Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

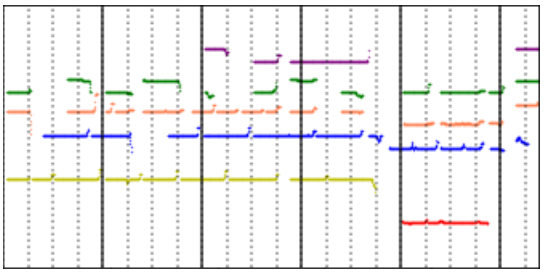
Μεταγραφή Μουσικής με χρήση Μεθόδων Βαθιάς Μάθησης

Μελέτη και υλοποίηση

Διπλωματικη Εργασια

του

ΒΑΣΙΛΕΙΟΥ Γ. ΛΕΥΚΟΒΙΤΣ



Επιβλέπων: Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.
Συνεπιβλέπων: Γεώργιος Αλεξανδρίδης
Εργαστηριακό Διδακτικό Προσωπικό Ε.Μ.Π.

Αθήνα, Μάρτιος 2023

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Μεταγραφή Μουσικής με χρήση Μεθόδων Βαθιάς Μάθησης

Μελέτη και υλοποίηση

Διπλωματικη Εργασια

του

ΒΑΣΙΛΕΙΟΥ Γ. ΛΕΥΚΟΒΙΤΣ

Επιβλέπων: Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.
Συνεπιβλέπων: Γεώργιος Αλεξανδρίδης
Εργαστηριακό Διδακτικό Προσωπικό Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 20η Μαρτίου 2023.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Γεώργιος Στάμου	Στέφανος Κόλλιας	Αθανάσιος Βουλόδημος
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2023

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

.....
Βασίλειος Γ. Λεύκοβιτς

20 Μαρτίου 2023

Περίληψη

Η παρούσα εργασία πραγματεύεται την εξερεύνηση μεθόδων μεταγραφής μουσικής, χρησιμοποιώντας συνελικτικά νευρωνικά δίκτυα. Στο πρώτο, θεωρητικό, μέρος, περιγράφονται κάποιες εισαγωγικές έννοιες τόσο σχετικά με τα ίδια τα νευρωνικά δίκτυα, όσο και με έννοιες εξειδικευμένες στον χώρο της επεξεργασίας μουσικής και ήχου. Στη συνέχεια, περιγράφεται η δομή και το περιεχόμενο του συνόλου δεδομένων Guitarset, καθώς και οι μετρικές οι οποίες χρησιμοποιούνται στα επόμενα κεφάλαια. Στο τέταρτο κεφάλαιο, περιγράφονται τα 4 διαφορετικά μοντέλα τα οποία εξετάστηκαν, ακολουθούμενα από τη σύγκριση των αποτελεσμάτων του κάθε μοντέλου. Τέλος, στον επίλογο γίνονται κάποιες προτάσεις για μελλοντικές επεκτάσεις των εξεταζόμενων μοντέλων και αλγορίθμων που χρησιμοποιήθηκαν.

Λέξεις Κλειδιά

Συνελικτικά Νευρωνικά Δίκτυα, Ήχος, Μουσική, Μεταγραφή Μουσικής, Επαναλαμβανόμενα Μοντέλα, Βαθιά Μάθηση

Abstract

This dissertation thesis deals with the exploration of music transcription methods, using convolutional neural networks. In the first, theoretical, part, some introductory concepts regarding neural networks themselves are described, as well as concepts specialized in the field of music and audio processing. Next, the structure and contents of Guitarset dataset are analysed, as well as the metrics that will be used in the following chapters. In the fourth chapter, four different models are being described, followed by the comparison of the results between them. Finally, in the conclusions, some suggestions are made for future extensions of the considered models and algorithms used, are formulated.

Keywords

CNN, Sound, Music, Music Transcription, Reccurent Models, Deep Learning

στην οικογένειά μου

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον κ. Γεώργιο Στάμου, Καθηγητή Ε.Μ.Π., για την ευκαιρία που μου έδωσε να ασχοληθώ με το αντικείμενο που πραγματεύεται η παρούσα διπλωματική εργασία. Επίσης θέλω να ευχαριστήσω τους κ. κ. Στέφανο Κόλλια, Καθηγητή Ε.Μ.Π. και Αθανάσιο Βουλόδημο, Επίκουρο Καθηγητή Ε.Μ.Π., που μου έκαναν την τιμή να συμμετέχουν στην τριμελή επιτοπή εξέτασης, καθώς και τους κ. κ. Γεώργιο Αλεξανδρίδη, Εργαστηριακό και Διδακτικό Προσωπικό Ε.Μ.Π., και Έντμοντ-Γρηγόρη Ντερβάκο, Υποψήφιο Διδάκτορα Ε.Μ.Π., για τη συνολική συνεισφορά και αρωγή τους. Επιπλέον, θα ήθελα να ευχαριστήσω το Εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης (AILS) της Σχολής Ηλεκτρολόγων Μηχανικών & Μηχανικών Ηλεκτρονικών Υπολογιστών του Ε.Μ.Π., για την πρόσβαση στην απαραίτητη υπολογιστική υποδομή για την εκτέλεση του πειραματικού μέρους της εργασίας, όσο και για τη διδασκαλία πλήθους μαθημάτων του Μεταπτυχιακού "Επιστήμη Δεδομένων και Μηχανική Μάθηση". Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου, η οποία υπήρξε στήριγμα, μεταξύ άλλων, και στην παρούσα διπλωματική εργασία.

Αθήνα, Μάρτιος 2023

Βασίλειος Γ. Λεύκοβιτς

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Η μουσική αποτελεί, σε μικρότερο ή μεγαλύτερο βαθμό, μέρος της καθημερινότητας πλήθους ανθρώπων. Ένα από τα βασικά εργαλεία των μουσικών οι οποίοι συνθέτουν και αποδίδουν μουσικά έργα, είναι η παρτιτούρα, στην κλασσική της αλλά και σε πιο ευανάγνωστες για αρχάριους μουσικούς μορφές (ταμπλατούρα). Μια παρτιτούρα μπορεί να παρέχει έναν αριθμό πληροφορίας σε έναν μουσικό που προσπαθεί να αποδώσει το μουσικό έργο που αυτή καταγράφει, όπως είναι, οι νότες, η διάρκεια τις κάθε νότας, οι παύσεις που ενδέχεται να υπάρχουν ανάμεσα σε αυτές, οι συγχορδίες που συνοδεύουν τη μελωδία και άλλα χρήσιμα στοιχεία τα οποία ο συνθέτης του μουσικού έργου έχει προβλέψει.

Κατά αυτόν τον τρόπο, η παρτιτούρα, αποτελεί έναν τρόπο επικοινωνίας και μετάδοσης της σύνθεσης μεταξύ των μουσικών, με έναν συνεπή και συνεκτικό τρόπο. Παρόλα αυτά, ανάλογα το είδος και την εποχή σύνθεσης της μουσικής, η εύρεση παρτιτούρας για κάποιο συγκεκριμένο έργο, μπορεί να είναι πολύ δύσκολη υπόθεση. Αυτό μπορεί να συμβαίνει είτε επειδή ο ίδιο ο συνθέτης δεν δημιούργησε την παρτιτούρα του έργου του, είτε επειδή ακόμη και να υπήρχε, έχει χαθεί στο χρόνο. Σε τέτοια περίπτωση, η εύρεση ενός τρόπου παρασκευής της παρτιτούρας αυτής από την ίδιο την ηχογράφηση της σύνθεσης, θα ήταν ένα πολύτιμο εργαλείο για οποιονδήποτε επιθυμεί να αποδώσει τη σύνθεση μουσικά.

Παραδοσιακά, ο τρόπος παρασκευής της παρτιτούρας είναι η ακουστική μεταφορά του έργου σε πεντάγραμμο, για το οποίο απαιτείται αρκετός χρόνος, καθώς και επαρκής εκπαίδευση από το μέρος του μουσικού που θα την πραγματοποιήσει. Το γεγονός αυτό αποτελεί και το κύριο κίνητρο της παρούσας διπλωματικής, αφού η επιδίωξη κατασκευής ενός μοντέλου, το οποίο θα μπορεί να υπερβεί τα προαναφερθέντα εμπόδια, θα ήταν ένα βήμα προς την διευκόλυνση της εκμάθησης μουσικής, καθώς και της ταχύτερης μεταφοράς ενός έργου σε παρτιτούρα.

Προσπερνώντας τη χρηστικότητα της ύπαρξης μιας τέτοιας αρχιτεκτονικής, η αυτόματη μεταγραφή μουσικής μέσω βαθιάς μάθησης, αποτελεί και ένα δύσκολο πρόβλημα προς επίλυση. Σύμφωνα με την τρέχουσα στάθμη της τεχνικής, την περίοδο συγγραφής της διπλωματικής, καμία από τις μετρικές που θα εξεταστούν δεν έχει επιτύχει απόδοση μεγαλύτερη από 80%. Αυτό σημαίνει, ότι υπάρχει ακόμα χώρος για εξέλιξη, καθώς και πειραματισμό, ώστε να παρουσιαστούν κάποιες εναλλακτικές μέθοδοι.

1.1 Αντικείμενο της διπλωματικής

Η παρούσα διπλωματική, έχει ως σκοπό να πραγματευτεί το απαιτητικό πρόβλημα της αυτόματης μεταγραφής μουσικής μέσω βαθιάς μάθησης. Συγκεκριμένα, θα χρησιμοποιηθεί η συλλογή δεδομένων Guitarset [1], η οποία περιέχει ηχογραφήσεις κιθάρας, κάθε μια από τις οποίες περιέχει τη μελωδία, καθώς και τις συγχορδίες που την συνοδεύουν. Σε αυτή τη συλλογή δεδομένων, θα εφαρμοστούν 3 διαφορετικές προσεγγίσεις, οι 2 εκ των οποίων αποτελούν μεθόδους που έχουν περιγραφεί στη βιβλιογραφία. Συνοπτικά :

1. Η πρώτη μέθοδος [2], αποτελεί μια από τις πρώτες απόπειρες για μεταγραφή μουσικής, και χρησιμοποιεί ένα βασικό συνελικτικό δίκτυο. Αυτή η προσέγγιση, παρότι απλοϊκή, εισήγαγε τον αρμονικό μετασχηματισμό σταθερού-Q, ο οποίος μετασχηματίζει το εισαγόμενο ηχητικό σήμα, σε απεικόνιση συχνότητας-χρόνου, και θα περιγραφεί αναλυτικότερα σε επόμενο κεφάλαιο. Το μεγαλύτερο μειονέκτημα της μεθόδου, έγκειται στο γεγονός ότι αντιμετωπίζει το πρόβλημα ως μια ενιαία εργασία, ενώ οι μεταγενέστερες μέθοδοι χωρίζουν το πρόβλημα σε επιμέρους εργασίες.

2. Η δεύτερη προσέγγιση [3], επιχειρεί τον χωρισμό του προβλήματος σε επιμέρους εργασίες, και χρησιμοποιεί επίσης ένα σύστημα από συνελικτικά δίκτυα. Εμφανίζει, όπως θα φανεί σε επόμενο κεφάλαιο, αρκετά βελτιωμένη απόδοση, σε όλες τις μετρικές, σε σχέση με την προηγούμενη μέθοδο, και θέτει τη στάθμη της τεχνικής για το συγκεκριμένο πρόβλημα.

