National Technical University of Athens
School of Electrical and Computer Engineering
Division of Signals, Control and Robotics
Speech and Language Processing group

Self-supervised Music Audio Representation Learning and Domain Adaptation Across Diverse Music Datasets

Diploma Thesis

of

ANGELOS-NIKOLAOS KANATAS

Supervisor: Alexandros Potamianos
Associate Professor, NTUA

Athens, June 2025

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Signals, Control and Robotics
Speech and Language Processing group

Self-supervised Music Audio Representation Learning and Domain Adaptation Across Diverse Music Datasets

Diploma Thesis

of

ANGELOS-NIKOLAOS KANATAS

Supervisor: Alexandros Potamianos
Associate Professor, NTUA

Approved by the examination committee on 1 July 2025.

(Signature)	*(Signature)*	*(Signature)*
.....	.....	.....
Alexandros Potamianos	Athanasios Rontogiannis	Costas Tzafestas
Associate Professor, NTUA	Associate Professor, NTUA	Associate Professor, NTUA

Athens, June 2025

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Signals, Control and Robotics
Speech and Language Processing group

(Signature)

.....
Angelos-Nikolaos Kanatas

Electrical & Computer Engineering Graduate, NTUA

Περίληψη

Αυτή η διπλωματική εργασία επικεντρώνεται στη μάθηση αναπαραστάσεων μουσικής μέσω αυτο-επιβλεπόμενης μάθησης και στην προσαρμογή υπολογιστικών μοντέλων σε ποικίλες μουσικές παραδόσεις. Πρόσφατες εξελίξεις στα foundation μοντέλα για μουσική έχουν βελτιώσει σημαντικά τη μάθηση αναπαραστάσεων ήχου και τα έχουν φέρει στο επίκεντρο της ανάκτησης μουσικής πληροφορίας (music information retrieval – MIR). Ωστόσο, η αποτελεσματικότητά τους παραμένει περιορισμένη για μη Δυτικές μουσικές παραδόσεις, καθώς έχουν εκπαιδευτεί κυρίως σε Δυτικά είδη μουσικής. Στη μελέτη αυτή, προτείνουμε το `CultureMERT-95M`, ένα πολυπολιτισμικά προσαρμοσμένο μοντέλο που στοχεύει στη βελτίωση μάθησης αναπαραστάσεων για ποικίλες υποεκπροσωπούμενες μουσικές κουλτούρες. Για τον σκοπό αυτό, εφαρμόζουμε μια μέθοδο συνεχούς προ-εκπαίδευσης (continual pre-training - CPT) δύο σταδίων, η οποία ενσωματώνει επαναθέρμανση και εκ νέου μείωση του ρυθμού μάθησης, επιτρέποντας σταθερή προσαρμογή με περιορισμένους υπολογιστικούς πόρους. Η συνεχής προ-εκπαίδευση του `MERT-95M` σε πολυπολιτισμικό σύνολο δεδομένων 650 ωρών, που περιλαμβάνει Ελληνικές, Τουρκικές και Ινδικές μουσικές παραδόσεις, οδηγεί σε μέση βελτίωση 4.43% στη μετρική ROC-AUC σε διάφορες εργασίες αυτόματης ταξινόμησης μουσικής (music auto-tagging tasks) μη Δυτικών παραδόσεων, ξεπερνώντας προηγούμενες μεθόδους, με αμελητέα απώλεια απόδοσης σε Δυτικά benchmarks. Επιπλέον, διερευνούμε την τεχνική task arithmetic, μια εναλλακτική προσέγγιση που συγχωνεύει εξειδικευμένα μοντέλα ανά παράδοση στον χώρο των βαρών, παρουσιάζοντας συγκρίσιμη απόδοση στα μη Δυτικά σύνολα δεδομένων, χωρίς επιδείνωση στα Δυτικά. Τέλος, αναλύουμε τη διαπολιτισμική μεταφερσιμότητα (cross-cultural transferability) μεταξύ μοντέλων που έχουν προσαρμοστεί σε επιμέρους παραδόσεις, δείχνοντας ότι διαφέρουν ως προς τη δυνατότητα μεταφοράς τους σε άλλες μουσικές κουλτούρες, ένα εύρημα που συσχετίζεται επίσης με την ομοιότητα μεταξύ των δεδομένων που χρησιμοποιούμε, με βάση μετρικές ομοιότητας σε επίπεδο ακουστικών tokens. Παρατηρούμε ότι η συνεχής προ-εκπαίδευση σε σύνολο δεδομένων από διαφορετικές μη Δυτικές παραδόσεις οδηγεί στην καλύτερη συνολική απόδοση, ενισχύοντας τη διαπολιτισμική γενίκευση του μοντέλου. Η μελέτη αυτή συμβάλλει στην ανάπτυξη πιο πολιτισμικά ευαισθητοποιημένων υπολογιστικών μοντέλων μουσικής, ικανών να κατανοούν υποεκπροσωπούμενες μουσικές παραδόσεις.

Λέξεις Κλειδιά

Ανάκτηση Μουσικής Πληροφορίας, Μάθηση Αναπαραστάσεων Μουσικής, Υπολογιστική Εθνομουσικολογία, Αυτο-επιβλεπόμενη Μάθηση, Συνεχής Προ-εκπαίδευση, Προσαρμογή Πεδίου, Διαπολιτισμική Προσαρμογή, Μεταφορά Μάθησης, Συγχώνευση Μοντέλων, Βαθιά Μάθηση, Μοντελοποίηση Μη Δυτικής Μουσικής, Αυτόματη Ταξινόμηση

Abstract

This thesis focuses on self-supervised music audio representation learning and cross-cultural adaptation of music foundation models to diverse musical traditions. Recent advances in music foundation models have improved audio representation learning and have brought them to the forefront of music information retrieval (MIR). However, their effectiveness across diverse musical traditions remains limited, as they are primarily trained on Western-centric data, overlooking the diversity of global musical cultures. To address this, we introduce `CultureMERT-95M`, a multi-culturally adapted foundation model developed to enhance cross-cultural music representation learning and understanding. To achieve this, we propose a two-stage continual pre-training (CPT) strategy that integrates learning rate re-warming and re-decaying, enabling stable adaptation even with limited computational resources. Continually pre-training `MERT-95M` on a 650-hour multi-cultural data mix, comprising Greek, Turkish, and Indian music traditions, results in an average improvement of 4.9% in ROC-AUC and AP across diverse non-Western music auto-tagging tasks, surpassing prior state-of-the-art, with minimal forgetting on Western-centric benchmarks. We further investigate task arithmetic, an alternative approach to multi-cultural adaptation that merges culturally specialized models in the weight space. Task arithmetic performs on par with our multi-culturally trained model on non-Western auto-tagging tasks and shows no regression on Western datasets. Finally, we analyze cross-cultural transferability between single-culture adapted models (via CPT), showing that musical traditions differ in how well they transfer to others, a pattern that correlates with acoustic token-level similarity among cultures, using as metrics the cosine distance and Jensen-Shannon divergence computed over EnCodec-extracted token distributions. Our findings demonstrate that exposure to culturally diverse data through multi-cultural CPT enhances cross-cultural generalization and leads to improved overall performance. This study contributes to the development of more culturally aware foundation models for music that generalize across diverse underrepresented musical traditions and enable world music understanding.

Keywords

Music Information Retrieval, Music Representation Learning, Computational Ethnomusicology, Self-supervised Learning, Continual Pre-Training, Domain Adaptation, Cross-Cultural Adaptation, Transfer Learning, Cross-Cultural Transfer, Model Merging, Task Arithmetic, Deep Learning, Foundation Models, Non-Western Music Modeling, Automatic classification

to my grandmother Angeliki, in loving memory.

Ευχαριστίες

Με αυτή τη διπλωματική εργασία ολοκληρώνεται ένα μεγάλο ταξίδι, γεμάτο ώρες μελέτης, προσπάθειας, θυσιών, επιτυχιών, αλλά και δυσκολιών. Ιδιαίτερα, ο τελευταίος χρόνος υπήρξε αρκετά στρεσογόνος αλλά και καθοριστικός για εμένα, και είμαι πολύ χαρούμενος και ευχαριστημένος για την πορεία και την έκβασή του. Θα ήθελα να εκφράσω την ειλικρινή μου ευγνωμοσύνη στους γονείς μου, Μόσχα και Γιώργο, καθώς και σε όλη την οικογένειά μου, για τη διαρκή στήριξη και αγάπη τους όλα αυτά τα χρόνια. Ήταν πάντα δίπλα μου, πρόθυμοι να με ακούσουν, να με στηρίξουν και να με ενθαρρύνουν, ακόμη κι όταν οι ανησυχίες και οι συζητήσεις γύρω από τη σχολή και τη διπλωματική εργασία τους φαίνονταν ακατανόητες. Χωρίς τη σταθερή τους παρουσία, τίποτα από όλα αυτά δεν θα ήταν δυνατό. Ευχαριστώ θερμά τους φίλους μου και ιδιαίτερα την Μαίρη, για την υπομονή, την κατανόηση και την αμέριστη ψυχολογική υποστήριξή τους καθ' όλη τη διάρκεια αυτής της απαιτητικής περιόδου, που με βοήθησαν να διατηρήσω την ισορροπία μου και να κάνουν την καθημερινότητά μου πιο υποφερτή και αισιόδοξη. Χωρίς αυτούς, αυτό το ταξίδι δεν θα είχε την ίδια σημασία, ούτε το ίδιο νόημα. Θα ήθελα επίσης να ευχαριστήσω ιδιαιτέρως τον υποψήφιο διδάκτορα Χάρη Παπαϊωάννου για την άριστη συνεργασία, την πολύτιμη καθοδήγηση, τις εποικοδομητικές συμβουλές και τη συνεχή του υποστήριξη και υπομονή, που με βοήθησαν να εξελιχθώ σε έναν πιο προσεκτικό, ώριμο και ολοκληρωμένο ερευνητή. Ένα μεγάλο ευχαριστώ οφείλω και στους συμφοιτητές μου, για την ανταλλαγή γνώσεων, τις συζητήσεις και την αμοιβαία στήριξη που υπήρξαν καθοριστικές. Τέλος, εκφράζω τις θερμές μου ευχαριστίες στον επιβλέποντα καθηγητή μου, Αλέξανδρο Ποταμιάνο, για τις γόνιμες ιδέες, την αδιάκοπη επίβλεψη και τη συνεχή και ουσιαστική καθοδήγησή του καθ' όλη τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας. Αφιερώνω αυτή τη διπλωματική εργασία στη μνήμη της γιαγιάς μου, Αγγελικής, για την αμέριστη αγάπη και συμπαράσταση που μου πρόσφερε καθ' όλη τη διάρκεια των παιδικών μου χρόνων.

Αθήνα, Ιούνιος 2025

Άγγελος-Νικόλαος Κανατάς

Table of Contents

List of Figures

List of Tables

Κεφάλαιο 0

Εκτεταμένη Ελληνική Περίληψη

0.1 Εισαγωγή

0.1.1 Κίνητρο

Η μουσική αποτελεί θεμελιώδες στοιχείο του ανθρώπινου πολιτισμού, παρούσα καθολικά σε όλες τις κοινωνίες, εκφραζόμενη μέσα από ποικίλες μορφές μοναδικές σε κάθε παράδοση [8, 9, 10]. Οι ρόλοι της περιλαμβάνουν τη ρύθμιση των συναισθημάτων, την επικοινωνία και τον κοινωνικό δεσμό· παίζει ρόλο στην τέχνη, την ψυχαγωγία, τη λατρεία και τη διαφήμιση, και αποτελεί σημαντικό κλάδο της παγκόσμιας οικονομίας. Ο διττός αυτός ρόλος, ως πολιτισμικό αντικείμενο και ως οικονομικός παράγοντας, προσφέρει ευκαιρίες για όφελος της κοινωνίας, ενώ ταυτόχρονα θέτει μοναδικές τεχνικές προκλήσεις όταν συνδυάζεται με την τεχνητή νοημοσύνη (artificial intelligence – AI) [11]. Πέραν των πρακτικών εφαρμογών, η κατανόηση της σημασιολογίας της μουσικής μέσω βαθιάς μάθησης (deep learning – DL), με ιδιαίτερη έμφαση σε ερμηνεύσιμες προσεγγίσεις (explainable AI – XAI), μπορεί επίσης να συνεισφέρει σε θεωρητικά ευρήματα σε τομείς όπως η εθνομουσικολογία και η μουσική ανθρωπολογία, η θεωρία της μουσικής και η γνωσιακή μουσικολογία. Επιπλέον, παρότι η μουσική συχνά περιγράφεται ως «παγκόσμια γλώσσα», αυτή η αντίληψη παραμένει αντικείμενο συζήτησης μεταξύ των μελετητών: ορισμένα χαρακτηριστικά φαίνεται να υπερβαίνουν τα πολιτισμικά όρια [12], ωστόσο οι μουσικές παραδόσεις έχουν εξελιχθεί με διακριτά χαρακτηριστικά και πολιτισμικά θεμελιωμένη σημασιολογία [13, 14]. Αυτή η αλληλεπίδραση μεταξύ καθολικότητας και πολιτισμικής ιδιαιτερότητας αποτελεί μια σύνθετη πρόκληση, την οποία ο κλάδος της υπολογιστικής μουσικολογίας και οι σύγχρονες προσεγγίσεις τεχνητής νοημοσύνης μπορούν να διερευνήσουν μέσα από μια νέα οπτική [15].

Η ανάκτηση μουσικής πληροφορίας (music information retrieval – MIR) αναφέρεται στον ερευνητικό τομέα που επικεντρώνεται στην εξαγωγή και ανάλυση πληροφορίας από μουσικά δεδομένα [16, 11]. Οι υπολογιστικές μέθοδοι σε αυτό το πεδίο συνδυάζουν τεχνικές επεξεργασίας σήματος για την εξαγωγή χαρακτηριστικών από ηχητικά σήματα, με αλγορίθμους μηχανικής μάθησης (machine learning – ML) για την εκτέλεση εργασιών μουσικής κατανόησης (music understanding tasks), όπως η ταξινόμηση είδους (genre classification), η ανίχνευση ρυθμού (beat tracking), ο εντοπισμός τονικότητας (key detection), ο διαχωρισμός πηγών (source separation) και η αυτόματη ετικετοποίηση (automatic tagging), μεταξύ άλλων. Σε αντίθεση με την ομιλία και τη γλώσσα, η μουσική είναι τυπικά πολυφωνική, συχνά αποτελο-

ύμενη από πολλαπλές ταυτόχρονες «φωνές» και στρώματα οργάνων. Επιπλέον, η «σημασία» της συνήθως δεν βασίζεται σε άμεσες αναφορές σε αντικείμενα του πραγματικού κόσμου ή συγκεκριμένα γεγονότα, αλλά είναι αφηρημένη και διαμορφώνεται από το πολιτισμικό πλαίσιο. Ως εκ τούτου, η κατανόηση της μουσικής παρουσιάζει ιδιαίτερες προκλήσεις, καθώς ενσωματώνει περίπλοκες, διαπλεκόμενες έννοιες που σχετίζονται με τον άνθρωπο, όπως τα συναισθήματα, οι εμπειρίες, η έκφραση, η πολιτισμική ταυτότητα, το κοινωνικό και ιστορικό πλαίσιο, η επικοινωνία και η δημιουργικότητα. Επιπλέον, η μουσική έχει συνήθως μεγαλύτερη χρονική διάρκεια και υψηλότερο ρυθμό δειγματοληψίας (sample rate) από την ομιλία ή τον γενικό ήχο, γεγονός που καθιστά υπολογιστικά απαιτητική τη μοντελοποίηση ολόκληρων μουσικών κομματιών. Ένα βασικό εμπόδιο είναι ότι η απευθείας μοντελοποίηση του ήχου (raw audio) εισάγει εξαρτήσεις μεγάλης εμβέλειας (long-range dependencies), καθιστώντας δύσκολη τη μάθηση των σημασιολογικών ιδιοτήτων της μουσικής σε διαφορετικά επίπεδα.

Ο όρος «foundation model» (FM) εισήχθη για να περιγράψει οποιαδήποτε προ-εκπαιδευμένη και ευέλικτη αρχιτεκτονική μηχανικής μάθησης, η οποία, αντί να βελτιστοποιείται για έναν συγκεκριμένο σκοπό, λειτουργεί ως κεντρικό framework από το οποίο μπορούν να προκύψουν πολλαπλά εξειδικευμένα μοντέλα για ένα ευρύ φάσμα εργασιών (downstream tasks) [17]. Η ανάδυση των foundation models έχει τροφοδοτηθεί από τις εξελίξεις στη βαθιά μάθηση, συμπεριλαμβανομένων αρχιτεκτονικών καινοτομιών όπως ο Transformer [18], καθώς και από τις βελτιώσεις στο υπολογιστικό υλικό (hardware). Πρόσφατα, τα foundation models έχουν κάνει την εμφάνισή τους και στον τομέα της μουσικής [1, 19, 20, 11], προσφέροντας ισχυρές, γενικού σκοπού αναπαραστάσεις, μέσω της μάθησης από δεδομένα ήχου μεγάλης κλίμακας. Τα μοντέλα αυτά έχουν τη δυνατότητα να συλλαμβάνουν ευρεία μουσικά χαρακτηριστικά και έχουν επιδείξει state-of-the-art επιδόσεις σε πλήθος εργασιών κατανόησης μουσικής, μειώνοντας έτσι την ανάγκη για εξειδικευμένη εκπαίδευση ανά task. Αξιοποιώντας την αυτο-επιβλεπόμενη μάθηση (self-supervised learning – SSL) σε μη επισημασμένα μουσικά δεδομένα μεγάλης κλίμακας, τα foundation models αντιμετωπίζουν το πρόβλημα της έλλειψης δεδομένων, μειώνουν το κόστος επισημάνσεων και βελτιώνουν τη γενίκευση στην ανάκτηση μουσικής πληροφορίας [11].

Παρά την πρόοδο, τα περισσότερα υφιστάμενα foundation models για μουσική έχουν εκπαιδευτεί κυρίως σε σύνολα δεδομένων που προέρχονται από Δυτικές μουσικές κουλτούρες, γεγονός που περιορίζει την ικανότητά τους να αναπαριστούν ποικίλα μουσικά στυλ [21, 3]. Σημαντικό είναι επίσης ότι τα μοντέλα αυτά σπάνια αξιολογούνται με βάση την παγκόσμια μουσική ποικιλομορφία, αφήνοντας σε μεγάλο βαθμό ανεξερεύνητη τη γενικευσιμότητά τους σε διαφορετικές μουσικές παραδόσεις, ιδιαίτερα στις υποεκπροσωπούμενες. Πολλές από αυτές τις παραδόσεις, όπως η Τουρκική, η Ινδική και η Ελληνική παραδοσιακή μουσική, χαρακτηρίζονται από μοναδικές μελωδικές δομές, τροπικά ή τονικά συστήματα, και ιδιαίτερα ρυθμικά μοτίβα, τα οποία δεν αποτυπώνονται επαρκώς από τα υπάρχοντα μοντέλα [22, 23, 24]. Η αδυναμία μοντελοποίησης τέτοιων πολιτισμικά ειδικών στυλιστικών χαρακτηριστικών όχι μόνο περιορίζει την εφαρμοσιμότητα των music foundation models, για παράδειγμα, σε συστήματα σύστασης περιεχομένου (recommendation systems) που προσαρμόζονται σε συγκεκριμένες γεωγραφικές περιοχές [25], ή στη διατήρηση πολιτιστικής κληρονομιάς, αλλά επίσης παραβλέπει την πλούσια πολιτισμικά μουσική γνώση που είναι κρίσιμη για την πρόοδο της έρευνας στην ανάκτηση μουσικής πληροφορίας [11]. Κατά συνέπεια, καθίσταται επιτακτική η ανάγκη για την

ανάπτυξη πιο συμπεριληπτικών και πολιτισμικά ευαισθητοποιημένων υπολογιστικών μοντέλων [26], ικανών να γενικεύουν πέρα από τις Δυτικοκεντρικές παραδόσεις και να προσαρμόζονται αποτελεσματικά σε ποικίλες, υποεκπροσωπούμενες μουσικές κουλτούρες. Αυτή η κατεύθυνση έχει ήδη σημειώσει πρόοδο σε συγγενείς τομείς όπως η επεξεργασία φυσικής γλώσσας (natural language processing – NLP) [27] και η αναγνώριση ομιλίας (speech recognition) [28], μέσω της ανάπτυξης πολιτισμικά προσαρμοσμένων και πολυγλωσσικών foundation models.

Μια πολλά υποσχόμενη προσέγγιση για την αντιμετώπιση αυτών των προκλήσεων είναι η συνεχής προ-εκπαίδευση (continual pre-training – CPT), η οποία έχει αναδειχθεί ως μια αποτελεσματική και ολοένα πιο διαδεδομένη μέθοδος τόσο στα μεγάλα γλωσσικά μοντέλα (large language models – LLMs) [4, 29, 30] όσο και στην πολυτροπική μάθηση (multimodal learning) [31]. Επιτρέποντας στα μοντέλα να προσαρμόζονται σταδιακά σε νέα domains, tasks ή γλώσσες, το CPT αποφεύγει την ανάγκη για πλήρη επανεκπαίδευση, μια διαδικασία που συχνά είναι μη πρακτική και υπολογιστικά δαπανηρή [32, 4]. Σημαντικά, έχει αποδειχθεί ότι σε ορισμένες περιπτώσεις επιτυγχάνει απόδοση ισοδύναμη ή και ανώτερη από την εκπαίδευση από την αρχή (training from scratch) [33, 34], ενώ ταυτόχρονα οδηγεί σε ταχύτερη σύγκλιση [35] και μείωση του φαινομένου του catastrophic forgetting [36]. Το CPT έχει επίσης αρχίσει να εφαρμόζεται και στον τομέα του ήχου, με πρόσφατες μελέτες να τεκμηριώνουν την αποτελεσματικότητά του στην προσαρμογή μοντέλων αναγνώρισης ομιλίας τόσο σε γλώσσες υψηλών όσο και χαμηλών πόρων [37, 28, 38]. Επιπλέον, η συγχώνευση μοντέλων (model merging) [39, 40] έχει αναδειχθεί ως μια απλή αλλά αποτελεσματική τεχνική για την προσαρμογή προ-εκπαιδευμένων μοντέλων σε πολλαπλά domains, συνδυάζοντας domain-specific μοντέλα στον χώρο των βαρών (weight space), χωρίς να απαιτείται επιπλέον εκπαίδευση [41] ή πρόσβαση στα αρχικά δεδομένα εκπαίδευσης [42]. Μια ιδιαίτερα αξιοσημείωτη μέθοδος είναι το task arithmetic (TA) [5], η οποία κατασκευάζει *task vectors* υπολογίζοντας τη διαφορά μεταξύ των παραμέτρων ενός προσαρμοσμένου μοντέλου και του αντίστοιχου προ-εκπαιδευμένου. Αυτά τα task vectors μπορούν στη συνέχεια να ενσωματωθούν στο προ-εκπαιδευμένο μοντέλο μέσω αλγεβρικών πράξεων στον Ευκλείδειο χώρο, δημιουργώντας με αυτόν τον τρόπο ένα ενοποιημένο μοντέλο. Δεδομένης της έλλειψης πολιτισμικά ποικίλων επισημασμένων μουσικών δεδομένων, η συνεχής προ-εκπαίδευση προσφέρει μια υπολογιστικά αποδοτική λύση για την προσαρμογή των foundation models σε μη Δυτικές μουσικές παραδόσεις, χωρίς την ανάγκη πλήρους επανεκπαίδευσης. Παράλληλα, η τεχνική task arithmetic επιτρέπει την ομαλή συγχώνευση μοντέλων στον χώρο των βαρών, διευκολύνοντας την πολυπολιτισμική προσαρμογή και περιορίζοντας το φαινόμενο του catastrophic forgetting.

0.1.2 Συνεισφορά

Ενώ τόσο η συνεχής προ-εκπαίδευση όσο και η τεχνική task arithmetic έχουν μελετηθεί εκτενώς σε άλλους τομείς, η εφαρμογή τους στην ανάκτηση μουσικής πληροφορίας παραμένει σε μεγάλο βαθμό ανεξερεύνητη. Στην παρούσα διπλωματική εργασία καλύπτουμε αυτό το κενό, αξιοποιώντας τις δύο τεχνικές για την προσαρμογή του `MERT-v1-95M`[1], ενός μουσικού foundation μοντέλου [1], το οποίο έχει εκπαιδευτεί αρχικά σε 1.000 ώρες κυρίως Δυτικής μουσικής [1, 43]. Στόχος μας είναι να το προσαρμόσουμε σε μουσικές παραδόσεις από την

[1] https://huggingface.co/m-a-p/MERT-v1-95M

Ανατολική Μεσόγειο και την Ινδική υποήπειρο, διατηρώντας παράλληλα την απόδοσή του σε «Δυτικο»-κεντρικά benchmarks.

Μια σημαντική πρόκληση στην προσαρμογή των foundation models σε ποικίλα domains είναι η επίτευξη αποδοτικής προσαρμογής χωρίς την εμφάνιση του φαινομένου catastrophic forgetting [44], όπου η προγενέστερη γνώση ενδέχεται να «ξεχαστεί» όταν το μοντέλο εκπαιδεύεται σε νέα δεδομένα [45]. Για την αντιμετώπιση αυτού του προβλήματος, προτείνουμε μια υπολογιστικά αποδοτική στρατηγική συνεχούς προ-εκπαίδευσης δύο σταδίων, η οποία ενσωματώνει επανεκκίνηση του ρυθμού μάθησης (learning rate re-warming) [4], σταθεροποιώντας την εκπαίδευση και επιτρέποντας πιο ομαλή και αποτελεσματική προσαρμογή.

Η παρούσα διπλωματική εργασία συνεισφέρει στα εξής:

1. Από όσο γνωρίζουμε, πρόκειται για την πρώτη μελέτη που εξερευνά τη **συνεχή προ-εκπαίδευση (continual pre-training)** και την τεχνική **task arithmetic** για **διαπολιτισμική προσαρμογή (cross-cultural adaptation)** στην ανάκτηση μουσικής πληροφορίας, τεκμηριώνοντας την αποτελεσματικότητα αυτών των μεθόδων στην εκμάθηση αναπαραστάσεων μουσικού ήχου στο πλαίσιο των foundation models.

2. Προτείνουμε μια **στρατηγική συνεχούς προ-εκπαίδευσης δύο σταδίων**, η οποία σταθεροποιεί την εκπαίδευση, μειώνει το catastrophic forgetting και επιτρέπει αποτελεσματική προσαρμογή υπό περιορισμένους υπολογιστικούς πόρους.

3. Το πολυπολιτισμικά προσαρμοσμένο μοντέλο μας, `CultureMERT`, υπερβαίνει την απόδοση του αρχικού `MERT-v1` κατά **4,43%** κατά μέσο όρο στη μετρική ROC-AUC σε μη Δυτικά auto-tagging tasks, παρουσιάζοντας επίσης σταθερές μέσες βελτιώσεις σε άλλες μετρικές: 5,4% στο Average Precision (AP), 3,6/% στο Micro-F1, και 6,8% στο Macro-F1, με ελάχιστη απώλεια απόδοσης στα Δυτικά benchmarks.

4. Τα πολιτισμικά προσαρμοσμένα μοντέλα μας **ξεπερνούν προηγούμενα state-of-the-art** αποτελέσματα σε όλα τα μη Δυτικά auto-tagging tasks που εξετάζουμε.

5. Διερευνούμε τη **διαπολιτισμική μεταφερσιμότητα (cross-cultural transferability)**, αναλύοντας κατά πόσο τα μοντέλα που προσαρμόζονται σε δεδομένα από μία μόνο μουσική παράδοση (π.χ., Οθωμανική/Τουρκική κλασσική μουσική) μπορούν να γενικεύσουν σε άλλες (π.χ., Ελληνική παραδοσιακή μουσική). Τα αποτελέσματα δείχνουν ότι οι πολιτισμικά εξειδικευμένες προσαρμογές παρουσιάζουν διαφοροποιημένη ικανότητα μεταφοράς σε άλλες μουσικές παραδόσεις, ενώ το πολυπολιτισμικά προσαρμοσμένο μοντέλο επιτυγχάνει τη μεγαλύτερη γενίκευση στα σύνολα δεδομένων που μελετώνται.

Με την αντιμετώπιση αυτών των προκλήσεων, η παρούσα διπλωματική εργασία συμβάλλει στην ανάπτυξη πολιτισμικά ευαισθητοποιημένων foundation models για τη μουσική, τα οποία επιτρέπουν την κατανόηση της παγκόσμιας μουσικής ποικιλομορφίας και ενισχύουν τη διαπολιτισμική εκμάθηση αναπαραστάσεων μουσικής με βάση τον ήχο. Η μελέτη αυτή αναδεικνύει την αποτελεσματικότητα της συνεχούς προ-εκπαίδευσης ως προσέγγιση για διαπολιτισμική προσαρμογή στην ανάκτηση μουσικής πληροφορίας, καθιερώνοντας το `CultureMERT-95M` ως ένα state-of-the-art foundation μοντέλο για ποικίλες μουσικές παραδόσεις. Προς υποστήριξη

της περαιτέρω έρευνας στην εκμάθηση αναπαραστάσεων για υποεκπροσωπούμενες μουσικές κουλτούρες, δημοσιεύουμε το `CultureMERT-95M`, καθώς και την παραλλαγή του με χρήση της τεχνικής task arithmetic, `CultureMERT-TA-95M`.

0.2 Ανάκτηση Μουσικής Πληροφορίας

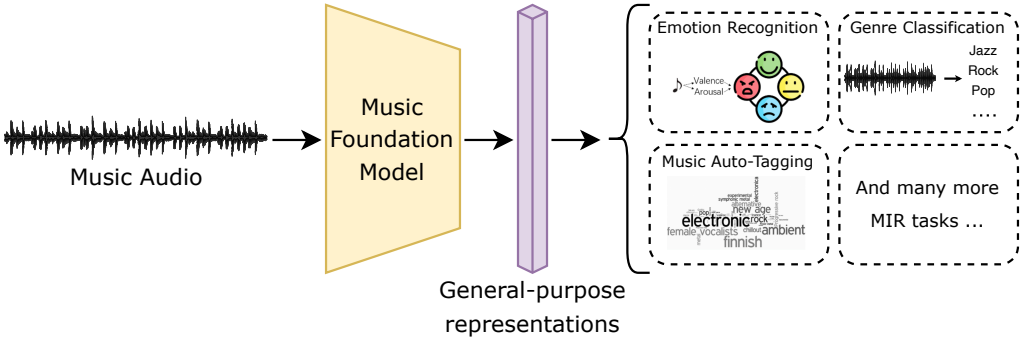
Η ανάκτηση μουσικής πληροφορίας είναι ένας διεπιστημονικός τομέας που επικεντρώνεται στην υπολογιστική ανάλυση, οργάνωση και διαχείριση δεδομένων που σχετίζονται με τη μουσική [46]. Ο όρος MIR χρησιμοποιείται μερικές φορές εναλλακτικά με τους όρους *music informatics* ή *music information processing* [47]. Τα τελευταία χρόνια, η έρευνα στον τομέα αυτό καθοδηγείται ολοένα και περισσότερο από τις εξελίξεις στη μηχανική μάθηση, και ιδίως στις μεθόδους βαθιάς μάθησης, οι οποίες έχουν επιφέρει αξιοσημείωτη πρόοδο σε πλήθος επιμέρους εφαρμογών.

0.2.1 Foundation Models στην Ανάκτηση Μουσικής Πληροφορίας

Πρώιμες μέθοδοι στην ανάκτηση μουσικής πληροφορίας βασίζονταν σε hand-crafted χαρακτηριστικά (π.χ. MFCCs, chroma features, constant-Q representations) και κλασσικούς αλγορίθμους μηχανικής μάθησης. Ωστόσο, η έλευση της βαθιάς μάθησης έφερε ριζικές αλλαγές στην εκμάθηση αναπαραστάσεων μουσικής, επιτυγχάνοντας αξιοσημείωτες επιδόσεις σε ποικίλες εφαρμογές [48]. Πιο πρόσφατα, η πιο επικρατούσα προσέγγιση για την εκμάθηση αναπαραστάσεων μουσικής βασίζεται στην αυτο-επιβλεπόμενη μάθηση (self-supervised learning – SSL), όπου τα μοντέλα εκπαιδεύονται σε proxy objectives που προκύπτουν απευθείας από τα ίδια τα δεδομένα εισόδου, εξαλείφοντας την ανάγκη για χειροκίνητη επισήμανση (labeling). Αυτή η προσέγγιση επιτρέπει την εξαγωγή πλούσιων και γενικεύσιμων αναπαραστάσεων, αξιοποιώντας αυτόματα παραγόμενα σήματα εποπτείας (self-supervision).

Πολλά μοντέλα βασισμένα στην αυτο-επιβλεπόμενη μάθηση έχουν επιδείξει ισχυρή απόδοση σε μια ευρεία γκάμα εργασιών ανάκτησης μουσικής πληροφορίας (downstream MIR tasks), μειώνοντας αποτελεσματικά το χάσμα με τις εποπτευόμενες προσεγγίσεις (supervised learning) [49, 50, 51, 52, 1, 53]. Τα μοντέλα αυτά προ-εκπαιδεύονται σε μεγάλες μουσικές συλλογές (music corpora), μαθαίνοντας γενικού σκοπού αναπαραστάσεις που μπορούν να μεταφερθούν αποτελεσματικά σε ποικίλες εφαρμογές MIR [11] (βλ. Σχήμα 1). Στη συνέχεια, προσαρμόζονται μέσω fine-tuning σε συγκεκριμένες εργασίες, χρησιμοποιώντας σημαντικά μικρότερα επισημασμένα σύνολα δεδομένων, επιτρέποντας έτσι αποδοτική μεταφορά γνώσης ακόμη και σε σενάρια περιορισμένων πόρων ή ελλιπούς εποπτείας.

Ένα κυρίαρχο pre-training paradigm είναι το *masked modeling* (MM), το οποίο προέρχεται από την προ-εκπαίδευση τύπου BERT στον τομέα της επεξεργασίας φυσικής γλώσσας [54]. Το MM βασίζεται στην τυχαία απόκρυψη (masking) τμημάτων της εισόδου και στην εκπαίδευση του μοντέλου να προβλέπει τα κρυμμένα στοιχεία με βάση τα συμφραζόμενα (context). Ένα χαρακτηριστικό παράδειγμα είναι το μοντέλο **MERT** [1] (Music undERstanding model with large-scale self-supervised Training), στο οποίο ένας encoder τύπου BERT, βασισμένος στην αρχιτεκτονική HuBERT [55], προ-εκπαιδεύεται σε μεγάλης κλίμακας ηχητικά δεδομένα μουσικής μέσω της μεθόδου masked language modeling (MLM). Το MERT υιοθετεί μια



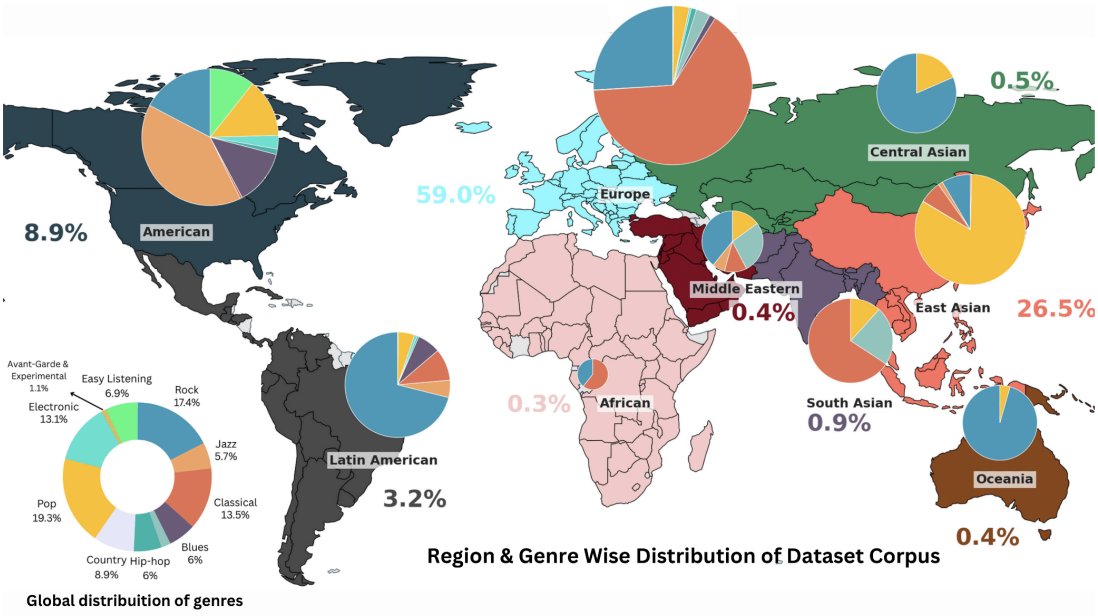
Σχήμα 1. *Τα foundation models μαθαίνουν αναπαραστάσεις γενικού σκοπού (general-purpose) μέσω μεγάλης κλίμακας προ-εκπαίδευσης σε ηχητικά δεδομένα, οι οποίες μπορούν στη συνέχεια να μεταφερθούν αποτελεσματικά σε ένα ευρύ φάσμα εφαρμογών ανάκτησης μουσικής πληροφορίας (MIR tasks).*

dual-teacher στρατηγική για τη δημιουργία σημάτων εποπτείας (supervision signals): έναν «ακουστικό» teacher, βασισμένο σε RVQ-VAE (συγκεκριμένα τον EnCodec audio tokenizer [2]), και έναν «μουσικό» teacher, βασισμένο στην ανακατασκευή μέσω του constant-Q transform (CQT). Ο συνδυασμός αυτός επιτρέπει στο μοντέλο να μαθαίνει τόσο ακουστικά όσο και αρμονικά χαρακτηριστικά της μουσικής. Το MERT διατίθεται σε δύο εκδοχές, με 95 και 330 εκατομμύρια παραμέτρους, και επιτυγχάνει state-of-the-art απόδοση σε 14 διαφορετικά music understanding tasks, επιβεβαιώνοντας την αποτελεσματικότητα της μεγάλης κλίμακας αυτο-επιβλεπόμενης προ-εκπαίδευσης και την ικανότητα ενοποίησης πολλαπλών εργασιών ανάκτησης μουσικής πληροφορίας σε ένα ενιαίο μοντέλο. Ένα άλλο παράδειγμα είναι το μοντέλο **MusicFM** [19], το οποίο βασίζεται στο MERT, αντικαθιστώντας τον εκπαιδεύσιμο ακουστικό audio tokenizer με έναν μη εκπαιδεύσιμο random projection quantizer, εμπνευσμένο από την αρχιτεκτονική του μοντέλου BEST-RQ [56]. Αυτή η προσέγγιση αφαιρεί την ανάγκη για ξεχωριστό στάδιο εκμάθησης αναπαραστάσεων (π.χ. μέσω RVQ ή k-means clustering), καθώς η διαδικασία του tokenization δεν απαιτεί εκπαίδευση. Συγκεκριμένα, το MusicFM προβάλλει τα φασματικά χαρακτηριστικά (log-mel features) σε λανθάνοντα χώρο (latent space) και τα διακριτοποιεί μέσω ενός τυχαία αρχικοποιημένου λεξικού (codebook). Παρά την απλότητά της, αυτή η προσέγγιση επιτυγχάνει υψηλή απόδοση, ιδιαίτερα όταν υπάρχουν επαρκή δεδομένα εκπαίδευσης.

