multi-label classification problem, where each music piece can be associated with multiple tags simultaneously.

This task serves as the primary evaluation context for the multicultural music representation methods developed in this dissertation. As a multi-label classification problem that captures multiple aspects of musical content, auto-tagging provides an ideal framework for assessing how well computational models represent diverse musical characteristics across traditions.

Several deep learning architectures have been proposed for music auto-tagging. Convolutional-based models, such as VGG-ish [37] and Musicnn [38], extract features from time-frequency representations, while end-to-end models like SampleCNN [69] and TCNN [70] process raw audio signals. Transformer-based approaches like AST [39] have recently demonstrated competitive performance

on auto-tagging benchmarks.

The evaluation of auto-tagging systems typically employs metrics designed for multi-label classification, including area under the receiver operating characteristic curve (ROC-AUC), average precision (AP), macro-F1, and micro-F1. These metrics account for both the binary nature of tag presence/absence and the potential imbalance in tag distributions across different musical traditions.

While auto-tagging has been extensively studied for commercial music, its application to traditional music from various regions presents additional challenges, including limited data, unique musical characteristics, and regionally specific tagging systems. In Chapter 4, we develop specialized approaches for tagging diverse musical traditions, exploring both transfer learning and few-shot learning, in order to address these challenges.

The auto-tagging task provides a practical framework for the cross-cultural representation learning approaches presented in subsequent chapters, allowing us to quantitatively assess how well different models capture the distinctive characteristics of diverse musical traditions.

2.10 Datasets and Models Used in this Work

This section presents the specific datasets and models used throughout this dissertation. By providing a comprehensive overview here, we can avoid duplicating this information in subsequent chapters.

2.10.1 Datasets

For our research on music representation learning across diverse traditions, we utilize a collection of datasets spanning different geographical regions and musical systems. These datasets are carefully selected to represent three distinct geographical regions: Europe and North America, the Eastern Mediterranean, and the Indian subcontinent, with each region represented by multiple datasets.

Figure 2.4 provides visual examples of the acoustic diversity captured in our dataset collection. The mel-spectrograms demonstrates some distinct patterns across musical traditions. Western music datasets (MagnaTagATune and FMA-medium) typically exhibit clear harmonic structures with regular temporal patterns, while Eastern Mediterranean music (Lyra and Turkish-makam) shows more complex modal characteristics with distinctive microtonal inflections visible in the frequency domain. Indian classical music recordings (Hindustani and Carnatic) demonstrate unique spectral signatures with prominent drone characteristics, complex ornamentations, and tradition-specific timbral qualities that reflect the use of traditional instruments and performance practices. These differences underscore the importance of developing computational approaches that can adapt to diverse musical characteristics rather than imposing uniform analytical frameworks.

For the cross-cultural music similarity study presented in Chapter 6, we expand this collection to include additional datasets representing Chinese traditional music (Jingju), Middle Eastern traditions (Arab-Andalusian), and Mediterranean music (CorpusCOFLA), providing a comprehensive sample of nine diverse musical traditions spanning Western, Middle Eastern, Mediterranean, Indian, and Chinese musical cultures.

As illustrated in Figure 2.5, the datasets that are utilized for automatic tagging exhibit characteristic long-tailed distributions, where a small number of tags account for the majority of annotations while many culturally significant attributes appear infrequently. This distribution pattern

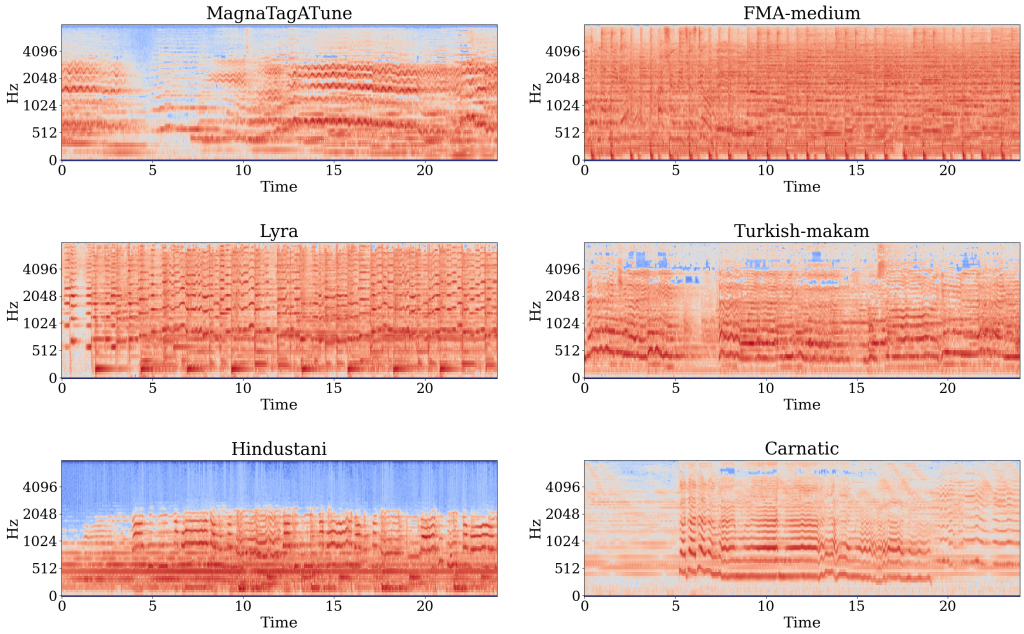

Figure 2.4. Mel-Spectrograms Across Musical Traditions. Representative mel-spectrograms from each dataset used in this dissertation, illustrating the diverse acoustic characteristics and temporal patterns across Western music (MagnaTagATune, FMA-medium), Eastern Mediterranean traditions (Lyra, Turkish-makam), and Indian classical music (Hindustani, Carnatic).

presents fundamental challenges for computational music analysis, as traditional supervised learning approaches often struggle with rare but important musical categories. The steep decline in tag frequencies across all traditions, from Western commercial music to traditional world music, motivates the development of few-shot learning methodologies that can effectively learn from limited examples, as explored in Chapter 4. Detailed information about the top 50 tags for each dataset can be found in Appendix A.

Western Music Datasets

MagnaTagATune (MTAT) The MagnaTagATune dataset [126] is widely used for music auto-tagging research. It consists of more than 25,000 audio recordings, with a total duration of approximately 210 hours. Each recording is annotated with a subset of 188 unique tags, though most research focuses on the top 50 most frequent tags, which include annotations for genre, instruments, and mood. This dataset primarily represents Western popular and classical music traditions.

FMA-medium The Free Music Archive (FMA) [114] is an open and accessible dataset used for various MIR tasks. The complete collection contains over 100,000 tracks organized in a hierarchical taxonomy of 161 genres. In our research, we use the FMA-medium subset, which consists of 25,000 tracks, each 30 seconds long, for a total duration of 208 hours. Like MTAT, FMA-medium primarily contains commercial music styles, including pop, rock, jazz, and electronic music.

Eastern Mediterranean Datasets

Lyra The Lyra dataset (see Chapter 3), developed as part of this dissertation, focuses on Greek traditional and folk music. It comprises 1,570 pieces with a total duration of 80 hours, making it

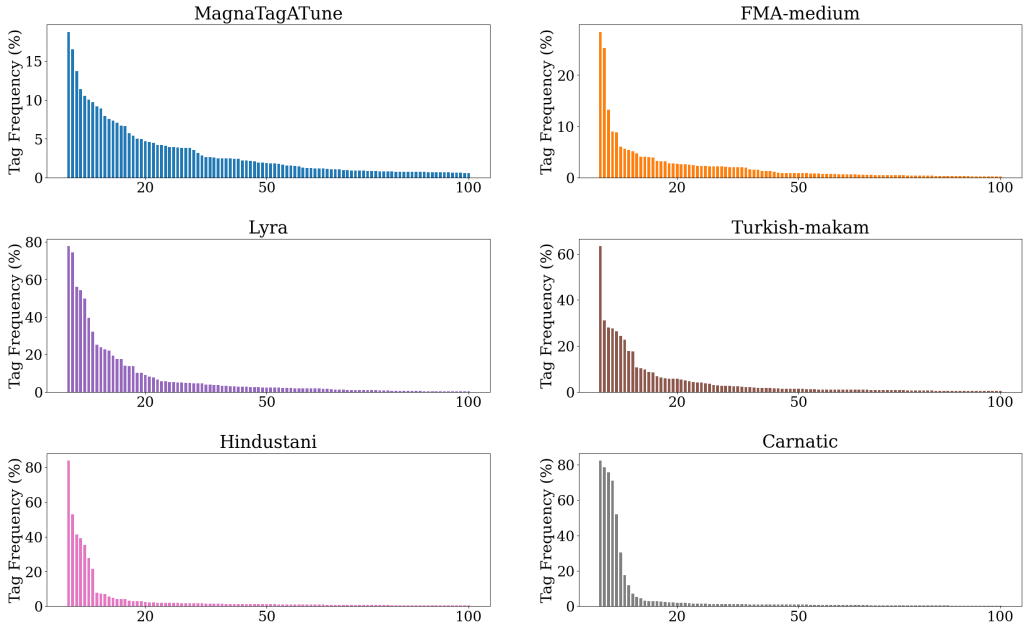

Figure 2.5. Tag Distribution Patterns Across Datasets. Frequency distribution of the most common tags in each dataset, demonstrating the long-tailed nature of musical annotations across all traditions. The steep decline in tag frequencies highlights the data scarcity challenges for rare but culturally significant musical attributes, motivating the few-shot learning approaches developed in this dissertation.

smaller but more focused than the other datasets in our collection. Lyra includes rich metadata regarding instrumentation, geography, and genre, with particularly fine-grained labeling focused on musicological aspects.

A distinguishing feature of the Lyra dataset is its homogeneous recording quality, as all content was collected from a documentary series presented by academics on Greek television. This ensures musicological soundness and consistent audio characteristics across the collection. The dataset focuses specifically on traditional and folk music, which offers unique perspectives by combining characteristics of both European and Eastern Mediterranean musical traditions.

Turkish-makam The Turkish makam corpus [84, 85] is part of the CompMusic project and includes thousands of audio recordings covering more than 5,000 works from hundreds of artists. For our research, we accessed 5,297 recordings with a total duration of 359 hours through the Dunya interface [88]. To maintain balance with other datasets, we limit each recording to a maximum of 150 seconds, resulting in a total duration of 215 hours. The annotations contain tags related to "makam" (modal structures), "usul" (rhythmic patterns), and "instruments."

Indian Classical Music Datasets

Hindustani The Hindustani corpus [19], also part of the CompMusic project, represents the classical tradition of North India. It includes 1,204 audio recordings with a total duration of 343 hours. To maintain consistency with other datasets, we limit each recording to a maximum of 780 seconds, resulting in approximately 206 hours of audio. The respective tags are related to "raga" (melodic frameworks), "tala" (rhythmic cycles), "instruments," and "form" (compositional structures).

Carnatic The Carnatic corpus [19] represents the classical tradition of South India and comprises 2,612 audio recordings with a total duration exceeding 500 hours. By limiting each recording to a maximum of 330 seconds, we reduce the total duration to 218 hours. As with the Hindustani dataset, the tags contain information about "raga," "tala," "instruments," and "form."

Additional Datasets for Cross-Cultural Similarity Study

For the comprehensive cross-cultural music similarity evaluation presented in Chapter 6, we incorporate additional datasets to provide broader cultural representation:

CorpusCOFLA The CorpusCOFLA dataset [35] focuses on flamenco music, a tradition with origins in Andalusia and diverse influences from Jewish, Arab, and Andalusian Gypsy cultures. The dataset consists of more than 1,800 audio recordings (95 hours) and contains metadata including a rich hierarchy of styles, performers, and editorial information.

Arab-Andalusian The Arab-Andalusian corpus [23] represents the musical tradition that developed in medieval Islamic territories of the Iberian Peninsula. This tradition combines Western and Eastern Mediterranean musical traits and has been preserved in North African countries. The dataset comprises 164 long recordings totaling approximately 125 hours with metadata including "nawba" (metrical mode), "tab" (melodic mode), and instrumentation.

Jingju The Jingju (Beijing Opera) corpus [127] represents a Chinese traditional performing art form combining musical and theatrical elements. It contains 864 recordings (71 hours) with metadata related to "shengqiang" (modal system) and "banshi" (metrical patterns).

Dataset Balance and Preprocessing

To ensure fair comparisons across datasets, we have taken several steps to balance and standardize our data. The Western datasets (MTAT and FMA-medium) have comparable total durations of approximately 210 hours, while the datasets from other regions (excluding Lyra) have been balanced to approximately 200-215 hours each by limiting individual recording lengths. We use consistent data splits across all experiments, following the protocol we establish in Section 4.3.

For our foundation model adaptation experiments, we further prepare the data by extracting 30-second segments from each training split of the traditional music datasets. To ensure balanced representation across traditions, we extract 200 hours each from the Turkish-makam, Carnatic, and Hindustani datasets, and 50 hours from Lyra (due to its smaller size), to create a combined 650-hour dataset integrating all four traditional music collections, which we utilize for multi-traditional continual pre-training.

For the cross-cultural similarity study, we select representative audio clips from each tradition, ensuring coverage across different musical characteristics while maintaining manageable dataset sizes for human annotation studies.

2.10.2 Models

Throughout this dissertation, we utilize several deep learning architectures for music analysis, focusing on models that can process audio spectrograms with minimal inductive bias. These models represent different architectural paradigms in deep learning: convolutional neural networks, musically-informed convolutional architectures, and Transformer-based approaches.

VGG-ish

VGG-ish is based on the Visual Geometry Group (VGG) network architecture [128], which was originally developed for image recognition. Our implementation follows the version described in [129], consisting of seven convolutional layers with 3×3 convolution filters and 2×2 max-pooling, followed by two fully-connected layers.

This model accepts mel-spectrograms as input, corresponding to 3.69-second audio chunks. Despite being adapted from computer vision, VGG-ish has proven effective for various MIR tasks, demonstrating the transferability of CNN architectures to audio spectrograms treated as image-like inputs.

The architecture consists of:

- Seven convolutional layers with increasing filter counts (32, 64, 128, 128, 256, 256, 512)
- 2×2 max-pooling after each convolutional layer
- Two fully-connected layers (2048 units and the output layer)
- ReLU activations and batch normalization throughout the network

Musicnn

Musicnn [68] is a music-specific CNN architecture designed to capture both temporal and timbral features from audio spectrograms. Unlike general-purpose CNN architectures adapted for audio, Musicnn incorporates domain knowledge about music signal characteristics directly into its architecture.

The key innovation of Musicnn is its first convolutional layer, which employs parallel vertical and horizontal filters:

- Vertical filters (with shapes like $M \times 1$, where M is the frequency dimension) capture timbral features across the frequency spectrum
- Horizontal filters (with shapes like $1 \times N$, where N spans the time dimension) capture temporal features and rhythmic patterns

These parallel filter paths are then concatenated and processed through additional 1D convolutional layers, followed by a pair of dense layers that summarize the extracted features and predict the relevant tags. Musicnn processes mel-spectrograms from 3-second audio chunks, capturing musically relevant patterns across both time and frequency dimensions.

Audio Spectrogram Transformer (AST)

The Audio Spectrogram Transformer (AST) [39] represents a more recent architectural paradigm based entirely on attention mechanisms. Adapted from the Vision Transformer architecture to audio spectrograms, AST demonstrates how Transformer models can effectively process music signals without relying on convolutional operations.

The key components of AST include:

- Spectrogram patching: The input mel-spectrogram is divided into 16×16 patches in both time and frequency dimensions
- Linear projection: Each patch is flattened and projected to a 768-dimensional embedding

- Positional encoding: A learnable positional embedding is added to each patch embedding to preserve spatial information
- Transformer encoder: The sequence of patch embeddings is processed through a standard Transformer encoder with multi-head self-attention
- Classification head: The encoder output is passed through a linear layer for final predictions

Following the recommendations of the original authors, we set the input length to 8 seconds for all AST experiments, allowing the model to capture longer-term temporal relationships than the CNN-based approaches.

2.10.3 Foundation Models

For our work on foundation model evaluation and adaptation, we utilize several state-of-the-art models to provide a comprehensive assessment of capabilities across diverse musical traditions:

MERT-95M

The Music undERstanding model with large-scale self-supervised Training (MERT) [40] employs a masked acoustic modeling approach similar to BERT in natural language processing [107]. The MERT-v1-95M variant has 12 Transformer encoder layers with approximately 95 million parameters. It is pre-trained on approximately 1,000 hours of music, primarily from commercial genres, using a masked spectrogram prediction objective. MERT accepts log-mel spectrograms as input and produces contextual representations that capture musical features at multiple levels of abstraction.

MERT-330M

The larger variant of MERT [40] scales up the model architecture to approximately 330 million parameters, with 24 Transformer encoder layers and wider attention heads. This model maintains the same masked acoustic modeling approach as MERT-95M but offers increased capacity for learning complex musical representations. The expanded parameter count potentially enables more nuanced modeling of diverse musical characteristics, though at increased computational cost.

CLAP-Music

Contrastive Language-Audio Pre-training for Music (CLAP-Music) [41] adapts the CLIP (Contrastive Language-Image Pre-training) framework to the audio domain, specifically focused on music. This model learns joint embeddings of music audio and textual descriptions through contrastive learning, aligning representations from both modalities. The audio encoder is based on a Vision Transformer architecture that processes mel-spectrograms, while the text encoder processes natural language descriptions of musical content. This multimodal approach offers advantages for analyzing diverse musical traditions, as the textual descriptions may help bridge gaps between different musical systems.

CLAP-Music&Speech

This variant of CLAP [41] extends the training data to include both music and speech content, creating a more generalized audio-language model. By incorporating speech alongside music during pre-training, this model potentially develops more robust representations of human-produced audio,

including singing and vocal techniques that vary across musical traditions. The additional speech data also provides broader exposure to diverse languages and acoustic environments, which may benefit generalization across different musical systems.

Qwen2-Audio

Qwen2-Audio [42] represents a unified audio understanding foundation model capable of processing both speech and music. This model employs a Transformer-based architecture with custom adaptations for audio signal processing. Qwen2-Audio is pre-trained on a diverse collection of audio data using multiple objectives, including masked acoustic modeling and contrastive learning. Its unified approach to audio understanding potentially enables better transfer between different audio domains, including across musical traditions.

CultureMERT

CultureMERT represents our culturally-adapted foundation model, developed through continual pre-training on diverse musical traditions (see Chapter 5). Built upon the MERT-95M architecture, CultureMERT incorporates learning from Greek, Turkish, and Indian musical traditions while maintaining performance on Western music benchmarks. This model serves as an example of how foundation models can be systematically adapted to better represent diverse musical cultures.

2.11 Summary

This chapter has provided a comprehensive background for the research presented in this dissertation, covering theoretical frameworks for comparative music analysis, fundamental concepts in music signal processing, deep learning approaches in MIR, methodological challenges in analyzing diverse musical systems, world music datasets and computational ethnomusicology, transfer learning and few-shot learning, foundation models in music, human perception and cross-cultural music similarity, and the specific datasets and models used throughout this dissertation.

Several interconnected themes emerge from this background that inform the subsequent chapters. Our approach emphasizes data-driven comparative analysis, focusing on studying relationships between musical systems using end-to-end deep learning models rather than imposing predetermined analytical frameworks. This methodology enables reduced methodological bias by avoiding feature engineering based on particular music theories and using data-driven approaches that minimize the risk of imposing inappropriate analytical assumptions on diverse musical traditions.

Central to our research is balancing specificity and generalization, exploring how to respect the distinctive elements of diverse musical traditions while enabling meaningful comparison and knowledge transfer across cultural boundaries. Given the data scarcity common in many traditional musical systems, we emphasize resource-conscious design through approaches like transfer learning and few-shot learning that can work effectively with limited annotated data, making it practical to develop models for diverse traditions. Our methodology pursues adaptive representation learning rather than imposing a single representational framework, developing approaches that can adjust to different musical traditions by learning appropriate representations from data rather than relying on predefined musical features.

A crucial aspect of this work involves human-centered evaluation, integrating human perception studies to validate computational approaches and ensure that similarity measures align with how

listeners from diverse cultural backgrounds actually perceive musical relationships. This human-in-the-loop validation becomes essential for developing culturally aware music technology systems that respect the perceptual dimensions most relevant to different cultural contexts.

Finally, we explore foundation model adaptation strategies for enhancing the capabilities of large-scale music models across diverse traditions, addressing their methodological limitations while leveraging their powerful representational capabilities. This includes systematic evaluation of how foundation model representations align with human cross-cultural music perception, providing crucial insights for developing more effective and culturally aware music AI systems.

These themes inform the research presented in subsequent chapters, which addresses the challenges of music representation learning across diverse traditions through dataset development, transfer learning, few-shot learning for low-resource scenarios, foundation model evaluation and adaptation strategies, and comprehensive evaluation against human cross-cultural music perception. By building on this background, the dissertation aims to advance more versatile and effective approaches to computational music analysis across diverse traditions while ensuring alignment with human perceptual understanding of musical similarity across cultures.

Chapter 3

The Lyra Dataset: A Resource for Greek Traditional and Folk Music

3.1 Motivation

As established in the previous chapters, computational approaches to music analysis have been predominantly developed for and evaluated on Western music traditions, creating significant gaps in our understanding and representation of diverse musical cultures. While several datasets for non-Western traditions have emerged in recent years, as discussed in Section 2.5, Greek traditional and folk music remains notably underrepresented in structured, high-quality datasets suitable for computational analysis. This gap is particularly significant given Greece's unique geographical and cultural position at the crossroads of Eastern and Western musical traditions, offering valuable perspectives that could inform computational approaches to cross-cultural music analysis.

The existing datasets for Greek music, such as the Greek Audio Dataset (GAD) [82] and its expanded version, the Greek Music Dataset (GMD) [83], have several limitations that hinder comprehensive computational analysis of Greek traditional music. First, these datasets cover a broad spectrum of Greek music, including contemporary pop and rock, rather than focusing specifically on traditional and folk genres. Second, they employ relatively coarse categorical labels that don't capture the rich musicological aspects of traditional music. Finally, the audio quality varies significantly across recordings, potentially introducing confounding factors in computational analysis that could mask the actual musical characteristics of interest.

To address these limitations and advance multicultural music representation learning, we developed the Lyra dataset [45], a focused collection of Greek traditional and folk music with consistent audio quality and fine-grained musicological annotations. This chapter describes the creation, structure, and characteristics of this dataset, which serves as a foundation for the cross-cultural music analysis approaches explored in subsequent chapters.

The Lyra dataset aims to make several specific contributions:

- Provide a high-quality resource for computational analysis of Greek traditional and folk music, enabling more inclusive approaches to music information retrieval that extend beyond Western-centric paradigms
- Capture the rich musicological diversity of Greek traditional music through fine-grained annotation of instrumentation, genres, geographical origins, and performance contexts
- Ensure consistent audio quality across the collection, minimizing technical variations that could interfere with the analysis of musical characteristics

Figure 3.1. Documentary Series Screenshot. Representative frame from "To Alati tis Gis - Salt of the Earth" showing the quality of the source material used for the Lyra dataset.

- Enable exploration of the unique aspects of Greek music that blend Eastern and Western influences, potentially revealing insights about musical characteristics that bridge different cultural traditions

- Support baseline computational tasks including genre, instrument, and regional classification, establishing performance benchmarks for future research

The development of the Lyra dataset represents the first step in our research agenda on multicultural music representation learning. By creating a structured, high-quality resource for an underrepresented musical tradition, we establish the foundation for subsequent investigations into cross-cultural knowledge transfer, few-shot learning for cultural adaptation, and foundation model evaluation and enhancement. The dataset's focus on Greek traditional music, with its blending of Eastern and Western characteristics, makes it particularly valuable for cross-cultural investigations, serving as a potential bridge between these broader musical spheres.

Unlike previous collection efforts [82, 83], our approach emphasizes musicological soundness, homogeneous recording quality, and detailed annotation, addressing the specific challenges that have limited computational analysis of non-Western traditions. The following sections detail our methodology for extracting and annotating this dataset from a documentary series, the resulting characteristics of the collection, and baseline classification experiments that demonstrate its utility for computational musicology.

3.2 Dataset Extraction and Description

3.2.1 Challenges and Methods

Large amounts of clean data is fundamental for current AI models to achieve their full potential. In this Chapter, we walk all the way, from the "data in-the-wild" multimedia content of a TV show to a fully annotated dataset, through a combination of machine automation and human evaluation/annotation processes. The consistency of the dataset and the richness of information

it provides are tested by developing and training models that perform three different classification tasks.

In the case of Greek traditional and folk music, there are few cases where metadata is combined with recordings in a structured manner. Additionally, there is a matter of quality of recordings as it is significantly affected by various factors, including the equipment used, the social occasion (e.g., during a festival or inside a studio) and the time period in which it took place, i.e., older recordings tend to be of lower quality.

An integration of dissimilar recordings, in terms of quality, can introduce significant deficiencies towards studying the musicological characteristics of world music with computational tools. In order to truncate the effect of the audio quality factor, we decided to incorporate the episodes from the Greek documentary series "To Alati tis Gis - Salt of the Earth" broadcasted by ERT (Hellenic Broadcasting Corporation), where primarily traditional and folk music is presented. The episodes were filmed during a 10 year period under strict production-level specifications, resulting to very clean and homogeneous audio content while significant wealth of information is provided by the presenter and the guests in the form of narrations between music performances. Figure 3.1 shows a representative frame from the documentary series, illustrating the professional production quality that characterizes the source material.

The presented dataset consists of both the multimedia content and the annotations of interest. The multimedia content is provided as start and end timestamps that correspond to a single music piece, as parts of a longer episode, which is available online. Regarding the annotations, a taxonomy of labels is defined, based on the potential purposes of studies that might involve this dataset, considering also what metadata information can be retrieved either directly from the source or be integrated by volunteer annotators during the data collection process.

The study of Greek traditional and folk music involves knowledge about (i) the instrumentation, (ii) the genres, (iii) the places of origin and (iv) the way listeners perceive this music in terms of "danceability", among others. While musical instruments, genres and geography are semantically well-defined, the same can not be claimed for listeners' perception. Having at hand the multimedia content, i.e., audio and video, can be helpful to this end. Annotation about whether a music piece is being danced during its live performance can reveal cultural characteristics regarding the way this piece is perceived by the community, because body movements play an important role in music perception [30].

As a result, the taxonomy consists of (i) the musical instruments participating in the performance of each music piece (singing voice is considered an instrument), (ii) the musical genres and sub-genres that are identified by musicologists in Greek music, (iii) the places of origin and (iv) whether the music piece is being danced during its performance.

Volunteer annotators, students of the Department of Music Studies, undertook the task of separating each episode in music pieces and also labeling each one of them according to the specified taxonomy. A helper website was utilized where the respective category labels were added. An account was created for each annotator for the label assignment task. Every piece was labeled by two annotators and the final labels are the set of them where both annotators agree. At the end, the dataset that contains the aforementioned annotations along with the timestamps and the respective video id for each music piece was extracted from the database of the helper site.

3.2.2 Dataset description

Lyra dataset is organized into a single table where each row corresponds to a music piece while the columns include the various metadata information. Table 3.1 demonstrates the metadata

name	# unique	multi-label	description
id	1570	No	unique identifier of the music piece
instruments	32	Yes	instruments participating in performance
genres	32	Yes	music style annotations
place	81	Yes (to contain regions)	full hierarchy of the place of origin
coordinates	81	No	latitude and longitude
is-danced	2	No	binary value (0 or 1)
youtube-id	74	No	id of the episode available online
start-ts	1570	No	start timestamp of the piece
end-ts	1570	No	end timestamp of the piece

Table 3.1. Metadata in the Lyra Dataset. Description of fields and content structure, showing information about the respective annotations.

categories.

Beginning with the simplest metadata categories, in terms of description, "id" is a unique identifier for each piece, generated by its title, replacing Greek with Latin characters and spaces with dashes. As expected, the number of unique values will be the same with the number of pieces, namely 1570. The same stands for "start-ts" and "end-ts" that denote the exact time (second) that a song starts and ends in the corresponding video. The duration of the music pieces sums to approximately 80 hours.

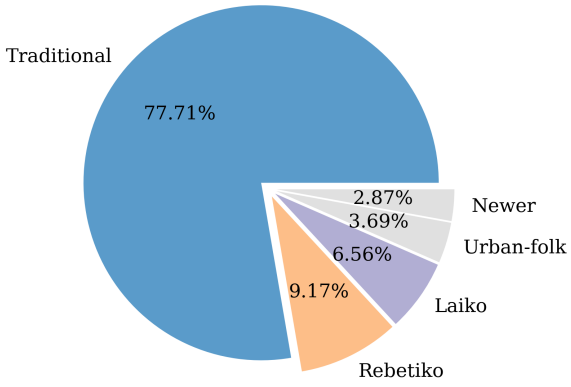

Figure 3.2. Genre Distribution in the Lyra Dataset. Relative frequencies of music genres, showing the predominance of traditional songs in the dataset.

The column "youtube-id" contains the id under which the video of the full episode is available online. The count of unique values are essentially the number of episodes that were used for the creation of the dataset. A typical duration of an episode is roughly a hundred minutes. The "is-danced" binary label informs about whether a music piece is being danced by the guests of the show. The music pieces annotated with "1" are approximately 51% while in the rest of them no dance performance occurs.

The classification of Greek music in "genres" is a work that requires one to take into account a number of socio-cultural and anthropo-geographical criteria. At an abstract level, we can distinguish the music of urban centers in contrast to the music of rural areas of Greece, with the former including rebetiko, laiko, urban-folk among others, while the latter, the music of rural areas, is

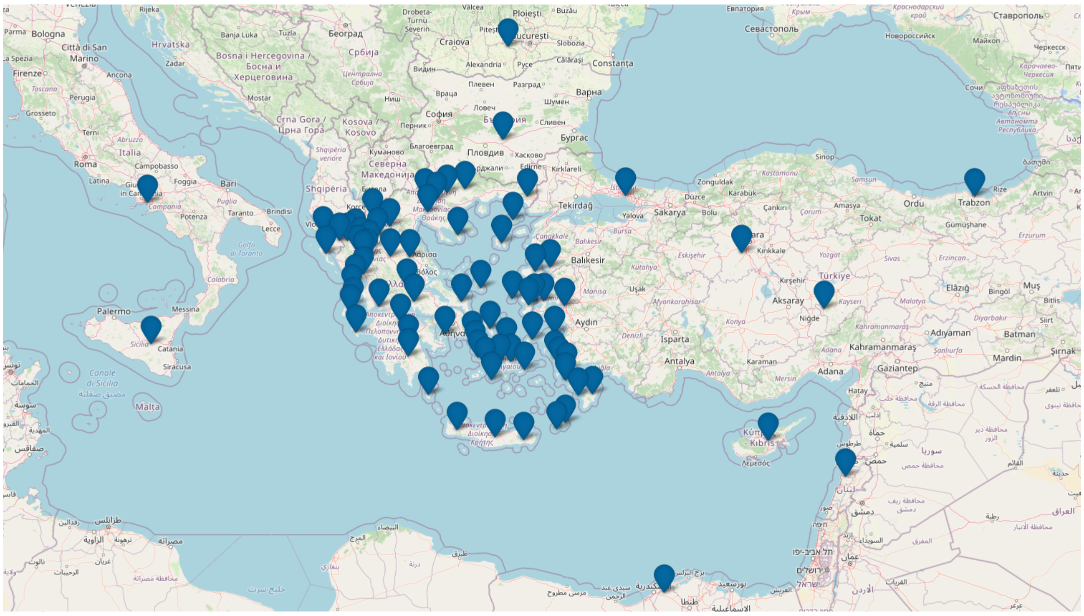

Figure 3.3. Geographical Distribution of Music Origins. Map visualization of all places represented in the dataset, highlighting the regional diversity of Greek musical traditions.

what we generally call traditional music. Figure 3.2 shows the frequencies of the genres in the dataset, with "traditional" being the dominant one constituting almost 78% of the total. Depending on the place of origin, the style of a traditional music piece varies accordingly and, thus, several sub-genres flourish, such as Epirotic for the songs originated from Epirus. The 32 unique values in this metadata category are separated into 5 distinct genres and 27 sub-genres.

From a musicological perspective, Greek traditional music can be divided into two large geographic areas, i.e., the island and the mainland Greece. Each one creates a distinctive musical feeling as there are large variations both on the rhythmic approach and the scales that are commonly used. For example, in islands we frequently come across music pieces with simple, fast rhythms while in the mainland more complex, slow rhythms are the norm.

The "place" (of origin) metadata category can be annotated with (i) a single label when the region from which a song derives is known, (ii) two labels when both region and a specific place are known and (iii) "None" denoting that there is not a specific place of origin for this piece. As an example, a music piece can be annotated with the region "Aegean sea" or with both "Aegean sea" and "Naxos", an island of the Aegean sea, if this knowledge is available. Specifically, from the 81 unique places in the dataset, 20 are regions and only half of them include the remaining 61. The most represented regions can be seen in Figure 3.4.

The exact latitude and longitude of each place is also available at the "coordinates" column. The music pieces that do not have an explicit place of origin, such as the ones that belong to the "laiko" genre, are accounted for approximately 23% of the total. Figure 3.3 shows the location of the 81 places that exist in the dataset. We may notice the constant ability of music to excess the borders; places where Greek culture thrived in the past and neighboring countries that share the same tradition, form a mosaic of people that communicated freely with each other in a musical way that has reached towards us.

Analogous connections, like the ones between genres and places, one expects to be observed between places and instruments as well. Indeed, for over 100 years, the established music ensembles of Greek traditional music are generally two, namely (i) those with the violin as leading instrument

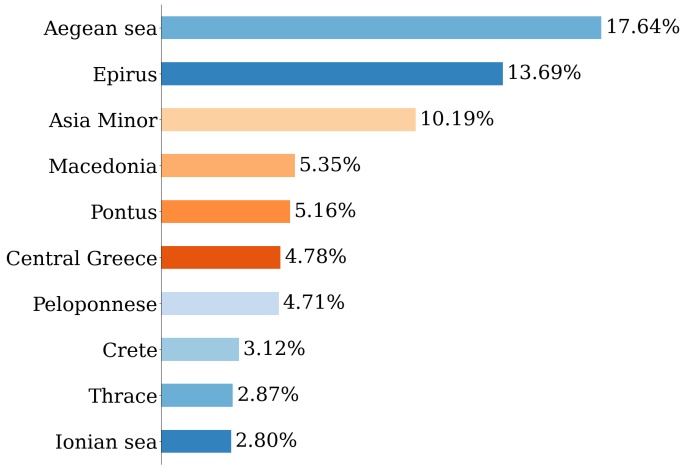

Figure 3.4. Regional Representation in the Dataset. Relative frequencies of the most represented geographical regions, demonstrating the distribution of musical samples across cultural areas in Greece.

(often substituted by lyra and santouri), which have a greater presence in island Greece and (ii) those with the klarino (Greek clarinet) as the leading instrument, which is dominant in the mainland.

In the popular and modern music domain, there is a great variety of instruments, but in most cases bouzouki, guitar, accordion and bass are common members of a laiko or rebetiko music ensemble. In the traditional music groups, the percussion and the laouto (Greek lute) are permanent companions of the leading instruments, offering melodic-harmonic background and rhythmic support. Of course, voice holds the main role in all kinds of performances.

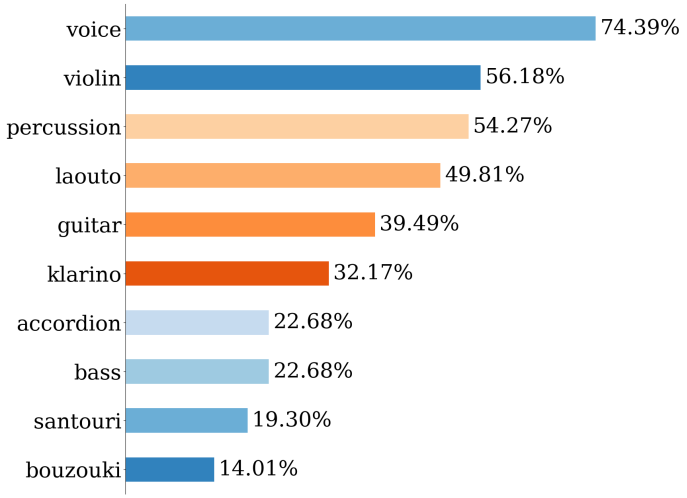

Figure 3.5. Instrument Frequency Distribution. Relative frequencies of the most common musical instruments in the dataset, illustrating the instrumental palette of Greek traditional music.

In Figure 3.5 one can see the frequencies of the most popular instruments in the dataset. Singing voice is evident in almost 75% of it and instruments like violin, percussion and laouto, that have presence in both islands and mainland as well, are following.

With regards to the music ensembles, it should be noted that 296 unique groups of instruments exist in the dataset, with the one constituted by voice, violin, percussion, laouto and klarino being

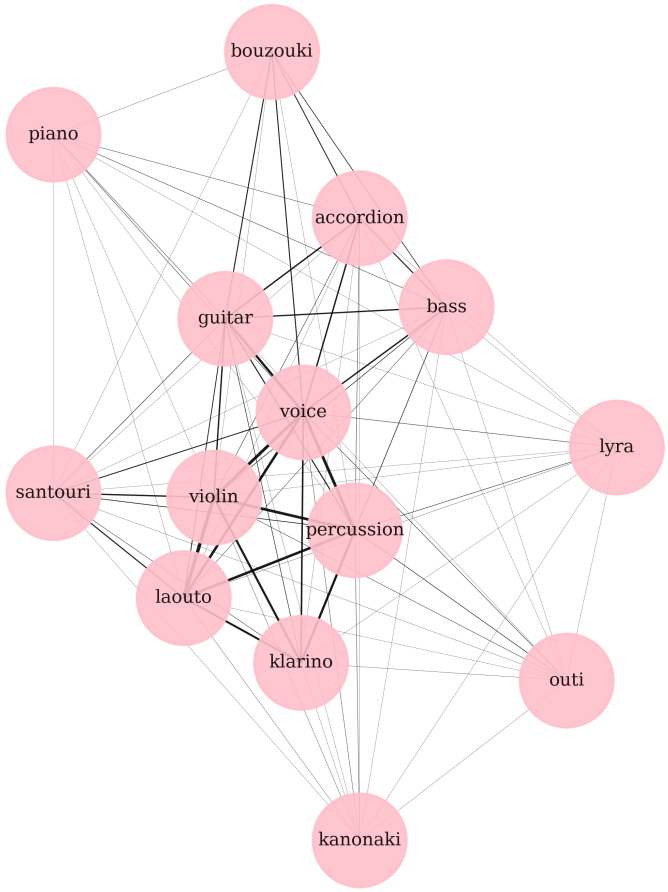

Figure 3.6. Instrument Co-occurrence Network. Graph visualization showing relationships between the fourteen most common instruments, with edge width proportional to co-occurrence frequency in music pieces.

by far the most popular by participating in the performance of around 12% of the music pieces. The co-occurrences of the most popular instruments can be seen in Figure 3.6 where the width of the graph edges is proportional to the number of pieces a pair of instruments co-occur in the dataset.

Sample rows of the dataset can be seen in Table 3.2. The dataset along with the baseline classification methods and the trained models are available online.[1]

The shared metadata should be considered as the version 1.0 of the dataset. In the next versions, it will be evolved towards two main directions, namely (i) the incorporation of more metadata categories such as annotations according to the content of the lyrics, the lyrics themselves as well as information about the types of the dances that occur and (ii) the addition of more music pieces by following the same process either for next episodes of the same documentary series or for other series that have a similar theme.

[1] https://github.com/pxaris/lyra-dataset

id	instruments	genres	place	coordinates	is-danced	youtube-id	start-ts	end-ts
alexandra	voice|violin| percussion| laouto| klarino	Traditional| Epirotic	Epirus| Zagori	39.8648| 20.9284	0	qr0wc1mLFUk	749	927
choros-tik	percussion| lyra	Traditional| Pontian	Pontus	40.9883| 39.7270	1	AwsOY3aLaIs	1731	1886
agiothodo-ritissa	voice|violin| santouri| percussion| laouto|guitar	Rebetiko	None	None	1	0cj8BNcAhg4	2632	2853
einai-arga-poly-arga	voice|piano| guitar|bass| bouzouki| accordion	Laiko	None	None	0	zkoqg3VRVLA	2365	2614

Table 3.2. Sample Entries from the Lyra Dataset. Representative rows demonstrating the multi-valued field structure with pipe-delimited values.

3.3 Baseline Classification

The audio recording of each music piece is represented using a Mel-scaled Spectrogram (mel-spectrogram): this is a spectrogram whose frequencies are converted to the mel scale according to the equation:

$$m = 2595 \log_{10}(1 + \frac{f}{700}) = 1127 \ln(1 + \frac{f}{700}), \tag{3.1}$$

where **m** is the frequency in Mels and **f** is the frequency in Hz. Mel-spectrograms are calculated per fix-sized segment duration of 10 seconds. A non-overlapping window of 50 milliseconds has been applied, therefore the Mel-spectrogram size is 200 windows $\times$ 128 frequency bins.

As a baseline classification approach, each 10-second Mel-spectrogram is classified to the aforementioned tasks (genre, place and instruments) using a Convolutional Neural Network (CNN). CNNs have been widely used in general audio [130], speech [131, 132] and music [133] classification tasks. In particular, we have adopted the following architecture: 4 convolutional layers of 5×5 kernels, single stride, and max pooling of size 2. The number of convolutional kernels (channels) are for the first layer 32, for the second 64, for the third 128 and for the fourth 256. The final output of the convolutional layers is passed through 3 fully connected layers, with the first having an output dimension of 1024, the second 256 and the third equal to the number of classes.

Note that fix-sized duration of segments is necessary, since audio recordings do not share the same size. Adopting a much longer segment would require zero padding for several mel-spectrograms and probably more CNN parameters. In addition, splitting the song into non-overlapping segments achieves some type of data augmentation. For two of the adopted classification tasks (namely genre and place), we have trained the CNNs using a multiclass, single-label setup, while for the instrument task, which is multi-label, we have trained multiple binary CNNs, one for each instrument, which have been evaluated separately.

After the training and validation procedure of each of the aforementioned CNNs, final testing was applied on the respective test recordings. For the test set, to avoid spreading pieces from the same broadcast across data splits, we separate training and test data on an episode level. From the 74 unique episodes, we randomly split 20% of them and use all the music pieces they include, namely 330, to form the test data. Obviously, this final testing needs to be carried out on a "song-level", not a 10-sec segment level. Towards this end, a simple aggregation method was adopted, by just averaging the posteriors of the individual segment decisions of the CNNs. This aggregated

Classifiers type	Task	F1 (%)
Multiclass classifiers	Genre	82.3
	Place	85.5
Binary classifiers for Instruments	voice	75.8
	violin	86.3
	percussion	97.1
	laouto	96.7
	guitar	80.2
	klarino	95.0

Table 3.3. Classification Performance on Training Data. F1 scores of various classifiers on 10-second segments using a 20% validation subset.

estimate was used as the final prediction and evaluated in the final testing.

3.4 Results

The performance results during the training of the baseline classifiers are shown in Table 3.3, computed on a validation subset that corresponds to 20% of the segments. For the multi-label task (instrument recognition), we show the F1 metric for each binary subtask separately, while for the single-label tasks (genre and place) we show the overall macro F1 for all classes. We remind that this evaluation is performed on the validation split of the 10-second data.

As soon as the 10-second classifiers are trained, they are applied on the whole recordings of the testing data, and a simple majority aggregation is performed to extract the final decision, as described in the previous Section. For the instrument classification task, we compute the Area Under the Curve (AUC) metric per label (binary classification subtask). The results are shown in Table 3.4.

Instrument	AUC (%)
voice	68.9
violin	85.2
percussion	95.1
laouto	93.8
guitar	73.5
klarino	90.9

Table 3.4. Instrument Classification Performance. Area Under the Curve (AUC) scores for instrument classifiers on the test data, showing recognition accuracy for different traditional Greek instruments.

Finally, the confusion matrices along with the respective F1 measures for the multiclass, single-label classification tasks of "genres" and "places" are shown in Figures 3.7 and 3.8. All genres have been taken into account for the respective classification, but not the sub-genres. On "places" task, the pieces have been classified to the 10 most common regions (including "None") plus the "other" category for the remaining. Genres classifier macro F1-score is 39.9% and micro F1-score is 87.2%, while for the places classifier the macro F1-score is 34.4% and the micro F1-score is 42.4%.

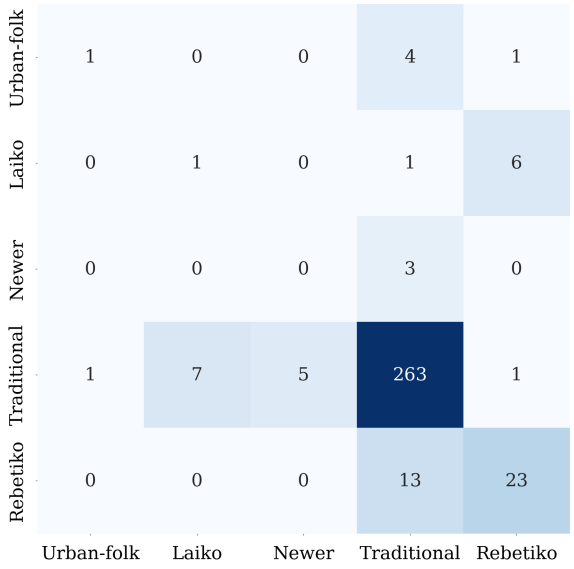

Figure 3.7. Genre Classification Performance. Confusion matrix for genre classes on test data, yielding Macro F1-score of 39.9% and Micro F1-score of 87.2%.

3.5 Discussion

A reason that the "voice" classifier has lower performance compared to the other ones may be of a musicological character. Indeed, while the presence of the rest of the instruments can depend significantly on the music style of a piece, the same does not apply to "voice" as it is the dominant musical instrument in any genre. Given the fact that the binary classifier is trained to recognize an instrument (evident in a part of a music piece) in each of the 10-second segments, regardless if it is present on it or not, we expect to move towards a space with latent musical features such as the music style, where "voice" may not be as discriminative as the rest of the instruments are.

With regards to confusion matrices, the misclassifications can be either due to imbalance between classes or statistical correlations across them. Specifically, for the "places" task, the confusions between regions that are geographically near may be justifiable, while for "genres" task the imbalance between the classes seems to have significantly affected the performance of the model at the least represented ones.

The classifier performance is improved in the work presented in Chapter 4, where a wide spectrum of models are utilized along with a cross-cultural transfer learning framework to further enhance their performance across diverse musical traditions.

3.6 Conclusions

Greek traditional and folk music integrates components of Eastern and Western idioms, providing interesting research directions in the field of computational ethnomusicology. We present "Lyra", a dataset of 1570 traditional and folk Greek music pieces that includes audio and video (timestamps and links to YouTube videos), along with annotations that describe aspects of particular interest for this dataset, including instrumentation, geographic information and labels of genre and subgenre, among others. The advantage of this dataset is that the entire content is harvested from web resources of a Greek documentary series that was produced by academics with specialization in this music and, therefore, includes high-quality and rich annotations extracted

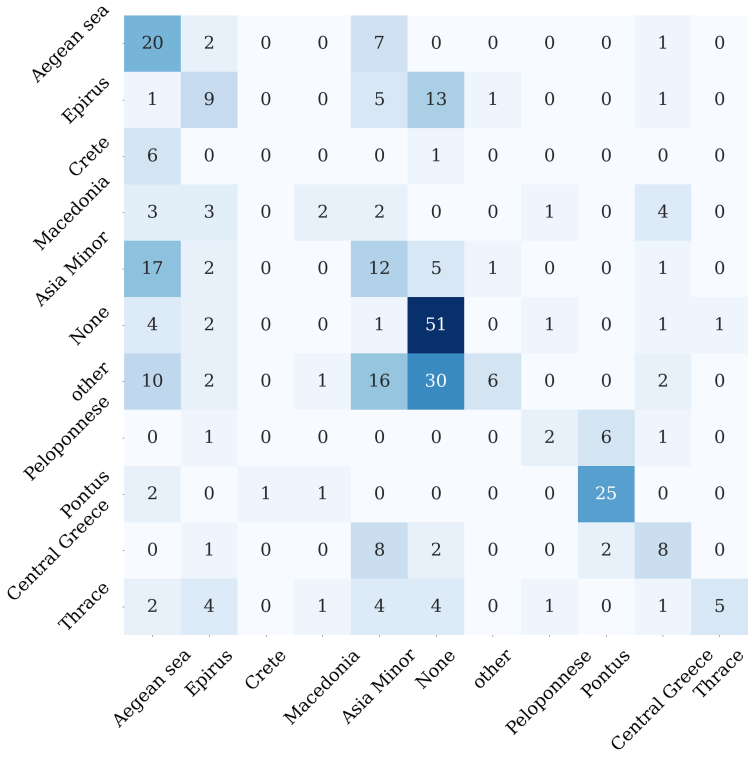

Figure 3.8. Geographical Classification Performance. Confusion matrix for place classes on test data, achieving Macro F1-score of 34.4% and Micro F1-score of 42.4%, showing regional identification challenges.

from the content of the shows. Additionally, the production of recordings and video material is professional-level, providing a common ground in terms of audio quality. Three baseline audio-based classification tasks are performed, namely instrument identification, place of origin and genre classification.

The presented results indicate that specialized tasks, that use the audio signal, can potentially provide valuable insight about several aspects of this music. The combination of video and audio signals allows possible experimentation on methods that process multimodal data. The Lyra dataset includes material that readily allows MIR tools to be employed for reaching valuable musicological results, and can potentially foster the expansion of MIR methods altogether.