3. Η τρίτη μέθοδος που εφαρμόστηκε, χρησιμοποιεί ένα τρέχον μοντέλο αιχμής [3], προσθέτοντας και έναν αναδρομικό χαρακτήρα. Στην ουσία, χρησιμοποιείται το μοντέλο της δεύτερης μεθόδου, όπου τα συνελικτικά δίκτυα δύο διαστάσεων, αντικαθίστανται με αναδρομικά δίκτυα τα οποία έχουν και χαρακτηριστικά συνέλιξης [4]. Κίνητρο για την συγκεκριμένη προσέγγιση, αποτελεί το γεγονός πως οι περισσότερες συνθέσεις ακολουθούν κάποια μουσική κλίμακα. Αυτό σημαίνει, ότι υπάρχει μια σαφής χρονική εξάρτηση ενός τμήματος της σύνθεσης από το προηγούμενό του, αφού θα ακολουθούν τους ίδιους "κανόνες".

4. Τέλος, υλοποιήθηκε και μια τροποποιημένη αρχιτεκτονική του τρέχοντος μοντέλου αιχμής, η οποία συμπεριλαμβάνει και τα offsets. Η αρχιτεκτονική αυτή, επιδιώκει να προσφέρει περισσότερη πληροφορία στο δίκτυο, η οποία μαζί με τα onsets, τα οποία ήδη παρέχονται, προσφέρει τόσο τους χρόνους έναρξης όσο και λήξης της κάθε νότας.

Το βασικό αντικείμενο της διπλωματικής είναι η εφαρμογή των προαναφερθέντων μοντέλων στην ίδια συλλογή δεδομένων, ώστε να απομακρυνθούν οποιοιδήποτε αστάθμητοι παράγοντες, και να γίνει μια αντικειμενικότερη σύγκριση των μεθόδων. Για παράδειγμα, το πρώτο μοντέλο [2], έχει εκπαιδευτεί σε 3 συλλογές δεδομένων στη βιβλιογραφία (Bach10 [5], Su [6] και MedleyDB [7]), αλλά όχι στο Guitarset [1]. Αντιστοίχως, το μοντέλο της δεύτερης μεθόδου, έχει εκπαιδευτεί σε 7 συνολικά datasets (MedleyDB [7], Guitarset [1], Molina [8], iKala [9], Maestro [10], Slakh [11] και Phoenicx [12]). Στην παρούσα εργασία, ο σκοπός είναι η σύγκριση των μεθόδων αυτών, αλλά και της τρίτης, αποκλειστικά στο Guitarset.

1.2 Οργάνωση της εργασίας

Η εργασία έχει οργανωθεί σε 3 βασικά τμήματα· το θεωρητικό μέρος, το πρακτικό μέρος και τον επίλογο, συγκροτώντας συνολικά 6 διαφορετικά κεφάλαια. Το πρώτο μέρος (Κεφάλαιο 2), περιέχει μια περιγραφή των βασικών θεωρητικών δομικών στοιχείων τα οποία απαιτούνται για την κατανόηση της εργασίας, ενώ στο δεύτερο (Κεφάλαια 3, 4, 5), περιγράφονται τα μοντέλα τα οποία αναλύθηκαν, καθώς σχολιάζονται και συγκρίνονται τα αποτελέσματά τους. Τέλος, στον επίλογο (Κεφάλαιο 6), παρουσιάζονται κάποιες πιθανές μελλοντικές επεκτάσεις των μεθόδων που μελετήθηκαν.

Μέρος Ι

Θεωρητικό Μέρος

Κεφάλαιο 2

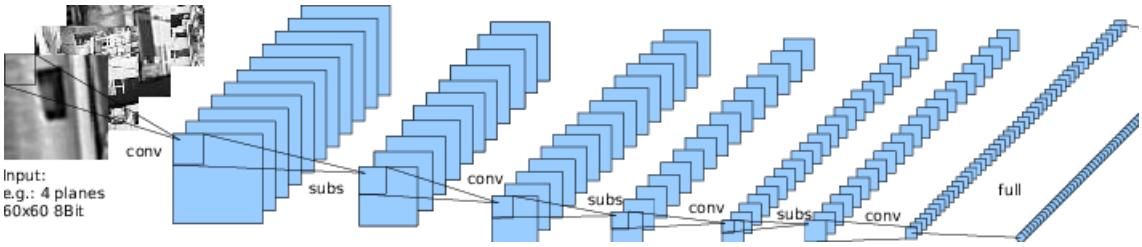
Θεωρητικό υπόβαθρο

Στο παρόν κεφάλαιο, παρουσιάζονται, αρχικά, κάποια βασικά στοιχεία για τους τύπους δικτύων που χρησιμοποιήθηκαν. Στη συνέχεια, αναφέρονται συνοπτικά κάποια τεχνικά και θεωρητικά στοιχεία που είναι χρήσιμα για την κατανόηση των επόμενων κεφαλαίων.

2.1 Συνελικτικά Νευρωνικά Δίκτυα

2.1.1 Συνελικτικά Επίπεδα

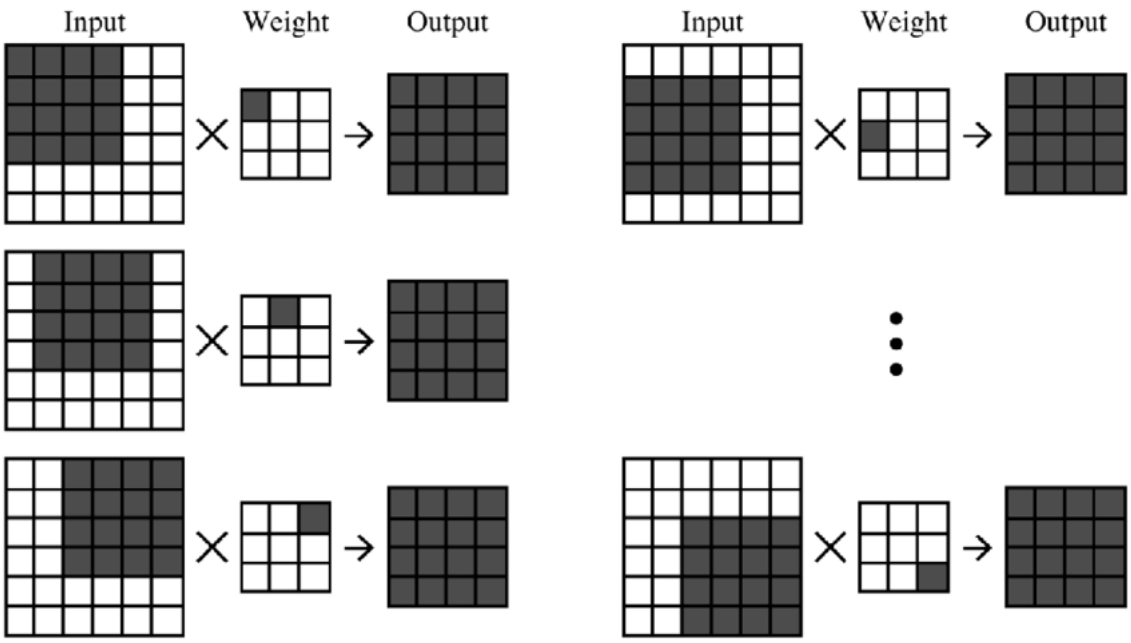
Τα συνελικτικά νευρωνικά δίκτυα (convolutional neural networks - CNNs) είναι ένας συγκεκριμένος τύπος νευρωνικών δικτύων, τα οποία ειδικεύονται στην εκπαίδευση μοντέλων που έχουν ως είσοδο εικόνες ή διαφόρων ειδών σήματα, εν προκειμένω ηχητικά. Στην καρδιά των CNNs, βρίσκονται τα συνελικτικά επίπεδα (Εικόνα 2.1), τα οποία διακρίνονται σε δισδιάστατα και τρισδιάστατα. Για τις ανάγκες της εργασίας, χρησιμοποιήθηκαν δισδιάστατα συνελικτικά επίπεδα, αλλά η ίδια αρχή λειτουργίας γενικεύεται και στα τρισδιάστατα.



Εικόνα 2.1: *Τυπική λειτουργία ενός CNN μοντέλου*

Η λειτουργία των συνελικτικών επιπέδων, βασίζεται σε "πυρήνες" (φίλτρα), τα οποία ουσιαστικά αποτελούν έναν τρισδιάστατο πίνακα με τις δύο εκ των τριών διαστάσεων να προσδίδονται ως παράμετροι στο στρώμα από την αρχιτεκτονική, και την τρίτη να συμβολίζει το βάθος του πίνακα εισόδου στο επίπεδο. Τα φίλτρα αυτά φέρουν ως στοιχεία του πίνακά τους εκπαιδεύσιμα βάρη, οι τιμές των οποίων προσδιρίζονται μέσω της προς τα πίσω διάδοσης του σφάλματος κατά τη διάρκεια της εκπαίδευσης του δικτύου. Μετά την αρχικοποίηση των βαρών, υπολογίζεται το εσωτερικό γινόμενο του φίλτρου με ένα τμήμα του πίνακα εισόδου, ίσων διαστάσεων. Έτσι, προκύπτει ένας γραμμικός συνδυασμός των βαρών του πυρήνα. Ο ίδιος υπολογισμός επαναλαμβάνεται για ολόκληρο τον πίνακα εισόδου, καθώς ο πυρήνας μετατοπίζεται πάνω στην είσοδο, κινούμενος βάσει ενός προκαθορισμένου βηματισμού, ο

οποίος έχει καθοριστεί με την αρχικοποίηση της αρχιτεκτονικής (Εικόνα 2.2).



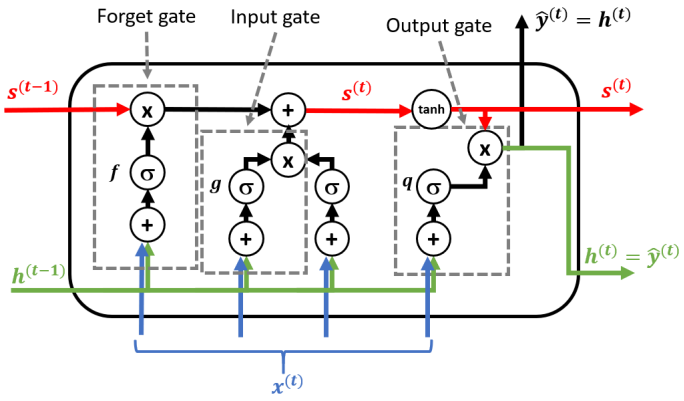
Εικόνα 2.2: *Λειτουργία των φίλτρων σε συνελικτικά επίπεδα*

Πραγματοποιώντας την παραπάνω διαδικασία με ένα συγκεκριμένο πυρήνα, θα παραχθεί μια δισδιάστατη έξοδος, αφού το εσωτερικό γινόμενο του φίλτρου με το αντίστοιχο τμήμα της εισόδου, θα παράγει κάθε φορά μονοδιάστατο αποτέλεσμα. Τα αποτελέσματα όλων των εσωτερικών γινομένων, τοποθετούνται στον χώρο ώστε να αντανακλούν τη θέση του τμήματος της εισόδου με το οποίο πολλαπλασιάζεται το φίλτρο, καθώς και τον βηματισμό του ίδιου του φίλτρου. Τα παραπάνω βήματα, μπορούν να πραγματοποιηθούν για έναν προκαθορισμένο αριθμό φίλτρων, κάθε ένα από τα οποία θα παράγει μια δισδιάστατη έξοδο, οι οποίες τελικά θα στοιβαχθούν ώστε να κατασκευαστεί το τελικό τρισδιάστατο αποτέλεσμα. Τέλος, συνηθίζεται να ακολουθείται το συνελικτικό στρώμα από κάποια συνάρτηση ενεργοποίησης, συνήθως σιγμοειδή ή ημι-γραμμική, ανάλογα με τις ανάγκες του μοντέλου.