Ωστόσο, υπάρχουν και άλλα είδη προ-εκπαίδευσης, όπως το *generative pre-training*. Αυτά τα foundation models εντάσσονται στην κατηγορία του auto-regressive predictive coding (APC), ενός παραδείγματος προ-εκπαίδευσης όπου το μοντέλο μαθαίνει να προβλέπει μελλοντικά tokens μέσα σε μια ακολουθία, χρησιμοποιώντας μια auto-regressive αρχιτεκτονική. Ένα χαρακτηριστικό παράδειγμα αποτελεί το **Jukebox** [20], ένας μεγάλης κλίμακας auto-regressive Transformer (5 δισεκατομμυρίων παραμέτρων), ο οποίος εκπαιδεύεται σε περισσότερα από 1,2 εκατομμύρια τραγούδια. Το Jukebox συμπιέζει το ηχητικό σήμα σε διακριτά codes χρησιμοποιώντας τρία ανεξάρτητα εκπαιδευμένα VQ-VAEs σε διαφορετικά temporal resolutions, με κάθε επίπεδο να χρησιμοποιεί λεξιλόγιο μεγέθους 2048. Παρόλο που σχεδιάστηκε κυρίως για music generation, το **JukeMIR** [57] έδειξε ότι οι ενδιάμεσες αναπαραστάσεις του Jukebox μπορούν να επαναχρησιμοποιηθούν για εργασίες κατανόησης μουσικής,

επιτυγχάνοντας ισχυρή απόδοση σε tasks όπως music auto-tagging, genre classification, key detection και emotion recognition. Ένα ακόμη παράδειγμα αποτελεί το **MusicGen** [58], το οποίο, αν και αρχικά σχεδιάστηκε και αυτό ως μοντέλο δημιουργίας μουσικής (music generation) με είσοδο κείμενο και μελωδία (text- and melody-conditioned), έχει επίσης αξιολογηθεί ως προς την ικανότητά του για εκμάθηση αναπαραστάσεων μουσικής (music representation learning). Το MusicGen χρησιμοποιεί έναν auto-regressive Transformer decoder, ο οποίος εκπαιδεύεται σε residual vector-quantized (RVQ) tokens που παράγονται από τον EnCodec tokenizer. Σε αντίθεση με μοντέλα όπως το Jukebox, τα οποία υλοποιούν πολυεπίπεδες ιεραρχίες με πολλαπλά priors, το MusicGen επεξεργάζεται επίπεδες ή διεμπλεκόμενες παράλληλες ροές από tokens που προέρχονται από πολλαπλά codebooks. Αυτός ο σχεδιασμός απλοποιεί τη διαδικασία εκπαίδευσης και βελτιώνει την υπολογιστική αποδοτικότητα, αποφεύγοντας το κόστος που συνεπάγεται η παραγωγή πολλαπλών ροών tokens από διαφορετικά codebooks. Οι ενδιάμεσες αναπαραστάσεις του MusicGen μπορούν να επαναχρησιμοποιηθούν σε πλήθος εφαρμογών ανάκτησης μουσικής πληροφορίας, όπως genre classification, key detection και music transcription.

0.2.2 Διαπολιτισμική Ανάκτηση Μουσικής Πληροφορίας



Σχήμα 2. *Παγκόσμια κατανομή μουσικών συλλογών δεδομένων (music corpora) ανά γεωγραφική περιοχή.*

Η ανάκτηση μουσικής πληροφορίας έχει παραδοσιακά επικεντρωθεί στην ανάλυση Δυτικών μουσικών ρεπερτορίων, με έμφαση κυρίως στην Ευρω-Αμερικανική ποπ και τη Δυτική κλασσική μουσική. Ένα αυξανόμενο σώμα ερευνητικών εργασιών αναδεικνύει την έντονη Δυτικοκεντρική προκατάληψη του πεδίου και υπογραμμίζει την ανάγκη για διαπολιτισμική διεύρυνση και δικαιότερη εκπροσώπηση [21, 11, 59, 60]. Για παράδειγμα, οι [61] παρουσίασαν το SAMBASET, ένα σύνολο δεδομένων διάρκειας άνω των 40 ωρών με Βραζιλιάνικη σάμπα, με στόχο να αντιπαρατεθούν στην κυρίαρχη Δυτικοκεντρική εστίαση του πεδίου. Υποστηρίζουν

ότι τα περισσότερα σύνολα δεδομένων, οι μεθοδολογίες και τα ερευνητικά συμπεράσματα στην ανάκτηση μουσικής πληροφορίας ενσωματώνουν σημαντικές πολιτισμικές προκαταλήψεις, με τη μη Δυτική μουσική να είναι συχνά υποεκπροσωπούμενη, ανεπαρκώς επισημασμένη ή ακόμη και εσφαλμένα κατηγοριοποιημένη. Αντίστοιχα, οι [3] ποσοτικοποιούν αυτή την προκατάληψη, δείχνοντας ότι μόλις το 5,7% των δεδομένων για δημιουργία μουσικής (music generation) προέρχεται από μη Δυτικές παραδόσεις, αναδεικνύοντας έτσι την έντονη υποεκπροσώπησή τους και την επείγουσα ανάγκη για πιο πολιτισμικά ποικίλα σύνολα δεδομένων (βλ. Σχήμα 2). Η αναγνώριση της Δυτικοκεντρικής προκατάληψης αποτελεί αναγκαίο πρώτο βήμα· ωστόσο, η ουσιαστική αντιμετώπισή της προϋποθέτει πρακτικές προσπάθειες για την ανάπτυξη κατάλληλων συνόλων δεδομένων, αναπαραστάσεων και μεθόδων αξιολόγησης προσαρμοσμένων σε ποικίλες (μη Δυτικές) μουσικές παραδόσεις.

Σύνολα Δεδομένων Πολλές πρωτοβουλίες έχουν αναδειχθεί για τη μείωση της πολιτισμικής προκατάληψης στην ανάκτηση μουσικής πληροφορίας, μέσω της δημιουργίας συνόλων δεδομένων από διαφορετικές γεωγραφικές περιοχές. Το CompMusic project [62] προσφέρει πάνω από 1300 ώρες μουσικών δεδομένων από παραδόσεις όπως η Ινδική (Ινδουστανική, Καρνατική), το Τουρκικό Μακάμ, η Αραβο-Ανδαλουσιανή μουσική και η Όπερα του Πεκίνου. Συμπληρωματικά, σύνολα δεδομένων όπως τα SAMBASET [61], corpusCOFLA [63], Nava Dastgāh [64], KritiSamhita [65] και Erkomaishvili Dataset [66] εστιάζουν στη Βραζιλιάνικη, Φλαμένκο, Περσική, Καρνατική και Γεωργιανή μουσική αντίστοιχα. Επιπλέον, έχουν προταθεί σύνολα δεδομένων όπως το Lyra για την Ελληνική παραδοσιακή μουσική [67], το CCMusic για Κινεζικές παραδόσεις [68], καθώς και δεδομένα για Αφρικανικά ιδιώματα όπως τα Sotho-Tswana [69] και Ndwom [70]. Τέλος, τα M4-RAG [71] και GlobalMood [15] προσφέρουν παγκόσμιας κλίμακας δεδομένα με πλούσιες πολυγλωσσικές και πολιτισμικές επισημάνσεις, ενισχύοντας την αντιπροσωπευτικότητα και τη διαπολιτισμική εγκυρότητα των δεδομένων.

Μεθοδολογίες Τα τελευταία χρόνια παρατηρείται αυξανόμενο ενδιαφέρον για την υπολογιστική μελέτη μη Δυτικών μουσικών παραδόσεων [72]. Ενδεικτικά παραδείγματα περιλαμβάνουν μελέτες αναγνώρισης Τουρκικών Μακάμ [73], ταξινόμησης Ινδικής μουσικής [74], καθώς και ανάλυσης της Ιρανικής [75], Κορεατικής [76] και Γκανέζικης [77] παραδοσιακής μουσικής. Η τελευταία μελέτη αποκαλύπτει μικροτονικές αποκλίσεις στις φωνητικές γραμμές των *seperewa* scales, αναδεικνύοντας τόσο τους περιορισμούς όσο και τις Δυτικοκεντρικές παραδοχές που ενσωματώνονται στα κλασσικά εργαλεία ανάκτησης μουσικής πληροφορίας. Παράλληλα, πρόσφατες εργασίες εξετάζουν τη διαπολιτισμική μεταφορά στην αυτόματη ταξινόμηση μουσικής (music auto-tagging) [78], καθώς και τεχνικές few-shot learning για σενάρια με χαμηλούς πόρους [79]. Επιπλέον, το CLaMP 3 [71] εισάγει πολυτροπικό και πολυγλωσσικό alignment για MIR tasks, πετυχαίνοντας state-of-the-art επιδόσεις σε εφαρμογές όπως η ανάκτηση μουσικής με βάση το κείμενο (text-to-audio retrieval). Τέλος, η πολιτισμική προσαρμογή σε μοντέλα music generation διερευνήθηκε μέσω αποδοτικής προσαρμογής των MusicGen [58] και Mustango [80], χρησιμοποιώντας low-resource πολιτισμικά σύνολα δεδομένων. Τα προσαρμοσμένα μοντέλα παρουσιάζουν βελτιωμένες επιδόσεις στην Ινδουστανική κλασική και την Τουρκική μουσική [3], αναδεικνύοντας τόσο τις δυνατότητες όσο και τις προκλήσεις της διαπολιτισμικής προσαρμογής των foundation models.

Προκλήσεις Παρά τις πρόσφατες προόδους, η ουσιαστική αντιμετώπιση της πολιτισμικής ανισορροπίας στο MIR δεν μπορεί να περιοριστεί στην απλή αύξηση της ποικιλομορφίας των συνόλων δεδομένων. Όπως τονίζουν οι [81], απαιτείται κριτική επανεξέταση των θεμελιωδών παραδοχών του πεδίου—επιστημολογικών, μεθοδολογικών και αξιακών—συμπεριλαμβανομένου του πώς ορίζεται, κατανοείται και μελετάται η μουσική σε διαφορετικά πολιτισμικά πλαίσια. Απαραίτητη είναι επίσης η διεπιστημονική σύνδεση με την εθνομουσικολογία και η ενεργή συμμετοχή ειδικών από τις αντίστοιχες μουσικές παραδόσεις. Στην πράξη, η έλλειψη δεδομένων περιορίζει τη δυνατότητα εκπροσώπησης πολλών πολιτισμών και οδηγεί σε unbalanced κατανομές μεταδεδομένων (metadata), οι οποίες επηρεάζουν την αξιολόγηση. Παράλληλα, τα μοντέλα που εκπαιδεύονται κυρίως σε Δυτικά δεδομένα ενσωματώνουν προκαταλήψεις, οι οποίες τα καθιστούν λιγότερο γενικεύσιμα σε μη Δυτικά μουσικά συστήματα. Για την υπέρβαση αυτών των περιορισμών απαιτούνται πολιτισμικά ευαισθητοποιημένα πρωτόκολλα αξιολόγησης, κατάλληλες μετρικές, και η από κοινού ανάπτυξη μοντέλων σε συνεργασία με local experts και practitioners, λαμβάνοντας υπόψη τις εννοιολογικές βάσεις και τις αξίες που διέπουν κάθε μουσική παράδοση.

0.3 Προτεινόμενη Μεθοδολογία

0.3.1 Σύνολα Δεδομένων και Αξιολόγηση

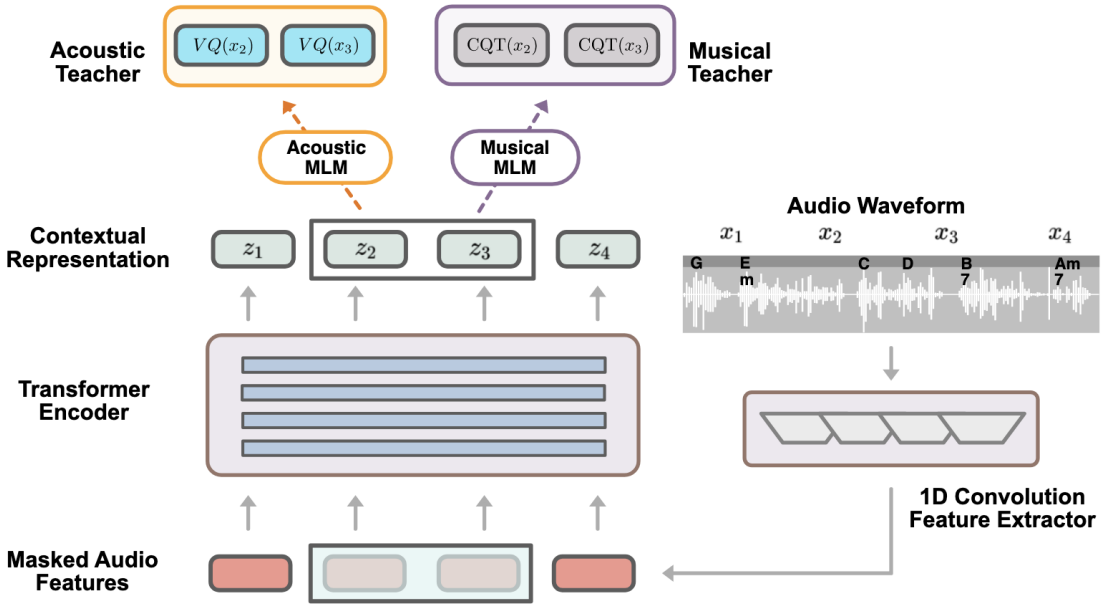
Για τα πειράματά μας, χρησιμοποιούμε μια ποικιλία από σύνολα μουσικών δεδομένων που καλύπτουν τόσο Δυτικές όσο και μη Δυτικές μουσικές παραδόσεις. Συγκεκριμένα, για την εκπροσώπηση της Δυτικής μουσικής χρησιμοποιούμε τα σύνολα MagnaTagATune (MTAT) [82] και FMA-medium [83]. Από την πλευρά των μη Δυτικών παραδόσεων, ενσωματώνουμε το σύνολο Lyra [67], το οποίο περιλαμβάνει Ελληνική παραδοσιακή μουσική, καθώς και τρεις συλλογές από τα CompMusic Corpora[2] [62]: την κλασσική Οθωμανική/Τουρκική μουσική (Turkish-makam) [84, 85], η οποία, μαζί με το Lyra, εκπροσωπεί την Ανατολική Μεσόγειο· καθώς και τις παραδόσεις της Ινδουστανικής (Hindustani) και Καρνατικής (Carnatic) κλασσικής μουσικής [86], που αντιστοιχούν στη Βόρεια και Νότια Ινδία αντίστοιχα.

Αξιολογούμε τα μοντέλα μας σε εργασίες αυτόματης επισημείωσης/ταξινόμησης μουσικής (music tagging), τόσο για Δυτικές όσο και για μη Δυτικές παραδόσεις, στο πλαίσιο διαπολιτισμικής αξιολόγησης (cross-cultural evaluation). Χρησιμοποιούμε καθιερωμένες μετρικές για multi-label classification, όπως το ROC-AUC, το average precision (AP), καθώς και τις μετρικές F1 (micro-averaged και macro-averaged). Ακολουθώντας τις προσεγγίσεις των [57, 1, 19], εφαρμόζουμε probing-based evaluation [87] αντί για πλήρες fine-tuning, διατηρώντας τα προ-εκπαιδευμένα μοντέλα «παγωμένα» και χρησιμοποιώντας τα ως deep feature extractors, ενώ εκπαιδεύουμε μόνο ένα MLP με ένα κρυφό επίπεδο 512 νευρώνων πάνω από αυτά. Για την προετοιμασία των δεδομένων μας για συνεχή προ-εκπαίδευση, εξάγουμε αποσπάσματα διάρκειας 30 δευτερολέπτων από κάθε σύνολο εκπαίδευσης (training set) των μη Δυτικών συνόλων δεδομένων. Δεδομένων των διαφορών στον συνολικό όγκο των συνόλων, εξισορροπούμε τη διάρκεια των δεδομένων εκπαίδευσης ανά πολιτισμό ώστε να διασφαλιστεί αναλογική εκπροσώπηση: εξάγουμε 200 ώρες από τα σύνολα Τουρκικού Μακάμ, Καρνατι-

[2] https://compmusic.upf.edu/corpora

κής και Ινδουστανικής μουσικής, και 50 ώρες από το Lyra, λόγω του μικρότερου μεγέθους του. Στη συνέχεια, συνδυάζουμε αυτά τα υποσύνολα για να κατασκευάσουμε ένα ενοποιημένο σύνολο 650 ωρών, που ενσωματώνει και τις τέσσερις παραδόσεις, με στόχο την πολυπολιτισμική εκπαίδευση του αρχικού μοντέλου.

0.3.2 Το μοντέλο MERT

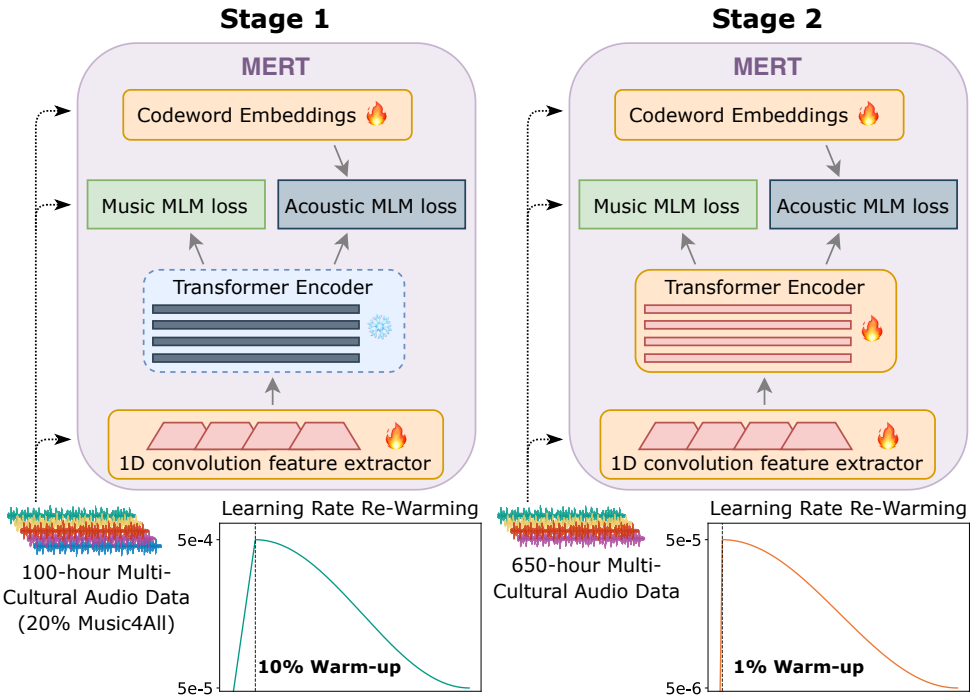


Σχήμα 3. *Η αρχιτεκτονική του μοντέλου MERT [1].*

Η συνεχής προ-εκπαίδευση που εφαρμόζουμε ακολουθεί το self-supervised masked language modeling (MLM) objective του $\texttt{MERT}^{\texttt{RVQ-VAE}}$, στο οποίο δύο teacher models παρέχουν τα pseudo-labels:

- (i) ένας *acoustic teacher*, συγκεκριμένα το EnCodec codec μοντέλο [2], το οποίο διακριτοποιεί τον ήχο σε tokens από $K = 8$ residual vector quantization (RVQ) codebooks, καθένα εκ των οποίων περιέχει $C = 1024$ codewords, και
- (ii) ένας *musical teacher*, βασισμένος σε ανακατασκευή constant-Q transform (CQT) spectrograms, ο οποίος μοντελοποιεί πληροφορία σχετική με το pitch και την αρμονική δομή.

Το `MERT-v1-95M` ακολουθεί την HuBERT αρχιτεκτονική [55], η οποία αποτελείται από έναν CNN-based feature extractor που επεξεργάζεται raw audio waveforms με sampling rate 24 kHz και τα μετατρέπει σε frame-level representations στα 75 Hz, καθώς και έναν 12-layer Transformer encoder που παράγει 768-dimensional contextual embeddings (βλ. Σχήμα 3). Κατά την εκπαίδευση, ένα υποσύνολο των frame embeddings καλύπτεται με μάσκα (masking) και το μοντέλο βελτιστοποιείται μέσω ενός multi-task learning (MTL) objective, το οποίο συνδυάζει masked acoustic token prediction και spectrogram reconstruction.



Σχήμα 4. ***Στρατηγική Συνεχούς Προ-εκπαίδευσης Δύο Σταδίων του*** *CultureMERT. Στο Στάδιο 1, εκπαιδεύεται ένα υποσύνολο των παραμέτρων (ο 1D CNN feature extractor και τα codeword embeddings) πάνω σε 100 ώρες δεδομένων για πολλαπλά epochs, με το 20% του συνόλου να αποτελείται από Δυτική μουσική. Στο Στάδιο 2, όλες οι παράμετροι του μοντέλου εκπαιδεύονται πάνω στο πλήρες σύνολο δεδομένων των 650 ωρών. Η διαδικασία επαναθέρμανσης του learning rate εφαρμόζεται και στα δύο στάδια, επιτρέποντας ομαλή και σταθερή προσαρμογή.*

0.3.3 Συνεχής Προ-εκπαίδευση Δύο Σταδίων

Για την προσαρμογή του μοντέλου MERT σε ποικίλες μουσικές παραδόσεις, εφαρμόζουμε continual pre-training, μια διαδικασία κατά την οποία ένα ήδη προ-εκπαιδευμένο μοντέλο συνεχίζει να εκπαιδεύεται σε νέα δεδομένα, με στόχο την προσαρμογή του σε διαφορετικό domain, διατηρώντας τη γνώση που έχει ήδη αποκτήσει, χωρίς ανάγκη πλήρους επανεκπαίδευσης. Στην περίπτωσή μας, αυτό συνεπάγεται τη συνεχή προ-εκπαίδευση του `MERT-v1-95M`, χρησιμοποιώντας το ίδιο pre-training objective, πάνω σε πολιτισμικά ετερογενή δεδομένα που εισάγουν σημαντική μετατόπιση στην κατανομή των δεδομένων (distribution shift), καθώς το μοντέλο είχε αρχικά εκπαιδευτεί κυρίως σε Δυτική μουσική [1, 43]. Δεδομένης αυτής της μετατόπισης, η αφελής συνέχιση της εκπαίδευσης, δηλαδή η ταυτόχρονη προσαρμογή όλων των παραμέτρων χωρίς επαναφορά του learning rate, μπορεί να οδηγήσει σε φαινόμενα catastrophic forgetting [44] και ανεπαρκή προσαρμογή [4], όπως επιβεβαιώνεται και από τα προκαταρκτικά μας πειράματα (βλ. Ενότητα 0.4.3). Για την αντιμετώπιση αυτών των φαινομένων, προτείνουμε μια **στρατηγική δύο σταδίων**, η οποία σταθεροποιεί τη διαδικασία εκπαίδευσης μέσω:

- (i) **επαναθέρμανσης (re-warming) και εκ νέου μείωσης του ρυθμού μάθησης** [4, 31, 34, 88, 89], και
- (ii) **σταδιακής προσαρμογής (staged adaptation)**.

Η μέθοδός μας παρουσιάζεται στο Σχήμα 4, όπου συγκεκριμένα απεικονίζεται η στρατηγική δύο σταδίων για τη συνεχή προ-εκπαίδευση του `CultureMERT`.

Σταδιακή Προσαρμογή (Staged Adaptation) Στα αρχικά μας πειράματα παρατηρήσαμε μια αρχική πτώση στην απόδοση κατά τη διάρκεια της συνεχούς προ-εκπαίδευσης, ακολουθούμενη από μια φάση αργής ανάκαμψης, φαινόμενο γνωστό ως *stability gap* [90]. Η αστάθεια αυτή οφείλεται στην απότομη προσαρμογή των παραμέτρων του μοντέλου σε μια αρκετά διαφορετική κατανομή δεδομένων (distribution shift). Για να μετριάσουμε αυτό το φαινόμενο, αντί να προσαρμόζουμε όλες τις παραμέτρους από την αρχή σε ολόκληρο το σύνολο δεδομένων, χωρίζουμε την εκπαίδευση σε δύο διαδοχικά στάδια, ώστε να επιτευχθεί ομαλότερη και πιο σταθερή προσαρμογή (βλ. Σχήμα 4):

- **Στάδιο 1 — Φάση Σταθεροποίησης**: Το μοντέλο εκπαιδεύεται αρχικά σε ένα μικρότερο υποσύνολο δεδομένων [90], με ενημέρωση μόνο του CNN-based feature extractor και του codeword embedding layer, ενώ ο Transformer encoder διατηρείται «παγωμένος». Για να αντιμετωπίσουμε το distribution shift και να περιορίσουμε το forgetting, ενσωματώνουμε ποσοστό 20% από δεδομένα του συνόλου Music4All [7], το οποίο είναι κατά βάση Δυτικής προέλευσης.

- **Στάδιο 2 — Πλήρης Προσαρμογή**: Ο Transformer encoder «ξεπαγώνεται» και συνεχίζεται η εκπαίδευση στο πλήρες σύνολο των 650 ωρών. Η ενσωμάτωση Δυτικών δεδομένων μπορεί να διατηρηθεί και σε αυτό το στάδιο, ώστε να περιοριστεί περαιτέρω το forgetting, ωστόσο εισάγει έναν συμβιβασμό (trade-off) μεταξύ πολιτισμικής προσαρμογής (plasticity) και διατήρησης πρότερης γνώσης (stability), μια πρόκληση γνωστή ως *stability–plasticity dilemma* [91, 92, 93] (βλ. Ενότητα 0.4.3).

Ο διαχωρισμός της εκπαίδευσης σε δύο στάδια μάς επιτρέπει να ελέγξουμε πιο αποτελεσματικά αυτήν τη δυναμική μεταξύ προσαρμοστικότητας και σταθερότητας, επιδιώκοντας μια λειτουργική ισορροπία ανάμεσα στην ενσωμάτωση νέας πολιτισμικής γνώσης και στη διατήρηση όσων έχει ήδη μάθει το προ-εκπαιδευμένο μοντέλο

Επαναθέρμανση (re-warming) του ρυθμού μάθησης Για τη βελτίωση της σταθερότητας κατά τη συνεχή προ-εκπαίδευση, εφαρμόζουμε *επαναθέρμανση* (re-warming) και εκ νέου *μείωση* του ρυθμού μάθησης (learning rate) και στα δύο στάδια. Η εκπαίδευση σε νέα δεδομένα μπορεί να οδηγήσει σε ασταθή σύγκλιση και φαινόμενα forgetting, εάν το learning rate δεν προσαρμοστεί κατάλληλα [4, 34]. Προηγούμενες μελέτες έχουν δείξει ότι το re-warming του learning rate είναι κρίσιμο για την επιτυχή προσαρμογή και τον μετριασμό του φαινομένου catastrophic forgetting [4, 31, 34]. Ο τρόπος με τον οποίο μεταβάλλεται το learning rate επηρεάζει καθοριστικά τα dynamics της εκπαίδευσης και την αποτελεσματικότητα της προσαρμογής, καθιστώντας την επαναθέρμανση απαραίτητη για αποδοτική μάθηση σε νέα δεδομένα. Πιο συγκεκριμένα, ακολουθούμε τα παρακάτω:

- Στο **Στάδιο 1**, υιοθετούμε ένα ελαφρώς πιο aggressive warm-up schedule, ώστε να ενισχύσουμε την αρχική προσαρμογή των low-level representations (CNN-based fea-

ture extractor και codeword embeddings), πριν ξεκινήσει η εκπαίδευση του Transformer encoder και η μάθηση πιο high-level αναπαραστάσεων.

- Στο **Στάδιο 2**, εφαρμόζουμε ένα ηπιότερο schedule, επιδιώκοντας ισορροπία ανάμεσα στο *plasticity* και *stability* κατά την πλήρη εκπαίδευση του μοντέλου, περιορίζοντας παράλληλα τις αστάθειες στην εκπαίδευση (training instabilities).

Ακολουθώντας τη μέθοδό μας, αναπτύσσουμε τα παρακάτω πολιτισμικά προσαρμοσμένα μοντέλα:

- (i) Ένα πολυπολιτισμικά προσαρμοσμένο μοντέλο, το `CultureMERT`, το οποίο εκπαιδεύεται σε ένα ενοποιημένο σύνολο δεδομένων που περιλαμβάνει και τις τέσσερις μη Δυτικές μουσικές παραδόσεις (Turkish-makam, Hindustani, Carnatic και Lyra).
- (ii) Τέσσερα μοντέλα προσαρμοσμένα σε δεδομένα μίας μόνο μουσικής παράδοσης, οδηγώντας στα `MakamMERT`, `HindustaniMERT`, `CarnaticMERT` και `LyraMERT`, αντίστοιχα.

0.3.4 Συγχώνευση Μοντέλων με την Τεχνική Task Arithmetic

Ως εναλλακτική της συνεχούς προ-εκπαίδευσης σε ένα ενοποιημένο σύνολο δεδομένων που περιλαμβάνει και τις τέσσερις μουσικές παραδόσεις, διερευνούμε τη μέθοδο task arithmetic [5] — μια τεχνική συγχώνευσης μοντέλων που συνδυάζει πολιτισμικά εξειδικευμένα μοντέλα στον χώρο των βαρών (weight space), με σκοπό την κατασκευή ενός ενιαίου πολυπολιτισμικού μοντέλου. Η μέθοδος task arithmetic βασίζεται στον αλγεβρικό συνδυασμό παραμέτρων μοντέλων, μέσω προσθαφαιρέσεων διανυσμάτων βαρών στον Ευκλείδειο χώρο. Συγκεκριμένα, αντιμετωπίζει τη αλγεβρική διαφορά μεταξύ ενός μοντέλου προσαρμοσμένου σε ένα task ή domain και της αρχικής του προ-εκπαιδευμένης εκδοχής ως ένα *task vector* στον χώρο των βαρών. Έχει αποδειχθεί ότι γραμμικοί συνδυασμοί τέτοιων task vectors μπορούν να κατευθύνουν αποτελεσματικά τη συμπεριφορά του μοντέλου και να επιτρέψουν μεταφορά γνώσης μεταξύ διαφορετικών domains [94, 5].

Στη δική μας περίπτωση, υπολογίζουμε τα *task vectors* ως την element-wise διαφορά μεταξύ των παραμέτρων των πολιτισμικά εξειδικευμένων μοντέλων — δηλαδή των single-culture continually pre-trained models — και του αρχικού μοντέλου `MERT-v1`. Πιο τυπικά, αν το αρχικό προ-εκπαιδευμένο μοντέλο έχει παραμέτρους θ_{pre} και το μοντέλο θ_i έχει προσαρμοστεί σε ένα πολιτισμικό σύνολο δεδομένων $\mathcal{D}_i$, τότε το task vector για τη συγκεκριμένη μουσική παράδοση i δίνεται από τη σχέση: $\tau_i = \theta_i - \theta_{\text{pre}}$. Για την πολυπολιτισμική προσαρμογή, κατασκευάζουμε ένα ενοποιημένο μοντέλο θ_{merged} μέσω *task arithmetic*, συγχωνεύοντας N μοντέλα που έχουν προσαρμοστεί σε μεμονωμένες μουσικές παραδόσεις, αθροίζοντας τα αντίστοιχα task vectors τ_i με συντελεστές βαρύτητας λ_i για καθένα από αυτά:

$$\theta_{\text{merged}} = \theta_{\text{pre}} + \sum_{i=1}^{N} \lambda_i \tau_i, \tag{1}$$

όπου $\lambda_i \in \mathbb{R}$ είναι βαθμωτοί υπερπαράμετροι (scalar hyperparameters) που ελέγχουν τη συνεισφορά κάθε task vector.

Συχνά εφαρμόζεται ένας ενιαίος συντελεστής λ για όλα τα task vectors, δηλαδή $\lambda_i = \lambda, \forall i$, οπότε η Εξίσωση 1 αναδιατυπώνεται ως:

$$\theta_{\text{merged}} = \theta_{\text{pre}} + \lambda \sum_{i=1}^{N} \tau_i. \tag{2}$$

Σε αυτό το πλαίσιο, συγχωνεύουμε τα $N = 4$ μοντέλα που έχουν προσαρμοστεί σε μεμονωμένες μουσικές παραδόσεις — τα `MakamMERT`, `HindustaniMERT`, `CarnaticMERT` και `LyraMERT` — για να κατασκευάσουμε ένα ενιαίο πολυπολιτισμικό μοντέλο, το οποίο αναφερόμαστε ως `CultureMERT-TA`. Λεπτομέρειες για την επιλογή του συντελεστή λ δίνονται στην Ενότητα 0.4.2.

0.4 Πειράματα και Αποτελέσματα

Όπως παρουσιάζεται στους Πίνακες 1 και 2, το `CultureMERT`, το οποίο έχει προσαρμοστεί μέσω πολυπολιτισμικής συνεχούς προ-εκπαίδευσης, υπερέχει σταθερά του αρχικού μοντέλου `MERT-v1` σε όλα τα μη Δυτικά music auto-tagging tasks, σε όλες τις μετρικές αξιολόγησης. Υπερβαίνει επίσης, κατά μέσο όρο, τα μοντέλα που έχουν προσαρμοστεί σε μεμονωμένες παραδόσεις, γεγονός που υποδηλώνει ότι η ενσωμάτωση πολιτισμικά ποικιλόμορφων δεδομένων κατά τη διάρκεια του continual pre-training ενισχύει την ποιότητα των αναπαραστάσεων για κάθε επιμέρους μουσική παράδοση, βελτιώνοντας έτσι τη γενίκευση. Αξιοσημείωτο είναι ότι το `CultureMERT` επιτυγχάνει αυτά τα αποτελέσματα με ελάχιστη απώλεια απόδοσης σε Δυτικά σύνολα αξιολόγησης (μέση μείωση ROC-AUC και AP κατά μόλις -0.05%), γεγονός που καταδεικνύει την αποτελεσματικότητα της προσέγγισής μας. Επιπλέον, παρουσιάζει καλύτερη διατήρηση της πρότερης γνώσης στη Δυτική μουσική, συγκριτικά με τα single-culture μοντέλα, τα οποία εμφανίζουν σημαντικότερες μειώσεις απόδοσης κατά την αξιολόγηση τους στα auto-tagging tasks των FMA-medium και MagnaTagATune (MTAT).