Beyond its immediate utility for Greek music research, the Lyra dataset serves as a foundation for the broader investigations presented in subsequent chapters of this dissertation. In Chapter 4, we utilize Lyra alongside other world music datasets to investigate cross-cultural knowledge transfer patterns between diverse musical traditions. Chapter 5 further employs the dataset in comprehensive evaluations of foundation models across multiple musical cultures, demonstrating how datasets like Lyra can be integrated into computational approaches for music representation learning.

Chapter 4

Learning Across Cultures

4.1 Motivation

As discussed in Chapters 1 and 2, the field of Music Information Retrieval (MIR) has traditionally focused on Western musical traditions, creating systems that may not adequately represent or analyze the distinctive characteristics of diverse musical systems worldwide. This chapter addresses this limitation through two complementary approaches to learning across musical cultures, each targeting different aspects of the challenge. The work presented here draws from our investigations into cross-cultural transfer learning [46] and multi-label few-shot learning for world music [47], which together provide comprehensive strategies for addressing the challenges of multicultural music representation learning.

4.1.1 Cross-Cultural Knowledge Transfer

The majority of pre-trained models in MIR have been developed on Western musical datasets, raising important questions about their applicability to diverse musical cultures. When analyzing world, folk, or traditional music, we must consider: what is the potential of models trained on Western music when applied to different musical cultures, and can models trained on specific non-Western traditions provide meaningful embeddings for cross-cultural analysis?

Transfer learning offers a promising approach for leveraging knowledge across musical traditions, potentially enabling models to benefit from patterns learned in different cultural contexts. While transfer learning has shown significant benefits in various domains, its effectiveness for cross-cultural music analysis remains largely unexplored. Prior research has shown that transfer learning can lead to significant performance improvements compared to training from scratch [134], but the patterns of transferability across diverse musical traditions have not been thoroughly examined.

The auto-tagging task, predicting tags related to genre, instrumentation, mood, and other attributes from audio signals, provides an ideal context for investigating cross-cultural knowledge transfer, as it captures multiple aspects of musical content [125] that may transfer differently across cultural boundaries.

Previous studies have applied deep learning models to specific musical traditions, including Indian classical music classification [135], Turkish makam recognition [136, 137], and Western music auto-tagging [66, 68]. However, comprehensive cross-domain knowledge transfer analysis across diverse musical cultures has been missing. This gap is particularly significant given the growing availability of datasets representing various musical traditions, including the Lyra dataset for Greek traditional music described in the previous chapter.

By systematically evaluating knowledge transfer across Western, Eastern Mediterranean, and

Indian musical traditions, we can derive insights about computational relationships between these cultures. These patterns may reveal which musical traditions share underlying features that facilitate knowledge transfer, potentially reflecting historical connections, geographic proximity, or parallel musical developments. This approach also addresses the Western-centric bias in MIR research, moving toward more versatile computational approaches that recognize the value of diverse musical knowledge.

4.1.2 Learning from Limited Examples

While transfer learning can effectively leverage knowledge when substantial annotated data is available, many musical traditions from various regions face significant challenges of data scarcity and tag imbalance. These challenges limit the applicability of conventional deep learning approaches, including transfer learning, which typically require abundant labeled examples. This is particularly problematic for underrepresented tags within established music domains and for emerging or niche musical traditions where comprehensive annotated datasets may be unavailable.

The ability to learn from limited examples is a remarkable feature of human cognition that enables rapid adaptation to new concepts and contexts [138]. Few-shot learning aims to bridge the gap between human and machine learning capabilities by developing methods that can generalize effectively from minimal examples [94]. Although few-shot learning has been applied in various domains including computer vision [96, 139], natural language processing [140, 141], and acoustic signal processing [142, 143], its application to multi-label music classification represents a novel contribution to the field.

In the cross-cultural transfer learning work presented in Section 4.2, we demonstrate that knowledge can be effectively transferred between musical traditions, but this approach remains limited to frequently occurring tags with sufficient training examples. Traditional multi-label classification methods [144], while effective in many scenarios, are not specifically designed for the extreme data scarcity often encountered in world music research. Few-shot learning [95, 145, 146] addresses these limitations by enabling the inclusion of underrepresented tags and musical cultures in computational models, even with minimal annotated examples.

The multi-label nature of music tagging presents additional challenges for few-shot learning, as each music piece can be associated with multiple tags simultaneously (genre, instrumentation, regional style). Existing approaches to multi-label few-shot learning often introduce significant complexity to the training process through additional modules [101, 102] or complex episode formation [105]. By developing more streamlined multi-label few-shot learning methods, we can improve the versatility of computational music analysis, enabling the representation of diverse musical characteristics that might otherwise be excluded due to data limitations.

Together these approaches, transfer learning and few-shot learning, provide complementary tools for cross-cultural music analysis, offering different strategies for addressing the challenges of learning across diverse musical traditions with varying data availability. This combined approach aligns with our goal of developing more inclusive and adaptive music representation learning methods that can effectively capture the rich diversity of global musical expressions.

4.2 Cross-Cultural Transfer Learning: Methodology

Building upon the motivations outlined in the previous section, we now detail our systematic approach to investigating knowledge transfer between musical traditions. This section describes

our experimental design, dataset selection, model architectures, and transfer learning methodology. The implementation of our work is available online.[1]

4.2.1 Experimental Setup

We selected six datasets representing three distinct geographic regions, with each region represented by two corpora. These datasets were described in detail in Chapter 2 and are briefly summarized here:

- **Western music**: MagnaTagATune [126] (25,000+ recordings, 210 hours, 50 most popular tags) and FMA-medium [114] (25,000 tracks, 208 hours, 20 hierarchical genre tags)
- **Eastern Mediterranean music**: Lyra (see Chapter 3; 1,570 pieces, 80 hours, 30 tags related to genre, place, and instruments) and Turkish-makam [84, 85] (5,297 recordings reduced to 215 hours, 30 tags related to makam, usul, and instruments)
- **Indian classical music**: Hindustani [19] (1,204 recordings reduced to 206 hours, 20 tags related to raga, tala, instruments, and form) and Carnatic [19] (2,612 recordings reduced to 218 hours, 20 tags with similar categories)

To ensure consistency, we balanced the datasets to have similar durations (except for Lyra) by setting maximum duration limits for recordings in the larger collections. We selected the top 50 tags for MagnaTagATune, 30 for Lyra and Turkish-makam, and 20 for the rest of the datasets. A detailed list of the most frequent tags per dataset can be seen at the Appendix A.

4.2.2 Models

We employed three model architectures that represent different approaches to music audio representation learning, all using mel-spectrograms as input but with varying architectural paradigms:

- **VGG-ish**: A 7-layer CNN with 3×3 convolution filters and 2×2 max-pooling, followed by two fully-connected layers, processing 3.69-second audio chunks [129].
- **Musicnn**: A music-inspired convolutional model with specialized vertical and horizontal filters in its first layer to capture timbral and temporal features respectively, processing 3-second audio chunks [68].
- **Audio Spectrogram Transformer (AST)**: An attention-based model that splits mel-spectrograms into 16×16 patches, projects them to embeddings, and processes them through a Transformer encoder, with an input length of 8 seconds [39].

These diverse architectures allow us to investigate whether patterns of cross-cultural knowledge transfer are consistent across different model designs or are model-dependent.

4.2.3 Transfer Learning Approach

The purpose of transfer learning is to improve the performance of models on target domains by transferring knowledge from different but related source domains [89]. In our study, we utilized parameter sharing, a model-based transfer learning technique [92], where a network trained on a

[1] https://github.com/pxaris/ccml

Model	**VGG-ish**		**Musicnn**		**AST**	
Metric / Dataset	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
MagnaTagATune	0.9123	0.4582	0.9019	0.4333	**0.9172**	**0.4654**
FMA-medium	**0.8889**	0.4949	0.8766	0.4473	0.8886	**0.5024**
Lyra	0.8097	0.4806	0.7391	0.4042	**0.8476**	**0.5333**
Turkish-makam	**0.8696**	0.5639	0.8505	0.5299	0.8643	**0.5669**
Hindustani	**0.8477**	**0.6082**	0.8471	0.6016	0.8307	0.5786
Carnatic	0.7392	0.4278	0.7496	0.4182	**0.7706**	**0.4394**

Table 4.1. Single-Domain Auto-Tagging Performance. ROC-AUC and PR-AUC scores of models trained and evaluated on the same musical tradition.

source task shares its parameters with a target network, which is then fine-tuned on the target task.

Our transfer learning methodology involved the following steps:

1. Train each model architecture on each single dataset to establish baseline performance for within-domain learning
2. For each source-target domain pair, initialize the target model with the parameters of the source-trained model
3. Apply two fine-tuning strategies: (a) fine-tuning only the output layer, which tests the direct transferability of learned representations, and (b) fine-tuning the whole network, which allows for more adaptation to the target domain
4. Evaluate performance on the target dataset using ROC-AUC and PR-AUC metrics
5. Compare cross-domain transfer performance to single-domain baseline performance
6. Aggregate results across all models and fine-tuning strategies to identify robust patterns of cross-cultural knowledge transfer

Following the domain adaptation literature [91, 147], we hypothesized that transfer learning performance would correlate with the similarity between musical traditions, with better transfer between more similar domains. By systematically evaluating all possible source-target pairs across our datasets, we could analyze which musical traditions show stronger knowledge transferability, potentially reflecting underlying similarities in musical characteristics.

This comprehensive approach allows us to investigate both the potential of Western-trained models when applied to different musical cultures and the capability of models trained on specific non-Western traditions to provide meaningful representations for cross-cultural analysis.

4.3 Cross-Cultural Transfer Learning: Experiments

As already mentioned, we use mel spectrograms as the input of all our models. In order to convert the audio recordings of the datasets to this representation, we use Librosa [148] to re-sample them to 16 kHz sample rate. Then, 512-point FFT with a 50% overlap is applied, the maximum frequency is set to 8 kHz and number of Mel bands to 128. Our intention, in this study, is not

the optimization of the performance of the single-domain tasks but rather studying the knowledge transfer across the domains. So, we keep our training setup as close as possible to the literature, at each single domain task, in order to have a sanity check for the implementation.

For VGG-ish and Musicnn models, we use a mixture of scheduled Adam [149] and stochastic gradient descent (SGD) for the optimization method, identical to what the authors at [129] have used. The batch size is set to 16 and the learning rate to $1e-4$ for both models while the maximum number of epochs are 200 for VGG-ish and 50 for Musicnn. With regards to the AST model, we follow the setup proposed in [39], namely batch size 12, Adam optimizer, learning rate scheduling that begins from $1e-5$ and is decreased by a factor of 0.85 every epoch after the 5th one as well as pre-trained on Imagenet Transformer weights.

All our models accept a fixed size audio chunk at their input but need to predict song-level tags. During the evaluation phase, we aggregate the tag scores across all chunks by averaging them to acquire the label scores for the whole audio. We use the area under receiver operating characteristic curve (ROC-AUC), a widely used evaluation metric on multi-label classification problems and the area under precision-recall curve (PR-AUC), a suitable metric for unbalanced datasets [150].

During transfer learning, we initialize all parameters of the target model, except for the output layer, from each source dataset and (i) allow only the output layer to be trained and (ii) train the whole network. In both settings, we use the same hyper-parameters and evaluation procedure with the single-domain setups across all datasets for each model architecture.

4.4 Cross-Cultural Transfer Learning: Results

The performance of the three models on all single-domain tasks can be seen in Table 4.1. The performance of the Musicnn and VGG-ish models on MagnaTagATune is similar to the reported metrics in [129], which indicates the validity of our implementation. In general, the AST model shows the best performance followed by VGG-ish and then Musicnn. This result should not be taken into account solidly, because no hyper-parameter tuning has been taken place for each domain and in order to keep the duration of the training to less than 24 hours for each task, the number of epochs for Musicnn was significantly less than VGG-ish. On the other hand, one should consider that the AST [39] and VGG-ish [129] models may, indeed, perform better for limited time resources.

In Table 4.2, one can see the ROC-AUC scores in all single-domain and cross-domain setups. The rows are the source datasets while the columns are the target datasets. A sub-table is constructed for each model architecture and for a transfer from domain A to B, the result of the fine-tuning of only the output layer ('output') as well as all the layers ('all') are reported. The single-domain setup is when source and target is the same dataset and, thus, only training of the whole network has meaning. The table is better parsed column-wise, e.g., by inspecting the results of VGG-ish model on MagnaTagATune when transferring knowledge from the other domains at the upper-left pair of columns in the table.

In order to aggregate all the cross-domain knowledge transfers, we follow the subsequent procedure: for each target task that consists of a specific model, target dataset and fine-tuning method, min-max normalization is applied to the $N-1$ transfer learning results, where N is the number of all datasets. The previous step leads to the construction of $M \times F$ matrices, M the number of the models and F the number of fine-tuning methods, where rows are the source domains, columns the target domains and diagonal elements are empty. Each cell has a value in the range $[0, 1]$, as a result of the normalization step, while the value 1 corresponds to the knowledge transfer that led to the best performance in the target domain. By calculating the element-wise mean of the

Target domain	**MagnaTag-ATune**		**FMA-medium**		**Lyra**		**Turkish-makam**		**Hindustani**		**Carnatic**	
trainable layer(s) / Source domain	output	all	output	all	output	all	output	all	output	all	output	all
VGG-ish												
MagnaTagATune	-	91.23	88.11	**92.39**	74.69	**85.40**	76.79	86.84	76.09	85.04	67.19	**74.71**
FMA-medium	85.82	**91.29**	-	88.89	68.56	84.04	75.40	**87.78**	75.77	84.39	67.03	74.56
Lyra	84.34	90.93	82.84	92.10	-	80.97	76.98	87.21	77.41	84.24	67.30	73.52
Turkish-makam	85.19	90.90	84.41	91.74	70.93	82.38	-	86.96	77.54	**85.32**	67.16	73.50
Hindustani	84.24	91.02	83.83	91.91	66.27	79.71	77.25	87.63	-	84.77	66.72	74.63
Carnatic	84.18	91.00	82.62	91.73	61.59	76.72	77.07	87.40	78.19	84.81	-	73.92
Musicnn												
MagnaTagATune	-	90.19	87.34	**91.03**	71.79	78.74	74.72	**85.96**	75.87	84.18	66.12	75.57
FMA-medium	85.52	**90.35**	-	87.66	65.94	77.59	75.51	85.13	73.16	**85.49**	66.38	75.77
Lyra	81.38	90.03	82.23	90.80	-	73.91	74.11	85.20	78.10	83.29	65.09	75.51
Turkish-makam	84.35	90.11	83.79	90.81	61.87	**79.83**	-	85.05	75.67	83.75	67.49	74.09
Hindustani	82.38	89.86	83.42	90.85	64.48	78.95	74.60	85.58	-	84.71	65.25	**76.95**
Carnatic	83.02	90.05	82.78	90.74	61.83	77.92	75.09	85.43	75.34	84.19	-	74.96
AST												
MagnaTagATune	-	**91.72**	89.25	91.99	75.68	83.77	76.28	87.20	74.67	**86.57**	66.03	75.43
FMA-medium	88.63	91.62	-	88.86	65.72	82.17	76.37	**87.43**	74.51	85.76	67.33	75.98
Lyra	87.49	91.44	87.44	**92.43**	-	**84.76**	77.08	86.80	72.24	83.73	68.47	76.59
Turkish-makam	87.33	91.40	86.31	91.95	72.70	77.95	-	86.43	70.13	83.56	67.10	75.23
Hindustani	87.40	91.35	87.11	92.26	71.74	84.60	75.70	86.90	-	83.07	67.75	75.85
Carnatic	87.42	91.45	86.83	91.75	63.33	81.44	76.87	87.14	74.11	82.91	-	**77.06**

Table 4.2. Cross-Domain Transfer Learning Performance. ROC-AUC scores (%) when applying transfer learning using three model architectures. Rows are the source domains and columns the target domains. After initializing the network with the parameters of the trained (at the source dataset) model, fine-tuning on the output layer as well as on the whole network is applied. The diagonal values (under the "all" columns) correspond to the respective single-domain models (no transfer learning) where the experimentation with only the output layer trainable has no meaning.

produced $M \times F$ matrices, we reach to the result that can be seen in Figure 4.2.

4.5 Cross-Cultural Transfer Learning: Analysis and Discussion

The results indicate that knowledge transfer both from Western to non-Western cultures and the opposite can be beneficial when deep learning models are used to perform automatic music tagging. Indeed, by inspecting Table 4.2, the general take-home message one should acquire is that regardless of the model architecture, all datasets have the potential to contribute as a source to a target domain by providing their deep audio embeddings. To investigate how valuable knowledge transfers from widely used datasets to non-Western music cultures can be, we focus on the last four datasets, i.e., the last eight columns of the table, and parse the two first rows, corresponding to MagnaTagATune and FMA datasets, at each model architecture. For instance, we notice that for Lyra, when Musicnn is used and fine-tuning only of the output layer is applied, the model coming

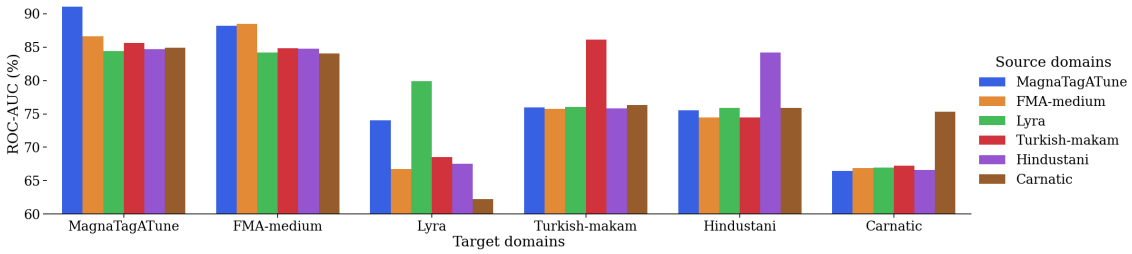

Figure 4.1. Cross-Domain Transfer Performance. Average ROC-AUC scores across three models for all cross-domain transfers with output layer fine-tuning, with highest bars in each group representing single-domain baseline performance.

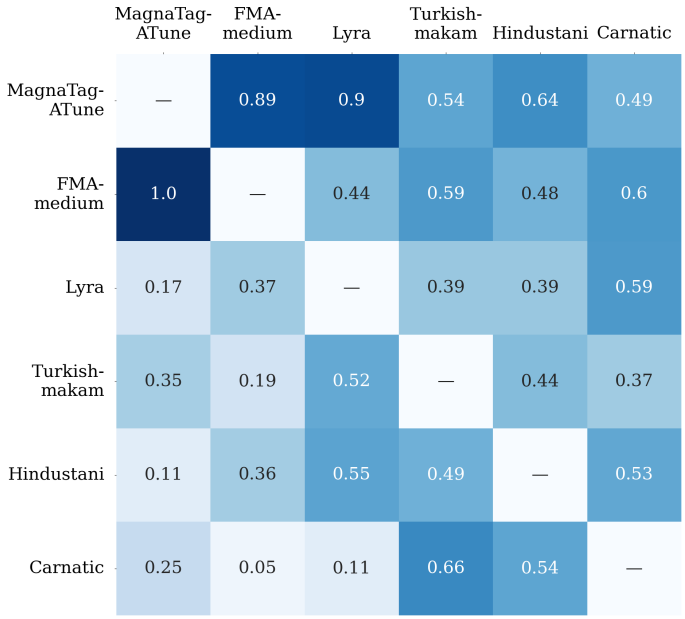

Figure 4.2. Cross-Cultural Transfer Learning Patterns. Heatmap of normalized knowledge transfer between source datasets (rows) and target datasets (columns), normalized and averaged across all models and fine-tuning methods to reveal cultural transferability.

from MagnaTagATune has the greater ROC-AUC score, namely 71.79%. Additionally, the AST model trained on the FMA-medium dataset, outperforms the others when totally fine-tuned to the Turkish-makam dataset, scoring 87.43%.

In order to study the inverse transfer direction, we center our interest to the first four columns of the entire table. Even though MagnaTagATune and FMA are almost always the best source for each other, the deep audio embeddings provided by the other datasets achieve competitive performance. For example, when MagnaTagATune is the target domain and fine-tuning is restricted to the output layer of the network, we observe that transferring from Turkish-makam leads to a performance that is comparable to the best source (FMA-medium) for all models.

By considering all cross-domain knowledge transfers, one can specify the best candidate to provide a trained model, with a specific architecture, for each target dataset. We, thus, notice that the model that is transferred from Hindustani outperforms the others at the Carnatic dataset, when fine-tuning on the whole Musicnn architecture is applied. A holistic picture of the cross-cultural music transfer learning is depicted in Figures 4.1 and 4.2.

In Fig. 4.1 the scores of all cross-domain transfers when fine-tuning the output layer, can be

seen, averaged across the three models. The uniformity of the performances of different sources at each target dataset can be examined. We, thus, recognize that the most unbalanced performances are spotted on the Lyra target domain, a result that is probably related to the smaller size of this dataset compared to the others. By exploring Fig. 4.2 in a column-wise fashion, we observe that for MagnaTagATune as the target domain, FMA-medium is the best source with a value equal to 1. This means that in all transfer learning setups, this source performed better than the others in this domain.

Both figures show that MagnaTagATune and FMA-medium perform consistently well across the domains, something that possibly indicates their appropriateness for the auto-tagging task. However, as we move to the Eastern cultures, we notice that their contribution is somehow decreased and other domains tend to contribute similarly or even more in those targets. The values at Fig. 4.2 should not be considered solidly as similarity metrics between the domains because other factors may also affect the results we notice. It is, although, a first step towards studying different music cultures using deep learning methods.

4.6 Label-Combination Prototypical Networks for Few-Shot Learning

Having explored cross-cultural transfer learning as one approach to learning across musical traditions, we now turn to the challenge of learning from limited examples, a critical issue for many world music contexts where annotated data is scarce. While transfer learning leverages knowledge from data-rich domains, it still requires sufficient examples of the target tags. For underrepresented tags or niche musical traditions, we need methods that can learn effectively from just a few examples.

In this section, we present Label-Combination Prototypical Networks (LC-Protonets), a novel approach designed specifically for multi-label few-shot learning scenarios. We first describe the foundation of our approach, Prototypical Networks, and their adaptation to multi-label settings, before introducing our proposed method. The implementation is available in an online repository.[2]

4.6.1 Prototypical Networks

Prototypical networks [95] are widely used in few-shot learning (FSL) and function by computing a prototype for each class, which represents the average embedding of the support items belonging to that class. Let S denote the support set, consisting of $N \times K$ examples, where N is the number of unique classes (referred to as the N-way) and K is the number of examples per class (referred to as the K-shot). The prototype for a class c, denoted as $\mathbf{p}_c$, is computed as the mean of the embedded support examples for that class:

$$\mathbf{p}_c = \frac{1}{K} \sum_{(\mathbf{x}_i, y_i) \in S} f_\theta(\mathbf{x}_i) \cdot \mathbb{1}_{y_i = c}, \tag{4.1}$$

where $f_\theta(\mathbf{x}_i)$ represents the embedding of input $\mathbf{x}_i$ through a mapping model, and $\mathbb{1}_{y_i=c}$ is an indicator function that equals 1 if the label y_i of $\mathbf{x}_i$ belongs to class c.

Once the prototypes are computed, a query set Q consisting of unseen examples is used to test the model. Each query item is classified based on the similarity to the prototypes, typically using

[2] https://github.com/pxaris/LC-Protonets

Euclidean or cosine distance. Specifically, the query sample $\mathbf{x}_q \in Q$ is assigned to the class whose prototype is the closest in the embedding space:

$$\hat{y}_q = \arg\min_c d(f_\theta(\mathbf{x}_q), \mathbf{p}_c), \tag{4.2}$$

where d refers to the chosen distance function. During training, cross-entropy loss and a Softmax function over the computed distances are used to learn the embeddings.

Episodic learning: Prototypical networks are trained using episodic learning, progressing through N-way K-shot episodes. In each episode, N classes are randomly sampled, and K support examples are drawn for each class to form the support set S. The query set Q consists of additional examples drawn from the same N classes. The model computes the prototypes from the support set, and the loss is calculated based on the classification accuracy of the query set. This episodic approach encourages the model to generalize better in few-shot settings by simulating small training tasks during learning.

Extension to multi-label setting: We adopt the term "ML-PNs" (multi-label Prototypical Networks) to refer to the extension of Prototypical Networks for the multi-label setting. This method follows the extension published in [104], where it is referred to as "Baseline"; however, we prefer a more explicit name here to enhance clarity. In this setting, where each sample may belong to multiple classes, each support item contributes to multiple prototypes. Let $\mathbf{y}_i$ be the set of labels for a given sample $\mathbf{x}_i$. For each label $y_{i,j} \in \mathbf{y}_i$, where j ranges from 1 to the number of classes N, the embedding $f_\theta(\mathbf{x}_i)$ is used to update the prototype corresponding to $y_{i,j}$. Consequently, the prototype for each class c is computed by averaging the embeddings of all support examples that belong to class c, even if they have additional labels.

In this setting, the Softmax function is replaced by a Sigmoid function, allowing the model to predict multiple labels for each query item. Binary cross-entropy loss is then used to optimize the model:

$$\mathcal{L} = -\sum_{q \in Q} \sum_c y_{q,c} \log(\hat{y}_{q,c}) + (1 - y_{q,c}) \log(1 - \hat{y}_{q,c}), \tag{4.3}$$

where $y_{q,c}$ represents the true label for class c for query q, and $\hat{y}_{q,c}$ is the predicted probability for that class.

4.6.2 LC-Protonets

Adapting few-shot learning to the multi-label regime presents a significant challenge, particularly because classes are correlated and each sample may belong to multiple classes. To address this, we propose Label-Combination Prototypical Networks (LC-Protonets), an approach to multi-label classification that extends Prototypical Networks in a simple yet effective way.

We consider multi-label classification as a problem where every combination of labels is a descriptive label. These combinations are all subsets of the label sets found in the support data, including the full label sets themselves. Hence, a support item $(\mathbf{x}_i, \mathbf{y}_i)$ with $\mathbf{y}_i = \{A, B, C\}$, defines and contributes to all the label combinations derived from the power set $\mathcal{P}$ of $\mathbf{y}_i$, excluding the empty set:

$$\begin{aligned} \mathcal{P}(\{A, B, C\}) = \{&\{A\}, \{B\}, \{C\}, \{A, B\}, \\ &\{A, C\}, \{B, C\}, \{A, B, C\}\}. \end{aligned} \tag{4.4}$$

Figure 4.3 illustrates an indicative example. The support set S consists of four items, each

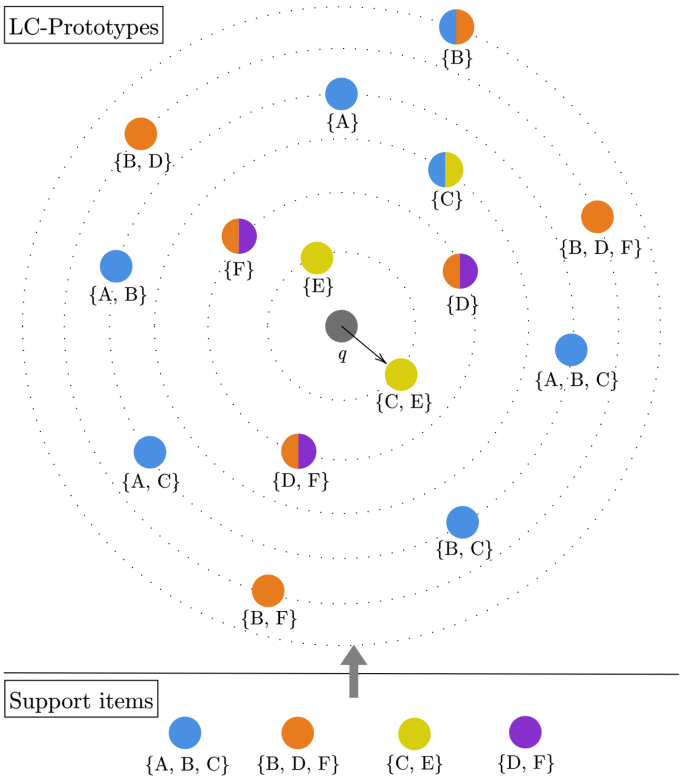

Figure 4.3. Label Combination Prototype Formation. Visualization of support set items (bottom) and derived LC-classes (top), with concentric circles showing equidistant LC-Prototypes representations from a query item q in the embedding space.

associated with a distinct set of labels. The set of label combinations, henceforth referred to as *LC-classes*, is defined as the union of the power sets of each $\mathbf{y}_i$ in the support set. Each LC-class is represented by an *LC-Prototype* (LCP), whose representation is computed by averaging the embeddings of the support items that include the corresponding LC-class in their power set, as shown in the color-coded example in the figure.

This approach addresses the multi-label classification problem as a mixture of few-shot and zero-shot learning scenarios. For instance, in Figure 4.3, the $\{B, D\}$ LC-class does not directly correspond to any support item, but its representation is inferred from items (one in this example) that include it in their label power sets. This introduces a zero-shot learning aspect. Meanwhile, the $\{B, D, F\}$ class has one corresponding item in the support set, enabling a few-shot scenario.

For a query item $q \in Q$, the distances to all LCPs are computed. Figure 4.3 provides a conceptual 2D representation of the embedding space, where concentric circles indicate equal distances from q to different LCPs. In the actual space, the LCPs with identical representations will occupy the same point.

In cases where a query item has equal distances to multiple LCPs, the one representing the largest number of labels is selected. That way, hierarchical relationships and strong correlation between the labels are supported. In the depicted example, both the $\{E\}$ and $\{C, E\}$ LCPs have the minimum distance to q, and the query is assigned to the $\{C, E\}$ LC-class. If E is hierarchically subordinate to C, the LCPs for $\{E\}$ and $\{C, E\}$ will share the same representation, but the model will consistently select the $\{C, E\}$ class. Even if C and E are not part of a formal hierarchy but they co-occur in the same support items, it is rational to assume strong correlation between them

and, again, assign the query to the $\{C, E\}$ class.

Training phase: Training is conducted using an episodic "N-way K-shot" approach. Here, N refers to the number of active labels (i.e., the number of singleton LC-classes in an episode) and K represents the number of items supporting each singleton LC-class. To prevent oversampling, an item sampled for a singleton class is also counted for any other active classes it belongs to, ensuring that the number of items for each class stays close to K.

Given the support items $(\mathbf{x}_i, \mathbf{y}_i) \in S$, the set of all LC-classes L is computed as:

$$L = \bigcup_{(\mathbf{x}_i, \mathbf{y}_i) \in S} \mathcal{P}(\mathbf{y}_i), \tag{4.5}$$

where $\mathcal{P}(\mathbf{y_i})$ is the power set of the labels of the i-th support item, excluding the empty set. The total number of LC-classes is given by the cardinality $|L|$ of the computed set[3].

We denote an LC-class as L_j, where $j = 1, 2, ..., |L|$. For each LC-class, one or more support items include it in the power set of their labels, forming the set $S_j \subseteq S$, defined as: $S_j = \{(\mathbf{x}_i, \mathbf{y}_i) \in S \mid L_j \in \mathcal{P}(\mathbf{y}_i)\}$.

The LCP representation $\mathbf{p}_j$ for the corresponding class is computed by averaging the embeddings of the items in S_j:

$$\mathbf{p}_j = \frac{1}{|S_j|} \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in S_j} f_\theta(\mathbf{x}_i), \tag{4.6}$$

where f_θ is the embedding mapping model with θ trainable parameters.

In each episode, a specified number of query items for each active label is sampled to form the query set Q. Given a query item $\mathbf{x}_i$ with a label set $\mathbf{y}_i$, its initial multi-hot label vector $\mathbf{z}_i \in \{0, 1\}^N$ is constructed such that $\mathbf{z}_i(k) = 1 \quad \text{if} \quad k \in \mathbf{y}_i, \forall k = 1, 2, \ldots, N$. The expanded multi-hot vector $\mathbf{z}_{mH_i} \in \{0, 1\}^{|L|}$ is then constructed by assigning a value of 1 to each of the item's LC-classes: $\mathbf{z}_{mH_i}(j) = 1 \quad \text{if} \quad L_j \in \mathcal{P}(\mathbf{y}_i), \quad \forall j = 1, 2, ..., |L|$. The loss function is based on the binary cross-entropy:

$$\begin{aligned} Loss(\mathbf{x}_i, \mathbf{z}_{mH_i}, \mathbf{p}) = -\mathbf{z}_{mH_i} \log(\sigma(-d(f_\theta(\mathbf{x}_i), \mathbf{p}))) + \\ (1 - \mathbf{z}_{mH_i}) \log(1 - \sigma(-d(f_\theta(\mathbf{x}_i), \mathbf{p}))), \end{aligned} \tag{4.7}$$

where d is the distance function, σ the sigmoid function and $\mathbf{p}$ the LCPs representations. We minimize the loss for all items in the query set Q.

Inference phase: At the inference phase, the support set S is created following a similar "N-way K-shot" setup used during training. The LC-classes L and their corresponding LCPs representations $\mathbf{p}_j$ (for $j = 1, 2, ..., |L|$) are computed. For a query instance, the distances to all LCPs are calculated, and the instance is assigned to the LC-class represented by the nearest LCP:

$$\hat{\mathbf{y}}_i = \arg \min_{L_j \in L} d(f_\theta(\mathbf{x}_i), \mathbf{p}_j). \tag{4.8}$$

It is important to note that the training phase of LC-Protonets closely follows the extension of Prototypical Networks [95] for the multi-label setting. However, in the latter, only singleton LC-classes are considered in L, making it a special case of the LC-Protonets approach. Another difference lies in the inference process, where the probabilities after a Sigmoid layer have to be utilized for classification as opposed to the direct approach adopted by the proposed method. LC-Protonets transforms the multi-label task to a single-label problem, where every combination of

[3]The scalability issues of the method in terms of LC-classes are being discussed in detail in Section 4.8.3

dataset	# recordings	# total tags	i^{th}: f(%)
MagnaTagATune	25863	188	50^{th}: 1.89%
FMA-medium	25000	151	20^{th}: 2.68%
Lyra	1570	146	30^{th}: 4.78%
Turkish-makam	5297	217	30^{th}: 2.83%
Hindustani	1204	273	20^{th}: 2.49%
Carnatic	2612	283	20^{th}: 2.03%

Table 4.3. Dataset Statistics and Tag Distribution. Number of recordings, total tags, and the relative frequency of the i^{th} most frequent (and last well-represented) tag for each dataset.

labels L_j is a descriptive label.

4.7 LC-Protonets: Experimental Design

4.7.1 Datasets and metrics

We incorporate a range of datasets from both mainstream and world music traditions for our study. For Western music, we use the MagnaTagATune dataset [126], a standard for auto-tagging, and the medium version of FMA [114]. For world music, we utilize the Lyra dataset (Chapter 3), along with the three datasets from the CompMusic Corpora [33]: Turkish-makam corpus [84, 85], Hindustani and Carnatic [19].

ML-FSL task	**5-way 3-shot**				**15-way 3-shot**			
classes type	*Base*		*Novel*		*Base*		*Novel*	
method / metric	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
ML-PNs	$\mathbf{65.21}_{4.33}$	$66.12_{4.41}$	$46.32_{3.88}$	$45.92_{3.53}$	$39.31_{1.66}$	$44.23_{2.11}$	$21.45_{1.3}$	$21.02_{1.22}$
One-vs.-Rest	$64.69_{4.19}$	$65.84_{4.16}$	$42.69_{3.53}$	$42.5_{3.4}$	$35.44_{1.64}$	$39.4_{1.79}$	$18.8_{1.45}$	$18.56_{1.41}$
LC-Protonets (ours)	$62.6_{4.93}$	$\mathbf{66.26}_{4.87}$	$\mathbf{47.89}_{6.69}$	$\mathbf{49.34}_{6.19}$	$\mathbf{42.84}_{2.71}$	$\mathbf{56.28}_{2.86}$	$\mathbf{28.5}_{3.61}$	$\mathbf{31.37}_{3.74}$
ML-FSL task	**30-way 3-shot**				**45-way 3-shot**		**60-way 3-shot**	
classes type	*Base*		*Base & Novel*		*Base & Novel*		*Base & Novel*	
method / metric	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
ML-PNs	$29.84_{1.14}$	$32.57_{1.2}$	$24.61_{0.83}$	$29.26_{1.27}$	$19.74_{0.66}$	$23.05_{1.02}$	$17.49_{0.62}$	$19.45_{0.67}$
One-vs.-Rest	$25.64_{1.21}$	$27.81_{1.35}$	$21.74_{0.98}$	$25.36_{1.29}$	$17.06_{0.76}$	$19.58_{1.0}$	$14.79_{0.68}$	$16.17_{0.86}$
LC-Protonets	$\mathbf{36.77}_{2.44}$	$\mathbf{50.77}_{1.79}$	$\mathbf{31.31}_{2.14}$	$\mathbf{52.65}_{1.95}$	$\mathbf{28.09}_{1.76}$	$\mathbf{50.28}_{2.06}$	$\mathbf{28.45}_{1.87}$	$\mathbf{46.48}_{1.82}$

Table 4.4. Performance Across ML-FSL Task Configurations. Macro-F1 (M-F1) and micro-F1 (m-F1) scores (%) with subscripted 95% confidence intervals for various "N-way" tasks, aggregated over all datasets and training setups with a consistent "3-shot" approach.

Table 4.3 provides details on the number of recordings, total tags, and the relative frequency of the last well-represented label for each dataset. We consider as well-represented the i most frequent tags for each dataset based on their successful inclusion in the supervised learning approach we followed in Section 4.2: 50 for MagnaTagATune, 30 for Lyra and Turkish-makam, and 20 for FMA-medium, Hindustani, and Carnatic. The data preparation for the automatic audio tagging task followed the same process described in Section 4.3. To use these datasets for few-shot learning, we

split the labels for training and testing, and we provide those splits to the public repository for reproducibility.

Since our method directly predicts a set of labels without assigning probabilities to individual labels, calculating metrics like Area Under the Curve (AUC) is not straightforward. Therefore, we selected Macro-F1 and Micro-F1 scores for evaluation. F1 score is the harmonic mean of the precision and recall scores and in its Macro- setting, the mean of the individual label scores is calculated. Micro-F1 computes metrics globally by aggregating true positives, false negatives, and false positives across all samples.

4.7.2 Backbone model

In few-shot learning, each sample is embedded into a feature space by a backbone model. Given the VGG-ish [151] model's ease of training and proven effectiveness in both supervised learning [46, 129] and ML-FSL tasks [104], we selected it as our backbone model. The architecture consists of a 7-layer Convolutional Neural Network (CNN) with 3×3 convolution filters and 2×2 max-pooling layers, followed by a couple of fully-connected layers. The model processes log mel-spectrograms as input features.

4.7.3 Comparative approaches

ML-PNs: Our method is compared to the extension of Prototypical Networks for multi-label classification, referred to as "ML-PNs". As described in Section 4.6.1, this approach uses a Sigmoid layer instead of Softmax for classification, and binary cross-entropy loss instead of categorical cross-entropy during training, compared to the standard single-label Prototypical Networks [95].

One-vs.-Rest: Another comparative approach is the "One-vs.-Rest" strategy introduced in [105]. In this method, the support set is divided into several subsets during training, where each subset focuses on a query's label along with $N-1$ other classes in an "N-way K-shot" format. The goal is to decompose the multi-label problem into multiple binary classification tasks.

4.7.4 Experimental setup

We split the labels of each dataset into training and testing sets. The training set is used both during the training phase of ML-FSL models and for pre-training the backbone VGG-ish model via supervised learning, while the testing set is used to form the novel classes for evaluation. For the audio recordings, we use the same splits as in previous studies [46, 129], with a split ratio of 0.7, 0.1, and 0.2 for the training, validation, and test sets, respectively.

To train the ML-FSL models, we employ three setups: (i) training from scratch with random weight initialization, (ii) full fine-tuning of a pre-trained backbone model, and (iii) fine-tuning only the last layer of the pre-trained backbone model. The model architecture remains the same across all setups, except in the fine-tuning cases where a VGG-ish model pre-trained on the well-represented tags is transferred as the backbone, excluding only the final classification layer.

In few-shot learning, *base* classes are those seen during training, while *novel* classes are unseen. We evaluate ML-FSL models on "Base", "Novel" and "Base & Novel" classes, with the latter including an equal mix of unseen and seen tags. This allows us to assess how well the model handles both seen and unseen classes during inference. Various values of N (the number of classes) are tested, while K is kept constant (3-shot) across the "N-way K-shot" ML-FSL tasks. Importantly, the same *base* classes used for training an ML-FSL model from scratch are also used for pre-training the

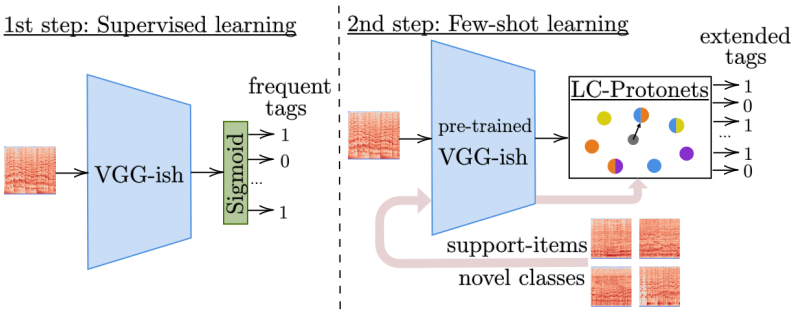

Figure 4.4. Two-Step Learning Framework. Process diagram showing supervised learning on well-represented tags in the first step, followed by LC-Protonets extension of the tag set using the previously trained model as backbone.

backbone model via supervised learning, ensuring that the *novel* classes remain unseen for models with a pre-trained backbone.

Cosine distance is used as the distance metric, and all methods are trained using episodic learning with a 10-way 3-shot setup. This setup is chosen to accommodate the low-resource nature of music data, as the absolute number of labels differs across domains. For instance, Hindustani, Carnatic, and FMA-medium datasets have only 20 labels in the training split. Additionally, selecting 3 examples per label allows under-represented labels to be included. 50 episodes are sampled for each epoch and 3 query items per label are utilized to compute the loss in each one of them.

The validation set is formed by holding out 5 classes from the training set during learning. The Adam optimizer [149] is used, and early stopping is applied based on the Macro-F1 score on the validation set. Regarding the input, the audio signal is sampled at 16 kHz, and a 512-point FFT with a 50% overlap is applied while the Mel bands are set to 128. During training, a random chunk of each audio recording is selected, while during testing, the average embedding of all chunks forms the representation of an instance.

4.7.5 Two-step learning method

In imbalanced datasets, it is common to encounter a large number of labels that occur infrequently, leading to a long-tailed label distribution. When training models using supervised learning, a threshold is often set to include only the most frequent categories. To address this limitation, we propose a two-step method that combines supervised and few-shot learning. Unlike ML-FSL setups that fine-tune a pre-trained backbone, this approach requires no fine-tuning.

As illustrated in Figure 4.4, the first step involves training a deep learning model on well-represented tags using supervised learning. The model is optimized with a Sigmoid classification layer and binary cross-entropy loss. In the second step, the pre-trained model is frozen and used as a backbone to map data samples into an embedding space defined by its penultimate layer. We extend the tag set and perform inference on any query item by applying LC-Protonets on top of the pre-trained model. Without additional training, LC-Protonets can classify previously unseen, under-represented labels, including those from the long tail of world music datasets, using just a few examples per label. In our experiments, we use 3 examples per label.

dataset	**MagnaTagATune**		**FMA-medium**		**Lyra**	
metric	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
method	*training from scratch*					
ML-PNs	$19.89_{0.72}$	$21.33_{1.03}$	$18.32_{0.54}$	$20.04_{0.6}$	$31.4_{1.28}$	$37.5_{1.71}$
One-vs.-Rest	$16.72_{0.86}$	$17.36_{1.13}$	$13.73_{0.84}$	$14.82_{0.51}$	$29.05_{0.41}$	$33.9_{1.32}$
LC-Protonets (ours)	$\mathbf{21.58}_{1.56}$	$\mathbf{29.49}_{2.12}$	$\mathbf{18.91}_{1.54}$	$\mathbf{34.75}_{1.73}$	$\mathbf{39.47}_{4.57}$	$\mathbf{59.82}_{2.76}$
method	*pre-trained backbone and full fine-tuning*					
ML-PNs	$25.0_{0.35}$	$27.63_{0.57}$	$22.08_{0.4}$	$23.71_{0.67}$	$35.64_{1.31}$	$41.72_{1.52}$
One-vs.-Rest	$19.82_{1.09}$	$20.96_{1.44}$	$18.7_{0.48}$	$19.64_{0.88}$	$29.02_{0.24}$	$33.36_{0.43}$
LC-Protonets	$\mathbf{33.66}_{1.41}$	$\mathbf{43.37}_{2.35}$	$\mathbf{33.37}_{0.98}$	$\mathbf{48.83}_{1.36}$	$\mathbf{45.29}_{2.42}$	$\mathbf{65.99}_{1.25}$
method	*pre-trained backbone and fine-tuning of the last layer*					
ML-PNs	$24.45_{0.59}$	$26.94_{0.91}$	$20.88_{0.37}$	$22.65_{0.46}$	$36.72_{0.91}$	$43.16_{0.95}$
One-vs.-Rest	$19.86_{2.31}$	$20.82_{2.38}$	$19.23_{0.44}$	$20.52_{0.64}$	$31.75_{1.18}$	$37.03_{1.65}$
LC-Protonets	$\mathbf{33.5}_{1.33}$	$\mathbf{43.27}_{1.98}$	$\mathbf{33.04}_{1.73}$	$\mathbf{48.68}_{1.94}$	$\mathbf{47.31}_{3.16}$	$\mathbf{68.58}_{1.01}$
method	*pre-trained backbone without any fine-tuning*					
ML-PNs	$13.62_{0.01}$	$14.3_{0.01}$	$10.58_{0.01}$	$11.18_{0.01}$	$28.93_{0.09}$	$33.25_{0.11}$
One-vs.-Rest	$13.62_{0.01}$	$14.31_{0.02}$	$10.58_{0.01}$	$11.19_{0.01}$	$28.9_{0.1}$	$33.22_{0.12}$
LC-Protonets	$\mathbf{33.52}_{1.19}$	$\mathbf{43.24}_{2.0}$	$\mathbf{33.73}_{1.27}$	$\mathbf{49.2}_{1.87}$	$\mathbf{47.32}_{3.76}$	$\mathbf{68.95}_{2.0}$
dataset	**Turkish-makam**		**Hindustani**		**Carnatic**	
metric	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
method	*training from scratch*					
ML-PNs	$20.29_{0.15}$	$22.12_{0.16}$	$18.16_{0.68}$	$23.89_{1.81}$	$20.12_{0.92}$	$30.11_{1.39}$
One-vs.-Rest	$20.3_{0.15}$	$22.12_{0.14}$	$17.86_{0.56}$	$22.77_{1.72}$	$\mathbf{20.23}_{1.53}$	$27.64_{1.07}$
LC-Protonets	$\mathbf{21.52}_{2.95}$	$\mathbf{37.41}_{2.68}$	$\mathbf{24.71}_{4.83}$	$\mathbf{50.82}_{3.49}$	$17.96_{0.47}$	$\mathbf{54.63}_{1.4}$
method	*pre-trained backbone and full fine-tuning*					
ML-PNs	$32.19_{1.43}$	$32.72_{1.46}$	$23.1_{0.53}$	$30.4_{0.92}$	$22.04_{0.79}$	$31.41_{2.38}$
One-vs.-Rest	$26.05_{1.42}$	$28.08_{1.63}$	$18.36_{0.93}$	$23.52_{1.66}$	$20.72_{0.44}$	$28.45_{0.97}$
LC-Protonets	$\mathbf{38.59}_{2.45}$	$\mathbf{57.31}_{2.18}$	$\mathbf{35.07}_{2.63}$	$\mathbf{59.03}_{2.33}$	$\mathbf{23.16}_{1.63}$	$\mathbf{63.69}_{2.57}$
method	*pre-trained backbone and fine-tuning of the last layer*					
ML-PNs	$28.52_{2.66}$	$30.73_{2.27}$	$22.84_{0.43}$	$31.16_{1.43}$	$21.38_{0.77}$	$29.46_{2.64}$
One-vs.-Rest	$28.95_{2.19}$	$30.77_{1.89}$	$20.18_{1.78}$	$26.29_{2.42}$	$20.73_{0.8}$	$28.44_{1.36}$
LC-Protonets	$\mathbf{38.52}_{2.08}$	$\mathbf{57.99}_{1.48}$	$\mathbf{34.64}_{2.13}$	$\mathbf{60.04}_{1.44}$	$\mathbf{23.25}_{0.69}$	$\mathbf{63.92}_{1.05}$
method	*pre-trained backbone without any fine-tuning*					
ML-PNs	$20.28_{0.1}$	$22.11_{0.12}$	$17.49_{0.12}$	$21.83_{0.15}$	$20.88_{0.07}$	$27.76_{0.05}$
One-vs.-Rest	$20.25_{0.06}$	$22.07_{0.05}$	$17.6_{0.12}$	$21.94_{0.18}$	$20.91_{0.05}$	$27.8_{0.07}$
LC-Protonets	$\mathbf{37.23}_{1.71}$	$\mathbf{56.83}_{0.83}$	$\mathbf{35.09}_{2.83}$	$\mathbf{59.69}_{1.81}$	$\mathbf{22.42}_{0.78}$	$\mathbf{63.66}_{1.32}$

Table 4.5. ML-FSL Performance Under Different Training Conditions. Macro-F1 (M-F1) and micro-F1 (m-F1) scores (%) with subscripted confidence intervals for a "30-way 3-shot" task across training scenarios: training from scratch, pre-trained backbone with full or partial fine-tuning, and no fine-tuning. Rows represent the multi-label few-shot learning methods, and columns correspond to the datasets.