2.1.2 Δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης

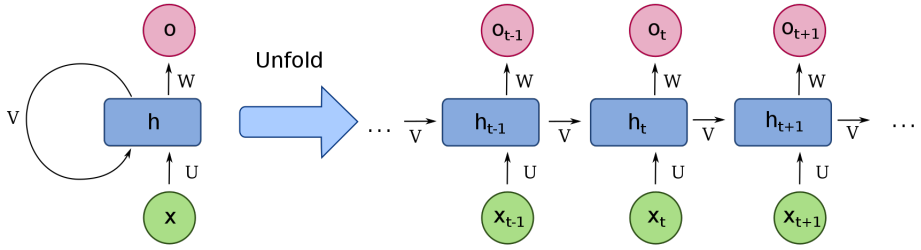
Τα δίκτυα μακράς βραχυπρόθεσμης μνήμης (long short-term memory - LSTM) εντάσσονται στην γενικότερη κατηγορία των αναδρομικών νευρωνικών δικτύων (recurrent neural networks - RNNs) και είναι σχεδιασμένα να δέχονται ακολουθίες δεδομένων. Η λειτουργία τους, βασίζεται στην αρχή ότι κάθε τμήμα της ακολουθίας είναι λογικό να επηρεάζει τα υπόλοιπα, ειδικά τα αμέσως επόμενα ή προηγούμενα, σε συγκεκριμένα προβλήματα. Τέτοιες εργασίες μπορεί να είναι η αναγνώριση φωνής ή η πρόβλεψη χρονοσειρών.

Κατά τη λειτουργία των αρχικών αναδρομικών δικτύων για κάθε τμήμα της ακολουθίας παράγονταν μια έξοδος και μία κρυφή κατάσταση, η οποία μεταφερόταν στο επόμενο τμήμα της ακολουθίας, συσχετίζοντας κατ' αυτόν τον τρόπο τα μέρη που την απαρτίζουν (Εικόνα 2.4). Η ανάγκη για ανάπτυξη των LSTM δικτύων προέκυψε από το φαινόμενο "εξαφανιζόμενων" κλίσεων, το οποίο επηρεάζει, μεταξύ άλλων, και τα RNN. Το συγκεκριμένο πρόβλημα, το οποίο συνήθως εντείνεται όσο μεγαλώνει το μήκος της ακολουθίας, προκαλεί τα βάρη να



Εικόνα 2.3: *Εσωτερική δομή ενός κελιού LSTM*

ανανεώνονται με πολύ μικρό ρυθμό, αυξάνοντας ιδιαιτέρως τον αριθμό των εποχών που απαιτούνται για την εκπαίδευση.



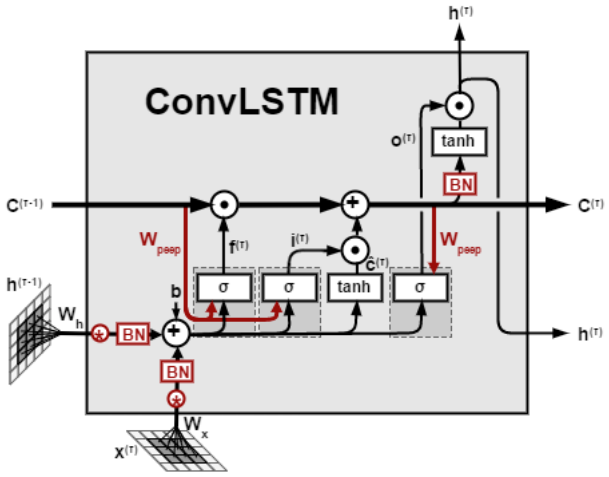
Εικόνα 2.4: *Εσωτερική δομή αναδρομικών δικτύων*

Το προαναφερθέν πρόβλημα, αποτέλεσε έναν από τους κύριους λόγους για την κατασκευή των LSTMs. Σε αυτού του είδους αναδρομικών μοντέλων, οι βασικότερες έννοιες είναι οι κρυφές καταστάσεις και οι καταστάσεις κελιού (Εικόνα 2.3). Η κρυφή κατάσταση κάθε βήματος της ακολουθίας ουσιαστικά περιέχει κωδικοποιημένη την πληροφορία του συγκεκριμένου μόνο βήματος και μεταφέρεται στο αμέσως επόμενο βήμα της ακολουθίας. Αντίθετα, η κατάσταση του κελιού περιέχει κωδικοποιημένη πληροφορία για όλα τα βήματα μέχρι και το παρόν και μεταφέρεται κάθε φορά στο επόμενο, κωδικοποιώντας, διαδοχικά, εντός της κάθε βήμα από το οποίο περνά.

Εκτός από τις καταστάσεις, υπάρχουν εντός της δομής των LSTMs οι πύλες εισόδου, εξόδου και λήθης. Σε όλες τις πύλες, υπάρχει σιγμοειδής συνάρτηση, με κύρια διαφορά ότι η πύλη λήθης λειτουργεί με απευθείας πολλαπλασιασμό με την κατάσταση του κελιού, πριν αυτό αλληλεπιδράσει με τις εξόδους των άλλων δύο πυλών. Αυτό έχει ως αποτέλεσμα, να δίνεται έλεγχος στο LSTM, μέσω της εκπαίδευσης, να ελαχιστοποιήσει στοιχεία της κατάστασης κελιού μέσω του πολλαπλασιασμού, κρίνοντας ποιες πληροφορίες του ιστορικού του δικτύου είναι χρήσιμες για τον σκοπό της εκπαίδευσης.

2.1.3 Συνελικτικά LSTM επίπεδα

Η συμβατική αρχιτεκτονική LSTM επιτρέπει την αντιμετώπιση πλήθους εργασιών οι οποίες απαιτούν χρονική συσχέτιση, εφόσον κάθε βήμα της ακολουθίας που εισέρχεται στο δίκτυο, μπορεί να περιγραφεί ως διάνυσμα, δηλαδή αντικείμενο μιας διάστασης. Το γεγονός αυτό περιορίζει τη διαστατικότητα των δεδομένων εισόδου, περιπλέκοντας τη χρήση των LSTMs σε προβλήματα τα οποία απαιτούν τόσο χρονική όσο και χωρική συσχέτιση. Ένα παράδειγμα τέτοιας εργασίας, εκτός από το αντικείμενο της παρούσας διπλωματικής, αποτελεί η πρόβλεψη βίντεο, όπου τα δεδομένα εισόδου μπορούν να περιγραφούν ως πλαίσια ενός βίντεο, τα οποία δεν είναι μονοδιάστατα. Μια λύση αποτελεί ο μετασχηματισμός κάθε βήματος της ακολουθίας σε μονοδιάστατο διάνυσμα, τεχνική η οποία διατηρεί κάποιο βαθμό χρονικής, αλλά όχι χωρικής, συσχέτισης.



Εικόνα 2.5: *Εσωτερική Δομή ενός ConvLSTM*

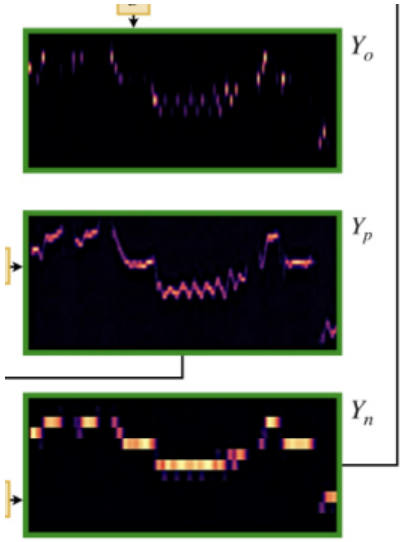
Για την αντιμετώπιση του παραπάνω προβλήματος εισήχθησαν τα συνελικτικά LSTM (Convolutional LSTMs - ConvLSTMs) [4], στα οποία η κρυφή κατάσταση, οι καταστάσεις κελιού και εισόδου, όπως και οι έξοδοι των πυλών είναι τρισδιάστατοι πίνακες. Στα ConvLSTMs πραγματοποιούνται λειτουργίες συνέλιξης στην κρυφή κατάσταση και στην κατάσταση εισόδου, έναντι των πολλαπλασιασμών με πίνακες βαρών που πραγματοποιούνται στα συμβατικά LSTMs.

2.2 Βασικές Έννοιες

2.2.1 Απεικονίσεις Onsets, Offsets, Νότες και Τόνος

Ένα μουσικό έργο, συνήθως, αλλά και εν προκειμένω, χαρακτηρίζεται από μια μελωδία καθώς και κάποια συνοδεία υπό τη μορφή συγχορδιών. Και τα δύο αυτά μέρη του μουσικού έργου, αποτελούνται, προφανώς, από νότες. Ουσιαστικά, στην παρούσα εργασία, η απεικόνιση των νοτών Y_n αποτελεί την συχνοτική απεικόνιση της μελωδίας, συναρτήσει του χρόνου. Αυτή η απεικόνιση, βρίσκεται, εννοιολογικά, πλησιέστερα στην ίδια την παρτιτούρα. Αυτό σημαίνει, ότι αν κάποιος μουσικός επιχειρούσε να παρασκευάσει μια παρτιτούρα από

μια ηχογράφηση, αυτό με το οποίο θα κατέληγε, θα έμοιαζε εξαιρετικά με το Y_n. Το χαρακτηριστικό αυτής της απεικόνισης, είναι ότι κάθε νότα χαρακτηρίζεται από μια ξεκάθαρη αρχή και ένα τέλος, συγκεκριμένης διάρκειας. Η εκκίνηση, αυτή, της κάθε νότας ονομάζεται onset, και η απεικόνιση μόνο των χρονικών στιγμών αυτών, ονομάζεται Y_o. Αντίστοιχα, η απεικόνιση των χρονικών στιγμών λήξης της κάθε νότας, ονομάζεται offset, και συμβολίζεται ως Y_f.