Dataset	**Turkish-makam**		**Hindustani**		**Carnatic**		**Lyra**		**FMA-medium**		**MTAT**		**Avg.**
Metrics	**ROC**	**AP**	**ROC**	**AP**	**ROC**	**AP**	**ROC**	**AP**	**ROC**	**AP**	**ROC**	**AP**	
MERT-v1	$83.2_{0.08}$	$53.3_{0.12}$	$82.4_{0.04}$	$52.9_{0.19}$	$74.9_{0.05}$	$39.7_{0.15}$	$85.7_{0.10}$	$56.5_{0.18}$	$90.7_{0.04}$	$48.1_{0.11}$	$89.6_{0.07}$	$35.9_{0.15}$	66.1
MakamMERT	$88.7_{0.11}$	$58.8_{0.22}$	$84.5_{0.16}$	$57.8_{0.18}$	$77.6_{0.14}$	$42.7_{0.16}$	$84.6_{0.12}$	$53.2_{0.17}$	$90.3_{0.12}$	$47.1_{0.16}$	$89.0_{0.07}$	$35.6_{0.12}$	67.5
CarnaticMERT	$88.4_{0.06}$	$58.4_{0.16}$	$87.0_{0.06}$	$60.2_{0.14}$	$78.8_{0.13}$	$\mathbf{44.0}_{0.17}$	$85.4_{0.11}$	$55.8_{0.16}$	$90.2_{0.10}$	$46.7_{0.09}$	$89.2_{0.10}$	$35.3_{0.11}$	68.3
HindustaniMERT	$88.3_{0.12}$	$58.2_{0.16}$	$87.4_{0.11}$	$60.3_{0.16}$	$77.0_{0.12}$	$42.7_{0.16}$	$84.2_{0.13}$	$52.0_{0.15}$	$90.2_{0.13}$	$46.1_{0.10}$	$89.1_{0.09}$	$35.8_{0.13}$	67.6
LyraMERT	$86.7_{0.07}$	$56.8_{0.13}$	$85.9_{0.08}$	$57.4_{0.13}$	$76.4_{0.09}$	$40.1_{0.13}$	$85.0_{0.11}$	$53.5_{0.14}$	$90.0_{0.08}$	$46.0_{0.16}$	$88.9_{0.05}$	$35.1_{0.14}$	66.8
CultureMERT	$\mathbf{89.6}_{0.09}$	$60.6_{0.21}$	$\mathbf{88.2}_{0.20}$	$\mathbf{63.5}_{0.24}$	$\mathbf{79.2}_{0.18}$	$43.1_{0.22}$	$86.9_{0.10}$	$56.7_{0.20}$	$90.7_{0.09}$	$48.1_{0.13}$	$89.4_{0.09}$	$35.9_{0.16}$	**69.3**
CultureMERT-TA	$89.0_{0.12}$	$\mathbf{61.0}_{0.18}$	$87.5_{0.10}$	$59.3_{0.13}$	$79.1_{0.11}$	$43.3_{0.13}$	$\mathbf{87.3}_{0.08}$	$\mathbf{57.3}_{0.19}$	$90.8_{0.06}$	$49.1_{0.15}$	$89.6_{0.10}$	$36.4_{0.14}$	69.1
(Previous) SOTA	87.7 [78]	57.7 [78]	86.5 [78]	63.1 [78]	77.0 [78]	43.9 [78]	85.4 [78]	54.3 [78]	**92.4** [78]	**53.7** [78]	**92.7** [95]	**41.4** [57]	-

Πίνακας 1. ***Αποτελέσματα Αξιολόγησης (ROC-AUC και AP) των Προεκπαιδευμένων και Πολιτισμικά Προσαρμοσμένων Μοντέλων MERT σε Διάφορες Εργασίες Αυτόματης Ταξινόμησης Μουσικής (1/2).*** *Αναφέρονται μέσοι όροι από πέντε random seeds, με τις αντίστοιχες τυπικές αποκλίσεις ως δείκτες. Η στήλη «Avg.» αντιπροσωπεύει τη μέση απόδοση σε όλα τα σύνολα δεδομένων και τις μετρικές αξιολόγησης για κάθε μοντέλο.*

Dataset	Turkish-makam		Hindustani		Carnatic		Lyra		FMA-medium		MTAT		Avg.
Metrics (F1)	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	
MERT-v1	73.0	38.9	71.1	33.2	80.1	30.0	72.4	42.6	57.0	36.9	**35.7**	21.2	49.3
MakamMERT	**77.5**	44.0	74.0	37.6	81.0	31.4	70.8	40.2	57.2	35.4	34.2	20.5	50.3
CarnaticMERT	76.8	44.0	76.2	46.3	81.6	32.4	72.9	42.8	57.3	35.3	33.3	22.5	51.8
HindustaniMERT	76.5	43.9	**78.9**	46.9	81.0	**33.0**	70.1	40.6	55.1	33.8	34.4	20.9	51.3
LyraMERT	75.9	42.1	75.9	44.9	80.9	29.6	71.3	41.1	56.2	33.9	33.8	21.2	50.6
CultureMERT	77.4	**45.8**	77.8	**50.4**	**82.7**	32.5	**73.1**	43.1	58.3	36.6	35.6	**22.9**	**52.9**
CultureMERT-TA	76.9	45.4	74.2	45.0	82.5	32.1	73.0	**45.3**	**59.1**	**38.2**	**35.7**	21.5	52.4

Πίνακας 2. ***Αποτελέσματα Αξιολόγησης (Micro-F1 και Macro-F1) των Προ-εκπαιδευμένων και Πολιτισμικά Προσαρμοσμένων Μοντέλων MERT σε Διάφορες Εργασίες Αυτόματης Ταξινόμησης Μουσικής (2/2).*** *Η στήλη «Avg.» παρουσιάζει τη μέση απόδοση κάθε μοντέλου, υπολογισμένη ως ο μέσος όρος επί όλων των συνόλων δεδομένων και των δύο μετρικών.*

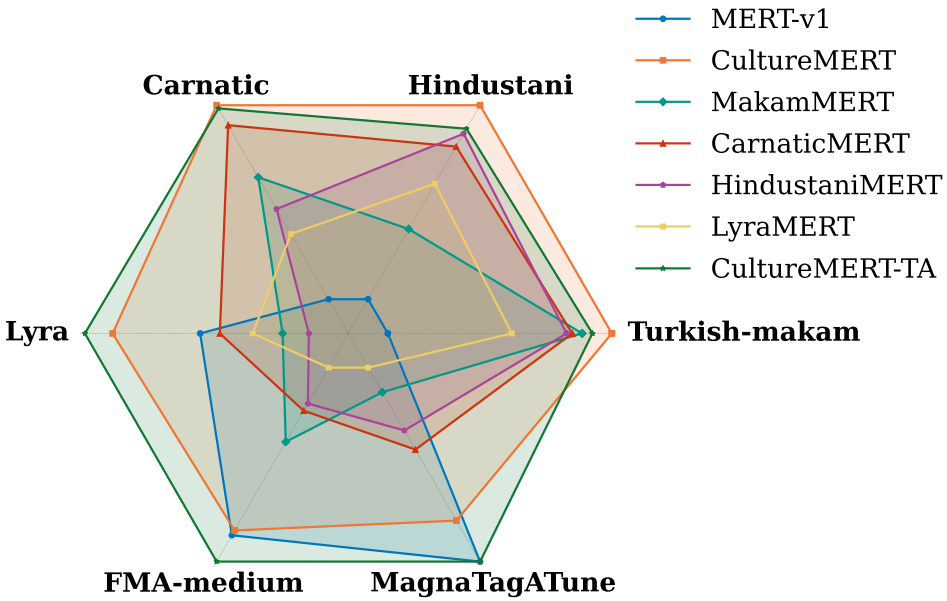
Παρατηρούμε επίσης ότι τα μοντέλα που έχουν προσαρμοστεί σε μία μόνο μουσική παράδοση (single-culture adapted models) τείνουν να επιτυγχάνουν βέλτιστη απόδοση στα αντίστοιχα in-domain tasks, ιδιαίτερα σε well-resourced παραδόσεις, επιβεβαιώνοντας την αποτελεσματικότητα της συνεχούς προ-εκπαίδευσης για domain-specific adaptation [32]. Η τάση αυτή παρατηρείται σε όλες τις μετρικές αξιολόγησης που χρησιμοποιούμε. Αξιοσημείωτο είναι ότι ακόμη και η προσαρμογή σε σύνολα δεδομένων με περιορισμένους πόρους (low-resource adaptation), όπως στην περίπτωση του `LyraMERT` που εκπαιδεύτηκε σε μόλις 50 ώρες Ελληνικής παραδοσιακής μουσικής, οδηγεί σε σημαντικές βελτιώσεις σε άλλα μη Δυτικά tasks. Το γεγονός αυτό υποδεικνύει ότι ακόμη και περιορισμένη έκθεση σε diverse δεδομένα κατά τη διάρκεια της εκπαίδευσης μπορεί να ενισχύσει σημαντικά τη διαπολιτισμική γενίκευση (cross-cultural generalization) πέραν των Δυτικών datasets.

Επιπλέον, η πολυπολιτισμική συγχώνευση μοντέλων μέσω της τεχνικής *task arithmetic* επιτυγχάνει συγκρίσιμες επιδόσεις με το `CultureMERT` στα μη Δυτικά σύνολα δεδομένων, ενώ το υπερβαίνει στα Δυτικά tasks και στο Lyra, καταδεικνύοντας ότι η συγχώνευση πολιτισμικά εξειδικευμένων μοντέλων στον χώρο των βαρών μπορεί να αποτελέσει μια αποτελεσματική, χωρίς επιπλέον εκπαίδευση (training-free), εναλλακτική — υπό την προϋπόθεση ότι τα επιμέρους μοντέλα είναι ήδη διαθέσιμα. Αξιοσημείωτο είναι επίσης ότι το task arithmetic υπερβαίνει, κατά μέσο όρο, ακόμη και το αρχικό προ-εκπαιδευμένο μοντέλο στα Δυτικά tasks, ενισχύοντας περαιτέρω την ικανότητά του να επιτυγχάνει ισορροπία μεταξύ αποτελεσματικής προσαρμογής και διατήρησης πρότερης γνώσης. Τέλος, τα `CultureMERT` και `CultureMERT-TA` υπερβαίνουν τα προηγούμενα SOTA αποτελέσματα της βιβλιογραφίας σε όλα τα μη Δυτικά music auto-tagging tasks, ως προς τις μετρικές ROC-AUC και AP, με την καλύτερη παραλλαγή του task arithmetic να επιτυγχάνεται για $\lambda = 0.2$ (βλ. Σχήμα 7). Ενδιαφέρον παρουσιάζει ότι μόνο τα πολυπολιτισμικά μοντέλα, `CultureMERT` και `CultureMERT-TA`, ξεπερνούν το αρχικό `MERT-v1` στο Lyra auto-tagging task, έστω και με τη μικρότερη διαφορά συγκριτικά με τα υπόλοιπα tasks. Η παρατήρηση αυτή συνάδει με το γεγονός ότι το `MERT-v1`, προ-εκπαιδευμένο σε Δυτική μουσική, αποτελεί ήδη ένα ισχυρό baseline για το Lyra, υπερβαίνοντας τα προηγούμενα

SOTA αποτελέσματα· γεγονός που ενδεχομένως αντανακλά ομοιότητες ανάμεσα στην Ελληνική παραδοσιακή μουσική και τις Δυτικές μουσικές παραδόσεις. Συνολικά, τα ευρήματά μας ενισχύουν περαιτέρω την αποτελεσματικότητα της πολυπολιτισμικής προσαρμογής, ιδίως σε σενάρια με περιορισμένους πόρους εκπαίδευσης και απαιτήσεις για διαπολιτισμική γενίκευση.

0.4.1 Διαπολιτισμική Γενίκευση και Μεταφορά Γνώσης

Όπως φαίνεται στο Σχήμα 5, η συνεχής προ-εκπαίδευση σε μία μουσική παράδοση μπορεί να ενισχύσει την απόδοση και σε άλλες, αν και σε διαφορετικό βαθμό. Ενδεικτικά, παρατηρείται ισχυρή διαπολιτισμική μεταφορά μεταξύ της Τουρκικής (Turkish-makam) και της Καρνατικής (Carnatic) παράδοσης, καθώς τα μοντέλα που έχουν προσαρμοστεί στη μία γενικεύουν αποτελεσματικά και στην άλλη. Παρόμοια μεταφορά γνώσης παρατηρείται και μεταξύ της Καρνατικής (Carnatic) και της Ινδουστανικής (Hindustani) μουσικής. Συγκεκριμένα, το μοντέλο που έχει προσαρμοστεί στην Καρνατική επιτυγχάνει υψηλές επιδόσεις σε όλες τις μετρικές κατά την αξιολόγησή του στο Hindustani auto-tagging task, ενώ το μοντέλο της Hindustani παρουσιάζει ελαφρώς καλύτερη απόδοση στις μετρικές F1 όταν εφαρμόζεται στην Καρνατική μουσική (βλ. Πίνακες 1 και 2). Η αμοιβαία αυτή μεταφερσιμότητα ενισχύει την εγγύτητα μεταξύ των δύο μουσικών παραδόσεων, ιδίως λόγω της κοινής χρήσης των *raga* και *tala* [24], παρά τις διαφορές στη μορφολογία της εκτέλεσης, τις μελωδικές κινήσεις και τα όργανα που χρησιμοποιούνται.



Σχήμα 5. ***Διαπολιτισμική Μεταφερσιμότητα.*** *Απόδοση των προσαρμοσμένων μοντέλων βάσει της μετρικής ROC-AUC σε όλα τα σύνολα δεδομένων, αναδεικνύοντας τάσεις μεταφοράς γνώσης μεταξύ των μουσικών παραδόσεων που εξετάζονται. Το* `CultureMERT` *γενικεύει αποτελεσματικά σε μη Δυτικά datasets, ενώ η συγχώνευση μοντέλων μέσω task arithmetic επιτυγχάνει αντίστοιχη απόδοση στα ίδια σύνολα και υπερτερεί στα Δυτικά datasets (FMA-medium, MTAT) καθώς και στο Lyra.*

Σε γενικές γραμμές, παρατηρούμε ότι το cross-cultural transferability δεν είναι πάντοτε συμμετρικό. Για παράδειγμα, ενώ το μοντέλο που έχει προσαρμοστεί στην Carnatic μουσι-

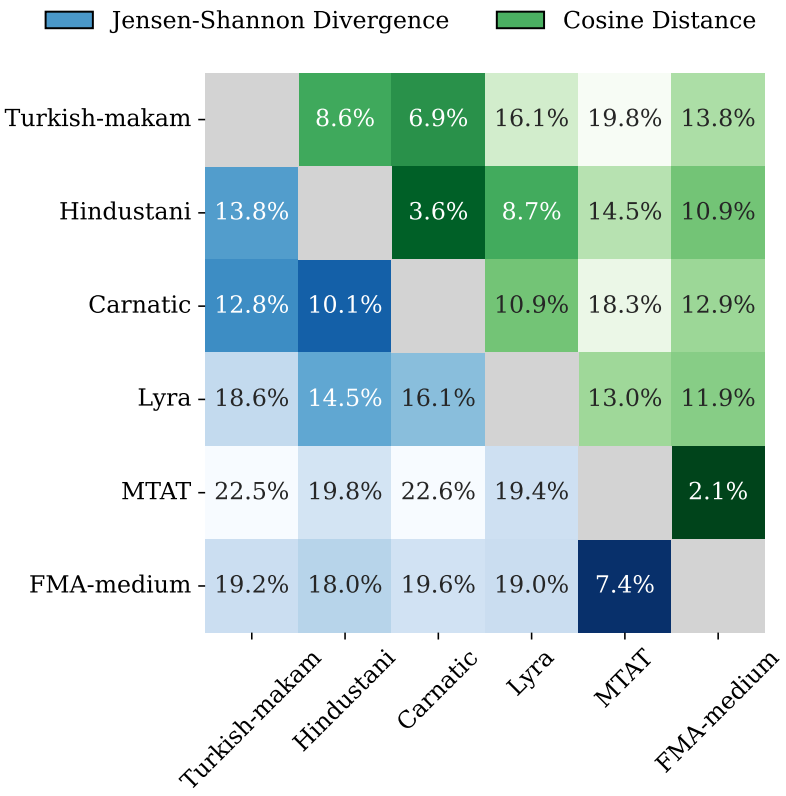
κή γενικεύει ικανοποιητικά στην Hindustani, η αντίστροφη κατεύθυνση οδηγεί σε ελαφρώς καλύτερες επιδόσεις μόνο σε ορισμένες μετρικές (π.χ., Macro-F1). Παρόμοιες ασυμμετρίες έχουν καταγραφεί και στη βιβλιογραφία για το cross-lingual transferability [96, 97, 98]. Ιδιαίτερο ενδιαφέρον παρουσιάζει το γεγονός ότι το μοντέλο που έχει προσαρμοστεί στην Καρνατική μουσική αναδεικνύεται ως το πιο σταθερά transferable μεταξύ όλων των *single-culture* μοντέλων, επιτυγχάνοντας τις υψηλότερες μέσες επιδόσεις σε πολλαπλές μη Δυτικές μουσικές παραδόσεις, σύμφωνα με τις μετρικές ROC-AUC, AP και F1. Εμφανίζει ισχυρή απόδοση όχι μόνο εντός του ευρύτερου Ινδικού πολιτισμικού πλαισίου (ανάμεσα στις παραδόσεις Carnatic και Hindustani), αλλά και γενικεύει αποτελεσματικά προς την Τουρκική κλασσική και την Ελληνική παραδοσιακή μουσική.

Η Ελληνική παραδοσιακή μουσική συνιστά μια ιδιαίτερη «πρόκληση», καθώς ενσωματώνει χαρακτηριστικά τόσο από μη Δυτικές όσο και από Δυτικές μουσικές παραδόσεις. Συγκεκριμένα, φέρει στοιχεία μελωδικού αυτοσχεδιασμού και μικροτονικής έκφρασης, κοινά με τις παραδόσεις της Τουρκικής Μακάμ και της Ινδουστανικής μουσικής, ενώ παράλληλα αξιοποιεί εναρμονίσεις επηρεασμένες από τη Δυτική κλασσική μουσική. Αυτός ο συνδυασμός τροπικών και τονικών συστημάτων έχει αναλυθεί εκτενώς στην εθνομουσικολογική βιβλιογραφία, ιδιαίτερα στο πλαίσιο του Ρεμπέτικου τραγουδιού, το οποίο συνδυάζει μελωδίες βασισμένες σε makam με αρμονικές πρακτικές της Δυτικής μουσικής σκέψης [99]. Στα πειράματά μας, παρατηρούμε ότι το `MERT-v1`, το οποίο έχει αρχικά εκπαιδευτεί σε Δυτική μουσική, ήδη λειτουργεί ως ένα ισχυρό baseline για το Lyra auto-tagging task. Επιπλέον, η περαιτέρω προσαρμογή του μοντέλου — είτε με δεδομένα από μία μόνο μουσική παράδοση είτε από ένα ετερογενές, πολυπολιτισμικό σύνολο — αποφέρει σταθερά τις μικρότερες βελτιώσεις στο Lyra, συγκριτικά με όλες τις άλλες μη Δυτικές παραδόσεις, σε όλες τις μετρικές αξιολόγησης. Το εύρημα αυτό υποδηλώνει ότι η μουσική δομή της Ελληνικής παραδοσιακής μουσικής ενδέχεται να ευθυγραμμίζεται, τουλάχιστον εν μέρει, με τις Δυτικές προκαταλήψεις που φέρει ήδη το αρχικό μοντέλο.

Όπως αναμενόταν, το προ-εκπαιδευμένο μοντέλο `MERT-v1` παρουσιάζει υψηλή απόδοση σε Δυτικοκεντρικά σύνολα δεδομένων, όπως τα MTAT και FMA-medium, γεγονός που ανταναχλά την αρχική προκατάληψή του υπέρ των Δυτικών μουσικών παραδόσεων. Αντιθέτως, τα μοντέλα που έχουν προσαρμοστεί σε επιμέρους μη Δυτικές παραδόσεις (π.χ., `MakamMERT`, `HindustaniMERT` κ.ά.) εμφανίζουν συχνά μειωμένη απόδοση σε αυτά τα Δυτικά benchmarks. Το φαινόμενο αυτό αναδεικνύει το σημαντικό domain shift μεταξύ Δυτικών και μη Δυτικών μουσικών συνόλων δεδομένων. Ωστόσο, η επίδραση αυτή μετριάζεται αισθητά από το `CultureMERT`, και ακόμη περισσότερο από το `CultureMERT-TA`, των οποίων η έκθεση σε ένα ευρύ φάσμα μουσικών παραδόσεων, είτε μέσω συνεχούς προ-εκπαίδευσης είτε μέσω συγχώνευσης μοντέλων, τούς επιτρέπει να διατηρούν αποτελεσματικότερη γενίκευση τόσο σε μη Δυτικές όσο και σε Δυτικές μουσικές παραδόσεις.

Ομοιότητα Μουσικών Παραδόσεων σε Επίπεδο Ακουστικών Tokens. Για να μελετήσουμε σε μεγαλύτερο βάθος τις διαπολιτισμικές ομοιότητες στα δεδομένα μας, αναλύουμε τη συχνότητα εμφάνισης κοινών ακουστικών *tokens* μεταξύ των συνόλων δεδομένων που αντιπροσωπεύουν τις μουσικές παραδόσεις που μελετάμε, χρησιμοποιώντας δύο μετρικές: την *απόκλιση Jensen-Shannon (JSD)* και την *απόσταση συνημιτόνου (cosine distance)* μετα-

ξύ των κατανομών tokens που εξάγονται από το μοντέλο EnCodec [2], το οποίο λειτουργεί ως audio tokenizer. Χαμηλότερες τιμές και στις δύο μετρικές υποδεικνύουν μεγαλύτερη ομοιότητα. Η ανάλυσή μας, όπως παρουσιάζεται στο Σχήμα 6, αποκαλύπτει έντονη ομοιότητα μεταξύ των μη Δυτικών μουσικών παραδόσεων, με ιδιαίτερα υψηλή εγγύτητα μεταξύ της Ινδουστανικής και της Καρνατικής μουσικής. Αντίθετα, τα Δυτικά σύνολα δεδομένων (MTAT, FMA-medium) παρουσιάζουν μεταξύ τους υψηλή ομοιότητα, αλλά εμφανίζουν σημαντικές αποκλίσεις σε σχέση με τις μη Δυτικές παραδόσεις. Η Ελληνική παραδοσιακή μουσική (Lyra), αν και είναι σχετικά πιο διακριτή, φαίνεται να είναι πιο κοντά με τις μη Δυτικές παραδόσεις παρά με τις Δυτικές.

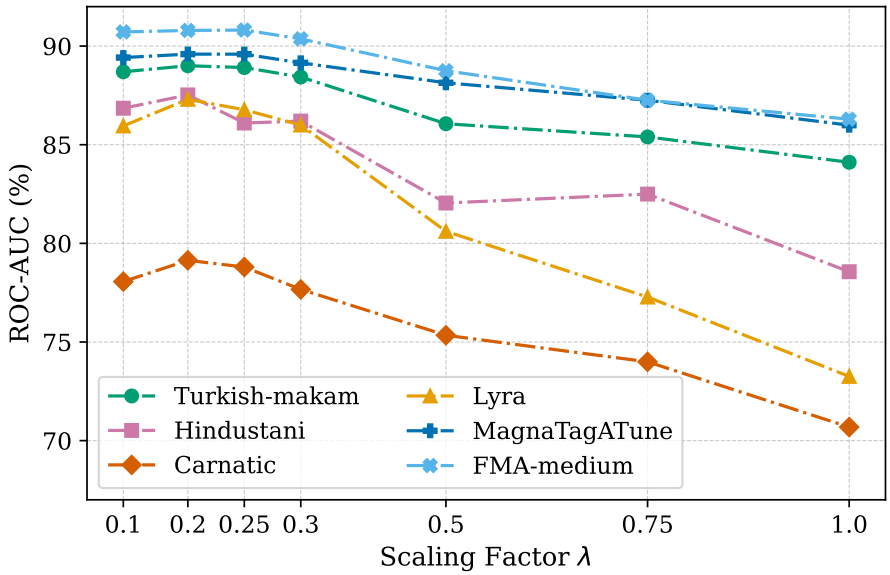


Σχήμα 6. ***Ομοιότητα Ακουστικών Tokens μεταξύ Μουσικών Παραδόσεων.*** *Ζεύγη ομοιότητας μεταξύ των κατανομών ακουστικών tokens, όπως εξάγονται από το EnCodec codec μοντέλο [2]. Οι τιμές ομοιότητας προκύπτουν ως μέσος όρος από 8 codebooks, καθένα εκ των οποίων περιέχει 1024 διακριτά tokens. Και οι δύο μετρικές παρουσιάζουν παρόμοιες τάσεις μεταξύ των dataset.*

Τα παραπάνω ευρήματα παρουσιάζουν ισχυρή συσχέτιση με τα αποτελέσματα διαπολιτισμικής μεταφοράς που αναλύθηκαν προηγουμένως. Αυτό υποδεικνύει ότι μετρικές ομοιότητας μπορούν να αξιοποιηθούν ως ενδείξεις θετικής μεταφοράς γνώσης μεταξύ μουσικών παραδόσεων. Η διαπίστωση αυτή έχει άμεσες πρακτικές προεκτάσεις: οι εν λόγω μετρικές μπορούν να καθοδηγήσουν τόσο την επιλογή όσο και την αναλογία των δεδομένων κατά την εκπαίδευση. Το εύρημα αυτό είναι ιδιαίτερα κρίσιμο σε σενάρια περιορισμένων πόρων, όπου μουσικές παραδόσεις με μικρή διαθεσιμότητα δεδομένων μπορούν να ενισχυθούν μέσω της αξιοποίησης συγγενών, πολιτισμικά συναφών παραδόσεων με μεγαλύτερη διαθεσιμότητα.

0.4.2 Επίδραση του Συντελεστή λ στην Τεχνική Task Arithmetic

Μια κρίσιμη παράμετρος στην τεχνική *task arithmetic* είναι η επιλογή του συντελεστή συγχώνευσης λ. Προηγούμενες μελέτες [6, 94] έχουν δείξει ότι μη βέλτιστες τιμές του λ μπορούν να υποβαθμίσουν σημαντικά την απόδοση κατά τη συγχώνευση. Για να μελετήσουμε την επίδραση του λ, πραγματοποιούμε συστηματική αξιολόγηση διαφορετικών τιμών ($\lambda \in \{0.1, 0.2, 0.25, 0.3, 0.5, 0.75, 1.0\}$), οι οποίες εφαρμόζονται ομοιόμορφα σε όλα τα *task vectors* — δηλαδή, χρησιμοποιείται ο ίδιος συντελεστής για κάθε επιμέρους μοντέλο — σύμφωνα με την απλοποιημένη διατύπωση της Εξίσωσης 2, συμπεριλαμβανομένης και της ειδικής περίπτωσης του *weight averaging* ($\lambda = 1/N = 0.25$). Σε συμφωνία με προηγούμενες παρατηρήσεις από τη βιβλιογραφία, διαπιστώνουμε ότι μη κατάλληλες τιμές, όπως $\lambda = 1.0$, οδηγούν σε χαμηλές επιδόσεις σε όλα τα benchmarks, όπως φαίνεται στο Σχήμα 7.



Σχήμα 7. ***Επίδραση του Συντελεστή Συγχώνευσης*** *λ* ***στην Απόδοση της Τεχνικής Task Arithmetic.*** *Οι τιμές της μετρικής ROC-AUC σε έξι διαφορετικές εργασίες αυτόματης ταξινόμησης μουσικής από ποικίλες μουσικές παραδόσεις αναδεικνύουν πώς η μεταβολή του λ επηρεάζει την απόδοση του task arithmetic κατά τη συγχώνευση των τεσσάρων single-culture adapted μοντέλων.*

0.4.3 Αποτελεσματικότητα της Προτεινόμενης Στρατηγικής Δύο Σταδίων

Παρατηρούμε ότι η απλή συνέχιση της εκπαίδευσης με τον μειωμένο ρυθμό μάθησης του αρχικού προ-εκπαιδευμένου μοντέλου, δηλαδή χωρίς *learning rate re-warming*, οδηγεί σε αναποτελεσματική προσαρμογή σε διαφορετικές μουσικές παραδόσεις, καθώς οι αναπαραστάσεις που έχει μάθει το αρχικό μοντέλο δεν μεταβάλλονται επαρκώς. Αυτό γίνεται ιδιαίτερα εμφανές στην περίπτωση της Τουρκικής Μακάμ μουσικής, όπου η συνεχής προ-εκπαίδευση σε ένα μόνο στάδιο χωρίς re-warming δεν οδηγεί σε καμία βελτίωση απόδοσης. Η εφαρμογή re-warming στο single-stage setup επιφέρει ήπια κέρδη προσαρμογής (+0.8%), αλλά συνοδεύεται από σημαντική απώλεια γνώσης (forgetting) στο MTAT auto-tagging task (−3.6%), ακόμη

και όταν ενσωματώνεται Δυτικά δεδομένα (20%) για την αντιμετώπιση αυτού του φαινομένου. Αντίθετα, η προτεινόμενη two-stage στρατηγική συνεχούς προ-εκπαίδευσης επιτυγχάνει σημαντικά καλύτερη προσαρμογή (+6.4%), ενώ ταυτόχρονα περιορίζει στο ελάχιστο το forgetting (−0.4%), αποδεικνύοντας την αποτελεσματικότητά της στην εξισορρόπηση μεταξύ plasticity και stability. Τέλος, εξετάζουμε τον ρόλο της ενσωμάτωσης δεδομένων που προέρχονται από Δυτικές μουσικές παραδόσεις ως μηχανισμό μετριασμού του φαινομένου catastrophic forgetting. Αν και η χρήση τέτοιων δεδομένων συμβάλλει στη διατήρηση της απόδοσης στο MTAT, εισάγει έναν συμβιβασμό μεταξύ προσαρμοστικότητας και διατήρησης πρότερης γνώσης: υπερβολικό replay Δυτικών δεδομένων μπορεί να εμποδίσει την αποτελεσματική πολιτισμική προσαρμογή. Τα αποτελέσματά μας (Πίνακας 3) υποδεικνύουν ότι ο περιορισμός του Western replay μόνο στο Στάδιο 1 προσφέρει τη βέλτιστη ισορροπία μεταξύ διατήρησης γνώσης και επιτυχούς προσαρμογής.

CPT Strategy	**Western Replay**	**Turkish-makam**	**MTAT**
MERT-v1 (Baseline)	-	83.2	**89.6**
Single-stage (w/ re-warming)	✓	83.8	86.0
Single-stage (w/o re-warming)	✓	83.0	87.5
Two-stage (*Ours*)	Stage 1	**89.6**	89.2
Two-stage (*Ours*)	Both stages	88.6	89.4

Πίνακας 3. ***Σύγκριση Στρατηγικών Συνεχούς Προ-εκπαίδευσης (CPT).*** *Τιμές ROC-AUC στα σύνολα δεδομένων Turkish-makam και MTAT. Η στρατηγική συνεχούς προ-εκπαίδευσης δύο σταδίων υπερτερεί της single-stage προσαρμογής, ενώ ο περιορισμός της ενσωμάτωσης Δυτικών δεδομένων (Western replay) μόνο στο Στάδιο 1 προσφέρει τον βέλτιστο συμβιβασμό μεταξύ πολιτισμικής προσαρμογής και διατήρησης πρότερης γνώσης. Σε όλα τα σενάρια CPT που προσθέτουμε Δυτικά δεδομένα, το 20% των συνολικών δεδομένων εκπαίδευσης προέρχεται από το σύνολο Music4All [7].*

0.5 Συμπεράσματα

0.5.1 Συζήτηση

Η παρούσα διπλωματική εργασία εξετάζει τη μάθηση αναπαραστάσεων για υποεκπροσωπούμενες μουσικές παραδόσεις και προτείνει το `CultureMERT-95M`, ένα πολυπολιτισμικά προσαρμοσμένο foundation model, το οποίο αναπτύχθηκε μέσω συνεχούς προ-εκπαίδευσης (continual pre-training) σε διάφορες μη Δυτικές μουσικές παραδόσεις. Συγκεκριμένα, προτείνουμε μια στρατηγική δύο σταδίων για τη συνεχή προ-εκπαίδευση, η οποία συνδυάζει «επαναθέρμανση» του ρυθμού μάθησης (learning rate re-warming) και σταδιακή προσαρμογή, επιτρέποντας σταθερή εκπαίδευση ακόμη και υπό περιορισμένους υπολογιστικούς πόρους. Τα πειραματικά αποτελέσματα δείχνουν ότι το `CultureMERT` υπερτερεί σταθερά του αρχικού μοντέλου `MERT-95M` σε διάφορες εργασίες ταξινόμησης μουσικής για μη Δυτικά datasets, ξεπερνώντας προηγούμενες state-of-the-art μεθόδους, ενώ ταυτόχρονα διατηρεί υψηλή απόδοση και σε Δυτικά benchmarks.

Επιπλέον, μελετάμε τη μεταφορά γνώσης μεταξύ μουσικών παραδόσεων, αναλύοντας την απόδοση πολιτισμικά εξειδικευμένων μοντέλων σε άλλες μουσικές παραδόσεις. Τα ευρήματα δείχνουν ότι η μεταφερσιμότητα (transferability) διαφέρει ανάλογα με την παράδοση και τα δεδομένα εκπαίδευσης, ανταγακλώντας γνωστές θεωρητικές συγγένειες από την εθνομουσικολογία, ενώ συσχετίζεται επίσης με μέτρα ομοιότητας μεταξύ των συνόλων δεδομένων που χρησιμοποιούμε σε επίπεδο ακουστικών token, προσφέροντας νέες υπολογιστικές προσεγγίσεις για τη χαρτογράφηση των σχέσεων μεταξύ μουσικών παραδόσεων. Η πολυπολιτισμική εκπαίδευση σε ενιαίο σύνολο δεδομένων που ενσωματώνει όλες τις μη Δυτικές παραδόσεις οδηγεί σε συνολικά βελτιωμένη απόδοση, ενισχύοντας τη γενίκευση σε ποικίλα ρεπερτόρια. Επιπλέον, εξετάζουμε τη μέθοδο task arithmetic ως εναλλακτική στρατηγική, συγχωνεύοντας πολιτισμικά εξειδικευμένα μοντέλα στον χώρο των βαρών. Η προσέγγιση αυτή αποδίδει συγκρίσιμα με το `CultureMERT` σε μη Δυτικά datasets, ενώ σε ορισμένες περιπτώσεις ξεπερνά ακόμη και το αρχικό προ-εκπαιδευμένο μοντέλο σε Δυτικά σύνολα δεδομένων.

Συνολικά, η εργασία αυτή συμβάλλει στην ανάπτυξη πολιτισμικά ευαισθητοποιημένων foundation models για τη μουσική και αποτελεί την πρώτη μελέτη που εφαρμόζει και αξιολογεί τεχνικές συνεχούς προ-εκπαίδευσης και συγχώνευσης μοντέλων στο πεδίο της ανάκτησης μουσικής πληροφορίας. Η μελέτη μας εντάσσεται σε μια ευρύτερη προσπάθεια ανάπτυξης υπολογιστικών μεθόδων που σέβονται την πολιτισμική ποικιλομορφία και στοχεύουν στην ενοποιημένη μάθηση καθολικών αναπαραστάσεων της μουσικής, ανοίγοντας προοπτικές για περαιτέρω έρευνα στο πεδίο της διαπολιτισμικής ανάκτησης μουσικής πληροφορίας.

0.5.2 Μελλοντικές Κατευθύνσεις και Προεκτάσεις

Η παρούσα εργασία ανοίγει πολλαπλές προοπτικές για μελλοντική έρευνα και επεκτάσεις. Μία βασική κατεύθυνση αφορά την κλιμάκωση της προσέγγισής μας σε περισσότερες μουσικές παραδόσεις, μεγαλύτερα μοντέλα (π.χ. MERT-330M), καθώς και τη διερεύνηση εναλλακτικών αρχιτεκτονικών, όπως τα MusicFM [19], MuQ [100] και SoniDo [101]. Μελλοντικές μελέτες θα πρέπει επίσης να διερευνήσουν πώς η επιλογή των μουσικών παραδόσεων και οι αναλογίες των αντίστοιχων δεδομένων κατά τη (συνεχή) προ-εκπαίδευση επηρεάζουν τη γενίκευση των μοντέλων σε διαφορετικά tasks. Επιπλέον, η προσέγγισή μας μπορεί να επεκταθεί σε άλλες κατηγορίες εφαρμογών, πέραν της αυτόματης ταξινόμησης μουσικής, όπως beat tracking, emotion recognition και source separation, καθώς και σε πολυτροπικά zero-shot σενάρια και μοντέλα που συνδυάζουν μουσική και φυσική γλώσσα. Για παράδειγμα, η ενσωμάτωση και περαιτέρω εκπαίδευση του `CultureMERT` σε τέτοια frameworks (όπως τα MuLan [95], MusiLingo [102] και CLaMP 3 [71]) θα μπορούσε να ενισχύσει τη γενίκευση σε υποεκπροσωπούμενες μουσικές παραδόσεις, σε εφαρμογές όπως η ανάκτηση μουσικών αποσπασμάτων βάσει φυσικής γλώσσας. Τέλος, η ερμηνευσιμότητα και επεξηγησιμότητα των μοντέλων παραμένει κρίσιμο ζητούμενο. Μελλοντική έρευνα θα μπορούσε να επικεντρωθεί σε τεχνικές επεξηγησιμότητας (XAI) για την κατανόηση και αιτιολόγηση των αποφάσεων των μοντέλων, ιδίως σε πολιτισμικά ευαίσθητα πλαίσια, καθιστώντας τις αποφάσεις της τεχνητής νοημοσύνης πιο διαφανείς και περισσότερο εναρμονισμένες με την ανθρώπινη μουσική αντίληψη.

Chapter 1

Introduction

1.1 Motivation

Music is a fundamental aspect of human culture, universally present across societies while taking diverse forms and expressions unique to each tradition [8, 9, 10]. Its functions include emotional regulation, communication, and social bonding; it plays roles in art, entertainment, worship, and advertising, and it constitutes a major global industry. This dual role, as both a cultural artifact and an economic driver, presents opportunities to benefit society while also posing unique technical challenges when combined with artificial intelligence (AI) [11]. Beyond practical applications, comprehension of music's semantics through deep learning (DL), particularly emphasizing interpretable approaches, can also significantly contribute to theoretical insights across various fields, including ethnomusicology and music anthropology, music theory, and the study of music cognition. For instance, by analyzing vast amounts of music data, computational models can uncover musical patterns and cultural influences in music evolution, and relate these findings to broader social and historical contexts. Furthermore, while music is often described as a "universal language", this notion remains debated among scholars: certain elements may transcend cultural boundaries [12], yet musical traditions have evolved with distinct characteristics and culturally grounded semantics [13, 14]. This interplay between universality and cultural specificity poses a complex challenge that music informatics and modern AI approaches can provide a new perspective to explore [15].

Music information retrieval (MIR) refers to the field of research focusing on extracting and analyzing information from music data [16, 11]. Computational methods in music typically employ signal processing techniques to extract relevant features from audio signals, which are then utilized by machine learning (ML) models for music understanding tasks, such as genre classification, beat tracking, key detection, source separation, and music auto-tagging, among other tasks. Unlike speech and language, music is typically polyphonic, often comprising multiple concurrent "voices" or instrumental layers, which makes it fundamentally a multi-stream signal. Moreover, musical "meaning" is not usually grounded in direct references to real-world objects or concrete events, but is instead abstract and often shaped by cultural context. Thus, music understanding presents significant challenges due to the intricate and interwoven human-related concepts embedded within the sequence of music signals, such as emotions, experiences, expressions, cultural

identity, societal and historical contexts, communication, and creativity. Additionally, music typically has a much longer duration and a higher sample rate compared to speech or general audio, making it computationally demanding to model entire pieces effectively. The key bottleneck is that modeling raw audio directly introduces extremely long-range dependencies, making it computationally challenging to learn the high-level semantics of music.

The term "foundation model" (FM) was introduced to refer to any pre-trained, versatile machine learning architecture that, rather than being optimized for just one specific purpose, functions as a central framework from which multiple specialized models can be derived for a wide variety of downstream tasks [17]. The emergence of foundation models has been driven by advancements in deep learning, including architectural innovations such as the Transformer [18], as well as improvements in computational hardware. Foundation models have recently emerged in the music domain [1, 19, 20, 11, 103, 58], offering powerful general-purpose representations learned from large-scale audio data. These models capture broad musical characteristics and have demonstrated state-of-the-art performance across a range of music understanding tasks, reducing the need for task-specific training. By leveraging self-supervised learning (SSL) on large amounts of unlabelled music data, foundation models address data scarcity, reduce annotation costs, and improve generalization in music information retrieval [11]. In general, SSL has emerged as a promising paradigm in representation learning, enabling models to learn meaningful representations from large unannotated datasets, without requiring explicit labels, by leveraging the inherent structures present in the data [104]. Similarly, it facilitates the extraction of generalizable knowledge from extensive unlabelled datasets, thereby enhancing model performance on downstream tasks where labeled data is scarce.

Despite these advances, most existing foundation models for music have been trained primarily on Western-centric datasets, limiting their ability to represent diverse musical styles [21, 3]. Critically, these models are also rarely evaluated on the world's musical diversity, leaving their generalization ability to diverse musical traditions, especially underrepresented ones, largely unexplored. Many musical traditions, including Turkish, Indian, and Greek traditional music, feature unique melodic structures, modal or tonal systems, and rhythmic patterns that are not adequately captured by these models [22, 23, 24]. Unlike Western classical and popular music, which primarily rely on equal temperament and harmony-based composition, these traditions incorporate microtonal intervals, distinct rhythmic cycles, and melodic improvisation. Failing to model such culture-specific stylistic elements not only narrows the applicability of music foundation models, for example, in region-specific recommendation systems [25] or cultural heritage preservation, but also overlooks rich, culturally specific knowledge crucial for advancing MIR research [11]. This bias reflects a broader historical trend of Westernisation, where the dominance of Western music in computational models risks marginalizing and displacing local traditions. Accordingly, there is an urgent need to develop more inclusive and culturally aware computational models [26], capable of generalizing beyond Western-centric traditions and adapting effectively to diverse underrepresented musical cultures, a direction that has also gained traction in other domains, such as natural language processing (NLP) [27] and

speech recognition [28], through culturally adapted and multilingual foundation models.