4.8 LC-Protonets: Performance Evaluation

4.8.1 ML-FSL tasks

In Table 4.4, the aggregated results of the LC-Protonets method and the two comparative approaches are presented. These results were calculated by averaging performance across all datasets and training setups: from scratch, full fine-tuning, and fine-tuning of the last layer. Each experiment was run five times with different random seeds, and the 95% confidence intervals are reported. While label splits remained consistent across runs, different active classes were sampled at each epoch during training, and different support items were selected in each run. We present both macro-F1 and micro-F1 scores for different numbers of labels, ranging from 5 to 60. The evaluations were performed on "Base" and "Novel" classes for smaller numbers of classes, and on "Base & Novel" classes for larger numbers.

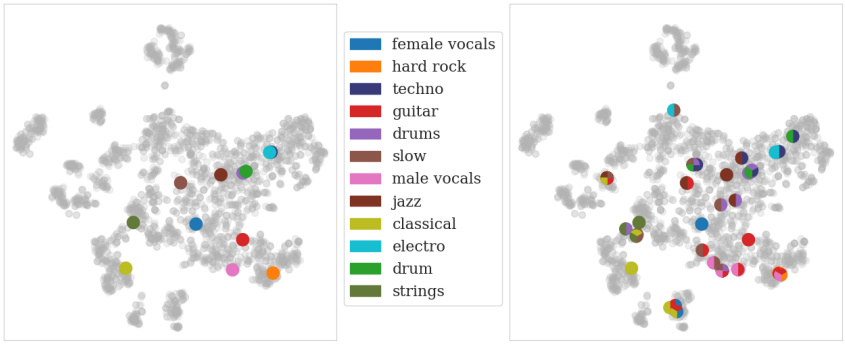

Figure 4.5. Prototype Embedding Visualization. t-SNE visualization of query items (in grey) and prototype embeddings (in distinct colors) for a "12-way 5-shot" ML-FSL task on the MagnaTagATune dataset; the left panel shows prototypes generated by the "ML-PNs" method (one per class), while the right panel displays those formed using the "LC-Protonets" method, where different colors within each prototype indicate the specific label combination it represents.

LC-Protonets outperformed other methods in nearly all tasks, except for the 5-way 3-shot task with base classes, where ML-PNs performed better in terms of macro-F1 score. In the 15-way 3-shot task, LC-Protonets showed superior performance on base classes and widened this gap further when novel classes were used. As the number of classes increased, LC-Protonets demonstrated substantially better performance compared to the other approaches.

In terms of confidence intervals, we observed wider ranges for few-shot conditions, such as the 5-way task, due to the random sampling of a small number of active classes in imbalanced datasets. Additionally, LC-Protonets' reliance on support set sampling for deriving LC-classes leads to wider confidence intervals compared to the other methods.

Figure 4.5 highlights the differences between the prototypes formed by ML-PNs and LC-Protonets. While ML-PNs create one prototype per class, LC-Protonets populate the embedding space with representations derived from the power sets of the support item labels. We believe this enhanced *positive sampling* of the feature space contributes to the significant performance improvement seen in the results.

Table 4.5 presents the results of the 30-way 3-shot task on "Base & Novel" classes for each dataset and training setup. When training from scratch, the LC-Protonets method showed improvement in all cases except for the macro-F1 evaluation on the Carnatic dataset, where One-vs.-Rest performed better. The difference between LC-Protonets and the comparative approaches was more evident

in the micro-F1 scores, as also noted in the aggregated results.

When using a pre-trained backbone model followed by full fine-tuning with episodic learning, the performance of all methods significantly improved across all datasets. However, the gap between LC-Protonets and the comparative approaches widened further. For instance, in MagnaTagATune, ML-PNs improved from 19.89% to 25.0%, while LC-Protonets increased from 21.58% to 33.66% in macro-F1 score. When only the last layer of the pre-trained model was fine-tuned, performance remained similar across datasets, with the exception of Lyra, where this approach led to slightly better results.

Finally, the last three rows of the upper and lower parts of Table 4.5 report the performance when a pre-trained backbone model was used without any fine-tuning. Interestingly, LC-Protonets maintained performance levels similar to those seen in the fine-tuning setups. This suggests that the method relies more on the quality of the representations provided by the backbone model than on episodic learning. By contrast, the performance of the comparative approaches dropped significantly, as they encountered challenges with multi-label classification without training, often assigning all labels to all samples. This can be seen in the very narrow confidence intervals of ML-PNs on the FMA-medium dataset, for example. More results, with regards to the proposed method and the comparative approaches, can be found at Appendix B.

Comparisons with state-of-the-art methods are not possible due to the lack of ML-FSL results for MIR and these datasets. Moreover, the literature commonly reports ROC-AUC for multi-label tasks instead of Macro-F1. State-of-the-art models also focus on top-N labels, unlike our method which targets under-represented classes.

dataset	**MagnaTagATune**				**FMA-medium**			
# tags: original / extended	50		80		20		40	
method / metric	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
VGG-ish	$26.74_{0.63}$	$42.29_{0.58}$	-	-	$36.90_{0.76}$	$59.61_{0.84}$	-	-
VGG-ish & LC-Protonets	$33.09_{0.83}$	$39.28_{1.77}$	$26.4_{0.26}$	$37.31_{0.47}$	$40.94_{2.0}$	$53.51_{0.73}$	$29.12_{1.44}$	$45.37_{1.71}$
dataset	**Lyra**				**Turkish-makam**			
# tags: original / extended	30		60		30		60	
method / metric	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
VGG-ish	$30.48_{1.23}$	$67.14_{1.17}$	-	-	$44.95_{0.82}$	$79.11_{0.98}$	-	-
VGG-ish & LC-Protonets	$47.32_{3.76}$	$68.95_{2.0}$	$46.05_{2.8}$	$69.03_{2.21}$	$37.23_{1.71}$	$56.83_{0.83}$	$30.07_{1.63}$	$56.22_{1.42}$
dataset	**Hindustani**				**Carnatic**			
# tags: original / extended	20		35		20		40	
method / metric	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
VGG-ish	$46.07_{1.12}$	$76.60_{1.29}$	-	-	$35.49_{1.54}$	$84.82_{1.71}$	-	-
VGG-ish & LC-Protonets	$40.69_{1.83}$	$64.38_{1.2}$	$31.33_{2.01}$	$58.38_{2.41}$	$32.1_{1.47}$	$64.84_{1.51}$	$18.13_{0.6}$	$64.25_{0.82}$

Table 4.6. Two-Step Learning Method Performance. Macro-F1 (M-F1) and micro-F1 (m-F1) scores (%) with subscripted 95% confidence intervals, comparing the "VGG-ish" model on well-represented tags against the "VGG-ish & LC-Protonets" method on both standard and extended tag sets.

4.8.2 Two-step learning method

The results of the proposed two-step learning method are shown in Table 4.6. For each dataset, two tag counts are used. The smaller number, such as 20 for FMA-medium, corresponds to the well-represented tags on which the VGG-ish model was trained, while the larger number, 40, represents

the extended tag set. The performance of both the "VGG-ish" model and the "VGG-ish & LC-Protonets" method on the well-represented tags is reported, and for the latter, its performance on the extended tag set is also included.

When examining the macro-F1 scores for both methods on the smaller set of tags, we observe similar performance across most datasets. The key architectural difference between the two approaches is the replacement of the VGG-ish Sigmoid classification layer with the LC-Protonets framework, which classifies an unknown sample to the label combination represented by the nearest LCP. The utilization of the LCPs offers a straightforward way to expand the number of labels. For instance, the tags in MagnaTagATune can be extended from 50 to 80, in Hindustani from 20 to 35, and doubled for the other datasets.

There is a relatively small drop in macro-F1 performance as the number of tags increases significantly. For example, in the Turkish-makam dataset, the macro-F1 score drops from 37.23% to 30.07% as the number of tags rises from 30 to 60. An exception is the Lyra dataset, where performance on the extended tag set remains nearly identical to the well-represented tags, likely due to stronger correlations between tags in Lyra compared to the other datasets.

4.8.3 Scalability

As the LC-classes are derived from the power sets of the sample labels, the number of LC-Prototypes increases significantly as the number of classes N grows. This results in a corresponding increase in inference time, as the distances from all LCPs must be computed for each query item. In Figure 4.6, the number of classes N is shown on the x-axis, while the left y-axis represents the number of LC-Prototypes, and the right y-axis shows the inference time per query item (in milliseconds), averaged across all datasets.

When we focus on the dashed blue line in the figure, showing the original method's inference time, we observe that when N increases from 20 to 30, the number of LCPs rises by a factor of about 15, from 487 to 7853, while inference time increases from 21 to 306 milliseconds. As the number of classes continues to grow, the number of LCPs increases substantially, reaching 53640 for 60 tags, and the inference time also rises to 2170 milliseconds.

Optimization approach: To address these scalability issues, we have developed an optimization that significantly improves inference efficiency while maintaining identical classification results. Our key insight is that multiple LC-classes often share identical LCP representations despite representing different label combinations. This occurs because the same set of support items contributes to multiple label combinations derived from their power sets. For example, if a support item with labels $\{A, B, C\}$ is the only item contributing to both $\{A, B\}$ and $\{B, C\}$ LCPs, these LCPs will have identical representations.

We exploit this redundancy by maintaining a dictionary structure that maps unique LCP representations to their corresponding sets of LC-classes:

$$\text{UniqueLCPs} = \{\mathbf{p}_m \mapsto \{L_j \mid \mathbf{p}_j = \mathbf{p}_m\}\}, \tag{4.9}$$

where $j = 1, 2, ..., |L|$ and $m = 1, 2, ..., M$ with M being the number of unique LCPs and $M \ll |L|$.

During inference, instead of computing distances between a query item and all possible $|L|$ LCPs, we only compute distances to the M unique LCP representations. For the nearest unique LCP, we then select the label combination with the maximum cardinality, consistent with our original approach.

As shown with the solid blue line in Figure 4.6, this optimization yields dramatic speed im-

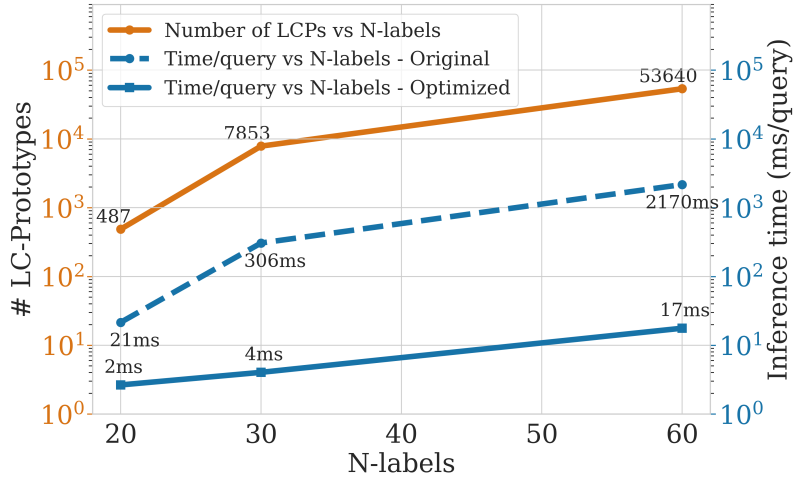

Figure 4.6. LC-Protonets Scalability Analysis. Relationship between number of labels (x-axis), LC-Prototypes (left y-axis), and inference time per item (right y-axis) on logarithmic scale. The dashed blue line shows the original method's inference time, while the solid blue line shows the optimized approach, demonstrating significant computational improvements.

provements: for datasets with 20 labels, inference time drops from 21ms to just 2ms (a 10× improvement), and for 60 labels, from over 2,000ms to only 17ms (over 100× improvement), all while producing identical classification results. We believe that this optimization addresses the primary scalability limitation of the method, making it practical for deployment across a wider range of application scenarios.

The inference process runs only during testing and not during model training. The average training time across all three methods was under an hour, with no significant differences between them. The trainable parameters amount to 3.66 million for training from scratch or full fine-tuning, and 262,000 for fine-tuning only the last layer. The experiments were conducted on an NVIDIA RTX A5000 GPU.

4.9 Conclusions

This chapter has explored two complementary approaches to address the challenge of developing computational models for analyzing diverse musical traditions: transfer learning across different musical systems and few-shot learning for scenarios with limited data. Both approaches contribute to our goal of creating more versatile computational methods while offering insights into how knowledge can be effectively shared across different musical traditions.

4.9.1 Cross-Cultural Transfer Learning

Our systematic investigation of knowledge transfer between musical traditions has yielded several important findings. First, we demonstrated that deep audio embedding models can benefit from knowledge transfer from Western to non-Western musical traditions and vice versa. This bidirectional transferability demonstrates mutual benefit between different musical systems and suggests that models trained on various musical traditions can contribute valuable knowledge to MIR systems.

By aggregating performance across three model architectures and different fine-tuning strate-

gies, we identified patterns of transferability that may reflect underlying similarities between musical cultures. These patterns could potentially be interpreted as computational similarity metrics between traditions, offering insights into musical relationships that reflect historical connections, geographic proximity, or parallel developments.

While Western datasets like MagnaTagATune and FMA-medium performed consistently well as source domains across various target traditions, we also observed that as we moved to Eastern Mediterranean and Indian musical traditions, other non-Western domains contributed similarly or sometimes more effectively to these targets. This suggests that the suitability of source domains varies across musical traditions, and that leveraging knowledge from diverse traditions may lead to more robust and broadly applicable computational models.

4.9.2 Label-Combination Prototypical Networks

To address the challenge of limited annotated data in world music collections, we introduced Label-Combination Prototypical Networks (LC-Protonets), a novel approach for multi-label few-shot learning. LC-Protonets consistently outperformed comparative approaches across diverse music datasets and ML-FSL tasks by creating prototypes for label combinations rather than individual labels.

Our experiments with different training setups revealed that utilizing pre-trained models as backbones significantly benefits all ML-FSL methods. Notably, LC-Protonets showed particular strength in using pre-trained embeddings even without fine-tuning, unlike comparative approaches. This enabled the development of a two-step learning method that can successfully expand a dataset's tag set by leveraging models pre-trained on well-represented tags, providing a practical pathway for including underrepresented musical characteristics in computational models.

Regarding the method's scalability issues, our optimized implementation addresses its computational complexity challenges by efficiently identifying unique prototypes, significantly reducing inference time while maintaining identical classification results. This optimization makes the approach practical for large label sets typically encountered in world music collections, though opportunities remain for further enhancing robustness against support set sampling variability in future work.

4.9.3 Synthesis and Future Directions

The two approaches explored in this chapter, transfer learning and few-shot learning, complement each other in addressing different aspects of learning across different musical traditions. Where transfer learning leverages knowledge from data-rich domains to enhance performance on common tasks with sufficient examples, few-shot learning enables effective learning even with minimal annotated data, particularly for underrepresented tags and traditions.

Interestingly, both approaches highlight the value of pre-trained models in cross-cultural music analysis, though they utilize these models differently. Transfer learning initializes models with parameters learned from a source domain, while LC-Protonets can directly leverage the embedding space of pre-trained models without additional training. This suggests that end-to-end deep learning models with minimal inductive bias can learn representations that transfer effectively across different musical traditions and learning paradigms.

The performance improvements observed when combining both approaches, using transfer learning to obtain better feature representations and few-shot learning to adapt these representations to new tags with limited examples, point toward integrated frameworks that could address

the full spectrum of data availability scenarios in multicultural music analysis.

Future research directions include exploring semantic similarities between labels across domains, incorporating additional datasets and model architectures, and investigating different tasks such as mode estimation that may reveal deeper cross-cultural musical connections. For few-shot learning, exploring different backbone architectures would further enhance the applicability of the LC-Protonets method to diverse musical contexts.

Together, these approaches advance computational methodologies for analyzing diverse musical traditions while facilitating cross-cultural comparison and knowledge transfer. By enabling effective learning across cultural boundaries and from limited examples, these methods enhance the versatility of music representation learning, making computational approaches more accessible for underrepresented musical traditions and characteristics.

The emergence of foundation models in music presents both new opportunities and challenges for multicultural representation learning. While the approaches developed in this chapter demonstrate effective strategies for working with conventional deep learning models, the next chapter investigates how these principles extend to large-scale foundation models and explores novel adaptation strategies that can enhance their cross-cultural capabilities while preserving their general musical knowledge.

Chapter 5

Foundation Models for Diverse Music Cultures

5.1 Motivation

The question of music's universality has long been debated among scholars. While certain musical elements appear to transcend cultural boundaries, musical traditions have evolved with distinctive characteristics and semantic content that reflect their cultural contexts [2, 5, 9, 10]. This tension between universal features and cultural specificity presents a complex challenge for computational music analysis, particularly as foundation models emerge as a transformative paradigm in artificial intelligence.

Foundation models have revolutionized multiple AI domains by learning general-purpose representations from large-scale data that can be adapted to diverse downstream tasks [106]. In music and audio, models like MERT [40], CLAP [41], and Qwen-Audio [42] have demonstrated impressive capabilities across various MIR tasks, from beat tracking to automatic tagging [43, 113]. These models implicitly claim a form of universality through their general-purpose nature, yet they have been predominantly trained on Western-centric data, raising critical questions about their ability to represent diverse musical traditions effectively.

The research presented in previous chapters has demonstrated both the potential and limitations of cross-cultural knowledge transfer in music analysis. Transfer learning revealed patterns of knowledge transferability between musical traditions, while few-shot learning provided strategies for addressing data scarcity in world music collections. Foundation models potentially offer more powerful general-purpose representations that could enhance cross-cultural music analysis, but their effectiveness across diverse traditions remains largely unexplored.

This chapter addresses this gap through two complementary investigations, drawing from our comprehensive evaluation of foundation models across world music corpora [48] and the collaborative development of CultureMERT for cross-cultural adaptation[1] [49]. First, Sections 5.2 through 5.4 present a systematic evaluation of foundation models across culturally diverse music corpora, assessing their cross-cultural capabilities under different resource constraints. By systematically evaluating these models' performance on Western popular, Greek traditional, Turkish makam, and Indian classical music traditions, we quantitatively assess their cross-cultural capabilities and contribute to broader discussions about the universality of musical representations. This evaluation employs three complementary methodologies that assess foundation models under different conditions: probing (using models as frozen feature extractors with trainable classifiers), supervised fine-tuning (adapting specific model layers to target domains), and multi-label few-shot learning (testing performance in low-resource scenarios common with world music collections).

[1] My contributions to this collaborative work focused on experimental design, cultural adaptation evaluation, and cross-cultural analysis frameworks.

Building on the insights from this evaluation, Sections 5.5 through 5.6 explore adaptation strategies to enhance the cultural inclusivity of foundation models. Despite their advances in music understanding, most existing foundation models have been trained primarily on Western-centric datasets, limiting their ability to represent diverse musical styles [13]. Many musical traditions, including Turkish makam, Indian classical, and Greek traditional music, feature distinctive melodic structures, modal systems, and rhythmic patterns that may not be adequately captured by Western-trained models [152–154].

This limitation has significant implications beyond academic interest. The inadequate representation of diverse musical traditions narrows the applicability of music foundation models for practical applications like region-specific recommendation systems [155] and cultural heritage preservation. It also overlooks the rich, culturally specific knowledge embedded in diverse musical traditions that could advance MIR research more broadly [43]. There is thus an urgent need to develop more inclusive and culturally aware computational models capable of generalizing beyond Western-centric traditions [44, 156].

To address these challenges, we introduce CultureMERT, a culturally adapted foundation model developed through continual pre-training (CPT), which has shown effectiveness in adapting large language models to new domains and languages [119, 157]. By enabling incremental adaptation without full retraining, CPT offers a computationally efficient pathway for enhancing cultural inclusivity while mitigating catastrophic forgetting [158]. We also explore task arithmetic [120] as an alternative approach, which combines domain-specific adaptations in weight space without requiring additional training or access to the original training data.

Together, these investigations contribute to our understanding of foundation models for music representation learning across cultures and advance toward more inclusive computational approaches to music analysis that respect and preserve the rich diversity of global musical expressions.

5.2 Multi-Method Evaluation Framework for Foundation Models

Having established the importance of evaluating foundation models across diverse musical cultures, we now present our comprehensive methodological framework for assessing these models' cross-cultural capabilities. This framework enables systematic comparison of state-of-the-art foundation models across Western and non-Western musical traditions under varying resource constraints, providing insights into both their strengths and limitations for cross-cultural music analysis.

Our methodological framework systematically evaluates whether foundation models can effectively represent musical characteristics across diverse cultural traditions. As shown in Figure 5.1, we employ three complementary methodologies: probing (*Prob.*), supervised fine-tuning (*SFT*), and multi-label few-shot learning (*ML-FSL*). Probing trains only an MLP classifier on frozen model representations, while SFT makes the model's last layers trainable alongside the MLP. ML-FSL extracts representations from three contexts, i.e., pretrained model (PT), trained probing model (*Prob.*) and fine-tuned model (SFT) to evaluate performance on extended tag sets under data scarcity conditions.

The implementation is being made available[2] for reproducibility and to promote research on world music.

[2] https://github.com/pxaris/FM-music-tagging

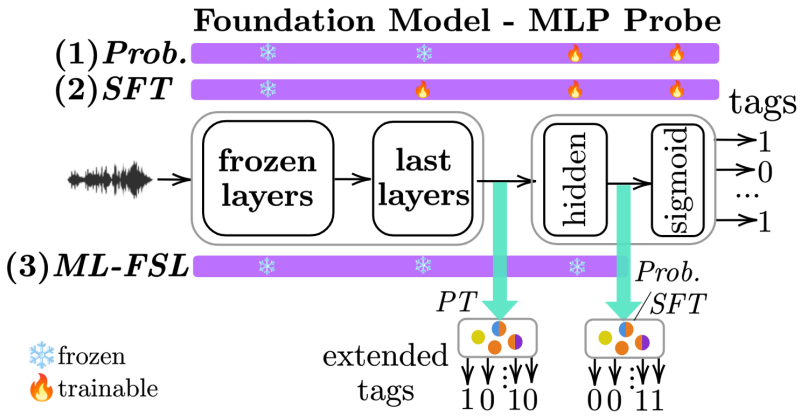

Figure 5.1. Multi-Method Evaluation Framework. Architectural overview showing three methodologies: (1) Probing (*Prob.*), (2) Supervised Fine-Tuning (*SFT*), and (3) Multi-Label Few-Shot Learning (*ML-FSL*). The diagram indicates feature extraction points used by *ML-FSL* from either Pre-Trained (*PT*), trained *Prob.* or *SFT* models.

5.2.1 Models

For our evaluation, we selected five state-of-the-art audio models spanning different architectures, pre-training approaches, and parameter scales:

MERT. We evaluate two variants of MERT [40]: MERT-95M[3] and MERT-330M[4] with 95M and 330M parameters respectively. These transformer-based models employ masked acoustic modeling, using an acoustic and a musical teacher, during pre-training. MERT-95M consists of 12 layers, while MERT-330M has 24 layers.

LAION-CLAP. We include two variants: CLAP-Music[5] (CLAP-M), trained exclusively on music data, and CLAP-Music&Speech[6] (CLAP-M&S), which incorporates additional speech data [41]. Both utilize HTS-AT [159] for audio encoding, a transformer-based model with 4 groups of swin-transformer blocks [160], with 68M audio-specific parameters within a larger 194M parameter model.

Qwen2-Audio. The largest model in our evaluation framework, Qwen2-Audio[7] [115], contains 637M audio-specific parameters within an 8.4B parameter architecture and features 32 transformer layers [161] in its audio tower.

VGG-ish. As a baseline comparison, we include VGG-ish [128, 129], a 3.6M parameter end-to-end model trained via supervised learning on mel-spectrograms to predict tags. For VGG-ish, we report results from Chapter 4, where the same experimental setup is used, rather than running new experiments.

5.2.2 Datasets

Our evaluation spans diverse traditions from six music datasets. As in the previous chapter, we utilize MagnaTagATune [126] (25,863 clips) and FMA-medium [114] (25,000 tracks) for Western music. For world music traditions, we incorporate the Lyra dataset (see Chapter 3) with 1,570 recordings of Greek folk music, and three collections from the CompMusic project [33]: the Turkish-

[3] https://huggingface.co/m-a-p/MERT-v1-95M
[4] https://huggingface.co/m-a-p/MERT-v1-330M
[5] https://huggingface.co/laion/larger_clap_music
[6] https://huggingface.co/laion/larger_clap_music_and_speech
[7] https://huggingface.co/Qwen/Qwen2-Audio-7B

makam corpus [84, 85] (5,297 recordings) as well as Hindustani [19] (1,204 recordings) and Carnatic [19] (2,612 recordings) of Indian classical music.

Following the same process with Chapter 4, we set maximum audio durations to achieve similar sizes between datasets and prepare their metadata for the auto-tagging task. For Probing and Supervised Fine-Tuning, we use the standard tag sets, i.e., 50 tags for MagnaTagATune, 30 for Lyra and Turkish-makam, and 20 for the rest of the datasets. Our ML-FSL experiments use extended tag sets that include previously unseen classes, summing up to: 80 tags for MagnaTagATune, 60 for Lyra and Turkish-makam, 40 for FMA-medium and Carnatic, and 35 for Hindustani, consistent with Section 4.7.

5.2.3 Evaluation methodologies

Probing. Our first methodology (*Prob.*) evaluates how well foundation models inherently represent musical characteristics across cultures. We employ probing, where the model remains frozen while only training a classifier on top of the extracted representations. Specifically, we implement a shallow Multi-layer Perceptron (MLP) with a single hidden layer of 512 units followed by a sigmoid classification layer, optimized with binary cross-entropy loss.

Supervised Fine-Tuning. To evaluate adaptation potential, we implement targeted supervised fine-tuning (*SFT*) by unfreezing a subset of model parameters. For MERT-95M, we unfreeze the last two transformer layers, while for MERT-330M only the last layer. For both CLAP models, we unfreeze the last group of swin-transformer blocks of the audio encoder along with the normalization and two projection layers. In Qwen2-Audio, we fine-tune the last layer of the audio tower along with the normalization layer before multi-modal projection. These choices were constrained by RAM limitations affecting both trainable parameters and hyperparameter tuning. We use the same trainable MLP Probe architecture as in the Probing experiments, initializing it with the weights learned during that phase. This weight initialization strategy helps maintain previously learned knowledge while adapting to new domains, mitigating potential catastrophic forgetting issues [162]. We also employ learning rate warmup and cosine scheduling to ensure stable adaptation [163].

Multi-Label Few-Shot Learning. Our third methodology (*ML-FSL*) evaluates performance in low-resource scenarios by employing the optimized version of LC-Protonets that is detailed in subsection 4.8.3. We extract representations from three different contexts: directly from the pre-trained model (*PT*), from the hidden layer of the trained MLP Probe (*Prob.*), and from the fine-tuned model (*SFT*). Notably, this methodology involves no additional training during few-shot evaluation; the model acts as a frozen feature extractor that maps both the few examples and the unknown items to an embedding space where classification occurs utilizing the LC-Protonets approach.

5.3 Foundation Models Evaluation: Experimental Setup

Experiments and resources. We conducted 5 runs with different random seeds for both Probing and ML-FSL tasks, but a single run for SFT due to computational constraints. SFT trainable parameters varied: 14M for MERT-95M, 13M for MERT-330M, 25M for CLAP models, and 56M for Qwen2-Audio. All experiments ran on an NVIDIA RTX A5000 GPU, and we used Qwen2-Audio in half-precision (FP16) in all our methodologies to fit in this card. Most SFT training completed within 24 hours, with only 3 out of 30 experiments extending to about 36 hours.

Dataset processing. We standardized Turkish-makam, Hindustani, and Carnatic datasets to approximately 200 hours each, matching MagnaTagATune and FMA-medium durations, while

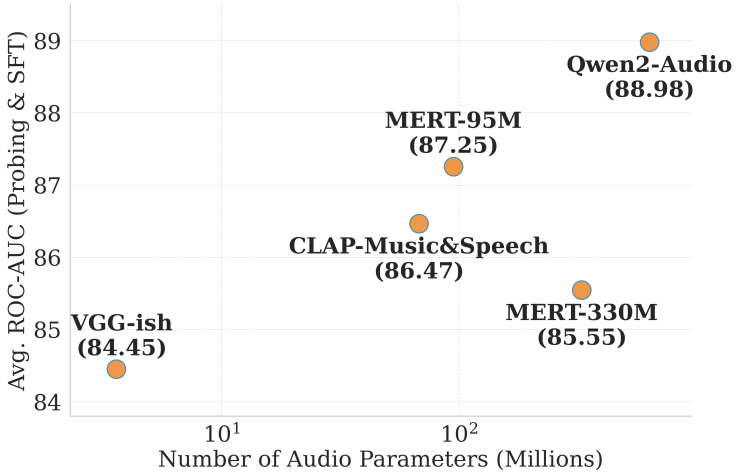

Figure 5.2. Model Size vs. Performance Relationship. Correlation between model size (audio-specific parameters on logarithmic scale) and mean ROC-AUC (%) across all datasets, revealing efficiency-performance trade-offs in foundation models.

Lyra remained at its original 80 hours. We followed the training, validation, and test splits from [46, 129]. For ML-FSL, evaluation items came exclusively from test sets to prevent data leakage, as in Section 4.7.

Model-specific configurations. Each foundation model required specific preprocessing: MERT models use 30-second windows at 24kHz, CLAP models 10-second windows at 48kHz, and Qwen2-Audio 30-second windows at 16kHz. All audio was converted to mono and resampled to the model's required rate.

Representation extraction strategies. For MERT models, we extract representations by summing the average, across time, hidden states of the last four layers of the models. For CLAP models, we extract them from the audio projection layer which takes as input the average pooled layer representation of the last hidden state. For Qwen2-Audio, we use the last hidden state embeddings averaged across all layers of the whole model, when passing a simple text prompt that includes nothing but the respective tags for audio processing, i.e., `<|audio_bos|><|AUDIO|><|audio_eos|>`. These representation extraction strategies, number of fine-tuned layers, and other design choices of our method were optimized through preliminary experiments.

Hyperparameters. For Probing, we used Adam optimizer [149] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$) with learning rate 10^{-3}, batch size 16, early stopping patience 10, and maximum 200 epochs. For SFT, we used AdamW [164] with identical β parameters but learning rate 10^{-4}, model-specific batch sizes (to fit maximum available resources) with gradient accumulation to simulate batch size 16 across all setups, patience 5, and maximum 30 epochs. We applied learning rate warmup and cosine scheduling for the first 5% of SFT epochs. ML-FSL evaluations used cosine distance with an N-way K-shot setup, with N being the number of extended tags per dataset and K equal to 3 examples per label in all experiments. We also attempted Low-Rank Adaptation [165] initially but abandoned it due to extensive hyperparameter tuning requirements across our 5×6 experimental matrix.

Evaluation metrics. For the Probing and SFT methodologies, we report area under the receiver operating characteristic curve (ROC-AUC) and mean average precision (mAP). These metrics are particularly well-suited for multi-label classification tasks [150] and are consistent with prior work in music tagging [46, 129]. For ML-FSL evaluation, we report macro-F1 (M-F1) and micro-F1

Model	**Params** Audio/Total	**ROC-AUC** (%)		**mAP** (%)	
VGG-ish [46]	3.6M/3.6M	84.45		50.56	
		Prob.	*SFT*	*Prob.*	*SFT*
MERT-95M	95M/95M	$87.25_{0.32}$	87.26	$52.25_{0.42}$	52.68
MERT-330M	330M/330M	$85.40_{0.68}$	85.69	$49.62_{0.83}$	50.47
CLAP-M	68M/194M	$71.52_{1.14}$	78.96	$29.98_{1.07}$	40.41
CLAP-M&S	68M/194M	$86.78_{0.31}$	86.15	$53.12_{0.87}$	51.99
Qwen2-Audio	637M/8.40B	$\mathbf{88.59}_{0.47}$	**89.37**	$\mathbf{56.48}_{0.63}$	**58.73**

Table 5.1. Foundation Model Performance Comparison. Average Probing and SFT task performance across all datasets, with values averaged over multiple runs (standard deviations as subscripts). Bold values indicate best performance per column.

(m-F1) scores, which align with the LC-Protonets evaluation framework (Section 4.6). F1 score is the harmonic mean of the precision and recall scores. Macro-F1 gives equal weight to all classes, while micro-F1 accounts for class imbalance by calculating metrics globally across all instances.

5.4 Foundation Models Evaluation: Results and Analysis

5.4.1 Probing and Supervised Fine-Tuning

Table 5.1 presents the performance of the evaluated foundation models averaged across all datasets for both Probing and SFT tasks. Overall, Qwen2-Audio achieves the highest performance with 88.59% ROC-AUC and 56.48% mAP in Probing, further improving to 89.37% ROC-AUC and 58.73% mAP after fine-tuning. This is followed by MERT-95M and CLAP-Music&Speech with comparable performance, while CLAP-Music shows significantly lower performance without speech data in its training corpus.

Figure 5.2 illustrates the relationship between model size (audio-specific parameters) and ROC-AUC performance, averaged across datasets and both Probing and SFT tasks. A generally positive correlation is revealed, with similar trends observed in both methodologies. Qwen2-Audio (637M parameters) consistently outperforms smaller models, achieving 88.98% average ROC-AUC score. Surprisingly, MERT-95M (87.25%) outperforms the much larger MERT-330M (85.55%). This is worth noting as [116] reported that both models performed on par for auto-tagging tasks, suggesting that our common representation extraction strategy for both MERT models may not optimally leverage the larger model's capacity. Another potential explanation is that MERT-95M has been trained on open data whereas MERT-330M has been trained with additional proprietary data with a strong Western bias [40].

When examining Probing performance across individual datasets, in Table 5.2, we observe a consistent pattern of decreasing performance for music traditions that are culturally distant from the data used to pre-train the respective foundation models. Western music datasets (MagnaTagATune and FMA-medium) consistently achieve the highest performance across all models, with ROC-AUC values reaching 96.60% for Qwen2-Audio on FMA-medium. Greek (Lyra) and Turkish (makam) music datasets show moderate performance, while Indian classical music (Hindustani and Carnatic) datasets typically exhibit the lowest performance. This cultural performance gap is espe-

Model	MagnaTagATune		FMA-medium		Lyra	
	ROC-AUC	mAP	ROC-AUC	mAP	ROC-AUC	mAP
VGG-ish [46]	91.23	45.82	88.89	49.49	80.97	48.06
Probing (Prob.)						
MERT-95M	$90.46_{0.10}$	$44.16_{0.21}$	$91.68_{0.08}$	$51.43_{0.43}$	$85.61_{0.66}$	$53.34_{0.61}$
MERT-330M	$89.66_{0.16}$	$41.73_{0.59}$	$90.78_{0.11}$	$48.85_{0.32}$	$84.65_{0.78}$	$51.81_{0.59}$
CLAP-M	$80.07_{0.21}$	$25.82_{0.13}$	$77.42_{0.15}$	$22.89_{0.38}$	$64.18_{1.29}$	$31.16_{0.43}$
CLAP-M&S	$92.41_{0.05}$	$\mathbf{48.54}_{0.16}$	$94.05_{0.08}$	$59.13_{0.54}$	$87.25_{0.18}$	$56.94_{0.51}$
Qwen2-Audio	$91.17_{0.13}$	$45.58_{0.21}$	$96.60_{0.07}$	$73.38_{0.28}$	$86.44_{0.81}$	$53.50_{0.65}$
Supervised Fine-Tuning (SFT)						
MERT-95M	90.62	44.52	91.70	51.74	84.89	53.62
MERT-330M	89.55	41.93	91.12	49.56	84.74	52.54
CLAP-M	88.54	39.26	88.37	42.04	71.97	38.14
CLAP-M&S	91.77	47.54	92.86	57.11	85.35	52.86
Qwen2-Audio	92.03	48.27	**97.02**	**75.94**	**87.57**	**57.04**
(Previous) SOTA	**92.7**	46.54	92.4	53.7	85.4	54.3

Model	Turkish-makam		Hindustani		Carnatic	
	ROC-AUC	mAP	ROC-AUC	mAP	ROC-AUC	mAP
VGG-ish [46]	86.96	56.39	84.77	60.82	73.92	42.78
Probing (Prob.)						
MERT-95M	$\mathbf{88.22}_{0.23}$	$57.89_{0.34}$	$86.59_{0.52}$	$60.26_{0.56}$	$80.96_{0.35}$	$46.41_{0.35}$
MERT-330M	$85.37_{0.64}$	$52.45_{1.12}$	$84.23_{1.36}$	$58.78_{2.08}$	$77.73_{1.03}$	$44.07_{0.31}$
CLAP-M	$77.31_{0.51}$	$38.77_{1.00}$	$68.69_{4.05}$	$33.43_{4.21}$	$61.47_{0.60}$	$27.83_{0.30}$
CLAP-M&S	$86.49_{0.27}$	$54.69_{0.36}$	$82.61_{1.14}$	$55.70_{3.29}$	$77.85_{0.13}$	$43.73_{0.35}$
Qwen2-Audio	$86.64_{0.42}$	$53.38_{0.79}$	$\mathbf{88.45}_{0.83}$	$62.42_{0.99}$	$82.22_{0.56}$	$50.59_{0.88}$
Supervised Fine-Tuning (SFT)						
MERT-95M	87.50	**57.91**	88.20	61.47	80.64	46.83
MERT-330M	86.17	53.80	85.49	61.33	77.05	43.66
CLAP-M	79.82	42.49	75.65	45.01	69.39	35.51
CLAP-M&S	86.69	54.93	83.73	56.91	76.51	42.58
Qwen2-Audio	87.95	56.10	88.32	**64.35**	**83.35**	**50.66**
(Previous) SOTA	87.7	57.7	86.5	63.1	77.0	43.9

Table 5.2. Dataset-Specific Model Performance. Detailed ROC-AUC and AP scores for each dataset-model combination. For Probing, values are averaged over multiple runs with subscripted standard deviations, while SFT results are from single runs. Bold values indicate best performance per metric and dataset. SOTA values are from [166] for MagnaTagATune and [46] for the rest of the datasets.

Model	M-F1			m-F1		
VGG-ish [47]	30.18			55.09		
	PT	*Prob.*	*SFT*	*PT*	*Prob.*	*SFT*
MERT-95M	$23.90_{1.52}$	$28.05_{1.74}$	$28.28_{1.80}$	$46.59_{1.57}$	$52.16_{1.43}$	$52.56_{1.63}$
MERT-330M	$23.03_{1.12}$	$28.48_{1.40}$	$28.51_{1.28}$	$45.11_{1.29}$	$51.78_{1.51}$	$51.80_{1.46}$
CLAP-M	$17.71_{1.20}$	$18.43_{1.40}$	$21.58_{1.13}$	$38.80_{1.37}$	$39.97_{1.20}$	$46.57_{1.20}$
CLAP-M&S	$\mathbf{28.23}_{1.36}$	$29.22_{1.09}$	$30.27_{1.90}$	$\mathbf{51.59}_{1.54}$	$53.32_{1.31}$	$54.43_{1.27}$
Qwen2-Audio	$25.98_{1.36}$	$\mathbf{30.96}_{1.26}$	$\mathbf{32.00}_{1.41}$	$49.97_{1.41}$	$\mathbf{55.66}_{0.82}$	$\mathbf{56.85}_{1.23}$

Table 5.3. ML-FSL Performance on Extended Tag Sets. Macro-F1 (M-F1) and micro-F1 (m-F1) scores averaged across datasets (with subscripted standard deviations) in three contexts (PT, Prob., SFT), demonstrating how foundation models perform with limited supervision on rare tags. Bold indicates best performance per column.

cially pronounced for CLAP-Music, where the ROC-AUC drops from 80.07% for MagnaTagATune to 61.47% for Carnatic.

Applying Supervised Fine-Tuning (*SFT*) generally improves performance across all models and datasets, with an average gain of 1-2% in ROC-AUC for most models. Notably, CLAP-Music shows the largest improvement with *SFT*, indicating greater adaptation potential despite lower absolute performance. For other models, the modest gains suggest that they require broader fine-tuning to further shift their pre-trained representations towards different cultures.

Importantly, our approaches achieve state-of-the-art performance in five out of six datasets, with MagnaTagATune being the only exception. However, their consistent performance decrease towards diverse cultures, suggests that their representations are still biased toward Western musical traditions.

5.4.2 Multi-label few-shot learning

Table 5.3 presents the ML-FSL evaluation results averaged across all datasets using extended tag sets. The results show consistent performance improvements moving from pre-trained models (PT) to trained probing models (Prob.) and then to supervised fine-tuned models (SFT) across all foundation models. The substantial gap between macro-F1 and micro-F1 metrics indicates considerable class imbalance in the extended tag sets, while the increased standard deviation stems from the support set sampling which can significantly impact the classification performance.

Qwen2-Audio demonstrates the best overall performance in the ML-FSL task with 32.00% macro-F1 and 56.85% micro-F1 after fine-tuning, followed closely by CLAP-Music&Speech with 30.27% macro-F1 and 54.43% micro-F1. Notably, even the best foundation model's performance (Qwen2-Audio) is comparable to a VGG-ish feature extractor trained via supervised learning on standard tags for each dataset. This stands in contrast to the Probing and SFT settings (Table 5.1), where foundation models clearly outperform VGG-ish, showing that ML-FSL tasks remain challenging for them despite their extensive pre-training. Supervised learning of a VGG-ish model on extended tag sets has not been conducted in the literature, likely due to the scarcity of examples for infrequent tags.

When examining the ML-FSL results per dataset in Table 5.4, we observe that only on Western datasets (MagnaTagATune and FMA-medium) does the best foundation model (Qwen2-Audio) achieve significantly better performance than the VGG-ish baseline. For Turkish-makam, VGG-ish representations actually outperform foundation models, while for Lyra, Hindustani, and Carnatic,

Model	MagnaTagATune		FMA-medium		Lyra	
	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
VGG-ish [47]	26.40	37.31	29.12	45.37	46.05	69.03
Pre-Trained models (PT)						
MERT-95M	$18.76_{1.04}$	$28.37_{1.38}$	$16.24_{0.64}$	$35.37_{0.94}$	$46.87_{2.59}$	$66.07_{2.25}$
MERT-330M	$18.17_{0.78}$	$26.99_{1.36}$	$16.24_{0.69}$	$31.15_{1.51}$	$44.22_{1.45}$	$65.48_{1.57}$
CLAP-M	$13.10_{0.84}$	$20.00_{1.15}$	$9.65_{0.29}$	$19.77_{1.31}$	$33.56_{2.88}$	$57.14_{1.49}$
CLAP-M&S	$25.90_{0.55}$	$36.55_{0.61}$	$28.78_{1.66}$	$42.95_{2.02}$	$48.03_{2.02}$	$69.04_{1.54}$
Qwen2-Audio	$21.29_{0.51}$	$32.09_{0.26}$	$29.76_{2.23}$	$47.50_{1.86}$	$39.99_{1.05}$	$64.24_{1.07}$
Trained Probing models (Prob.)						
MERT-95M	$23.77_{0.85}$	$34.71_{1.03}$	$24.62_{1.19}$	$42.96_{1.30}$	$45.80_{2.76}$	$68.16_{1.81}$
MERT-330M	$24.48_{0.59}$	$34.78_{1.45}$	$25.21_{0.76}$	$40.65_{1.76}$	$47.92_{3.26}$	$\mathbf{70.15}_{2.18}$
CLAP-M	$14.84_{0.49}$	$22.67_{1.00}$	$11.55_{0.50}$	$22.72_{1.48}$	$34.85_{4.03}$	$57.73_{1.37}$
CLAP-M&S	$26.90_{0.47}$	$37.62_{0.93}$	$31.14_{1.28}$	$46.53_{1.59}$	$47.10_{0.89}$	$69.77_{0.53}$
Qwen2-Audio	$26.79_{0.40}$	$37.65_{0.21}$	$39.49_{1.02}$	$56.30_{0.82}$	$42.52_{1.81}$	$67.10_{1.13}$
Supervised Fine-Tuned models (SFT)						
MERT-95M	$24.46_{0.79}$	$35.28_{0.90}$	$24.94_{1.18}$	$42.78_{1.44}$	$45.51_{3.74}$	$67.93_{2.72}$
MERT-330M	$23.78_{0.65}$	$33.67_{0.91}$	$24.94_{1.21}$	$39.95_{1.77}$	$\mathbf{48.50}_{2.75}$	$70.06_{2.23}$
CLAP-M	$22.15_{0.51}$	$32.67_{1.22}$	$19.61_{0.79}$	$34.81_{0.99}$	$30.46_{2.04}$	$55.86_{2.02}$
CLAP-M&S	$26.28_{0.50}$	$37.23_{1.09}$	$30.27_{1.56}$	$46.57_{1.61}$	$48.09_{4.74}$	$69.93_{2.28}$
Qwen2-Audio	$\mathbf{27.67}_{0.25}$	$\mathbf{38.57}_{0.18}$	$\mathbf{40.10}_{1.29}$	$\mathbf{57.17}_{0.95}$	$44.13_{2.45}$	$68.34_{2.38}$

Model	Turkish-makam		Hindustani		Carnatic	
	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
VGG-ish [47]	**30.07**	**56.22**	31.33	58.38	18.13	**64.25**
Pre-Trained models (PT)						
MERT-95M	$20.69_{1.77}$	$40.95_{1.80}$	$25.87_{2.45}$	$51.50_{1.92}$	$14.97_{0.64}$	$57.26_{1.10}$
MERT-330M	$20.14_{2.01}$	$39.71_{1.95}$	$25.08_{1.40}$	$50.14_{1.08}$	$14.32_{0.41}$	$57.21_{0.28}$
CLAP-M	$14.33_{1.10}$	$32.12_{1.37}$	$21.06_{1.63}$	$47.38_{1.60}$	$14.55_{0.43}$	$56.42_{1.30}$
CLAP-M&S	$24.19_{1.73}$	$47.13_{2.22}$	$26.29_{1.20}$	$54.50_{1.57}$	$16.19_{1.02}$	$59.38_{1.30}$
Qwen2-Audio	$19.89_{1.71}$	$42.27_{1.88}$	$28.42_{1.96}$	$55.92_{1.70}$	$16.55_{0.69}$	$57.82_{1.67}$
Trained Probing models (Prob.)						
MERT-95M	$26.14_{1.73}$	$50.00_{0.70}$	$30.75_{2.95}$	$56.41_{2.18}$	$17.25_{0.98}$	$60.70_{1.55}$
MERT-330M	$26.97_{1.61}$	$50.47_{1.13}$	$29.25_{1.55}$	$53.77_{1.82}$	$17.06_{0.61}$	$60.85_{0.69}$
CLAP-M	$16.68_{0.81}$	$36.00_{1.22}$	$18.77_{1.42}$	$44.96_{0.95}$	$13.87_{1.16}$	$55.74_{1.15}$
CLAP-M&S	$25.58_{1.59}$	$49.70_{1.39}$	$28.11_{1.38}$	$56.43_{2.19}$	$16.46_{0.92}$	$59.88_{1.25}$
Qwen2-Audio	$26.09_{1.65}$	$51.59_{1.20}$	$31.62_{1.26}$	$60.08_{0.40}$	$19.25_{1.40}$	$61.24_{1.14}$
Supervised Fine-Tuned models (SFT)						
MERT-95M	$26.16_{1.87}$	$49.76_{1.54}$	$30.40_{2.15}$	$56.39_{1.68}$	$18.18_{1.08}$	$63.19_{1.48}$
MERT-330M	$26.84_{1.51}$	$50.29_{1.25}$	$30.56_{1.31}$	$55.25_{1.58}$	$16.43_{0.27}$	$61.57_{1.04}$
CLAP-M	$20.66_{1.69}$	$45.80_{1.13}$	$21.95_{1.31}$	$50.74_{1.14}$	$14.63_{0.45}$	$59.53_{0.67}$
CLAP-M&S	$28.91_{1.75}$	$53.87_{1.56}$	$31.27_{2.47}$	$57.41_{0.74}$	$16.82_{0.37}$	$61.55_{0.34}$
Qwen2-Audio	$27.61_{2.37}$	$53.98_{1.55}$	$\mathbf{32.52}_{1.23}$	$\mathbf{60.26}_{0.89}$	$\mathbf{19.97}_{0.87}$	$62.76_{1.43}$

Table 5.4. Dataset-Specific ML-FSL Performance. Detailed macro-F1 (M-F1) and micro-F1 (m-F1) scores on extended tag sets for each individual dataset across three contexts. Values are means with subscripted standard deviations. Bold indicates best performance per column.

the results are comparable. This pattern provides additional clear evidence of the implicit Western-centric bias integrated into models due to their pre-training data.

5.5 CultureMERT: A Multi-Culturally Adapted Foundation Model

Building upon the insights from our comprehensive evaluation of foundation models, we now present a novel approach to enhance their cultural inclusivity developed in collaboration with Angelos-Nikolaos Kanatas. The evaluation results have clearly demonstrated both the potential and limitations of existing foundation models when applied to diverse musical traditions. In particular, we observed a consistent performance gap for culturally distant traditions and in low-resource scenarios, highlighting the need for dedicated adaptation strategies.

The technical implementation of the continual pre-training methodology presented in this section was primarily developed by Angelos-Nikolaos Kanatas, with my contributions focusing on the experimental design, cultural adaptation evaluation framework, and cross-cultural analysis presented in Sections 5.5.4 and 5.6.