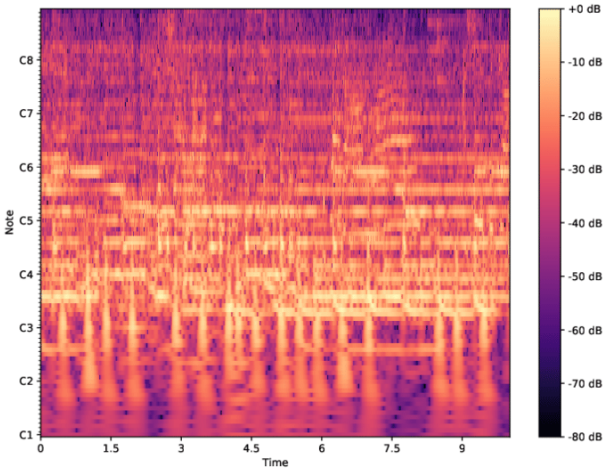
Εικόνα 2.6: *Σύγκριση των απεικονίσεων Τόνου, Onset και Νότας*

Εκτός από τις ίδιες τις νότες, και τα onsets τους, υπάρχει και ένας επιπλέον παράγοντας που αξίζει να μελετηθεί. Σε όλα τα όργανα, και ειδικότερα στην κιθάρα την οποία αφορά το Guitarset, υπάρχει περίπτωση να πραγματοποιηθεί κάποια απόκλιση από την προβλεπόμενη νότα, συχνά επιτηδευμένα. Σε αυτήν τη λογική, βασίζονται τεχνικές όπως vibrato, η οποία είναι η ελαφριά συχνοτική διακύμανση γύρω από μια συγκεκριμένη νότα, ή το slide, όπου για να μεταβεί κανείς από τη μια νότα στην άλλη, περνάει από όλες τις ενδιάμεσες νότες πολύ γρήγορα. Αυτές οι διακυμάνσεις, δεν θα απεικονίζονταν στο Y_n, αφού συνήθως έχουν πολύ μικρή διάρκεια ή πλάτος διακύμανσης, και δεν έχουν ως σκοπό την αλλαγή αλλά την εύηχη αλλοίωση της νότας. Η απεικόνιση τόνου, Y_p, μπορεί να χαρακτηριστεί ως η Y_n, με τις προαναφερθείσες συμπληρωματικές πληροφορίες (Εικόνα 2.6). Σε κάθε περίπτωση, η Y_p είναι μια λεπτομερέστερη περιγραφή της απεικόνισης των νοτών, η οποία όμως περιέχει κάποιες πληροφορίες οι οποίες δεν θα υπήρχε ανάγκη να καταγραφούν στην παρτιτούρα του έργου.

2.2.2 Αρμονικός Μετασχηματισμός Σταθερού-Q

Ο αρμονικός μετασχηματισμός σταθερού-Q (Harmonic Constant-Q Transform - HCQT), αποτελεί κεντρικό άξονα όλων των μοντέλων, αφού είναι ο μηχανισμός μέσω του οποίου το μονοδιάστατο ηχητικό σήμα μετατρέπεται σε τρισδιάστατη απεικόνιση $H[h, f, t]$, όπου f ο άξονας της συχνότητας, t ο άξονας του χρόνου και h ο αρμονικός άξονας. Βασίζεται στον μετασχηματισμό σταθερού-Q (Constanct-Q Transform - CQT) [13], ο οποίος μπορεί να

περιγραφεί ως μια σειρά συχνοτικών φίλτρων, κάθε ένα από τα οποία αντιστοιχεί έναν "κάδο", ο οποίος περιγράφει συχνότητα πολλαπλάσια του αμέσως προηγούμενού του.



Εικόνα 2.7: *Απεικόνιση CQT*

Κατά τον μετασχηματισμό CQT, ο ν-οστός κάδος αντιστοιχεί σε συχνότητα $f_n = 2^{1/C} f_{n-1}$, όπου C ο αριθμός των "κάδων" ανά οκτάβα, το οποίο είναι παράμετρος που μπορεί να επιλεχθεί κάθε φορά. Έτσι, αναδρομικά, προκύπτει ότι:

$$f_n = 2^{n/C} * f_{min} \tag{2.1}$$

, όπου f_{min} η συχνότητα που αντιστοιχεί στο χαμηλότερο, συχνοτικά, "κάδο".

Η απεικόνιση CQT, είναι ταυτόσημη με την αρμονική συνιστώσα $h = 1$ του H, δηλαδή $H[h = 1, f, t]$. Χρησιμοποιώντας το γεγονός ότι κάθε νότα έχει αρμονικές με συχνότητες πολλαπλάσιες της δικής της συχνότητας, μπορεί να υπολογιστεί ο αρμονικός μετασχηματισμός για δεδομένο h, ως:

$$f_n = h * 2^{n/C} * f_{min}, h = [0.5, 1, 2, 3, 4, 5...] \tag{2.2}$$

Στη συνέχεια, μετά τον υπολογισμό του μετασχηματισμού της Εξίσωσης 2.2 για έναν ορισμένο αριθμό h, η διαδικασία ολοκληρώνεται με το αρμονικό στοίβαγμα των συνιστωσών για κάθε διαφορετικό h, ώστε να σχηματιστεί ο τρισδιάστατος πίνακας $H[h, f, t]$. Αυτός ο μετασχηματισμός, υπερέχει του απλού CQT, αφού αυτός μπορεί να συλλάβει μόνο άρτια πολλαπλάσια της συχνότητας του κάθε "κάδου", όπως προκύπτει από την Εξίδωση 2.1. Η μέθοδος HCQT, μπορεί να συλλάβει όχι μόνο τα άρτια, αλλά και τα περιττά πολλαπλάσια, διευρύνοντας το περιεχόμενο της αρμονικής πληροφορίας που παρέχεται στο νευρωνικό δίκτυο. Ο συγκεκριμένος μετασχηματισμός τοποθετείται, για τους παραπάνω λόγους, στην αρχή κάθε νευρωνικού δικτύου που θα εξεταστεί.

Μέρος II

Πρακτικό Μέρος

Κεφάλαιο 3

Περιγραφή Dataset και Μετρικών

3.1 Guitarset

Όπως έχει αναφερθεί στην εισαγωγή της παρούσας διπλωματικής, η συλλογή δεδομένων που χρησιμοποιήθηκε για την εκπαίδευση, καθώς και τη σύγκριση όλων των διαφορετικών μοντέλων, είναι το Guitarset [1]. Για τον σκοπό της δημιουργίας του συγκεκριμένου συνόλου δεδομένων, εργάστηκαν 6 έμπειροι κιθαρίστες, από τους οποίους ζητήθηκε να εκτελέσουν 30 διαφορετικές συνθέσεις, για τις οποίες τους ζητήθηκε να παράσχουν συνοδεία με τη μορφή συγχορδιών, και μελωδική γραμμή πάνω από τις συγχορδίες που επέλεξαν.

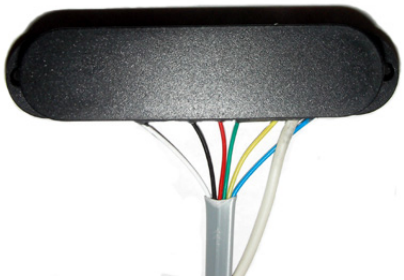
Για κάθε μια από τις συνθέσεις, είχαν διαθέσιμο μετρονόμο, ο οποίος προσέδιδε τον ρυθμό της σύνθεσης, καθώς και ηχογραφημένα ντραμς και μπάσο.Έτσι, το συγκεκριμένο σύνολο δεδομένων, αποτελείται από 360 ηχογραφήσεις κιθάρας, οι μισές από τις οποίες αποτελούν την εκτέλεση κάποιας μελωδίας, ενώ οι άλλες μισές την συνοδεία μέσω συγχορδιών, των αντίστοιχων μελωδιών. Η κάθε μια από αυτές τις ηχογραφήσεις, διατίθεται 4 εκδοχές, ανάλογα με το μέσο ηχογράφησης που χρησιμοποιήθηκε για την δημιουργία της:

1. Εξαφωνικές Ηχογραφήσεις (Hexaphonic Recordings)
2. Εξαφωνικές Debleeded Ηχογραφήσεις (Hexaphonic Debleeded Recordings)
3. Μονοκαναλικές Ηχογραφήσεις (Mono Recordings)
4. Εξαφωνικές-Μονοκαναλικές Μεικτές Ηχογραφήσεις (Hexaphonic-Mono Mix Recordings)

3.1.1 Αρχεία Ήχου

Για τις εξαφωνικές ηχογραφήσεις, χρησιμοποιήθηκε εξαφωνικός μαγνήτης, ο οποίος στερεώθηκε κάτω από τις χορδές της ακουστικής κιθάρας, με τρόπο τέτοιο ώστε κάθε ένας από τους 6 μικρούς μαγνήτες που τον αποτελούν να βρίσκεται ακριβώς κάτω από μια διαφορετική χορδή. Κάθε ένας από τους μικρούς μαγνήτες εντός του εξαφώνου, κάθε ένας από τους οποίους αντιστοιχεί σε μια χορδή, έχει μια ξεχωριστή έξοδο, γεγονός που απομονώνει το ήχο κάθε χορδής σε μεγαλύτερο βαθμό από ότι ένας συμβατικός μαγνήτης. Η εξαφωνική εκδοχή της κάθε ηχογράφησης σχηματίζεται συνδυάζοντας τις 6 αυτές εξόδους σε ένα αρχείο ήχου.

Η εξαφωνική debleeded εκδοχή προκύπτει εφαρμόζοντας debleeding στις αρχικές εξαφωνικές ηχογραφήσεις, χρησιμοποιώντας τον αλγόριθμο KAMIR [14]. Είναι χρήσιμο να



Εικόνα 3.1: *Παράδειγμα Εξαφωνικού Μαγνήτη Κιθάρας*

εφαρμοστεί debleeding στις αρχικές ηχογραφήσεις, αφού υπάρχει περίπτωση να ηχογραφηθεί κάποια χορδή σε μαγνήτη διαφορετικό από αυτόν που της αντιστοιχεί, ακόμη και χρησιμοποιώντας εξαφωνικό pickup. Αξίζει να σημειωθεί, ότι το παραπάνω φαινόμενο θα ήταν εξαιρετικά ενισχυμένο στην περίπτωση χρήσης ενός συμβατικού μαγνήτη κιθάρας.

Οι μονοκαναλικές ηχογραφήσεις πραγματοποιήθηκαν παράλληλα με τις εξαφωνικές, τοποθετώντας ένα Neumann U87 πυκνωτικό μικρόφωνο σε απόσταση περίπου 30 εκατοστών από το 18^{ο} τάστο της κιθάρας. Αυτός ο τρόπος ηχογράφησης είναι και ο πιο συνηθισμένος σε συνθήκες ηχογράφησης ακουστικής κιθάρας για κάποιο έργο, οπότε έχει αξία ως συνεισφορά στη συλλογή δεδομένων. Τέλος, υπάρχει και η εξαφωνική-μονοκαναλική μεικτή εκδοχή της κάθε ηχογράφησης, η οποία ουσιαστικά συνδυάζει τα 6 κανάλια του εξαφώνου με το 1 κανάλι του πυκνωτικού μικροφώνου, ώστε να παραχθεί ένα μεικτό αρχείο ήχου.