One promising avenue for addressing these challenges is continual pre-training (CPT), which has emerged as an effective and increasingly popular approach in large language models (LLMs) [4, 29, 30, 105, 32, 88, 33, 34, 106, 107, 108] and multimodal learning [31]. By enabling models to incrementally adapt to new domains, tasks, or languages, CPT avoids the need for full re-training, which is often impractical and computationally expensive [32, 4, 88, 106, 37]. Notably, it has been shown to match, or even surpass, training from scratch in some cases [33, 34], while also leading to faster convergence [35] and mitigating catastrophic forgetting [36]. CPT has also gained traction in the audio domain, with recent work demonstrating its effectiveness in adapting pre-trained speech models to both high- and low-resource languages [37, 28, 109, 38, 110].

Additionally, model merging [39, 40, 111, 112] has proven to be a simple yet effective technique for adapting pre-trained models across multiple domains by combining domain-specific parameters in weight space, without requiring additional training [41] or access to the original training data [42]. A notable method within this paradigm is task arithmetic (TA) [5], which constructs *task vectors* by computing the difference between the parameters of an adapted model and its pre-trained counterpart, encoding domain-specific knowledge. These task vectors can then be integrated into the pre-trained model via algebraic operations in Euclidean space to create a unified model from multiple independently adapted models, offering a computationally efficient alternative to multi-task learning (MTL) [6]. Task arithmetic provides a modular framework for editing pre-trained models and fusing knowledge across domains, enabling generalization across diverse tasks while preserving information from both individual adaptations and the original pre-training.

Given the scarcity of culturally diverse, annotated music data, CPT provides a computationally efficient solution for adapting foundation models to non-Western traditions without requiring full re-training, while task arithmetic enables seamless model merging in weight space, facilitating multi-cultural adaptation while mitigating catastrophic forgetting.

1.2 Research Objective and Contributions

While both continual pre-training and task arithmetic have been widely explored in other domains, their application to MIR remains largely unexplored. We bridge this gap by leveraging these techniques to adapt the `MERT-v1-95M`[1] music foundation model [1], originally trained on 1K hours of predominantly Western music [1, 43], to diverse musical cultures from the Eastern Mediterranean and the Indian subcontinent, while preserving performance on "Western"-centric benchmarks.

This process can be naturally framed as a domain adaptation (DA) [113] task, in which a model trained on a Western source domain is adapted to perform effectively on culturally distinct target domains. In this context, continual pre-training and model merging can be viewed as strategies within the broader DA framework. While conventional DA approaches

[1] https://huggingface.co/m-a-p/MERT-v1-95M

often involve aligning feature distributions or fine-tuning on the target domain, CPT entails further training a pre-trained model, typically using the same self-supervised objective, on domain-specific data to incrementally adapt it to a new target domain, whereas TA enables multi-domain merging directly in parameter space without access to the original training data.

A major challenge in adapting foundation models to diverse domains is ensuring efficient adaptation while avoiding catastrophic forgetting [44], where previously learned knowledge can be "lost" when the model is trained on new data [45]. To address this, we propose a computationally efficient two-stage continual pre-training approach that integrates learning rate re-warming [4] and staged adaptation, stabilizing training and enabling smoother, effective adaptation.

This thesis makes the following contributions:

1. To the best of our knowledge, this is the first study to explore **continual pre-training** and **task arithmetic** for cross-cultural adaptation in MIR, demonstrating their effectiveness in music audio representation learning within the context of music foundation models.

2. We propose a **two-stage continual pre-training strategy** that stabilizes training, mitigates catastrophic forgetting, and facilitates effective adaptation under constrained computational resources.

3. Our multi-cultural model, `CultureMERT`, outperforms the original `MERT-v1` by an average of **4.43%** in ROC-AUC across diverse non-Western music tagging tasks, alongside consistent average improvements of 5.4% in AP, 3.6% in Micro-F1, and 6.8% in Macro-F1, while exhibiting minimal forgetting on Western benchmarks.

4. Our culturally adapted models **surpass previous state-of-the-art** results across all evaluated non-Western music tagging tasks.

5. We explore **cross-cultural transferability**, analyzing how models adapted to one musical tradition (e.g., Turkish Makam) generalize to others (e.g., Greek folk). Our results indicate that single-culture adaptations exhibit varying degrees of transfer across cultural domains, with the multi-culturally adapted model yielding the best generalization across cultures.

By addressing these challenges, this thesis contributes to the development of culturally aware foundation models for music that enable world music understanding and enhance cross-cultural music representation learning. Our work highlights the efficacy of continual pre-training for cross-cultural adaptation in MIR, establishing `CultureMERT-95M` as a state-of-the-art foundation model for diverse musical traditions. To support reproducibility and further research on world music representation learning, we publicly release `CultureMERT-95M`, along with the task arithmetic variant, `CultureMERT-TA-95M`.

1.3 Outline

The remainder of this thesis is organized as follows:

- Chapter 2 reviews relevant literature on music representation learning and foundation models for music, discusses MIR tasks and evaluation paradigms, and examines prior work and challenges in cross-cultural MIR.
- Chapter 3 presents the methodology of this thesis, including a detailed overview of the MERT architecture and its pre-training objective, the proposed two-stage continual pre-training strategy for cultural adaptation, and task arithmetic as an alternative approach to multi-cultural adaptation through model merging. This chapter also describes the datasets used, the experimental setup, and the probing-based evaluation protocol.
- Chapter 4 presents the experimental results and provides an in-depth discussion. It evaluates the impact of continual pre-training and task arithmetic on culturally diverse music tagging tasks, analyzes cross-cultural transferability patterns, and demonstrates the effectiveness of our proposed approach through detailed ablation studies.
- Chapter 5 concludes the thesis by summarizing the main findings, highlighting the limitations of the current study, and outlining promising extensions, directions, and research avenues for future work.

Chapter 2

Music Information Retrieval

Music information retrieval (MIR) is a multidisciplinary field that focuses on the computational analysis, organization, and manipulation of music-related data [46]. The term MIR is sometimes used interchangeably with *music informatics* or *music information processing* [47]. It encompasses a wide range of tasks, including indexing, retrieval, recommendation, transcription, source separation, and music generation. In recent years, MIR research has become increasingly driven by advances in machine learning (ML), particularly deep learning (DL) methods, which have enabled significant progress across many subfields. In this chapter, we first review recent advances in music audio representation learning and foundation models for music, followed by a discussion of MIR tasks and evaluation paradigms. We conclude with an overview of cross-cultural MIR, highlighting emerging datasets, methodologies, and ongoing challenges in this important area.

2.1 Music Representation Learning

The objective of representation learning is to discover and learn meaningful features that facilitate downstream tasks while remaining robust to the complex variations inherent in natural data [114]. Representation learning has gained popularity across many domains due to its effectiveness, computational efficiency, and the simplicity of reusing pre-trained model representations as features for a wide range of downstream tasks. This objective is central to advancing artificial intelligence (AI), as learning good representations reduces the need for task-specific engineering and enables knowledge transfer across tasks. A variety of model architectures, training paradigms, and modalities have been employed to learn representations that perform well on tasks such as classification, retrieval, and generation. In the context of music informatics, such learned representations can be leveraged for downstream MIR tasks such as automatic classification, recommendation, generation, and emotion recognition, among others. Ideally—especially in the case of foundation models—these representations should capture broad musical features such as rhythm, melody, dynamics, timbre, pitch, and harmony, as well as higher-level abstractions, including compositional structure, arrangement, and cultural context. This makes them essential for models designed to support a wide range of music understanding tasks, where a deep, holistic representation of musical content is crucial.

Early music information retrieval relied on hand-crafted features (e.g., MFCCs, chroma

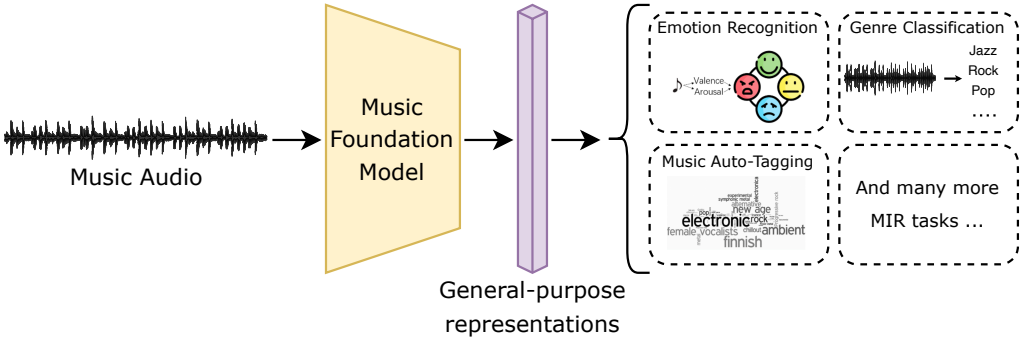


Figure 2.1. *Foundation models for music learn general-purpose representations through large-scale pre-training, which can then be transferred to a wide range of downstream MIR tasks.*

features, or constant-Q representations) and classical machine learning algorithms. However, deep learning has since revolutionized music representation learning, achieving remarkable success [48]. In particular, convolutional neural networks (CNNs) trained on labeled music tasks, such as music auto-tagging and genre classification, have become widely adopted. For example, musicnn [115] offers pre-trained VGG-like CNNs that achieve high accuracy on benchmarks such as MagnaTagATune. Transfer learning [116, 117] has further improved performance by fine-tuning models pre-trained on large audio corpora, significantly boosting results on downstream MIR tasks including instrument classification and genre recognition. This approach also enables effective use of smaller annotated datasets, which is particularly advantageous given the high cost and difficulty of manual labeling, especially in diverse or low-resource musical domains. More recently, an alternative strategy for learning music representations leverages self-supervised learning (SSL), where models are trained on proxy objectives derived directly from the input data, eliminating the need for manual annotation. This approach enables the extraction of rich representations using automatically generated supervision signals. Several SSL-based models for music have demonstrated strong performance across a range of downstream MIR tasks, effectively closing the gap with supervised approaches [49, 50, 51, 52, 1, 53].

2.2 Foundation Models for Music

Recent research has increasingly focused on developing foundation models for music [11, 103], inspired by analogous advances in natural language processing (NLP) and computer vision [17]. These models are typically pre-trained on large-scale music corpora using SSL, learning general-purpose representations that can transfer effectively to a wide range of MIR tasks [11] (see Figure 2.1). After pre-training, they are fine-tuned and evaluated on downstream tasks using much smaller labeled datasets, enabling effective transfer even in low-resource or annotation-scarce scenarios. Notably, [101] shows that intermediate features extracted from such pre-trained models can serve as general-purpose boosters for task-specific models, particularly when training data is limited or computational resources are constrained.

One dominant pre-training paradigm is *masked modeling (MM)*, adapted from BERT-style pre-training in NLP [54]. MM works by randomly masking portions of the input and training the model to infer the missing segments from surrounding context. This approach, commonly employed in Transformer models, allows for effective modeling of sequential data and long-range dependencies, as in the case of music audio signals. Music audio, typically sampled at rates up to 48 kHz and spanning several minutes or more, results in extremely long sequences that pose significant computational challenges, particularly during model training. To address this, models employ tokenizers to compress audio signals into compact latent representations with shorter sequence lengths.

In general, audio tokenization refers to the process of transforming raw audio into sequences of "meaningful" units (i.e., tokens or codes) that can be used for language modeling (e.g., masked prediction or auto-regressive generation), downstream tasks (e.g., text-to-speech synthesis [118, 119]), or audio compression. Unlike text, where tokens are inherently discrete (e.g., words, subwords, or characters), audio is a continuous waveform that lacks natural segmentation boundaries. Earlier approaches relied on handcrafted features such as spectrograms and cepstral coefficients (e.g., MFCCs), whereas more recent methods learn token representations directly from data using models like vector-quantized autoencoders [2, 120, 121], which primarily capture acoustic representations, or self-supervised audio encoders [1, 122, 123, 124], which aim to learn higher-level semantic representations.

Therefore, input sequences are typically tokenized into either continuous embeddings or discrete integer tokens. The latter are often produced by neural audio codecs (NACs), learnable vector-quantized models that discretize audio into compact, semantically meaningful tokens at specific temporal resolutions [125]. These tokens can serve as inputs or targets for masked prediction and generation tasks in music foundation models [11]. Notable examples include EnCodec [2], SoundStream [121], Descript Audio Codec (DAC) [120], SpeechTokenizer [126], X-Codec [126], and, more recently, SemantiCodec [124], Pyramid-Codec [127], and SAT [128], which combine fine-grained acoustic fidelity with the ability to capture long-range semantic information via hierarchical tokenization. Depending on the tokenization strategy, the masked prediction objective is formulated as either a regression task, in the case of continuous masked modeling where models predict continuous targets using typically the mean squared error (MSE) loss, or as a classification task, as in discrete masked modeling where models predict token indices, typically using cross-entropy losses. Finally, the choice of tokenization strategy can significantly impact the trade-off between training efficiency, data requirements, and model capacity, an important yet underexplored aspect in the development of music FMs. In particular, the segmentation of raw audio into tokens is highly non-trivial, and future work may benefit from exploring adaptive, content-aware tokenization schemes that better capture the temporal variability of music (see Section 5.3 for a suggested research direction on this topic).

Overall, recent work highlights a clear trend: neural audio codecs—and audio tokenization more broadly—play a central role in state-of-the-art music foundation models by providing compact, high-fidelity representations that support both language model-based audio generation and understanding tasks, aligning the training paradigm of audio models with the masked or auto-regressive language modeling paradigms used in NLP.

Below, we review key foundation model architectures that have driven recent progress in music audio representation learning, focusing on their pre-training paradigms (e.g., masked modeling, multimodal learning, generative modeling), architectural choices, and audio tokenization strategies.

MERT [1] (Music undERstanding model with large-scale self-supervised Training) is a prominent example, a BERT-style encoder built upon the HuBERT architecture [55] and pre-trained on large-scale music audio using a masked language modeling (MLM) objective. MERT employs a dual-teacher strategy to generate pseudo-labels: an RVQ-VAE-based "acoustic" teacher (specifically, the EnCodec audio tokenizer [2]) and a "musical" teacher based on constant-Q transform (CQT) reconstruction. This combination enables the model to learn both acoustic and harmonic characteristics of music. Available in 95M and 330M parameter variants, MERT achieves state-of-the-art performance across 14 diverse music understanding tasks, demonstrating the effectiveness of large-scale self-supervised pre-training for unifying multiple MIR tasks in a single model. For further architectural and pre-training details, refer to Section 3.2.

Music2Vec [123] is a lightweight and efficient self-supervised framework for learning music audio representations directly from raw waveforms. Inspired by the data2vec architecture [129], it adopts a teacher–student setup where both models share the same architecture, comprising a 1-D CNN feature extractor followed by a 12-layer Transformer encoder. The CNN maps 16 kHz audio into 50 Hz feature sequences, which are then passed through the Transformer. The student is trained on masked input segments to predict the contextualized representations generated by the teacher model, which are derived from the outputs of all 12 Transformer layers. Importantly, the teacher's parameters are updated via exponential moving average of the student's weights. Music2Vec is trained from scratch on 1k hours of music audio and achieves competitive performance on several MIR tasks, including music auto-tagging, genre classification, emotion regression, and key detection, despite having fewer than 2% of the parameters of large models like Jukebox [20].

MusicFM [19] builds upon MERT by replacing its learned tokenization mechanism with a non-trainable random projection quantizer, inspired by BEST-RQ [56]. This approach removes the need for a separate representation learning stage (e.g., RVQ or k-means clustering), as the tokenization process is entirely training-free. Specifically, MusicFM maps log-mel spectral features into a latent space via random projection and discretizes them using a randomly initialized codebook. Despite its simplicity, this tokenization strategy achieves strong performance—particularly when sufficient training data is available—across both sequence-level and token-level MIR tasks, including beat tracking, chord recognition, and music tagging. These results suggest that random quantization can be a computationally efficient and effective alternative to learned audio tokenizers for foundation models in music informatics.

MuQ [100] is a self-supervised foundation model for music audio representation learning that introduces a novel tokenization method, Mel residual vector quantization (Mel-RVQ). Unlike prior models that rely on random projections (e.g., MusicFM) or heavy neural codecs (e.g., EnCodec in MERT), MuQ employs a lightweight residual quantizer trained on Mel-spectrograms using simple linear projections. This approach provides both computational efficiency and stable token generation, addressing the initialization sensitivity of random quantizers and the resource demands of neural codecs. The model is trained to predict these discrete Mel-RVQ tokens using a Conformer-based encoder under a masked language modeling objective. Despite using only 0.9K hours of pre-training data, MuQ outperforms larger models like MERT and MusicFM on a wide range of downstream MIR tasks, including genre classification, instrument and key detection, and music structure analysis. Further scaling to 160K hours and incorporating iterative training improves performance even more. Additionally, MuQ supports multimodal extension through MuQ-MuLan, a joint music-text embedding model that achieves state-of-the-art results on the MagnaTagATune zero-shot music tagging task.

MULE [130] presents a comparative study of supervised and unsupervised strategies for pre-training audio models specifically for music understanding. It explores how dataset domain (music vs. general audio) and training strategy (supervised vs. unimodal contrastive unsupervised) affect the transferability of audio embeddings across a wide range of music tasks. Notably, it demonstrates that supervised training on large-scale expert-annotated music data achieves state-of-the-art performance across diverse tagging tasks, while unsupervised learning on in-domain music audio yields highly generalizable embeddings with strong performance on novel tasks.

Generative Pre-Training Generative modeling has also shown promise in music representation learning. A seminal example is Jukebox [20], a large-scale (5B parameter) auto-regressive Transformer model trained on over 1.2 million songs. Jukebox compresses raw audio into discrete codes using three separately trained VQ-VAEs at different temporal resolutions (each with a vocabulary size of 2048). To model long-range musical structure, separate auto-regressive priors are trained for each level, and audio is generated hierarchically from coarse to fine resolution, with each level conditioned on the upsampled codes of the coarser level. The model enables controllable generation by conditioning on metadata such as genre, artist, timing, and optionally unaligned lyrics. While primarily designed for music generation, JukeMIR [57] demonstrated that intermediate representations from Jukebox can be repurposed for music understanding tasks, achieving strong performance in music auto-tagging, genre classification, key detection, and emotion recognition.

Building upon this, SoniDo [101] proposes a two-stage hierarchical foundation model tailored for both music understanding and generation. It uses a hierarchically quantized VAE (HQ-VAE) in the first stage to extract coarse-to-fine latent tokens, and models their distribution in the second stage using a stack of sparse Transformer decoders conditioned on CLAP embeddings for multimodal alignment. Importantly, the token hierarchy in SoniDo is jointly trained, unlike Jukebox, where each level is independently trained, which

allows richer inter-level dependencies. SoniDo extracts task-agnostic representations from intermediate Transformer layers, supporting a broad range of downstream tasks including music auto-tagging, music transcription, music source separation, and music mixing.

Furthermore, while originally developed as a text- and melody-conditioned music generation model, MusicGen [58] has also been evaluated for music representation learning. It employs a single-stage auto-regressive Transformer decoder trained on residual vector-quantized (RVQ) tokens from EnCodec. In contrast to models like Jukebox or SoniDo, which use multi-level hierarchies with multiple priors, MusicGen processes flattened or interleaved parallel token streams from multiple codebooks. This design simplifies training and improves efficiency by avoiding the computational costs of generating multiple codebook streams. As demonstrated in the SoniDo paper, intermediate representations from MusicGen can be repurposed for downstream MIR tasks such as genre classification, key detection, and music transcription.

Additionally, AudioLM [131] introduces a multi-stage auto-regressive framework that models long-term structure and fine-grained detail by combining semantic tokens extracted from w2v-BERT (k-means-clustered representations) [122] with coarse and fine acoustic tokens derived from the SoundStream neural audio codec. Originally proposed for speech and piano music continuation, AudioLM achieves coherent generation over long timescales without requiring textual supervision, showcasing the power of hybrid tokenization in audio foundation models. Building on this framework, MusicLM [132] extends AudioLM by incorporating text conditioning via MuLan embeddings, enabling high-fidelity and semantically aligned music generation from textual descriptions. MusicLM demonstrates improvements in both audio quality and adherence to text prompts, and supports additional conditioning modalities such as melody inputs.

Most recently, YuE [133] introduced a family of open-source foundation models for long-form lyrics-to-song generation. Built on the LLaMA2 architecture and trained on trillions of tokens, YuE uses a two-stage hierarchical modeling framework with semantic-acoustic fused tokenization and track-decoupled next-token prediction. The model enables up to five-minute coherent generation with fine-grained control over lyrics, structure, and style, while also demonstrating strong performance in music understanding tasks such as the MARBLE benchmark, matching or exceeding previous state-of-the-art methods.

These foundation models fall under the umbrella of auto-regressive predictive coding (APC), a pre-training paradigm where a model learns to predict future tokens (i.e., discrete audio codes) in a sequence using an auto-regressive architecture.

Multimodal Learning Multimodal approaches extend music foundation models by incorporating modalities beyond audio, such as natural language. Typically, the text modality is processed using pre-trained embeddings from large NLP encoders or decoders. For example, LLark [134] is an instruction-tuned multimodal model that integrates a pretrained Jukebox-based audio encoder with a LLaMA-2 language model via a projection module, following a prefix tuning strategy. It supports flexible (audio, text) input prompts and produces text outputs, enabling tasks like music captioning, genre identification, and compositional reasoning. Evaluated on music understanding, captioning, and reasoning tasks,

LLark achieves strong zero-shot performance, showcasing the benefits of instruction-tuned training on open-source music datasets. Similarly, the JMLA model [135] also targets zero-shot music tagging via a joint music and language attention mechanism to address the *open-set* music tagging problem. It connects a pre-trained masked audio encoder to a Falcon-7B decoder using a perceiver resampler and dense cross-attention layers across encoder-decoder layers, allowing multi-level semantic information exchange. By leveraging GPT-processed music descriptions for training, JMLA achieves competitive zero-shot performance across multiple benchmarks, demonstrating the effectiveness of tight cross-modal alignment in music tagging.

Furthermore, MusiLingo [102] bridges music and language with prefix tuning by aligning MERT music embeddings with a frozen large language model (Vicuna-7B) via a linear projection, enabling captioning and instruction-following for music-related queries. Mustango [80] further explores text-to-music generation through controllable diffusion modeling, using musically enriched prompts to guide generation via chord, tempo, and key information. LTU (Listen, Think, and Understand) [136] is a multimodal instruction-following model for general audio understanding. It integrates a CAV-MAE-pretrained AST audio encoder with a Vicuna-7B language model via LoRA adapters, and is trained using a four-stage perception-to-understanding curriculum on the OpenAQA-5M dataset, which combines closed- and open-ended (audio, question, answer) pairs. LTU demonstrates strong generalization across audio tasks and exhibits emerging reasoning abilities, including step-by-step inference, explanation, and uncertainty awareness.

Qwen-Audio [137] and its successor Qwen2-Audio [138] are large-scale audio-language models that integrate a pre-trained audio encoder with a frozen large language model to support universal audio understanding. Both models adopt a multi-task training strategy across over 30 speech, sound, and music tasks, using next-token prediction. To mitigate interference from heterogeneous datasets, Qwen-Audio conditions the decoder on hierarchical label sequences to balance shared knowledge with task-specific signals, while Qwen2-Audio replaces these with natural language prompts to improve generalization and alignment. It scales up training data and demonstrates strong performance on tasks such as genre classification, instrument recognition, and emotion description, achieving state-of-the-art results on the AIR-Bench music subset without task-specific fine-tuning.

Additionally, contrastive learning-based (CL) models such as MuLan [95], MusCALL [139], and CLAP [140] align audio and text representations into a shared embedding space using paired audio-text data, enabling zero-shot tagging and cross-modal retrieval. MuLan and CLAP scale to large datasets and leverage strong pre-trained encoders, while MusCALL emphasizes lightweight training and robust contrastive alignment, introducing a content-aware loss and exploring audio self-supervision through SimCLR. By leveraging natural language supervision instead of fixed label taxonomies, these models offer greater flexibility and semantic grounding for downstream MIR tasks.

In addition to text, some multimodal models also incorporate visual information, such as album artwork, music video frames, or other associated visuals, as in recent models like M^2UGen [141], V2Meow [142], and VidMuse [143], to bridge the gap between visual and audio modalities. Finally, AnyGPT [144] further pushes the boundary by enabling

any-to-any multimodal generation across music, speech, image, and text using a unified framework based on discrete tokenization. By leveraging modality-specific tokenizers and a shared language model backbone (LLaMA-2 7B), AnyGPT achieves zero-shot generalization across diverse modality combinations without architectural changes.

We should note that while foundation models for music can be designed to process various modalities—such as symbolic music representations (e.g., MIDI, ABC notations, tokenized sequences), sheet music, and others—this thesis focuses primarily on music **audio** representation learning, where the musical modality is the raw or preprocessed audio signal. Therefore, our emphasis is on models that operate directly on audio input, aiming to learn representations that capture both low-level acoustic features and high-level musical semantics. This focus aligns with recent trends in MIR research and self-supervised learning [11]. Nonetheless, we acknowledge that many non-audio musical modalities remain underexplored in the development of foundation models [11], and future work could benefit from addressing this gap. For example, ChatMusician [145] is a recent LLM trained on ABC notation—a compact, text-compatible symbolic music format—that demonstrates strong symbolic music understanding and generation capabilities. Built on LLaMA 2 via continual pre-training and fine-tuning, it shows that treating music as a second language allows LLMs to reason about music theory, compose structured scores, and outperform GPT-4 on a college-level symbolic music benchmark.

Future work may benefit from integrating such symbolic modalities with audio-based models in a multimodal setting, combining the structural "clarity" of symbolic representations with the expressive nuance and dynamics of raw audio signals. This could involve developing unified multimodal representations that incorporate symbolic music, audio, text, and music score images within FM architectures [146]. Such fusion may enable models to better capture the multifaceted nature of music by leveraging diverse sources of information and mitigating the limitations of relying on a single modality. Very recently, the UniMuLM framework [147] was proposed to address this challenge by unifying symbolic music, waveform audio, and textual instructions into a single language model. It introduces a bar-level tokenizer that explicitly aligns symbolic and waveform representations. UniMuLM demonstrates strong performance across diverse music tasks, including captioning, continuation, inpainting, and music question answering, underscoring the potential of multimodal integration for advancing music understanding and generation. Similarly, Seed-Music [148] is a unified multimodal music generation framework that combines text, audio, and symbolic modalities. It combines MuLan-based text embeddings (including phonemes for lyrics), an auto-regressive (LM-based) audio token generator, and a symbolic lead sheet generator to produce high-quality, controllable music conditioned on multimodal inputs. In parallel, incorporating domain knowledge, such as music theory, notation, and structural semantics, into foundation model design could guide learning toward musically meaningful abstractions, potentially moving beyond architectural conventions inherited from other domains.

2.3 MIR Tasks and Evaluation

Music information retrieval encompasses a wide range of tasks involving both audio and symbolic music. As noted earlier, this thesis primarily focuses on the raw audio signal as the musical modality. Common MIR tasks[1] include:

- **Tonality and Harmony:** mode recognition, chord recognition, key estimation.
- **Melody and Pitch:** melody extraction, pitch and multi-pitch estimation, note tracking, automatic music transcription.
- **Rhythm:** onset detection, beat and downbeat tracking, metre estimation, tempo estimation.
- **Timbre and Instrumentation:** musical instrument identification, playing technique detection.
- **Temporal Alignment:** score following, audio-to-score alignment.
- **Temporal Segmentation:** music/non-music detection, structural segmentation/structure analysis, time boundary identification.
- **Source Separation:** musical instrument source separation, harmonic-percussive source separation.
- **Performance-related Understanding:** performer identification, technique identification, performance assessment, difficulty estimation.
- **Clip-level Classification:** auto-tagging[2], genre classification, mood/emotion recognition.
- **Retrieval and Similarity:** audio identification, audio matching, cover song detection.
- **Vocal Understanding:** singer identification, vocal technique detection, automatic lyrics transcription and alignment, singing transcription, vocal source separation, lyrics interpretation.
- **Multimodal Understanding:** cross-modal retrieval and recommendation, music captioning, music instruction following, music question answering.
- **Music Generation:** text-to-music generation, symbolic music generation, monophonic/polyphonic music generation, conditioned generation (chord sequences, melody, video, text descriptions), melody harmonization, lyrics to singing, singing voice synthesis, singing voice conversion.

[1] This list is indicative; constructing a complete MIR taxonomy is challenging and beyond the scope of this thesis.

[2] Automatic music tagging, which refers to the automatic assignment of descriptive metadata to audio tracks, encompassing attributes such as genre, mood, instrumentation, tempo, language, and even contextual information such as geographic location, is typically referred to as *music auto-tagging*.

MIR-based evaluation can be broadly categorized into two paradigms: **probing-based** and **language-based** evaluation. The probing setup is typically employed for unimodal models that learn audio representations, where the audio encoder is treated as a feature extractor—either frozen or fine-tuned—and a lightweight probing head, typically a shallow multilayer perceptron (MLP), is trained on top with labeled data to perform the downstream task. In contrast, multimodal models that incorporate both audio and text modalities can be evaluated via natural language prompting: the model is provided with a task-specific instruction (e.g., *"What is the key of this song?"*), and its generated response (e.g., *"This song is in F minor."*) is mapped to a corresponding label for scoring—typically in the case of closed-ended tasks.

Recently, several promising evaluation protocols have emerged for benchmarking learned music representations. The most comprehensive among them is the MARBLE protocol [149], which introduces a unified framework for evaluating music audio representations across a wide range of MIR tasks. MARBLE organizes its evaluation into a four-level hierarchical taxonomy: (i) **acoustic-level**, encompassing tasks such as singer identification, instrument classification, and source separation that focus on low-level signal features; (ii) **performance-level**, targeting expressive elements like vocal techniques and ornamentation; (iii) **score-level**, covering tasks such as pitch tracking, beat tracking, melody extraction, chord estimation, and lyrics transcription; and (iv) **high-level description**, which includes tasks like key detection, genre classification, music tagging, and emotion recognition. The benchmark spans 18 tasks across 12 publicly available datasets and aims to provide a standardized, reproducible, and fair evaluation protocol for pre-trained music models.

In the domain of multimodal foundation models, which extend beyond traditional MIR tasks, a broad spectrum of evaluation tasks has also been explored to assess music understanding. A particularly prominent one is **cross-modal retrieval**, which serves as a standard benchmark for evaluating alignment capabilities across modalities. This task involves retrieving samples in one modality (e.g., audio clips) based on a query in a different modality (e.g., text or video), thereby testing the model's ability to associate and align semantically related content across modalities. Another complementary evaluation approach centers around language generation, often used for open-ended tasks. This paradigm is suited to models that take audio or audio-text pairs as input and generate free-form textual outputs. Typical tasks include music captioning, commonly evaluated on the MusicCaps dataset [150] or custom ad-hoc datasets, and music question answering, where performance is assessed using automatic metrics (e.g., BLEU, METEOR, ROUGE), human evaluation, or large language model (LLM)-based scoring.

However, evaluation in MIR remains an open and challenging research problem. Despite significant advances in model architectures and datasets, the field continues to grapple with issues such as inconsistent evaluation protocols across studies—including differences in datasets, metrics, experimental settings, and even task formulations—alongside concerns of data leakage across train-test splits and model bias [151, 152]. Furthermore, existing unified benchmarks for music representation learning are predominantly Western-centric and lack cultural diversity, limiting the ability to assess the generalizability of learned

representations across diverse musical traditions. Finally, comprehensive evaluation of learned representations should go beyond downstream task-specific probing to assess internal structural properties of the latent space, such as robustness, invariance, safety, and interpretability, as well as alignment with human preferences. While such holistic evaluation protocols are gaining traction in other domains, notably with recent efforts proposing standardized metrics to quantify informativeness, equivariance, invariance, and disentanglement of representations [153], they remain largely unexplored in the context of music representation learning.

Another key consideration is the role of temporal resolution and granularity, along with the distinction between sequence-level and token-level downstream MIR tasks, which is essential in music representation learning. *Temporal resolution* refers to how finely the input audio is segmented or represented over time (e.g., the number of tokens/frames per second in language modeling approaches), determining how precisely a model can capture the timing and dynamics of musical events at a fine temporal level. Downstream tasks in MIR differ in their temporal granularity requirements: *sequence-level* tasks, such as music auto-tagging or genre classification, require global understanding of the entire clip, while *token- or frame-level* tasks demand fine-grained temporal precision, as in downbeat tracking, chord recognition, or instrument source separation. Importantly, temporal granularity requirements vary not only across tasks, but also across musical cultures, where distinct rhythmic structures, ornamentation, and expressive timing necessitate varying levels of sensitivity. Designing a general-purpose foundation model that can flexibly adapt to such variations remains a core challenge, particularly in cross-cultural contexts where both task semantics and temporal structures differ significantly.

We should note that this study centers on music understanding and does not address music generation, which is beyond our current scope.

2.4 Cross-Cultural MIR

MIR has traditionally centered on the analysis of Western musical forms, notably, mainstream Euro-American popular music and Western classical repertoire. A growing body of work highlights the strong Western bias in MIR research and emphasizes the need for cross-cultural broadening [21, 11, 59, 60]. For instance, [61] introduced SAMBASET, a 40+ hour dataset of Brazilian samba music, specifically to challenge the dominant "Western" focus. They argue that most MIR datasets, methodologies, and conclusions embed substantial cultural bias, with non-Western music often being underrepresented, poorly labeled, or even mislabeled. Similarly, [3] quantify this bias by showing that only 5.7% of music generation data derives from non-Western traditions, underscoring their underrepresentation and the urgent need for more culturally diverse datasets (see Figure 2.2). While identifying Western bias is a necessary first step, addressing it requires practical efforts to develop datasets, representations, and evaluations tailored to non-Western musical traditions.

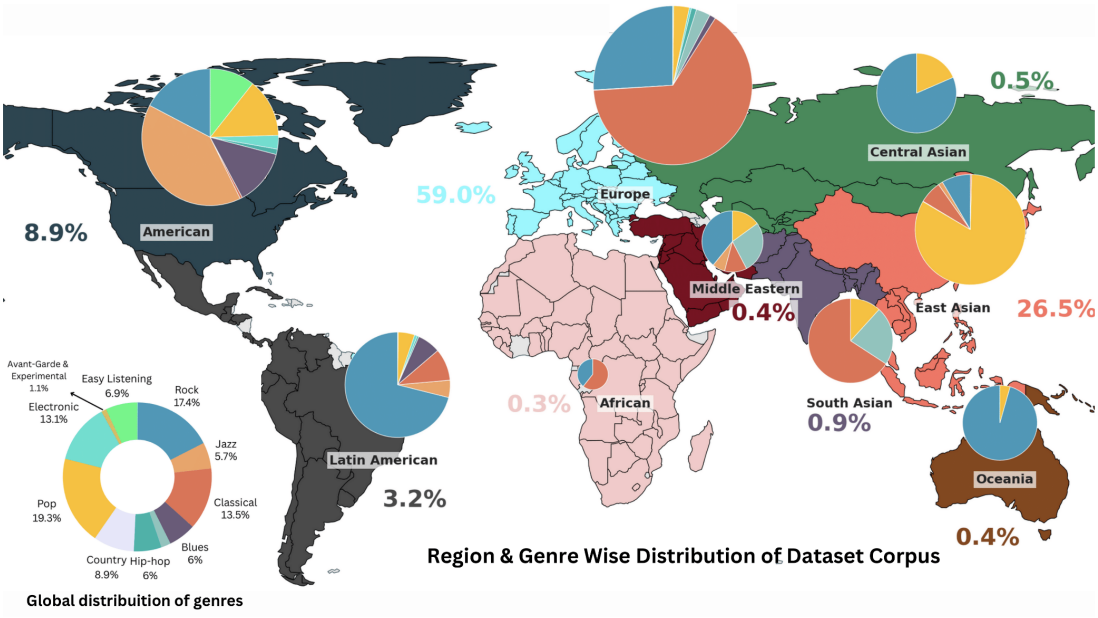


Figure 2.2. *Global distribution of music corpora by region. Pie charts illustrate genre composition within each region.*[3] *Reproduced from [3].*

Datasets Several initiatives have emerged to bridge this gap by curating culturally specific music corpora from diverse regions. The CompMusic project [62] has been central to this effort, offering an extensive collection of over 1,300 hours of music corpora spanning various non-Western traditions. This includes large annotated datasets for Indian classical music, specifically Hindustani and Carnatic [86, 155], as well as corpora for Turkish Makam [84, 85], Beijing Opera [156], and Arab-Andalusian [157] music. Complementing this, KritiSamhita [65] offers a tonic-annotated Carnatic vocal dataset, while [158] present a 191-hour Hindustani classical dataset labeled by raga and tonic.

As previously noted, for Latin-American music, [61] introduced SAMBASET, a 40-hour dataset of Brazilian *samba de enredo* recordings with rich metadata and beat/downbeat annotations. Additional Latin-American datasets include annotated recordings of Uruguayan *candombe* drumming for beat and downbeat tracking [159]; the Latin Music Database (LMD) [160], which contains over 3,200 full-length recordings across ten Latin genres for genre classification; the Brazilian Music Dataset (BMD) [161], focused on regional Brazilian styles such as *repente*, *brega*, and MPB (Brazilian popular music); and the Brazilian Rhythmic Instruments Dataset (BRID) [162], a dataset of Brazilian rhythmic instrument recordings across styles like *samba*, *samba de enredo*, *partido-alto*, and *capoeira*, designed for rhythmic pattern and microtiming analysis. For classical flamenco music from Southern Spain, the corpusCOFLA [63] offers over 1,800 commercial recordings (approximately 95 hours) selected from canonical anthologies, accompanied by editorial metadata and several test collections. These include manual and automatic annotations for tasks such as vocal detection, vocal pitch extraction, automatic singing transcription, repeated melodic

[3] However, we acknowledge that genre categorization is subjective and music genre perception varies significantly across cultures [154].

pattern discovery, melodic similarity, and style classification, supporting computational analysis of flamenco's rich melodic ornamentation and vocal expressivity.