The overall framework is illustrated in Figure 5.3, which depicts the two-stage continual pre-training strategy for `CultureMERT`. In the following section, we first review the architecture and pre-training objective of MERT, and then present our CPT strategy for cultural adaptation. Finally, we investigate task arithmetic, an alternative approach to multicultural adaptation that merges culturally specialized models in weight space to construct a unified multicultural model, `CultureMERT-TA`.

To support research on world music representation learning, we publicly release `CultureMERT-95M`[8] and `CultureMERT-TA-95M`[9], fostering the development of more culturally aware music foundation models.

5.5.1 MERT Pre-Training Objective

Our continual pre-training follows the self-supervised masked language modeling objective of $\texttt{MERT}^{\texttt{RVQ-VAE}}$ [40], which uses two teacher models: (i) an acoustic teacher (EnCodec codec model [167]) that discretizes audio into tokens from $K = 8$ residual vector quantization codebooks, and (ii) a musical teacher based on Constant-Q Transform spectrogram reconstruction.

`MERT-v1-95M` follows the HuBERT architecture [168] with a CNN-based feature extractor and 12-layer Transformer encoder. The training objective combines masked acoustic token prediction and spectrogram reconstruction:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{RVQ}} + \mathcal{L}_{\text{CQT}}, \tag{5.1}$$

where $\mathcal{L}_{\text{RVQ}}$ is the acoustic MLM loss using Noise Contrastive Estimation, and $\mathcal{L}_{\text{CQT}}$ is the CQT reconstruction loss minimizing mean squared error between predicted and ground-truth features.

5.5.2 Two-Stage Continual Pre-Training Strategy

To adapt the MERT foundation model to diverse musical traditions, we employ continual pre-training, which extends the training of a pre-trained model on new data, aiming to adapt it to a shifted domain or task while retaining prior knowledge, without re-training from scratch. In

[8] https://huggingface.co/ntua-slp/CultureMERT-95M
[9] https://huggingface.co/ntua-slp/CultureMERT-TA-95M

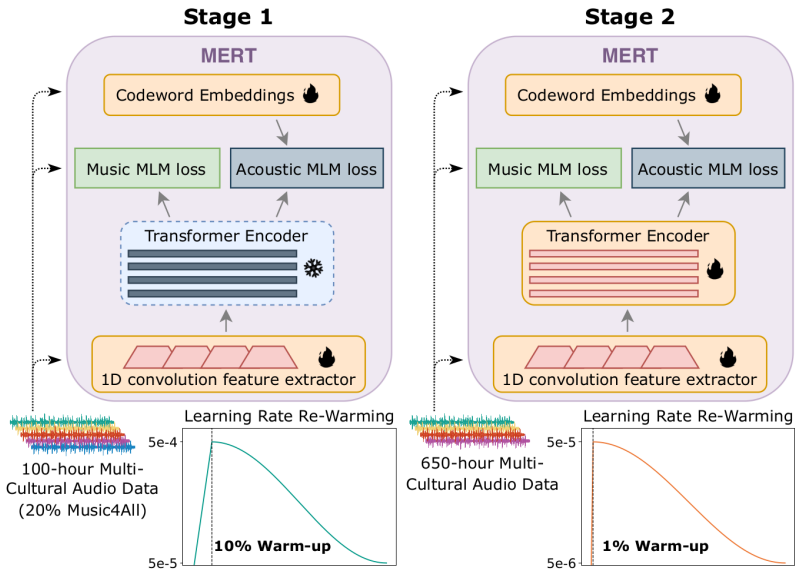

Figure 5.3. Two-Stage Continual Pre-Training Strategy for `CultureMERT`. In Stage 1, a subset of parameters is trained on 100h of multi-cultural data with 20% Western music for stabilization. In Stage 2, all parameters are unfrozen and trained on the full 650h dataset. Learning rate re-warming and re-decaying is applied in both stages.

our case, this involves continually pre-training the `MERT-v1-95M` model on culturally diverse data that introduce a significant distribution shift, as it was initially trained on predominantly Western music [40, 169].

Given this shift, naively continuing to train the model can lead to catastrophic forgetting [162] and poor adaptation [119], as confirmed by preliminary experiments (see Table 5.5). To address this, we propose a **two-stage** strategy that stabilizes training through: (i) learning rate re-warming and re-decaying [119, 163, 170], and (ii) staged adaptation.

To mitigate the *stability gap* observed during continual pre-training [171, 172], we split training into two stages, as illustrated in Figure 5.3:

Stage 1 - Stabilization Phase: Train on a smaller data subset, updating only the CNN-based feature extractor and codeword embedding layer while keeping the Transformer encoder frozen. To reduce distribution gap and mitigate forgetting [170], we incorporate 20% Music4All data [173] (primarily Western) into the pre-training mix.

Stage 2 - Full Adaptation: Unfreeze the Transformer encoder and continue training on the full dataset.

This approach balances *plasticity* (adaptation to non-Western traditions) and *stability* (retaining knowledge on Western datasets), addressing the *stability-plasticity dilemma* [174, 175].

Learning Rate Re-Warming: We apply learning rate re-warming and re-decaying in both stages, as prior work has shown this is crucial for preventing poor convergence and mitigating catastrophic forgetting during continual pre-training [119, 163, 170].

Following this strategy, we develop: (i) a **multi-culturally adapted model**, `CultureMERT`, trained on a diverse mix spanning all four non-Western musical traditions; and (ii) **single-culture adapted models** (`MakamMERT`, `HindustaniMERT`, `CarnaticMERT`, `LyraMERT`).

CPT Strategy	Western Replay	Turkish-makam	MTAT
MERT-v1 (Baseline)	-	83.2	**89.6**
Single-stage	✓	83.8	86.0
Single-stage (no re-warm)	✓	83.0	87.5
Two-stage (*Ours*)	Stage 1	**89.6**	89.2
Two-stage (*Ours*)	Both stages	88.6	89.4

Table 5.5. CPT Strategy Comparison. ROC-AUC scores on Turkish-makam and MTAT datasets. Two-stage CPT outperforms single-stage adaptation, with Western replay limited to Stage 1 yielding the best trade-off between cultural adaptation and knowledge retention.

5.5.3 Task Arithmetic for Cross-Cultural Adaptation

As an alternative to continual pre-training, we explore task arithmetic [120], which combines culturally specialized models in weight space. We obtain *task vectors* by computing the element-wise difference between single-culture adapted models and the base `MERT-v1` model: $\tau_i = \theta_i - \theta_0$.

For multi-cultural adaptation, we construct a unified model by merging task vectors:

$$\theta' = \theta_0 + \sum_{i=1}^{N} \lambda_i \tau_i, \tag{5.2}$$

where λ_i controls each task vector's contribution. When $\lambda = 1/N$, this simplifies to weight averaging [176, 177].

5.5.4 Experimental Implementation

Implementation Details: We initialize models from the publicly available `MERT-v1-95M` pre-trained checkpoint. Training uses 5-second audio segments randomly cropped from 30-second pre-training data, with the EnCodec neural codec model [167] remaining frozen throughout continual pre-training [40]. We apply in-batch noise mixture augmentation and pre-layer normalization [178] for stable training.

Probing-Based Evaluation: Following [40, 109, 113], we adopt probing-based evaluation, keeping pre-trained models frozen while training only a shallow MLP for sequence-level tasks. Our evaluation follows the MARBLE protocol [116] for both Western and non-Western music tagging tasks. Audio files are segmented into 30-second chunks with predictions aggregated by averaging.

Training Configuration: We develop both multi-culturally adapted models (`CultureMERT`) trained on diverse mixes spanning all four non-Western traditions, and single-culture adapted models for each tradition individually. Training follows the two-stage approach with appropriate data allocation: 100 hours in Stage 1 and the full 650-hour dataset in Stage 2 for multi-cultural adaptation, with proportionally scaled configurations for single-culture models.

Dataset	MagnaTagATune		FMA-medium		Lyra		Avg.
Metrics	ROC	AP	ROC	AP	ROC	AP	
MERT-v1	$89.6_{0.07}$	$35.9_{0.15}$	$90.7_{0.04}$	$48.1_{0.11}$	$85.7_{0.10}$	$56.5_{0.18}$	66.1
MakamMERT	$89.0_{0.07}$	$35.6_{0.12}$	$90.3_{0.12}$	$47.1_{0.16}$	$84.6_{0.12}$	$53.2_{0.17}$	67.5
CarnaticMERT	$89.2_{0.10}$	$35.3_{0.11}$	$90.2_{0.10}$	$46.7_{0.09}$	$85.4_{0.11}$	$55.8_{0.16}$	68.3
HindustaniMERT	$89.1_{0.09}$	$35.8_{0.13}$	$90.2_{0.13}$	$46.1_{0.10}$	$84.2_{0.13}$	$52.0_{0.15}$	67.6
LyraMERT	$88.9_{0.05}$	$35.1_{0.14}$	$90.0_{0.08}$	$46.0_{0.16}$	$85.0_{0.11}$	$53.5_{0.14}$	66.8
CultureMERT	$89.4_{0.09}$	$35.9_{0.16}$	$90.7_{0.09}$	$48.1_{0.13}$	$86.9_{0.10}$	$56.7_{0.20}$	**69.3**
CultureMERT-TA	$89.6_{0.10}$	$36.4_{0.14}$	$90.8_{0.06}$	$49.1_{0.15}$	$\mathbf{87.3}_{0.08}$	$\mathbf{57.3}_{0.19}$	69.1
(Previous) SOTA	**92.7** [166]	**41.4** [109]	**92.4** [46]	**53.7** [46]	85.4 [46]	54.3 [46]	-

Dataset	Turkish-makam		Hindustani		Carnatic		Avg.
Metrics	ROC	AP	ROC	AP	ROC	AP	
MERT-v1	$83.2_{0.08}$	$53.3_{0.12}$	$82.4_{0.04}$	$52.9_{0.19}$	$74.9_{0.05}$	$39.7_{0.15}$	66.1
MakamMERT	$88.7_{0.11}$	$58.8_{0.22}$	$84.5_{0.16}$	$57.8_{0.18}$	$77.6_{0.14}$	$42.7_{0.16}$	67.5
CarnaticMERT	$88.4_{0.06}$	$58.4_{0.16}$	$87.0_{0.06}$	$60.2_{0.14}$	$78.8_{0.13}$	$\mathbf{44.0}_{0.17}$	68.3
HindustaniMERT	$88.3_{0.12}$	$58.2_{0.16}$	$87.4_{0.11}$	$60.3_{0.16}$	$77.0_{0.12}$	$42.7_{0.16}$	67.6
LyraMERT	$86.7_{0.07}$	$56.8_{0.13}$	$85.9_{0.08}$	$57.4_{0.13}$	$76.4_{0.09}$	$40.1_{0.13}$	66.8
CultureMERT	$\mathbf{89.6}_{0.09}$	$60.6_{0.21}$	$\mathbf{88.2}_{0.20}$	$\mathbf{63.5}_{0.24}$	$\mathbf{79.2}_{0.18}$	$43.1_{0.22}$	**69.3**
CultureMERT-TA	$89.0_{0.12}$	$\mathbf{61.0}_{0.18}$	$87.5_{0.10}$	$59.3_{0.13}$	$79.1_{0.11}$	$43.3_{0.13}$	69.1
(Previous) SOTA	87.7 [46]	57.7 [46]	86.5 [46]	63.1 [46]	77.0 [46]	43.9 [46]	-

Table 5.6. Evaluation Results of Pre-Trained and Adapted MERT Models. ROC-AUC and AP scores across datasets (with standard deviations as subscripts), highlighting the impact of multi-cultural CPT (CultureMERT) and task arithmetic on cross-cultural adaptation and transfer. The "Avg." column represents the average performance across all datasets and evaluation metrics for each model.

5.6 CultureMERT: Performance Evaluation and Cross-Cultural Analysis

The following analysis examines the cross-cultural capabilities and transfer patterns of the adapted models, building on the evaluation framework developed for assessing cultural adaptation effectiveness.

As shown in Table 5.6, `CultureMERT`, adapted via multi-cultural continual pre-training, consistently outperforms the original `MERT-v1` model across all non-Western tasks and evaluation metrics, achieving an average improvement of 4.9%. It also surpasses the single-culture adapted models on average, suggesting that incorporating culturally diverse data during CPT benefits all non-Western traditions by improving the quality of representations for each individual culture, thereby enhancing generalization. Notably, `CultureMERT` achieves this with minimal forgetting on Western benchmarks (0.05% average drop across ROC-AUC and AP), demonstrating the efficacy of our approach. We further observe that single-culture adapted models tend to perform best on their respective in-domain tasks for well-resourced traditions, reaffirming the effectiveness of CPT for domain-specific adaptation [157]. However, even low-resource adaptation, as in the case

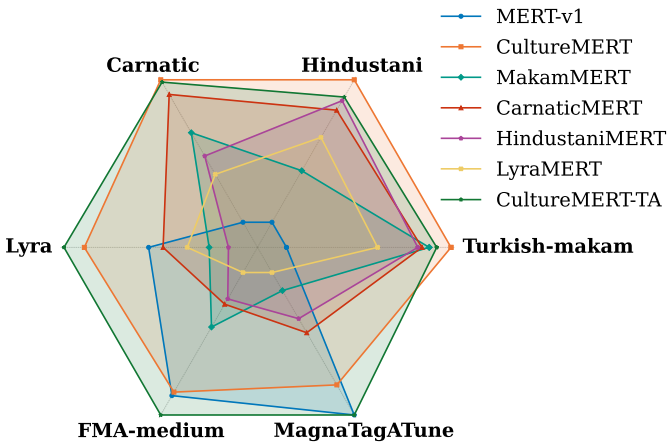

Figure 5.4. Cross-Cultural Transferability. Relative ROC-AUC performance across datasets, highlighting key trends in cross-cultural transfer. `CultureMERT` generalizes well to non-Western datasets, while task arithmetic performs on par in these settings and even surpasses both the pre-trained and multi-culturally adapted models on Western benchmarks (FMA-medium, MTAT) and Lyra.

of `LyraMERT` trained on just 50 hours, leads to noticeable gains across other non-Western tasks, indicating that even limited cultural exposure can significantly boost cross-cultural generalization.

Moreover, task arithmetic performs comparably to `CultureMERT` on non-Western tasks and even surpasses it on Western benchmarks and Lyra, demonstrating that weight-space merging of culturally specialized models can serve as an effective, training-free alternative to multi-cultural CPT—provided such models are available. Interestingly, it also outperforms the unadapted base model by 0.4% on average across Western tasks. Notably, only the multi-cultural models, `CultureMERT` and `CultureMERT-TA`, outperform `MERT-v1` on Lyra, where the latter already serves as a strong baseline. This further underscores the effectiveness of multi-cultural adaptation, particularly in low-resource and transfer settings. Finally, `CultureMERT` and `CultureMERT-TA` surpass previous state-of-the-art (SOTA) results on all non-Western music tagging tasks, with the best task arithmetic variant obtained using $\lambda = 0.2$.

Cross-Cultural Transfer

As illustrated in Figure 5.4, continual pre-training on one musical tradition can benefit others to varying degrees, revealing asymmetries in cross-cultural transfer effectiveness. For instance, we observe strong transfer between Turkish-makam and Carnatic music, with models adapted to either tradition generalizing well to the other. This aligns with their shared theoretical foundations as modal frameworks that emphasize microtonality and improvisation, serving similar roles in their respective cultures [179]. Additionally, the strong performance of the Carnatic-adapted model on the Hindustani domain reinforces the musical proximity between these traditions, particularly in their shared use of *raga* (melodic mode) and *tala* (rhythmic framework) [154]. Interestingly, the model adapted to Carnatic music appears to be the most consistently transferable among single-culture adaptations, achieving strong results not only within Indian classical traditions but also generalizing well to Turkish-makam and Lyra.

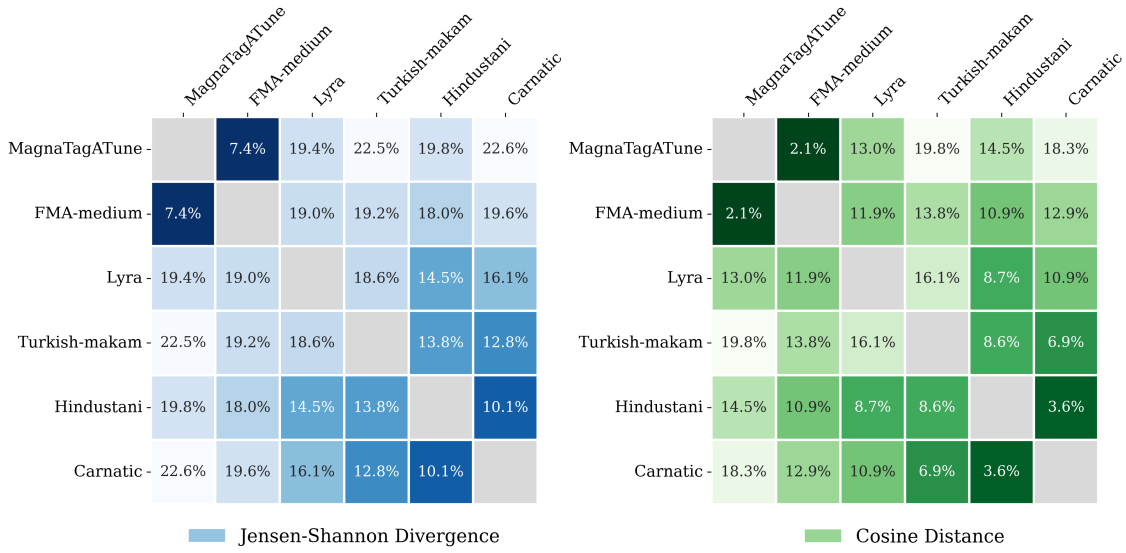

Figure 5.5. Token Similarity Across Cultures. Pairwise similarity between token distributions extracted from the EnCodec codec model [167]. Similarity scores are averaged across 8 codebooks, each containing 1024 discrete codewords (acoustic pseudo-tokens).

Token-Level Culture Similarity

To further examine cross-cultural similarities in our data, we analyze token overlap across musical traditions using both the Jensen-Shannon divergence (JSD) and cosine distance between token distributions extracted from the EnCodec model [167], which serves as our audio tokenizer. Lower values in both metrics indicate greater similarity. Our analysis, as shown in Figure 5.5, reveals strong token-level similarity among non-Western traditions, particularly between Hindustani and Carnatic music. In contrast, Western datasets (MTAT, FMA-medium) are highly similar to each other but notably dissimilar from non-Western traditions. Greek traditional music (Lyra), while distinct, aligns more closely with non-Western traditions than Western ones.

Interestingly, these findings correlate with our results on cross-cultural transfer, suggesting that token-level similarity metrics can serve as predictors of positive cross-cultural transfer. This insight has practical implications: such similarity metrics can guide the selection and refinement of pre-training data mixtures during CPT, or inform the adjustment of arithmetic operations when merging models via task arithmetic. Similar approaches for quantifying language similarity and predicting positive cross-lingual transfer, based on the similarity of extracted linguistic or acoustic tokens, have been explored in both the text [180, 181] and speech domains [182].

5.7 Conclusions

Our investigation of foundation models for diverse music cultures has revealed both the potential and limitations of current approaches, while also demonstrating effective strategies for enhancing versatility in music representation learning.

5.7.1 Foundation Models Evaluation: Key Findings

In our comprehensive evaluation of state-of-the-art foundation models across culturally diverse music corpora, we found that these models achieved better performance than previous approaches

for world music analysis, demonstrating impressive cross-cultural transfer capabilities. However, we also identified clear indicators of Western-centric bias, particularly in challenging low-resource scenarios.

The multi-label few-shot learning tasks particularly revealed these limitations. When faced with these challenging scenarios, foundation models performed on par with significantly smaller and simpler models, with performance notably degrading further on non-Western datasets. This finding underscores the limitations of current foundation models in representing the distinctive characteristics of diverse musical traditions, despite their general-purpose capabilities.

5.7.2 CultureMERT: Advancing Cross-Cultural Adaptation

To address the limitations identified in our evaluation, we developed CultureMERT, a multi-culturally adapted music foundation model created through continual pre-training on diverse non-Western musical traditions. Our two-stage CPT strategy, incorporating learning rate re-warming and staged adaptation, enabled stable training even under constrained computational resources.

Cross-cultural evaluation demonstrated that CultureMERT consistently outperformed the original pre-trained model across diverse non-Western music tagging tasks while preserving performance on Western benchmarks. This finding confirms the potential of continual pre-training for enhancing the cultural inclusivity of foundation models without sacrificing their general capabilities.

We also explored task arithmetic as an alternative approach to cross-cultural adaptation, finding that it offers a strong alternative to multi-cultural CPT by effectively merging culturally specialized models in weight space and mitigating catastrophic forgetting. This computationally efficient approach to model merging provides another pathway for enhancing the versatility of foundation models, particularly in scenarios where access to original training data or computational resources is limited.

5.7.3 Synthesis and Future Directions

Both approaches, foundation model evaluation and adaptation, have advanced our understanding of cross-cultural music representation learning. Our evaluation framework provides a systematic methodology for assessing the universality of music representations, while our adaptation strategies offer practical approaches for enhancing the cultural inclusivity of foundation models.

Despite these advances, several limitations and challenges remain. The frozen EnCodec tokenizer, trained on Western music, may be suboptimal for encoding culturally diverse musical languages, motivating future work on adapting or re-training audio tokenizers for diverse traditions. Additionally, future research could extend our methodological framework by incorporating Low-Rank Adaptation (LoRA) and implementing broader supervised fine-tuning to investigate further cultural adaptation.

Other promising directions include scaling to additional musical cultures, extending evaluation beyond sequence-level classification tasks, exploring mode estimation tasks that compare key in Western cultures with makam or raga recognition in other traditions, and conducting fine-grained ablations to better understand the adaptation process.

The work presented in this chapter contributes to the development of more inclusive and culturally aware computational models for music analysis, advancing toward the goal of truly universal music representations that can respect and preserve the rich diversity of global musical expressions. By combining the powerful representational capabilities of foundation models with effective adaptation strategies, we can enhance cross-cultural music understanding while maintaining performance

across diverse musical traditions.

However, a fundamental question remains unanswered: how well do these computational advances actually align with human perception of musical relationships across cultures? While our technical evaluations demonstrate improved performance on standardized benchmarks, the ultimate validation of cross-cultural music representation learning lies in its alignment with how humans actually perceive and understand musical similarity across diverse traditions. The next chapter addresses this crucial gap by presenting the first comprehensive evaluation of computational music similarity methods, including both the foundation models evaluated in this chapter and traditional signal processing approaches, against human cross-cultural music perception.

Chapter 6

Cross-Cultural Music Similarity: Bridging Human Perception and Computational Methods

The preceding chapters have developed a comprehensive framework for multicultural music representation learning, from dataset creation and methodological innovations to foundation model evaluation and adaptation. While these technical advances demonstrate improved computational performance across diverse musical traditions, a fundamental question remains: how well do these computational approaches align with human perception of musical relationships across cultures?

This chapter addresses this crucial validation gap by presenting the first systematic evaluation of computational music similarity methods against human perception. Building upon the foundation models evaluated in Chapter 5, the transfer learning insights from Chapter 4, and the diverse musical datasets established throughout this dissertation, we provide empirical evidence of how well computational approaches capture the nuanced ways humans perceive musical similarity across cultural boundaries.

6.1 Motivation

The assessment of musical similarity across cultural boundaries represents one of the most fundamental yet challenging problems in music information retrieval and computational music analysis. Music similarity assessment underlies numerous MIR applications, from recommendation systems and playlist generation to musicological analysis and content organization [11, 183, 184]. The complexity of this task becomes particularly pronounced when considering diverse cultural traditions, where conventional Western-centric approaches may fail to capture the nuanced relationships between musical styles, instruments, and aesthetic principles that define different musical cultures [15, 32].

Traditional computational approaches to music similarity have predominantly relied on either signal processing features or learned representations from deep neural networks. Signal processing features offer the advantage of interpretability through their connection to established music theory concepts such as rhythm, melody, harmony, and timbre [185, 186]. These features provide direct insights into which musical dimensions drive similarity assessments, enabling musicologists and system developers to understand and validate computational decisions. However, they often incorporate Western musical assumptions that may not generalize effectively across cultures. For instance, standard chroma features assume 12-tone equal temperament, potentially missing the microtonal ornamentations essential to many non-Western traditions [32]. Similarly, conventional rhythmic features may struggle with the complex asymmetrical patterns found in Eastern

Mediterranean and Indian music traditions.

Conversely, deep learning approaches have demonstrated impressive performance on various MIR tasks [109, 129], leveraging large-scale datasets to learn complex patterns that traditional hand-crafted features might miss. Recent advances in foundation models for audio and music, including models like MERT [40], CLAP [41], and other large-scale pre-trained architectures, offer new possibilities for cross-cultural music understanding. These models, trained on diverse audio data, potentially capture richer and more culturally aware representations compared to traditional approaches. However, such learned representations often lack interpretability and may inadvertently perpetuate cultural biases present in their training data [14, 44], which predominantly consists of Western commercial music. This creates a potential "self-reinforcing cycle" where established Western-centric datasets lead to specialized algorithms, which in turn encourage more similar data collection patterns [77].

The emergence of foundation models in music AI has introduced both opportunities and challenges for cross-cultural music similarity assessment. While these models demonstrate impressive capabilities across various MIR benchmarks, their alignment with human perception of musical similarity, particularly across cultural boundaries, remains largely unexplored. Most existing benchmarks and evaluation frameworks focus on Western musical contexts, an issue addressed in Chapter 5, with crucial questions remaining about how well these sophisticated computational approaches capture the nuanced ways humans perceive musical relationships across diverse traditions.

A critical gap exists in current MIR research: the systematic evaluation of computational similarity measures against human perception across diverse musical cultures. While several studies have examined specific aspects of cross-cultural music analysis [78], comprehensive comparisons between human judgments, interpretable signal processing features, and state-of-the-art foundation models are lacking. This gap is particularly problematic because the ultimate goal of music similarity systems is to align with human perception and serve users across different cultural contexts. Without empirical validation against human cross-cultural music perception, we cannot assess whether computational advances actually improve our ability to capture meaningful musical relationships as understood by human listeners.

The challenge is further complicated by the multi-dimensional nature of musical similarity. Research examining similarity perception across musical styles [122] has found that human judgments are context-specific and roughly equivalent between trained musicians and non-musicians, with ratings primarily based on surface features such as dynamics, articulation, texture, and contour rather than deeper structural relationships. Early work investigating statistical features and perceived similarity of folk melodies [121] found that frequency-based musical properties could account for moderate amounts (40%) of listeners' similarity ratings, with descriptive variables like melodic predictability and rhythmic variability achieving slightly better performance (55%). These findings suggest that humans may perceive similarity along various dimensions, including overall musical characteristics, cultural identity, personal preference for recommendations, timbral qualities, melodic relationships, or rhythmic patterns, while that each of these dimensions may be weighted differently across cultural contexts.

The assumption of music as a "universal language" has been increasingly challenged by research demonstrating cultural specificity in musical perception and understanding [2, 10]. Cultural context influences auditory perception and aesthetic appraisal, leading to diverse "listening frameworks" and "musical ontologies" that shape how different communities understand and categorize musical experience. This cultural conditioning of musical perception has profound implications for

developing computational approaches that can effectively assess similarity across diverse musical traditions.

Moreover, the question of which musical dimensions are most predictive of human similarity judgments across cultures remains open. While musicological theory suggests the importance of melody, rhythm, harmony, and timbre, empirical validation of these theoretical frameworks across diverse cultural contexts is limited. Large-scale evaluations have provided some insights into the relationship between computational approaches and human perception. A comprehensive cross-site evaluation [123] compared acoustic techniques against subjective measures across 400 popular artists, demonstrating that acoustic measures could achieve agreement with ground truth data comparable to internal agreement between different subjective sources. Recent work [124] evaluated audio representations against human timbre similarity ratings, finding that style embeddings from foundation models like CLAP achieved superior performance compared to traditional signal processing features.

The research presented in this chapter addresses these fundamental challenges by providing the first comprehensive evaluation of computational music similarity methods against human cross-cultural music perception. Building upon the foundations established in previous chapters, we provide crucial empirical validation of how well computational approaches to multicultural music representation actually align with human perceptual understanding.

The contributions include: (1) a novel dataset of human similarity judgments across multiple cultural dimensions collected from 125 participants with diverse backgrounds, evaluating 1,130 audio pairs from nine musical datasets; (2) systematic evaluation of both traditional signal processing features and modern foundation models against human perception using multiple evaluation metrics; and (3) analysis of the interpretable factors that drive similarity perception across cultures, providing insights for developing more effective music AI systems. The complete dataset and implementation are made available for reproducibility[1].

This human-centered evaluation represents an essential step toward developing culturally aware music AI systems that can effectively serve diverse global populations while enhancing rather than diminishing musical cultural diversity.

6.2 Human Similarity Study

To understand how humans perceive musical similarity across different cultural traditions, we conducted a comprehensive online survey collecting similarity judgments from participants with diverse musical backgrounds and cultural origins. This section describes our survey methodology, the summary statistics of our study, and the participant demographics.

6.2.1 Survey Design and Methodology

Audio Dataset Selection: Our study encompasses nine musical datasets representing diverse cultural traditions as described in Section 2.10. From each dataset, we selected 52 representative audio clips of 20-second duration, resulting in 468 total clips spanning diverse instrumentation, vocal styles, and musical structures. The selection process prioritized musical diversity within each tradition while ensuring audio quality suitable for perceptual evaluation.

Pairwise Comparison Framework: Following established methodologies in music perception research [187, 188], we employed a pairwise comparison approach where participants evaluated

[1] https://github.com/pxaris/CCMSim

Category	Count
Total Participants	125
Unique Annotated Audio Pairs	1,130
Unique Annotated Audio Clips	463
Annotated Audio Pairs per Participant	10
Annotated Similarity Types per Audio Pair	3
Participants Unique Countries Of Origin	21
Participants Unique Music Training Levels	13
Participants Unique Familiar Music Cultures	58

Table 6.1. Summary Statistics of the Human Annotation Study. Comprehensive overview of participant demographics, study design parameters, and data collection metrics.

randomly selected pairs of 20-second audio clips. Each participant assessed 10 unique pairs, with pairs distributed to ensure comprehensive coverage across all dataset combinations while avoiding participant fatigue.

For each audio pair, participants provided ratings on three distinct similarity dimensions using a 9-point Likert scale, from 1 to 5 with step 0.5:

1. **Overall Musical Similarity**: "How similar are the two audio clips overall?"
2. **Cultural Similarity**: "How similar are the two audio clips in their cultural characteristics?"
3. **Recommendation-level Similarity**: "How probable is it for you to put the two audio clips in the same playlist?"

The three-dimensional framework allows examination of how different aspects of similarity align or diverge, particularly important for cross-cultural analysis where musical and cultural similarity may not coincide [189].

Survey Implementation: The survey was implemented as a web-based application ensuring cross-platform compatibility and ease of access. Participants were recruited through academic networks, social media, and music communities, with an emphasis on achieving demographic diversity.

6.2.2 Participant Demographics and Data Statistics

Our study collected responses from 125 participants, resulting in 1,130 unique annotated pairs covering 463 unique audio clips. Each participant annotated 10 audio pairs on three similarity dimensions. As shown in Table 6.1, the annotators came from 21 countries and reported 13 distinct levels of music training, while 58 different music cultures were identified as familiar by at least one participant.

Age and Gender Distribution: Figure 6.1 shows the participants' age distribution spanning from 18-64 years, with the majority falling within the 25-44 age range. The 25-34 group represents the largest segment (50 participants), followed by 35-44 (37) and 18-24 (29). With regards to the gender, there are 75 male participants (60.0%), 42 female participants (33.6%), and 8 participants identifying as other or preferring not to disclose (6.4%).

Musical Training and Expertise: Figure 6.2 presents the distribution of participants' musical backgrounds. Participants selected from predefined categories designed for this study, with an additional free-text option for custom responses. Music enthusiasts (18.2%), advanced amateur musicians (17.5%), and amateur musicians (16.9%) represent the largest groups. Notably, 13.0% of

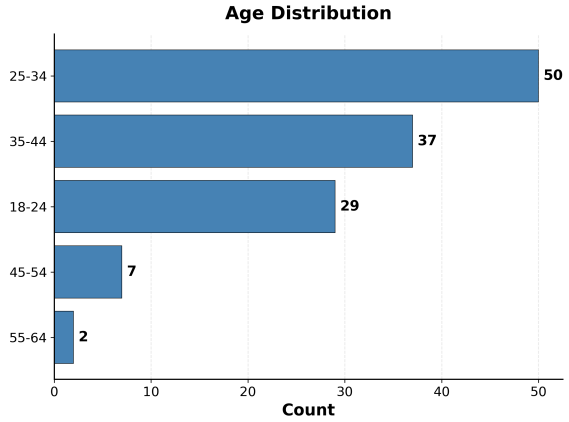

Figure 6.1. Age Distribution of Study Participants. Demographic breakdown of the 125 participants across different age ranges.

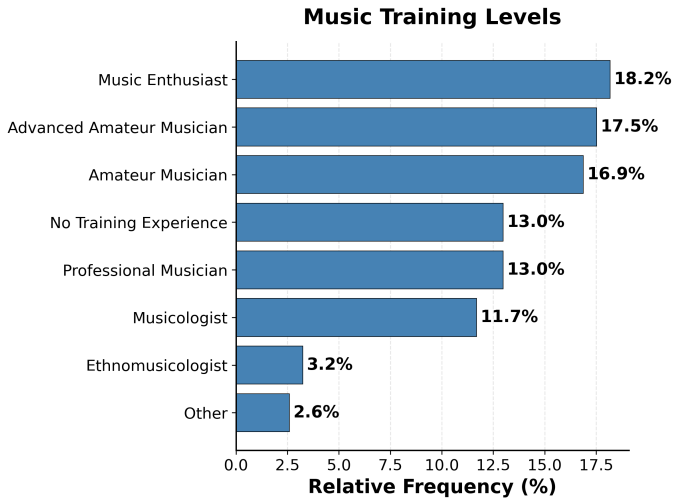

Figure 6.2. Distribution of Participants by Music Training Level. Participants' self-reported musical background and experience levels.

participants had no formal musical training, while 26.0% were professional musicians, musicologists, or ethnomusicologists, providing both expert validation and general population perspectives.

Cultural and Geographic Diversity: There are 21 unique countries of origin for the participants. Greece provided the largest group (62.4%), followed by China (5.6%), Italy (4.8%), France (4.0%), and the United Kingdom (4.0%). Despite the Greek majority, representation from diverse regions including Asia, Europe, and North America ensures cross-cultural validity.

Musical Cultural Familiarity: Figure 6.3 shows participants' self-reported familiarity with different musical traditions. Each participant was able to select multiple cultures, and the result is a long-tailed distribution with 58 distinct values. Greek music shows the highest familiarity (19.6%, consistent with participant distribution), followed by United States (13.4%) and United Kingdom (12.3%) music.

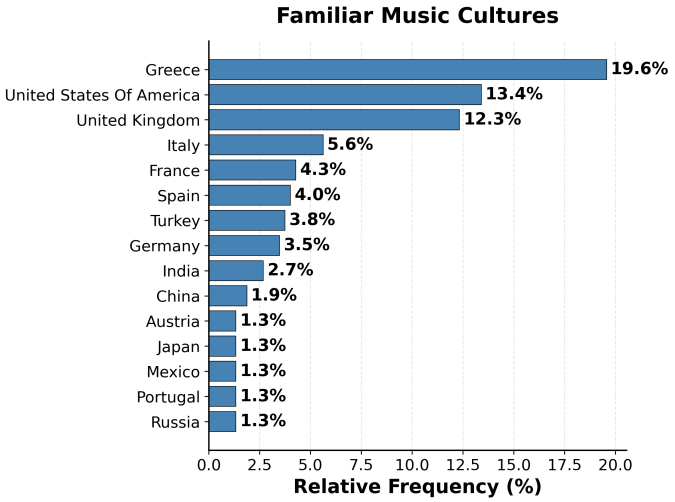

Figure 6.3. Participants' Familiarity with Musical Cultures. Distribution of participant self-reported familiarity with musical traditions (top-15 values).

6.3 Signal Processing Features for Cross-Cultural Music Analysis

Music similarity assessment across cultures requires capturing the multi-dimensional nature of musical perception while addressing the inherent challenges of cross-cultural analysis. Traditional approaches often focus on single dimensions or simple feature concatenation, missing nuanced relationships between musical aspects [190]. In this work we utilize a multi-dimensional framework treating rhythm, melody, harmony, and timbre as distinct but complementary dimensions, each employing specialized feature extraction and similarity computation methods that preserve unique characteristics and temporal dynamics. Complete mathematical formulations and implementation details are provided in the Appendix C.

6.3.1 Multi-Dimensional Feature Framework

Our framework addresses the complexity of cross-cultural music similarity by utilizing four musical dimensions:

Melody Analysis employs the PYIN algorithm [56] for robust F0 extraction from polyphonic audio, treating fundamental frequency as the dominant melodic skeleton. For robust cross-cultural melodic analysis, we implement dual-resolution pitch class representations. Given clean F0 extraction $\mathbf{f}_0^{\text{clean}}$ from PYIN, we compute MIDI-like numbers as:

$$m = 12 \log_2 \left(\frac{f}{440} \right) + 69, \tag{6.1}$$

where f represents the fundamental frequency in Hz. Quarter-tone and semitone pitch classes are calculated as:

$$pc_{\text{quarter}} = \lfloor (2m) \bmod 24 \rfloor, \tag{6.2}$$

$$pc_{\text{semi}} = \lfloor m \bmod 12 \rfloor, \tag{6.3}$$

where pc_{quarter} provides 24 bins per octave for microtonal analysis and pc_{semi} provides 12 bins for traditional Western analysis. Melodic intervals I_i are computed in quarter-tone units between consecutive F0 values to capture microtonal ornamentations:

$$I_i = 24 \log_2 \left(\frac{f_{0,i+1}}{f_{0,i}} \right), \tag{6.4}$$

where $f_{0,i}$ and $f_{0,i+1}$ are consecutive fundamental frequency values.

Rhythm Analysis combines tempo and beat tracking using dynamic programming-based algorithms [60], onset detection through complex domain methods [191, 192], beat interval analysis for rhythmic regularity assessment, and tempogram analysis [193] capturing tempo variations across time. This unified approach measures both local rhythmic events and global temporal structure.

Harmony Analysis integrates CENS (chroma energy normalized statistics) features [194] at both 24-bin (quarter-tone) and 12-bin (semitone) resolutions for cross-cultural harmonic analysis, chord recognition through major/minor triad template matching, key estimation via the Krumhansl-Schmuckler algorithm [195], chord transition matrix analysis, and Tonnetz tonal centroid features [196]. This combination captures both local harmonic content and global tonal structure while accommodating different tuning systems.

Timbre Analysis combines 13 Mel-frequency cepstral coefficients (MFCCs) [59, 185] with their temporal dynamics through delta features, spectral shape characteristics including centroid, rolloff, bandwidth, and contrast [197], spectral flatness [198], and RMS energy analysis. Rather than temporal averaging, we preserve timbral complexity through comprehensive statistical feature vectors. For any time series $\mathbf{s} = [s_1, s_2, \ldots, s_{N_t}]$ where N_t represents the number of time frames, we compute:

$$\boldsymbol{\sigma}(\mathbf{s}) = [\mu_s, \sigma_s, \tilde{s}, q_{25}(s), q_{75}(s), \Delta_s, \min(s), \max(s)]^T, \tag{6.5}$$

where μ_s is the mean, σ_s is the standard deviation, $\tilde{s}$ is the median, $q_{25}(s)$ and $q_{75}(s)$ are the 25th and 75th percentiles, Δ_s is the range, and $\min(s)$, $\max(s)$ are the minimum and maximum values. This 8-dimensional representation captures distribution characteristics while avoiding information loss from temporal averaging.

6.3.2 Similarity Computation and Integration

Each dimension employs multi-component similarity measures using cosine similarity and statistical comparisons. Melody similarity S_{melody} emphasizes interval patterns:

$$S_{\text{melody}} = 0.3 \left(\frac{S_{pc,\text{quarter}} + S_{pc,\text{semi}}}{2} \right) + 0.4 S_I + 0.3 S_{stats}, \tag{6.6}$$

where $S_{pc,\text{quarter}}$ and $S_{pc,\text{semi}}$ represent quartertone and semitone pitch class similarities, S_I measures melodic interval pattern similarity, and S_{stats} compares F0 statistical characteristics. Weights are designed to emphasize key-invariant interval patterns (0.4) while balancing pitch class and statistical similarities (0.3 each), with pitch class weight equally divided between quartertone and semitone representations.

The rhythm similarity S_{rhythm} combines four equally weighted components:

$$S_{\text{rhythm}} = \frac{1}{4}(S_{\text{tempo}} + S_{\text{onset}} + S_{\text{beat}} + S_{\text{tempogram}}), \tag{6.7}$$

where S_{tempo} measures tempo similarity, S_{onset} captures onset pattern similarity, S_{beat} evaluates

beat interval consistency, and $S_{\text{tempogram}}$ compares tempo evolution patterns.

Harmony similarity S_{harmony} integrates five complementary harmonic aspects:

$$S_{\text{harmony}} = \frac{1}{5}(S_{\text{chroma}} + S_{\text{key}} + S_{\text{chord\_dist}} + S_{\text{chord\_trans}} + S_{\text{tonnetz}}), \tag{6.8}$$

where S_{chroma} compares chroma profiles, S_{key} measures key similarity, $S_{\text{chord\_dist}}$ evaluates chord distribution similarity, $S_{\text{chord\_trans}}$ analyzes chord transition patterns, and S_{tonnetz} compares tonal centroid features.

Timbre similarity S_{timbre} prioritizes MFCC characteristics and temporal dynamics:

$$S_{\text{timbre}} = 0.45S_{\text{MFCC}} + 0.45S_{\text{dynamics}} + 0.1S_{\text{spectral}}, \tag{6.9}$$

where S_{MFCC} measures MFCC distribution similarity, S_{dynamics} captures temporal evolution patterns through delta features, and S_{spectral} compares spectral shape characteristics.

The overall signal processing similarity S_{SP} integrates all four dimensions with equal weighting:

$$S_{\text{SP}} = \frac{1}{4}(S_{\text{melody}} + S_{\text{rhythm}} + S_{\text{harmony}} + S_{\text{timbre}}). \tag{6.10}$$

This multi-dimensional framework establishes a way for comparing signal processing features against both human perception and foundation model representations. It incorporates several adaptations for cross-cultural analysis including quarter-tone resolution for microtonal systems, statistical rather than averaged representations to preserve temporal complexity, and multiple complementary features within each domain. However, limitations remain particularly regarding Western bias in harmonic analysis components.

6.4 Foundation Models for Cross-Cultural Music Representation

We utilize the five state-of-the-art foundation models evaluated in Section 5.2 of the previous chapter along with the culturally-adapted models we introduced in Section 5.5.

Specifically, we selected:

- the **MERT-95M**[2] and **MERT-330M**[3] models, which employ masked acoustic modeling with dual teacher supervision from both acoustic and musical perspectives [40];
- the **CLAP-Music**[4], which specializes in musical content through music-only training data, and the **CLAP-Music&Speech**[5], which incorporates both music and speech data for potentially more robust audio representations;
- **Qwen2-Audio**[6], that represents the largest model in our evaluation; and
- **CultureMERT-95M**[7], developed through a two-stage continual pre-training strategy starting from MERT-95M, and **CultureMERT-TA-95M**[8], its alternative that was created using task arithmetic.

[2] https://huggingface.co/m-a-p/MERT-v1-95M
[3] https://huggingface.co/m-a-p/MERT-v1-330M
[4] https://huggingface.co/laion/larger_clap_music
[5] https://huggingface.co/laion/larger_clap_music_and_speech
[6] https://huggingface.co/Qwen/Qwen2-Audio-7B
[7] https://huggingface.co/ntua-slp/CultureMERT-95M
[8] https://huggingface.co/ntua-slp/CultureMERT-TA-95M

Regarding the audio preprocessing and the representation extraction strategies, we follow the implementation described in Section 5.3 along with the existing research [199].

- **MERT models (MERT-95M and MERT-330M):** We use 30-second windows at 24kHz sampling rate, converted to mono channel. Representations are extracted by averaging hidden states across the middle four layers (layers 4-7 for MERT-95M, layers 10-13 for MERT-330M), then computing temporal means.
- **CLAP models (CLAP-Music and CLAP-Music&Speech):** We employ 10-second windows at 48kHz sampling rate. Representations are extracted from the audio projection layer processing average-pooled final hidden states.
- **Qwen2-Audio:** We use 30-second windows at 16kHz sampling rate. Representations are obtained by averaging last hidden state embeddings across all audio tower layers using minimal text prompts (`<|audio_bos|><|AUDIO|><|audio_eos|>`) to activate audio understanding while minimizing text biases.
- **CultureMERT models (CultureMERT-95M and CultureMERT-TA-95M):** We follow the same preprocessing as MERT models with 30-second windows at 24kHz sampling rate, converted to mono channel. Representation extraction follows the MERT strategy as well, averaging middle-layer representations to leverage both original MERT capabilities and learned cultural adaptations.

6.4.1 Similarity Computation from Foundation Model Representations

For each foundation model, we compute pairwise similarity between audio representations using cosine similarity. Specifically, given foundation model representations $\mathbf{r}_1, \mathbf{r}_2 \in \mathbb{R}^d$ for two audio clips, similarity is computed as:

$$S_{FM}(\mathbf{r}_1, \mathbf{r}_2) = \frac{\mathbf{r}_1 \cdot \mathbf{r}_2}{||\mathbf{r}_1||_2 \cdot ||\mathbf{r}_2||_2}.$$

This similarity measure is computed for all audio pairs in our dataset, enabling comparison with human annotations across the three similarity dimensions (overall musical, cultural, and recommendation-level).

The evaluation framework provides insights into how different foundation models capture cross-cultural musical relationships, complementing the interpretable signal processing analysis and establishing a direct comparison with human similarity perception.

6.5 Methodology

This section details our experimental methodology for evaluating computational similarity measures against human perception across diverse musical cultures. We systematically compare signal processing features and foundation model representations using multiple evaluation metrics and comprehensive data preprocessing approaches.

6.5.1 Data Preparation and Normalization

Our experimental framework addresses individual rating scale differences and handles outliers through a two-stage normalization process designed to ensure fair comparison across computational methods and human judgments.

Human Annotation Normalization

To mitigate individual participant biases in similarity rating scales, we apply mean-centering normalization to human annotations. For each participant p, we compute their global mean rating μ_p across all similarity dimensions and audio pairs:

$$\mu_p = \frac{1}{N_p} \sum_{i=1}^{N_p} s_{p,i}, \tag{6.11}$$

where N_p is the total number of ratings provided by participant p, and $s_{p,i}$ represents their i-th similarity rating. Each participant's ratings are then mean-centered:

$$s_{p,i}^{centered} = s_{p,i} - \mu_p. \tag{6.12}$$

This approach ensures that participants with consistently high or low rating tendencies contribute equally to the evaluation, addressing systematic biases in individual rating behaviors [200].

After mean-centering, we aggregate multiple participant ratings for each audio pair by computing the mean across annotators, then apply robust min-max normalization using the 5th and 95th percentiles (p_5 and p_{95}) of the mean-centered distribution:

$$s_{human}^{norm} = \text{clip}\left(\frac{s_{human}^{centered} - p_5}{p_{95} - p_5}, 0, 1 \right), \tag{6.13}$$

where the clip function constrains values to the $[0, 1]$ range, effectively handling outliers beyond the percentile bounds.

Computational Similarity Normalization

For computational similarities from both signal processing features and foundation models, we apply the same robust min-max normalization strategy. Foundation model similarities, originally computed using cosine similarity in the $[-1, 1]$ range, are first transformed to $[0, 1]$ using:

$$s_{cosine}^{scaled} = \frac{1 + \cos(\theta)}{2}, \tag{6.14}$$

where θ represents the angle between feature vectors. This transformation preserves the relative ordering while ensuring compatibility with human ratings.

Subsequently, all computational similarities undergo robust normalization using their respective 5th and 95th percentiles, ensuring consistent preprocessing across human annotations and computational methods.

6.5.2 Evaluation Metrics

We employ a comprehensive evaluation framework using five complementary metrics, each capturing different aspects of similarity alignment between computational methods and human perception [124].

Kendall's Tau (τ): Measures rank correlation using concordant and discordant pairs, providing a robust estimate of ranking agreement that is less sensitive to outliers:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}, \tag{6.15}$$

where n_c and n_d represent the number of concordant and discordant pairs, respectively, and n is the total number of observations [201].

Spearman Rank Correlation (ρ): Evaluates monotonic relationships by comparing rank orderings, making it robust to non-linear transformations of similarity scales:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \tag{6.16}$$

where d_i represents the difference between ranks for the i-th observation [202].

Normalized Discounted Cumulative Gain (NDCG): Adapted from information retrieval, NDCG evaluates the quality of similarity rankings by emphasizing correct identification of highly similar pairs. For each query audio, we rank all other clips according to computational similarity and compare against human similarity as relevance scores:

$$\text{NDCG} = \frac{\sum_{i=1}^{n} \frac{2^{rel_i} - 1}{\log_2(i+1)}}{\sum_{i=1}^{n} \frac{2^{rel_i^*} - 1}{\log_2(i+1)}}, \tag{6.17}$$

where rel_i is the relevance of the item at position i (computational method similarity), and rel_i^* represents the ideal ranking (human similarity) [203].