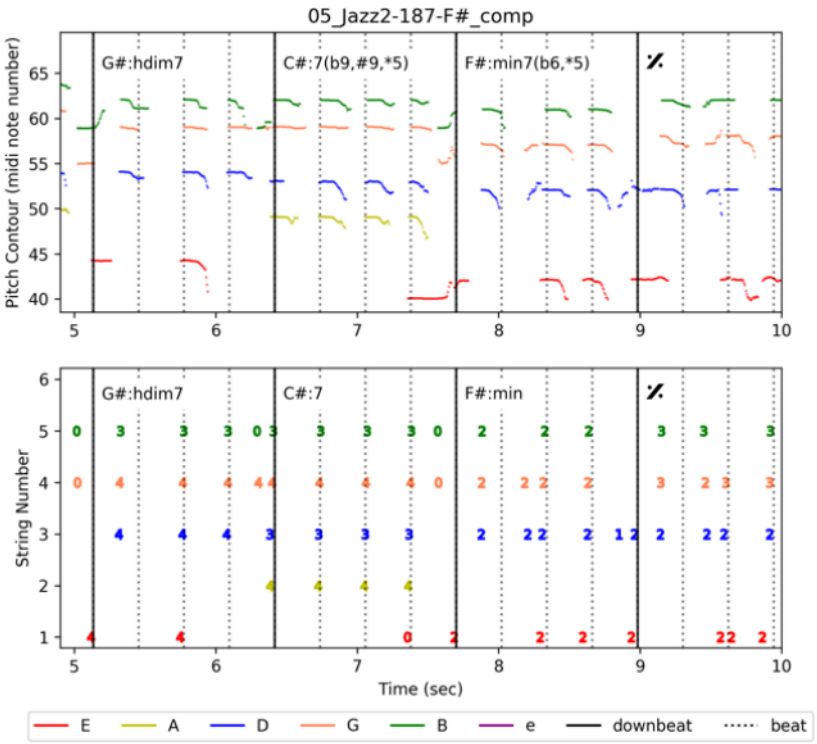
3.1.2 Annotations

Τα αρχεία ήχου που περιλαμβάνονται στη συλλογή δεδομένων αποτελούν την είσοδο των μοντέλων που θα συγκριθούν. Αντίστοιχα, τα στόχοι ή ετικέτες που θα χρησιμοποιηθούν στο μοντέλο, είναι τα annotations. Μπορούν να θεωρηθούν ως μια απεικόνιση συχνότητας - χρόνου, για κάθε ηχογράφηση και διακρίνονται σε 4 διαφορετικούς τύπους, ανάλογα με το τι περιγράφουν:

1. Annotation Τόνου
2. Annotation Νότας
3. Onsets Annotation
4. Offsets Annotation

Τα παραπάνω annotations, αντιστοιχούν 1 προς 1 με τις αντίστοιχες απεικονίσεις οι οποίες είχαν περιγραφεί στο προηγούμενο Κεφάλαιο, και αντιστοιχούν στα Y_p, Y_n, Y_o και Y_f. Αναλόγως της αρχιτεκτονικής και των απαιτήσεων του κάθε μοντέλου, μπορεί να χρησιμοποιηθούν όλοι οι τύποι απεικονίσεων ή κάποιο υποσύνολο αυτών. Στην περίπτωση του Deep Salience μοντέλου [2], χρησιμοποιούνται μόνο τα Annotations τόνου, καθιστώντας αυτή τη μέθοδο την πιο απλοϊκή που θα εξεταστεί.

Για την κατασκευή των Annotations επιστρατεύτηκαν οι εξαφωνικές ηχογραφήσεις, αφού παρέχουν διαφορετικό κανάλι για κάθε χορδή. Πραγματοποιήθηκε προεπεξεργασία στις



Εικόνα 3.2: *Annotations Τόνου μαζί με την αντίστοιχη ταμπλατούρα*

ηχογραφήσεις εφαρμόζοντας debleeding [14] και δημιουργώντας μια αρχική απεικόνιση Y_n, μέσω της PYIN βιβλιοθήκης [15], η οποία τέθηκε προς χειροκίνητη επικύρωση. Έτσι τα annotations των onsets, offsets και τόνου προκύπτουν με ημιαυτόματο τρόπο, αφού επικυρώνονται τελικά χειροκίνητα. Αφού δημιουργήθηκαν τα annotations για κάθε ένα από τα 6 κανάλια, συναθροίστηκαν ώστε να σχηματίσουν τα τελικά annotations τα οποία παρέχονται στο Guitarset.

3.2 Μετρικές

Για την αξιολόγηση και τη σύγκριση των διάφορων μοντέλων χρειάζεται, προφανώς, η επιστράτευση ορισμένων μετρικών. Λόγω της φύσης του προβλήματος της μεταγραφής μουσικής, θα συνέφερε, επιπλέον, η εύρεση μετρικών οι οποίες εξειδικεύονται στον ήχο. Έχοντας αυτά κατά νου, καθώς και το περιεχόμενο σχετικών δημοσιεύσεων, χρησιμοποιήθηκε η βιβλιοθήκη mir_eval [16]. Η συγκεκριμένη βιβλιοθήκη περιλαμβάνει ένα σύνολο μετρικών οι οποίες σχετίζονται άμεσα με τη μουσική, συνυπολογίζοντας στοιχεία όπως τα onsets, offsets και τη συχνοτική διαφορά της εξόδου του δικτύου από τις ετικέτες.

3.2.1 Πιστότητα σε Επίπεδο Πλαισίου

Για όλα τα μοντέλα που εκπαιδεύτηκαν, χρησιμοποιήθηκε απώλεια διασταυρούμενες εντροπίας ως συνάρτηση βελτιστοποίησης. Αναλόγως το μοντέλο, μπορούσε να είναι μια

μόνο συνάρτηση της εξόδου του δικτύου σε σχέση με την απεικόνιση Y_p (Deep Salience) είτε άθροισμα επιμέρους απωλειών διασταυρούμενης εντροπίας όλων ή μέρους των απεικονίσεων Y_p, Y_n, Y_o και Y_f. Παρόλα αυτά, για την σωστή εκτίμηση της ποιότητας των προβλέψεων του κάθε δικτύου, είναι χρήσιμο να γίνει μια διαφορετική τύπου σύγκριση, αφού η ελαχιστοποίηση της απώλειας διασταυρούμενης εντροπίας δεν εγγυάται την εγγύτητα της πρόβλεψης με τον στόχο, με μουσικούς όρους.

Η πρώτη, και απλούστερη, μετρική που χρησιμοποιήθηκε, είναι η πιστότητα σε επίπεδο πλαισίου (frame level accuracy - FLA) . Μάλιστα, στο Deep Salience μοντέλο είναι η μόνη μετρική που χρησιμοποιήθηκε, αφού η αρχιτεκτονική είναι σχεδιασμένη να προβλέπει μόνο την Y_p ή απεικόνιση τόνου. Η FLA υπολογίζεται ως εξής (Εξίσωση 3.1):

$$Accuracy = \frac{TP}{TP + FP + FN} \tag{3.1}$$

όπου TP - true positive (αληθώς θετικά), FP - false positive (ψευδώς Θετικά) και FN - false negative (ψευδώς αρνητικά).

Ο παραπάνω υπολογισμός της FLA υπονοεί τη δυαδική φύση των ετικετών αλλά και των εξόδων του δικτύου, ώστε να ορίζονται μεγέθη όπως τα αληθώς θετικά. Στο στάδιο της προεπεξεργασίας κάθε μοντέλου, μετατρέπονται τα annotations σε απεικονίσεις συχνότητας-χρόνου, δημιουργώντας δυαδικούς "κάδους" και στις δύο αυτές διαστάσεις. Καθώς το μοντέλο εκπαιδεύεται ελαχιστοποιώντας την απώλεια διασταυρούμενης εντροπίας, οι τιμές των στοιχείων των εξόδων διαμορφώνονται ανάμεσα στο 0 και το 1. Αυτές οι τιμές μπορούν να ερμηνευτούν ως πιθανότητες, και να τεθεί ένα κατώφλι το οποίο όταν μια τιμή το ξεπερνά να θεωρείται 1 (και αντίθετα να θεωρείται 0). Κατά αυτόν τον τρόπο, διαμορφώνονται τόσο οι έξοδοι όσο και οι ετικέτες ως δυαδικές απεικονίσεις, καθιστώντας δυνατή την αποτίμηση της Εξίσωσης 3.1. Τέλος, να σημειωθεί ότι η προαναφερόμενη Εξίσωση αποτιμάται για οποιαδήποτε εκ των 4 απεικονίσεων (νότες, τόνους, onsets & offsets), αλλά στη συγκεκριμένη εργασία υπολογίζεται για τις Y_p και Y_n, και ονομάζεται πιστότητα τόνου και νότας, αντίστοιχα.

3.2.2 Μετρικές F

Η μετρική FLA είναι ένα άρτιο σημείο εκκίνησης για την αξιολόγηση των προβλέψεων κάθε μοντέλου, αλλά εμφανίζει κάποια μειονεκτήματα. Αρχικά, αφορά, μόνο την Y_p ή Y_n απεικόνιση, χωρίς να περιέχεται πληροφορία σχετικά με τα σημεία έναρξης των νοτών ή να υπάρχει για χρονική ανοχή. Είναι λογικό, εξάλλου, να θεωρείται μια πρόβλεψη λιγότερο σωστή αν βρίσκεται σε μεγαλύτερη χρονική απόσταση από την πραγματική νότα. Ένα άλλο στοιχείο μιας καλής μετρικής είναι ότι θα έπρεπε, ιδανικά, να υπάρχει κάποιο συχνοτικό εύρος γύρω από την κάθε συχνότητα στόχο, γύρω από το οποίο η πρόβλεψη να θεωρείται σωστή. Αυτό έχει μουσικό νόημα, αφού, ιδιαίτερα όταν οι συχνοτικοί "κάδοι" έχουν μικρό εύρος, η πολύ μικρή απόκλιση της πρόβλεψης από τη συχνότητα στόχο μπορεί να είναι ανεπαίσθητη στο ανθρώπινο αυτί. Τέλος, μέσω της πιστότητας δεν ελέγχεται η σωστή διάρκεια της κάθε νότας. Αυτό σημαίνει ότι μπορεί σε μια περίπτωση να έχει αποτύχει να προβλεφθεί σωστά ένας μόνο "κάδος" εντός της διάρκειας μια νότας και να υπάρξει πολύ υψηλή πιστότητα. Στην πραγματικότητα, όμως, θα έχει μεταγραφεί η εκτέλεση της νότας δύο αντί της μιας

φοράς, όπως θα έπρεπε.

Για την επίλυση των παραπάνω προβλημάτων, χρησιμοποιήθηκαν οι μετρικές F και Fno. Και οι δύο αυτές μετρικές αποτελούν μετρικές σε επίπεδο νότας, αντί πλαισίου ή χρονικού "κάδου". Αυτό έχει ως αποτέλεσμα, να αξιολογούνται οι προβλέψεις βάσει ολόκληρων των νοτών που προβλέπουν, κάτι που προσδίδει μεγαλύτερο μουσικό νόημα. Η μετρική F αξιολογεί ως σωστή μια νότα, εφόσον τηρεί ορισμένες προϋποθέσεις που αφορούν τα onsets, offsets και τον τόνο της νότας, δηλαδή το συχνοτικό "κάδο" στον οποίο έχει γίνει η πρόβλεψη από το μοντέλο. Οι συνθήκες που πρέπει να ικανοποιούνται ταυτόχρονα ώστε μια νότα να θεωρηθεί ότι έχει προβλεφθεί σωστά, είναι οι εξής:

- Το onset να έχει προβλεφθεί εντός χρονικού διαστήματος εύρους 50ms γύρω από το onset αναφοράς.
- Η συχνότητα που έχει προβλεφθεί για μια νότα να βρίσκεται εντός εύρους συχνότητας $\frac{1}{4}$ τόνου (50 cents) από τη συχνότητα αναφοράς.
- Το offset να έχει προβλεφθεί εντός χρονικού διαστήματος εύρους 50ms ή ίσου με το 20% της συνολικής διάρκειας της νότας γύρω από το offset αναφοράς, ανάλογα με το ποιο εκ των δύο μεγεθών είναι μεγαλύτερο.