For Iranian traditional music, the Nava *Dastgāh* dataset [64] provides 1,786 solo vocal excerpts (totalling approximately 55 hours), categorized into seven canonical *dastgāhs* and labeled by expert performers. In addition, the KUG *Dastgāhi* Corpus (KDC) [163] offers a growing collection of well-curated and annotated audio for computational analysis of Persian modal music. Furthermore, a curated corpus of traditional Georgian *a cappella* vocal music is provided by the Erkomaishvili Dataset [66], comprising 101 historic three-voice overdubbed recordings performed by master chanter Artem Erkomaishvili. The dataset includes transcriptions in Western staff notation (in MusicXML), F0 trajectories for all three vocal parts, and manually annotated note and rest onset positions. The recordings exhibit not equal-tempered tuning and abundant pitch slides, reflecting the distinctive tuning practices and harmonic thinking of Georgian polyphony. These properties make the dataset a valuable resource for MIR tasks such as multi-pitch estimation, onset detection, source separation, and score-to-audio alignment, while also supporting musicological research on traditional Georgian vocal music. Moreover, [67] introduced Lyra, an 80-hour corpus of Greek traditional and folk music, annotated with rich metadata on genre, place of origin, and instrumentation. For Chinese musical traditions, [68] proposed CCMusic, a unified database integrating multiple Chinese music datasets, both published and unpublished, with standardized structure, annotations, and evaluation protocols. In the Scandinavian context, [164] released a corpus of Norwegian Hardanger fiddle recordings, annotated with precise note and beat onsets by expert performers. Emerging efforts are also addressing the under-representation of African music traditions in MIR. The Sotho-Tswana dataset [69] is a multimodal collection of music video clips annotated for genre, sentiment, lyrics, and visual features. Likewise, the Ndwom dataset [70] contains curated Akan music videos, spanning genres such as Highlife, Gospel, Soul, and Asakaa, with multimodal annotations (audio, lyrics, and video frames), curated and labeled by native Akan speakers.

Finally, very recently, [71] introduced M4-RAG, a web-scale dataset comprising 2.31 million music–text pairs, covering 160,000 hours of audio from 1.8 million tracks, with rich metadata and multilingual annotations across 27 languages and 194 countries. They also released WikiMT-X, a benchmark designed to support evaluation in multilingual and cross-modal MIR, addressing critical gaps in globally representative, high-quality training and evaluation resources. In parallel, the GlobalMood benchmark [15] provides a novel cross-cultural dataset for music emotion recognition, comprising 1,180 songs from 59 countries and nearly one million mood ratings elicited through a bottom-up, participant-driven tagging approach across five culturally and linguistically distinct regions. Fine-tuning multimodal models like CLAP on this culturally grounded dataset significantly improves their alignment with human judgments. It also demonstrates that mood perception is culturally grounded and highlights the need for localized descriptors and multilingual annotation pipelines for building more representative and equitable music foundation models.

Methodologies Furthermore, recent years have seen growing interest in the computational analysis of non-Western musical traditions [72]. Notable work includes studies on

Turkish makam recognition [73, 165], Indian classical music classification [74, 158], and analysis of Iranian [75] and Korean [76] traditional music. Additionally, [77] present a computational pitch analysis of traditional Ghanaian *seperewa* (Akan harp-lute) songs, revealing systematic microtonal deviations from equal temperament in vocal tracks and highlighting both the limitations and the implicit Western-centric biases of standard MIR tools and assumptions when applied to culturally specific musical systems such as Ghanaian *seperewa* scales. While some recent efforts have explored music auto-tagging in cross-cultural transfer settings [78] and addressed challenges in low-resource and imbalanced world music datasets through few-shot learning [79], comprehensive evaluations or adaptations of foundation models in cross-cultural low-resource MIR contexts remain scarce. Recently, CLaMP 3 [71] proposed a framework to align multiple musical modalities and multilingual text in a shared representation space via contrastive learning, enabling cross-modal alignment and generalization to unseen languages for MIR tasks, and demonstrated state-of-the-art performance on tasks such as text-to-audio and text-to-symbolic music retrieval. Recent efforts have also explored cultural adaptation in music generation. In particular, [3] investigate parameter-efficient fine-tuning (PEFT) approaches and demonstrate that culturally adapting MusicGen [58] and Mustango [80] with low-resource corpora improves performance on Hindustani classical and Turkish Makam music, highlighting both the potential and the challenges of cross-cultural adaptation.

Challenges Still, despite these promising developments, addressing cultural imbalance in MIR requires more than simply diversifying datasets, as [81] argue. The field must critically reflect on its foundational assumptions, epistemological, ontological, methodological, and axiological, by reconsidering what music is, how it is understood, and how it should be studied in ways that acknowledge and respect diverse cultural knowledge systems. The very definition of music, the boundaries between music and other cultural expression, and the values attached to musical sound can differ radically across cultures. They also call for greater interdisciplinarity with ethnomusicology, as well as the inclusion of domain experts, to incorporate non-Western musical concepts and values. In practice, several recurring challenges hinder progress in cross-cultural MIR. Data scarcity is foremost: many musical cultures lack large annotated corpora, resulting in long-tail label distributions and unseen tags during evaluation. Model bias is another key issue: models trained on Western-centric data often reflect WEIRD (Western, Educated, Industrialized, Rich, Democratic) assumptions [26], and may also encode inductive biases implicitly tuned to Western musical structures, genres, and semantics, thereby limiting their generalizability to non-Western traditions. These biases are often compounded, as Western-oriented data and modeling assumptions mutually reinforce one another, producing systems that underperform, or even misrepresent, music from other cultural contexts. To address these limitations, culturally diverse benchmarks, evaluation metrics, and protocols are essential for uncovering and mitigating bias in computational models and for quantifying progress toward truly universal music representations. This includes rethinking ground-truth annotations, adapting evaluation criteria to align with the epistemologies of different musical traditions, and co-developing models in collaboration with local experts and practitioners.

Chapter 3

Methodology

In this chapter, we first review the architecture and pre-training objective of **MERT** (**M**usic und**ER**standing model with large-scale self-supervised **T**raining), and then present our continual pre-training strategy for cultural adaptation. Finally, we investigate task arithmetic, an alternative approach to multi-cultural adaptation that merges culturally specialized models in weight space to construct a unified multi-cultural model, `CultureMERT-TA`. Our experiments are conducted on a diverse collection of Western and non-Western music datasets, with model evaluation performed on corresponding music auto-tagging tasks, as detailed in the following section.

3.1 Datasets

For our experiments, we use a diverse set of music datasets spanning both Western and non-Western traditions. Specifically, we adopt the MagnaTagATune (MTAT) [82] and FMA-medium [83] datasets to represent "Western"[1] music. For non-Western traditions, we incorporate the Lyra corpus [67], featuring Greek traditional and folk music, along with three collections from the CompMusic Corpora[2] [62]: Turkish-makam [84, 85], which, together with Lyra, represent music of the Eastern Mediterranean; and Hindustani and Carnatic music [86], representing North and South Indian classical traditions, respectively.

We assess our models on both Western and non-Western music tagging tasks for cross-cultural evaluation, using standard multi-label classification metrics, including the area under the receiver operating characteristic curve (ROC-AUC), average precision (AP), and F1 scores (both micro-averaged and macro-averaged). Following [78, 79], we utilize the top-k tags relevant to each dataset: 50 tags for MTAT (spanning *genre*, *instruments*, and *mood*), 20 hierarchical *genre* tags for FMA-medium, 30 tags for Turkish-makam (covering *makam*, *usul*, and *instruments*), 20 tags for Hindustani and Carnatic (primarily reflecting *raga*, *tala*, *instruments*, and *forms*), and 30 tags for Lyra (related to *genre*, *place*, and *instruments* metadata).

All audio is resampled to 24 kHz, and we adopt the same data splits as [78]. To prepare our data for continual pre-training, we extract 30-second segments from each training split of the non-Western datasets. Given the varying dataset sizes, we balance the pre-training

[1] We use the term "Western" to refer to music styles predominantly rooted in Western cultures, including pop, rock, and Western classical.

[2] https://compmusic.upf.edu/corpora

duration across cultures to ensure proportional representation by extracting 200 hours each from the Turkish-makam, Carnatic, and Hindustani datasets, and 50 hours from Lyra due to its smaller size. Additionally, we combine these subsets to construct a unified 650-hour dataset integrating all four traditions for multi-cultural continual pre-training. While Lyra is of limited volume compared to the other datasets, its inclusion in the multi-cultural data mix serves two key purposes: (i) our preliminary experiments indicate that even a small amount of diverse data enhances overall generalization performance, and (ii) it ensures the multi-cultural model is exposed to all non-Western traditions it is evaluated on, maintaining consistency in evaluation.

3.2 MERT Pre-Training Objective

Our continual pre-training objective follows the self-supervised masked language modeling (MLM) objective of $\texttt{MERT}^{\texttt{RVQ-VAE}}$, where two teacher models provide the pseudo-labels:

- (i) an *acoustic teacher*, the EnCodec codec model [2], which discretizes/tokenizes audio into tokens from $K = 8$ residual vector quantization (RVQ) codebooks, each containing $C = 1024$ codewords, and
- (ii) a *musical teacher*, based on constant-Q transform (CQT) spectrogram reconstruction, encoding pitch and harmonic structure.

`MERT-v1-95M` follows the HuBERT architecture [55], comprising a CNN-based feature extractor that encodes raw 24 kHz waveforms into 75 Hz frame-level representations, followed by a 12-layer Transformer encoder, producing 768-dimensional contextual embeddings (see Figure 3.1). During training, a subset of frame embeddings is masked, and the model is optimized using a multi-task learning (MTL) objective, combining masked acoustic token prediction and spectrogram reconstruction.

The overall training objective is:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{RVQ}} + \mathcal{L}_{\text{CQT}}, \tag{3.1}$$

where the acoustic MLM loss $\mathcal{L}_{\text{RVQ}}$ encourages the model to predict masked RVQ-VAE tokens from K codebooks, using a noise-contrastive estimation (NCE) [166] loss:

$$\mathcal{L}_{\text{RVQ}} = \sum_{k=1}^{K} \sum_{t \in M} \log p_\theta(c_{t,k} | \boldsymbol{x}'_t), \tag{3.2}$$

with M denoting the set of masked time frames, $c_{t,k}$ the *ground-truth* discrete codeword from the k-th codebook at time frame t extracted via the EnCodec tokenizer, and p_θ the model's predicted token distribution:

$$p_\theta(c | \boldsymbol{x}'_t) = \frac{\exp(\text{sim}(T(\boldsymbol{o}_t), \boldsymbol{e}_c)/\tau)}{\sum_{c'=1}^{C} \exp(\text{sim}(T(\boldsymbol{o}_t), \boldsymbol{e}_{c'})/\tau)}. \tag{3.3}$$

Here, $\boldsymbol{x}'_t$ is the masked input feature, $\boldsymbol{o}_t$ is the model's output representation, $T(\boldsymbol{o}_t)$ projects

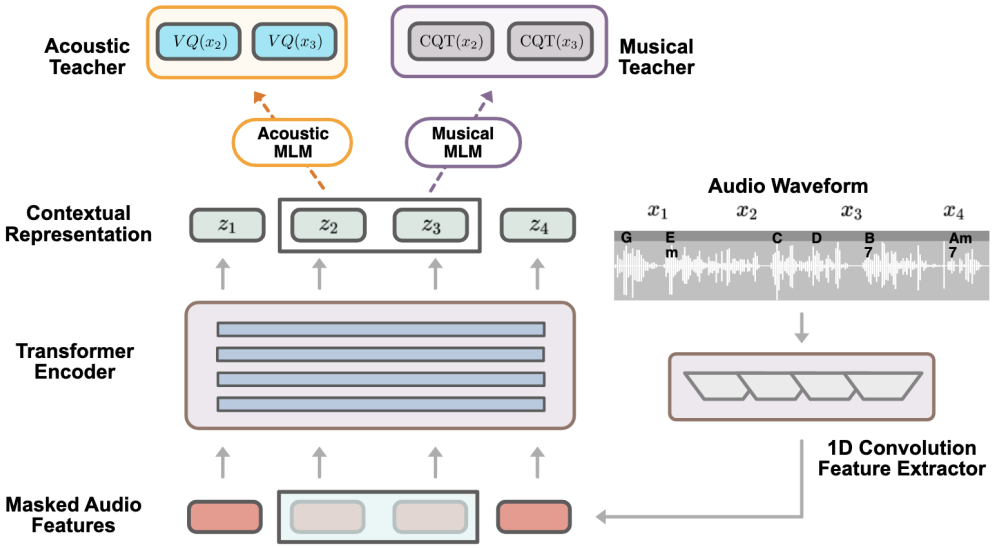


Figure 3.1. *MERT Pre-Training Framework [1].*

it to the codeword embedding space, $\boldsymbol{e}_c$ is the embedding of codeword $c \in \mathcal{C}_k$, where $k \in \{1, \ldots, K\}$, $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity, and $\tau = 0.1$ is a temperature scaling parameter.

The musical MLM CQT reconstruction loss $\mathcal{L}_{\mathrm{CQT}}$ minimizes the mean squared error (MSE) between the model's predicted $\hat{\mathbf{z}}_{\mathrm{CQT},t}$ and ground-truth $\mathbf{z}_{\mathrm{CQT},t}$ frame-level CQT features:

$$\mathcal{L}_{\mathrm{CQT}} = \sum_{t \in M} \|\mathbf{z}_{\mathrm{CQT},t} - \hat{\mathbf{z}}_{\mathrm{CQT},t}\|_2^2 . \tag{3.4}$$

The hyperparameter α controls the relative importance of the acoustic MLM token prediction loss $\mathcal{L}_{\mathrm{RVQ}}$ and the musical MLM spectrogram reconstruction loss $\mathcal{L}_{\mathrm{CQT}}$. By jointly optimizing these objectives, MERT pre-training balances acoustic and musical representation learning.

3.3 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 our case, this involves continually pre-training the `MERT-v1-95M` model, using the same pre-training objective, on culturally diverse data that introduce a significant distribution shift, as it was initially trained on predominantly Western music [1, 43]. Given this shift, naively continuing to train the model, i.e., adapting all parameters at once without resetting the learning rate, can lead to catastrophic forgetting [44] and poor adaptation [4], as confirmed by our preliminary experiments (see Section 4.6). To address this, we propose a **two-stage** strategy that stabilizes training through:

- (i) **learning rate re-warming and re-decaying** [4, 31, 34, 88, 89], and

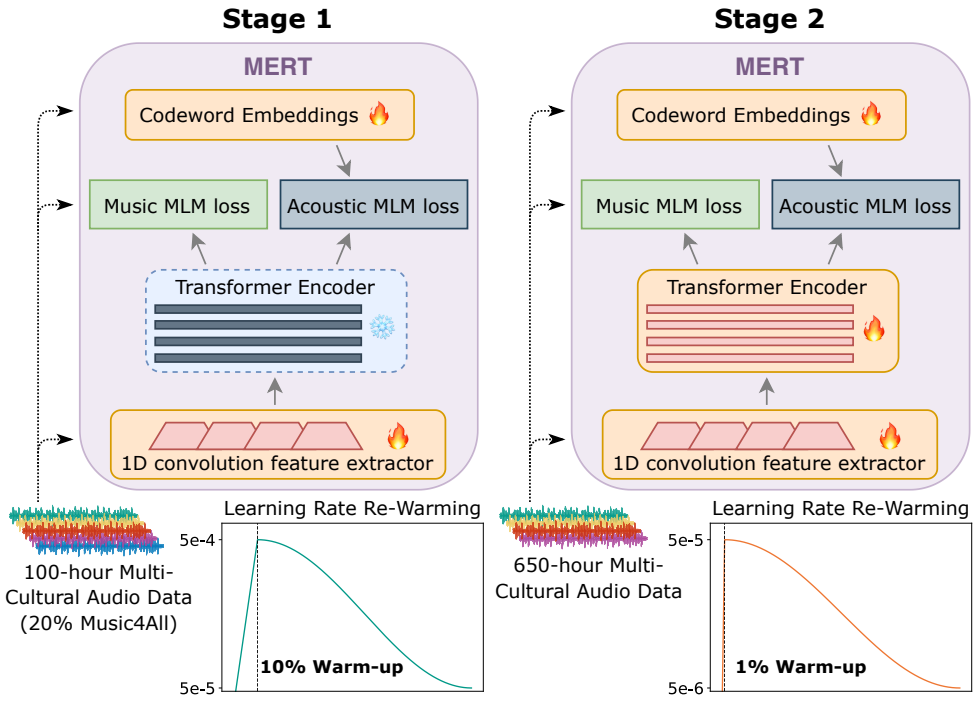


Figure 3.2. ***Two-Stage Continual Pre-Training Strategy for*** `CultureMERT`. *In Stage 1, a subset of parameters (the 1D CNN feature extractor and codeword embeddings) is trained on 100 hours of multi-cultural data for multiple epochs, with 20% Western music for stabilization. In Stage 2, all parameters are unfrozen and trained on the full 650-hour dataset. Learning rate re-warming and re-decaying is applied in both stages for smooth and stable adaptation.*

- (ii) **staged adaptation**.

The overall framework of our approach is illustrated in Figure 3.2, which depicts the two-stage continual pre-training strategy for `CultureMERT`.

Staged Adaptation In our preliminary experiments, we observed an initial performance drop during CPT, followed by a slow recovery phase, a phenomenon known as the *stability gap* [90, 167, 31]. This instability arises due to the abrupt adaptation of model parameters to a substantially shifted data distribution, which can temporarily degrade previously learned representations before stabilizing. To mitigate this, rather than full-parameter adaptation on the entire dataset in a single epoch, which induces a large plasticity gradient for a long period [167], we split training into two stages to reduce instability and ensure smoother adaptation, as illustrated in Figure 3.2:

- **Stage 1** *Stabilization Phase*: We first train on a smaller subset of the data [90], updating only the CNN-based feature extractor and the codeword embedding layer while keeping the Transformer encoder frozen. To reduce the distribution gap and mitigate forgetting [88, 109, 35], we incorporate a fraction of Music4All data [7], which is primarily of Western origin, into the pre-training mix, accounting for 20% of the total training data (*Western replay*).

- **Stage 2** *Full Adaptation*: We unfreeze the Transformer encoder and continue training on the full dataset. While *Western replay* (e.g., including a portion of Music4All data) can also be applied at this stage to further mitigate forgetting, it introduces a trade-off between cultural adaptation and knowledge preservation, i.e., the *stability–plasticity dilemma* [91] (see Section 4.6).

This two-stage approach is particularly motivated by computational constraints, specifically the batch size mismatch between pre-training and adaptation. `MERT-v1-95M` was originally trained with batch sizes of 1.5 hours per step, whereas we use a significantly smaller effective batch size of 160 seconds per step due to memory limitations. Training with this reduced batch size directly on the entire dataset with full-parameter adaptation resulted in unstable training and frequent crashes, degrading performance on both Western and non-Western benchmarks.

By structuring adaptation in two stages, we strike to balance *plasticity* (adaptation to non-Western traditions) and *stability* (retaining knowledge on Western datasets), a challenge known as the *stability-plasticity dilemma* [91, 92, 93]. Intuitively, the initial *stabilization phase* allows lower-level acoustic representations, captured by the CNN-based feature extractor and the codeword embeddings, to adapt first and calibrate to the shifted distribution before updating high-level Transformer representations.

Learning Rate Re-Warming and Re-Decaying To further improve adaptation stability during continual pre-training, we apply learning rate re-warming and re-decaying in both stages. Continual pre-training on a shifted distribution can lead to poor convergence and forgetting if the learning rate is not adjusted properly [4, 34]. Prior work has shown that resetting the learning rate schedule, i.e., *re-warming* the model, during continual pre-training is crucial for effective adaptation and mitigating catastrophic forgetting [4, 31, 34, 88, 89]. The learning rate schedule significantly impacts the training dynamics and efficacy of CPT and re-warming is necessary for efficient adaptation to new data.

We adopt a two-phase learning rate schedule comprising a linear warm-up followed by cosine annealing (see Figure 3.3), following prior work [4, 88]. The learning rate η_t at timestep t is defined as:

(1) Linear warm-up (for $t \leq T_{\text{warmup}}$):

$$\eta_t = \eta_{\max} \cdot \frac{t}{T_{\text{warmup}}} \tag{3.5}$$

(2) Cosine annealing (for $t_{\text{ann}} \leq t \leq t_{\text{end}}$):

$$\eta_t = \eta_{\min} + \frac{\eta_{\max} - \eta_{\min}}{2} \cdot \left(\cos\left(\pi \cdot \frac{t - t_{\text{ann}}}{t_{\text{end}} - t_{\text{ann}}} \right) + 1 \right) \tag{3.6}$$

where:

$\eta_{\max}$ is the maximum learning rate, $\eta_{\min}$ is the minimum learning rate, T_{warmup} is the warm-up duration, $t_{\text{ann}} = T_{\text{warmup}}$ is the start of cosine annealing, and $t_{\text{end}} = T_{\text{ann}} + t_{\text{ann}}$ is the total training duration.

Warm-up durations in prior work typically range between 0.1%–2% of total training steps. In the audio domain, HuBERT base uses a warm-up phase of 8% during initial pre-training. Models using shorter warm-up phases tend to forget and adapt more quickly in early training due to the faster learning rate increase. However, over longer training durations, this effect becomes less impactful on overall forgetting and adaptation, as noted in [4]. The choice of learning rate re-warming strategy depends on the training objective and the task at hand. The selection of warm-up duration and maximum learning rate also reflects a trade-off between stability and plasticity. Carefully designed warm-up and decay schemes are crucial for effective continual adaptation [31]. In our preliminary experiments, we extensively tested different warm-up and decay durations, as well as learning rate values; the final values used in each stage are reported in Section 3.5.3. Finally, the learning rate is typically annealed down to 0.1× the maximum learning rate, consistent with prior cosine decay schedules, where the maximum learning rate is initialized to match the $\eta_{\max}$ of the original pre-trained model.

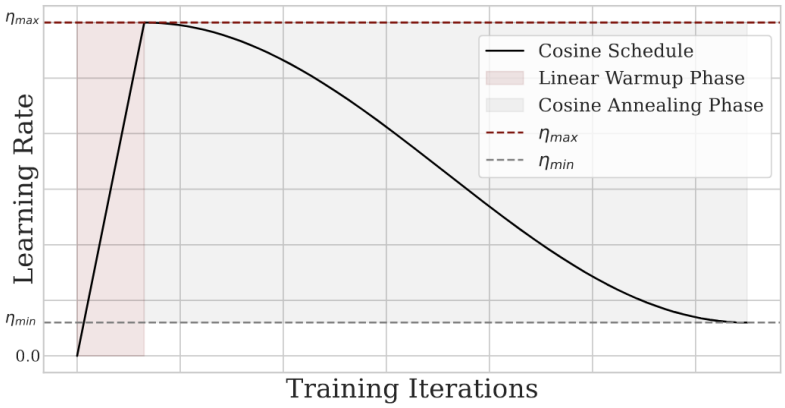


Figure 3.3. ***Linear warm-up and cosine annealing schedule.*** *Reproduced from [4].*

- In **Stage 1**, we adopt a moderately aggressive warm-up and decay schedule to encourage early adaptation of low-level representations.
- In **Stage 2**, a less aggressive schedule balances plasticity and stability during full-model training, reducing also training instabilities.

Following this two-stage CPT strategy, we develop two types of culturally adapted models:

- (i) a **multi-culturally adapted model**, `CultureMERT`, trained on a culturally diverse mix spanning all four non-Western musical traditions (Turkish-makam, Hindustani, Carnatic, and Lyra); and
- (ii) **single-culture adapted models**, each continually pre-trained on data from a single tradition, resulting in `MakamMERT`, `HindustaniMERT`, `CarnaticMERT`, and `LyraMERT`.

3.4 Task Arithmetic for Cross-Cultural Adaptation

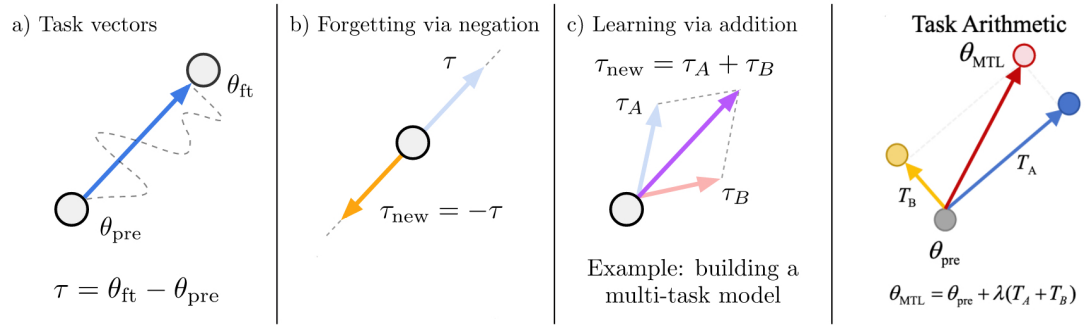


Figure 3.4. ***Merging Models via Task Arithmetic.*** *Adapted from [5] and [6].*

As an alternative to continual pre-training on multi-cultural data, we explore task arithmetic [5], a model merging method that combines culturally specialized models in weight space to construct a unified multi-cultural model. Model merging [40] has recently emerged as a promising approach for integrating multiple independently fine-tuned or task-specific models into a single model without requiring access to training data or additional re-training. Task arithmetic operates by algebraically merging model parameters through weight vector addition and subtraction in Euclidean space, as illustrated in Figure 3.4. Specifically, it treats the difference between a task-adapted model and its pre-trained base as a *task vector* in weight space. Linear combinations of such task vectors have been shown to effectively steer model behavior and enable knowledge transfer across domains [94, 5]. For example, adding a task vector to a base model can improve its performance on the corresponding task, while combining multiple task vectors supports multi-task generalization. Notably, task vectors exhibit a form of *compositionality*: expressions such as $\tau_D = \tau_C + (\tau_B - \tau_A)$ can yield improved performance on a target task D, even without direct training data—enhancing domain generalization and revealing analogical structure in model space (e.g., *"A is to B as C is to D"*). Furthermore, negating task vectors enables the removal of specific "behaviors" from a model, offering a mechanism for targeted forgetting. These properties highlight task arithmetic as an efficient, modular, and data-free strategy for domain adaptation (DA), and a lightweight tool for model editing.

In our setting, we obtain *task vectors* by computing the element-wise difference between the parameters of the single-culture continually pre-trained models—i.e., the culturally specialized models—and those of the `MERT-v1` model. Formally, given the initial pre-trained model with parameters θ_{pre} and a continually pre-trained model θ_i adapted to a cultural dataset $\mathcal{D}_i$, the task vector for culture i is given by $\tau_i = \theta_i - \theta_{pre}$, capturing the parameter shift induced by culture-specific adaptation. For multi-cultural adaptation, we construct a unified model θ_{merged} by merging N single-culture adapted models via task arithmetic, summing their respective task vectors τ_i with corresponding scaling factors λ_i:

$$\theta_{merged} = \theta_{pre} + \sum_{i=1}^{N} \lambda_i \tau_i, \tag{3.7}$$

where $\lambda_i \in \mathbb{R}$ are scalar hyperparameters that control the contribution of each task vector,

typically determined using held-out validation sets.

Prior work on task arithmetic typically uses a single scaling factor λ for all task vectors, i.e., $\lambda_i = \lambda$, $\forall i$, reducing Equation 3.7 to:

$$\theta_{\text{merged}} = \theta_{\text{pre}} + \lambda \sum_{i=1}^{N} \tau_i. \tag{3.8}$$

In the special case where $\lambda = 1/N$, this further simplifies to:

$$\theta_{\text{merged}} = \theta_{\text{pre}} + \frac{1}{N} \sum_{i=1}^{N} (\theta_i - \theta_{\text{pre}}) = \frac{1}{N} \sum_{i=1}^{N} \theta_i, \tag{3.9}$$

which corresponds to *weight averaging* [111, 112, 39], where the adapted models are merged by directly averaging their parameters.

Here, we merge $N = 4$ single-culture adapted models—`MakamMERT`, `HindustaniMERT`, `CarnaticMERT`, and `LyraMERT`—to construct a unified multi-cultural model, referred to as `CultureMERT-TA`. Details on the choice of scaling factor λ are provided in Section 4.4.

3.5 Experimental Setup

3.5.1 Implementation Details

In all continual pre-training setups, we initialize our models from the publicly available `MERT-v1-95M`[3] pre-trained checkpoint. Training was conducted using the FAIRSEQ[4] framework on a single NVIDIA GeForce GTX TITAN X GPU with 12 GB of memory. All models were trained with half-precision (FP16), using 5-second audio segments as input context, randomly cropped from the extracted 30-second pre-training audio data. The weight of the acoustic loss in the pre-training objective is set to $\alpha = 10.0$. The EnCodec neural audio codec (NAC) model [2], which tokenizes audio into discrete codewords, remains frozen during continual pre-training, as in [1]. To enhance representation robustness, we apply *in-batch noise mixture augmentation* with a mixup probability of 0.5 (see Section 4.6.2 for a complete discussion), and use pre-layer normalization (Pre-LN) [168] for training stability, following [1]. The impact of mixup augmentation is further examined in Section 4.6.2 (Table 4.4). Other training settings mirror those of the `MERT-v1-95M` setup; further details are provided in Appendix A.1.

3.5.2 Probing-Based Evaluation

Following [57, 1, 19], we adopt a probing-based [87] evaluation rather than fine-tuning, keeping the pre-trained models frozen as deep feature extractors while training only a shallow multilayer perceptron (MLP) with a single 512-dimensional hidden layer for sequence-level tasks. Our evaluation follows the MARBLE protocol [149] under constrained settings, and we apply it to both Western and non-Western music tagging tasks for cross-cultural

[3] https://huggingface.co/m-a-p/MERT-v1-95M
[4] https://github.com/facebookresearch/fairseq

evaluation. To process long-duration audio files, we segment them into 30-second chunks using a sliding window approach and aggregate the chunk-level predictions by averaging to obtain the final prediction for the entire audio file. For Turkish-makam, Hindustani, and Carnatic tasks, we apply a maximum duration cut[5] as in [78]. Evaluation hyperparameters are detailed in Appendix A.2.

3.5.3 Continual Pre-Training Settings

Multi-Cultural CPT In Stage 1, training runs for 2,250 steps with a 10% linear warm-up period, using 100 hours of the dataset, where 20% of the mix consists of Western music from Music4All [7]. Optimization follows AdamW [169] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e^{-5}$. Training employs an effective batch size of 32 recordings (160 seconds) with gradient accumulation over 8 steps. The maximum learning rate is set to $\eta_{\max} = 5e^{-4}$, followed by a cosine decay to a minimum of $\eta_{\min} = 5e^{-5}$. Gradient clipping is applied with a norm of 1.0 to prevent exploding gradients. In Stage 2, training extends to 14,625 steps with a 1% warm-up period, using the full 650-hour dataset. Optimization follows AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 1e^{-5}$, maintaining the same batch size as Stage 1. The learning rate decays from a maximum value of $\eta_{\max} = 5e^{-5}$ to $\eta_{\min} = 5e^{-6}$. Gradient clipping remains at 1.0.

Single-Culture CPT In Stage 1, we train on 60 hours, with 20% of the mix allocated to Music4All, for a total of 1,350 training steps. In Stage 2, we expand training to the full 200-hour dataset for 4,500 steps. We employ the same optimizers, batch size settings, and learning rate schedules as in the multi-cultural CPT. For Lyra, due to its smaller size (50 hours), we train on 20 hours in Stage 1 (450 steps) and then on the full dataset in Stage 2 (1,125 steps).

[5] This is done to ensure comparability with previous state-of-the-art results reported in [78].

Chapter 4

Results and Discussion

In this chapter, we present and analyze the empirical findings of our study. We begin with the evaluation results of our culturally adapted models across diverse music auto-tagging tasks, highlighting the effectiveness of multi-cultural continual pre-training and model merging via task arithmetic. We then examine patterns of cross-cultural transfer, focusing on how cultural proximity influences cross-cultural generalization of single-culture adapted models. This is followed by an analysis of token-level similarity across musical traditions as a potential predictor of positive transferability. We also explore the sensitivity of task arithmetic to the scaling factor and layer-wise probing performance. Finally, we provide an in-depth evaluation and analyze the training dynamics of our proposed two-stage adaptation strategy through detailed ablations, demonstrating its effectiveness in mitigating catastrophic forgetting, stabilizing training, and enhancing cultural adaptation.

4.1 Evaluation Results

As shown in Tables 4.1 and 4.2, `CultureMERT`, adapted via multi-cultural continual pre-training, consistently outperforms the initial `MERT-v1` model across all non-Western tasks and evaluation metrics. 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 computed for each individual culture, thereby enhancing generalization. This finding aligns with observations in multilingual NLP and speech recognition, where pre-training on diverse multilingual corpora, such as in XLM-R [170] and XLS-R [171], has been shown to improve cross-lingual transfer and performance [97, 172], particularly in low-resource or unseen language settings [98]. Notably, `CultureMERT` achieves this with minimal forgetting on Western benchmarks (-0.05% average drop in ROC-AUC and AP), demonstrating the efficacy of our approach. Furthermore, it shows better retention of prior Western knowledge compared to single-culture models, which suffer greater performance drops when evaluated on FMA-medium and MagnaTagATune (MTAT).

We further observe that single-culture adapted models tend to achieve the best performance on their respective in-domain tasks, particularly for well-resourced traditions, reaffirming the effectiveness of continual pre-training for domain-specific adaptation [32]. This trend holds consistently across all evaluation metrics. Interestingly, even low-resource

Dataset	Turkish-makam		Hindustani		Carnatic		Lyra		FMA-medium		MTAT		Avg.
Metrics	ROC	AP	ROC	AP	ROC	AP	ROC	AP	ROC	AP	ROC	AP	
MERT-v1	$83.2_{0.08}$	$53.3_{0.12}$	$82.4_{0.04}$	$52.9_{0.19}$	$74.9_{0.05}$	$39.7_{0.15}$	$85.7_{0.10}$	$56.5_{0.18}$	$90.7_{0.04}$	$48.1_{0.11}$	$89.6_{0.07}$	$35.9_{0.15}$	66.1
MakamMERT	$88.7_{0.11}$	$58.8_{0.22}$	$84.5_{0.16}$	$57.8_{0.18}$	$77.6_{0.14}$	$42.7_{0.16}$	$84.6_{0.12}$	$53.2_{0.17}$	$90.3_{0.12}$	$47.1_{0.16}$	$89.0_{0.07}$	$35.6_{0.12}$	67.5
CarnaticMERT	$88.4_{0.06}$	$58.4_{0.16}$	$87.0_{0.06}$	$60.2_{0.14}$	$78.8_{0.13}$	$\mathbf{44.0}_{0.17}$	$85.4_{0.11}$	$55.8_{0.16}$	$90.2_{0.10}$	$46.7_{0.09}$	$89.2_{0.10}$	$35.3_{0.11}$	68.3
HindustaniMERT	$88.3_{0.12}$	$58.2_{0.16}$	$87.4_{0.11}$	$60.3_{0.16}$	$77.0_{0.12}$	$42.7_{0.16}$	$84.2_{0.13}$	$52.0_{0.15}$	$90.2_{0.13}$	$46.1_{0.10}$	$89.1_{0.09}$	$35.8_{0.13}$	67.6
LyraMERT	$86.7_{0.07}$	$56.8_{0.13}$	$85.9_{0.08}$	$57.4_{0.13}$	$76.4_{0.09}$	$40.1_{0.13}$	$85.0_{0.11}$	$53.5_{0.14}$	$90.0_{0.08}$	$46.0_{0.16}$	$88.9_{0.05}$	$35.1_{0.14}$	66.8
CultureMERT	$\mathbf{89.6}_{0.09}$	$60.6_{0.21}$	$\mathbf{88.2}_{0.20}$	$\mathbf{63.5}_{0.24}$	$\mathbf{79.2}_{0.18}$	$43.1_{0.22}$	$86.9_{0.10}$	$56.7_{0.20}$	$90.7_{0.09}$	$48.1_{0.13}$	$89.4_{0.09}$	$35.9_{0.16}$	**69.3**
CultureMERT-TA	$89.0_{0.12}$	$\mathbf{61.0}_{0.18}$	$87.5_{0.10}$	$59.3_{0.13}$	$79.1_{0.11}$	$43.3_{0.13}$	$\mathbf{87.3}_{0.08}$	$\mathbf{57.3}_{0.19}$	$90.8_{0.06}$	$49.1_{0.15}$	$89.6_{0.10}$	$36.4_{0.14}$	69.1
(Previous) SOTA	87.7 [78]	57.7 [78]	86.5 [78]	63.1 [78]	77.0 [78]	43.9 [78]	85.4 [78]	54.3 [78]	**92.4** [78]	**53.7** [78]	**92.7** [95]	**41.4** [57]	-

Table 4.1. ***Evaluation Results (ROC-AUC and AP) of Pre-Trained and Culturally Adapted MERT Models on Diverse Music Auto-Tagging Tasks (1/2).*** *We report averages across five random seeds with standard deviations as subscripts. The "Avg." column represents the average performance across all datasets and evaluation metrics for each model. The results highlight the impact of multi-cultural CPT (`CultureMERT`) and multi-cultural model merging via task arithmetic (`CultureMERT-TA`) on cross-cultural adaptation and transfer.*

adaptation, as in the case of `LyraMERT` trained on just 50 hours of Greek folk music, leads to noticeable gains across other non-Western tasks, indicating that even limited cultural exposure can significantly enhance cross-cultural generalization beyond Western datasets. This finding is consistent with recent work in low-resource NLP [32] and speech recognition [38], where in the latter continually pre-training with as little as 10 hours of target language data yielded substantial improvements over unadapted models. Improvements in Macro-F1 scores across non-Western datasets are particularly noteworthy, highlighting that cross-cultural adaptation not only enhances overall accuracy but also improves recognition of less frequent tags among the top-k most common labels used in our evaluation. This is particularly important for ethnomusicological datasets [79], where even within the top-k evaluated tags, frequency distributions remain imbalanced and capturing a wider diversity of musical concepts is crucial.