Triplet Agreement: Evaluates whether computational methods preserve relative similarity orderings within triplets of audio clips. For each triplet (a, b, c) where human judgments indicate $s_{human}(a, b) > s_{human}(a, c)$ by a margin $\epsilon = 0.1$, we assess whether the computational method produces the same ordering:

$$\text{Agreement} = \frac{1}{N} \sum_{triplets} \mathbf{1}[\text{sign}(\Delta s_{comp}) = \text{sign}(\Delta s_{human})], \tag{6.18}$$

where $\Delta s_{comp} = s_{comp}(a, b) - s_{comp}(a, c)$ and $\Delta s_{human} = s_{human}(a, b) - s_{human}(a, c)$ represent the similarity differences for computational and human judgments respectively, and N is the total number of valid triplets. This metric is particularly relevant for retrieval applications where relative ranking matters more than absolute similarity values [204].

6.5.3 Feature Contribution Analysis

To understand the relative importance of different musical dimensions in predicting human perception and foundation model behavior, we employ linear regression analysis using signal processing features as predictors.

Linear Regression Framework

We formulate the prediction task as a multiple linear regression problem where signal processing features serve as independent variables predicting both human similarity ratings and foundation model similarities as dependent variables:

$$\begin{aligned} s_{predicted} = \beta_0 &+ \beta_1 f_{melody} + \beta_2 f_{rhythm} \\ &+ \beta_3 f_{harmony} + \beta_4 f_{timbre} + \epsilon, \end{aligned} \tag{6.19}$$

where β_i represents the regression coefficients indicating feature importance, and ϵ is the residual error term.

Multiple Random Split Strategy

To ensure robust coefficient estimates and assess the stability of feature importance rankings, we employ multiple train-test splits with different random seeds. For each target variable (three human dimensions and seven foundation models), we perform 5 iterations with different random partitions of the data into training (80%) and testing (20%) sets.

For each iteration k, we train a linear regression model on the training set and evaluate it on the held-out test set. We then compute coefficient means and standard deviations across the 5 iterations:

$$\bar{\beta}_i = \frac{1}{K} \sum_{k=1}^{K} \beta_{i,k}, \tag{6.20}$$

$$\sigma_{\beta_i} = \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} (\beta_{i,k} - \bar{\beta}_i)^2}, \tag{6.21}$$

where $K = 5$ represents the number of random splits, and $\beta_{i,k}$ is the coefficient for feature i in iteration k. Model performance is assessed using Mean Absolute Error (MAE) on the held-out test sets, averaged across all iterations:

$$\text{MAE} = \frac{1}{K} \sum_{k=1}^{K} \text{MAE}_k, \tag{6.22}$$

$$\text{MAE}_k = \frac{1}{n_{test,k}} \sum_{i=1}^{n_{test,k}} |s^{(i)}_{predicted,k} - s^{(i)}_{target,k}|, \tag{6.23}$$

where $n_{test,k}$ is the number of test samples in iteration k.

6.5.4 Ensemble Methods

To investigate the potential for combining signal processing features and foundation model representations for improved human similarity prediction, we implement ensemble regression using both linear and gradient boosting approaches. We follow the same multiple random split strategy as in the feature contribution analysis, performing 5 iterations with different random partitions of the data into training (80%) and testing (20%) sets to ensure robust performance estimates.

Linear Regression: Serves as a baseline ensemble approach, combining all available signal processing and foundation model features in a single linear model [205]. Feature importance is directly interpretable through regression coefficients, enabling analysis of which feature combinations contribute most effectively to human similarity prediction.

LightGBM: A gradient boosting framework that can capture non-linear relationships between features and targets [206]. We employ early stopping with validation monitoring to prevent overfitting, using parameters optimized for regression tasks including a learning rate of 0.05 and maximum of 1000 boosting rounds with early stopping after 50 rounds without improvement.

Performance Evaluation

Ensemble models are evaluated using the same comprehensive metric suite as individual computational methods (Kendall's Tau, Spearman correlation, NDCG, and Triplet Agreement), enabling

direct comparison of improvement over single-method approaches. Performance metrics are averaged across the 5 random split iterations to provide robust estimates with confidence intervals.

Feature importance analysis identifies which combinations of signal processing and foundation model features contribute most effectively to human similarity prediction. The ensemble framework provides insights into the complementary strengths of interpretable signal processing features and learned foundation model representations, establishing whether hybrid approaches can achieve superior alignment with human cross-cultural music perception compared to individual computational methods.

6.5.5 Cross-Cultural Analysis Framework

To examine cross-cultural patterns in computational similarity measures, we aggregate individual pair similarities at the dataset level, creating 9×9 similarity matrices representing relationships between musical traditions from both human and computational perspectives.

For each dataset pair (i, j), we compute mean similarity across all annotated pairs belonging to those traditions, providing a macro-level view of cultural relationships. For computational methods, we also evaluate cultural discrimination using the complete 468×468 similarity matrix (approximately 100,000 unique pairs), which provides a more comprehensive assessment of cross-cultural patterns beyond the human-annotated subset.

We analyze within-tradition versus between-tradition discrimination using a distance-based separation ratio that measures how well computational methods distinguish between different musical cultures.

Distance-Based Separation Ratio

Since computational similarities are proximity measures, we convert them to distance measures for more intuitive separation analysis. For a similarity matrix $\mathbf{S}$, we compute the corresponding distance matrix as $\mathbf{D} = \mathbf{1} - \mathbf{S}$, where $\mathbf{1}$ is a matrix of ones with the same dimensions as $\mathbf{S}$.

We then calculate the mean intra-tradition distance (diagonal elements) and mean inter-tradition distance (off-diagonal elements):

$$D_{\text{intra}} = \frac{1}{N} \sum_{i=1}^{N} D_{ii}, \tag{6.24}$$

$$D_{\text{inter}} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} D_{ij}, \tag{6.25}$$

where N is the number of musical traditions (datasets).

The distance-based separation ratio is defined as:

$$R_{\text{separation}} = \frac{D_{\text{inter}}}{D_{\text{intra}}}. \tag{6.26}$$

Higher separation ratios indicate better discrimination between musical cultures, as they reflect larger distances between different traditions relative to cohesion within each tradition. Methods that fail to capture cultural distinctions would show similar distances both within and between datasets, yielding ratios close to 1.

We compute separation ratios for both human-annotated pairs (to enable direct comparison with human judgments) and all possible audio pairs (to better assess computational methods' performance), providing complementary perspectives on cross-cultural discrimination capabilities.

6.6 Results and Discussion

This section presents our comprehensive evaluation of computational similarity measures against human perception across diverse musical cultures. We begin with analysis of human similarity judgments, followed by systematic comparison of signal processing features and foundation models, discriminability assessment, interpretable feature analysis, and ensemble regression results.

6.6.1 Human User Study Results

Our human similarity study collected 1,130 unique audio pair annotations from 125 participants across three similarity dimensions: overall musical similarity, cultural similarity, and recommendation-level similarity. The results reveal important insights into how humans perceive cross-cultural musical relationships and the consistency of different similarity conceptualizations.

Inter-Dimensional Correlation Analysis

The Spearman correlations between the three human similarity dimensions reveal strong relationships, indicating substantial overlap while still allowing participants to distinguish meaningfully between different aspects of similarity. Overall musical similarity correlates highly with both cultural similarity ($\rho = 0.78$) and recommendation-level similarity ($\rho = 0.77$), while cultural and recommendation-level similarities also show strong correlation ($\rho = 0.74$). These high correlations demonstrate that while the three dimensions capture closely related aspects of musical similarity, they also provide complementary information about human perception to a certain degree. This can have potential benefits for music information retrieval systems that target specific dimensions of similarity.

Cross-Cultural Dataset Relationships

Figure 6.4 presents the dataset-level cultural similarity matrix, revealing systematic patterns in how participants perceive relationships between different musical traditions. We focus on cultural similarity as it most directly captures participants' perception of tradition-based relationships, distinct from purely musical or personal preference considerations. Several notable clusters emerge from the data:

Indian Classical Music Cluster: Hindustani and Carnatic traditions show the highest mutual similarity (0.69), reflecting their shared historical and theoretical foundations. Both traditions also show moderate similarity with other non-Western traditions.

Mediterranean/Middle Eastern Cluster: Arab-Andalusian, Lyra (Greek) and Turkish-makam exhibit elevated mutual similarities, with Arab-Andalusian showing particularly strong connections to Turkish-makam (0.65). This clustering reflects historical cultural exchanges across the Mediterranean region.

Western Music Separation: MagnaTagATune and FMA-medium show high mutual similarity (0.49) and generally lower similarities with traditional music datasets, suggesting that participants clearly distinguish Western popular/commercial music from traditional world music.

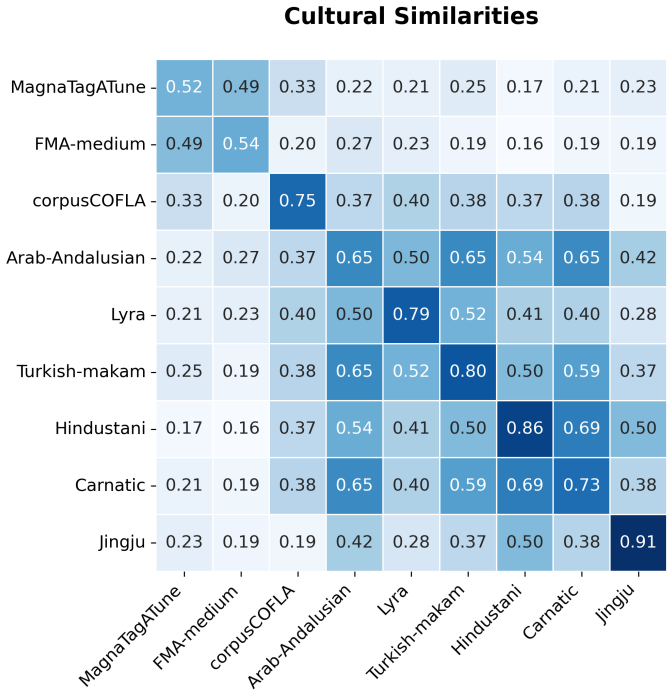

Figure 6.4. Cultural Similarity Matrix Across Datasets. Heat map visualization of human-perceived cultural similarity ratings. Values represent mean cultural similarity ratings aggregated across all participant annotations for pairs between and within datasets. Clear cultural clusters emerge, with higher similarities (darker blue) indicating stronger cultural relationships.

Chinese Opera and Flamenco Distinctiveness: Jingju (Beijing Opera) shows the highest within-tradition similarity (0.91) and generally lower cross-cultural similarities, indicating its unique characteristics that participants found difficult to relate to other traditions. Similarly, corpusCOFLA (Flamenco) demonstrates high within-tradition similarity (0.75) while showing substantially lower cross-cultural similarities (0.40 or below), reflecting its distinctive musical characteristics that participants perceived as culturally distinct from other traditions.

These patterns align with musicological understanding of cultural relationships and historical influences, validating the meaningfulness of human cross-cultural similarity judgments.

Multidimensional Scaling Visualization

The MDS [207] visualization in Figure 6.5 provides a spatial representation of musical traditions based on recommendation-level similarity distances. The two-dimensional projection reveals meaningful cultural groupings and relationships. The dotted circles around each point represent the internal diversity of each musical tradition, with larger circles indicating datasets where participants perceived greater variation within the tradition itself (diagonal elements on a cross-cultural matrix).

The visualization shows clear separation between Western commercial music (MagnaTagATune, FMA-medium) on the right side and traditional world music on the left. Within the traditional music cluster, we observe sub-groupings that correspond to geographical and cultural relationships: Indian classical traditions (Hindustani, Carnatic) cluster in the upper left, while Mediterranean/Middle Eastern traditions (corpusCOFLA, Arab-Andalusian, Lyra, Turkish-makam) group in the center and lower left. Jingju appears distinctly separated at the top, consistent with its

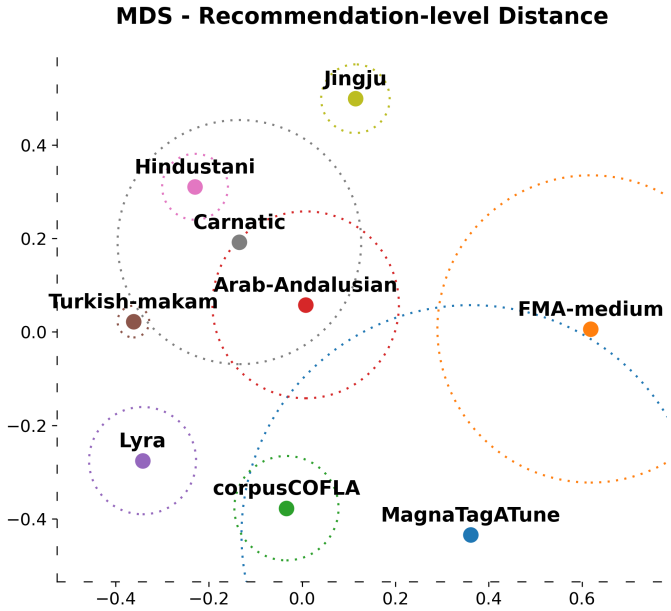

Figure 6.5. Multidimensional Scaling Visualization of Musical Datasets. 2D projection based on recommendation-level similarity distances derived from human annotations, revealing cultural clustering patterns across nine musical traditions. Dotted circles around each dataset represent internal diversity (inverse self-similarity) within each musical tradition.

unique musical characteristics.

Within vs. Cross-Dataset Similarity Distributions

Figure 6.6 compares the distributions of similarity ratings for within-dataset pairs versus cross-dataset pairs across all three similarity dimensions. The results demonstrate clear discrimination between musical cultures.

For all three dimensions, within-dataset similarities show significantly higher means (Overall: $\mu = 0.650$, Cultural: $\mu = 0.732$, Recommendation: $\mu = 0.701$) compared to cross-dataset similarities (Overall: $\mu = 0.361$, Cultural: $\mu = 0.360$, Recommendation: $\mu = 0.368$). Cultural similarity shows the largest separation ($\Delta\mu = 0.372$), followed by recommendation-level similarity ($\Delta\mu = 0.333$) and overall similarity ($\Delta\mu = 0.289$).

The clear separation between within and cross-dataset distributions confirms that participants consistently recognize and distinguish between different musical cultures, providing a strong foundation for evaluating computational methods' ability to capture cross-cultural musical relationships.

6.6.2 Signal Processing Features and Foundation Models vs. Human Perception

We systematically evaluated both signal processing features and foundation models against human similarity judgments using five complementary metrics across the three similarity dimensions. This comprehensive evaluation provides insights into which computational approaches best align with human cross-cultural music perception.

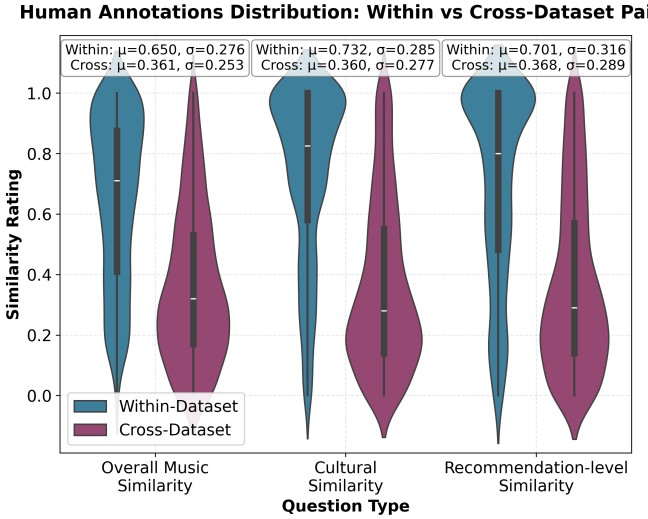

Figure 6.6. Distribution Comparison of Human Similarity Ratings. Violin plots with embedded box plots comparing within-dataset pairs versus cross-dataset pairs across the three similarity dimensions. Statistical parameters (μ: mean, σ: standard deviation) demonstrate clear separation between within and cross-cultural similarities.

Comprehensive Performance Evaluation

Table 6.2 presents the complete evaluation results for all computational methods across the five metrics and three similarity dimensions. The results reveal distinct performance patterns between signal processing features and foundation models.

Signal Processing Features Performance: Among signal processing features, melody consistently demonstrates superior performance across all metrics and similarity dimensions. Melody achieves the best MAE scores (29.5-30.9%) and shows the strongest correlations with human judgments (Spearman $\rho = 0.14 - 0.15$, Kendall $\tau = 0.12 - 0.13$). This confirms melody's central role in human music similarity perception across cultures.

In contrast, rhythm, harmony, and timbre features show limited alignment with human perception, with correlations near zero or slightly negative. Rhythm and harmony features particularly struggle, suggesting that our signal processing implementations may not adequately capture the complex rhythmic and harmonic relationships that humans perceive across different musical traditions.

Foundation Models Performance: Foundation models generally outperform signal processing features across most metrics, with CLAP-Music&Speech emerging as the top performer, achieving the highest triplet agreement (62.6-64.9%) and NDCG scores (88.0-89.8%). However, melody features remain competitive, achieving the best MAE scores (29.5-30.9%) and strong correlation performance. The superior performance of CLAP-Music&Speech over CLAP-Music highlights the synergy between melody and speech modalities, as both capture complementary aspects of musical expression, melodic patterns and intonation patterns [3, 208], that are particularly valuable for cross-cultural music understanding given the vocal traditions prominent in many world music cultures.

This performance advantage, however, comes with a trade-off in cultural discriminability. While CLAP-Music&Speech excels at aligning with human similarity judgments, Qwen2-Audio demonstrates superior cultural boundary detection (Table 6.3), suggesting that models optimized for universal musical understanding may sacrifice some discriminative power between cultural tradi-

Method	Triplet Agr. ↑ (%)			NDCG ↑ (%)			Spearman ρ ↑ (-1, 1)			Kendall τ ↑ (-1, 1)			MAE ↓ (%)		
Similarity type	**Overall**	**Cultural**	**Recomm.**	**Overall**	**Cultural**	**Recomm.**	**Overall**	**Cultural**	**Recomm.**	**Overall**	**Cultural**	**Recomm.**	**Overall**	**Cultural**	**Recomm.**
Signal Processing Features															
Melody	61.5	61.1	60.7	88.4	87.6	86.8	0.15	**0.14**	**0.15**	**0.14**	**0.12**	**0.13**	**29.5**	**30.5**	**30.9**
Rhythm	51.3	52.1	50.3	85.8	84.0	84.0	-0.00	-0.01	-0.02	-0.00	-0.01	-0.02	32.5	34.3	34.6
Harmony	51.8	50.8	50.7	85.3	83.4	83.6	0.02	-0.00	0.02	0.02	0.00	0.02	32.1	33.5	34.3
Timbre	54.2	54.7	55.6	86.1	84.8	85.3	-0.03	0.04	0.04	-0.03	0.03	0.03	35.2	36.4	36.3
Foundation Models															
MERT-95	59.8	59.7	60.0	88.2	87.1	87.3	0.06	0.09	0.10	0.05	0.08	0.08	31.3	32.4	32.3
CultureMERT	56.3	57.0	57.4	86.8	86.2	86.4	0.04	0.08	0.08	0.03	0.06	0.07	33.0	34.1	34.4
CultureMERT-TA	55.1	55.8	56.5	86.6	86.0	86.4	0.02	0.06	0.06	0.01	0.05	0.05	33.6	34.6	34.8
MERT-330	57.6	57.3	58.7	87.8	86.5	86.9	0.08	0.05	0.09	0.06	0.04	0.08	35.0	35.6	35.7
CLAP-Music	55.6	56.0	54.8	86.8	85.3	84.8	0.05	0.03	-0.01	0.04	0.02	-0.01	40.9	41.7	41.6
CLAP-Music&Speech	**64.9**	**62.6**	**64.9**	**89.8**	**88.0**	**88.6**	**0.16**	0.11	0.14	**0.14**	0.09	0.12	29.6	30.8	**30.9**
Qwen2-Audio	58.4	58.0	59.5	88.0	86.5	86.9	0.05	0.06	0.08	0.04	0.05	0.08	36.7	37.3	37.3

Table 6.2. Comprehensive Evaluation of Signal Processing Features and Foundation Models. Performance comparison against human similarity judgments across three similarity dimensions (overall musical, cultural, and recommendation-level). Values are shown as percentages (%) for Triplet Agreement, NDCG, and MAE, and as correlation values for Spearman and Kendall metrics. Arrows indicate whether higher (↑) or lower (↓) values represent better performance, with best performance within each similarity dimension and metric shown in bold.

tions.

MERT-95 demonstrates consistent performance across all similarity dimensions, while the larger MERT-330 model shows mixed results, sometimes underperforming its smaller counterpart, consistent with our findings in Section 5.4. The culturally adapted CultureMERT variants underperform their base MERT-95 model, which is logical given our participant pool's predominantly Western musical backgrounds. Since CultureMERT was specifically adapted toward non-Western cultures (Greek, Turkish, Indian traditions), its lower alignment with our listener judgments reflects the influence of listener cultural background on evaluation results rather than model inadequacy.

Comparative Analysis Across Methods

Figure 6.7 provides a radar plot comparison of the top-performing methods, averaged across similarity dimensions and normalized for visualization. The radar plot clearly illustrates CLAP-Music&Speech's superior performance across most metrics, particularly excelling in Triplet Agreement and NDCG measures. Melody emerges as the strongest signal processing feature, showing balanced performance across all evaluation dimensions. The plot also reveals that foundation models generally maintain more consistent performance profiles compared to signal processing features, which show greater variability across different metrics.

The correlation-based metrics (Spearman ρ and Kendall τ) show lower absolute values across all methods, reflecting the inherent challenges in capturing the complex, non-linear relationships that characterize human cross-cultural music similarity perception. However, the relative rankings remain consistent, with CLAP-Music&Speech and melody maintaining their leading positions. These results demonstrate that while foundation models achieve superior alignment with human perception, the gap between computational methods and human judgment remains substantial, indicating significant room for improvement in cross-cultural music similarity modeling.

6.6.3 Cross-Cultural Discriminability Analysis

To assess how well computational methods distinguish between different musical traditions, we analyze their cross-dataset discriminability using distance-based separation ratios. We compare these ratios with the respective human perception values to provide insights into how effectively computational approaches can capture cultural boundaries.

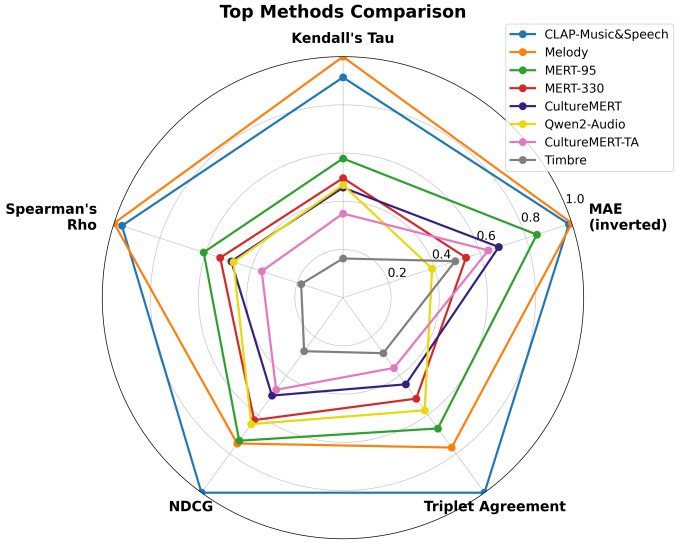

Figure 6.7. Radar Plot Comparison of Top-Performing Computational Methods. Performance visualization averaged across three similarity dimensions, with metrics normalized to [0,1] scale where higher values indicate better performance (MAE is inverted).

Separation Ratio Results

Table 6.3 presents the distance-based separation ratios for humans and all computational methods across both annotated pairs and the complete audio dataset. The results reveal significant differences in cultural discrimination capabilities between human perception and computational approaches, highlighting a fundamental trade-off between universal musical understanding and cultural discriminability.

Human Baseline Performance: Human similarity judgments demonstrate superior cultural discrimination, with the Cultural similarity dimension achieving the highest separation ratio (2.361), followed by Recommendation-level (2.106) and Overall musical similarity (1.803). This confirms that humans consistently recognize and distinguish between different musical traditions, with cultural similarity showing the strongest discrimination as expected.

Signal Processing Features: Among signal processing features, melody again emerges as the most discriminative (1.276 for annotated pairs), consistent with its superior performance in direct human alignment metrics. Rhythm, harmony, and timbre show limited discrimination capabilities, with ratios close to 1.0, indicating they struggle to distinguish between musical cultures effectively. Notably, rhythm shows a separation ratio slightly below 1.0 for annotated pairs (0.989), indicating limited cultural discrimination capability in this subset.

Foundation Models Performance: Foundation models substantially outperform signal processing features in cultural discrimination, but reveal an important trade-off between universal musical understanding and cultural discriminability. Qwen2-Audio achieves the highest separation ratios among all computational methods (1.579 for annotated pairs, 1.602 for all pairs), demonstrating superior ability to distinguish between musical traditions. In contrast, CLAP-Music&Speech, while excelling at human similarity alignment, shows more modest discrimination performance (1.366 for annotated pairs), suggesting that models optimized for universal cross-cultural musical understanding may sacrifice some discriminative power between cultural boundaries.

The culturally adapted CultureMERT variants underperform their base MERT-95 model, which may reflect the significant influence of listener cultural background on evaluation results. Since

Method	Annotated Pairs	All Pairs
Human Similarities		
Overall Music	1.803	—
Cultural	2.361	—
Recommendation-level	2.106	—
Signal Processing Features		
Melody	1.276	1.180
Rhythm	0.989	1.037
Harmony	1.018	1.031
Timbre	1.025	1.018
Foundation Models		
MERT-95	1.280	1.209
CultureMERT	1.259	1.180
CultureMERT-TA	1.262	1.169
MERT-330	1.401	1.298
CLAP-Music	1.415	1.217
CLAP-Music&Speech	1.366	1.318
Qwen2-Audio	**1.579**	**1.602**

Table 6.3. Cross-Cultural Discrimination Analysis Using Distance-Based Separation Ratios. Comparison of cultural boundary detection capabilities between humans and all computational methods. Higher values indicate better discrimination between musical traditions. Annotated pairs use only human-annotated audio pairs (1, 130), while all pairs use the complete similarity matrix ($\sim$ 100k pairs).

CultureMERT was specifically adapted toward non-Western cultures (Greek, Turkish, Indian traditions) and our participant pool predominantly consists of listeners with Western musical backgrounds, this apparent underperformance could indicate that the models have developed different similarity representations that may not align with our predominantly Western evaluation benchmark.

Cross-Cultural Similarity Patterns

Figure 6.8 presents the dataset-level similarity matrix for Qwen2-Audio, the best-performing method in terms of cultural discrimination. This visualization can be directly compared with Figure 6.4, which shows human cultural similarity perceptions for the same dataset pairs.

The Qwen2-Audio similarity matrix reveals several compelling patterns that partially echo human perception patterns. The model demonstrates strong within-tradition clustering (diagonal values), with Jingju and Hindustani showing the highest self-similarity (0.89), followed by corpus-COFLA (0.86), Turkish-makam (0.85) and Lyra (0.85). This ranking partially aligns with human judgments, where these traditions also demonstrate high within-tradition similarity.

Notably, Qwen2-Audio shows exceptional discrimination for Jingju (Beijing Opera), which emerges as the most distinctive tradition in the computational similarity space. Jingju's cross-cultural similarities are substantially lower than all other traditions, ranging from 0.57-0.67 compared to the 0.75-0.85 range typical for other cross-cultural pairs. This strong discrimination may reflect the model's training on diverse audio data that includes Chinese traditional music, resulting in embedding space organization that treats Chinese traditional opera as a distinct category. However, this specialization toward specific cultural discrimination may come at the cost of broader cross-cultural alignment capabilities, as evidenced by its lower performance in human similarity

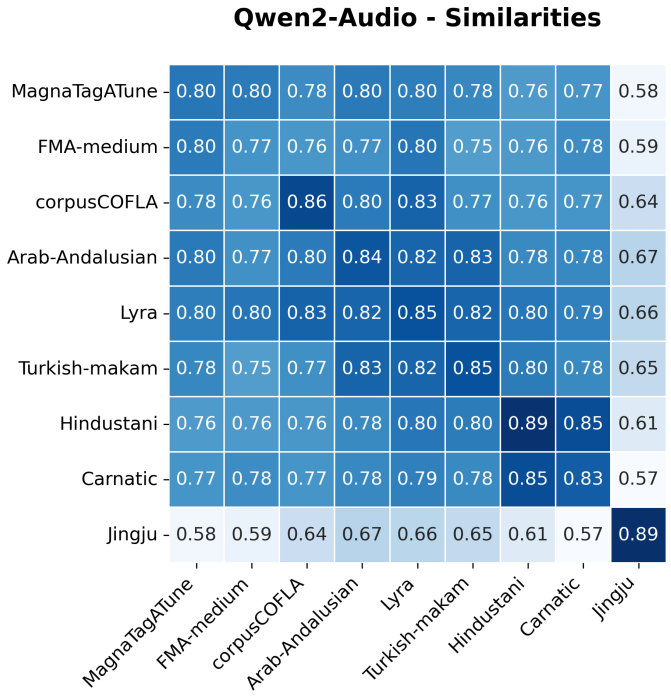

Figure 6.8. Cross-Cultural Similarity Matrix for Qwen2-Audio Foundation Model. Heat map visualization showing computational similarity patterns between musical traditions as captured by the Qwen2-Audio model. Darker colors indicate higher similarities.

prediction compared to more universal models like CLAP-Music&Speech.

Mirroring human perception patterns, the Indian classical music pair (Hindustani-Carnatic) shows elevated mutual similarity, the Mediterranean/Middle Eastern cluster (Arab-Andalusian, Lyra, Turkish-makam) exhibits clear interconnections, and the corpusCOFLA shows more distinctive characteristics. In contrast to human perception, Western commercial music traditions (MagnaTagATune-FMA-medium) demonstrate high similarity with non-Western traditions except for Jingju.

The distance-based separation ratio, while providing valuable insights into cultural discrimination capabilities, treats all cross-cultural distinctions equally, which may be overly restrictive for traditions that share significant musical characteristics. Future evaluation frameworks should consider incorporating cultural proximity into discrimination metrics to provide more nuanced assessment of cross-cultural music understanding.

Detailed results on cross-cultural similarity patterns for both human annotations and computational methods can be found in Appendix D.

6.6.4 Feature Contribution Analysis

To understand which musical dimensions drive similarity perception in both human judgments and foundation model representations, we performed linear regression analysis using signal processing features as predictors. This interpretability analysis reveals the relative importance of melody, rhythm, harmony, and timbre in explaining similarity patterns across different evaluation contexts.

Figure 6.9 presents the linear regression weights when predicting human similarity judgments and foundation model similarities using the four signal processing features. The analysis reveals distinct patterns in how different targets weight various musical dimensions, providing insights into

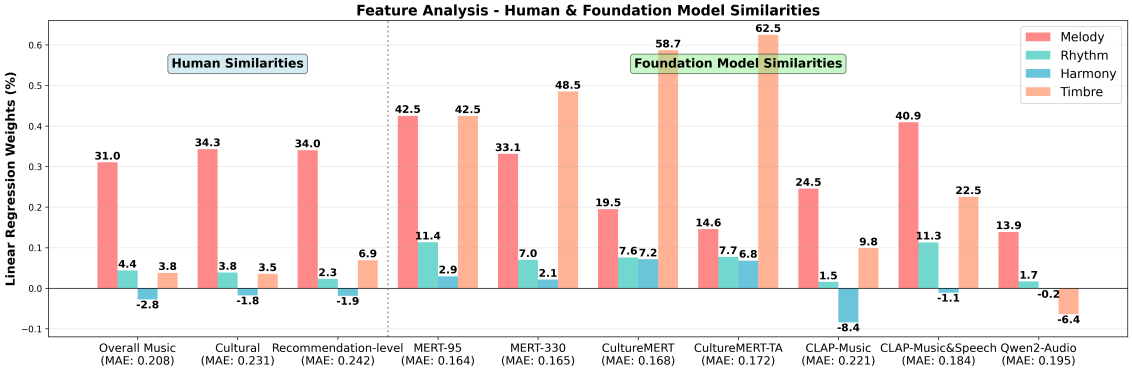

Figure 6.9. Linear Regression Weights for Signal Processing Features. Bar charts showing the contribution of signal processing features (melody, rhythm, harmony, timbre) in predicting human similarity judgments and foundation model similarities. Positive weights indicate that higher feature similarity contributes to higher predicted similarity, with MAE values in parentheses indicating prediction accuracy.

the underlying factors that drive similarity perception.

Human Similarity Patterns

Human similarity judgments across all three dimensions show remarkably consistent patterns in their relationship to signal processing features. Melody emerges as the dominant predictive factor, with weights of 31.0% (Overall Music), 34.3% (Cultural), and 34.0% (Recommendation-level), confirming melody's central role in human cross-cultural music perception. This consistency across similarity dimensions suggests that melodic content serves as a fundamental basis for how humans assess musical relationships, regardless of whether they focus on overall musical characteristics, cultural identity, or personal preference.

Rhythm shows modest positive contributions (4.4%, 3.8%, 2.3%) across all human dimensions, while harmony exhibits small negative weights (-2.8%, -1.8%, -1.9%), suggesting that harmonic similarity may actually decrease perceived overall similarity in cross-cultural contexts. This counterintuitive finding may reflect the Western bias inherent in our harmonic feature extraction, where Western chord templates and key estimation algorithms may not adequately capture the harmonic relationships that humans perceive in non-Western musical traditions.

Timbre contributes positively but modestly to human similarity perception (3.8%, 3.5%, 6.9%), with the highest weight for recommendation-level similarity, suggesting that timbral characteristics may play a larger role in personal preference judgments than in overall musical or cultural assessments.

Foundation Model Patterns

Foundation models demonstrate markedly different feature weighting patterns compared to human perception, revealing how learned representations prioritize different musical aspects when operating in a multicultural similarity space. Most notably, timbre emerges as the dominant factor for several foundation models, with particularly high weights for CultureMERT-TA (62.5%), CultureMERT (58.7%), and MERT-330 (48.5%). This emphasis on timbral features reveals an important insight: when computational models are required to establish similarity relationships across diverse musical cultures, they gravitate toward universal acoustic characteristics that remain

Method	Triplet Agr. ↑ (%)			NDCG ↑ (%)			Spearman ρ ↑ (-1, 1)			Kendall τ ↑ (-1, 1)			MAE ↓ (%)		
Similarity type	**Overall**	**Cultural**	**Recomm.**	**Overall**	**Cultural**	**Recomm.**	**Overall**	**Cultural**	**Recomm.**	**Overall**	**Cultural**	**Recomm.**	**Overall**	**Cultural**	**Recomm.**
Linear Regression	67.0	**66.7**	**65.1**	**92.5**	**91.4**	**90.9**	**0.19**	**0.15**	**0.18**	0.18	**0.14**	**0.17**	**19.7**	**22.2**	**23.0**
LightGBM	**67.2**	63.8	64.4	92.2	90.6	90.1	**0.19**	0.12	0.13	**0.19**	0.12	0.13	19.8	**22.2**	23.2

Table 6.4. Ensemble Regression Results Combining Signal Processing Features and Foundation Models. Performance evaluation of ensemble methods for predicting human similarity judgments. Values are shown as percentages (%) for Triplet Agreement, NDCG, and MAE, and as correlation values for Spearman and Kendall metrics. Arrows indicate whether higher (↑) or lower (↓) values represent better performance, with best performance within each similarity dimension and metric shown in bold.

consistent across traditions, rather than culture-specific patterns like the melodic ones, that may vary significantly between musical systems.

MERT-95 shows a more balanced approach, equally weighting melody (42.5%) and timbre (42.5%) contributions. CLAP models exhibit distinct behavior, with CLAP-Music&Speech showing strong melody emphasis (40.9%) similar to human patterns, while CLAP-Music displays more modest feature contributions overall. Qwen2-Audio presents an extreme pattern with a small positive weight for melody (13.9%), a negative timbre weight (-6.4%), and near-zero rhythm and harmony weights. This pattern may indicate that Qwen2-Audio's representations capture musical relationships through dimensions not adequately represented by our four-feature decomposition, reflecting the model's multimodal architecture and diverse training objectives.

6.6.5 Ensemble Methods for Human Similarity Prediction

To leverage the complementary strengths of signal processing features and foundation model representations, we developed ensemble regression methods that combine all computational approaches to predict human similarity judgments. This analysis explores whether integrating interpretable musical features with learned representations can achieve superior alignment with human cross-cultural music perception.

Ensemble Performance Results

Table 6.4 presents the performance of both linear regression and LightGBM ensemble methods across all evaluation metrics and similarity dimensions. The ensemble methods achieve remarkable improvements compared to individual approaches shown in Table 6.2. Linear regression ensemble achieves triplet agreement scores of 65.1-67.0% compared to the best individual method (CLAP-Music&Speech) at 62.6-64.9%. Similarly, NDCG scores reach 90.9-92.5% versus the previous best of 88.0-89.8%. The correlation metrics also show consistent improvements, with Spearman correlations reaching 0.15-0.19 compared to individual method maximums of 0.11-0.16.

Most significantly, the ensemble methods achieve substantial reductions in prediction error, with MAE values of 19.7-23.2% representing improvements of approximately 6-7 percentage points over the best individual methods (melody features at 29.5-30.9% MAE). This represents a relative error reduction of roughly 25-30%, demonstrating that combining multiple computational approaches provides complementary information for predicting human similarity judgments.

Linear regression demonstrates superior performance across most metrics, particularly excelling in ranking-based measures (NDCG, correlation metrics) and achieving the best MAE scores. LightGBM shows marginally better performance only in overall music similarity triplet agreement, with differences being minimal across other metrics. This demonstrates that the different computational

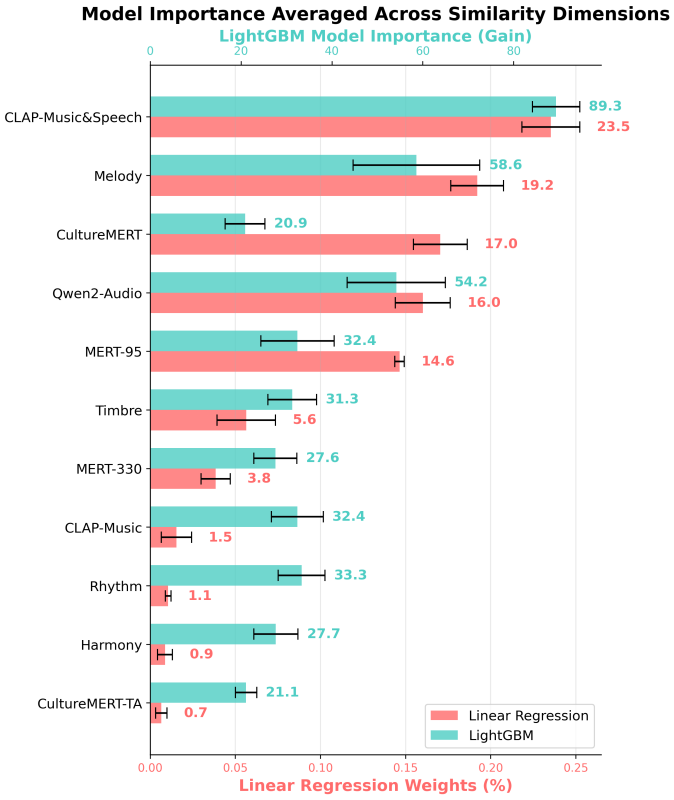

Figure 6.10. Computational Model Importance in Ensemble Methods. Feature importance visualization averaged across similarity dimensions and sorted by linear regression coefficients. The dual x-axes accommodate the different scales of coefficient values (linear regression) and gain scores (LightGBM), showing the relative contribution of each computational method to ensemble performance.

approaches provide complementary information that linear combination can effectively leverage to predict human cross-cultural music similarity judgments.

Computational Model Contribution Analysis

Figure 6.10 reveals the relative importance of different computational methods within the ensemble approaches, averaged across the three similarity dimensions and sorted by linear regression importance. Linear regression coefficients represent the direct contribution of each feature to the prediction, while LightGBM gain scores measure the total improvement in splitting criterion achieved by each feature across all decision tree splits, providing a measure of feature utility in the gradient boosting framework.

CLAP-Music&Speech emerges as the dominant contributor in both ensemble methods, with the highest linear regression coefficient (23.5%) and highest LightGBM gain score (89.3). This finding validates our earlier observation that this model achieves the best alignment with human cross-cultural music perception. Among signal processing features, melody maintains its position as the most important traditional feature (linear coefficient: 19.2%, LightGBM gain: 58.6), confirming its fundamental role in human music similarity perception across cultures.

Foundation models show varied contributions, with CultureMERT (17.0%, 20.9) and Qwen2-Audio (16.0%, 54.2) providing meaningful but secondary contributions. Interestingly, while MERT-95 achieves reasonable individual performance, its ensemble importance is more modest (14.6%,

32.4), suggesting its representations overlap considerably with other foundation models. The signal processing features beyond melody show limited but non-zero contributions, indicating they provide some unique information not captured by the foundation models.

These ensemble results demonstrate that combining diverse computational approaches yields substantial improvements in predicting human cross-cultural music similarity judgments, achieving performance levels that demonstrate the potential for practical culturally aware music AI applications.

6.7 Conclusions

This chapter has presented the first comprehensive evaluation of computational music similarity methods against human cross-cultural music perception, spanning nine diverse musical traditions and encompassing both interpretable signal processing features and state-of-the-art foundation models. Our evaluation reveals that foundation models outperform traditional signal processing features, with CLAP-Music&Speech achieving the highest individual performance (62.6-64.9% triplet agreement, 88.0-89.8% NDCG). Among signal processing features, melody consistently emerges as the dominant factor in predicting human similarity judgments, confirming its universal importance across cultures and providing quantitative validation of musicological understanding.

However, our findings reveal a fundamental trade-off between universal musical understanding and cultural discriminability in computational models. While CLAP-Music&Speech excels at aligning with human similarity perception, Qwen2-Audio demonstrates superior cultural boundary detection (separation ratios up to 1.60 vs. 1.37), suggesting that models optimized for universal cross-cultural understanding may sacrifice discriminative power between cultural traditions. This highlights the challenge of developing systems that can both capture cross-cultural musical relationships and maintain cultural distinctiveness.

The cultural discrimination analysis underlines significant gaps between human and computational approaches. While humans demonstrate strong cultural awareness with separation ratios of 1.80-2.36, computational methods achieve more modest discrimination (1.03-1.60). Importantly, the apparent underperformance of culturally adapted models like CultureMERT reflects the influence of listener cultural background on evaluation results, as these models were adapted toward non-Western cultures while most participants reported Western musical backgrounds.

Feature contribution analysis uncovers fundamental differences between human and computational processing strategies: humans consistently prioritize melodic content across all similarity dimensions, while foundation models tend to emphasize timbral characteristics that remain consistent across traditions. Most encouragingly, ensemble methods combining signal processing features with foundation model representations achieve substantial improvements, reaching 65.1-67.0% triplet agreement and reducing prediction errors by 25-30% compared to individual methods.

These findings establish human perceptual validation as an essential component of multicultural music representation learning and provide actionable insights for developing more culturally aware music AI systems. The comprehensive evaluation framework and ensemble approaches developed in this study offer templates for future research advancing the alignment between computational music analysis and human cross-cultural music perception.

Chapter 7

Conclusions

This dissertation has explored representation learning across diverse musical traditions for music signal analysis through a series of interconnected studies. Beginning with the development of the Lyra dataset for Greek traditional and folk music, progressing through investigations of transfer learning and few-shot learning approaches, evaluating and adapting foundation models for diverse musical traditions, and culminating in a comprehensive evaluation of computational music similarity methods against human cross-cultural music perception, this research has contributed to advancing representation learning for varied musical systems. This concluding chapter synthesizes the key contributions and findings, reflects critically on the research journey, acknowledges limitations, and discusses future directions.

7.1 Summary of Contributions

The primary contributions of this dissertation span dataset development, methodological innovations, model evaluation, adaptation strategies, and human-computational alignment studies, collectively addressing the research questions established in the introduction.

7.1.1 Addressing Data Availability for Diverse Musical Traditions (RQ1)

The Lyra dataset represents our comprehensive response to the fundamental challenge of data scarcity in computational analysis of traditional music. This collection of Greek traditional and folk music, comprising 1,570 pieces with approximately 80 hours of high-quality recordings, demonstrates a methodology for creating culturally-grounded datasets that can support sophisticated computational analysis while respecting musicological integrity. The dataset's development process emphasized consistent recording quality through systematic collection from academic documentary sources, ensuring that technical variations would not confound the analysis of musical characteristics.

The rich metadata annotations covering instrumentation, geographic origin, genre, and subgenre provide a level of detail rarely available in world music datasets, enabling fine-grained computational analysis of traditional music practices. The academic rigor embedded in the annotations, derived from expert presentations and documentaries, ensures that the computational analysis is grounded in authentic musicological knowledge rather than external categorizations. The multimodal access through timestamped YouTube links enables researchers to examine both audio and video content, supporting analysis of performance practices and cultural contexts that are often absent from purely audio-based datasets.

Beyond its immediate utility for Greek music research, the Lyra dataset establishes a replicable methodology for dataset development in computational ethnomusicology, demonstrating how documentary sources can be systematically leveraged to create high-quality resources for underrepresented musical traditions.

7.1.2 Understanding Cross-Cultural Knowledge Transfer (RQ2)

Our systematic investigation of knowledge transfer between musical traditions provides the first comprehensive analysis of how computational knowledge moves across diverse musical systems. Through evaluation of multiple deep audio embedding models across musical corpora spanning Western, Mediterranean, and Indian traditions, this research reveals that cross-cultural knowledge transfer in music exhibits complex patterns that reflect both shared musical elements and culture-specific characteristics.

The bidirectional nature of effective knowledge transfer challenges assumptions about the primacy of Western-trained models for world music analysis. Models trained on traditional music systems can provide valuable initializations for Western music tasks, demonstrating that diverse musical traditions contain computational knowledge that benefits broader musical understanding. The asymmetric patterns of transfer effectiveness reveal that geographic and historical proximity often correlates with successful knowledge transfer, with Indian traditions showing particularly strong bidirectional transferability.

These findings establish that computational models can reveal meaningful relationships between musical systems that complement traditional musicological comparative studies with quantitative, data-driven insights. The patterns of transfer effectiveness provide a new lens for understanding musical relationships across cultures, suggesting computational approaches to comparative musicology that could inform both technical and theoretical understanding of musical systems.

7.1.3 Learning from Limited Examples in Musical Contexts (RQ3)

The development of LC-Protonets addresses the pervasive challenge of data scarcity in world music research through a novel approach to multi-label few-shot learning. By creating prototypes for label combinations rather than individual labels, this methodology enables computational models to learn from the power set of available annotations, significantly expanding the effective training signal from limited examples.

The integration with pre-trained embedding spaces demonstrates how few-shot learning can leverage the representational power of models trained on larger datasets while adapting to the specific characteristics of traditional music contexts. The two-step learning framework shows particular promise for expanding the coverage of computational models to include rare but culturally significant musical attributes that would otherwise be excluded due to data limitations.

The computational optimization that addresses scalability concerns makes the approach practical for real-world deployment, while the consistent performance improvements across diverse music datasets demonstrate the generalizability of the approach beyond the specific contexts in which it was developed. This contribution provides a pathway for including underrepresented musical elements in computational models, supporting more comprehensive and inclusive approaches to music analysis.

7.1.4 Evaluating Foundation Models Across Musical Traditions (RQ4)

Our comprehensive evaluation of state-of-the-art music foundation models across diverse musical traditions provides crucial insights into both the potential and limitations of current approaches to universal music representation. The multi-faceted evaluation framework, employing probing, supervised fine-tuning, and few-shot learning methodologies, reveals that foundation models demonstrate impressive cross-cultural capabilities compared to previous approaches while simultaneously exhibiting clear Western-centric biases.

The systematic comparison across six musical corpora representing different traditions shows that larger models typically demonstrate better generalization capabilities, but performance consistently declines for culturally distant traditions. The few-shot learning evaluation proves particularly revealing, showing that foundation models struggle with the kind of low-resource scenarios common in world music research, often performing no better than much smaller, specialized models.

These findings establish important benchmarks for the field while highlighting the gap between the universality claims of foundation models and their actual performance across diverse musical contexts. The evaluation framework itself represents a methodological contribution, providing templates for future assessment of cross-cultural music representation capabilities.

7.1.5 Adapting Foundation Models to Become Multicultural (RQ5)

CultureMERT represents our systematic approach to enhancing the cultural awareness of foundation models through continual pre-training on diverse musical traditions. The two-stage adaptation strategy addresses the fundamental challenge of catastrophic forgetting while enabling stable acquisition of new cultural knowledge. Training on a carefully curated 650-hour dataset comprising Greek, Turkish, and Indian music demonstrates that foundation models can be effectively enhanced to better represent non-Western traditions while preserving their general capabilities.

The exploration of task arithmetic as an alternative adaptation approach provides a resource-efficient method for combining cultural adaptations without requiring simultaneous access to all datasets. This modular approach to cultural adaptation offers practical advantages for scenarios where data sharing or computational resources are constrained, while achieving comparable performance to continual pre-training approaches.

The consistent improvements across diverse music tagging tasks, with an average ROC-AUC improvement of 4.43% compared to the original foundation model, demonstrate the practical value of cultural adaptation while establishing that such improvements can be achieved without sacrificing performance on Western benchmarks.

7.1.6 Bridging Human Perception and Computational Music Similarity (RQ6)

The human perception study represents a crucial validation step for the computational advances developed throughout this dissertation. By collecting similarity judgments from 125 participants across diverse backgrounds on 1,130 audio pairs spanning nine musical traditions, this investigation provides the first empirical assessment of how well our multicultural representation learning approaches align with human cross-cultural music perception.