Οι συνθήκες των onsets και offsets εκφράζουν την ελάχιστη χρονική διαφορά που επιτρέπεται να εμφανίζει η πρόβλεψη σε σχέση με τον στόχο, ώστε το αποτέλεσμα να είναι πρακτικά το ίδιο, σε ότι αφορά τη μουσική σύνθεση και την μεταγραφή της. Το κριτήριο της διαφοράς όσο αφορά τη συχνότητα, τίθεται στο $\frac{1}{4}$ του τόνου, αφού , τουλάχιστον στη δυτική μουσική, θεωρείται ότι η ελάχιστη απόσταση μεταξύ 2 νοτών είναι $\frac{1}{2}$ τόνος.

Κεφάλαιο 4

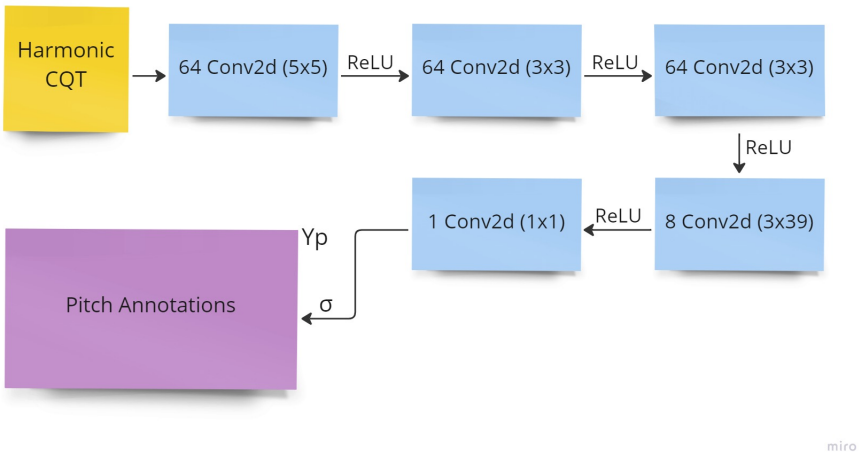
Υλοποίηση

4.1 Deep Salience

4.1.1 Περιγραφή Μοντέλου

Το πρώτο μοντέλο το οποίο εξετάστηκε (Εικόνα 4.1), αποτελεί την πιο απλοϊκή προσέγγιση του προβλήματος προς επίλυση [2]. Η αρχιτεκτονική περιλαμβάνει ένα σχετικά απλό CNN μοντέλο, με το οποίο επιχειρείται η πρόβλεψη της Y_p, δηλαδή της απεικόνισης τόνου του κάθε δείγματος ήχου εισόδου. Η μόνη μετρική που χρησιμοποιήθηκε εν προκειμένω, ήταν η πιστότητα τόνο αφού οι μετρικές F προϋποθέτουν την πρόβλεψη επιπλέον απεικονίσεων.

Αρχικά, πραγματοποιούνται 5x5 συνελίξεις στα δεδομένα εισόδου, παράγοντας έξοδο βάθους 128, όπως προκύπτει από το πλήθος των φίλτρων του συνελικτικού στρώματος. Τέλος, οι έξοδοι από το πρώτο συνελικτικό στρώμα, διέρχονται από ReLU συνάρτηση ενεργοποίησης, πριν εισέλθουν στο επόμενο στρώμα. Με την ίδια λογική δομούνται τα επόμενα συνελικτικά στρώματα, με συνελίξεις διαστάσεων 5x5, 3x3, 3x3, 39x3, 1x1 και πλήθος φίλτρων 64, 64, 64, 8, 1 αντίστοιχα. Στο τελευταίο στρώμα, η συνάρτηση ενεργοποίησης διαφοροποιείται από τη ReLU, και χρησιμοποιείται η λογιστική.

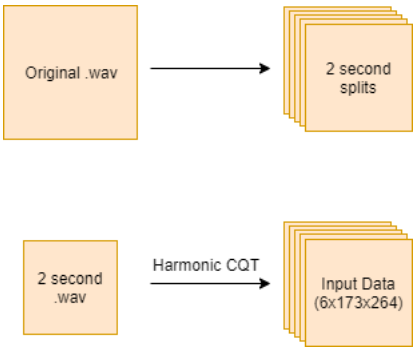


Εικόνα 4.1: *Αρχιτεκτονική Deep Salience CNN Μοντέλου*

4.1.2 Προεπεξεργασία

Πριν την είσοδό τους στο εκπαιδεύσιμο δίκτυο, πραγματοποιήθηκαν κάποιοι μετασχηματισμοί τόσο στα δεδομένα εισόδου, όσο και στα annotations τόνου τα οποία αποτέλεσαν τους στόχους της εκπαίδευσης. Στα αρχεία ήχου, εφαρμόστηκε, μετά από επαύξηση δεδομένων, ο μετασχηματισμός HCQT, αφού πρώτα κάθε αρχείο χωριστεί σε ίσαξια τμήματα 2 δευτερολέπτων, εφαρμόζοντας το κατάλληλο "γέμισμα", όπου αυτό χρειάζεται. Κατά τον HCQT, χρησιμοποιήθηκαν 6 αρμονικές, όπως προτείνεται από την αρχική δημοσίευση της μεθόδου. Για την επαύξηση των δεδομένων, στο συγκεκριμένο, όπως και στα άλλα, μοντέλο, χρησιμοποιήθηκε Reverb και Gain, τα οποία είναι 2 εφέ που χρησιμοποιούνται στη μουσική κατά κόρον. Το πρώτο μπορεί να παρομοιαστεί με ηχώ, ενώ το δεύτερο με επιτηδευμένη παραμόρφωση του ήχου.

Για την εκπαίδευση όλων των δικτύων, χρησιμοποιήθηκαν τα εξαφωνικά και debleeded εξαφωνικά αρχεία ήχου, ώστε να υπάρχει σύγκριση επί ίσοις όροις. Ο διαχωρισμός των δεδομένων σε σύνολα εκπαίδευσης επικύρωσης και ελέγχου πραγματοποιήθηκε με αναλογία 80:10:10, για τον οποίο αξίζει να σημειωθεί μια λεπτομέρεια: όπως έχει διευκρινιστεί σε προηγούμενο κεφάλαιο, κάθε αρχείο μελωδίας έχει και ένα αρχείο με τη μουσική συνοδεία συγχορδιών. Έτσι, ο διαχωρισμός έγινε με τέτοιο τρόπο, ώστε αν ένα εκ των δύο αρχείων του ζεύγους μελωδίας-συνοδείας ανήκε σε ένα στο ένα σύνολο, τότε να ανήκει και το άλλο. Με αυτόν τον τρόπο, αποφεύγεται η παροχή επιπλέον πληροφορίας για δείγμα του συνόλου επικύρωσης ή εκπαίδευσης κατά την εκπαίδευση, αλλά παρέχεται και η πλήρης πληροφορία κάθε φορά (Εικόνα 4.2).

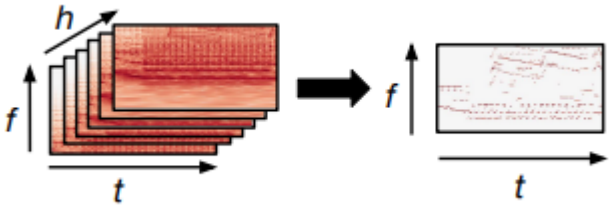


Εικόνα 4.2: *Περιγραφή της διαδικασίας προεπεξεργασίας*

Μεταβαίνοντας στα Annotations τόνου Y_p, τα οποία αποτέλεσαν τους στόχους του δικτύου, ήταν απαραίτητη η δημιουργία κατάλληλων "κάδων" τόσο στον άξονα της συχνότητας, όσο και στον άξονα του χρόνου. Κάτι τέτοιο είναι απαραίτητο, ώστε να υπάρχει η ίδια διάσταση στους στόχους, όπως και στις εξόδους του δικτύου, οι οποίες είναι δισδιάστατες (Εικόνα 4.3). Η ανάλυση που επιλέχθηκε για τα "κάδους" είναι 3 ανά ημιτόνιο και 173 ανά 2 δευτερόλεπτα στη διάσταση του χρόνου.

4.1.3 Μετα-επεξεργασία

Μετά την εκπαίδευση του μοντέλου, όλα τα στοιχεία των εξόδων του βρίσκονται εντός του εύρους 0 και 1, λόγω της λογιστικής συνάρτησης ενεργοποίησης που εφαρμόστηκε μετά το



Εικόνα 4.3: *Αναπαράσταση των δεδομένων εισόδου (αριστερά) και των στόχων (δεξιά)*

τελευταίο συνελικτικό στρώμα. Παρόλα αυτά, οι αναπαραστάσεις Y_p, που είναι οι στόχοι, είναι δυαδικές απεικονίσεις οι οποίες περιλαμβάνουν είτε 0 είτε 1. Επιπλέον, τα στοιχεία των εξόδων του δικτύου, μπορούν να ερμηνευτούν ως πιθανότητες το συγκεκριμένο στοιχείο να είναι 0 ή 1. Αυτό σημαίνει, ότι απαιτείται κάποιο βήμα μετα-επεξεργασίας, ώστε οι έξοδοι του δικτύου να μετατραπούν σε δυαδικές απεικονίσεις, για τον υπολογισμό του πιστότητας του τόνου. Ο πιο απλός τρόπος, θα ήταν η χρήση μιας τιμής κατωφλίου, βάσει της οποίας θα τίθεται ένα στοιχείο ίσο με 1 όταν την υπερβαίνει (και 0 αντίθετα).

Η ερμηνεία των στοιχείων της πρόβλεψης του μοντέλου ως πιθανότητα, θα υπονοούσε την επιλογή του παραπάνω ορίου ως 0.5 ή 50%. Κάτι τέτοιο, παρότι διαισθητικά ορθό, ίσως περιόριζε την απόδοση του μοντέλου. Για αυτόν το λόγο, θεωρήθηκε ορθότερο να θεωρηθεί ως μια παράμετρος του σταδίου μεταεπεξεργασίας και να εξερευνηθεί πλήθος τιμών ανάμεσα στο 0 και το 1 για αυτό το κατώφλι. Αυτό θα είναι και ένα από τα αποτελέσματα το οποίο θα συζητηθεί στην επόμενη ενότητα.