Moreover, multi-cultural model merging via task arithmetic achieves comparable performance 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, task arithmetic also outperforms the original pre-trained model on average across Western tasks, further reinforcing its ability to balance adaptation and retention. Finally, `CultureMERT` and `CultureMERT-TA` surpass previous state-of-the-art (SOTA) results (ROC-AUC and AP) on all non-Western music tagging tasks, with the best task arithmetic variant obtained using $\lambda = 0.2$ (see Figures 4.4 and 4.6). Notably, only the multi-cultural models, `CultureMERT` and `CultureMERT-TA`, outperform the original `MERT-v1` on Lyra, albeit with the smallest margin compared to other non-Western tasks. This observation aligns with the fact that `MERT-v1`, pre-trained on Western mu-

Dataset	Turkish-makam		Hindustani		Carnatic		Lyra		FMA-medium		MTAT		Avg.
Metrics (F1)	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	
MERT-v1	73.0	38.9	71.1	33.2	80.1	30.0	72.4	42.6	57.0	36.9	**35.7**	21.2	49.3
MakamMERT	**77.5**	44.0	74.0	37.6	81.0	31.4	70.8	40.2	57.2	35.4	34.2	20.5	50.3
CarnaticMERT	76.8	44.0	76.2	46.3	81.6	32.4	72.9	42.8	57.3	35.3	33.3	22.5	51.8
HindustaniMERT	76.5	43.9	**78.9**	46.9	81.0	**33.0**	70.1	40.6	55.1	33.8	34.4	20.9	51.3
LyraMERT	75.9	42.1	75.9	44.9	80.9	29.6	71.3	41.1	56.2	33.9	33.8	21.2	50.6
CultureMERT	77.4	**45.8**	77.8	**50.4**	**82.7**	32.5	**73.1**	43.1	58.3	36.6	35.6	**22.9**	**52.9**
CultureMERT-TA	76.9	45.4	74.2	45.0	82.5	32.1	73.0	**45.3**	**59.1**	**38.2**	**35.7**	21.5	52.4

Table 4.2. ***Evaluation Results (Micro-F1 and Macro-F1) of Pre-Trained and Culturally Adapted MERT Models on Diverse Music Auto-Tagging Tasks (2/2).*** *The "Avg." column represents the average performance across all datasets and both Micro-F1 and Macro-F1 for each model. The results further highlight the impact of multi-cultural CPT (`CultureMERT`) and multi-cultural model merging via task arithmetic (`CultureMERT-TA`) on cross-cultural adaptation and transfer.*

sic, already serves as a strong baseline for Lyra, surpassing previous SOTA, potentially reflecting certain underlying similarities between Greek folk music and Western musical traditions. Overall, these results further underscore the effectiveness of multi-cultural adaptation, especially in low-resource and transfer settings.
We next analyze these quantitative findings in greater depth by examining patterns of cross-cultural transfer and cross-cultural generalization of culturally adapted models.

4.2 Cross-Cultural Transfer

As illustrated in Figures 4.1 and 4.2, continual pre-training on one musical tradition can benefit others to varying degrees, revealing differing levels of 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 both *maqam* (Makam) and *raga* (Carnatic) being modal systems that emphasize microtonal pitch variation, ornamentation, and improvisation, serving similar functions within their respective musical cultures [173].

Additionally, we observe strong cross-cultural transfer between the Carnatic and Hindustani traditions. Specifically, the Carnatic-adapted model achieves high scores across all metrics when evaluated on the Hindustani auto-tagging task, while the Hindustani-adapted model shows slightly stronger transfer in F1 scores when evaluated on Carnatic music (see Tables 4.1 and 4.2). This mutual transferability reinforces the musical proximity between these traditions, particularly in their shared use of *raga* (melodic mode) and *tala* (rhythmic framework) [24], despite differences in performance structure, melodic movements, and the types of instruments used. Moreover, despite some shared musical characteristics, such as modal improvisation and microtonality, Turkish-makam models do not generalize well to Hindustani music. This gap highlights that theoretical similarity may not necessarily

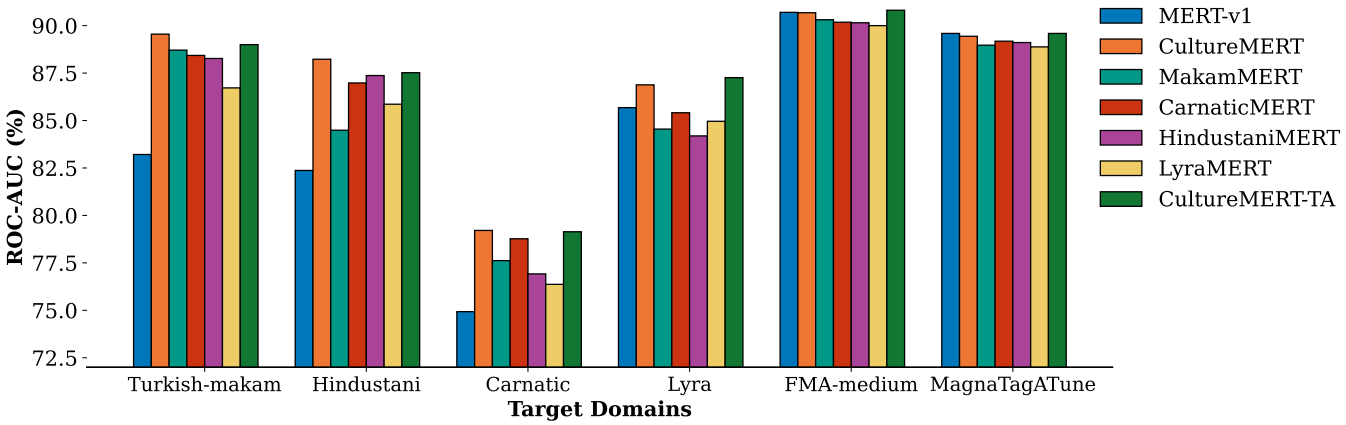


Figure 4.1. ***ROC-AUC Comparison Across Culturally Adapted Models on Diverse Music Auto-Tagging Tasks.*** *Continual pre-training on multi-cultural data (`CultureMERT`) consistently achieves the highest performance across most datasets, particularly for non-Western traditions, surpassing both single-culture adaptations and model merging via task arithmetic (`CultureMERT-TA`). However, the latter demonstrates particularly strong results on Lyra and Western-centric auto-tagging tasks.*

translate to practical transferability. In general, we observe that cross-cultural transfer is not always symmetric. For instance, while the Carnatic-adapted model generalizes well to Hindustani music, the reverse direction yields slightly better results in certain metrics (e.g., Macro-F1). Such asymmetries have also been observed in cross-lingual transfer research [96, 97, 98]. We encourage further exploration of (a)symmetric cross-cultural transfer patterns in the context of music representation learning.

Interestingly, the model adapted to Carnatic music appears to be the most consistently transferable among all single-culture adaptations, achieving the highest average scores across multiple non-Western traditions in ROC-AUC, AP, and F1 metrics. It performs strongly not only within Indian classical traditions but also generalizes well to Turkish-makam and Lyra, suggesting a particularly robust capacity for cross-cultural generalization. This observation aligns with findings in cross-lingual NLP [98], where certain high-resource "super-donor" languages consistently boost performance across diverse low-resource languages, often irrespective of linguistic proximity.

Greek traditional and folk music presents a unique challenge, as it theoretically shares elements with both non-Western and Western traditions. It exhibits melodic improvisation similar to Turkish-makam and Hindustani music, while also employing harmonic accompaniment influenced by Western classical and folk traditions. This blending of modal and tonal frameworks has been extensively discussed in ethnomusicological studies, particularly in the context of "Rebetiko", which integrates makam-based melodies with Western chordal harmony [99]. In our experiments, we observe that `MERT-v1`, originally trained on Western music, already serves as a strong baseline when evaluated on the Lyra auto-tagging task. Moreover, adapting the initial model, whether using data from a single tradition or from a diverse multi-cultural mix, consistently yields the smallest gains on Lyra across all evaluation metrics among the non-Western tasks. This suggests that the underlying musical structure of Greek folk music may partially align with the Western biases already

present in the foundation model.

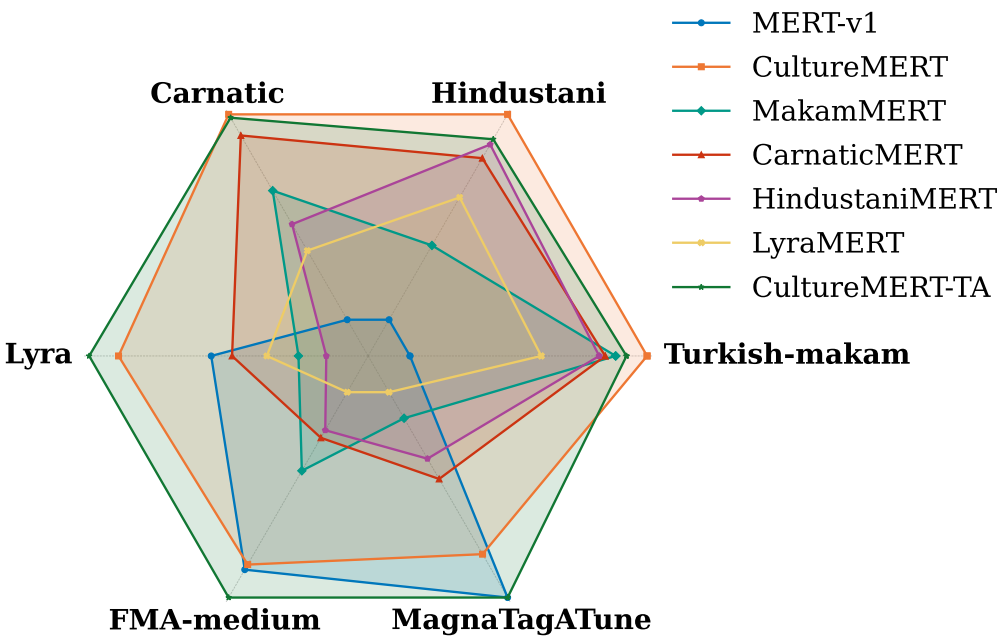


Figure 4.2. ***Cross-Cultural Transferability.*** *Relative ROC-AUC performance across datasets, highlighting key trends in cross-cultural transfer.* `CultureMERT` *generalizes well to non-Western datasets, while task arithmetic performs on par in these settings and even surpasses both the pre-trained and multi-culturally adapted models on Western benchmarks (FMA-medium, MTAT) and Lyra.*

As expected, the pre-trained `MERT-v1` model performs strongly on Western-centric datasets such as MTAT and FMA-medium, reflecting its initial training bias toward Western musical traditions. In contrast, models adapted to individual non-Western cultures (e.g., `MakamMERT`, `HindustaniMERT`, etc.) often exhibit reduced performance on these Western benchmarks. This highlights the substantial domain shift between Western and non-Western musical audio representations. However, this effect is substantially mitigated by `CultureMERT`, and even more so by `CultureMERT-TA`, whose exposure to a diverse range of musical traditions during continual pre-training or model merging enables them to better retain generalization across both non-Western and Western domains.

The efficacy of task arithmetic in cross-cultural transfer mirrors recent findings in cross-lingual transfer learning. Notably, [94] demonstrated that combining language- and task-specific models via arithmetic operations significantly improves performance across both high-resource and low-resource languages.

Overall, these results emphasize that cultural proximity, shared musical structures, and the internal diversity of traditions all play critical roles in cross-cultural transferability. It is important to note, however, that our analysis is based on continual pre-training starting from a Western-biased foundation model, rather than training from scratch for each tradition. Thus, the observed cross-cultural transfer patterns may partly reflect how these musical cultures are projected into the representational space shaped by prior Western-centric pre-training.

4.3 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 codec model [2], which serves as our audio tokenizer. These pseudo-tokens represent discrete acoustic representations that are also used as masked prediction targets in the acoustic MLM pre-training objective of MERT (Section 3.2). Lower values in both metrics indicate greater similarity. Our analysis, as shown in Figure 4.3, 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 folk music (Lyra), while distinct, aligns more closely with non-Western traditions than Western ones. These findings underscore the need for cultural adaptation to address distributional shifts in audio representations.

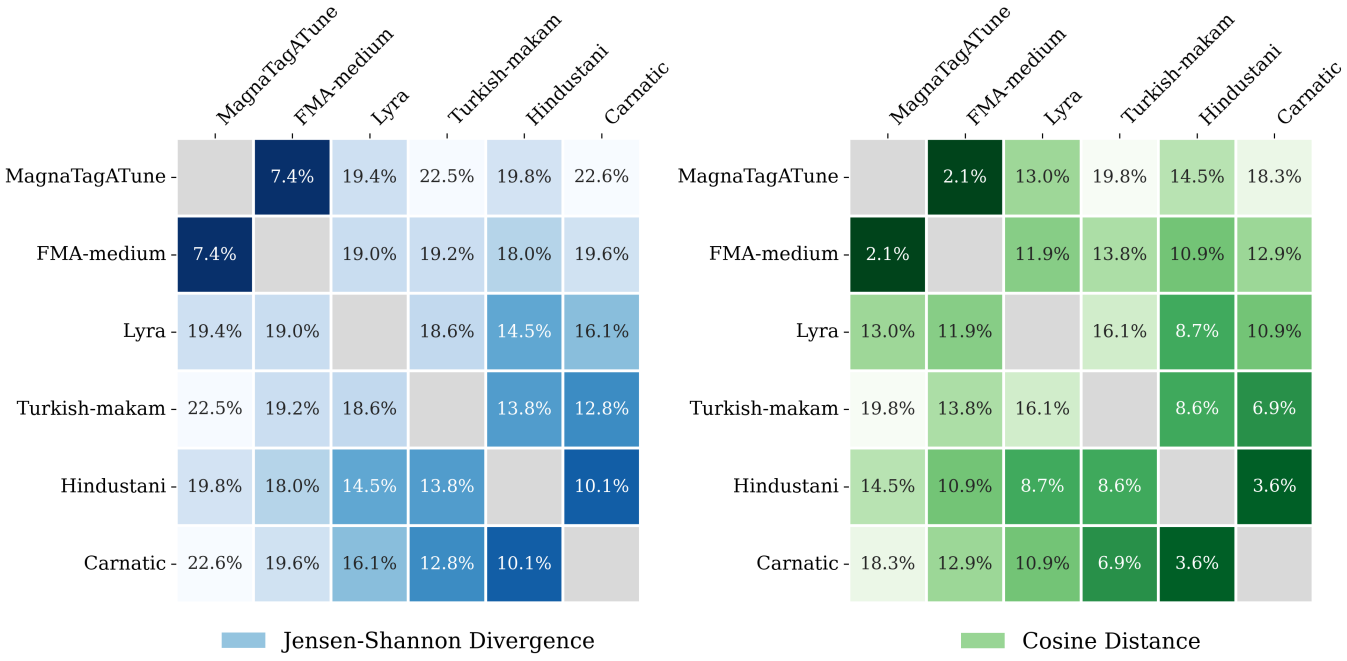


Figure 4.3. ***Token Similarity Across Cultures.*** *Pairwise similarity between acoustic token distributions extracted from the EnCodec NAC model [2]. Similarity scores are averaged across* 8 *codebooks, each containing* 1024 *discrete codewords (acoustic pseudo-tokens). Both measures—JSD and cosine distance—show consistent trends across cultures.*

Interestingly, these findings correlate with our results on cross-cultural transfer (Section 4.2), suggesting that token-level similarity metrics can serve as predictors of **positive cross-cultural transfer**. This insight has practical implications: such similarity metrics can help guide the selection and refinement of mixture proportions of pre-training data during CPT, or inform the adjustment of arithmetic operations when combining models via task arithmetic. This is particularly valuable in low-resource scenarios, where limited data for an underrepresented culture can be complemented by leveraging similar, higher-resource cultures as effective "donors", an approach supported by recent findings in low-resource

speech recognition [38] and NLP [98]. 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 [105, 32, 174, 96, 98] and speech domains [38]. Finally, these observations are consistent with known ethnomusicological similarities and resonate with findings from prior work in multilingual adaptation [38, 175], where cross-lingual similarities, and, in turn, positive cross-lingual transfer, were often associated with historical, structural, and social proximity between languages, such as shared linguistic roots or sustained contact.

We should note that the EnCodec audio tokenizer used for extracting acoustic token distributions was originally trained on Western musical data and was kept frozen during our experiments. While this could introduce a Western-centric bias in how token similarities are measured, it is actually aligned with the inductive biases of the `MERT-v1` foundation model, which was also trained predominantly on Western music. Following arguments made in recent low-resource speech adaptation research [38], using a similarity measure grounded in the pre-trained model's internal representations, rather than relying on external notions of similarity, is often more predictive of positive transfer in CPT settings. Thus, despite potential biases, the EnCodec-derived token similarity remains a suitable and meaningful predictor of cross-cultural transferability in our setting. Additionally, an alternative strategy could involve following the Acoustic Token Distribution Similarity (ATDS) approach proposed in [38], by extracting frame-level contextual embeddings from a transformer layer of the pre-trained model, clustering them to induce deep semantic tokens, and computing token distribution similarities based on the resulting semantic token frequency vectors. Such model-internal representations could offer an even more tailored and task-specific measure of cultural proximity, closely aligned with the inductive biases of the model being adapted. Furthermore, it would be interesting to derive semantic token similarities from our culturally adapted models using the same approach, and compare them to those from the original `MERT-v1` pre-trained model and the EnCodec acoustic tokens, to better understand how cultural adaptation shifts internal representations of musical proximity.

However, it remains an open question which aspects of similarity most effectively predict cross-domain or cross-cultural transfer relative to the target task. This echoes recent findings in multilingual NLP, where the most predictive notion of similarity varies depending on the downstream task. For instance, [96] demonstrate that syntactic similarity best predicts cross-lingual transfer in POS tagging and parsing, while lexical and n-gram overlap are stronger predictors for topic classification. Their study also considers a wide array of similarity measures, including grammatical structure, phonological and phonetic features, phylogenetic relatedness, geographic proximity, and dataset-level token overlap. Inspired by this, we suggest that future work in music understanding should investigate which dimensions of similarity, whether captured via acoustic tokens, semantic representations, or external musicological knowledge, best align with task-specific performance in transfer settings. Moreover, diverse and representative music corpora could support the construction of phylogenetic trees or networks based on various similarity measures to further explore cross-cultural relationships. Finally, exploring audio tokenizers trained on globally diverse corpora may also yield a more holistic and "unbiased" view of cross-cultural proximity.

4.4 Task Arithmetic Scaling Factor

A key consideration in task arithmetic is the choice of the scaling factor λ, which controls the balance between task vectors. Prior work [6, 94] has shown that suboptimal values can significantly degrade performance in multi-task model merging. We systematically evaluate different values of a shared scaling factor $\lambda \in \{0.1, 0.2, 0.25, 0.3, 0.5, 0.75, 1.0\}$, applied uniformly across all task vectors, following the simplified formulation in Equation 3.8, including the special case of weight averaging ($\lambda = 0.25$). Consistently with prior observations, we find that ill-suited values, such as $\lambda = 1.0$, result in poor performance across all benchmarks, as shown in Figures 4.4 and 4.6.

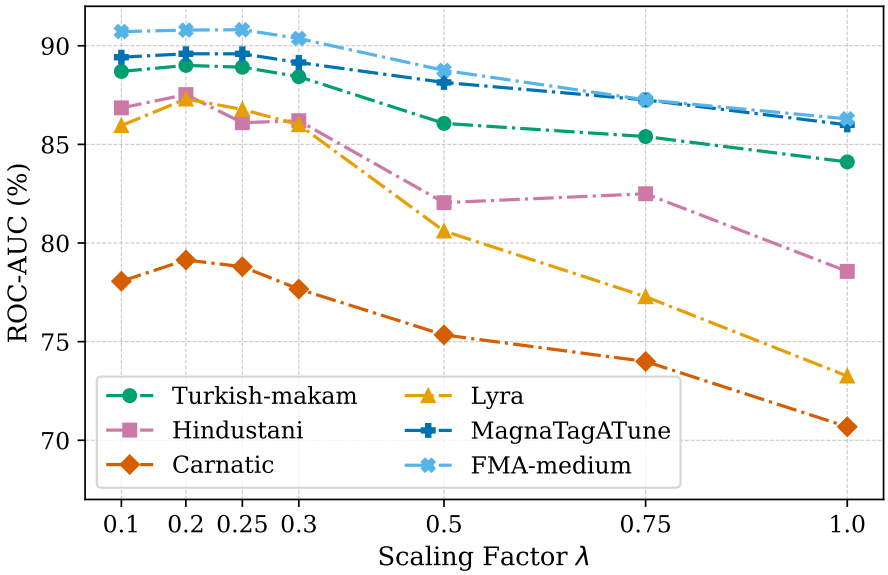


Figure 4.4. ***Effect of Scaling Factor λ on Task Arithmetic Performance.*** *The ROC-AUC scores across six diverse music tagging tasks demonstrate how varying λ impacts task arithmetic when merging the four non-Western single-culture adapted models.*

AWD [176] suggests that this sensitivity arises because inter-task interference is amplified when task vectors are scaled up. This is consistent with our observations: as the scaling factor λ increases, performance systematically degrades across all evaluated auto-tagging tasks (Figure 4.4). Such interference stems from task vectors not being orthogonal [176], a hypothesis further supported by measuring the cosine similarity between task vectors (Figure 4.5). In line with this, other studies have noted that naive linear merging via task arithmetic can suffer from parameter conflicts. For instance, TIES-Merging [177] identifies two key sources of cross-task interference: (a) redundant small-magnitude parameters that introduce noise when merged, and (b) sign conflicts where models "disagree" on the direction of parameter changes.

While multi-cultural continual pre-training jointly learns representations across multiple musical cultures, task arithmetic offers a post-hoc merging strategy by combining culturally specialized models without requiring additional training or access to original data, provided such models are already trained. In our experiments, the best-performing task arithmetic variant was obtained with a scaling factor of $\lambda = 0.2$, a result that was

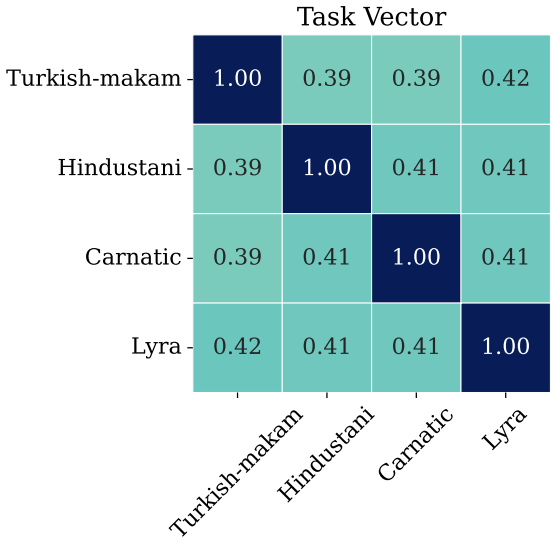


Figure 4.5. ***Cosine Similarity Between Task Vectors.*** *The values highlight significant overlap (non-orthogonality) among task vectors, which contributes to inter-task interference during model merging with task arithmetic.*

consistent across all evaluated music auto-tagging tasks (Figure 4.4). Although task arithmetic offers a strong alternative, its effectiveness depends critically on the careful tuning of scaling factors and may be more sensitive to inter-task interference, particularly when task vectors are highly correlated. In future work, we plan to investigate more robust model merging methods and task arithmetic variants that better mitigate parameter interferences and task conflicts, including adaptive scaling strategies, task- and layer-weighted merging, and interference-robust optimization techniques (see Section 5.3 for suggested future work).

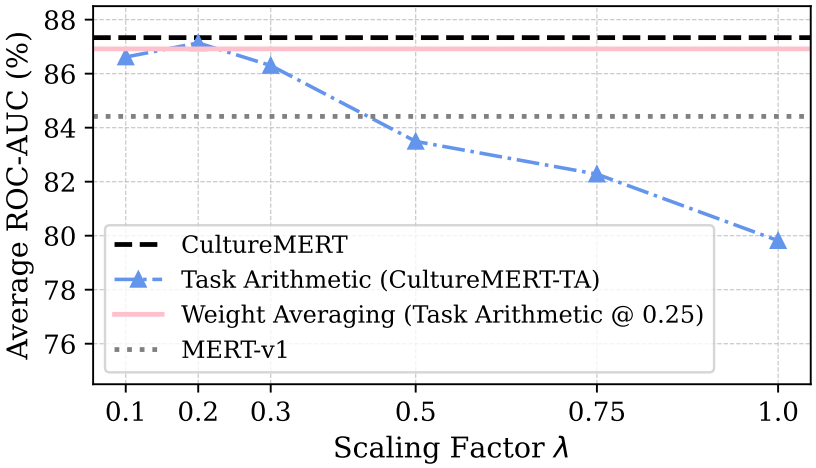


Figure 4.6. ***Task Arithmetic vs. Multi-Cultural CPT.*** *Average ROC-AUC performance across benchmarks for different task arithmetic scaling factors, compared against multi-cultural continual pre-training (`CultureMERT`) and the pre-trained baseline (`MERT-v1`). The best average task arithmetic performance is achieved with a scaling factor of* $\lambda = 0.2$.

4.5 Layer-wise Cultural Encoding in `CultureMERT`

Probing-based evaluation (for details see Appendix A.2) reveals that different transformer layers in `CultureMERT` provide the most effective representations for different cultural auto-tagging tasks (see Figure 4.7). Interestingly, these optimal layers not only vary across tasks but also differ from those observed in `MERT-v1`, indicating a "reorganization" of cultural knowledge during multi-cultural continual pre-training. This suggests that culturally relevant information may be encoded at varying depths within the multi-culturally adapted model, rather than being uniformly distributed across layers. Such variation reflects the diversity of musical representations learned through CPT and highlights the importance of task-aware feature extraction and layer-wise selection for culture-specific modeling. The following list summarizes the transformer layers that yielded the best performance via probing for each evaluation dataset:

- **Turkish-makam:** `layer 6`
- **Hindustani:** `layer 7`
- **Carnatic:** `layer 7`
- **Lyra:** `layer 8`
- **FMA-medium:** weighted sum over all layers (`all`)
- **MagnaTagATune:** weighted sum over all layers (`all`)

As shown in Figure 4.7, intermediate transformer layers in `CultureMERT` yield the best representations for most tasks, particularly for non-Western traditions. The optimal layers for these datasets range from layers 5 to 8, indicating that culture-specific information may be encoded at the mid-network layers after continual pre-training. Interestingly, the Hindustani and Carnatic auto-tagging tasks both achieve peak performance at `layer 7`, aligning with prior observations of cross-cultural similarity and transfer between these two traditions. In contrast, Western benchmarks such as FMA-medium and MTAT exhibit more uniform performance across layers. The learnable weighted sum over all layers (`all`) performs robustly across all datasets and achieves the best results on Western-centric benchmarks. This pattern may reflect the impact of multi-cultural continual pre-training on a Western-biased model: while Western benchmarks retain broadly distributed representations from the base model, non-Western datasets benefit from localized adaptation, with culture-specific features emerging more strongly at specific intermediate layers. Notably, representations extracted from the lower layers of the transformer encoder, and especially from the pre-transformer 1-D CNN feature extractor (`layer 0`), result in consistently lower performance across all evaluated tasks, suggesting that they primarily encode low-level acoustic characteristics insufficient for semantic-level tasks. Finally, top transformer layers (e.g., `layer 11` and `layer 12`) underperform slightly across most datasets, likely because they specialize more in the masked modeling pre-training objective rather than providing features relevant for downstream tasks.

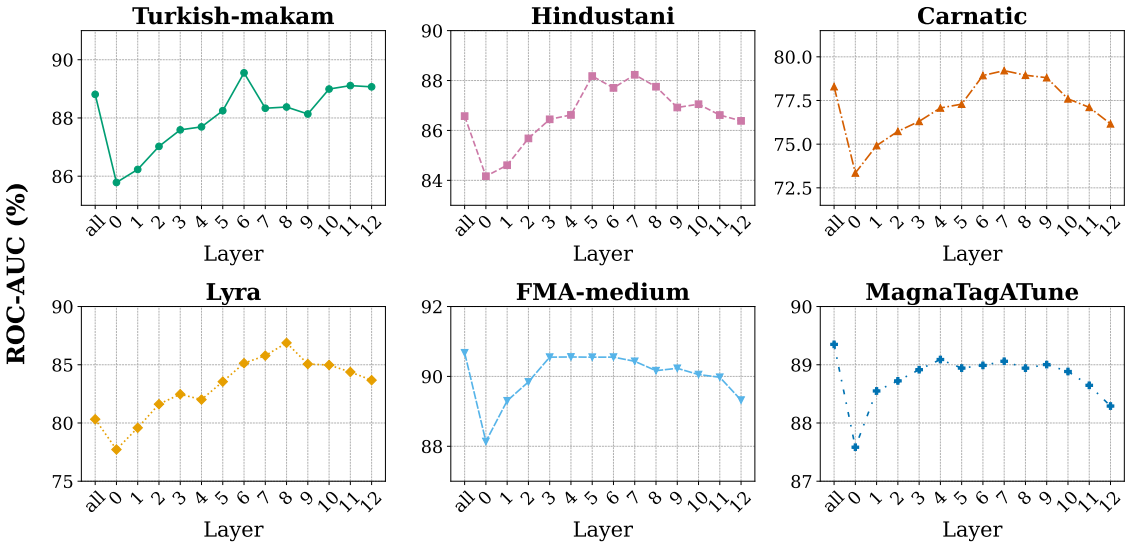


Figure 4.7. ***Layer-wise Probing Performance of `CultureMERT` across Datasets.*** *ROC-AUC scores across layers for each evaluation dataset, obtained via probing of representations extracted from the frozen backbone.*

These findings are consistent with prior work in speech representation learning [178], computer vision [179], and NLP [180, 181], where task- or language-specific information tends to be concentrated at different layers of transformer models, and the most informative and transferable features often lie in the middle of the network. A similar trend was also observed in the original `MERT-v1` model across different music understanding tasks, showing also that intermediate layers tended to cluster music examples by genre or content, whereas the top layers focused more on the MLM-style pre-training task [1]. Our findings also resonate with recent layer-wise analyses in music foundation models such as MuQ [100], which show that acoustic tasks (e.g., pitch or instrument classification) tend to benefit from lower layers, while semantic tasks (e.g., genre classification or structure analysis) perform best at higher layers. In contrast, comprehensive tasks such as music tagging distribute across layers. These observations suggest that different MIR tasks, and by extension, culturally diverse musical content, may require representations extracted from different network depths. Such insights reinforce the importance of task-specific representation selection and motivate future work exploring more adaptive, culturally aware, and **interpretable** layer aggregation strategies to further illuminate how musical and culturally-specific attributes are distributed across the network [182].

4.6 Two-Stage Adaptation Strategy

In this section, we evaluate the effectiveness of our proposed two-stage continual pre-training strategy, which incorporates learning rate re-warming, and compare it against single-stage CPT baselines that perform full-parameter adaptation in a single step, with and without re-warming. Our empirical analysis supports the core design choices of the two-stage approach, focusing on three key aspects: (i) mitigation of catastrophic forgetting, (ii) effectiveness of cultural adaptation, and (iii) training stability.

4.6.1 Mitigating Catastrophic Forgetting

Figure 4.8 shows the ROC-AUC performance on the Western MagnaTagATune (MTAT) dataset during two-stage continual pre-training. We observe a performance drop in Stage 1, followed by gradual recovery in Stage 2, evidence of catastrophic forgetting due to a distributional shift, as well as the effectiveness of staged adaptation. Specifically, Stage 1 plays a critical role in mitigating forgetting by adapting only low-level representations using a re-warmed learning rate, while keeping the Transformer encoder frozen. This controlled adaptation sets the stage for smoother full-parameter training in Stage 2, where a decayed learning rate promotes more stable optimization. The initial drop in Stage 1 is partially alleviated by *Western replay*, i.e., injecting training data resembling the original pre-training distribution of `MERT-v1`. As shown in Table 4.3, this two-stage setup, particularly when Western replay is restricted to Stage 1, yields the best trade-off between cultural adaptation and knowledge retention.

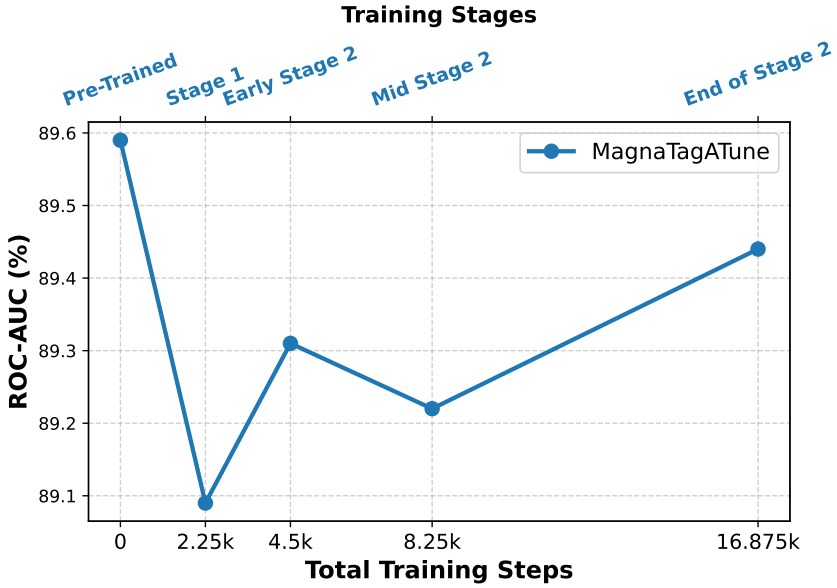


Figure 4.8. ***Catastrophic Forgetting on the MTAT Dataset.*** *ROC-AUC performance during two-stage continual pre-training shows an initial drop in Stage 1, followed by recovery in Stage 2. This demonstrates how staged adaptation with learning rate re-warming and Western replay (20%) mitigates catastrophic forgetting.*

4.6.2 Cross-Cultural Adaptation Effectiveness

We observe that naively continuing pre-training with the reduced learning rate from the original pre-trained model (i.e., without re-warming) fails to adapt effectively to different cultures, as the pre-trained representations are not sufficiently shifted. This is evident in the Turkish-makam case, where single-stage CPT without learning rate re-warming yields no performance improvement. Applying re-warming in the single-stage setup results in minor adaptation gains (+0.8%), but this comes at the expense of substantial forgetting on the MTAT task (−3.6%), even when Western replay (20%) is included. In contrast, the two-stage CPT strategy achieves significantly greater adaptation (+6.4%) while minimizing

forgetting (−0.4%), highlighting its effectiveness in balancing plasticity and stability. Finally, we examine the role of Western data replay as a mechanism for mitigating forgetting. While incorporating Western data helps preserve performance on MTAT, it introduces a stability–plasticity trade-off: excessive replay can inhibit effective cultural adaptation. Our results (Table 4.3) suggest that restricting replay to Stage 1 offers the best overall balance between knowledge retention and effective adaptation.

CPT Strategy	Western Replay	Turkish-makam	MTAT
MERT-v1 (Baseline)	-	83.2	**89.6**
Single-stage (w/ re-warming)	✓	83.8	86.0
Single-stage (w/o re-warming)	✓	83.0	87.5
Two-stage (*Ours*)	Stage 1	**89.6**	89.2
Two-stage (*Ours*)	Both stages	88.6	89.4

Table 4.3. ***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. All CPT setups involving Western replay sample 20% of the total training data from the Music4All dataset [7].*

To further compare in-depth the two-stage approach versus single-stage full-parameter adaptation, we examine the training loss dynamics during CPT.

Musical MLM loss. Figure 4.10 shows the musical CQT MLM loss curves across different cultural adaptations and CPT strategies. Subfigure 4.10b compares two-stage and single-stage CPT on the multi-cultural dataset. The two-stage strategy consistently converges to a lower loss and demonstrates substantially less variance throughout training. In contrast, single-stage CPT exhibits noisier optimization dynamics, with slower convergence and a higher final loss. These observations support our hypothesis that Stage 1 enables smoother adaptation by first calibrating lower-level features to new data. This stage likely reduces representational shock, facilitating more stable and effective full-parameter training in Stage 2. In Subfigure 4.10a, we observe that convergence behavior varies across musical cultures: `CarnaticMERT` and `LyraMERT` reach the lowest final loss, followed by `HindustaniMERT` and `MakamMERT`. This variation suggests varying degrees of alignment between the original pre-trained model and each target cultural distribution. Notably, Greek and South Indian traditions appear more aligned, possibly due to greater overlap in pitch structure or spectral content as captured by the CQT representation.

Acoustic MLM loss. Figure 4.11 presents the corresponding results for acoustic MLM loss. Once again, two-stage CPT outperforms the single-stage baseline in both convergence speed and final loss values. Interestingly, the breakdown by codebook in Figure 4.11a reveals that codebooks 0–2, particularly codebook 0, consistently reach lower loss values, whereas deeper codebooks (e.g., codebooks 6 and 7) plateau at significantly higher levels. Since we use a pre-trained EnCodec tokenizer to discretize each waveform into eight parallel streams of quantized indices, one per residual codebook, these trends

reflect the relative predictability of each token stream. More specifically, the results suggest that tokens from later codebooks are far less predictable. This observation aligns with the structure of residual vector quantization. Lower codebooks capture coarse-grained acoustic features such as pitch and broad timbral envelopes, as well as rich semantic information like melody and vocal content [133]. In contrast, higher codebooks encode fine-grained, high-frequency residuals that are more variable [183]. These deep codebook tokens exhibit higher entropy and lower contextual redundancy, making them harder to model under a masked language modeling (MLM) objective. Moreover, since acoustic MLM loss is computed independently for each codebook and summed to obtain the final loss, this hierarchy in representational difficulty becomes particularly evident. This behavior was also reflected in preliminary experiments, where training with only the first four codebooks (codebook 0-3) yielded nearly identical downstream performance. This further motivated our exploration of using only four randomly selected codebooks per batch during multi-cultural continual pre-training (Table 4.4). Despite the reduced supervision signal per batch, this variant achieves comparable downstream performance with lower GPU memory utilization.

In-Batch Noise Mixture Probability	Acoustic Target Class	Turkish-makam	
		ROC-AUC	AP
0.5	$1024 \times 8^{\text{ all codebooks}}$	**89.55**	**60.62**
✗	$1024 \times 8^{\text{ all codebooks}}$	88.71	59.54
0.5	$1024 \times 4^{\text{ random codebooks}}$	88.45	59.24
✗	$1024 \times 4^{\text{ random codebooks}}$	88.14	58.31

Table 4.4. ***Mixup Augmentation and Codebook Usage Ablation.*** *This ablation study examines the effect of in-batch noise mixture augmentation and acoustic target class selection during multi-cultural continual pre-training, evaluated on the Turkish-makam auto-tagging task. Using a 0.5 probability for mixup consistently improves performance. Sampling four randomly selected codebooks per batch (instead of predicting targets from all 8 codebooks) offers a more memory-efficient alternative with only minor performance degradation, albeit with slower convergence [1] due to reduced supervision per update step.*

We also explore the impact of **in-batch noise mixture augmentation** introduced in the original MERT framework, which adds short, randomly selected audio segments from the same batch to the input waveform with a fixed probability. This augmentation encourages the model to learn robust, invariant representations by exposing it to perturbed inputs during pre-training. In our multi-cultural continual pre-training setting, we apply this augmentation with a 0.5 probability and observe consistent gains across downstream tasks (see Table 4.4). We hypothesize the mixup acts as a regularizer that enhances generalization, especially when learning from diverse cultural audio sources in multi-cultural pre-training.