The results validate several key hypotheses while revealing important gaps. The superior performance of foundation models over signal processing features (with CLAP-Music&Speech achieving 62.6-64.9% triplet agreement) confirms the value of the approaches evaluated in Chapter 5. The

emergence of melody as the most predictive traditional feature validates musicological understanding of its central role in music.

However, the study reveals a critical tension in multicultural music representation: the trade-off between universal musical understanding and cultural discriminability. This finding has profound implications for how we conceptualize "universal" music representations, suggesting that truly effective cross-cultural systems may need to balance these competing objectives rather than optimize for one.

The discovery that humans prioritize melodic content while foundation models emphasize timbral characteristics represents a significant finding for the field. This misalignment suggests that current pre-training objectives and evaluation metrics may not capture the perceptual dimensions most salient to human listeners across cultures. This insight directly informs the design of future foundation models and evaluation frameworks developed in this dissertation.

Most encouragingly, the success of ensemble methods (achieving 25-30% error reduction) demonstrates that the diverse computational approaches developed throughout this dissertation can provide complementary information. The finding that interpretable signal processing features remain crucial when combined with sophisticated learned representations points towards the usage of hybrid approaches for music analysis.

This human-centered validation establishes that the technical advances explored in previous chapters have genuine perceptual relevance while revealing the fundamental challenge of aligning computational optimization with human cross-cultural music understanding, a challenge that defines the future research agenda for culturally aware music AI systems.

7.2 Synthesis of Findings

Examining the research contributions holistically reveals several cross-cutting insights about the nature of multicultural music representation learning and the challenges inherent in developing computational approaches that can effectively bridge cultural boundaries in music analysis.

7.2.1 The Challenge of Musical Knowledge Transfer

The investigation of cross-cultural transfer reveals that musical knowledge transfer operates according to complex patterns that resist simple explanations based solely on geographic proximity or historical connections. While Indian traditions do show strong bidirectional transferability, the patterns of transfer effectiveness suggest that computational similarity between musical traditions may capture aspects of musical relationships that are not immediately apparent through traditional musicological analysis.

The asymmetric nature of many transfer relationships indicates that musical traditions may share certain computational features while differing in others, creating selective patterns of knowledge transfer that could inform our understanding of both musical universals and cultural specificity. The finding that models trained on multiple traditional musical systems show enhanced generalization capabilities suggests that diversity in training data may be more important than scale for developing broadly applicable music representations.

These patterns of computational knowledge transfer find validation in the human perception study, where participants demonstrate sophisticated understanding of cross-cultural musical relationships that may transcend simple geographic boundaries. The fact that ensemble methods combining different computational approaches achieve superior alignment with human perception

suggests that humans may integrate multiple types of musical information when making similarity judgments across cultures.

7.2.2 Resource Constraints and Methodological Innovation

The consistent theme of resource constraints across different aspects of multicultural music research has driven methodological innovations that extend beyond their immediate applications. Few-shot learning approaches like LC-Protonets demonstrate that meaningful progress can be achieved even with severely limited annotated data, while transfer learning shows that knowledge from resource-rich domains can effectively bootstrap learning in low-resource contexts.

The exploration of continual pre-training and task arithmetic for foundation model adaptation reveals different strategies for balancing computational efficiency with adaptation effectiveness. Continual pre-training offers superior performance when computational resources are available, while task arithmetic provides comparable results with significantly reduced resource requirements.

This methodological diversity suggests that successful multicultural music representation learning requires a toolkit of complementary approaches rather than a single universal solution. This finding is reinforced by the human perception study, where ensemble models combining several computational methods achieve superior alignment with human judgment compared to any individual method.

7.2.3 The Question of Musical Universality

The research provides nuanced evidence regarding the possibility of universal music representations that can effectively capture diverse musical traditions. Current foundation models demonstrate impressive cross-cultural capabilities that suggest some shared representational structures across musical traditions, drawing from existing statistical universals in music. However, the consistent performance degradation for culturally distant traditions, particularly in challenging few-shot learning scenarios, indicates substantial culture-specific elements that resist universal representation.

The human perception study adds crucial perspective to this question by revealing that while humans can make meaningful similarity judgments across diverse musical traditions, they also demonstrate strong cultural awareness with clear discrimination between different musical systems. The finding that melody emerges as a universal predictor of human similarity judgments across cultures supports the existence of some musical universals, while the cultural discrimination patterns confirm the importance of culture-specific musical knowledge.

The success of multicultural training approaches in enhancing cross-cultural generalization suggests that universality in music representation may be achievable, but only through deliberate inclusion of diverse musical traditions in the training process rather than through post-hoc adaptation of Western-centric models. This finding has important implications for the development of future music foundation models and suggests that achieving universal music representation requires fundamental changes in how these models are conceived and trained.

7.2.4 Human-Computational Alignment in Cross-Cultural Contexts

The systematic comparison of computational approaches against human cross-cultural music perception reveals both encouraging alignments and significant gaps that have implications across all aspects of this research. The validation that foundation models outperform signal processing

features confirms the value of learned representations explored in earlier chapters, while the discovery of fundamental processing differences, i.e., humans prioritizing melodic content versus models emphasizing timbral characteristics, reveals a critical misalignment that arises questions about the optimization goals of music AI systems.

The tension between universal musical understanding and cultural discriminability emerges as a central challenge for multicultural music representation. This trade-off manifests differently across models: some excel at cross-cultural similarity alignment while others better maintain cultural boundaries. This finding suggests that the pursuit of universal music representations may inherently conflict with preserving cultural distinctiveness, requiring careful consideration of which objective to prioritize in different application contexts.

The influence of evaluation context, particularly listener cultural background, adds complexity to assessing multicultural music systems. The performance of culturally adapted models varies significantly depending on the cultural context of evaluation, highlighting that effectiveness cannot be measured independently of who the listener is. This finding has profound implications for how we design evaluation studies and interpret results in cross-cultural music research.

These alignment challenges have immediate practical implications for music technology deployment. Current systems relying primarily on foundation model representations may miss perceptual dimensions most salient to human listeners across cultures. However, the success of ensemble approaches in bridging this gap demonstrates that combining interpretable features with learned representations offers a pathway toward more perceptually grounded systems.

The broader pattern that emerges is one of incomplete but improvable alignment: while computational methods lag behind human cultural discrimination capabilities, the complementary strengths of different approaches suggest that careful combination strategies can significantly advance human-computational alignment in cross-cultural music understanding.

7.3 Critical Reflection

This research journey has evolved from addressing practical challenges in computational analysis of traditional music to engaging with fundamental questions about the nature of musical representation across cultural boundaries and its relationship to human perception. The progression from dataset development through methodological innovation to foundation model adaptation and human-computational alignment studies reflects both the natural evolution of the research questions and a deepening understanding of the complexity inherent in multicultural music representation learning.

The methodological choices made throughout this research involved significant trade-offs that shaped both the scope and the conclusions of the work. The decision to focus primarily on audio-based representations, while enabling direct comparison across musical traditions, necessarily limited the depth of cultural understanding that could be incorporated into the computational models. Similarly, the emphasis on classification tasks provided clear evaluation metrics but may have underexplored other aspects of musical understanding that are equally important for cross-cultural analysis.

The inclusion of human perception studies in the final phase of this research proved transformative for understanding the effectiveness and limitations of computational approaches. The discovery that current computational methods emphasize different musical dimensions than humans suggests that future research in multicultural music representation learning must consider human perceptual validation as a central rather than peripheral concern.

The evolution of my understanding throughout this research process reflects the inherent complexity of bridging computational and cultural approaches to music analysis. Early assumptions about the transferability of techniques from natural language processing proved insufficient for addressing the unique challenges of musical representation across cultures. The recognition that musical semantic spaces resist the kind of alignment successful in cross-lingual NLP led to the development of more nuanced approaches that respect both shared and distinctive aspects of musical traditions.

Perhaps most significantly, the research has highlighted the importance of approaching multicultural music representation learning as both a technical and cultural challenge that must ultimately be validated against human perceptual understanding. While computational advances can provide powerful tools for music analysis, their effectiveness ultimately depends on their ability to capture and respect the cultural knowledge embedded in different musical traditions while aligning with how humans actually perceive and categorize musical relationships across cultures.

The human perception study revealed that achieving this alignment requires not just technical sophistication but also careful consideration of which musical dimensions computational systems prioritize and how these align with human perceptual strategies. The promising results of the ensemble methods suggest that future advances may come from hybrid approaches that leverage both the interpretability of traditional signal processing and the pattern recognition capabilities of modern deep learning systems.

7.4 Limitations and Challenges

Despite the contributions outlined above, this dissertation faces several important limitations that constrain the generalizability of its findings and highlight areas requiring further research.

7.4.1 Dataset and Cultural Representation Limitations

The Lyra dataset, while providing a valuable resource for Greek traditional music research, reflects several constraints that limit its broader applicability. The reliance on documentary source material, though ensuring musicological soundness, introduces potential selection biases toward performances deemed worthy of documentation by the creators of the source material. This selection process may inadvertently emphasize certain aspects of Greek traditional music while underrepresenting others, such as informal or ritual contexts where much traditional music naturally occurs. The geographic and genre categorizations employed in the dataset, while detailed and musicologically informed, necessarily simplify the complex reality of regional variations and fusion practices that characterize living musical traditions. Furthermore, the use of YouTube links for audio access, while providing multimodal capabilities, creates potential sustainability issues if the source videos become unavailable, highlighting the broader challenge of creating stable, long-term resources for computational ethnomusicology.

Beyond the Lyra dataset specifically, this research faces broader limitations in cultural representation and scope. While the cross-cultural similarity study expands coverage to nine musical traditions, the investigation still focuses primarily on a subset of global musical diversity, leaving vast areas unexplored. Major musical traditions from sub-Saharan Africa, indigenous Americas, Southeast Asia, and other regions remain underrepresented, limiting the generalizability of findings about cross-cultural transfer patterns and adaptation strategies. This limited scope reflects broader challenges in computational ethnomusicology, where resource constraints and data availability often determine which traditions can be included in comparative studies.

The framing of musical traditions within geographical and cultural categories, while necessary for systematic investigation, risks oversimplifying the complex reality of musical cultures. The use of terms like "Western" and "non-Western" music, though common in computational research, creates artificial dichotomies that may not reflect the fluid, interconnected nature of musical traditions. Musical cultures exist on a continuum shaped by historical exchanges, migration patterns, and cultural adaptation rather than strict geographical divisions. The computational approach employed in this research, while providing valuable quantitative insights, operates within a data-driven framework that may not capture the historical, theoretical, and cultural knowledge embedded within specific musical traditions.

The human perception study, while including participants from 21 countries with diverse musical backgrounds, exhibits a notable imbalance with the majority (62.4%) from Greece and other European countries, with relatively few participants from the cultures represented in the musical datasets. This participant distribution limits the ability to measure computational models' cultural bias by evaluating their alignment with listeners from different musical backgrounds, an analysis that a more culturally diverse participant pool would enable.

7.4.2 Methodological and Technical Constraints

The methodological approaches employed throughout this research reflect broader limitations in current computational approaches to multicultural music analysis. The utilization of supervised learning, particularly classification tasks, while providing clear evaluation metrics, may not capture the more nuanced aspects of musical understanding that are central to cross-cultural music analysis. Many aspects of musical meaning, including improvisation, ornamentation, and contextual interpretation, resist categorical classification and may require alternative computational frameworks that are not fully explored in this research. Furthermore, the traditional evaluation metrics, designed primarily for commercial music applications, may inadequately reflect the cultural significance of correctly identifying rare but important musical attributes in traditional music contexts.

The cross-cultural similarity study, while providing valuable insights into human-computational alignment, is constrained to three similarity dimensions (overall musical, cultural, and recommendation-level) and may not capture other important aspects of cross-cultural music perception. The use of pairwise similarity judgments, while enabling systematic comparison, may not fully capture the complex, context-dependent nature of musical similarity perception that varies based on listening purpose, cultural background, and individual experience.

The temporal constraints of the research necessitated certain simplifications that may have limited the depth of investigation. Computational constraints required limiting the scale of fine-tuning experiments and the exploration of alternative adaptation strategies. The context length limitations during model training and evaluation, typically constrained to short audio segments of 5-30 seconds, may prevent the capture of longer-term musical structures that are crucial for understanding many traditional music forms. This limitation is particularly significant for musical traditions that employ extended improvisational sections or complex structural organizations that unfold over longer time periods.

The foundation models evaluated were primarily trained on commercial music datasets, potentially limiting their understanding of traditional world music characteristics. Some signal processing features exhibit tendencies toward Western musical concepts that may not adequately capture relationships important in non-Western traditions. These limitations affect both individual model performance and the ensemble methods that combine different computational approaches.

7.4.3 Foundation Model Architecture Limitations

The foundation models evaluated and adapted in this research carry inherent limitations that reflect broader challenges in developing universal music representations. The audio tokenizers used in these models, such as EnCodec [167], are trained predominantly on commercial music and may introduce systematic biases in the initial representation of diverse musical characteristics. This limitation affects the representational granularity available for non-Western traditions, as the tokenizer may not adequately capture microtonal inflections, complex rhythmic patterns, or distinctive timbral characteristics of traditional instruments.

The models operate primarily at the audio signal level without incorporating the broader contextual knowledge, including performance practices, cultural meanings, and historical contexts, that shapes musical understanding within specific traditions. This limitation reflects a fundamental challenge in current approaches to music AI, where models excel at pattern recognition in acoustic signals but struggle to incorporate the cultural and contextual knowledge that is essential for meaningful musical understanding across traditions.

The finding that foundation models tend to emphasize timbral characteristics while humans prioritize melodic content reveals a fundamental architectural limitation that affects human-computational alignment. Current pre-training objectives and architectures may be inherently biased toward learning spectral patterns that are more easily captured through self-supervised learning objectives, potentially missing the melodic relationships that are most salient to human perception across cultures.

The adaptation strategies explored in this research, while effective within their scope, may be insufficient for fully addressing the representational challenges posed by diverse musical traditions. The two-stage continual pre-training approach, while computationally efficient and practically necessary under resource constraints, may not be optimal for larger computational budgets or different model architectures. The exploration of adaptation strategies was limited to specific architectural approaches and may not generalize to other foundation model designs or scaling regimes.

7.4.4 Evaluation and Validation Challenges

The evaluation frameworks employed in this research face fundamental challenges that limit the confidence with which conclusions can be drawn about cross-cultural music representation learning. The lack of standardized benchmarks for many musical traditions complicates comparative evaluation and may lead to conclusions that are not robust across different evaluation contexts. The datasets used for evaluation, while representing diverse traditions, vary significantly in size, annotation quality, audio quality, and cultural coverage, making direct comparisons across traditions potentially misleading.

The human perception study, while providing crucial insights into human-computational alignment, faces limitations in participant recruitment, cultural representation, and evaluation scope. The cross-cultural similarity task, while systematic and replicable, represents only some aspects of musical understanding and may not capture other dimensions of cross-cultural music perception that may be equally important for developing culturally aware music AI systems.

The influence of cultural background on evaluation outcomes represents a crucial finding that affects the interpretation of all cross-cultural music research. The discovery that culturally adapted models may appear to underperform when evaluated by participants from different cultural backgrounds highlights the importance of considering participant demographics in cross-cultural eval-

uation studies. This finding suggests that the effectiveness of computational approaches to multicultural music representation cannot be assessed independently of the cultural context of the evaluation, adding another layer of complexity to developing universal music AI systems.

The risk of data leakage, even under seemingly valid experimental protocols, represents a significant concern for the validity of cross-cultural transfer learning results. Subtle overlaps in musical artists, recording conditions, instrumentation, or cultural contexts between training and evaluation domains may introduce spurious correlations that inflate performance estimates. The standard practice of using dataset-provided train/test splits, while following established protocols, may not adequately control for these forms of contamination in cross-cultural scenarios where the boundaries between musical traditions are often fluid and overlapping.

The representativeness of the datasets used for different musical traditions introduces additional validation concerns. The utililzed datasets, while carefully curated and culturally informed, may not fully capture the diversity within each tradition, potentially missing crucial aspects such as regional variations, contemporary adaptations, or specific performance contexts. This limitation affects the generalizability of findings and may lead to conclusions about musical traditions that are based on incomplete representations of their full diversity.

7.4.5 Theoretical and Interpretive Limitations

This research operates within a computational framework that, while providing valuable quantitative insights, has inherent limitations in addressing the theoretical and interpretive aspects of cross-cultural music analysis. The data-driven approach, while reducing certain forms of analytical bias, may miss important aspects of musical understanding that require cultural knowledge and theoretical frameworks specific to individual traditions. The computational patterns of similarity and transfer identified in this research, while meaningful from a technical perspective, may not align with musicological understanding of relationships between musical traditions.

The emphasis on automatic tagging tasks and similarity assessment, while providing clear evaluation metrics, represents only a subset of musical understanding and may not capture other dimensions of musical cognition and cultural meaning that are equally important for cross-cultural music analysis. The research does not adequately address questions of musical aesthetics, spiritual or ritual significance, or the social functions of music within different cultural contexts, all of which are crucial for comprehensive understanding of musical traditions.

The temporal scope of this research, conducted over a specific period with particular datasets and computational tools, limits the stability and generalizability of conclusions. The rapid evolution of foundation models and computational approaches means that specific technical findings may become obsolete, while the broader insights about cross-cultural representation learning and human-computational alignment may have more lasting value. This limitation highlights the need for continued research that can adapt to evolving computational capabilities while maintaining focus on the fundamental challenges of multicultural music representation and its alignment with human perception.

7.5 Future Directions

Building upon the contributions and addressing the limitations identified in this research, several promising directions emerge for advancing multicultural music representation learning and improving human-computational alignment in cross-cultural music understanding.

The expansion of dataset coverage represents a fundamental priority for enabling world music representation learning. Recent initiatives in computational ethnomusicology, such as the development of large-scale traditional music corpora [33] and community-driven annotation projects [209], provide models for creating more comprehensive datasets that include underrepresented musical traditions from Africa, East Asia, and the Americas. Future dataset development should emphasize multimodal collections that integrate audio, video, lyrics, cultural context, and performance practice information, following emerging trends in multimodal machine learning [210]. The development of annotation schemes that capture tradition-specific musical attributes and concepts, moving beyond conventional commercial music categories, requires close collaboration between computational researchers and cultural practitioners to ensure authenticity and relevance.

The need for more culturally diverse participant pools in human perception studies represents a critical direction for future research. With larger and more representative participant populations from different musical traditions, researchers could measure computational models' cultural bias by evaluating their alignment with listeners from diverse musical backgrounds. This would enable investigation of whether computational models exhibit systematic biases toward certain cultural contexts and how these biases could be mitigated through improved training or adaptation strategies.

Methodological innovations building upon the approaches developed in this dissertation offer several promising directions. The finding that humans prioritize melodic content while foundation models emphasize timbral characteristics suggests the need for pre-training objectives that better capture melodic relationships across cultures. Self-supervised learning approaches specifically designed for music understanding across different traditions could potentially reduce reliance on annotated data while capturing the distinctive characteristics of diverse musical systems [211–213]. Recent advances in adaptive architectures, including mixture-of-experts models and dynamic neural networks [214, 215], suggest possibilities for developing models that can dynamically adapt to different musical traditions while maintaining computational efficiency.

The success of ensemble methods in achieving superior human-computational alignment suggests that future research should explore more sophisticated approaches to combining different computational methods. Advanced ensemble architectures that can learn optimal combinations of interpretable features and learned representations, rather than relying on simple linear combinations, may achieve even better alignment with human perception while maintaining interpretability. Multi-task learning approaches that simultaneously optimize for human similarity prediction and traditional music analysis tasks could produce models that better balance multiple aspects of musical understanding.

Cross-modal learning approaches that integrate audio analysis with contextual knowledge, visual information, and textual descriptions align with recent trends in multimodal foundation models [216, 217] and could provide more comprehensive approaches to cultural music understanding. The integration of cultural metadata, performance practice information, and contextual knowledge into computational models represents a promising direction for developing more culturally aware music AI systems that go beyond purely acoustic analysis.

The development of foundation models specifically designed for multicultural music representation requires fundamental changes in current approaches to model development. Rather than adapting Western-centric models post-hoc, future work should focus on developing foundation models pre-trained from the outset on diverse musical traditions, following recent trends in multilingual language models [29]. The creation of audio tokenizers specifically designed for diverse musical traditions could address the limitations of current systems [167, 218] in representing micro-

tonal inflections, complex rhythmic patterns, and distinctive timbral characteristics of traditional instruments.

Advanced mechanisms that allow models to adapt to cultural context represent an emerging area of research in culturally-aware AI systems [219]. Future foundation models could incorporate explicit cultural conditioning mechanisms that enable them to adjust their processing based on the cultural context of the input music, potentially improving both technical performance and cultural appropriateness.

The development of evaluation frameworks that better capture the multi-dimensional nature of cross-cultural music perception represents another important direction. Current evaluation approaches, while systematic, may not adequately assess the complex ways in which humans perceive musical relationships across cultures. Future frameworks should consider incorporating cultural proximity into discrimination metrics, recognizing that some cultural boundaries are more permeable than others, and developing evaluation approaches that can assess model performance across different aspects of musical understanding beyond similarity assessment.

The translation of technical advances into practical applications with cultural and societal impact represents a crucial direction for future work. Cultural heritage preservation applications could leverage the computational tools developed in this research to support documentation and analysis of endangered musical traditions, building upon recent initiatives in digital cultural heritage [220]. Educational applications that help students understand diverse musical traditions and their relationships could democratize access to cross-cultural musical knowledge while respecting cultural specificity.

Cross-cultural recommendation systems that facilitate meaningful discovery across musical traditions while respecting their distinctive characteristics represent an emerging area of research in culturally aware music information retrieval [221, 222]. The insights from the human perception study regarding the importance of melodic content and the effectiveness of ensemble methods could inform the development of recommendation systems that better align with how humans actually perceive musical relationships across cultures.

Creative tools that support cross-cultural music creation and collaboration, while preserving distinctive cultural characteristics, represent an emerging area of research in AI-assisted creativity [14]. Such tools could leverage the computational understanding of cross-cultural musical relationships developed in this research while ensuring that they enhance cultural diversity in music creation.

Finally, the establishment of ongoing evaluation frameworks that can continuously assess the alignment between computational approaches and human cross-cultural music perception represents an important infrastructure need for the field. Regular evaluation campaigns, similar to those established in other areas of AI research, could track progress in developing more culturally aware and perceptually aligned music AI systems while providing standardized benchmarks for comparing different approaches.

7.6 Closing Thoughts

This dissertation has addressed fundamental challenges in developing computational representations that can effectively capture the rich diversity of musical traditions worldwide while aligning with human cross-cultural music perception. Through systematic investigation spanning dataset development, methodological innovation, comprehensive evaluation, adaptive enhancement, and human perceptual validation, this research has contributed to advancing the field of multicul-

tural music representation learning while revealing both the potential and limitations of current computational approaches.

The journey from the Lyra dataset through cross-cultural transfer learning, few-shot learning methodologies, foundation model adaptation, and human-computational alignment studies reflects an evolving understanding of how computational models can bridge cultural boundaries in music representation. The research demonstrates that while truly universal music representations that perfectly align with human perception remain an aspirational goal, significant progress can be achieved through approaches that respect both cross-cultural commonalities and culture-specific characteristics.

The findings of this research have implications beyond the technical domain of music information retrieval. As music technologies increasingly mediate how we discover, create, and share music globally, the development of more culturally aware computational approaches that align with human perceptual understanding becomes essential for preserving the rich diversity of human musical expression. The computational tools, methodological frameworks, and evaluation approaches developed in this dissertation provide pathways for ensuring that technological advances in music AI celebrate rather than homogenize the world's musical heritage while serving users in ways that respect their perceptual understanding of musical relationships.

Looking toward the future, the research presented here represents foundational work in an emerging field that sits at the intersection of computational intelligence, cultural understanding, and human perception. The methodologies, insights, and open-source resources contributed by this dissertation provide building blocks for future research that can further advance our ability to computationally represent and analyze the full spectrum of human musical expression while ensuring that these representations align with how humans actually perceive and understand music across cultures.

It is my hope that this work will inspire continued exploration of multicultural music representation learning with explicit attention to human-computational alignment, advancing both technical capabilities and cultural understanding in our rapidly evolving technological landscape.

Beyond the technical contributions, I hope this work serves to highlight the fundamental role that music plays in human development, a universal truth that transcends cultural boundaries, as Plato observed:

Ἆρ᾽ οὖν, ἦν δ᾽ ἐγώ, ὦ Γλαύκων, τούτων ἕνεκα κυριωτάτη ἐν μουσικῇ τροφή, ὅτι μάλιστα καταδύεται εἰς τὸ ἐντὸς τῆς ψυχῆς ὅ τε ῥυθμὸς καὶ ἁρμονία, καὶ ἐρρωμενέστατα ἅπτεται αὐτῆς φέροντα τὴν εὐσχημοσύνην, καὶ ποιεῖ εὐσχήμονα, ἐάν τις ὀρθῶς τραφῇ, εἰ δὲ μή, τοὐναντίον;[1]

- Πλάτων, Πολιτεία, Γ.401d

[1] *"And is it not for this reason, Glaucon, that education in music is most sovereign, because more than anything else rhythm and harmony find their way to the inmost soul and take strongest hold upon it, bringing with them and imparting grace, if one is rightly trained, and otherwise the contrary?"*, Plato, Republic, 3.401d

Appendices

Appendix A

Tags Distribution per Dataset

MagnaTagATune		FMA-medium		Lyra	
guitar	18.76%	Rock	28.41%	genres–Traditional	77.71%
classical	16.52%	Electronic	25.26%	instrument–Voice	74.39%
slow	13.71%	Punk	13.28%	instrument–Violin	56.18%
techno	11.42%	Experimental	9.0%	instrument–Percussion	54.27%
strings	10.55%	Hip-Hop	8.8%	instrument–Laouto	49.81%
drums	10.05%	Folk	6.08%	instrument–Guitar	39.43%
electronic	9.74%	Garage	5.67%	instrument–Klarino	32.17%
rock	9.17%	Instrumental	5.4%	genres–Nisiotiko	25.29%
fast	8.92%	Indie-Rock	5.17%	place–None	23.76%
piano	7.95%	Pop	4.74%	instrument–Accordion	22.68%
ambient	7.56%	Chip Music	4.12%	instrument–Bass	21.97%
beat	7.37%	International	4.07%	instrument–Santouri	19.3%
violin	7.06%	Ambient Electronic	4.02%	genres–Aegean	17.64%
vocal	6.69%	IDM	3.95%	place–Aegean-sea	17.64%
synth	6.64%	Soundtrack	3.28%	instrument–Bouzouki	14.01%
female	5.7%	Techno	3.21%	genres–Epirotic	13.69%
indian	5.39%	Downtempo	3.16%	place–Epirus	13.69%
opera	5.01%	House	2.79%	genres–Mikrasiatiko	10.19%
male	4.95%	Chiptune	2.78%	place–Asia-minor	10.19%
singing	4.68%	Trip-Hop	2.68%	genres–Rebetiko	9.17%
vocals	4.58%	Hardcore	2.63%	instrument–Oud	8.15%
no vocals	4.48%	Post-Punk	2.58%	instrument–Lyra	7.83%
harpsichord	4.23%	Psych-Rock	2.5%	genres–Laiko	6.56%
loud	4.2%	Classical	2.48%	instrument–Kanonaki	5.8%
quiet	4.08%	Dubstep	2.31%	instrument–Piano	5.8%
flute	3.96%	Metal	2.31%	genres–Macedonian	5.35%
woman	3.93%	Singer-Songwriter	2.3%	place–Macedonia	5.35%
male vocal	3.87%	Avant-Garde	2.24%	genres–Pontian	5.16%
no vocal	3.85%	Glitch	2.2%	place–Pontus	5.16%
pop	3.85%	Lo-Fi	2.2%	genres–Central-Greek	4.78%
soft	3.81%	Power-Pop	2.18%	place–Central-Greece	4.78%
sitar	3.58%	Loud-Rock	2.14%	genres–Peloponnesian	4.71%
solo	3.19%	Post-Rock	2.07%	place–Peloponnese	4.71%
man	2.87%	Experimental Pop	2.05%	instrument–Mandolin	4.52%
classic	2.67%	Old-Time / Historic	2.04%	instrument–Ney	3.95%
choir	2.66%	Dance	2.02%	place–Smyrni	3.89%
voice	2.57%	Noise	1.95%	instrument–Baglamas	3.76%
new age	2.51%	Ambient	1.63%	genres–Urban-folk	3.69%
dance	2.51%	Noise-Rock	1.58%	instrument–Tambouras	3.31%
female vocal	2.49%	Jazz	1.54%	instrument–Tsampouna	3.18%
male voice	2.49%	Chill-out	1.3%	genres–Cretan	3.12%
beats	2.45%	Rap	1.27%	place–Crete	3.12%
harp	2.41%	Electroacoustic	1.27%	genres–Newer	2.87%
cello	2.22%	Progressive	1.11%	genres–Ionian	2.8%
no voice	2.22%	Sound Collage	0.97%	place–Ionian-sea	2.8%
weird	2.15%	Reggae - Dub	0.93%	genres–Thracian	2.68%
country	2.09%	Improv	0.92%	place–Thrace	2.68%
female voice	1.95%	Shoegaze	0.92%	genres–Vlachic	2.61%
metal	1.95%	Balkan	0.88%	place–Metsovo	2.42%
choral	1.89%	Drum & Bass	0.87%	genres–Urban-light	2.36%

Table A.1. Top 50 Label Frequencies by Dataset (Part 1 of 2). Relative frequencies (%) of the most common labels in MagnaTagATune, FMA-medium, and Lyra datasets, highlighting the long-tailed distributions.

Turkish-makam		Hindustani		Carnatic	
instrument–Voice	63.33%	instrument–Voice	83.9%	instrument–Voice	82.35%
instrument–Kanun	31.09%	instrument–Tabla	53.03%	instrument–Violin	78.45%
instrument–Tanbur	27.93%	form–Khayal	41.33%	instrument–Mridangam	75.65%
instrument–Ney	27.56%	instrument–Harmonium	39.25%	form–Kriti	70.87%
instrument–performing orchestra	26.38%	tala–Teentaal	35.35%	tala–adi	51.88%
instrument–Oud	24.36%	instrument–Tambura	27.88%	instrument–Ghatam	30.32%
instrument–Classical kemençe	22.79%	tala–Ektaal	21.58%	instrument–Khanjira	17.65%
instrument–Cello	17.83%	instrument–Pakhavaj	7.88%	tala–rupaka	11.98%
instrument–Violin	17.62%	instrument–Sarangi	7.3%	tala–mishra chapu	7.27%
makam–Hicaz	10.63%	form–Dhrupad	7.05%	form–Varnam - Tana Varnam	5.21%
usul–Aksak	10.38%	form–Thumri	5.56%	form–Alapana	4.67%
instrument–Percussion	9.75%	tala–Jhaptaal	4.98%	tala–khanda chapu	3.22%
usul–Düyek	8.61%	instrument–Sitar	4.23%	form–Pallavi	2.99%
usul–Aksaksemai	8.53%	raga–Bhairabi	4.23%	instrument–Morsing	2.91%
makam–Nihavent	6.85%	form–Bhajan	4.07%	raga–ragamalika	2.87%
makam–Hüzzam	6.3%	raga–Yaman kalyan	3.32%	raga–thodi	2.6%
instrument–Clarinet	5.89%	form–Tarana	3.07%	form–Thillana	2.53%
usul–Curcuna	5.72%	tala–Rupak	2.99%	form–Mangalam	2.22%
instrument–Bendir	5.71%	instrument–Bansuri	2.9%	raga–bhairavi	2.18%
makam–Uşşak	5.71%	raga–Khamaj	2.49%	raga–kalyani	2.03%
makam–Kürdilihicazkar	5.39%	raga–Bageshree	2.32%	raga–kamas	1.99%
makam–Rast	5.0%	raga–Malkauns	2.32%	raga–saurashtram	1.95%
instrument–Kudüm	4.64%	instrument–Violin	2.07%	raga–sankarabharanam	1.65%
instrument–Viola	4.36%	raga–Des	2.07%	raga–kamavardani	1.45%
usul–Yürüksemai	4.17%	raga–Todi	1.99%	tala–atta	1.45%
usul–Sofyan	4.13%	raga–Marwa	1.91%	raga–behag	1.42%
makam–Segah	3.85%	raga–Miya malhar	1.91%	instrument–Thavil	1.38%
usul–Ağıraksak	3.54%	tala–Jhoomra	1.91%	raga–begada	1.34%
makam–Hüseyni	3.09%	raga–Lalat	1.83%	raga–mohanam	1.26%
usul–Devr-i Kebir	2.83%	instrument–Sarod	1.74%	form–Javali	1.23%
usul–Senginsemai	2.75%	raga–Ahir bhairav	1.74%	raga–sindhubhairavi	1.23%
instrument–Daire	2.64%	tala–Sooltal	1.74%	form–Thiruppugazh	1.19%
makam–Hicazkar	2.62%	instrument–Santoor	1.66%	instrument–Tambura	1.19%
usul–Semai	2.51%	raga–Bilaskhani todi	1.66%	raga–saveri	1.19%
makam–Mahur	2.47%	raga–Darbari	1.58%	raga–kamboji	1.15%
usul–Hafif	2.34%	raga–Mishra piloo	1.58%	raga–kapi	1.15%
instrument–Double bass	2.19%	raga–Bhairav	1.49%	raga–riti gaula	1.15%
usul–Nimsofyan	2.1%	raga–Hamsadhvani	1.49%	form–Tani avartanam	1.11%
makam–Suzinak	1.89%	raga–Bihag	1.41%	raga–Purvikalyani	1.11%
instrument–Strings	1.76%	raga–Basant	1.33%	raga–nata	1.11%
makam–Karcığar	1.68%	raga–Puriya dhanashree	1.33%	raga–surati	1.11%
makam–Muhayyer	1.68%	raga–Bhoop	1.24%	raga–hamsadhvani	1.07%
usul–Türkaksağı	1.67%	raga–Mishra maand	1.24%	raga–madyamavati	1.07%
makam–Saba	1.65%	tala–Chautal	1.24%	form–Bhajan	1.03%
usul–Muhammes	1.55%	form–Dadra	1.16%	raga–hindolam	1.03%
usul–Serbest	1.48%	instrument–Shehnai	1.16%	raga–karaharapriya	1.0%
instrument–Accordion	1.4%	raga–Abhogi	1.16%	form–Keertana (Devara nama)	0.96%
makam–Beyati	1.4%	raga–Jog	1.16%	raga–ananda bhairavi	0.96%
makam–Acemaşiran	1.37%	raga–Madhuvanti	1.16%	raga–kanada	0.96%
makam–Neva	1.35%	raga–Sohini	1.16%	raga–mukhari	0.96%

Table A.2. Top 50 Label Frequencies by Dataset (Part 2 of 2). Relative frequencies (%) of the most common labels in Turkish-makam, Hindustani, and Carnatic datasets, showing similar long-tailed patterns across non-Western traditions.

The above tables present the relative frequencies of the top 50 labels in each of the six datasets examined in this work. These distributions highlight the long-tailed nature of musical attributes across diverse cultural traditions, with a small number of tags accounting for the majority of annotations while many of them appear only rarely. This imbalance demonstrates the need for specialized methods to handle rare but semantically important musical attributes.

Appendix B

Supplementary Material for Multi-Label Few-Shot Learning

The following table shows the number of: prototypes N_P, unique items in the support set $|S|$ and query set $|Q|$, true labels per item N_ℓ/item and predicted labels per item $N_{\hat{\ell}}$/item across all datasets and methods. The ML-FSL task follows a "30-way 3-shot" setup, including both "Base" and "Novel" classes, with the model trained "from scratch" and evaluated on each dataset and method. The table presents the average values over 5 runs, along with the standard deviation. The variation in the number of prototypes N_P for the LC-Protonets method arises from sampling different support items at each run, as their label combination power sets result in varying LC-Prototypes. Notably, the proposed method maintains the number of predicted labels $N_{\hat{\ell}}$/item close to the true value across datasets, unlike the comparative approaches.

dataset	method	N_P	$\lvert S\rvert$	$\lvert Q\rvert$	$\frac{N_\ell}{item}$	$\frac{N_{\hat{\ell}}}{item}$
MagnaTag-ATune	**ML-PNs**	30	49	3761	2.3	15.8±1.6
	One-vs.-Rest	30				18.8±2.2
	LC-Protonets	**708±366**				**2.3±0.2**
FMA-medium	**ML-PNs**	30	60	4669	1.8	13.3±0.3
	One-vs.-Rest	30				17.4±0.9
	LC-Protonets	**115±11**				**2.2±0.1**
Lyra	**ML-PNs**	30	32	298	6.0	22.4±3.8
	One-vs.-Rest	30				27.3±4.4
	LC-Protonets	**3361±831**				**5.2±0.1**
Turkish-makam	**ML-PNs**	30	42	1015	3.7	30.0±0.0
	One-vs.-Rest	30				29.9±0.2
	LC-Protonets	**15875±2647**				**4.0±0.4**
Hindu-stani	**ML-PNs**	30	51	186	3.7	25.3±3.7
	One-vs.-Rest	30				27.8±3.2
	LC-Protonets	**1225±87**				**3.9±0.1**
Carnatic	**ML-PNs**	30	53	469	4.8	19.4±4.3
	One-vs.-Rest	30				27.6±3.5
	LC-Protonets	**1438±113**				**4.8±0.1**

Table B.1. ML-FSL Method Operational Metrics. Comparison of prototype count, support-/query characteristics and number of predicted labels, across methods and datasets for a "30-way 3-shot" task with both base and novel classes.

The following tables present the complete performance results for multi-label few-shot learning methods across all six music datasets. Specifically, Macro-F1 (M-F1) and micro-F1 (m-F1) scores (%) with 95% confidence intervals are reported, when utilizing (i) training from scratch, (ii) a pre-trained backbone model followed by full fine-tuning, (iii) a pre-trained backbone model followed by fine-tuning of the final layer, and (iv) a pre-trained backbone model without any fine-tuning. Rows represent the multi-label few-shot learning methods, and columns represent the domains. For each domain and method, evaluation is performed on a "15-way 3-shot" task with "Novel" classes, and a "45-way 3-shot" task with "Base & Novel" classes. It is observed that using a pre-trained backbone improves the performance of all methods, with LC-Protonets proving to be the best approach in almost all setups and domains. Also, when no fine-tuning is performed, the comparative approaches show a significant drop in performance, while LC-Protonets maintain similar or even better results compared to the respective fine-tuned models.

dataset	**MagnaTagATune**				**FMA-medium**			
ML-FSL task	*15-way 3-shot*		*45-way 3-shot*		*15-way 3-shot*		*45-way 3-shot*	
metric	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
method	*training from scratch*							
ML-PNs	20.9±1.61	20.53±1.9	15.69±0.71	16.89±0.91	**23.58±1.2**	22.61±1.06	13.67±0.44	14.91±0.41
One-vs.-Rest	19.16±0.79	18.8±0.79	13.4±0.58	13.78±0.77	18.22±2.3	17.93±2.3	10.2±0.37	10.66±0.5
LC-Protonets	**25.04±3.25**	**27.04±3.0**	**18.77±1.48**	**27.69±1.77**	20.01±2.28	**23.42±2.1**	**14.19±0.77**	**28.83±1.73**
method	*pre-trained backbone and full fine-tuning*							
ML-PNs	28.51±0.36	27.26±0.43	21.02±0.16	22.71±0.11	26.34±1.42	25.39±1.39	16.63±0.42	17.59±0.53
One-vs.-Rest	22.19±0.91	21.9±0.97	15.81±0.81	16.67±0.99	23.79±1.24	23.15±1.74	13.81±0.44	14.27±0.76
LC-Protonets	**40.4±1.33**	**42.85±1.37**	**29.52±1.54**	**40.4±2.01**	**30.58±3.45**	**33.23±3.55**	**26.63±1.6**	**42.66±1.11**
method	*pre-trained backbone and fine-tuning of the last layer*							
ML-PNs	28.1±0.72	26.88±0.8	20.52±0.37	22.02±0.58	25.39±1.16	24.55±1.14	15.9±0.32	17.05±0.45
One-vs.-Rest	24.5±3.3	23.82±3.19	16.06±1.39	16.56±1.64	24.34±1.25	23.42±1.67	14.27±0.43	15.18±0.67
LC-Protonets	**40.68±1.46**	**42.99±1.09**	**29.14±1.62**	**40.11±2.09**	**31.0±2.47**	**33.68±2.41**	**26.37±1.24**	**42.65±0.71**
method	*pre-trained backbone without any fine-tuning*							
ML-PNs	14.57±0.02	14.6±0.01	10.9±0.02	11.4±0.02	13.71±0.01	13.72±0.01	7.9±0.01	8.35±0.01
One-vs.-Rest	14.56±0.06	14.59±0.06	10.9±0.02	11.39±0.02	13.71±0.03	13.72±0.03	7.89±0.01	8.35±0.01
LC-Protonets	**40.6±0.88**	**42.95±1.13**	**29.19±1.65**	**40.13±2.02**	**31.52±2.45**	**33.81±2.71**	**26.82±1.5**	**43.23±0.8**

dataset	**Lyra**				**Turkish-makam**			
ML-FSL task	*15-way 3-shot*		*45-way 3-shot*		*15-way 3-shot*		*45-way 3-shot*	
metric	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
method	*training from scratch*							
ML-PNs	30.26±4.73	29.97±4.77	26.22±1.39	30.82±1.72	13.35±0.11	13.42±0.11	14.96±0.08	16.58±0.09
One-vs.-Rest	24.99±3.54	25.0±3.41	24.08±0.98	27.67±1.54	13.24±0.09	13.31±0.11	14.85±0.07	16.43±0.06
LC-Protonets	**46.39±7.11**	**48.82±8.16**	**42.81±5.47**	**60.47±5.99**	**14.39±2.93**	**17.39±2.66**	**18.85±1.93**	**37.0±2.59**
method	*pre-trained backbone and full fine-tuning*							
ML-PNs	34.33±1.41	34.2±1.16	30.81±1.22	36.03±1.44	17.74±0.93	17.4±0.63	23.89±0.52	24.26±0.5
One-vs.-Rest	26.02±3.95	25.77±3.51	24.22±0.72	27.32±0.86	15.22±0.7	15.12±0.57	18.49±0.88	20.41±1.16
LC-Protonets	**57.2±6.75**	**59.69±6.25**	**50.17±2.54**	**68.49±2.56**	**23.09±1.45**	**24.13±1.12**	**32.31±0.9**	**56.45±0.64**
method	*pre-trained backbone and fine-tuning of the last layer*							
ML-PNs	35.52±2.09	35.34±2.04	31.35±0.4	37.15±0.51	16.35±1.44	16.15±1.23	21.05±2.12	22.95±1.89
One-vs.-Rest	28.18±1.91	28.1±1.9	27.16±1.11	31.33±1.36	16.23±0.63	16.03±0.7	20.5±1.69	22.35±1.77
LC-Protonets	**61.49±4.62**	**63.77±4.11**	**52.36±2.94**	**70.88±2.55**	**24.26±2.01**	**25.45±1.95**	**33.12±0.81**	**57.87±0.7**
method	*pre-trained backbone without any fine-tuning*							
ML-PNs	23.45±0.08	23.53±0.08	23.71±0.11	27.13±0.15	13.32±0.11	13.38±0.11	14.96±0.06	16.57±0.07
One-vs.-Rest	23.49±0.13	23.56±0.12	23.68±0.04	27.08±0.09	13.22±0.07	13.28±0.07	14.84±0.04	16.42±0.05
LC-Protonets	**64.13±2.21**	**65.64±2.6**	**52.4±2.5**	**70.78±2.24**	**26.08±2.65**	**26.29±2.36**	**32.44±1.2**	**56.34±0.92**

Table B.2. ML-FSL Performance Across Training Conditions (Part 1 of 2). Macro-F1 and micro-F1 scores (%) with confidence intervals for MagnaTagATune, FMA-medium, and Lyra datasets across four training scenarios and three ML-FSL methods.

dataset	**Hindustani**				**Carnatic**			
ML-FSL task	*15-way 3-shot*		*45-way 3-shot*		*15-way 3-shot*		*45-way 3-shot*	
metric	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
method	*training from scratch*							
ML-PNs	13.34±0.98	13.54±0.76	-	-	**12.87±1.16**	**13.05±1.18**	**14.18±0.48**	22.48±1.66
One-vs.-Rest	13.03±0.4	13.19±0.46	-	-	12.79±0.14	12.86±0.17	14.05±1.12	19.71±0.69
LC-Protonets	**20.18±7.89**	**25.74±8.76**	-	-	9.35±3.25	11.13±2.87	13.05±0.9	**54.64±1.65**
method	*pre-trained backbone and full fine-tuning*							
ML-PNs	16.43±1.55	15.85±1.28	-	-	**13.13±0.41**	**13.16±0.41**	15.23±0.45	22.72±1.77
One-vs.-Rest	14.4±2.24	14.1±1.57	-	-	13.03±0.21	13.08±0.2	14.66±0.28	20.81±1.28
LC-Protonets	**21.82±3.78**	**29.78±5.23**	-	-	10.13±4.22	11.86±4.57	**17.18±1.35**	**62.7±2.95**
method	*pre-trained backbone and fine-tuning of the last layer*							
ML-PNs	17.2±1.81	16.2±1.48	-	-	12.83±0.34	12.91±0.29	14.98±0.81	21.51±2.65
One-vs.-Rest	15.61±1.89	15.04±1.41	-	-	**13.4±0.62**	**13.44±0.62**	14.37±0.53	20.52±0.91
LC-Protonets	**26.08±3.08**	**32.22±4.42**	-	-	10.82±3.66	11.43±3.78	**16.94±1.32**	**63.34±1.81**
method	*pre-trained backbone without any fine-tuning*							
ML-PNs	12.7±0.47	12.85±0.48	-	-	12.72±0.2	12.77±0.2	14.47±0.06	19.82±0.09
One-vs.-Rest	12.84±0.21	12.97±0.22	-	-	**12.74±0.12**	**12.78±0.12**	14.55±0.04	19.91±0.06
LC-Protonets	**26.06±3.18**	**31.43±3.02**	-	-	11.3±3.91	10.91±3.34	**17.02±0.94**	**62.35±2.36**

Table B.3. ML-FSL Performance Across Training Conditions (Part 2 of 2). Macro-F1 and micro-F1 scores (%) with confidence intervals for Turkish-makam, Hindustani, and Carnatic datasets. Note: "-" denotes that there are not enough data samples for the "N-way K-shot" setup in the dataset.

Appendix C

Signal Processing Feature Implementation Details

This appendix provides complete mathematical formulations and implementation details for all signal processing features described in the respective Section of the main paper.

A multi-dimensional framework is developed treating rhythm, melody, harmony, and timbre as distinct but complementary dimensions. Each employs specialized feature extraction and similarity computation methods preserving unique characteristics and temporal dynamics.

C.1 Melody Feature Analysis

Melody analysis extracts melodic content from polyphonic audio by treating fundamental frequency (F0) as the dominant frequency skeleton capturing prominent melodic content. The framework uses dual-resolution pitch class analysis with both semitone (Western music) and quarter-tone (multi-cultural) representations for comprehensive melodic characterization.

C.1.1 Feature Extraction

Let $y[n]$ denote the discrete audio signal with sampling rate $f_s = 22050$ Hz and hop length $H = 512$ samples.

F0 Extraction: PYIN algorithm for robust F0 estimation [56]:

$$\mathbf{f}_0, \mathbf{v}, \mathbf{p} = \text{PYIN}(y, f_{\min}, f_{\max}, f_s, H), \tag{C.1}$$

where $f_{\min} = 65.4$ Hz (C2), $f_{\max} = 2093$ Hz (C7), $\mathbf{v}$ is voicing probability, and $\mathbf{p}$ contains confidence values. Clean F0 extraction:

$$\mathbf{f}_0^{\text{clean}} = \mathbf{f}_0[\mathbf{v} \wedge \text{isfinite}(\mathbf{f}_0)]. \tag{C.2}$$

Dual-Resolution Pitch Class Analysis: For each frequency $f \in \mathbf{f}_0^{\text{clean}}$, compute MIDI-like number:

$$m = 12 \log_2 \left(\frac{f}{440} \right) + 69. \tag{C.3}$$

Quarter-tone pitch classes (24 bins per octave):

$$pc_{\text{quarter}} = \lfloor (2m) \bmod 24 \rfloor. \tag{C.4}$$

Semitone pitch classes (12 bins per octave):

$$pc_{\text{semi}} = \lfloor m \bmod 12 \rfloor. \tag{C.5}$$

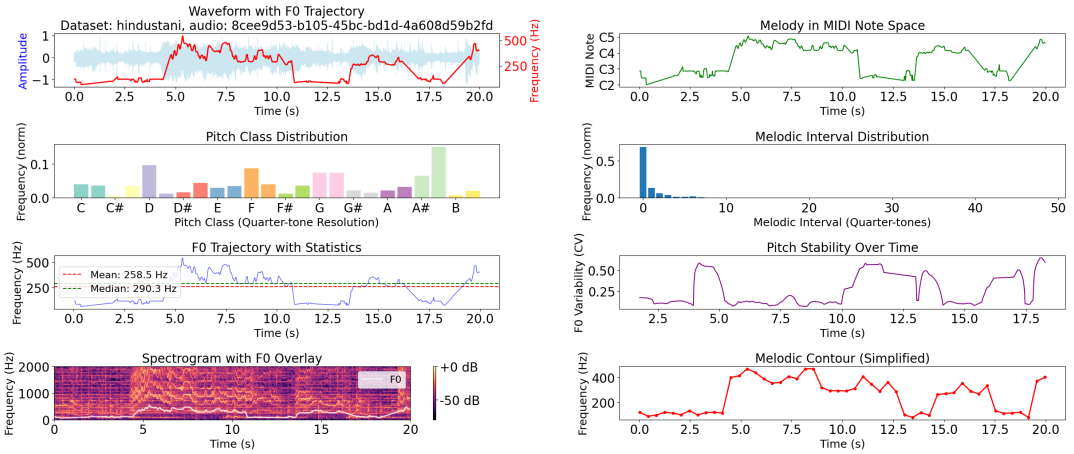

Figure C.1. Melody Feature Extraction on Hindustani Classical Music. Comprehensive visualization of melodic analysis components. Top left: waveform with F0 trajectory (red) capturing microtonal ornamentations. Top right: melody representation in MIDI space showing pitch structure. Middle: 24-pitch class distribution with detected micro-tones and melodic interval distribution dominated by small intervals. Bottom: pitch stability and contour analysis demonstrating both local ornamental details and global melodic structure.