4.1.4 Αποτελέσματα

4.1.4.1 Εκπαίδευση Μοντέλου

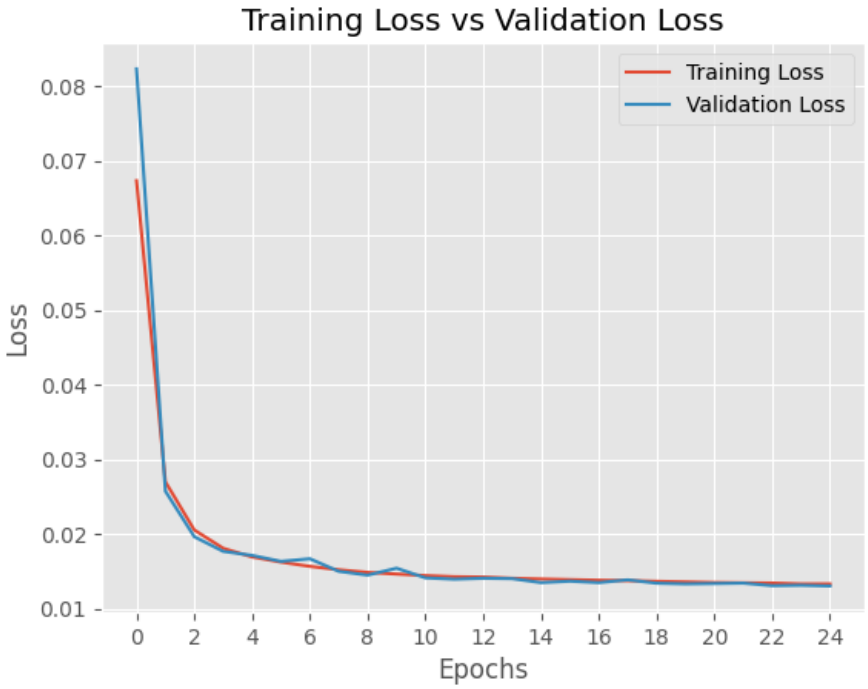
Προτού αξιολογηθεί το μοντέλο για την απόδοσή του είναι απαραίτητη, προφανώς, η εκπαίδευσή του. Ανάμεσα στους 4 τύπους ηχογραφήσεων για κάθε απόδοση, βάσει του μέσου ηχογράφησης, αποφασίστηκε η χρήση των εξαφωνικών και των debleeded εξαφωνικών ηχογραφήσεων. Η ίδια η εκπαίδευση, πραγματοποιήθηκε με τα εξής χαρακτηριστικά (Πίνακας 4.1):

Υπερπαράμετρος	Τιμή
Μέγεθος Δέσμης	64
Ρυθμός Μάθησης	10^{-4}
Εποχές	20
Βελτιστοποίηση	Adam [17]
Συνάρτηση Απώλειας	Διασταυρούμενη Εντροπία

Πίνακας 4.1: *Υπερπαράμετροι του μοντέλου Deep Salience*

Εκτός από τα προαναφερθέντα χαρακτηριστικά, χρησιμοποιήθηκε τεχνική πρόορου τερματισμού, ώστε να αποθηκευτούν τα βάρη του μοντέλου που εμφάνισαν τη βέλτιστη απόδοση βάσει της απώλειας διασταυρούμενης εντροπίας στο σύνολο δεδομένων επικύρωσης.

Όπως παρατηρείται στην Εικόνα 4.4, οι δύο απώλειες, εκπαίδευσης και επικύρωσης,



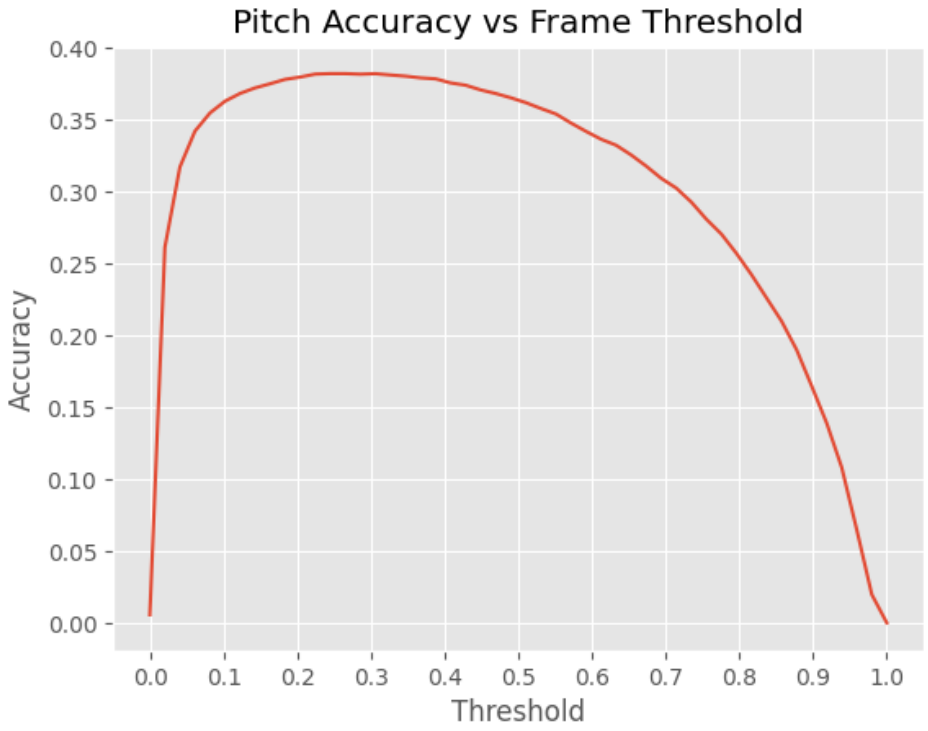
Εικόνα 4.4: *Απώλεια εκπαίδευσης (κόκκινο) και επικύρωσης (μπλέ) για την εκπαίδευση του Deep Salience μοντέλου*

παραμένουν κοντά μεταξύ τους, με την απώλεια επικύρωσης να είναι σταθερά μεγαλύτερη, όπως αναμένεται. Παρόλα αυτά, δεν παρατηρείται το φαινόμενο υπερ-προσαρμογής, όπως φαίνεται και από την απόδοση στο σύνολο επικύρωσης. Με τις παραπάνω υπερπαραμέτρους και την αρχιτεκτονική του μοντέλου, φαίνεται να ελαχιστοποιείται, λοιπόν, η απώλεια διασταυρούμενης εντροπίας εντός των 20 εποχών που επιλέχθηκαν.

4.1.4.2 Επικύρωση Μοντέλου

Όπως σημειώθηκε, η διαδικασία της εκπαίδευσης, μπορεί να περιγραφεί ως η βελτιστοποίηση της συνάρτησης απώλειας που επιλέχθηκε, εδώ της διασταυρούμενης εντροπίας. Παρόλα αυτά, ο σκοπός είναι η βελτιστοποίηση πιστότητας τόνου. Για να πραγματοποιηθεί αυτό, όπως έχει ήδη ειπωθεί, είναι απαραίτητο να τεθεί ένα κατώφλι άνω του οποίου οι τιμές της εξόδου τίθενται ίσες με 1, αλλιώς 0.

Για την επιλογή του κατάλληλου κατωφλίου διατυπώνεται ένα νέο πρόβλημα βελτιστοποίησης, το οποίο σχετίζεται, αυτή τη φορά, με την πιστότητα τόνου. Προς επίλυση αυτού, υπολογίζεται η πιστότητα για 50 ισαπέχουσες τιμές του ορίου, από το 0 ως το 1. Το εύρος αυτό, καθορίζεται, ουσιαστικά, από τη σιγμοειδή συνάρτηση ενεργοποίησης για την παραγωγή της εξόδου. Ακολουθώντας αυτήν τη μεθοδολογία, εντοπίζεται η μέγιστη ακρίβεια **38.21%** για όριο **0.245**, χρησιμοποιώντας, για την επιλογή του, το σύνολο δεδομένων επικύρωσης (Εικόνα 4.5). Τέλος, χρησιμοποιώντας το όριο που προέκυψε μέσω της διαδικασίας επικύρωσης, πραγματοποιήθηκε αξιολόγηση του μοντέλου στο σύνολο δεδομένων ελέγχου ώστε να υπάρχει πιο αντικειμενική εικόνα της απόδοσής του.



Εικόνα 4.5: *Πιστότητα τόνου του μοντέλου συναρτήσει του κατωφλίου*

4.1.4.3 Αξιολόγηση Μοντέλου

Αφού επιλέχθηκε το κατώφλι που θα χρησιμοποιηθεί, μένει να εφαρμοστεί η αξιολόγηση του μοντέλου στο σύνολο ελέγχου, το οποίο δεν έχει αλληλεπιδράσει με κανέναν τρόπο με το ίδιο το μοντέλο. Αυτό, διασφαλίζει την, όσο το δυνατόν, αντικειμενικότερη αξιολόγηση.

Σύνολο Ελέγχου	Πιστότητα τόνου
Εξαφωνικό	37.39%
Εξαφωνικό Debleeded	37.25%
Εξαφωνικό και Εξαφωνικό Debledeed	37.32%
Μονοκάναλο	36.04%
Μείξη μονοκάναλου και εξαφωνικού	36.78%

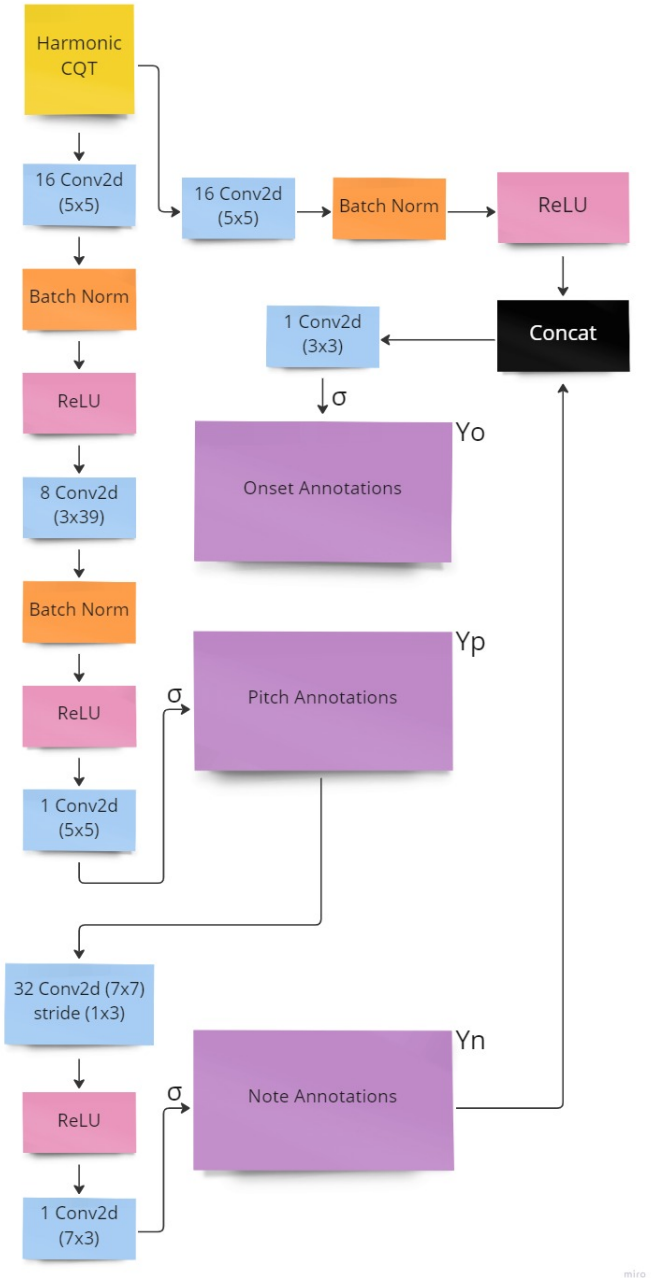
Πίνακας 4.2: *Σύγκριση της πιστότητας τόνου σε διαφορετικά σύνολα ελέγχου*

Όπως έχει διατυπωθεί στην περιγραφή του Guitarset, υπάρχουν 4 τύποι ηχογραφήσεων του κάθε μουσικού έργου. Για κάθε έναν από τους τύπους, παρουσιάζεται η πιστότητα τόνου στον Πίνακα 4.2, το οποίο κυμαίνεται μεταξύ του 36% και 38%. Επιπλέον, έχει πραγματοποιηθεί αξιολόγηση και στο Εξαφωνικό και Εξαφωνικό Debledeed σύνολο ελέγχου, το οποίο έχει την ίδια δομή με τα δεδομένα εκπαίδευσης και επικύρωσης του μοντέλου.