4.6.3 Training Stability

We finally examine training stability by analyzing the gradient norm during continual pre-training. As shown in Figure 4.9, the two-stage CPT strategy exhibits significantly more stable gradient dynamics compared to the single-stage variant. In the multi-cultural adaptation setup, Stage 1 rapidly stabilizes gradients at low magnitudes, which enables Stage 2 to proceed with smoother full-parameter updates. In contrast, single-stage CPT displays sharp oscillations and frequent gradient spikes, particularly during the early stages of training. This instability often results in gradient explosions, which in turn cause frequent crashes and degraded convergence.

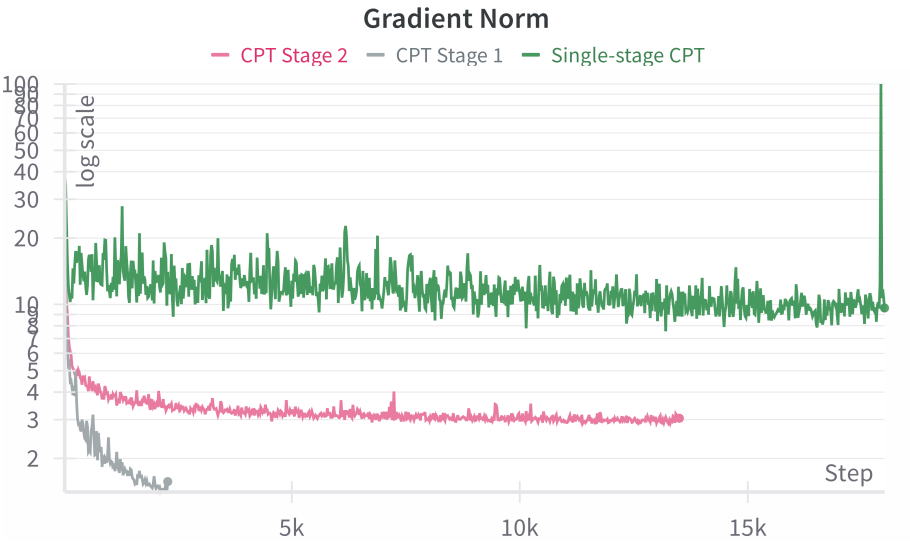
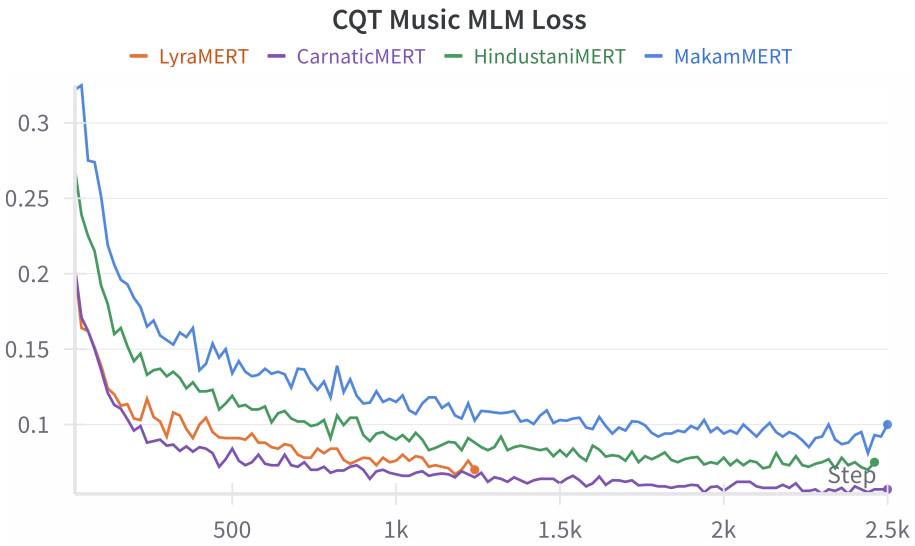
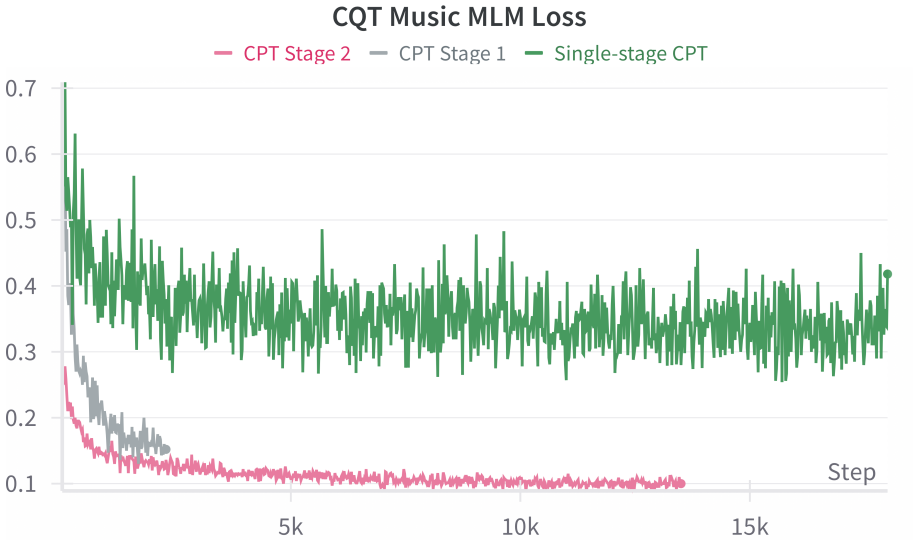


Figure 4.9. ***Gradient Norm Comparison: Two-stage vs. Single-stage CPT.*** *The two-stage CPT strategy stabilizes gradient updates more effectively, maintaining consistently lower and smoother gradient norms throughout training. In contrast, single-stage CPT exhibits sharp oscillations and occasional spikes, indicating unstable optimization and potential gradient explosions that can lead to training crashes.*

Overall, our proposed two-stage training strategy with re-warmed learning rates proves crucial in our setting for maintaining training stability when adapting to culturally diverse data distributions, without exhibiting forgetting.

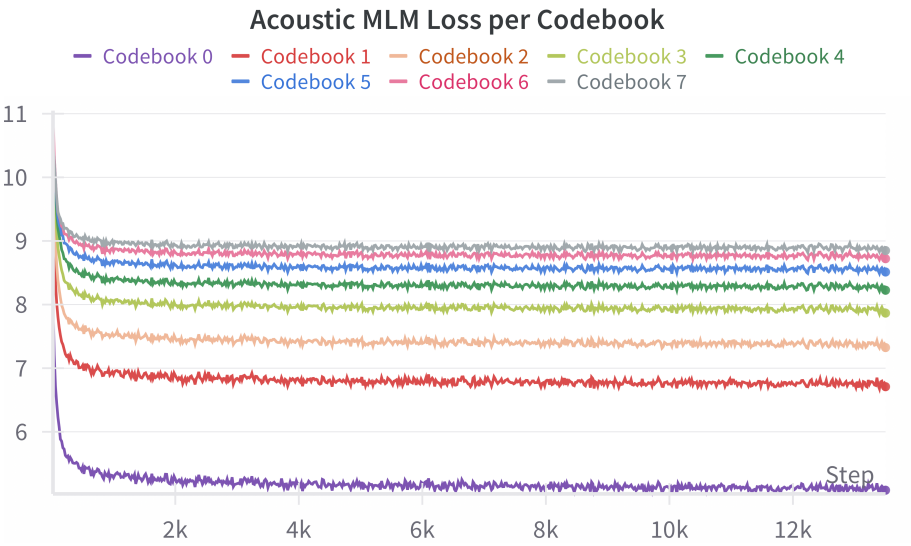


(a) ***Musical MLM Loss Across Cultures.*** *CQT reconstruction loss curves during two-stage CPT (Stage 2) across different single-culture adaptations. Convergence behavior varies by dataset, with Carnatic and Lyra achieving the lowest final loss.*

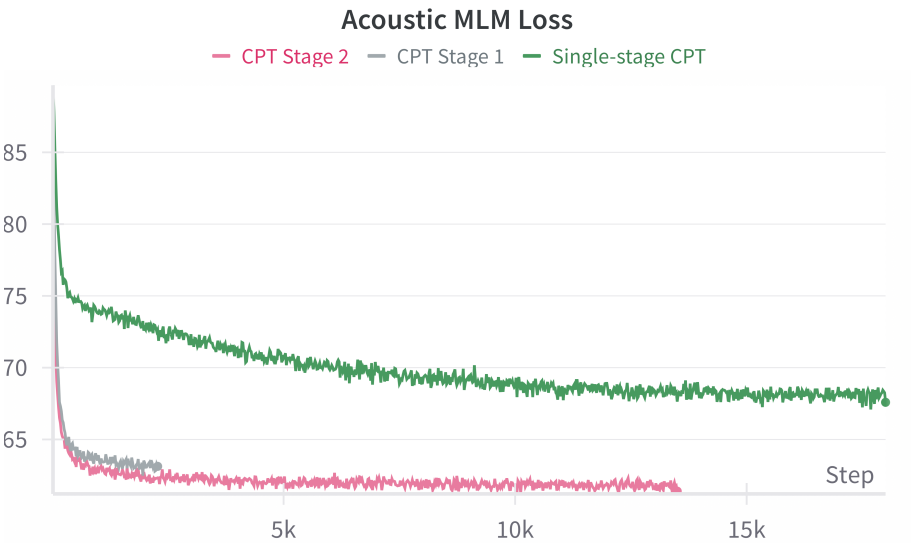


(b) ***Two-stage vs. Single-stage Multi-Cultural CPT.*** *The two-stage approach achieves lower and more stable loss due to the initial Stage 1 stabilization phase, which enables smoother full-parameter training in Stage 2. In contrast, full-parameter adaptation at once exhibits greater fluctuations and higher final loss, indicating less stable convergence.*

Figure 4.10. ***Musical MLM Loss During Continual Pre-Training.*** *Subfigure (a) shows loss curves for two-stage CPT across different cultures, while (b) compares overall training dynamics between single-stage and two-stage CPT on the multi-cultural dataset.*



(a) ***Acoustic MLM Loss per Codebook.*** *Loss curves for individual codebooks reveal varying convergence behavior in multi-cultural CPT, with earlier codebooks (e.g., codebooks 0–2) achieving lower losses, while deeper codebooks (e.g., codebooks 6 and 7) plateau at higher values, indicating greater modeling difficulty.*



(b) ***Two-stage vs. Single-stage Multi-Cultural CPT.*** *The second stage in the two-stage approach enables further loss reduction, as Stage 1 calibrates representations for smoother initialization. In contrast, single-stage adaptation exhibits higher final loss and slower convergence.*

Figure 4.11. ***Acoustic MLM Loss During Continual Pre-Training.*** *Subfigure (a) illustrates loss behavior across individual EnCodec codebooks, while (b) compares overall training dynamics between single-stage and two-stage CPT on the multi-cultural dataset.*

Chapter 5

Conclusions, Limitations, and Future Work

5.1 Conclusions

In this thesis, we explore cross-cultural music representation learning and introduce `CultureMERT-95M`, a multi-culturally adapted foundation model developed via continual pre-training (CPT) on diverse non-Western musical traditions. We propose a two-stage CPT strategy that incorporates learning rate re-warming and staged adaptation, enabling stable training under limited computational resources. Our results demonstrate that `CultureMERT` consistently outperforms the initial pre-trained `MERT-95M` model across diverse non-Western music tagging tasks, surpassing previously reported state-of-the-art, while preserving performance on "Western"-centric benchmarks.

We further examine how models adapted to specific cultural datasets perform and transfer to other cultural domains. Interestingly, cross-cultural evaluation reveals that transferability varies across musical traditions and aligns with known theoretical similarities from ethnomusicology, offering a novel computational perspective on cultural relationships. Notably, these results correlate with token-level similarity metrics between cultural datasets, such as Jensen-Shannon divergence and cosine distance computed over acoustic token distributions extracted from the EnCodec codec model, suggesting that such metrics can predict positive cross-cultural transfer, in line with findings from prior work in text and speech domains. These similarity metrics may also serve as heuristics for refining the composition of pre-training data in continual pre-training, tailored to the target evaluation task or cultural context.

Continual pre-training on a culturally diverse dataset comprising all studied non-Western traditions (i.e., `CultureMERT`) consistently yields the best overall performance, enhancing cross-cultural generalization compared to single-culture adaptations. Additionally, we investigate task arithmetic, which offers a strong alternative to multi-cultural CPT, effectively merging culturally specialized models, obtained via single-culture continual pre-training, in weight space and mitigating catastrophic forgetting. Task arithmetic, `CultureMERT-TA`, performs on par with `CultureMERT` on non-Western tasks, while also demonstrating strong performance on Western datasets, interestingly even surpassing the original pre-trained model in some cases.

Overall, this investigation contributes to the development of culturally aware foundation models for music and is, to our knowledge, the first to apply and validate continual

pre-training and model merging techniques—originally introduced in other domains—in the context of music audio representation learning, paving the way toward universal music representations. Our study aligns with the broader goal of creating computational methods that respect cultural diversity and ethical considerations, while offering a lens for comparison and knowledge transfer across world music traditions. Finally, this work lays a foundation for future cross-cultural MIR research, encouraging the development of inclusive foundation models that generalize across underrepresented musical traditions.

5.2 Limitations

5.2.1 Model and Scaling Considerations

While our exploration shows promising results, several limitations remain. The MERT model relies on the frozen EnCodec audio tokenizer for its self-supervised acoustic MLM pre-training, which is trained on Western music, making it potentially suboptimal for encoding culturally diverse musical languages. This limitation could affect the representational granularity for non-Western traditions, motivating future work on adapting or re-training audio tokenizers to better align with cultural diversity. Furthermore, we propose a computationally efficient two-stage continual pre-training strategy. However, while effective and crucial under constrained resources, future work could explore whether such staged adaptation remains necessary when scaling to larger computational budgets (e.g., increased batch sizes) or model sizes. Additionally, our continual pre-training strategy is specifically tailored to the MERT architecture, and future work should explore extending and applying the proposed two-stage CPT framework to other foundation models for music. This study also did not explore the effects of pre-training with more extensive datasets; a more fine-grained investigation into the impact of data volume, increased context lengths, and scaling model size (e.g., using the 330M MERT variant), could yield deeper insights into data–model trade-offs. Following [1], the input context length during pre-training was limited to 5 seconds, constraining the model's ability to capture the long-range dependencies inherent in music signals. This limitation may hinder performance on tasks requiring extended musical context (for example, music structure analysis), highlighting the need for future research into long-sequence modeling in FMs. Moreover, examining the impact of cultural composition in the pre-training data, such as training on specific subsets of musical traditions (e.g., Carnatic-Hindustani data mix), would help clarify how different cultural combinations influence transferability and performance in cross-cultural adaptation. Analyzing transfer behavior and synergy among cultural pairs, especially in relation to the token-level similarity metrics we examined, could illuminate which traditions benefit most from co-training, particularly in low-resource settings where underrepresented traditions may gain from leveraging culturally related data.

5.2.2 Datasets and Evaluation

Our investigation focuses on four non-Western musical traditions, Carnatic, Hindustani, Turkish-makam, and Greek folk, leaving other genres within the CompMusic Corpora

[184, 62], such as Beijing Opera and Andalusi classical music, unexplored. Furthermore, our evaluation is limited to sequence-level tasks, focusing specifically on the music automatic tagging task. However, many MIR applications require predictions at finer (e.g., frame-level) temporal resolutions. In this context, robust foundation models for music should also demonstrate strong capabilities in token-level classification tasks [19], such as beat/downbeat tracking, structure analysis, and chord recognition, which also typically require modeling longer-term temporal contexts. For non-Western traditions, several existing CompMusic datasets offer opportunities for such evaluations: the Carnatic Music Rhythm Dataset and Hindustani Music Rhythm Dataset include time-aligned taala and taal cycle markers, respectively, which are useful for rhythm analysis; the Mridangam Stroke Dataset provides individual percussive stroke recordings suitable for stroke classification tasks; the Tabla Solo Dataset offers time-aligned syllabic scores and audio recordings of solo performances, facilitating studies in syllabic percussion patterns and structural segmentation; and the Turkish Makam Melodic Phrase Dataset and Annotated Jingju Arias Dataset contain structural phrase-level annotations for segmentation and phrase boundary detection. Additional culturally diverse datasets reviewed in Section 2.4, including those from Chinese, Persian, African, and other musical traditions, present further opportunities for expanding cross-cultural evaluation. Collectively, these resources open up promising directions for extending foundation model evaluation beyond tagging in culturally diverse settings.

It is also important to acknowledge that evaluation practices in MIR have been criticized for inconsistencies in experimental protocols, data leakage, and weak construct validity, factors that can undermine the generalizability and interpretability of models' performance [151]. In particular, data leakage can significantly impact a model's performance by artificially inflating its evaluation results. While we follow standard evaluation protocols by adhering to dataset-provided train/test splits for continual pre-training and probing, recent work on transfer learning has shown that subtle forms of data leakage may still arise even under seemingly valid partitioning strategies [152]. For example, overlaps in musical artists, recording conditions, or culturally specific instrumentation between training and evaluation domains may unintentionally introduce spurious correlations, leading to shortcut learning [185]. As such, exploring more robust evaluation designs that avoid any cross-contamination between training and testing domains remains an important open challenge.

5.2.3 Cultural Framing and Interpretive Scope

We acknowledge the limitations of framing music within a "Western" versus "non-Western" dichotomy. While such terminology is commonly used in computational research for convenience, it risks oversimplifying the diversity of musical traditions. The concept of "non-Western" inherently groups together vastly different cultures, where categorizing "Western" music as a singular entity neglects its own internal diversity. Musical cultures exist on a continuum shaped by historical exchanges and regional adaptations rather than strict geographical divisions, and such classifications should be interpreted with caution.

Additionally, this work does not aim to establish or analyze cross-cultural similarities from an ethnomusicological perspective. Computational approaches inherently operate within a data-driven framework and are not explicitly informed by the historical, theoretical, or cultural knowledge embedded in the studied traditions. Nevertheless, our study may offer a novel lens on cross-cultural transfer patterns that emerge directly from the data through deep learning methods, contributing also to the development of more robust and culturally aware computational models for music. Moreover, our analysis on cross-cultural transferability should be considered in light of potential limitations in the representativeness and coverage of the datasets used. For example, as highlighted by [186], the datasets used in this thesis for Carnatic and Hindustani music may lack crucial aspects of these traditions, including instrumental compositions, improvisational elements, and performance context, leading to an incomplete representation in computational studies.

5.3 Future Work

Several interesting avenues for future research and potential extensions of this work can be identified:

- **Scale up** to more cultures, more data, and larger models (MERT-330M large). Experiment with other model architectures beyond MERT, for example, MusicFM [19], MuQ [100], SoniDo [101], and YuE [133]. Conduct more ablations and explore scaling laws in music audio foundation models.

- Train from scratch **multicultural foundation models** for music by exploring large-scale pre-training strategies across diverse musical traditions, and compare their performance to continual pre-training and model merging approaches. Future work should also examine how cultural selection and data mixture proportions during (continual) pre-training influence downstream generalization, akin to recent work in cross-lingual transfer [97, 98, 172]. This includes analyzing **cross-cultural transfer** dynamics at scale and identifying potential "super-donor" and "super-recipient" musical traditions that consistently enhance or benefit from transfer.

- Evaluate culturally adapted models beyond sequence-level tasks, such as music auto-tagging. Extend to **token-level tasks** (e.g., beat tracking, source separation) and **multimodal tasks** (e.g., music–language models such as MusiLingo [102], LLark [134], and CLaMP 3 [71]) in **zero-shot settings** and culturally diverse contexts. A promising direction is to construct a joint music–text embedding model using our culturally adapted music encoders as audio backbones, fine-tuned via contrastive learning in the style of MuLan [95]. Replacing the frozen MERT audio encoder in frameworks like CLaMP 3 with `CultureMERT` could improve generalization to under-represented musical cultures. Such models could support zero-shot cultural music understanding tasks such as culturally informed captioning, retrieval, and cross-modal tagging. Moreover, they would advance interpretability and open new directions for cross-cultural music–language alignment. Additionally, since the use of pre-trained

encoders is common in modern generative models, `CultureMERT` could also serve as a conditioning audio encoder in generative frameworks such as MusicGen or SoniDo.

- Future work could explore **parameter-efficient fine-tuning (PEFT)** techniques (e.g., adapter-based methods such as LoRA [187]) as lightweight alternatives to full-parameter adaptation in CPT, or as complementary approaches, particularly for supervised fine-tuning (SFT) after the CPT phase, instead of relying solely on probing-based evaluation [188]. In this context, few-shot and low-resource setups could be further investigated through multi-label few-shot learning approaches, such as *"LC-Protonets"* [79], which has demonstrated strong performance under challenging evaluation settings on low-resource world music tagging tasks. Additionally, future work could explore **mixed-objective** training strategies that combine self-supervised learning and supervised fine-tuning in a joint multi-task learning framework, similar to the M2DS2 [189] and UDALM [190] approaches proposed in the speech and text domains.

- Adapt existing **audio tokenizers** to non-Western music. Current models (for example, EnCodec used in this study, DAC [120]) are trained on Western data, which may limit their effectiveness in encoding culturally diverse musical languages, especially when integrated into model architectures as audio tokenizers [144, 191, 58, 131], or as acoustic teachers [1, 192]. Recently, UniCodec [193] introduced a single-domain-adaptive codebook and domain Mixture-of-Experts strategy to unify audio modeling across speech, music, and environmental sound, representing a promising direction for **culturally adaptive tokenization**.

- Explore different **music teachers** for self-supervision: The current 336-bin CQT reconstruction loss with 48 bins/octave encodes Western assumptions like octave equivalence and equal temperament, which may bias self-supervision against non-Western traditions with microtonality or unequal tuning. Future work could explore culturally appropriate alternatives such as trainable filterbanks [194], non-stationary Gabor transforms with variable resolution [195], or data-driven frequency scales derived from pitch distributions in each culture [196], to provide richer and less biased supervisory signals for cross-cultural music representation learning.

- Current audio tokenization schemes (e.g., NACs such as EnCodec [2], or wav2vec2-style feature extractors [1]) tokenize audio into fixed-length frames (e.g., 13.3 ms per token at a 75 Hz frame rate), imposing a **uniform** temporal resolution that may fail to capture the adaptive timing structures characteristic of many musical traditions. This rigidity is particularly limiting in expressive performance styles involving tempo rubato, improvisation, or free rhythm, especially in non-Western traditions, where timing naturally deviates from a strict metrical grid. Moreover, fixed-rate tokenization can over-segment "musically predictable" or slow-evolving passages (e.g., a sustained tone or steady rhythmic pulse), while lacking the flexibility to compress or expand time **dynamically** based on musical content, thereby introducing unnecessary computational overhead in regions with low information density. Many music

traditions also differ in their temporal structure and rhythmic granularity, further motivating the need for adaptive tokenization strategies. An interesting and novel research direction would be to adapt the **Byte Latent Transformer (BLT)** [197] to music, enabling variable-length adaptive music tokenization with *entropy patching*. BLT could allow models to allocate more attention to complex or unpredictable musical segments (e.g., ornaments, modulations), better handling regions of different information densities, and improving **representation fidelity**. Moreover, it could learn variable-length patterns specific to each tradition and avoid fixed-token biases from Western-trained audio tokenizers. Finally, BLT's representation may be **more interpretable** in terms of musical structure compared to low-level acoustic tokens. This interpretability could be explored by analyzing whether the model's learned patch boundaries align with known musical event boundaries, such as phrase transitions, onsets, or changes in instrumentation or harmony. Such an architecture could be adapted and utilized for both music understanding tasks and generation.

- Investigate **advanced model merging** techniques and task arithmetic variants that address task conflicts and parameter interference (e.g., AdaMerging [6], TIES [177], DELLA [198], CART [111], AWD [176], Adaptive Projective Gradient Descent [199], OPCM [200], TSV-Merge [201], AdaRank [202]) to combine specialized models more effectively. These methods utilize **adaptive, task- and layer-wise scaling factors** for task arithmetic, rather than a fixed global λ, whose selection is often sensitive—as demonstrated in our findings—thus better managing merging conflicts. Additionally, they incorporate techniques such as pruning, orthogonalization and disentanglement, and low-rank subspace methods to mitigate parameter interference and improve multi-task compatibility. These techniques could also be applied to merge models trained on different music understanding tasks (for example, combining beat tracking and music auto-tagging expert models).

- Explore **dynamic data mixtures** during continual pre-training (e.g., DoGE [203], DoReMi [204], RegMix [205], D-CPT [206]) to balance domain influence. Additionally, compare multi-cultural CPT with **curriculum/incremental learning** strategies [31] progressively adapting the model to different musical cultures (e.g., Lyra $\rightarrow$ Hindustani $\rightarrow$ Carnatic $\rightarrow$ ...). In such staged settings, **infinite or meta learning rate schedules** [31] could be employed to improve training stability and reduce unwanted forgetting. Moreover, curriculum-based CPT could be evaluated against model merging approaches, such as task arithmetic explored in this thesis, which in general eliminate the need for multiple adaptation steps [207], or **hybrid approaches** such as MagMax [208], BECAME [209], and Branch-and-Merge [210], which combine sequential adaptation with model merging to consolidate cross-cultural knowledge more effectively and mitigate catastrophic forgetting.

- Extend to training and adaptation of **music generation models** [58, 3, 101, 133] and explore the applicability of our paradigm in the context of **multimodal models** (e.g., music–language models) in cross-cultural settings.

- Despite their effectiveness, foundation models for music often function as black boxes. However, **interpretability** is essential for fair, ethical, and trustworthy AI, especially in cultural contexts. As shown in Section 4.5, we observed that different transformer layers in `CultureMERT` yielded the best results when evaluated via probing on music tagging tasks across different musical cultures, suggesting that culturally relevant information may be encoded at varying depths within the adapted multi-cultural model. Possible research directions on the interpretability of music FMs include using probing classifiers and intervention techniques [211], saliency maps, attention analysis, and clustering learned embeddings (e.g., by abstract cross-cultural musical concepts, or culture-specific elements). Furthermore, recently, [212] proposed transforming CLAP audio embeddings into sparse, concept-based representations aligned with human-interpretable audio concepts, shown to retain or even improving performance on downstream tasks. Applying similar post-hoc transformations to `CultureMERT` could provide insight into how cultural information is semantically encoded and which interpretable dimensions drive predictions across musical traditions. Moreover, exploring **explainable AI (XAI)** techniques tailored to the music domain, such as counterfactual explanations (audio-level, latent-space, or generative), could further enhance our understanding of these models and make AI decisions more transparent and aligned with human musical understanding [158].

Chapter 6

Ethics Statement and Responsible Use

Careful consideration is advised before deploying the models developed and presented in this work in real-world contexts, as they may still reflect cultural and dataset-specific biases. Some of the datasets used are not publicly available and were accessed under research-use agreements. Any released models should not be used for commercial or generative applications without explicit attention to cultural representation, appropriate licensing, and the consent of the relevant communities or dataset curators.

Appendices

Appendix A

Training and Evaluation Settings for `MERT-v1-95M`

This appendix details the training settings and hyperparameters used for training the `MERT-v1-95M` model, as well as the downstream evaluation on music auto-tagging tasks under the MARBLE constrained protocol.

A.1 Training Settings

The core settings for training `MERT-v1-95M` are summarized as follows:

- **Audio sampling rate:** 24 kHz
- **Input segment length:** 5 seconds, randomly cropped during training
- **Feature extractor:** 7-layer 1D CNN with GELU activations and GroupNorm in the first layer; no internal dropout. Architecture: [(512,10,5)] + [(512,3,2)]×4 + [(512,2,2)]×2, producing frame-level representations at 75 Hz
- **Transformer encoder:** 12-layer Transformer with 768-dimensional embeddings, 12 attention heads, and 3072-dimensional feed-forward layers
- **Projection layers:** CNN outputs are projected to the transformer input dimension (768) via a linear layer; transformer outputs are projected to a 64-dimensional space to match the dimensionality of codeword embeddings; separate final projections are used for each codebook
- **Positional embeddings:** One convolutional positional embedding layer with 128 filters and 16 groups
- **Masking:** Random masking applied at the frame level (after the CNN feature extractor), with 80% probability and a mask length of 5 frames; loss is computed only on masked frames
- **CQT prediction:** Auxiliary objective to predict CQT spectrograms with 336 bins from masked frames
- **Gradient scaling:** Feature extractor gradients are scaled by a factor of 0.1 during training

- **Dropout settings:** 0.1 for input, features, encoder layers, and attention modules; 0.0 for activation; encoder layerdrop probability is set to 0.05.
- **Codebook usage:** All 8 codebooks are jointly predicted during masked prediction, instead of randomly accessing a subset of codebooks per batch, leading to faster convergence. An ablation on random 4-codebook sampling is presented in Section 4.6.2 (Table 4.4).
- **Temperature scaling:** Fixed at 0.1 for contrastive prediction loss
- **Data augmentation:** In-batch noise mixture augmentation applied with probability 0.5

A.2 Evaluation Settings

Evaluation follows the MARBLE probing-based constrained protocol: backbone models are frozen, probing heads are shallow (single-layer MLP), and hyperparameters are selected from a restricted grid search space. Performance on auto-tagging tasks is measured using macro-averaged ROC-AUC, mean Average Precision (mAP), Micro-F1, and Macro-F1 scores. The detailed evaluation settings are:

- **Classifier:** Single-layer MLP with 512 hidden units and ReLU activation, trained on top of frozen MERT representations
- **Backbone feature extraction:** Task-specific selection of either a single transformer layer or a learnable weighted sum over all layers (see Section 4.5 for details and discussion on task-specific feature selection)
- **Batch size:** 64
- **Learning rate:** Chosen per task from {5e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2}
- **Dropout probability:** 0.2
- **Optimizer:** Adam [213] with default parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$)
- **Learning rate scheduler:** ReduceLROnPlateau, with patience typically set to 3 epochs (task-specific)
- **Early stopping:** Enabled, with patience typically set to 10 epochs (task-specific)
- **Training epochs:** Up to 100

Hyperparameters and training strategies that are explicitly discussed in Section 3 (e.g., optimizer settings and learning rate schedules during continual pre-training) are not repeated here for brevity.

Bibliography

[1] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos *et al.*, "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.

[2] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Trans. Mach. Learn. Res.*, vol. 2023, 2023.

[3] A. Mehta, S. Chauhan, A. Djanibekov, A. Kulkarni, G. Xia, and M. Choudhury, "Music for all: Exploring multicultural representations in music generation models," *CoRR*, vol. abs/2502.07328, 2025.

[4] A. Ibrahim, B. Thérien, K. Gupta, M. L. Richter, Q. G. Anthony, E. Belilovsky, T. Lesort, and I. Rish, "Simple and scalable strategies to continually pre-train large language models," *Trans. Mach. Learn. Res.*, vol. 2024, 2024.

[5] G. Ilharco, M. T. Ribeiro, M. Wortsman, L. Schmidt, H. Hajishirzi, and A. 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.

[6] E. Yang, Z. Wang, L. Shen, S. Liu, G. Guo, X. Wang, and D. Tao, "Adamerging: Adaptive model merging for multi-task learning," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[7] I. A. P. Santana, F. Pinhelli, J. Donini, L. G. Catharin, R. B. Mangolin, Y. M. e Gomes da Costa, V. D. Feltrim, and M. A. 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.

[8] N. Jacoby, R. Polak, J. A. Grahn, D. J. Cameron, K. M. Lee, R. Godoy, E. A. Undurraga, T. Huanca, T. Thalwitzer, N. Doumbia *et al.*, "Commonality and variation in mental representations of music revealed by a cross-cultural comparison of rhythm priors in 15 countries," *Nature Human Behaviour*, vol. 8, no. 5, May 2024, pp. 846–877, publisher: Nature Publishing Group.

[9] S. A. Mehr, M. Singh, D. Knox, D. M. Ketter, D. Pickens-Jones, S. Atwood, C. Lucas, N. Jacoby, A. A. Egner, E. J. Hopkins *et al.*, "Universality and diversity in human song," *Science*, vol. 366, no. 6468, Nov. 2019, p. eaax0868.

[10] N. Jacoby, E. H. Margulis, M. Clayton, E. Hannon, H. Honing, J. Iversen, T. R. Klein, S. A. Mehr, L. Pearson, I. Peretz *et al.*, "Cross-Cultural Work in Music Cognition: Challenges, Insights, and Recommendations," *Music perception*, vol. 37, no. 3, Feb. 2020, pp. 185–195.

[11] Y. Ma, A. Øland, A. Ragni, B. M. D. Sette, C. Saitis, C. Donahue, C. Lin, C. Plachouras, E. Benetos, E. Quinton *et al.*, "Foundation models for music: A survey," *CoRR*, vol. abs/2408.14340, 2024.

[12] P. E. Savage, S. Brown, E. Sakai, and T. E. Currie, "Statistical universals reveal the structures and functions of human music," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 29, Jul. 2015, pp. 8987–8992.

[13] S. E. Trehub, J. Becker, and I. Morley, "Cross-cultural perspectives on music and musicality," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 370, no. 1664, Mar. 2015, p. 20140096.

[14] E. H. Margulis, P. C. M. Wong, C. Turnbull, B. M. Kubit, and J. D. McAuley, "Narratives imagined in response to instrumental music reveal culture-bounded intersubjectivity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 119, no. 4, Jan. 2022, p. e2110406119.

[15] H. Lee, E. Çelen, P. Harrison, M. Anglada-Tort, P. van Rijn, M. Park, M. Schönwiesner, and N. Jacoby, "Globalmood: A cross-cultural benchmark for music emotion recognition," 2025.

[16] J. S. Downie, "Music information retrieval," *Annual Review of Information Science and Technology*, vol. 37, no. 1, Jan. 2003, pp. 295–340.

[17] R. Bommasani, D. A. Hudson, E. Adeli, R. B. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *CoRR*, vol. abs/2108.07258, 2021.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," Aug. 2023, arXiv:1706.03762 [cs].

[19] M. Won, Y. Hung, and D. Le, "A foundation model for music informatics," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*. IEEE, 2024, pp. 1226–1230.

[20] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *CoRR*, vol. abs/2005.00341, 2020.

[21] E. Gómez, P. Herrera, and F. Gómez-Martin, "Computational Ethnomusicology: perspectives and challenges," *Journal of New Music Research*, vol. 42, no. 2, June 2013, pp. 111–112.

[22] T. Lidy, C. N. S. Jr., O. Cornelis, F. Gouyon, A. Rauber, C. A. A. Kaestner, and A. 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," *Signal Process.*, vol. 90, no. 4, 2010, pp. 1032–1048.

[23] G. Plaja-Roglans, T. Nuttall, L. Pearson, X. Serra, and M. Miron, "Repertoire-specific vocal pitch data generation for improved melodic analysis of carnatic music," *Trans. Int. Soc. Music. Inf. Retr.*, vol. 6, no. 1, 2023, pp. 13–26.

[24] G. K. Koduri, M. Miron, J. Serrà, and X. 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*, A. Klapuri and C. Leider, Eds. University of Miami, 2011, pp. 263–268.

[25] A. Ferraro, G. Ferreira, F. Diaz, and G. Born, "Measuring commonality in recommendation of cultural content to strengthen cultural citizenship," *Trans. Recomm. Syst.*, vol. 2, no. 1, 2024, pp. 10:1–10:32.

[26] A. Holzapfel, B. L. Sturm, and M. Coeckelbergh, "Ethical dimensions of music information retrieval technology," *Trans. Int. Soc. Music. Inf. Retr.*, vol. 1, no. 1, 2018, pp. 44–55.

[27] C. C. Liu, I. Gurevych, and A. Korhonen, "Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art," *CoRR*, vol. abs/2406.03930, 2024.

[28] M. DeHaven and J. Billa, "Improving low-resource speech recognition with pretrained speech models: Continued pretraining vs. semi-supervised training," *CoRR*, vol. abs/2207.00659, 2022.

[29] D. M. Alves, J. Pombal, N. M. Guerreiro, P. H. Martins, J. Alves, M. A. Farajian, B. Peters, R. Rei, P. Fernandes, S. Agrawal *et al.*, "Tower: An open multilingual large language model for translation-related tasks," *CoRR*, vol. abs/2402.17733, 2024.

[30] L. Voukoutis, D. Roussis, G. Paraskevopoulos, S. Sofianopoulos, P. Prokopidis, V. Papavasileiou, A. Katsamanis, S. Piperidis, and V. Katsouros, "Meltemi: The first open large language model for greek," *CoRR*, vol. abs/2407.20743, 2024.

[31] V. Udandarao, K. Roth, S. Dziadzio, A. Prabhu, M. Cherti, O. Vinyals, O. J. Hénaff, S. Albanie, Z. Akata, and M. Bethge, "A practitioner's guide to real-world continual multimodal pretraining," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, Eds., 2024.

[32] S. Gururangan, A. Marasovic, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. 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*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 8342–8360.

[33] K. Fujii, T. Nakamura, M. Loem, H. Iida, M. Ohi, K. Hattori, H. Shota, S. Mizuki, R. Yokota, and N. Okazaki, "Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities," *CoRR*, vol. abs/2404.17790, 2024.

[34] K. Gupta, B. Thérien, A. Ibrahim, M. L. Richter, Q. Anthony, E. Belilovsky, I. Rish, and T. Lesort, "Continual pre-training of large language models: How to (re)warm your model?" *CoRR*, vol. abs/2308.04014, 2023.

[35] W. Zheng, W. Pan, X. Xu, L. Qin, L. Yue, and M. Zhou, "Breaking language barriers: Cross-lingual continual pre-training at scale," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Y. Al-Onaizan, M. Bansal, and Y. Chen, Eds. Association for Computational Linguistics, 2024, pp. 7725–7738.

[36] A. Cossu, A. Carta, L. C. Passaro, V. Lomonaco, T. Tuytelaars, and D. Bacciu, "Continual pre-training mitigates forgetting in language and vision," *Neural Networks*, vol. 179, 2024, p. 106492.

[37] K. Nowakowski, M. Ptaszynski, K. Murasaki, and J. Nieuwazny, "Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining," *Inf. Process. Manag.*, vol. 60, no. 2, 2023, p. 103148.