Normalized pitch class histograms:

$$h_{pc,\text{quarter}}[k] = \sum_{i=1}^{|\mathbf{f}_0^{\text{clean}}|} \mathbf{1}[pc_{\text{quarter},i} = k], \quad k = 0, 1, \ldots, 23 \tag{C.6}$$

$$h_{pc,\text{semi}}[k] = \sum_{i=1}^{|\mathbf{f}_0^{\text{clean}}|} \mathbf{1}[pc_{\text{semi},i} = k], \quad k = 0, 1, \ldots, 11 \tag{C.7}$$

$$\mathbf{h}^*_{pc,\text{quarter}} = \frac{\mathbf{h}_{pc,\text{quarter}}}{||\mathbf{h}_{pc,\text{quarter}}||_1}, \quad \mathbf{h}^*_{pc,\text{semi}} = \frac{\mathbf{h}_{pc,\text{semi}}}{||\mathbf{h}_{pc,\text{semi}}||_1}. \tag{C.8}$$

Melodic Interval Analysis: Consecutive F0 intervals in quarter-tone units:

$$I_i = 24 \log_2 \left(\frac{f_{0,i+1}}{f_{0,i}} \right), \quad i = 1, \ldots, |\mathbf{f}_0^{\text{clean}}| - 1. \tag{C.9}$$

Interval histogram $\mathbf{h}_I \in \mathbb{R}^{49}$ for intervals $[-48, +48]$ quarter-tones:

$$h_I[k] = \sum_i \mathbf{1}[|\text{clip}(I_i, -48, 48)| = k], \quad k = 0, 1, \ldots, 48 \tag{C.10}$$

$$\mathbf{h}^*_I = \frac{\mathbf{h}_I}{||\mathbf{h}_I||_1}. \tag{C.11}$$

Using absolute value $|I_i|$ creates symmetric representation focusing on interval magnitude and we apply a rational clipping in a range of 4 octaves.

Statistical Characterization: F0 distribution statistics:

$$\mu_{f0} = \frac{1}{|\mathbf{f}_0^{\text{clean}}|} \sum_{i=1}^{|\mathbf{f}_0^{\text{clean}}|} f_{0,i} \quad \text{(mean)}, \tag{C.12}$$

$$\sigma_{f0} = \sqrt{\frac{1}{|\mathbf{f}_0^{\text{clean}}|} \sum_{i=1}^{|\mathbf{f}_0^{\text{clean}}|} (f_{0,i} - \mu_{f0})^2} \quad \text{(std)}, \tag{C.13}$$

$$\Delta_{f0} = \max(\mathbf{f}_0^{\text{clean}}) - \min(\mathbf{f}_0^{\text{clean}}) \quad \text{(range)}, \tag{C.14}$$

$$\tilde{f}_0 = \text{median}(\mathbf{f}_0^{\text{clean}}) \quad \text{(median)}. \tag{C.15}$$

C.1.2 Similarity Computation

Four weighted similarity components incorporating dual-resolution pitch class analysis:

$$S_{\text{melody}} = 0.15 \cdot S_{pc,\text{quarter}} + 0.15 \cdot S_{pc,\text{semi}} + 0.4 \cdot S_I + 0.3 \cdot S_{stats}. \tag{C.16}$$

Pitch Class Similarities: Cosine similarity of normalized histograms for both resolutions:

$$S_{pc,\text{quarter}} = \frac{\mathbf{h}^*_{pc,\text{quarter},1} \cdot \mathbf{h}^*_{pc,\text{quarter},2}}{||\mathbf{h}^*_{pc,\text{quarter},1}||_2 \cdot ||\mathbf{h}^*_{pc,\text{quarter},2}||_2}, \tag{C.17}$$

$$S_{pc,\text{semi}} = \frac{\mathbf{h}^*_{pc,\text{semi},1} \cdot \mathbf{h}^*_{pc,\text{semi},2}}{||\mathbf{h}^*_{pc,\text{semi},1}||_2 \cdot ||\mathbf{h}^*_{pc,\text{semi},2}||_2}. \tag{C.18}$$

Interval Similarity: Excluding zero interval for melodic motion focus:

$$\mathbf{h}^*_{I,\text{motion}} = \frac{[\mathbf{h}^*_I]_{k=1}^{48}}{||[\mathbf{h}^*_I]_{k=1}^{48}||_1}, \tag{C.19}$$

$$S_I = \frac{\mathbf{h}^*_{I,\text{motion},1} \cdot \mathbf{h}^*_{I,\text{motion},2}}{||\mathbf{h}^*_{I,\text{motion},1}||_2 \cdot ||\mathbf{h}^*_{I,\text{motion},2}||_2}. \tag{C.20}$$

Statistical Similarity: Normalized differences of F0 statistics:

$$S_\mu = 1 - \frac{|\mu_{f0,1} - \mu_{f0,2}|}{\max(\mu_{f0,1}, \mu_{f0,2})}, \tag{C.21}$$

$$S_\sigma = 1 - \frac{|\sigma_{f0,1} - \sigma_{f0,2}|}{\max(\sigma_{f0,1}, \sigma_{f0,2})}, \tag{C.22}$$

$$S_{\tilde{f}} = 1 - \frac{|\tilde{f}_{0,1} - \tilde{f}_{0,2}|}{\max(\tilde{f}_{0,1}, \tilde{f}_{0,2})}, \tag{C.23}$$

$$S_\Delta = 1 - \frac{|\Delta_{f0,1} - \Delta_{f0,2}|}{\max(\Delta_{f0,1}, \Delta_{f0,2})}, \tag{C.24}$$

$$S_{\text{range}} = \frac{1}{2}(S_\Delta + S_\mu), \tag{C.25}$$

$$S_{stats} = \frac{1}{4}(S_\mu + S_\sigma + S_{\tilde{f}} + S_{\text{range}}). \tag{C.26}$$

The dual-resolution framework combines the robustness of semitone analysis for Western musical patterns with the sensitivity of quarter-tone analysis for microtonal ornamentations, making it suitable for cross-cultural melodic similarity assessment. It handles, also, polyphonic F0 challenges through statistical distributions and interval patterns, making it robust to discontinuous F0 trajectories while preserving melodic characteristics.

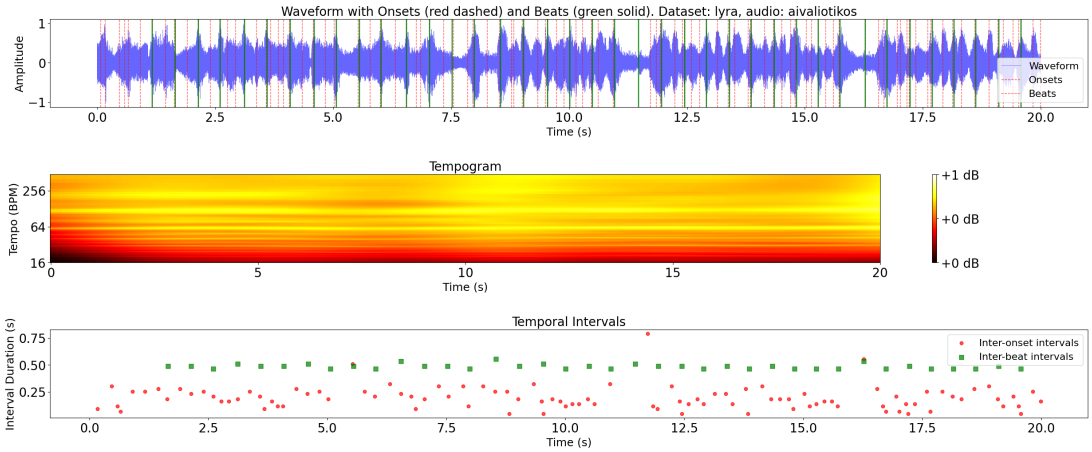

Figure C.2. Rhythm Feature Extraction on Lyra Dataset Example. Multi-panel visualization of rhythmic analysis components. Top: waveform with detected onsets (red dashed) and beats (green solid). Middle: tempogram showing tempo consistency around 32-64 BPM. Bottom: temporal intervals between onsets (red dots) and beats (green squares) demonstrating rhythmic regularity and variation patterns.

C.2 Rhythm Feature Analysis

Rhythm analysis captures the temporal foundation of music through pulse, meter, timing patterns, and rhythmic density. Our framework uses a four-component similarity measure combining tempo tracking, onset detection, beat analysis, and tempogram features.

C.2.1 Feature Extraction

Let $y[n]$ denote the discrete audio signal with sampling rate $f_s = 22050$ Hz and hop length $H = 512$ samples.

Tempo and Beat Tracking: We use librosa's dynamic programming-based beat tracking [60] to extract tempo T (BPM) and beat locations $B = \{b_1, b_2, \ldots, b_N\}$:

$$T, B = \text{beat\_track}(y, f_s, H). \tag{C.27}$$

Figure C.2 shows successful beat tracking on traditional Lyra music despite complex ornamental patterns.

Onset Detection: Onset detection identifies musical event beginnings using the complex domain algorithm [191, 192]:

$$O = \text{onset\_detect}(y, f_s, H). \tag{C.28}$$

Key rhythmic features from onset times $O = \{o_1, o_2, \ldots, o_M\}$:

$$\text{IOI} = \{o_{i+1} - o_i : i = 1, \ldots, M-1\} \quad \text{(inter-onset intervals)}, \tag{C.29}$$

$$\rho = \frac{M}{L/f_s} \quad \text{(onset density)}, \tag{C.30}$$

$$\mu_{IOI} = \frac{1}{M-1} \sum_{i=1}^{M-1} (o_{i+1} - o_i), \tag{C.31}$$

$$\sigma_{IOI} = \sqrt{\frac{1}{M-1} \sum_{i=1}^{M-1} (o_{i+1} - o_i - \mu_{IOI})^2}, \tag{C.32}$$

where L is audio length, ρ is onset density, μ_{IOI} is mean inter-onset interval, and σ_{IOI} measures timing variability.

Beat Interval Analysis: Inter-beat intervals from detected positions:

$$t_B = \text{frames\_to\_time}(B, f_s, H), \tag{C.33}$$

$$\text{IBI} = \{t_{B,i+1} - t_{B,i} : i = 1, \ldots, N-1\}, \tag{C.34}$$

$$\mu_{IBI} = \frac{1}{N-1} \sum_{i=1}^{N-1} (t_{B,i+1} - t_{B,i}). \tag{C.35}$$

Tempogram Analysis: Tempo information across time-frequency bins k [193]:

$$\mathbf{T}(k, n) = \text{tempogram}(y, f_s, H). \tag{C.36}$$

We extract the tempo profiles by averaging across time:

$$\mathbf{p}_T(k) = \frac{1}{N_t} \sum_{n=1}^{N_t} \mathbf{T}(k, n). \tag{C.37}$$

C.2.2 Similarity Computation

Four equally weighted similarity components:

$$S_{\text{rhythm}} = \frac{1}{4}(S_{\text{tempo}} + S_{\text{onset}} + S_{\text{beat}} + S_{\text{tempogram}}). \tag{C.38}$$

Tempo Similarity: Normalized difference:

$$S_{\text{tempo}} = 1 - \frac{|T_1 - T_2|}{\max(T_1, T_2)}. \tag{C.39}$$

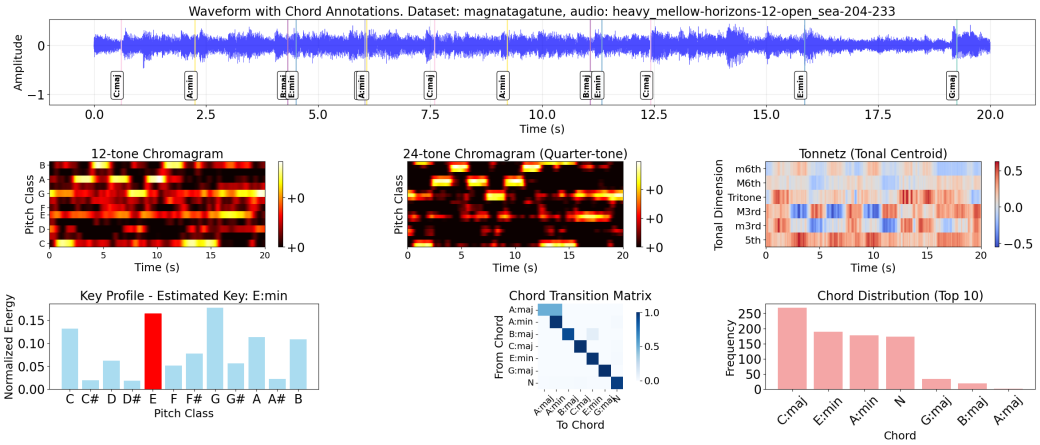

Figure C.3. Harmony Feature Extraction on MagnaTagATune Example. Comprehensive harmonic analysis visualization. Top: waveform with chord annotations from template matching. Middle: 12-tone chromagram, 24-tone chromagram, and Tonnetz tonal centroids capturing harmonic content. Bottom: key profile analysis (E:minor), chord transition matrix, and chord distribution showing harmonic relationships and progressions.

Onset Pattern Similarity: Three onset measures:

$$S_\rho = 1 - \frac{|\rho_1 - \rho_2|}{\max(\rho_1, \rho_2)}, \tag{C.40}$$

$$S_\mu = 1 - \frac{|\mu_{IOI,1} - \mu_{IOI,2}|}{\max(\mu_{IOI,1}, \mu_{IOI,2})}, \tag{C.41}$$

$$S_\sigma = 1 - \frac{|\sigma_{IOI,1} - \sigma_{IOI,2}|}{\max(\sigma_{IOI,1}, \sigma_{IOI,2})}, \tag{C.42}$$

$$S_{\text{onset}} = \frac{1}{3}(S_\rho + S_\mu + S_\sigma). \tag{C.43}$$

Beat Pattern Similarity: Mean beat interval comparison:

$$S_{\text{beat}} = 1 - \frac{|\mu_{IBI,1} - \mu_{IBI,2}|}{\max(\mu_{IBI,1}, \mu_{IBI,2})}. \tag{C.44}$$

Tempogram Similarity: Cosine similarity of normalized tempo profiles:

$$\mathbf{p}^*_{T,1} = \frac{\mathbf{p}_{T,1}}{||\mathbf{p}_{T,1}||_1}, \quad \mathbf{p}^*_{T,2} = \frac{\mathbf{p}_{T,2}}{||\mathbf{p}_{T,2}||_1}, \tag{C.45}$$

$$S_{\text{tempogram}} = \frac{\mathbf{p}^*_{T,1} \cdot \mathbf{p}^*_{T,2}}{||\mathbf{p}^*_{T,1}||_2 \cdot ||\mathbf{p}^*_{T,2}||_2}. \tag{C.46}$$

C.3 Harmony Feature Analysis

Harmony analysis captures vertical music structure through complementary representations addressing local chord structures and global tonal relationships. The framework combines chroma features, chord recognition, key estimation, and tonal centroids.

C.3.1 Feature Extraction

Let $y[n]$ denote the discrete audio signal with sampling rate $f_s = 22050$ Hz and hop length $H = 512$ samples.

Chroma Features: CENS features at 24-bin (quarter-tone) and 12-bin (semitone) resolutions [194]:

$$\begin{aligned}\mathbf{C}_{24} &= \text{chroma\_cens}(y, f_s, H, \\ &\qquad \text{bins\_per\_octave} = 24, \text{n\_chroma} = 24), \end{aligned} \tag{C.47}$$

$$\mathbf{C}_{12} = \text{chroma\_cens}(y, f_s, H, \text{n\_chroma} = 12), \tag{C.48}$$

where $\mathbf{C}_{24} \in \mathbb{R}^{24 \times N_t}$ and $\mathbf{C}_{12} \in \mathbb{R}^{12 \times N_t}$.

Chord Templates: Major and minor triad templates for chromatic pitch classes $\mathcal{N} = \{\text{C}, \text{C}\#, \text{D}, \text{D}\#, \text{E}, \text{F}, \text{F}\#, \text{G}, \text{G}\#, \text{A}, \text{A}\#, \text{B}\}$:

$$\mathbf{t}_{i,\text{maj}}[k] = \begin{cases} 1 & \text{if } k \in \{i, (i+4) \bmod 12, (i+7) \bmod 12\} \\ 0 & \text{otherwise} \end{cases}, \tag{C.49}$$

$$\mathbf{t}_{i,\text{min}}[k] = \begin{cases} 1 & \text{if } k \in \{i, (i+3) \bmod 12, (i+7) \bmod 12\} \\ 0 & \text{otherwise} \end{cases}. \tag{C.50}$$

Plus no-chord template: $\mathbf{t}_N = \frac{1}{12}\mathbf{1}_{12}$.

Chord Recognition: Template matching for each time frame:

$$\mathbf{c}_n^* = \frac{\mathbf{C}_{12}(\cdot, n)}{||\mathbf{C}_{12}(\cdot, n)||_1}, \tag{C.51}$$

$$P_{i,n} = \frac{\mathbf{c}_n^* \cdot \mathbf{t}_i}{||\mathbf{c}_n^*||_2 \cdot ||\mathbf{t}_i||_2 + \epsilon}, \tag{C.52}$$

$$\hat{c}_n = \arg\max_i P_{i,n}, \tag{C.53}$$

where $\epsilon = 10^{-8}$ and i indexes 49 templates (24 major + 24 minor + 1 no-chord).

Key Estimation: Krumhansl-Schmuckler algorithm [195]:

$$\mathbf{p}_{\text{obs}} = \frac{1}{N_t} \sum_{n=1}^{N_t} \mathbf{C}_{12}(\cdot, n), \tag{C.54}$$

$$\mathbf{p}_{\text{obs}}^* = \frac{\mathbf{p}_{\text{obs}}}{||\mathbf{p}_{\text{obs}}||_1}. \tag{C.55}$$

Theoretical key profiles:

$$\begin{aligned}\mathbf{k}_{\text{maj}} &= [6.35, 2.23, 3.48, 2.33, 4.38, 4.09, \\ &\qquad 2.52, 5.19, 2.39, 3.66, 2.29, 2.88]^T, \end{aligned} \tag{C.56}$$

$$\begin{aligned}\mathbf{k}_{\text{min}} &= [6.33, 2.68, 3.52, 5.38, 2.60, 3.53, \\ &\qquad 2.54, 4.75, 3.98, 2.69, 3.34, 3.17]^T. \end{aligned} \tag{C.57}$$

Key estimation through profile correlation:

$$r_{i,\text{maj}} = \text{corr}(\mathbf{p}^*_{\text{obs}}, \text{circshift}(\mathbf{k}^*_{\text{maj}}, i)), \tag{C.58}$$

$$r_{i,\text{min}} = \text{corr}(\mathbf{p}^*_{\text{obs}}, \text{circshift}(\mathbf{k}^*_{\text{min}}, i)), \tag{C.59}$$

$$\hat{k} = \arg\max_{i,m} r_{i,m}. \tag{C.60}$$

Chord Transition Matrix: From chord sequence $\{\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_{N_t}\}$:

$$T_{ij} = \sum_{n=1}^{N_t-1} \mathbf{1}[\hat{c}_n = i \wedge \hat{c}_{n+1} = j], \tag{C.61}$$

$$\mathbf{T}^*_{ij} = \frac{T_{ij}}{\sum_k T_{ik}}. \tag{C.62}$$

Tonnetz Features: Tonal centroid mapping to geometric space [196]:

$$\mathbf{X} = \text{tonnetz}(\text{harmonic}(y), f_s, H) \in \mathbb{R}^{6\times N_t}. \tag{C.63}$$

C.3.2 Similarity Computation

Five equally weighted harmony components:

$$S_{\text{harmony}} = \frac{1}{5}(S_{\text{chroma}} + S_{\text{key}} + S_{\text{chord\_dist}} + S_{\text{chord\_trans}} + S_{\text{tonnetz}}). \tag{C.64}$$

Chroma Similarity: 24-bin chroma profiles:

$$\bar{\mathbf{c}}_1 = \frac{1}{N_{t,1}} \sum_{n=1}^{N_{t,1}} \mathbf{C}_{24,1}(\cdot, n), \quad \bar{\mathbf{c}}_2 = \frac{1}{N_{t,2}} \sum_{n=1}^{N_{t,2}} \mathbf{C}_{24,2}(\cdot, n),, \tag{C.65}$$

$$\bar{\mathbf{c}}^*_1 = \frac{\bar{\mathbf{c}}_1}{||\bar{\mathbf{c}}_1||_1}, \quad \bar{\mathbf{c}}^*_2 = \frac{\bar{\mathbf{c}}_2}{||\bar{\mathbf{c}}_2||_1}, \tag{C.66}$$

$$S_{\text{chroma}} = \frac{\bar{\mathbf{c}}^*_1 \cdot \bar{\mathbf{c}}^*_2}{||\bar{\mathbf{c}}^*_1||_2 \cdot ||\bar{\mathbf{c}}^*_2||_2}. \tag{C.67}$$

Key Similarity: Profile and estimated key combination:

$$S_{\text{profile}} = \frac{\mathbf{p}^*_{\text{obs},1} \cdot \mathbf{p}^*_{\text{obs},2}}{||\mathbf{p}^*_{\text{obs},1}||_2 \cdot ||\mathbf{p}^*_{\text{obs},2}||_2}, \tag{C.68}$$

$$S_{\text{estimated}} = S_{\text{CoF}}(\hat{k}_1, \hat{k}_2), \tag{C.69}$$

$$S_{\text{key}} = \frac{1}{2}(S_{\text{profile}} + S_{\text{estimated}}). \tag{C.70}$$

Chord Distribution Similarity: Occurrence frequencies:

$$\mathbf{d}_1[i] = \frac{|\{n : \hat{c}_{1,n} = i\}|}{N_{t,1}}, \quad \mathbf{d}_2[i] = \frac{|\{n : \hat{c}_{2,n} = i\}|}{N_{t,2}}, \tag{C.71}$$

$$S_{\text{chord\_dist}} = \frac{\mathbf{d}_1 \cdot \mathbf{d}_2}{||\mathbf{d}_1||_2 \cdot ||\mathbf{d}_2||_2}. \tag{C.72}$$

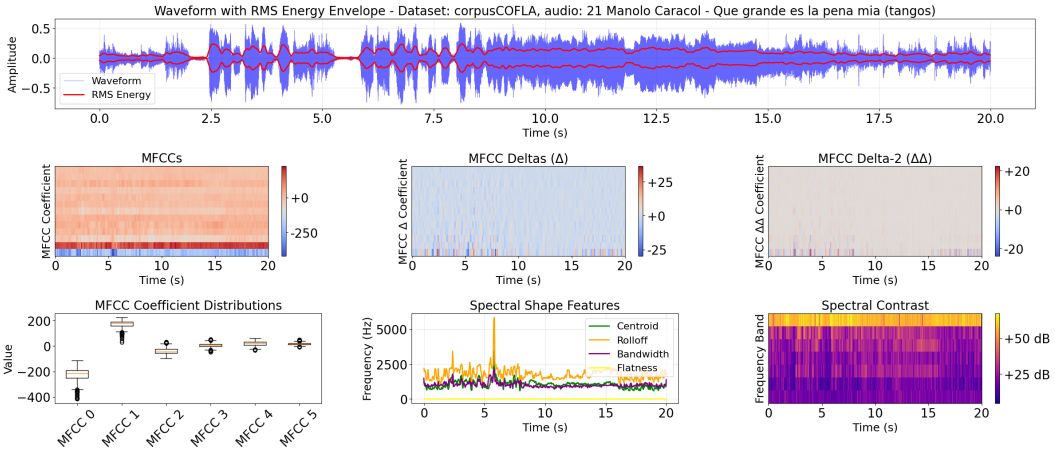

Figure C.4. Timbre Feature Extraction on CorpusCOFLA Example. Comprehensive timbral analysis visualization. Top: waveform with RMS energy envelope. Middle: MFCC coefficients and delta features capturing timbral evolution over time. Bottom: MFCC distributions (left), spectral shape features (middle), and spectral contrast analysis (right).

Chord Transition Similarity: Transition pattern comparison:

$$\mathbf{f}_1 = \text{flatten}(\mathbf{T}_1^*[\mathbf{M}]), \quad \mathbf{f}_2 = \text{flatten}(\mathbf{T}_2^*[\mathbf{M}]), \tag{C.73}$$

$$S_{\text{chord\_trans}} = \frac{\mathbf{f}_1 \cdot \mathbf{f}_2}{||\mathbf{f}_1||_2 \cdot ||\mathbf{f}_2||_2}, \tag{C.74}$$

where $\mathbf{M} = (\mathbf{T}_1^* > 0) \vee (\mathbf{T}_2^* > 0)$.

Tonnetz Similarity: Averaged tonal centroids across time:

$$\bar{\mathbf{x}}_1 = \frac{1}{N_{t,1}} \sum_{n=1}^{N_{t,1}} \mathbf{X}_1(\cdot, n), \quad \bar{\mathbf{x}}_2 = \frac{1}{N_{t,2}} \sum_{n=1}^{N_{t,2}} \mathbf{X}_2(\cdot, n), \tag{C.75}$$

$$S_{\text{tonnetz}} = \frac{\bar{\mathbf{x}}_1 \cdot \bar{\mathbf{x}}_2}{||\bar{\mathbf{x}}_1||_2 \cdot ||\bar{\mathbf{x}}_2||_2}. \tag{C.76}$$

Note that the computed harmony features introduce a Western bias as a result of both the utilized chord templates - i.e., major/minor triads - and the theoretical profiles that were used on key estimation.

C.4 Timbre Feature Analysis

Timbre analysis captures perceptual qualities distinguishing sounds of equal pitch, loudness, and duration. The framework preserves temporal complexity and statistical richness of spectral features through statistical distributions rather than temporal averaging.

C.4.1 Feature Extraction

Let $y[n]$ denote the discrete audio signal with sampling rate $f_s = 22050$ Hz and hop length $H = 512$ samples.

MFCC Features: 13 Mel-frequency cepstral coefficients for spectral envelope [59, 185]:

$$\mathbf{M} = \text{mfcc}(y, f_s, H, \text{n\_mfcc} = 13) \in \mathbb{R}^{13 \times N_t}. \tag{C.77}$$

Temporal dynamics through delta features:

$$\mathbf{M}_{\Delta} = \text{delta}(\mathbf{M}) \quad \text{(first-order differences)}, \tag{C.78}$$

$$\mathbf{M}_{\Delta\Delta} = \text{delta}(\mathbf{M}, \text{order} = 2) \quad \text{(second-order differences)}. \tag{C.79}$$

Spectral Shape Features: Multiple spectral characteristics [197]:

$$\mathbf{c}_s = \text{spectral\_centroid}(y, f_s, H) \quad \text{(brightness)}, \tag{C.80}$$

$$\mathbf{r}_s = \text{spectral\_rolloff}(y, f_s, H) \quad \text{(85\% energy point)}, \tag{C.81}$$

$$\mathbf{b}_s = \text{spectral\_bandwidth}(y, f_s, H) \quad \text{(frequency spread)}, \tag{C.82}$$

$$\mathbf{K}_s = \text{spectral\_contrast}(y, f_s, H) \in \mathbb{R}^{7 \times N_t} \quad \text{(peaks vs valleys)}. \tag{C.83}$$

Spectral Texture Features: Texture and energy characteristics:

$$\mathbf{f}_s = \text{spectral\_flatness}(y, H) \quad [198], \tag{C.84}$$

$$\mathbf{e}_{\text{rms}} = \text{rms}(y, H) \quad \text{(energy dynamics)}. \tag{C.85}$$

C.4.2 Statistical Feature Representation

Statistical feature vectors preserve temporal dynamics. For time series $\mathbf{s} = [s_1, s_2, \ldots, s_{N_t}]$:

$$\boldsymbol{\sigma}(\mathbf{s}) = [\mu_s, \sigma_s, \tilde{s}, q_{25}(s), q_{75}(s), \Delta_s, \min(s), \max(s)]^T, \tag{C.86}$$

where:

$$\mu_s = \frac{1}{N_t} \sum_{i=1}^{N_t} s_i \quad \text{(mean)}, \tag{C.87}$$

$$\sigma_s = \sqrt{\frac{1}{N_t} \sum_{i=1}^{N_t} (s_i - \mu_s)^2} \quad \text{(standard deviation)}, \tag{C.88}$$

$$\tilde{s} = \text{median}(\mathbf{s}) \quad \text{(median)}, \tag{C.89}$$

$$q_{25}(s), q_{75}(s) = \text{percentile}(\mathbf{s}, 25), \text{percentile}(\mathbf{s}, 75), \tag{C.90}$$

$$\Delta_s = \max(\mathbf{s}) - \min(\mathbf{s}) \quad \text{(range)}. \tag{C.91}$$

This 8-dimensional vector $\boldsymbol{\sigma}(\mathbf{s}) \in \mathbb{R}^8$ captures distribution characteristics while avoiding information loss.

C.4.3 Similarity Computation

Three weighted timbre components focusing more on the MFCCs and their dynamics in contrast to the spectral features that are better suited to monophonic signals:

$$S_{\text{timbre}} = 0.45 \cdot S_{\text{MFCC}} + 0.45 \cdot S_{\text{dynamics}} + 0.1 \cdot S_{\text{spectral}}. \tag{C.92}$$

MFCC Distribution Similarity: For each coefficient i:

$$\boldsymbol{\sigma}_i^{(1)} = \boldsymbol{\sigma}(\mathbf{M}_1(i,\cdot)), \quad \boldsymbol{\sigma}_i^{(2)} = \boldsymbol{\sigma}(\mathbf{M}_2(i,\cdot)), \tag{C.93}$$

$$s_i = \frac{\boldsymbol{\sigma}_i^{(1)} \cdot \boldsymbol{\sigma}_i^{(2)}}{||\boldsymbol{\sigma}_i^{(1)}||_2 \cdot ||\boldsymbol{\sigma}_i^{(2)}||_2}. \tag{C.94}$$

Exponential weighting for lower-order coefficients as they are more important:

$$\mathbf{w} = \exp(-0.1 \cdot [0, 1, 2, \ldots, 12]^T), \quad \mathbf{w}^* = \frac{\mathbf{w}}{||\mathbf{w}||_1}, \tag{C.95}$$

$$S_{\text{MFCC}} = \sum_{i=0}^{12} w_i^* \cdot s_i. \tag{C.96}$$

MFCC Dynamics Similarity: Temporal evolution patterns:

$$S_\Delta = \text{cosine\_sim}(\text{flatten}(\mathbf{M}_{\Delta,1}), \text{flatten}(\mathbf{M}_{\Delta,2})), \tag{C.97}$$

$$S_{\Delta\Delta} = \text{cosine\_sim}(\text{flatten}(\mathbf{M}_{\Delta\Delta,1}), \text{flatten}(\mathbf{M}_{\Delta\Delta,2})), \tag{C.98}$$

$$S_{\text{dynamics}} = \frac{1}{2}(S_\Delta + S_{\Delta\Delta}). \tag{C.99}$$

Spectral Similarity: Six spectral characteristics:

$$S_{\text{centroid}} = \text{cosine\_sim}(\boldsymbol{\sigma}(\mathbf{c}_{s,1}), \boldsymbol{\sigma}(\mathbf{c}_{s,2})), \tag{C.100}$$

$$S_{\text{rolloff}} = \text{cosine\_sim}(\boldsymbol{\sigma}(\mathbf{r}_{s,1}), \boldsymbol{\sigma}(\mathbf{r}_{s,2})), \tag{C.101}$$

$$S_{\text{bandwidth}} = \text{cosine\_sim}(\boldsymbol{\sigma}(\mathbf{b}_{s,1}), \boldsymbol{\sigma}(\mathbf{b}_{s,2})), \tag{C.102}$$

$$S_{\text{contrast}} = \text{cosine\_sim}(\text{flatten}(\mathbf{K}_{s,1}), \text{flatten}(\mathbf{K}_{s,2})), \tag{C.103}$$

$$S_{\text{flatness}} = \text{cosine\_sim}(\boldsymbol{\sigma}(\mathbf{f}_{s,1}), \boldsymbol{\sigma}(\mathbf{f}_{s,2})), \tag{C.104}$$

$$S_{\text{rms}} = \text{cosine\_sim}(\boldsymbol{\sigma}(\mathbf{e}_{\text{rms},1}), \boldsymbol{\sigma}(\mathbf{e}_{\text{rms},2})), \tag{C.105}$$

$$S_{\text{spectral}} = \frac{1}{6}(S_{\text{centroid}} + S_{\text{rolloff}} + S_{\text{bandwidth}} + S_{\text{contrast}} + S_{\text{flatness}} + S_{\text{rms}}). \tag{C.106}$$

Cosine similarity is computed as:

$$\text{cosine\_sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{||\mathbf{a}||_2 \cdot ||\mathbf{b}||_2}. \tag{C.107}$$

The cosine similarity range $[-1, 1]$ is normalized to $[0, 1]$ for consistency with other similarity measures:

$$\text{normalized\_sim} = \frac{1 + \text{cosine\_sim}(\mathbf{a}, \mathbf{b})}{2}. \tag{C.108}$$

Appendix D

Human Perception and Computational Methods Cross-cultural Similarities - Detailed Results

We utilize 9 diverse musical datasets - MagnaTagATune, MagnaTagATune, FMA-medium, corpusCOFLA, Arab-Andalusian, Lyra, Turkish-makam, Hindustani, Carnatic, and Jingju - and we use a subset of 52 music pieces per dataset, keeping a 20-second clip for each one.

The unique audio pairs that were annotated through the user study were 1,130, and thus we computed the dataset-level similarities for the human ratings and all the computational methods used in our research.

D.1 Human Perception

The dataset-level similarity matrices of the Figures D.1, D.2 and D.3 show the human-perceived relationships between musical traditions. The values represent the mean similarity ratings aggregated across all participant annotations for pairs between different datasets and darker colors indicate higher similarities.

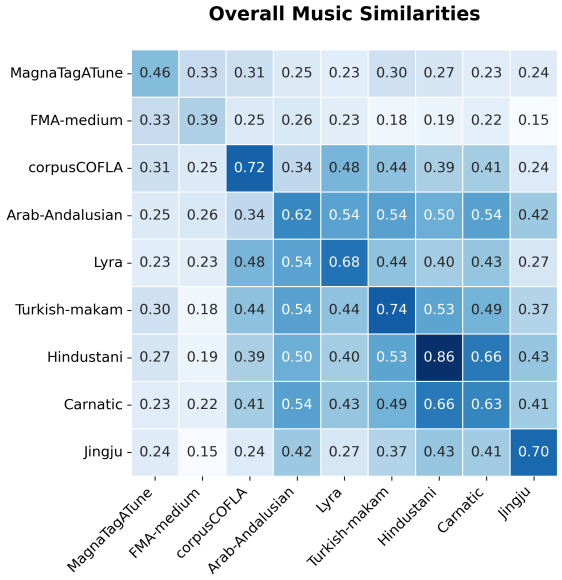

Figure D.1. Overall Music Similarity Matrix Across Datasets. Heat map visualization of human-perceived overall musical similarity ratings aggregated across all participant annotations, showing cross-cultural musical relationships as evaluated by human listeners.

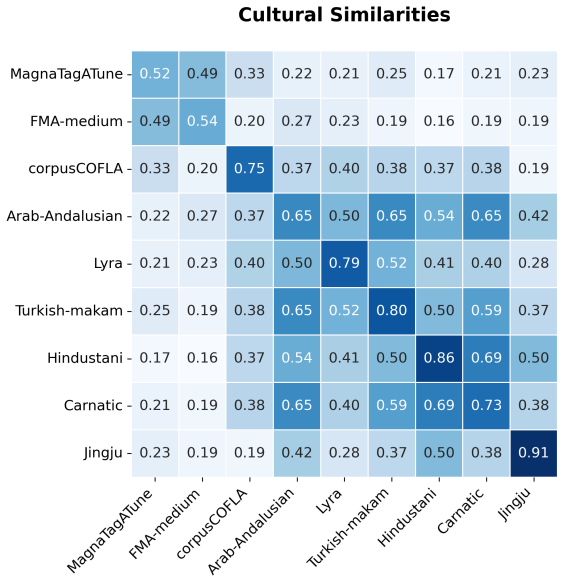

Figure D.2. Cultural Similarity Matrix Across Datasets. Heat map visualization of human-perceived cultural similarity ratings, revealing how participants assess cultural relationships and boundaries between different musical traditions represented in the study.

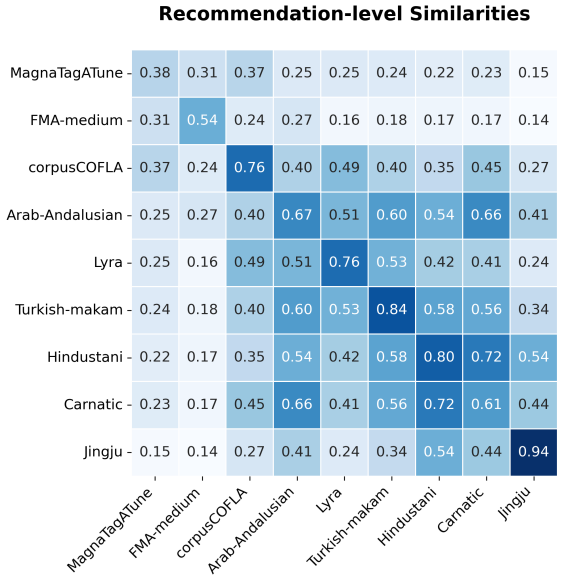

Figure D.3. Recommendation-Level Similarity Matrix Across Datasets. Heat map visualization of human-perceived recommendation-level similarity ratings, showing which musical traditions participants would consider suitable for personal recommendation contexts.

D.2 Signal Processing features

The Figures D.4, D.5, D.6, D.7 and D.8 show the dataset-level similarity for each Signal Processing feature dimension computed in our research, as well as when averaging all dimensions.

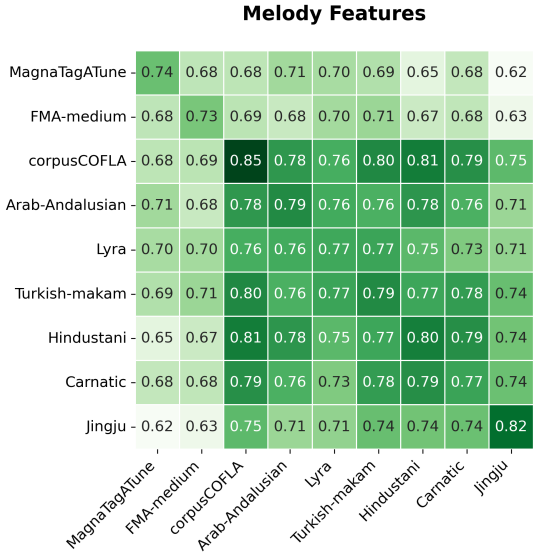

Figure D.4. Melody Similarity Matrix Across Datasets. Computational similarity matrix based on melody features extracted using pitch tracking and melodic analysis.

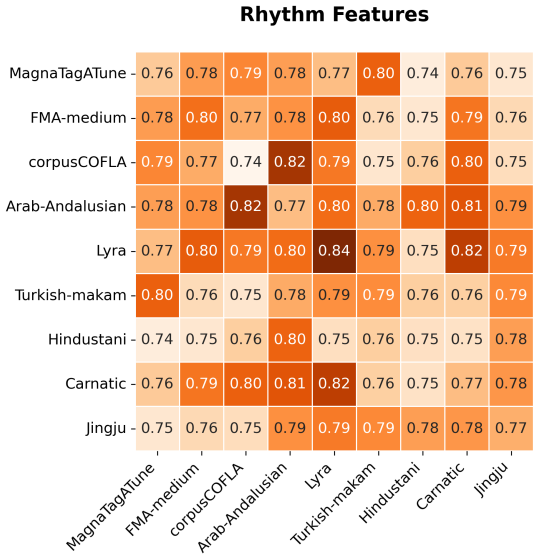

Figure D.5. Rhythm Similarity Matrix Across Datasets. Computational similarity matrix based on rhythm features extracted through onset detection and tempo analysis.

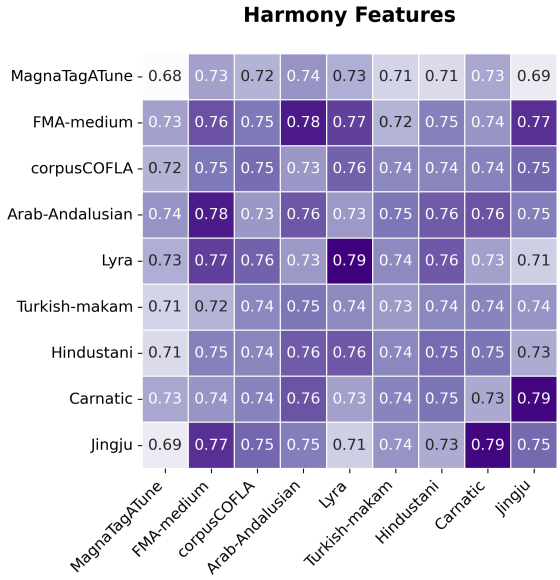

Figure D.6. Harmony Similarity Matrix Across Datasets. Computational similarity matrix based on harmony features including chromagrams and chord analysis.

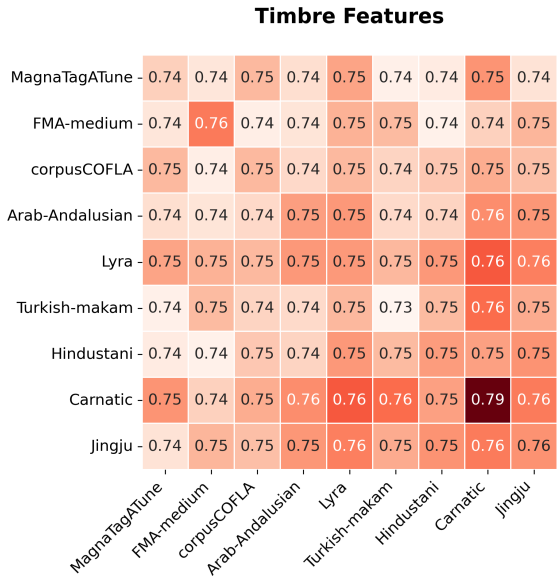

Figure D.7. Timbre Similarity Matrix Across Datasets. Computational similarity matrix based on timbral features including MFCCs and spectral characteristics.

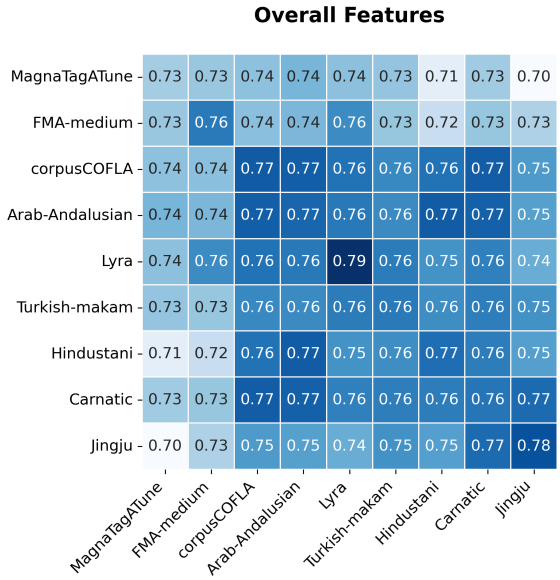

Figure D.8. Overall Similarity Matrix Averaging Signal Processing Features. Computational similarity matrix combining rhythm, melody, harmony, and timbre similarities through equal-weight averaging, providing a comprehensive signal processing perspective on cross-cultural musical relationships.

D.3 Foundation Models

Figures D.9, D.10, D.11, D.12, D.13, D.14 and D.15 show the dataset-level similarity matrices for each Foundation model utilized in our research.

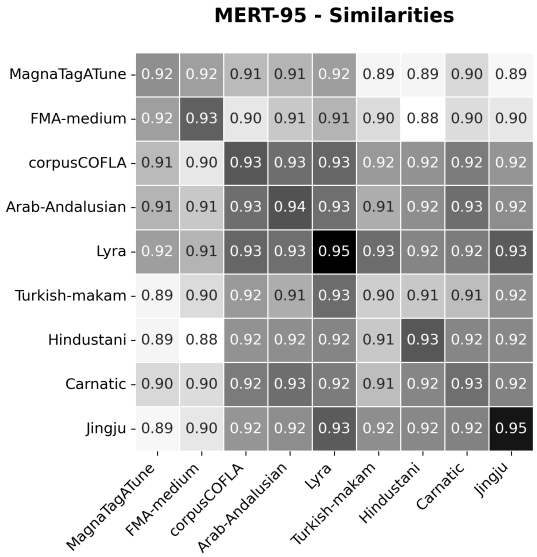

Figure D.9. MERT-95M Foundation Model Similarity Matrix. Computational similarity matrix derived from MERT-95M embeddings.

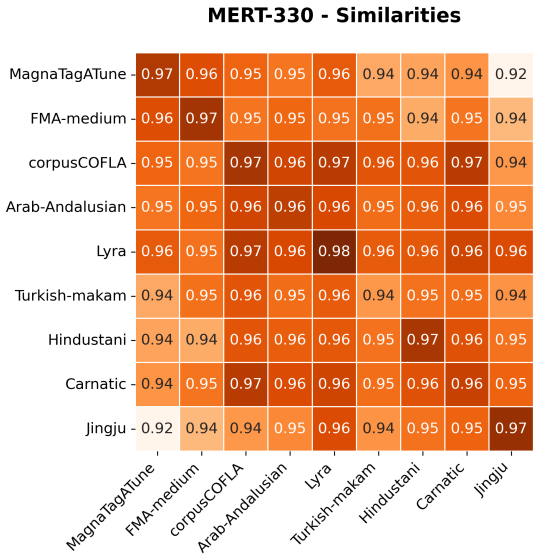

Figure D.10. MERT-330M Foundation Model Similarity Matrix. Computational similarity matrix derived from MERT-330M embeddings.

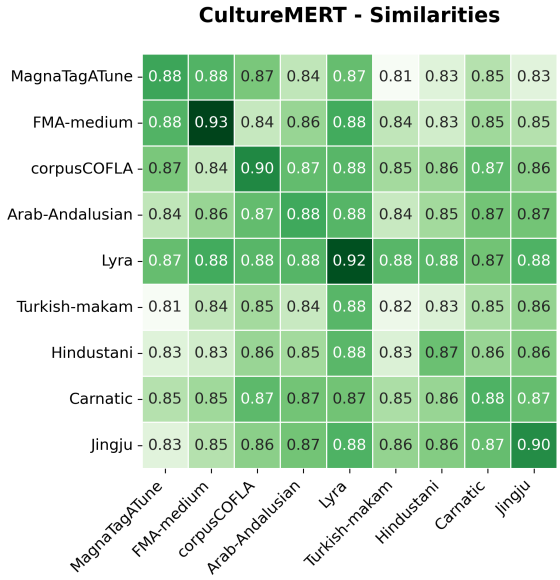

Figure D.11. CultureMERT Foundation Model Similarity Matrix. Computational similarity matrix derived from CultureMERT embeddings.

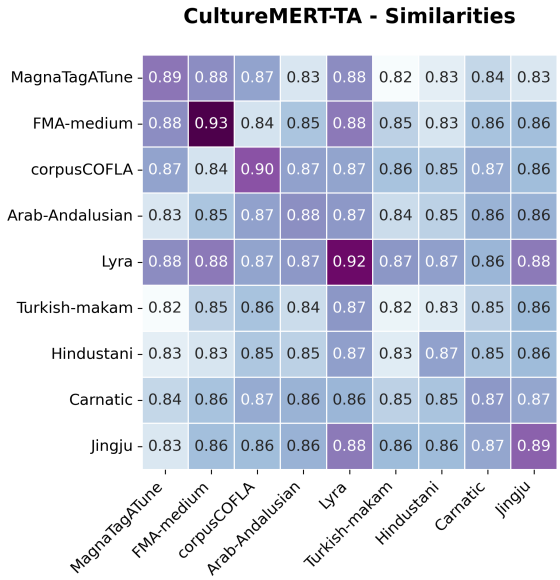

Figure D.12. CultureMERT-TA Foundation Model Similarity Matrix. Computational similarity matrix derived from CultureMERT-TA embeddings.

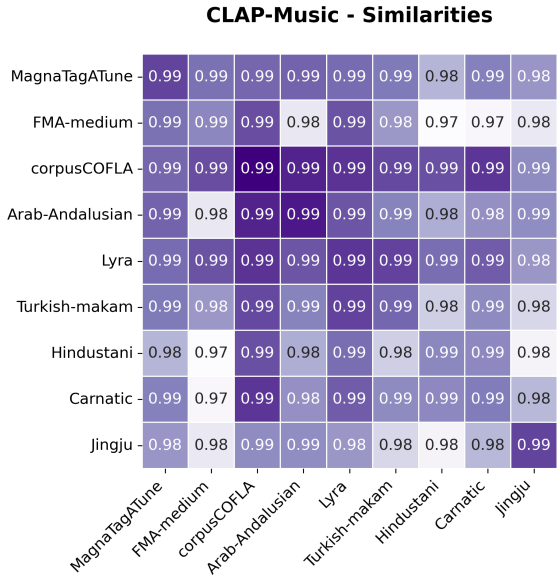

Figure D.13. CLAP-Music Foundation Model Similarity Matrix. Computational similarity matrix derived from CLAP-Music embeddings.

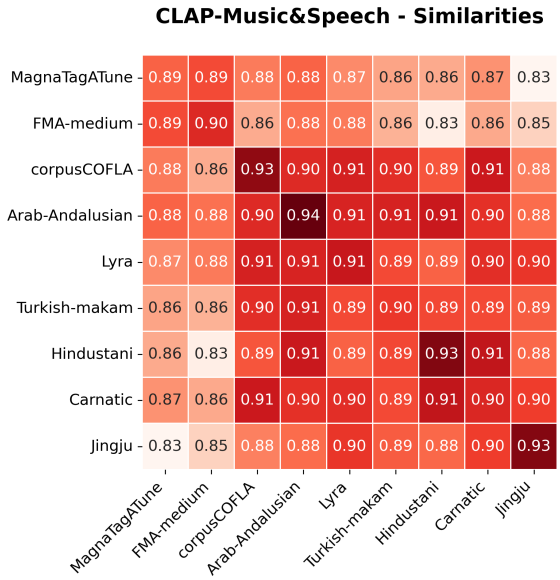

Figure D.14. CLAP-Music&Speech Foundation Model Similarity Matrix. Computational similarity matrix derived from CLAP-Music&Speech embeddings.

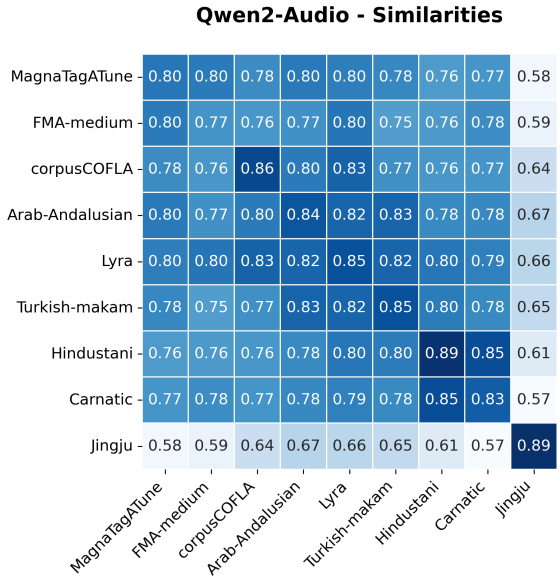

Figure D.15. Qwen2-Audio Foundation Model Similarity Matrix. Computational similarity matrix derived from Qwen2-Audio embeddings.

Bibliography

[1] Charles Darwin. *The descent of man, and selection in relation to sex*. Princeton University Press, 1871.

[2] Samuel A. Mehr, Manvir Singh, Dean Knox, Daniel M. Ketter, Daniel Pickens-Jones, S. Atwood, Christopher Lucas, Nori Jacoby, Alena A. Egner, Erin J. Hopkins, Rhea M. Howard, et al. "Universality and diversity in human song". In: *Science* 366 (2019).

[3] Aniruddh D Patel. *Music, language, and the brain*. Oxford university press, 2010.

[4] Jacques Attali. *Noise: The political economy of music*. Vol. 16. Manchester University Press, 1985.

[5] Sandra E. Trehub, Judith Becker, and Iain Morley. "Cross-cultural perspectives on music and musicality". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370 (2015).

[6] Joseph P Swain. "The concept of musical syntax". In: *The Musical Quarterly* 79.2 (1995), pp. 281–308.

[7] Sheng-Kuan Chung. "Digital storytelling in integrated arts education". In: *The International Journal of Arts Education* 4.1 (2006), pp. 33–63.

[8] Leonard B Meyer. *Emotion and meaning in music*. University of chicago Press, 1956.

[9] Elizabeth Hellmuth Margulis, Patrick C. M. Wong, Cara Turnbull, Benjamin M Kubit, and J. Devin McAuley. "Narratives imagined in response to instrumental music reveal culture-bounded intersubjectivity". In: *Proceedings of the National Academy of Sciences of the United States of America* 119 (2022).

[10] Patrick E. Savage, Steven Brown, Emi Sakai, and Thomas E. Currie. "Statistical universals reveal the structures and functions of human music". In: *Proceedings of the National Academy of Sciences* 112 (2015), pp. 8987–8992.

[11] Markus Schedl, Emilia Gómez, and Julián Urbano. "Music Information Retrieval: Recent Developments and Applications". In: *Found. Trends Inf. Retr.* 8.2-3 (2014), pp. 127–261.

[12] Meinard Müller. *Fundamentals of Music Processing - Audio, Analysis, Algorithms, Applications*. Springer, 2015.

[13] Emilia Gómez, Perfecto Herrera, and Francisco Gómez-Martin. "Computational Ethnomusicology: perspectives and challenges". In: *Journal of New Music Research* 42.2 (June 2013), pp. 111–112.

[14] Atharva Mehta, Shivam Chauhan, Amirbek Djanibekov, Atharva Kulkarni, Gus Xia, and Monojit Choudhury. "Music for All: Representational Bias and Cross-Cultural Adaptability of Music Generation Models". In: *NAACL (Findings)*. Association for Computational Linguistics, 2025, pp. 4569–4585.

[15] Xavier Serra, Michela Magas, Emmanouil Benetos, Magdalena Chudy, Simon Dixon, Arthur Flexer, Emilia Gómez, Fabien Gouyon, Perfecto Herrera, Sergi Jorda, et al. *Roadmap for music information research*. 2013.

[16] Justin London, Nori Jacoby, and Rainer Polak. "Theoretical and Practical Aspects of Cross-Cultural Corpus Studies: Two Case Studies from Mali". In: *The Oxford Handbook of Music and Corpus Studies*. Oxford University Press, 2022. ISBN: 9780190945442. DOI: 10.1093/oxfordhb/9780190945442.013.32.

[17] Thanos Fouloulis, Aggelos Pikrakis, and Emilios Cambouropoulos. "Traditional asymmetric rhythms: A refined model of meter induction based on asymmetric meter templates". In: *Proceedings of the third international workshop on folk music analysis*. 2013, pp. 28–32.

[18] M Kemal Karaosmanoğlu, Barış Bozkurt, Andre Holzapfel, and Nilgün Doğrusöz Dişiaçık. "A symbolic dataset of Turkish makam music phrases". In: *Proceedings of Fourth International Workshop on Folk Music Analysis (FMA2014)*. 2014, pp. 10–14.

[19] Ajay Srinivasamurthy, Gopala Krishna Koduri, Sankalp Gulati, Vignesh Ishwar, and Xavier Serra. "Corpora for Music Information Research in Indian Art Music". In: *ICMC*. Michigan Publishing, 2014.

[20] Baris Bozkurt, Ruhi Ayangil, and André Holzapfel. "Computational Analysis of Turkish Makam Music: Review of State-of-the-Art and Challenges". In: *Journal of New Music Research* 43 (2014), pp. 23–3.

[21] Gopala K. Koduri, Joan Serrà, and Xavier Serra. "Characterization of Intonation in Carnatic Music by Parametrizing Pitch Histograms". In: *International Society for Music Information Retrieval Conference*. 2012.

[22] Žanna Pärtlas. "Theoretical Approaches to Heterophony." In: *Res Musica* 8 (2016).

[23] Rafael Caro Repetto, Niccolò Pretto, Amin Chaachoo, Barış Bozkurt, and Xavier Serra. "An open corpus for the computational research of Arab-Andalusian music". In: *Proceedings of the 5th International Conference on Digital Libraries for Musicology*. 2018, pp. 78–86.

[24] Xia Gong, Yuxiang Zhu, Haidi Zhu, and Haoran Wei. "ChMusic: A Traditional Chinese Music Dataset for Evaluation of Instrument Recognition". In: *2021 4th International Conference on Big Data Technologies*. 2021, pp. 184–189.

[25] Pratik M. Joshi, Sebastin Santy, Amarjit Budhiraja, Kalika Bali, and Monojit Choudhury. "The State and Fate of Linguistic Diversity and Inclusion in the NLP World". In: *Annual Meeting of the Association for Computational Linguistics*. 2020.

[26] Emily M. Bender and Batya Friedman. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science". In: *Transactions of the Association for Computational Linguistics* 6 (2018), pp. 587–604.

[27] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021).

[28] Sebastian Ruder, Ivan Vulic, and Anders Søgaard. "A Survey of Cross-lingual Word Embedding Models". In: *J. Artif. Intell. Res.* 65 (2019), pp. 569–631.

[29] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. "Unsupervised Cross-lingual Representation Learning at Scale". In: *ACL*. Association for Computational Linguistics, 2020, pp. 8440–8451.

[30] Marc Leman and Pieter-Jan Maes. "The role of embodiment in the perception of music". In: *Empirical Musicology Review* 9.3-4 (2015), pp. 236–246.

[31] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[32] Maria Panteli, Emmanouil Benetos, and Simon Dixon. "A review of manual and computational approaches for the study of world music corpora". In: *Journal of New Music Research* 47.2 (2018), pp. 176–189.

[33] Xavier Serra. "Creating Research Corpora for the Computational Study of Music: the case of the CompMusic Project". In: *Semantic Audio*. Audio Engineering Society, 2014.

[34] Peter Van Kranenburg, Martine De Bruin, and Anja Volk. "Documenting a song culture: The Dutch Song Database as a resource for musicological research". In: *International Journal on Digital Libraries* 20.1 (2019), pp. 13–23.

[35] Nadine Kroher, José-Miguel Díaz-Báñez, Joaquin Mora, and Emilia Gómez. "Corpus COFLA: A research corpus for the computational study of flamenco music". In: *Journal on Computing and Cultural Heritage (JOCCH)* 9.2 (2016), pp. 1–21.

[36] Sebastian Rosenzweig, Frank Scherbaum, David Shugliashvili, Vlora Arifi-Müller, and Meinard Müller. "Erkomaishvili Dataset: A curated corpus of traditional Georgian vocal music for computational musicology". In: *Transactions of the International Society for Music Information Retrieval* 3.1 (2020).

[37] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. "CNN architectures for large-scale audio classification". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Mar. 2017, pp. 131–135.

[38] Jordi Pons and Xavier Serra. "musicnn: Pre-trained convolutional neural networks for music audio tagging". In: *arXiv preprint arXiv:1909.06654* (Sept. 2019).

[39] Yuan Gong, Yu-An Chung, and James R. Glass. "AST: Audio Spectrogram Transformer". In: *Interspeech*. ISCA, 2021, pp. 571–575.

[40] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger B. Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. "MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training". In: *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL: https://openreview.net/forum?id=w3YZ9MSlBu.

[41] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. "Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation". In: *ICASSP*. IEEE, 2023, pp. 1–5.

[42] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. "Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models". In: *CoRR* abs/2311.07919 (2023).

[43] Yinghao Ma, Anders Øland, Anton Ragni, Bleiz MacSen Del Sette, Charalampos Saitis, Chris Donahue, Chenghua Lin, Christos Plachouras, Emmanouil Benetos, Elio Quinton, et al. "Foundation Models for Music: A Survey". In: *CoRR* abs/2408.14340 (2024).

[44] André Holzapfel, Bob L. Sturm, and Mark Coeckelbergh. "Ethical Dimensions of Music Information Retrieval Technology". In: *Trans. Int. Soc. Music. Inf. Retr.* 1 (2018), pp. 44–55.

[45] Charilaos Papaioannou, Ioannis Valiantzas, Theodore Giannakopoulos, Maximos A. Kaliakatsos-Papakostas, and Alexandros Potamianos. "A Dataset for Greek Traditional and Folk Music: Lyra". In: *ISMIR*. 2022, pp. 377–383.

[46] Charilaos Papaioannou, Emmanouil Benetos, and Alexandros Potamianos. "From West to East: Who Can Understand the Music of the Others Better?" In: *ISMIR*. 2023, pp. 311–318.

[47] Charilaos Papaioannou, Emmanouil Benetos, and Alexandros Potamianos. "LC-Protonets: Multi-Label Few-Shot Learning for World Music Audio Tagging". In: *IEEE Open Journal of Signal Processing* 6 (2025), pp. 138–146.

[48] Charilaos Papaioannou, Emmanouil Benetos, and Alexandros Potamianos. "Universal Music Representations? Evaluating Foundation Models on World Music Corpora". In: *ISMIR*. (accepted for publication). 2025.

[49] Angelos-Nikolaos Kanatas, Charilaos Papaioannou, and Alexandros Potamianos. "CultureMERT: Continual Pre-Training for Cross-Cultural Music Representation Learning". In: *ISMIR*. (accepted for publication). 2025.

[50] Bruno Nettl. *The study of ethnomusicology: Thirty-three discussions*. University of Illinois Press, 2015.

[51] Marcus T Pearce and Geraint A Wiggins. "Auditory expectation: the information dynamics of music perception and cognition". In: *Topics in cognitive science* 4.4 (2012), pp. 625–652.

[52] Curtis Roads. *The computer music tutorial*. MIT press, 1996.

[53] Keunwoo Choi, György Fazekas, Kyunghyun Cho, and Mark Sandler. "A tutorial on deep learning for music information retrieval". In: *arXiv preprint arXiv:1709.04396* (2017).

[54] Judith C Brown. "Calculation of a constant Q spectral transform". In: *The Journal of the Acoustical Society of America* 89.1 (1991), pp. 425–434.

[55] Sebastian Ewert. "Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features". In: *Proc. ISMIR*. 2011.

[56] Matthias Mauch and Simon Dixon. "pYIN: A fundamental frequency estimator using probabilistic threshold distributions". In: *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE. 2014, pp. 659–663.

[57] Sebastian Böck, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer. "madmom: A New Python Audio and Music Signal Processing Library". In: *ACM Multimedia*. ACM, 2016, pp. 1174–1178.

[58] Emilia Gómez. "Tonal description of music audio signals". In: *Department of Information and Communication Technologies* (2006).

[59] Beth Logan et al. “Mel frequency cepstral coefficients for music modeling.” In: *ISMIR*. Vol. 270. 1. 2000, p. 11.

[60] Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. “librosa: Audio and Music Signal Analysis in Python”. In: *SciPy*. scipy.org, 2015, pp. 18–24.

[61] Joan Serrà, Meinard Müller, Peter Grosche, and Josep Lluís Arcos. “Unsupervised Detection of Music Boundaries by Time Series Structure Features”. In: *AAAI*. AAAI Press, 2012, pp. 1613–1619.

[62] Andre Holzapfel and Yannis Stylianou. “Scale Transform in Rhythmic Similarity of Music”. In: *IEEE Trans. Speech Audio Process.* 19.1 (2011), pp. 176–185.

[63] Georgi Bogomilov Dzhambazov, Ajay Srinivasamurthy, Sertan Sentürk, and Xavier Serra. “On the use of note onsets for improved lyrics-to-audio alignment in turkish makam music”. In: *Devaney J, Mandel MI, Turnbull D, Tzanetakis G, editors. Proceedings of the 17th International Society for Music Information Retrieval Conference*. 2016.

[64] Steven Davis and Paul Mermelstein. “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”. In: *IEEE transactions on acoustics, speech, and signal processing* 28.4 (1980), pp. 357–366.

[65] Michael A Casey, Remco Veltkamp, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney. “Content-based music information retrieval: Current directions and future challenges”. In: *Proceedings of the IEEE* 96.4 (2008), pp. 668–696.

[66] Keunwoo Choi, George Fazekas, and Mark Sandler. “Automatic tagging using deep convolutional neural networks”. In: *arXiv preprint arXiv:1606.00298* (June 2016).

[67] Sander Dieleman and Benjamin Schrauwen. “End-to-end learning for music audio”. In: *ICASSP*. IEEE, 2014, pp. 6964–6968.

[68] Jordi Pons, Oriol Nieto, Matthew Prockup, Erik M. Schmidt, Andreas F. Ehmann, and Xavier Serra. “End-to-end Learning for Music Audio Tagging at Scale”. In: *ISMIR*. 2018, pp. 637–644.

[69] Jongpil Lee, Jiyoung Park, Keunhyoung Luke Kim, and Juhan Nam. “SampleCNN: End-to-end deep convolutional neural networks using very small filters for music classification”. In: *Applied Sciences* 8.1 (2018), p. 150.

[70] Ashutosh Pandey and DeLiang Wang. “TCNN: Temporal Convolutional Neural Network for Real-time Speech Enhancement in the Time Domain”. In: *ICASSP*. IEEE, 2019, pp. 6875–6879.

[71] Sebastian Böck, Florian Krebs, and Gerhard Widmer. “Joint Beat and Downbeat Tracking with Recurrent Neural Networks”. In: *ISMIR*. 2016, pp. 255–261.

[72] Keunwoo Choi, György Fazekas, Mark B. Sandler, and Kyunghyun Cho. “Transfer Learning for Music Classification and Regression Tasks”. In: *ISMIR*. 2017, pp. 141–149.

[73] Janne Spijkervet and John Ashley Burgoyne. “Contrastive Learning of Musical Representations”. In: *ISMIR*. 2021, pp. 673–681.

[74] Aäron van den Oord, Sander Dieleman, and Benjamin Schrauwen. “Transfer Learning by Supervised Pre-training for Audio-based Music Classification”. In: *ISMIR*. 2014, pp. 29–34.

[75] Meinard Müller, Andreas Arzt, Stefan Balke, Matthias Dorfer, and Gerhard Widmer. "Cross-modal music retrieval and applications: An overview of key methodologies". In: *IEEE Signal Processing Magazine* 36.1 (2018), pp. 52–62.

[76] Ajay Srinivasamurthy, André Holzapfel, and Xavier Serra. "In Search of Automatic Rhythm Analysis Methods for Turkish and Indian Art Music". In: *Journal of New Music Research* 43 (2014), pp. 114–94. URL: https://api.semanticscholar.org/CorpusID:37521886.

[77] Bob L. Sturm. "Classification accuracy is not enough - On the evaluation of music genre recognition systems". In: *J. Intell. Inf. Syst.* 41.3 (2013), pp. 371–406.

[78] Maria Panteli. "Computational analysis of world music corpora". PhD thesis. Queen Mary University of London, UK, 2018.

[79] Emilia Gómez and Perfecto Herrera. "Comparative Analysis of Music Recordings from Western and Non-Western traditions by Automatic Tonal Feature Extraction". In: *Empirical Musicology Review* 3 (2008), pp. 140–156.

[80] Justin Salamon, Sankalp Gulati, and Xavier Serra. "A Multipitch Approach to Tonic Identification in Indian Classical Music". In: *ISMIR*. FEUP Edições, 2012, pp. 499–504.

[81] George Tzanetakis, Ajay Kapur, W Andrew Schloss, and Matthew Wright. "Computational ethnomusicology". In: *Journal of interdisciplinary music studies* 1.2 (2007), pp. 1–24.

[82] Dimos Makris, Katia Lida Kermanidis, and Ioannis Karydis. "The greek audio dataset". In: *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer. 2014, pp. 165–173.

[83] Dimos Makris, Ioannis Karydis, and Spyros Sioutas. "The greek music dataset". In: *Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS)*. 2015, pp. 1–7.

[84] Burak Uyar, Hasan Sercan Atli, Sertan Şentürk, Barış Bozkurt, and Xavier Serra. "A corpus for computational research of Turkish makam music". In: *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*. 2014, pp. 1–7.

[85] Sertan Şentürk. "Computational analysis of audio recordings and music scores for the description and discovery of Ottoman-Turkish Makam music". PhD thesis. Universitat Pompeu Fabra, 2016.

[86] Rafael Caro Repetto and Xavier Serra. "Creating a Corpus of Jingju (Beijing Opera) Music and Possibilities for Melodic Analysis." In: *ISMIR*. 2014, pp. 313–318.

[87] Sergio Oramas, Mohamed Sordo, and Xavier Serra. "Automatic creation of knowledge graphs from digital musical document libraries". In: *Conference in Interdisciplinary Musicology (CIM 2014)*. 2014.

[88] Alastair Porter, Mohamed Sordo, and Xavier Serra. "Dunya: A system for browsing audio music collections exploiting cultural context". In: *Proceedings of the 14th Int. Society for Music Information Retrieval Conf., Curitiba, Brazil*. 2013.

[89] Sinno Jialin Pan and Qiang Yang. "A Survey on Transfer Learning". In: *IEEE Transactions on Knowledge and Data Engineering* 10 (Oct. 2010), pp. 1345–1359.

[90] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. "How transferable are features in deep neural networks?" In: *arXiv preprint arXiv:1411.1792* (Nov. 2014).

[91] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. "Simultaneous Deep Transfer Across Domains and Tasks". In: *arXiv preprint arXiv:1510.02192* (Oct. 2015).

[92] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. "A Comprehensive Survey on Transfer Learning". In: *arXiv preprint arXiv:1911.02685* (June 2020).

[93] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. "Learning Transferable Features with Deep Adaptation Networks". In: *arXiv preprint arXiv:1502.02791* (May 2015).

[94] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. "Generalizing from a Few Examples: A Survey on Few-shot Learning". In: *ACM Comput. Surv.* 53.3 (2021), 63:1–63:34.

[95] Jake Snell, Kevin Swersky, and Richard S. Zemel. "Prototypical Networks for Few-shot Learning". In: *NIPS*. 2017, pp. 4077–4087.

[96] Tsendsuren Munkhdalai and Hong Yu. "Meta Networks". In: *ICML*. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 2554–2563.

[97] Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks". In: *ICML*. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 1126–1135.

[98] Yu Wang, Justin Salamon, Mark Cartwright, Nicholas J. Bryan, and Juan Pablo Bello. "Few-shot Drum Transcription in Polyphonic Music". In: *ISMIR*. 2020, pp. 117–124.

[99] Yu Wang, Daniel Stoller, Rachel M. Bittner, and Juan Pablo Bello. "Few-Shot Musical Source Separation". In: *ICASSP*. IEEE, 2022, pp. 121–125.

[100] Hugo Flores García, Aldo Aguilar, Ethan Manilow, and Bryan Pardo. "Leveraging Hierarchical Structures for Few-Shot Musical Instrument Recognition". In: *ISMIR*. 2021, pp. 220–228.

[101] Amit Alfassy, Leonid Karlinsky, Amit Aides, Joseph Shtok, Sivan Harary, Rogério Schmidt Feris, Raja Giryes, and Alexander M. Bronstein. "LaSO: Label-Set Operations Networks for Multi-Label Few-Shot Learning". In: *CVPR*. Computer Vision Foundation / IEEE, 2019, pp. 6548–6557.

[102] Christian Simon, Piotr Koniusz, and Mehrtash Harandi. "Meta-Learning for Multi-Label Few-Shot Classification". In: *WACV*. IEEE, 2022, pp. 346–355.

[103] Mengting Hu, Shiwan Zhao, Honglei Guo, Chao Xue, Hang Gao, Tiegang Gao, Renhong Cheng, and Zhong Su. "Multi-Label Few-Shot Learning for Aspect Category Detection". In: *ACL/IJCNLP (1)*. Association for Computational Linguistics, 2021, pp. 6330–6340.

[104] Jinhua Liang, Huy Phan, and Emmanouil Benetos. "Learning from Taxonomy: Multi-Label Few-Shot Classification for Everyday Sound Recognition". In: *ICASSP*. IEEE, 2024, pp. 771–775.

[105] Kai-Hsiang Cheng, Szu-Yu Chou, and Yi-Hsuan Yang. "Multi-label Few-shot Learning for Sound Event Recognition". In: *MMSP*. IEEE, 2019, pp. 1–5.

[106] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, et al. "On the Opportunities and Risks of Foundation Models". In: *CoRR* abs/2108.07258 (2021).

[107] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL-HLT (1)*. Association for Computational Linguistics, 2019, pp. 4171–4186.

[108] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, and other. "Language Models are Few-Shot Learners". In: *NeurIPS*. 2020.

[109] Rodrigo Castellon, Chris Donahue, and Percy Liang. "Codified audio language modeling learns useful representations for music information retrieval". In: *ISMIR*. 2021, pp. 88–96.

[110] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. "Jukebox: A Generative Model for Music". In: *CoRR* abs/2005.00341 (2020).

[111] Matthew C. McCallum, Filip Korzeniowski, Sergio Oramas, Fabien Gouyon, and Andreas F. Ehmann. "Supervised and Unsupervised Learning of Audio Representations for Music Understanding". In: *ISMIR*. 2022, pp. 256–263.

[112] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Chenghua Lin, Xingran Chen, Anton Ragni, Hanzhi Yin, Zhijie Hu, Haoyu He, Emmanouil Benetos, Norbert Gyenge, Ruibo Liu, and Jie Fu. "MAP-Music2Vec: A Simple and Effective Baseline for Self-Supervised Music Audio Representation Learning". In: *CoRR* abs/2212.02508 (2022).

[113] Minz Won, Yun-Ning Hung, and Duc Le. "A Foundation Model for Music Informatics". In: *ICASSP*. IEEE, 2024, pp. 1226–1230.

[114] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. "FMA: A Dataset for Music Analysis". In: *ISMIR*. 2017, pp. 316–323.

[115] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. "Qwen2-Audio Technical Report". In: *CoRR* abs/2407.10759 (2024).

[116] Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Le Zhuo, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Deng, Ningzhi Wang, Chenghua Lin, Emmanouil Benetos, Anton Ragni, et al. "MARBLE: Music Audio Representation Benchmark for Universal Evaluation". In: *NeurIPS*. 2023.

[117] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. "MIR_EVAL: A Transparent Implementation of Common MIR Metrics." In: *ISMIR*. Vol. 10. 2014, p. 2014.

[118] Christos Plachouras, Pablo Alonso-Jiménez, and Dmitry Bogdanov. "mir_ref: A representation evaluation framework for music information retrieval tasks". In: *arXiv preprint arXiv:2312.05994* (2023).

[119] Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Gregory Anthony, Eugene Belilovsky, Timothée Lesort, and Irina Rish. "Simple and Scalable Strategies to Continually Pre-train Large Language Models". In: *Trans. Mach. Learn. Res.* 2024 (2024). URL: https://openreview.net/forum?id=DimPeeCxKO.

[120] Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. "Editing models with task arithmetic". In: *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL: https://openreview.net/forum?id=6t0Kwf8-jrj.

[121] Tuomas Eerola, Topi Järvinen, Jukka Louhivuori, and Petri Toiviainen. "Statistical features and perceived similarity of folk melodies". In: *Music Perception* 18.3 (2001), pp. 275–296.

[122] Alexandra Lamont and Nicola Dibben. "Motivic structure and the perception of similarity". In: *Music Perception* 18.3 (2001), pp. 245–274.

[123] Adam Berenzweig, Beth Logan, Daniel PW Ellis, and Brian Whitman. "A large-scale evaluation of acoustic and subjective music-similarity measures". In: *Computer Music Journal* (2004), pp. 63–76.

[124] Haokun Tian, Stefan Lattner, and Charalampos Saitis. "Assessing the Alignment of Audio Representations with Timbre Similarity Ratings". In: *ISMIR*. (accepted for publication). 2025.

[125] Keunwoo Choi. "Deep neural networks for music tagging". PhD thesis. Queen Mary University of London, UK, 2018.

[126] Edith Law, Kris West, Michael Mandel, Mert Bay, and J Stephen Downie. "Evaluation of algorithms using games: The case of music tagging". In: *10th International Society for Music Information Retrieval Conference, ISMIR 2009*. 2009, pp. 387–392.

[127] Rafael Caro Repetto and Xavier Serra. "Creating a corpus of jingju (beijing opera) music and possibilities for melodic analysis". In: (2014).

[128] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *ICLR*. 2015.

[129] Minz Won, Andres Ferraro, Dmitry Bogdanov, and Xavier Serra. "Evaluation of CNN-based Automatic Music Tagging Models". In: *CoRR* abs/2006.00751 (2020).

[130] Michalis Papakostas and Theodoros Giannakopoulos. "Speech-music discrimination using deep visual feature extractors". In: *Expert Systems with Applications* 114 (2018), pp. 334–344.

[131] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. "Speech emotion recognition using CNN". In: *Proceedings of the 22nd ACM international conference on Multimedia*. 2014, pp. 801–804.

[132] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. "Convolutional recurrent neural networks for music classification". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017, pp. 2392–2396.

[133] Mohsin Ashraf, Guohua Geng, Xiaofeng Wang, Farooq Ahmad, and Fazeel Abid. "A globally regularized joint neural architecture for music classification". In: *IEEE Access* 8 (2020), pp. 220980–220989.

[134] Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. "Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks". In: *arXiv preprint arXiv:1703.06345* (Mar. 2017).

[135] Akhilesh Kumar Sharma, Gaurav Aggarwal, Sachit Bhardwaj, Prasun Chakrabarti, Tulika Chakrabarti, Jemal H. Abawajy, Siddhartha Bhattacharyya, et al. "Classification of Indian Classical Music With Time-Series Matching Deep Learning Approach". In: *IEEE Access* 9 (2021), pp. 102041–102052.

[136] Emir Demirel, Baris Bozkurt, and Xavier Serra. "Automatic makam recognition using chroma features". In: *8th International Workshop on Folk Music Analysis*. 2018, pp. 19–24.

[137] Kaustuv Kanti Ganguli, Sertan Şentürk, and Carlos Guedes. "Critiquing Task-versus Goal-oriented Approaches: A Case for Makam Recognition". In: *Proceedings of the 23rd Int. Society for Music Information Retrieval Conf., Bengaluru, India*. Dec. 2022.

[138] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. "Human-level concept learning through probabilistic program induction". In: *Science* 350.6266 (2015), pp. 1332–1338.

[139] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. "Siamese neural networks for one-shot image recognition". In: *ICML deep learning workshop*. Vol. 2. 1. 2015.

[140] Vidur Joshi, Matthew E. Peters, and Mark Hopkins. "Extending a Parser to Distant Domains Using a Few Dozen Partially Annotated Examples". In: *ACL (1)*. Association for Computational Linguistics, 2018, pp. 1190–1199.

[141] Lukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. "Learning to Remember Rare Events". In: *ICLR*. 2017.

[142] Brenden M. Lake, Chia-ying Lee, James R. Glass, and Joshua B. Tenenbaum. "One-shot learning of generative speech concepts". In: *CogSci*. 2014.

[143] Sercan Ömer Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. "Neural Voice Cloning with a Few Samples". In: *NeurIPS*. 2018, pp. 10040–10050.

[144] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis P. Vlahavas. "Random k-Labelsets for Multilabel Classification". In: *IEEE Trans. Knowl. Data Eng.* 23.7 (2011), pp. 1079–1089.

[145] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. "Matching Networks for One Shot Learning". In: *NIPS*. 2016, pp. 3630–3638.

[146] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. "Learning to Compare: Relation Network for Few-Shot Learning". In: *CVPR*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 1199–1208.

[147] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. "Adapting Visual Category Models to New Domains". In: *Computer Vision – ECCV 2010*. Springer, 2010, pp. 213–226.

[148] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. "librosa: Audio and music signal analysis in python". In: *Proceedings of the 14th python in science conference*. Vol. 8. 2015, pp. 18–25.

[149] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *ICLR*. 2015.

[150] Jesse Davis and Mark Goadrich. "The relationship between Precision-Recall and ROC curves". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 233–240.

[151] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *ICLR*. 2015.

[152] Thomas Lidy, Carlos Nascimento Silla Jr., Olmo Cornelis, Fabien Gouyon, Andreas Rauber, Celso A. A. Kaestner, and Alessandro L. Koerich. "On the suitability of state-of-the-art music information retrieval methods for analyzing, categorizing and accessing non-Western and ethnic music collections". In: *Signal Process.* 90.4 (2010), pp. 1032–1048. DOI: 10.1016/J.SIGPRO.2009.09.014. URL: https://doi.org/10.1016/j.sigpro.2009.09.014.

[153] Genís Plaja-Roglans, Thomas Nuttall, Lara Pearson, Xavier Serra, and Marius Miron. "Repertoire-Specific Vocal Pitch Data Generation for Improved Melodic Analysis of Carnatic Music". In: *Trans. Int. Soc. Music. Inf. Retr.* 6.1 (2023), pp. 13–26. DOI: 10.5334/TISMIR.137. URL: https://doi.org/10.5334/tismir.137.

[154] Gopala K. Koduri, Marius Miron, Joan Serrà, and Xavier Serra. "Computational Approaches for the Understanding of Melody in Carnatic Music". In: *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011*. Ed. by Anssi Klapuri and Colby Leider. University of Miami, 2011, pp. 263–268. URL: http://ismir2011.ismir.net/papers/PS2-16.pdf.

[155] Andres Ferraro, Gustavo Ferreira, Fernando Diaz, and Georgina Born. "Measuring Commonality in Recommendation of Cultural Content to Strengthen Cultural Citizenship". In: *Trans. Recomm. Syst.* 2.1 (2024), 10:1–10:32. DOI: 10.1145/3643138. URL: https://doi.org/10.1145/3643138.

[156] Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. "Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art". In: *CoRR* abs/2406.03930 (2024). DOI: 10.48550/ARXIV.2406.03930. arXiv: 2406.03930. URL: https://doi.org/10.48550/arXiv.2406.03930.

[157] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault. Association for Computational Linguistics, 2020, pp. 8342–8360. DOI: 10.18653/V1/2020.ACL-MAIN.740. URL: https://doi.org/10.18653/v1/2020.acl-main.740.

[158] Andrea Cossu, Antonio Carta, Lucia C. Passaro, Vincenzo Lomonaco, Tinne Tuytelaars, and Davide Bacciu. "Continual pre-training mitigates forgetting in language and vision". In: *Neural Networks* 179 (2024), p. 106492. DOI: 10.1016/J.NEUNET.2024.106492. URL: https://doi.org/10.1016/j.neunet.2024.106492.

[159] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. "HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection". In: *ICASSP*. IEEE, 2022, pp. 646–650.

[160] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". In: *ICCV*. IEEE, 2021, pp. 9992–10002.

[161] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. 2017.

[162] James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. "Overcoming catastrophic forgetting in neural networks". In: *CoRR* abs/1612.00796 (2016).

[163] Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. "Continual Pre-Training of Large Language Models: How to (re)warm your model?" In: *CoRR* abs/2308.04014 (2023).

[164] Ilya Loshchilov and Frank Hutter. "Decoupled Weight Decay Regularization". In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[165] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "LoRA: Low-Rank Adaptation of Large Language Models". In: *ICLR*. OpenReview.net, 2022.

[166] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis. "MuLan: A Joint Embedding of Music Audio and Natural Language". In: *ISMIR*. 2022, pp. 559–566.

[167] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. "High Fidelity Neural Audio Compression". In: *Trans. Mach. Learn. Res.* 2023 (2023).

[168] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units". In: *IEEE ACM Trans. Audio Speech Lang. Process.* 29 (2021), pp. 3451–3460. DOI: 10.1109/TASLP.2021.3122291. URL: https://doi.org/10.1109/TASLP.2021.3122291.

[169] Dichucheng Li, Yinghao Ma, Weixing Wei, Qiuqiang Kong, Yulun Wu, Mingjin Che, Fan Xia, Emmanouil Benetos, and Wei Li. "Mertech: Instrument Playing Technique Detection Using Self-Supervised Pretrained Model with Multi-Task Finetuning". In: *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*. IEEE, 2024, pp. 521–525. DOI: 10.1109/ICASSP48485.2024.10447445. URL: https://doi.org/10.1109/ICASSP48485.2024.10447445.

[170] Jupinder Parmar, Sanjeev Satheesh, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. "Reuse, Don't Retrain: A Recipe for Continued Pretraining of Language Models". In: *CoRR* abs/2407.07263 (2024). DOI: 10.48550/ARXIV.2407.07263. arXiv: 2407.07263. URL: https://doi.org/10.48550/arXiv.2407.07263.

[171] Yiduo Guo, Jie Fu, Huishuai Zhang, Dongyan Zhao, and Yikang Shen. "Efficient Continual Pre-training by Mitigating the Stability Gap". In: *CoRR* abs/2406.14833 (2024). DOI: 10.48550/ARXIV.2406.14833. arXiv: 2406.14833. URL: https://doi.org/10.48550/arXiv.2406.14833.

[172] Matthias De Lange, Gido M. van de Ven, and Tinne Tuytelaars. "Continual evaluation for lifelong learning: Identifying the stability gap". In: *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL: https://openreview.net/forum?id=Zy350cRstc6.

[173] Igor André Pegoraro Santana, Fabio Pinhelli, Juliano Donini, Leonardo Gabiato Catharin, Rafael Biazus Mangolin, Yandre Maldonado e Gomes da Costa, Valéria Delisandra Feltrim, and Marcos Aurélio Domingues. "Music4All: A New Music Database and Its Applications". In: *2020 International Conference on Systems, Signals and Image Processing, IWSSIP 2020, Niterói, Brazil, July 1-3, 2020*. IEEE, 2020, pp. 399–404. DOI: 10.1109/IWSSIP48289.2020.9145170. URL: https://doi.org/10.1109/IWSSIP48289.2020.9145170.

[174] German Ignacio Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. "Continual lifelong learning with neural networks: A review". In: *Neural Networks* 113 (2019), pp. 54–71. DOI: 10.1016/J.NEUNET.2019.01.012. URL: https://doi.org/10.1016/j.neunet.2019.01.012.

[175] Dongwan Kim and Bohyung Han. "On the Stability-Plasticity Dilemma of Class-Incremental Learning". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023, pp. 20196–20204. DOI: 10.1109/CVPR52729.2023.01934. URL: https://doi.org/10.1109/CVPR52729.2023.01934.

[176] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. "Patching open-vocabulary models by interpolating weights". In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. Ed. by Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh. 2022. URL: http://papers.nips.cc/paper%5C_files/paper/2022/hash/bc6cddcd5d325e1c0f826066c1ad0215-Abstract-Conference.html.

[177] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time". In: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 23965–23998. URL: https://proceedings.mlr.press/v162/wortsman22a.html.

[178] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. "On Layer Normalization in the Transformer Architecture". In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 10524–10533. URL: http://proceedings.mlr.press/v119/xiong20b.html.

[179] Dilorom Karomat. "12 maqam system and its similarity with Indian Raga's (according to the Indian Manuscripts)". In: *Indian Musicological Society. Journal of the Indian Musicological Society* 36 (2006), p. 62.

[180] Evangelia Gogoulou, Timothée Lesort, Magnus Boman, and Joakim Nivre. "Continual Learning Under Language Shift". In: *Text, Speech, and Dialogue - 27th International Conference, TSD 2024, Brno, Czech Republic, September 9-13, 2024, Proceedings, Part I*. Ed. by Elmar Nöth, Ales Horák, and Petr Sojka. Vol. 15048. Lecture Notes in Computer Science. Springer, 2024, pp. 71–84. DOI: 10.1007/978-3-031-70563-2_6. URL: https://doi.org/10.1007/978-3-031-70563-2%5C_6.

[181] Verena Blaschke, Masha Fedzechkina, and Maartje ter Hoeve. "Analyzing the Effect of Linguistic Similarity on Cross-Lingual Transfer: Tasks and Experimental Setups Matter". In: *CoRR* abs/2501.14491 (2025). DOI: 10.48550/ARXIV.2501.14491. arXiv: 2501.14491. URL: https://doi.org/10.48550/arXiv.2501.14491.

[182] Nay San, Georgios Paraskevopoulos, Aryaman Arora, Xiluo He, Prabhjot Kaur, Oliver Adams, and Dan Jurafsky. "Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens". In: *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP, SIGTYPE 2024, St. Julian's, Malta, March 22, 2024*. Ed. by Michael Hahn, Alexey Sorokin, Ritesh Kumar, Andreas Scherbakov, Yulia Otmakhova, Jinrui Yang, Oleg Serikov, Priya Rani, Edoardo M. Ponti,

Saliha Muradoglu, Rena Gao, Ryan Cotterell, and Ekaterina Vylomova. Association for Computational Linguistics, 2024, pp. 100–112. URL: https://aclanthology.org/2024.sigtyp-1.13.

[183] Anja Volk, W Bas de Haas, and Peter Van Kranenburg. "Towards modelling variation in music as foundation for similarity". In: *Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music*. School of Music Studies, Aristotle University of Thessaloniki. 2012, pp. 1085–1094.

[184] Peter Knees and Markus Schedl. "A survey of music similarity and recommendation from music context data". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 10.1 (2013), pp. 1–21.

[185] George Tzanetakis and Perry Cook. "Musical genre classification of audio signals". In: *IEEE Transactions on speech and audio processing* 10.5 (2002), pp. 293–302.

[186] Olivier Lartillot and Petri Toiviainen. "A Matlab toolbox for musical feature extraction from audio". In: *International conference on digital audio effects*. Vol. 237. Bordeaux. 2007, p. 244.

[187] Elizabeth S Nawrot. "The perception of emotional expression in music: Evidence from infants, children and adults". In: *Psychology of music* 31.1 (2003), pp. 75–92.

[188] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. "Enabling factorized piano music modeling and generation with the MAESTRO dataset". In: *arXiv preprint arXiv:1810.12247* (2018).

[189] Anja Volk and Peter Van Kranenburg. "Melodic similarity among folk songs: An annotation study on similarity-based categorization in music". In: *Musicae Scientiae* 16.3 (2012), pp. 317–339.

[190] Rainer Typke. "Music retrieval based on melodic similarity". In: ASCI. 2007.

[191] Juan Pablo Bello, Chris Duxbury, Mike E. Davies, and Mark B. Sandler. "On the use of phase and energy for musical onset detection in the complex domain". In: *IEEE Signal Process. Lett.* 11.6 (2004), pp. 553–556.

[192] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler. "A tutorial on onset detection in music signals". In: *IEEE Transactions on speech and audio processing* 13.5 (2005), pp. 1035–1047.

[193] Peter Grosche, Meinard Müller, and Frank Kurth. "Cyclic tempogram—a mid-level tempo representation for musicsignals". In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2010, pp. 5522–5525.

[194] Meinard Müller and Sebastian Ewert. "Chroma Toolbox: Matlab Implementations for Extracting Variants of Chroma-Based Audio Features". In: *ISMIR*. University of Miami, 2011, pp. 215–220.

[195] Carol L Krumhansl and Edward J Kessler. "Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys." In: *Psychological review* 89.4 (1982), p. 334.

[196] Christopher Harte, Mark Sandler, and Martin Gasser. "Detecting harmonic change in musical audio". In: *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*. 2006, pp. 21–26.

[197] Anssi Klapuri and Manuel Davy. "Signal processing methods for music transcription". In: (2007).

[198] Shlomo Dubnov. "Generalization of spectral flatness measure for non-gaussian linear processes". In: *IEEE Signal Processing Letters* 11.8 (2004), pp. 698–701.

[199] Christos Plachouras. "Beyond Benchmarks: A Toolkit for Music Audio Representation Evaluation". MA thesis. Universitat Pompeu Fabra, 2023.

[200] Pigi Kouki, Shobeir Fakhraei, James Foulds, Magdalini Eirinaki, and Lise Getoor. "Hyper: A flexible and extensible probabilistic framework for hybrid recommender systems". In: *Proceedings of the 9th ACM Conference on Recommender Systems*. 2015, pp. 99–106.

[201] Maurice G Kendall. "A new measure of rank correlation". In: *Biometrika* 30.1-2 (1938), pp. 81–93.

[202] C. Spearman. "The Proof and Measurement of Association Between Two Things". In: *The American Journal of Psychology* 15.1 (1904), pp. 72–101.

[203] Kalervo Järvelin and Jaana Kekäläinen. "Cumulated gain-based evaluation of IR techniques". In: *ACM Transactions on Information Systems (TOIS)* 20.4 (2002), pp. 422–446.

[204] Kira Radinsky and Nir Ailon. "Ranking from pairs and triplets: Information quality, evaluation methods and query complexity". In: *Proceedings of the fourth ACM international conference on Web search and data mining*. 2011, pp. 105–114.

[205] Sanford Weisberg. *Applied linear regression*. Vol. 528. John Wiley & Sons, 2005.

[206] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems* 30 (2017).

[207] Joseph B Kruskal. "Nonmetric multidimensional scaling: a numerical method". In: *Psychometrika* 29.2 (1964), pp. 115–129.

[208] Yuto Ozaki, Adam Tierney, Peter Q. Pfordresher, John M. McBride, Emmanouil Benetos, Polina Proutskova, Gakuto Chiba, Fang Liu, Nori Jacoby, Suzanne C. Purdy, et al. "Globally, songs and instrumental melodies are slower and higher and use more stable pitches than speech: A Registered Report". In: *Science Advances* 10.20 (2024). DOI: 10.1126/sciadv.adm9797.

[209] Juan Sebastián Gómez Cañón, Nicolás Felipe Gutiérrez Páez, Lorenzo Porcaro, Alastair Porter, Estefanía Cano, Perfecto Herrera-Boyer, Aggelos Gkiokas, Patricia Santos, Davinia Hernández-Leo, Casper Karreman, and Emilia Gómez. "TROMPA-MER: an open dataset for personalized music emotion recognition". In: *J. Intell. Inf. Syst.* 60.2 (2023), pp. 549–570.

[210] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. "Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions". In: *ACM Comput. Surv.* 56.10 (2024), p. 264.

[211] Shuo Liu, Adria Mallol-Ragolta, Emilia Parada-Cabaleiro, Kun Qian, Xin Jing, Alexander Kathan, Bin Hu, and Björn W. Schuller. "Audio self-supervised learning: A survey". In: *Patterns* 3.12 (2022), p. 100616.

[212] Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe. "Self-Supervised Speech Representation Learning: A Review". In: *IEEE J. Sel. Top. Signal Process.* 16.6 (2022), pp. 1179–1210.

[213] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: *NeurIPS*. 2020.

[214] Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. "DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models". In: *ACL (1)*. Association for Computational Linguistics, 2024, pp. 1280–1297.

[215] Changho Hwang, Wei Cui, Yifan Xiong, Ziyue Yang, Ze Liu, Han Hu, Zilong Wang, Rafael Salas, Jithin Jose, Prabhat Ram, HoYuen Chau, Peng Cheng, Fan Yang, Mao Yang, and Yongqiang Xiong. "Tutel: Adaptive Mixture-of-Experts at Scale". In: *MLSys*. mlsys.org, 2023.

[216] Joshua Patrick Gardner, Simon Durand, Daniel Stoller, and Rachel M. Bittner. "LLark: A Multimodal Instruction-Following Language Model for Music". In: *ICML*. OpenReview.net, 2024.

[217] Fan Liu, Tianshu Zhang, Wenwen Dai, Chuanyi Zhang, Wenwen Cai, Xiaocong Zhou, and Delong Chen. "Few-shot adaptation of multi-modal foundation models: a survey". In: *Artif. Intell. Rev.* 57.10 (2024), p. 268.

[218] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. "High-Fidelity Audio Compression with Improved RVQGAN". In: *NeurIPS*. 2023.

[219] Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. "CultureLLM: Incorporating Cultural Differences into Large Language Models". In: *NeurIPS*. 2024.

[220] Florin Girbacia. "An Analysis of Research Trends for Using Artificial Intelligence in Cultural Heritage". In: *Electronics* 13.18 (2024), p. 3738.

[221] Eva Zangerle, Martin Pichl, and Markus Schedl. "User Models for Culture-Aware Music Recommendation: Fusing Acoustic and Cultural Cues". In: *Trans. Int. Soc. Music. Inf. Retr.* 3.1 (2020), pp. 1–16.

[222] Mario Casillo, Francesco Colace, Dajana Conte, Marco Lombardi, Domenico Santaniello, and Carmine Valentino. "Context-aware recommender systems and cultural heritage: a survey". In: *J. Ambient Intell. Humaniz. Comput.* 14.4 (2023), pp. 3109–3127.

List of Abbreviations

AI	Artificial Intelligence
AP	Average Precision
AST	Audio Spectrogram Transformer
BERT	Bidirectional Encoder Representations from Transformers
CENS	Chroma Energy Normalized Statistics
CLAP	Contrastive Language-Audio Pretraining
CNN	Convolutional Neural Network
CPT	Continual Pre-Training
CQT	Constant-Q Transform
FFT	Fast Fourier Transform
FMA	Free Music Archive
FSL	Few-Shot Learning
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
LC-Protonets	Label-Combination Prototypical Networks
LightGBM	Light Gradient-Boosting Machine
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MDS	Multidimensional Scaling
MERT	Music undERstanding model with large-scale self-supervised Training
MFCCs	Mel-Frequency Cepstral Coefficients
MIDI	Musical Instrument Digital Interface
MIR	Music Information Retrieval
MLFSL	Multi-Label Few-Shot Learning
MLM	Masked Language Modeling
MLP	Multi-Layer Perceptron
ML-PNs	Multi-Label Prototypical Networks
MTAT	MagnaTagATune
NDCG	Normalized Discounted Cumulative Gain
NLP	Natural Language Processing
PR-AUC	Precision-Recall - Area Under Curve
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
ROC-AUC	Receiver Operating Characteristic - Area Under Curve
RQ	Research Question
SFT	Supervised Fine-Tuning
SGD	Stochastic Gradient Descent
SOTA	State-of-the-art

STFT	Short-Time Fourier Transform
TA	Task Arithmetic
t-SNE	t-distributed Stochastic Neighbor Embedding
VGG	Visual Geometry Group