[38] N. San, G. Paraskevopoulos, A. Arora, X. He, P. Kaur, O. Adams, and D. 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*, M. Hahn, A. Sorokin, R. Kumar, A. Scherbakov, Y. Otmakhova, J. Yang, O. Serikov, P. Rani, E. M. Ponti, S. Muradoglu *et al.*, Eds. Association for Computational Linguistics, 2024, pp. 100–112.

[39] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. G. Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith *et al.*, "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*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 23 965–23 998.

[40] E. Yang, L. Shen, G. Guo, X. Wang, X. Cao, J. Zhang, and D. Tao, "Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities," *CoRR*, vol. abs/2408.07666, 2024.

[41] G. Stoica, D. Bolya, J. Bjorner, P. Ramesh, T. Hearn, and J. Hoffman, "Zipit! merging models from different tasks without training," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[42] X. Jin, X. Ren, D. Preotiuc-Pietro, and P. Cheng, "Dataless knowledge fusion by merging weights of language models," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[43] D. Li, Y. Ma, W. Wei, Q. Kong, Y. Wu, M. Che, F. Xia, E. Benetos, and W. 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.

[44] J. Kirkpatrick, R. Pascanu, N. C. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *CoRR*, vol. abs/1612.00796, 2016.

[45] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, 2024, pp. 5362–5383.

[46] M. Müller, *Information retrieval for music and motion*. Springer, 2007.

[47] M. Müller, *Fundamentals of Music Processing - Audio, Analysis, Algorithms, Applications*. Springer, 2015.

[48] E. J. Humphrey, J. P. Bello, and Y. LeCun, "Feature learning and deep architectures: new directions for music informatics," *J. Intell. Inf. Syst.*, vol. 41, no. 3, 2013, pp. 461–481.

[49] H. Wu, C. Kao, Q. Tang, M. Sun, B. McFee, J. P. Bello, and C. Wang, "Multi-task self-supervised pre-training for music classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, 2021, pp. 556–560.

[50] H. Zhu, Y. Niu, D. Fu, and H. Wang, "Musicbert: A self-supervised learning of music representation," in *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, H. T. Shen, Y. Zhuang, J. R. Smith, Y. Yang, P. César, F. Metze, and B. Prabhakaran, Eds. ACM, 2021, pp. 3955–3963.

[51] A. N. Carr, Q. Berthet, M. Blondel, O. Teboul, and N. Zeghidour, "Self-supervised learning of audio representations from permutations with differentiable ranking," *IEEE Signal Process. Lett.*, vol. 28, 2021, pp. 708–712.

[52] J. Spijkervet and J. A. Burgoyne, "Contrastive learning of musical representations," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, J. H. Lee, A. Lerch, Z. Duan, J. Nam, P. Rao, P. van Kranenburg, and A. Srinivasamurthy, Eds., 2021, pp. 673–681.

[53] A. Ragano, E. Benetos, and A. Hines, "Learning music representations with wav2vec 2.0," in *31st Irish Conference on Artificial Intelligence and Cognitive Science, AICS 2023, Letterkenny, Ireland, December 7-8, 2023*. IEEE, 2023, pp. 1–6.

[54] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186.

[55] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, 2021, pp. 3451–3460.

[56] C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu, "Self-supervised learning with random-projection quantizer for speech recognition," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 3915–3924.

[57] R. Castellon, C. Donahue, and P. Liang, "Codified audio language modeling learns useful representations for music information retrieval," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, J. H. Lee, A. Lerch, Z. Duan, J. Nam, P. Rao, P. van Kranenburg, and A. Srinivasamurthy, Eds., 2021, pp. 88–96.

[58] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.

[59] X. Serra, "A multicultural approach in music information research," in *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011*, A. Klapuri and C. Leider, Eds. University of Miami, 2011, pp. 151–156.

[60] G. Born, "Diversifying MIR: knowledge and real-world challenges, and new interdisciplinary futures," *Trans. Int. Soc. Music. Inf. Retr.*, vol. 3, no. 1, 2020, pp. 193–204.

[61] L. S. Maia, M. Fuentes, L. W. P. Biscainho, M. Rocamora, and S. Essid, "SAMBASET: A dataset of historical samba de enredo recordings for computational music analysis," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, A. Flexer, G. Peeters, J. Urbano, and A. Volk, Eds., 2019, pp. 628–635.

[62] X. Serra, "Creating research corpora for the computational study of music: the case of the compmusic project," in *AES International Conference on Semantic Audio 2014, London, UK, January 27-29, 2014*, C. Dittmar, G. Fazekas, and S. Ewert, Eds. Audio Engineering Society, 2014.

[63] N. Kroher, J. M. Díaz-Báñez, J. Mora, and E. Gómez, "Corpus COFLA: A research corpus for the computational study of flamenco music," *ACM Journal on Computing and Cultural Heritage*, vol. 9, no. 2, 2016, pp. 10:1–10:21.

[64] B. Baba Ali, A. Gorgan Mohammadi, and A. Faraji Dizaji, "Nava: A Persian Traditional Music Database for the Dastgah and Instrument Recognition Tasks," *Advanced Signal Processing*, vol. 3, no. 2, Nov. 2019, pp. 125–134, publisher: Vice Chancellery for Research and Technology, University of Tabriz.

[65] S. Konduri, K. V. Pendyala, and V. S. Pendyala, "KritiSamhita: A machine learning dataset of South Indian classical music audio clips with tonic classification," *Data in Brief*, vol. 55, Aug. 2024, p. 110730.

[66] S. Rosenzweig, F. Scherbaum, D. Shugliashvili, V. Arifi-Müller, and M. Müller, "Erkomaishvili dataset: A curated corpus of traditional georgian vocal music for computational musicology," *Trans. Int. Soc. Music. Inf. Retr.*, vol. 3, no. 1, 2020, pp. 31–41.

[67] C. Papaioannou, I. Valiantzas, T. Giannakopoulos, M. A. Kaliakatsos-Papakostas, and A. 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, December 4-8, 2022*, P. Rao, H. A. Murthy, A. Srinivasamurthy, R. M. Bittner, R. C. Repetto, M. Goto, X. Serra, and M. Miron, Eds., 2022, pp. 377–383.

[68] M. Zhou, S. Xu, Z. Liu, Z. Wang, F. Yu, W. Li, and B. Han, "Ccmusic: An open and diverse database for chinese music information retrieval research," *Trans. Int. Soc. Music. Inf. Retr.*, vol. 8, no. 1, 2025, pp. 22–38.

[69] O. Oguike and M. Primus, "A dataset for multimodal music information retrieval of Sotho-Tswana musical videos," *Data in Brief*, vol. 55, Aug. 2024, p. 110672.

[70] S. E. Moore, N. A. Asare, and S. K. Kubiti, "Ndwom: a multimodal music information retrieval dataset for akan musical videos," 2025.

[71] S. Wu, Z. Guo, R. Yuan, J. Jiang, S. Doh, G. Xia, J. Nam, X. Li, F. Yu, and M. Sun, "Clamp 3: Universal music information retrieval across unaligned modalities and unseen languages," *CoRR*, vol. abs/2502.10362, 2025.

[72] M. Panteli, "Computational analysis of world music corpora," Ph.D. dissertation, Queen Mary University of London, UK, 2018.

[73] E. Demirel, B. Bozkurt, and X. Serra, "Automatic makam recognition using chroma features," 2018.

[74] A. K. Sharma, G. Aggarwal, S. Bhardwaj, P. Chakrabarti, T. Chakrabarti, J. H. Abawajy, S. Bhattacharyya, R. Mishra, A. Das, and H. Mahdin, "Classification of indian classical music with time-series matching deep learning approach," *IEEE Access*, vol. 9, 2021, pp. 102 041–102 052.

[75] D. Ebrat and F. Didehvar, "Iranian modal music (dastgah) detection using deep neural networks," *CoRR*, vol. abs/2203.15335, 2022.

[76] D. Han, R. C. Repetto, and D. Jeong, "Finding tori: Self-supervised learning for analyzing korean folk song," in *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023, Milan, Italy, November 5-9, 2023*, A. Sarti, F. Antonacci, M. Sandler, P. Bestagini, S. Dixon, B. Liang, G. Richard, and J. Pauwels, Eds., 2023, pp. 440–447.

[77] K. L. Walls, I. R. Román, K. V. Ert, C. Harper, and L. Adu-Gilmore, "Analyzing pitch content in traditional ghanaian seperewa songs," *CoRR*, vol. abs/2411.08234, 2024.

[78] C. Papaioannou, E. Benetos, and A. 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, November 5-9, 2023*, A. Sarti, F. Antonacci, M. Sandler, P. Bestagini, S. Dixon, B. Liang, G. Richard, and J. Pauwels, Eds., 2023, pp. 311–318.

[79] C. Papaioannou, E. Benetos, and A. Potamianos, "LC-Protonets: Multi-label few-shot learning for world music audio tagging," *IEEE Open Journal of Signal Processing*, vol. 6, 2025, pp. 138–146.

[80] J. Melechovský, Z. Guo, D. Ghosal, N. Majumder, D. Herremans, and S. Poria, "Mustango: Toward controllable text-to-music generation," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, K. Duh, H. Gómez-Adorno, and S. Bethard, Eds. Association for Computational Linguistics, 2024, pp. 8293–8316.

[81] R. S. Huang, A. Holzapfel, B. L. T. Sturm, and A.-K. Kaila, "Beyond Diverse Datasets: Responsible MIR, Interdisciplinarity, and the Fractured Worlds of Music,"

Transactions of the International Society for Music Information Retrieval, vol. 6, no. 1, Jun. 2023.

[82] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music tagging," in *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009, Kobe International Conference Center, Kobe, Japan, October 26-30, 2009*, K. Hirata, G. Tzanetakis, and K. Yoshii, Eds. International Society for Music Information Retrieval, 2009, pp. 387–392.

[83] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," in *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, S. J. Cunningham, Z. Duan, X. Hu, and D. Turnbull, Eds., 2017, pp. 316–323.

[84] B. Uyar, H. S. Atli, S. Sentürk, B. Bozkurt, and X. Serra, "A corpus for computational research of turkish makam music," in *Proceedings of the 1st International Workshop on Digital Libraries for Musicology, DLfM@JCDL 2014, London, United Kingdom, September 12, 2014*, B. Fields and K. R. Page, Eds. ACM, 2014, pp. 1–7.

[85] S. Sentürk, "Computational analysis of audio recordings and music scores for the description and discovery of ottoman-turkish makam music," Ph.D. dissertation, Pompeu Fabra University, Spain, 2017.

[86] A. Srinivasamurthy, G. K. Koduri, S. Gulati, V. Ishwar, and X. Serra, "Corpora for music information research in indian art music," in *Music Technology meets Philosophy - From Digital Echos to Virtual Ethos: Joint Proceedings of the 40th International Computer Music Conference, ICMC 2014, and the 11th Sound and Music Computing Conference, SMC 2014, Athens, Greece, September 14-20, 2014*. Michigan Publishing, 2014.

[87] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

[88] J. Parmar, S. Satheesh, M. Patwary, M. Shoeybi, and B. Catanzaro, "Reuse, don't retrain: A recipe for continued pretraining of language models," *CoRR*, vol. abs/2407.07263, 2024.

[89] S. Hu, Y. Tu, X. Han, C. He, G. Cui, X. Long, Z. Zheng, Y. Fang, Y. Huang, W. Zhao *et al.*, "Minicpm: Unveiling the potential of small language models with scalable training strategies," *CoRR*, vol. abs/2404.06395, 2024.

[90] Y. Guo, J. Fu, H. Zhang, D. Zhao, and Y. Shen, "Efficient continual pre-training by mitigating the stability gap," *CoRR*, vol. abs/2406.14833, 2024.

[91] M. Mermillod, A. Bugaiska, and P. Bonin, "The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects," *Frontiers in Psychology*, vol. 4, Aug. 2013, publisher: Frontiers.

[92] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, 2019, pp. 54–71.

[93] D. Kim and B. 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. 20 196–20 204.

[94] M. Parovic, I. Vulic, and A. Korhonen, "Investigating the potential of task arithmetic for cross-lingual transfer," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 2: Short Papers, St. Julian's, Malta, March 17-22, 2024*, Y. Graham and M. Purver, Eds. Association for Computational Linguistics, 2024, pp. 124–137.

[95] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. W. Ellis, "Mulan: A joint embedding of music audio and natural language," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*, P. Rao, H. A. Murthy, A. Srinivasamurthy, R. M. Bittner, R. C. Repetto, M. Goto, X. Serra, and M. Miron, Eds., 2022, pp. 559–566.

[96] V. Blaschke, M. Fedzechkina, and M. ter Hoeve, "Analyzing the effect of linguistic similarity on cross-lingual transfer: Tasks and experimental setups matter," *CoRR*, vol. abs/2501.14491, 2025.

[97] D. Malkin, T. Limisiewicz, and G. Stanovsky, "A balanced data approach for evaluating cross-lingual transfer: Mapping the linguistic blood bank," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, M. Carpuat, M. de Marneffe, and I. V. M. Ruíz, Eds. Association for Computational Linguistics, 2022, pp. 4903–4915.

[98] V. Protasov, E. Stakovskii, E. Voloshina, T. Shavrina, and A. Panchenko, "Super donors and super recipients: Studying cross-lingual transfer between high-resource and low-resource languages," in *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, A. K. Ojha, C.-h. Liu, E. Vylomova, F. Pirinen, J. Abbott, J. Washington, N. Oco, V. Malykh, V. Logacheva, and X. Zhao, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 94–108.

[99] S. Delegos, "A Modal Heterotopia: Rethinking Makam Modality and Chordal Harmony in Interwar Rebetiko," *Yearbook for Traditional Music*, vol. 56, no. 1, Jul. 2024, pp. 81–104.

[100] H. Zhu, Y. Zhou, H. Chen, J. Yu, Z. Ma, R. Gu, Y. Luo, W. Tan, and X. Chen, "Muq: Self-supervised music representation learning with mel residual vector quantization," *CoRR*, vol. abs/2501.01108, 2025.

[101] W. Liao, Y. Takida, Y. Ikemiya, Z. Zhong, C. Lai, G. Fabbro, K. Shimada, K. Toyama, K. W. Cheuk, M. A. M. Ramírez *et al.*, "Music foundation model as generic booster for music downstream tasks," *CoRR*, vol. abs/2411.01135, 2024.

[102] Z. Deng, Y. Ma, Y. Liu, R. Guo, G. Zhang, W. Chen, W. Huang, and E. Benetos, "Musilingo: Bridging music and text with pre-trained language models for music captioning and query response," in *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, K. Duh, H. Gómez-Adorno, and S. Bethard, Eds. Association for Computational Linguistics, 2024, pp. 3643–3655.

[103] W. Li, Y. Cai, Z. Wu, W. Zhang, Y. Chen, R. Qi, M. Dong, P. Chen, X. Dong, F. Shi *et al.*, "A survey of foundation models for music understanding," *CoRR*, vol. abs/2409.09601, 2024.

[104] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, "Self-supervised representation learning: Introduction, advances, and challenges," *IEEE Signal Process. Mag.*, vol. 39, no. 3, 2022, pp. 42–62.

[105] E. Gogoulou, T. Lesort, M. Boman, and J. 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*, ser. Lecture Notes in Computer Science, E. Nöth, A. Horák, and P. Sojka, Eds., vol. 15048. Springer, 2024, pp. 71–84.

[106] H. Shi, Z. Xu, H. Wang, W. Qin, W. Wang, Y. Wang, and H. Wang, "Continual learning of large language models: A comprehensive survey," *CoRR*, vol. abs/2404.16789, 2024.

[107] A. Alexandrov, V. Raychev, D. I. Dimitrov, C. Zhang, M. T. Vechev, and K. Toutanova, "Bggpt 1.0: Extending english-centric llms to other languages," *CoRR*, vol. abs/2412.10893, 2024.

[108] D. Roussis, L. Voukoutis, G. Paraskevopoulos, S. Sofianopoulos, P. Prokopidis, V. Papavasileiou, A. Katsamanis, S. Piperidis, and V. Katsouros, "Krikri: Advancing open large language models for greek," 2025.

[109] H. Zhu, G. Cheng, J. Wang, W. Hou, P. Zhang, and Y. Yan, "Boosting cross-domain speech recognition with self-supervision," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 32, 2024, pp. 471–485.

[110] A. A. Attia, D. Demszky, T. Ògúnrèmí, J. Liu, and C. Y. Espy-Wilson, "Cpt-boosted wav2vec2.0: Towards noise robust speech recognition for classroom environments," *CoRR*, vol. abs/2409.14494, 2024.

[111] J. Choi, D. Kim, C. Lee, and S. Hong, "Revisiting weight averaging for model merging," *CoRR*, vol. abs/2412.12153, 2024.

[112] G. Ilharco, M. Wortsman, S. Y. Gadre, S. Song, H. Hajishirzi, S. Kornblith, A. Farhadi, and L. 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*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022.

[113] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," *CoRR*, vol. abs/2010.03978, 2020.

[114] Y. Bengio, A. C. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, 2013, pp. 1798–1828.

[115] J. Pons and X. Serra, "musicnn: Pre-trained convolutional neural networks for music audio tagging," *CoRR*, vol. abs/1909.06654, 2019.

[116] K. Choi, G. Fazekas, M. B. Sandler, and K. Cho, "Transfer learning for music classification and regression tasks," in *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, S. J. Cunningham, Z. Duan, X. Hu, and D. Turnbull, Eds., 2017, pp. 141–149.

[117] A. van den Oord, S. Dieleman, and B. Schrauwen, "Transfer learning by supervised pre-training for audio-based music classification," in *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, H. Wang, Y. Yang, and J. H. Lee, Eds., 2014, pp. 29–34.

[118] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," *CoRR*, vol. abs/2303.03926, 2023.

[119] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *CoRR*, vol. abs/2301.02111, 2023.

[120] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.

[121] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 30, 2022, pp. 495–507.

[122] Y. Chung, Y. Zhang, W. Han, C. Chiu, J. Qin, R. Pang, and Y. Wu, "w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021*. IEEE, 2021, pp. 244–250.

[123] Y. Li, R. Yuan, G. Zhang, Y. Ma, C. Lin, X. Chen, A. Ragni, H. Yin, Z. Hu, H. He *et al.*, "Map-music2vec: A simple and effective baseline for self-supervised music audio representation learning," *CoRR*, vol. abs/2212.02508, 2022.

[124] H. Liu, X. Xu, Y. Yuan, M. Wu, W. Wang, and M. D. Plumbley, "Semanticodec: An ultra low bitrate semantic audio codec for general sound," *IEEE J. Sel. Top. Signal Process.*, vol. 18, no. 8, 2024, pp. 1448–1461.

[125] P. Mousavi, G. Maimon, A. Moumen, D. Petermann, J. Shi, H. Wu, H. Yang, A. Kuznetsova, A. Ploujnikov, R. Marxer *et al.*, "Discrete audio tokens: More than a survey!" 2025.

[126] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "Speechtokenizer: Unified speech tokenizer for speech language models," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[127] J. Chen, Z. Dai, Z. Ye, X. Tan, Q. Liu, Y. Guo, and W. Xue, "Pyramidcodec: Hierarchical codec for long-form music generation in audio domain," in *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Y. Al-Onaizan, M. Bansal, and Y. Chen, Eds. Association for Computational Linguistics, 2024, pp. 4253–4263.

[128] K. Qiu, X. Li, H. Chen, J. Sun, J. Wang, Z. Lin, M. Savvides, and B. Raj, "Efficient autoregressive audio modeling via next-scale prediction," *CoRR*, vol. abs/2408.09027, 2024.

[129] A. Baevski, W. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 1298–1312.

[130] M. C. McCallum, F. Korzeniowski, S. Oramas, F. Gouyon, and A. F. Ehmann, "Supervised and unsupervised learning of audio representations for music understanding," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*, P. Rao,

H. A. Murthy, A. Srinivasamurthy, R. M. Bittner, R. C. Repetto, M. Goto, X. Serra, and M. Miron, Eds., 2022, pp. 256–263.

[131] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi *et al.*, "Audiolm: A language modeling approach to audio generation," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 31, 2023, pp. 2523–2533.

[132] A. Agostinelli, T. I. Denk, Z. Borsos, J. H. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, "Musiclm: Generating music from text," *CoRR*, vol. abs/2301.11325, 2023.

[133] R. Yuan, H. Lin, S. Guo, G. Zhang, J. Pan, Y. Zang, H. Liu, Y. Liang, W. Ma, X. Du *et al.*, "Yue: Scaling open foundation models for long-form music generation," *CoRR*, vol. abs/2503.08638, 2025.

[134] J. P. Gardner, S. Durand, D. Stoller, and R. M. Bittner, "Llark: A multimodal instruction-following language model for music," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

[135] X. Du, Z. Yu, J. Lin, B. Zhu, and Q. Kong, "Joint music and language attention models for zero-shot music tagging," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*. IEEE, 2024, pp. 1126–1130.

[136] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. R. Glass, "Listen, think, and understand," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[137] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *CoRR*, vol. abs/2311.07919, 2023.

[138] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin *et al.*, "Qwen2-audio technical report," *CoRR*, vol. abs/2407.10759, 2024.

[139] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Contrastive audio-language learning for music," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*, P. Rao, H. A. Murthy, A. Srinivasamurthy, R. M. Bittner, R. C. Repetto, M. Goto, X. Serra, and M. Miron, Eds., 2022, pp. 640–649.

[140] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP learning audio concepts from natural language supervision," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*. IEEE, 2023, pp. 1–5.

[141] A. S. Hussain, S. Liu, C. Sun, and Y. Shan, "M^2ugen: Multi-modal music understanding and generation with the power of large language models," *CoRR*, vol. abs/2311.11255, 2023.

[142] K. Su, J. Y. Li, Q. Huang, D. Kuzmin, J. Lee, C. Donahue, F. Sha, A. Jansen, Y. Wang, M. Verzetti *et al.*, "V2meow: Meowing to the visual beat via video-to-music generation," in *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, M. J. Wooldridge, J. G. Dy, and S. Natarajan, Eds. AAAI Press, 2024, pp. 4952–4960.

[143] Z. Tian, Z. Liu, R. Yuan, J. Pan, X. Huang, Q. Liu, X. Tan, Q. Chen, W. Xue, and Y. Guo, "Vidmuse: A simple video-to-music generation framework with long-short-term modeling," *CoRR*, vol. abs/2406.04321, 2024.

[144] J. Zhan, J. Dai, J. Ye, Y. Zhou, D. Zhang, Z. Liu, X. Zhang, R. Yuan, G. Zhang, L. Li *et al.*, "Anygpt: Unified multimodal LLM with discrete sequence modeling," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 9637–9662.

[145] R. Yuan, H. Lin, Y. Wang, Z. Tian, S. Wu, T. Shen, G. Zhang, Y. Wu, C. Liu, Z. Zhou *et al.*, "Chatmusician: Understanding and generating music intrinsically with LLM," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 6252–6271.

[146] S. Li, S. Ji, Z. Wang, S. Wu, J. Yu, and K. Zhang, "A survey on music generation from single-modal, cross-modal, and multi-modal perspectives," 2025.

[147] T. Tu, X. Liu, Y. Ma, J. Qi, and T.-S. Chua, "Unified music-language model for symbolic and waveform integration," 2025.

[148] Y. Bai, H. Chen, J. Chen, Z. Chen, Y. Deng, X. Dong, L. Hantrakul, W. Hao, Q. Huang, Z. Huang *et al.*, "Seed-music: A unified framework for high quality and controlled music generation," *CoRR*, vol. abs/2409.09214, 2024.

[149] R. Yuan, Y. Ma, Y. Li, G. Zhang, X. Chen, H. Yin, L. Zhuo, Y. Liu, J. Huang, Z. Tian *et al.*, "MARBLE: music audio representation benchmark for universal evaluation," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.

[150] S. Doh, K. Choi, J. Lee, and J. Nam, "Lp-musiccaps: Llm-based pseudo music captioning," in *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023, Milan, Italy, November 5-9, 2023*, A. Sarti, F. Antonacci, M. Sandler, P. Bestagini, S. Dixon, B. Liang, G. Richard, and J. Pauwels, Eds., 2023, pp. 409–416.

[151] B. L. T. Sturm and A. Flexer, "A review of validity and its relationship to music information research," in *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023, Milan, Italy, November 5-9, 2023*, A. Sarti, F. Antonacci, M. Sandler, P. Bestagini, S. Dixon, B. Liang, G. Richard, and J. Pauwels, Eds., 2023, pp. 47–55.

[152] A. Apicella, F. Isgrò, and R. Prevete, "Don't push the button! exploring data leakage risks in machine learning and transfer learning," *CoRR*, vol. abs/2401.13796, 2024.

[153] C. Plachouras, J. Guinot, G. Fazekas, E. Quinton, E. Benetos, and J. Pauwels, "Towards a Unified Representation Evaluation Framework Beyond Downstream Tasks," May 2025, arXiv:2505.06224 [cs].

[154] J. H. Lee, K. Choi, X. Hu, and J. S. Downie, "K-pop genres: A cross-cultural exploration," in *Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR 2013, Curitiba, Brazil, November 4-8, 2013*, A. de Souza Britto Jr., F. Gouyon, and S. Dixon, Eds., 2013, pp. 529–534.

[155] A. Srinivasamurthy, S. Gulati, R. C. Repetto, and X. Serra, "Saraga: Open Datasets for Research on Indian Art Music," *Empirical Musicology Review*, vol. 16, no. 1, Dec. 2021, pp. 85–98, number: 1.

[156] R. C. Repetto and X. Serra, "Creating a corpus of jingju (beijing opera) music and possibilities for melodic analysis," in *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, H. Wang, Y. Yang, and J. H. Lee, Eds., 2014, pp. 313–318.

[157] M. Sordo, A. Chaachoo, and X. Serra, "Creating corpora for computational research in arab-andalusian music," in *Proceedings of the 1st International Workshop on Digital Libraries for Musicology, DLfM@JCDL 2014, London, United Kingdom, September 12, 2014*, B. Fields and K. R. Page, Eds. ACM, 2014, pp. 1–3.

[158] P. Singh and V. Arora, "Explainable deep learning analysis for raga identification in indian art music," *CoRR*, vol. abs/2406.02443, 2024.

[159] L. O. Nunes, M. Rocamora, L. Jure, and L. W. P. Biscainho, "Beat and downbeat tracking based on rhythmic patterns applied to the uruguayan candombe drumming," in *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26-30, 2015*, M. Müller and F. Wiering, Eds., 2015, pp. 264–270.

[160] C. N. S. Jr., A. L. Koerich, and C. A. A. Kaestner, "The latin music database," in *ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, September 14-18, 2008*, J. P. Bello, E. Chew, and D. Turnbull, Eds., 2008, pp. 451–456.

[161] J. M. de Sousa, E. T. Pereira, and L. R. Veloso, "A robust music genre classification approach for global and regional music datasets evaluation," in *2016 IEEE International Conference on Digital Signal Processing, DSP 2016, Beijing, China, October 16-18, 2016*. IEEE, 2016, pp. 109–113.

[162] L. S. Maia, P. Tomaz Jr, M. Fuentes, M. Rocamora, L. Biscainho, M. Costa, and S. Cohen, "A novel dataset of brazilian rhythmic instruments and some experiments in computational rhythm analysis," in *2018 AES Latin American Congress of Audio Engineering*, 2018, pp. 53–60.

[163] B. Nikzat and R. C. Repetto, "KDC: an open corpus for computational research of dastg?hi music," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*, P. Rao, H. A. Murthy, A. Srinivasamurthy, R. M. Bittner, R. C. Repetto, M. Goto, X. Serra, and M. Miron, Eds., 2022, pp. 321–328.

[164] O. Lartillot, M. Johansson, A. Elowsson, L. Monstad, and M. Cyvin, "A dataset of norwegian hardanger fiddle recordings with precise annotation of note and beat onsets," *Trans. Int. Soc. Music. Inf. Retr.*, vol. 6, no. 1, 2023, pp. 186–202.

[165] K. K. Ganguli, S. Sentürk, and C. Guedes, "Critiquing task- versus goal-oriented approaches: A case for makam recognition," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*, P. Rao, H. A. Murthy, A. Srinivasamurthy, R. M. Bittner, R. C. Repetto, M. Goto, X. Serra, and M. Miron, Eds., 2022, pp. 369–376.

[166] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, ser. JMLR Proceedings, Y. W. Teh and D. M. Titterington, Eds., vol. 9. JMLR.org, 2010, pp. 297–304.

[167] M. D. Lange, G. M. van de Ven, and T. 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.

[168] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. 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*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 10 524–10 533.

[169] I. Loshchilov and F. 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.

[170] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 8440–8451.

[171] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "XLS-R: self-supervised cross-lingual speech representation learning at scale," in *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 2278–2282.

[172] Y. Fujinuma, J. L. Boyd-Graber, and K. Kann, "Match the script, adapt if multilingual: Analyzing the effect of multilingual pretraining on cross-lingual transferability," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022, pp. 1500–1512.

[173] D. Karomat, "12 maqam system and its similarity with indian raga's (according to the indian manuscripts)," *Indian Musicological Society. Journal of the Indian Musicological Society*, vol. 36, 2006, p. 62.

[174] J. Eronen, M. Ptaszynski, and F. Masui, "Zero-shot cross-lingual transfer language selection using linguistic similarity," *Inf. Process. Manag.*, vol. 60, no. 3, 2023, p. 103250.

[175] C. Macaire, D. Schwab, B. Lecouteux, and E. Schang, "Automatic speech recognition and query by example for creole languages documentation," in *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022, pp. 2512–2520.

[176] F. Xiong, R. Cheng, W. Chen, Z. Zhang, Y. Guo, C. Yuan, and R. Xu, "Multi-task model merging via adaptive weight disentanglement," *CoRR*, vol. abs/2411.18729, 2024.

[177] P. Yadav, D. Tam, L. Choshen, C. A. Raffel, and M. Bansal, "Ties-merging: Resolving interference when merging models," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.

[178] A. Pasad, J. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021*. IEEE, 2021, pp. 914–921.

[179] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 1691–1703.

[180] M. Tanti, L. van der Plas, C. Borg, and A. Gatt, "On the language-specificity of multilingual BERT and the impact of fine-tuning," in *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2021, Punta Cana, Dominican Republic, November 11, 2021*, J. Bastings, Y. Belinkov, E. Dupoux, M. Giulianelli, D. Hupkes, Y. Pinter, and H. Sajjad, Eds. Association for Computational Linguistics, 2021, pp. 214–227.

[181] N. F. Liu, M. Gardner, Y. Belinkov, M. E. Peters, and N. A. Smith, "Linguistic knowledge and transferability of contextual representations," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 1073–1094.

[182] R. Achtibat, S. M. V. Hatefi, M. Dreyer, A. Jain, T. Wiegand, S. Lapuschkin, and W. Samek, "Attnlrp: Attention-aware layer-wise relevance propagation for transformers," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

[183] A. Ziv, I. Gat, G. L. Lan, T. Remez, F. Kreuk, J. Copet, A. Défossez, G. Synnaeve, and Y. Adi, "Masked audio generation using a single non-autoregressive transformer," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[184] A. Porter, M. Sordo, and X. Serra, "Dunya: A system to browse audio music collections exploiting cultural context," in *Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR 2013, Curitiba, Brazil, November 4-8, 2013*, A. de Souza Britto Jr., F. Gouyon, and S. Dixon, Eds., 2013, pp. 101–106.

[185] R. Geirhos, J. Jacobsen, C. Michaelis, R. S. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nat. Mach. Intell.*, vol. 2, no. 11, 2020, pp. 665–673.

[186] L. Pearson, "Cultural Specificities in Carnatic and Hindustani Music: Commentary on the Saraga Open Dataset," *Empirical Musicology Review*, vol. 16, no. 1, Dec. 2021, pp. 166–171, number: 1.

[187] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[188] Y. Ding and A. Lerch, "Parameter-efficient transfer learning for music foundation models," *CoRR*, vol. abs/2411.19371, 2024.

[189] G. Paraskevopoulos, T. Kouzelis, G. Rouvalis, A. Katsamanis, V. Katsouros, and A. Potamianos, "Sample-efficient unsupervised domain adaptation of speech recognition systems: A case study for modern greek," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 32, 2024, pp. 286–299.

[190] C. Karouzos, G. Paraskevopoulos, and A. Potamianos, "UDALM: unsupervised domain adaptation through language modeling," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Association for Computational Linguistics, 2021, pp. 2579–2590.

[191] H. F. García, P. Seetharaman, R. Kumar, and B. Pardo, "Vampnet: Music generation via masked acoustic token modeling," in *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023, Milan, Italy, November 5-9, 2023*, A. Sarti, F. Antonacci, M. Sandler, P. Bestagini, S. Dixon, B. Liang, G. Richard, and J. Pauwels, Eds., 2023, pp. 359–366.

[192] L. Pepino, P. Riera, and L. Ferrer, "Encodecmae: Leveraging neural codecs for universal audio representation learning," *CoRR*, vol. abs/2309.07391, 2023.

[193] Y. Jiang, Q. Chen, S. Ji, Y. Xi, W. Wang, C. Zhang, X. Yue, S. Zhang, and H. Li, "Unicodec: Unified audio codec with single domain-adaptive codebook," *CoRR*, vol. abs/2502.20067, 2025.

[194] N. Zeghidour, O. Teboul, F. de Chaumont Quitry, and M. Tagliasacchi, "LEAF: A learnable frontend for audio classification," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[195] A. Holzapfel, G. A. Velasco, N. Holighaus, M. Dörfler, and A. Flexer, "Advantages of nonstationary gabor transforms in beat tacking," in *Proceedings of the 1st International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies*, ser. MIRUM '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 45–50.

[196] F. Yesiler, B. Bozkurt, and X. Serra, "Makam Recognition Using Extended Pitch Distribution Features And Multi-Layer Perceptrons," Jul. 2018, publisher: Zenodo.

[197] A. Pagnoni, R. Pasunuru, P. Rodríguez, J. Nguyen, B. Muller, M. Li, C. Zhou, L. Yu, J. Weston, L. Zettlemoyer *et al.*, "Byte latent transformer: Patches scale better than tokens," *CoRR*, vol. abs/2412.09871, 2024.

[198] P. T. Deep, R. Bhardwaj, and S. Poria, "Della-merging: Reducing interference in model merging through magnitude-based sampling," *CoRR*, vol. abs/2406.11617, 2024.

[199] Y. Wei, A. Tang, L. Shen, C. Yuan, and X. Cao, "Modeling multi-task model merging as adaptive projective gradient descent," *CoRR*, vol. abs/2501.01230, 2025.

[200] A. Tang, E. Yang, L. Shen, Y. Luo, H. Hu, B. Du, and D. Tao, "Merging models on the fly without retraining: A sequential approach to scalable continual model merging," *CoRR*, vol. abs/2501.09522, 2025.

[201] A. A. Gargiulo, D. Crisostomi, M. S. Bucarelli, S. Scardapane, F. Silvestri, and E. Rodolà, "Task singular vectors: Reducing task interference in model merging," *CoRR*, vol. abs/2412.00081, 2024.

[202] C. Lee, J. Choi, C. Lee, D. Kim, and S. Hong, "Adarank: Adaptive rank pruning for enhanced model merging," *CoRR*, vol. abs/2503.22178, 2025.

[203] S. Fan, M. Pagliardini, and M. Jaggi, "DOGE: domain reweighting with generalization estimation," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

[204] S. M. Xie, H. Pham, X. Dong, N. Du, H. Liu, Y. Lu, P. Liang, Q. V. Le, T. Ma, and A. W. Yu, "Doremi: Optimizing data mixtures speeds up language model pretraining," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.

[205] Q. Liu, X. Zheng, N. Muennighoff, G. Zeng, L. Dou, T. Pang, J. Jiang, and M. Lin, "Regmix: Data mixture as regression for language model pre-training," *CoRR*, vol. abs/2407.01492, 2024.

[206] H. Que, J. Liu, G. Zhang, C. Zhang, X. Qu, Y. Ma, F. Duan, Z. Bai, J. Wang, Y. Zhang *et al.*, "D-CPT law: Domain-specific continual pre-training scaling law for large language models," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, Eds., 2024.

[207] R. Chitale, A. Vaidya, A. Kane, and A. Ghotkar, "Task arithmetic with lora for continual learning," *CoRR*, vol. abs/2311.02428, 2023.

[208] D. Marczak, B. Twardowski, T. Trzcinski, and S. Cygert, "MAGMAX: leveraging model merging for seamless continual learning," in *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXV*, ser. Lecture Notes in Computer Science, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., vol. 15143. Springer, 2024, pp. 379–395.

[209] M. Li, Y. Lu, Q. Dai, S. Huang, Y. Ding, and H. Lu, "BECAME: BayEsian Continual Learning with Adaptive Model MErging," Apr. 2025, arXiv:2504.02666 [cs].

[210] A. Alexandrov, V. Raychev, M. Mueller, C. Zhang, M. T. Vechev, and K. Toutanova, "Mitigating catastrophic forgetting in language transfer via model merging," in *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Y. Al-Onaizan, M. Bansal, and Y. Chen, Eds. Association for Computational Linguistics, 2024, pp. 17 167–17 186.

[211] W. Ma, X. Li, and G. Xia, "Do music LLMs learn symbolic concepts? a pilot study using probing and intervention," in *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.

[212] A. Zhang, E. Thomaz, and L. Lu, "Transformation of audio embeddings into interpretable, concept-based representations," 2025.

[213] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

List of Abbreviations

AI	Artificial Intelligence
AP	Average Precision
APC	Auto-regressive Predictive Coding
CL	Contrastive Learning
CNN	Convolutional Neural Network
CPT	Continual Pre-Training
CQT	Constant-Q Transform
DA	Domain Adaptation
DL	Deep Learning
FM	Foundation Model
HQ-VAE	Hierarchically Quantized-Variational AutoEncoder
JSD	Jensen-Shannon Divergence
LLM	Large Language Model
MIR	Music Information Retrieval
ML	Machine Learning
MLM	Masked Language Modeling
MLP	Multilayer Perceptron
MM	Masked Modeling
MTL	Multi-Task Learning
MSE	Mean Squared Error
NAC	Neural Audio Codec
NCE	Noise-Contrastive Estimation
NLP	Natural Language Processing
PEFT	Parameter-Efficient Fine-Tuning
ROC-AUC	Area Under the Receiver-Operating Characteristic Curve
RVQ	Residual Vector Quantization
SOTA	State-of-the-Art
SSL	Self-supervised Learning
TA	Task Arithmetic
VQ-VAE	Vector Quantized-Variational AutoEncoder
XAI	Explainable Artificial Intelligence