4.2 Basic Pitch

4.2.1 Περιγραφή Μοντέλου

Το δεύτερο, και τελευταίο μοντέλο της βιβλιογραφίας, που εξετάστηκε, αποτελεί το Basic Pitch [3] (Εικόνα 4.6), το οποίο είναι πιο σύνθετο από το Deep Salience, τόσο ως προς την αρχιτεκτονική, όσο και τα στάδια της προ και μετα-επεξεργασίας. Η προσέγγιση αυτή, συμπεριλαμβάνει τις απεικονίσεις των Onsets (Y_o), Τόνου (Y_p) και Νοτών (Y_n), και αξιολογεί το μοντέλο τόσο βάσει της πιστότητας τόνου όσο και βάσει των δύο μετρικών F που έχουν ήδη αναφερθεί.



Εικόνα 4.6: *Αρχιτεκτονική Basic Pitch Μοντέλου*

Το πρώτο τμήμα του μοντέλου, εμφανίζει εξαιρετική ομοιότητα στην δομή με το Deep Salience, μιας και προέκυψε με αυτό κατά νου. Αφού πραγματοποιηθεί ο HCQT μετασχηματισμός, γίνονται 5x5 συνελίξεις με βάθος 16, ακολουθούμενες από ένα στρώμα κανονικοποίησης δέσμης και τέλος ReLU συνάρτηση ενεργοποίησης. Στη συνέχεια ακολουθεί η ίδια διαδικασία, διαφέροντας στο βάθος το οποίο είναι 8, και στον πυρήνα, ο οποίος είναι διάστασης 3x39. Τέλος, πραγματοποιείται συνέλιξη βάθους 1, με πυρήνα 5ξ5 η οποία ακολουθείται από σιγμοειδή ενεργοποίηση. Η έξοδος του πρώτου αυτού κλάδου του δικτύου, συγκρίνεται με τα Annotations τόνου Y_p, τα οποία επιχειρεί να προβλέψει.

Για τον πρώτο κλάδο της αρχιτεκτονικής, η δομή είναι ίδιας μορφής με το μοντέλο Deep Salience, με αρκετά λιγότερες παραμέτρους. Αυτό επιλέχθηκε, ώστε να είναι το μοντέλο πιο αποδοτικό και να μειωθεί ο χρόνος εκπαίδευσης και πρόβλεψης. Η ιδιαιτερότητα του Basic Pitch, έγκειται στη χρήση της πρόβλεψης του Y_p, ώστε να προβλεφθούν εν συνεχεία τα Y_n και τα Y_o. Επιπλέον, έχει διαισθητική σημασία και η σειρά με την οποία επιλέχθηκαν οι προβλέψεις, αφού το Y_p εμπεριέχει όλη την πληροφορία του Y_n και με τη σειρά του το Y_n όλη την πληροφορία του Y_o.

Συνεχίζοντας στον δεύτερο κλάδο του μοντέλου, η πρόβλεψη για το Y_p, υποβάλλεται σε 7x7 συνελίξεις βάθους 32 με βηματισμό 1x3, προτού εισέλθει από συνάρτηση ενεργοποίησης ReLU. Ο βηματισμός στα προηγούμενα στρώματα συνελίξεων δεν αναφέρθηκε επειδή ήταν παντού ίσος με 1. Η επιλογή του βηματισμού ίσου με 3 στη διάσταση των συχνοτήτων, πραγματοποιείται ώστε να υποτριπλασιαστεί η διαστατικότητα όσον αφορά τις συχνότητες, ώστε να καθρεφτίζεται ο επίσης υποτριπλασιασμός των "κάδων" των annotations νότας Y_n σε σχέση με τα annotations τόνου Y_p. Τέλος, πραγματοποιούνται 7x3 συνελίξεις βάθους 1 και εφαρμόζεται σιγμοειδής συνάρτηση ενεργοποίησης ώστε να εξαχθεί η πρόβλεψη για το Y_o.

Στον τρίτο, και τελικό, κλάδο του μοντέλου, η πρόβλεψη για το Y_n στοιβάζεται με τα αρχικά δεδομένα εισόδου του δικτύου, αφού αυτά έχουν περάσει από 5x5 συνελίξεις βάθους 32, και βηματισμού 1x3, έπειτα στρώμα κανονικοποίησης δέσμης και τελικά ReLU συνάρτηση ενεργοποίησης. Ο βηματισμός, όπως και σε προηγούμενο στάδιο, επιλέχθηκε 1x3, ώστε να υποτριπλασιαστεί η διάσταση των συχνοτήτων. Αφού στοιβαχθούν, υπόκεινται σε 3x3 συνελίξεις βάθους 1, και τελικά προκύπτει η πρόβλεψη για το Y_o αφού εφαρμοστεί σιγμοειδής ενεργοποίηση.

Για την εκπαίδευση του δικτύου, εφόσον υπάρχουν 3 έξοδοι, έπρεπε να χρησιμοποιηθεί μια κατάλληλη συνάρτηση βελτιστοποίησης, η οποία θα συμπεριελάμβανε όλες τις προβλέψεις του δικτύου. Η συνάρτηση που επιλέχθηκε είναι το άθροισμα των επιμέρους απωλειών διασταυρούμενης εντροπίας για κάθε μια εκ των προβλέψεων (Εξίσωση 4.1).

$$L = L_p + L_n + L_o \tag{4.1}$$

4.2.2 Προ-επεξεργασία

Το στάδιο της προ-επεξεργασίας, εμφανίζει πολλά κοινά σημεία με την προ-επεξεργασία στην περίπτωση του Deep Salience. Τα annotations τόνου υπόκεινται σε ανάλυση 3 "κάδων" ανά ημιτόνιο, και πραγματοποιείται η ίδια επαύξηση δεδομένων και ο HCQT μετασχηματισμός, ο οποίος εισήχθη εννοιολογικά από το πρώτο μοντέλο. Οι δύο νέοι τύποι annotations,

Y_n και Y_o, που εισάγονται, αναλύονται σε "κάδους" του ενός ημιτονίου. Έτσι, πραγματοποιώντας τον ίδιο τεμαχισμό των αρχικών αρχείων ήχου σε τμήματα 2 δευτερολέπτων, προκύπτουν οι παρακάτω διαστάσεις για τους 3 διαφορετικούς τύπους annotations (Πίνακας 4.3)

Annotation	Συμβολισμός	Διάσταση
Τόνος	Y_p	(264, 173)
Νότα	Y_n	(88, 173)
Onset	Y_o	(88, 173)

Πίνακας 4.3: *Διαστάσεις annotations για το Basic Pitch μοντέλο*

Η διάσταση του χρόνου τεμαχίζεται σε 173 "κάδους" ανά 2 δευτερόλεπτα, για όλα τα annotations. Κατά τη διέλευση των δεδομένων εντός του μοντέλου, εφαρμόζεται, στα συνελικτικά στρώματα, κατάλληλο γέμισμα, ώστε να μην αλλάζουν οι διαστάσεις των δεδομένων, εξαιρώντας τις 2 περιπτώσεις όπου γίνεται επιτηδευμένα εφαρμόζοντας τον βηματισμό 1x3. Τέλος, εφαρμόζεται ο ίδιος τεμαχισμός σε σύνολα δεδομένων εκπαίδευσης, επαλήθευσης και ελέγχου (80:10:10), διαχωρίζοντας τα δεδομένα, πάντα, ώστε αν υπάρχει η μελωδία μια σύνθεσης σε ένα σύνολο, να υπάρχει και η αντίστοιχη συνοδεία.

4.2.3 Μετα-επεξεργασία

Για κάθε αρχείο ήχου το οποίο εισέρχεται εντός του δικτύου, πραγματοποιούνται 3 προβλέψεις, όπως έχει σημειωθεί. Η πρώτη είναι αυτή του Y_p, και βάσει αυτής υπολογίστηκε η πιστότητα τόνου, ακολουθώντας την ίδια διαδικασία με το Deep Salience. Όπως θα παρουσιαστεί στην ενότητα των Αποτελεσμάτων, εξερευνήθηκε ένα σύνολο τιμών κατωφλίου, πάνω από τις οποίες τα στοιχεία της εξόδου τίθενται ίσα με 1, αλλιώς 0. Από αυτές τις τιμές, επιλέχθηκε εκείνη που μεγιστοποιεί τη μετρική της πιστότητας τόνου.

Η διαφοροποίηση του Basic Pitch, όσο αφορά το στάδιο της μετα-επεξεργασίας, προκύπτει στη χρήση των προβλέψεων για τα τις νότες (Y_n) και onsets (Y_o). Για τη χρήση αυτών, επιστρατεύεται ένας αλγόριθμος εμπνευσμένος από το Onsets and Frames [10], ο οποίος εισάγει 4 διαφορετικές παραμέτρους προς εξερεύνηση. Αυτές είναι οι:

1. Onset Threshold ή Κατώφλι Έναρξης Νότας
2. Note Threshold ή Κατώφλι Νότας
3. Minimum Note Length ή Ελάχιστη Διάρκεια Νότας
4. Tolerance ή Ανοχή

Αρχικά, διατηρούνται στην έξοδο Y_o τα στοιχεία με τιμή ανώτερη του ορίου έναρξης νότας, που έχει τεθεί ως υπερπαράμετρος. Αυτά τα στοιχεία, και οι συντεταγμένες τους $(f_i, t_i)_o$, θεωρούνται σημεία έναρξης νότας, ενώ τα υπόλοιπα απορρίπτονται. Για κάθε ένα από τα onsets και κατά φθίνουσα σειρά στιγμών έναρξης t_i, δημιουργείται η κάθε νότα, θεωρώντας τα σημεία του Y_n μέρος της νότας εφόσον έχουν τιμή μεγαλύτερη του κατωφλίου νότας, $t > t_i$ και μέχρι η τιμή να μειωθεί κάτω του ορίου αυτού για περισσότερους "κάδους" από αυτούς

που ορίζονται από την ανοχή. Εννοιολογικά, τα στοιχεία των εξόδων μπορούν να θεωρηθούν πιθανότητες ύπαρξης ή μη της συχνότητας που αντιστοιχεί στον "κάδο" τους, αιτιολογώντας έτσι την ύπαρξη κατωφλίων, όπως υπήρχαν και στο Deep Salience μοντέλο.

Εικόνα 4.7: *Απώλεια εκπαίδευσης (κόκκινο) και επικύρωσης (μπλέ) για την εκπαίδευση του Basic Pitch μοντέλου*

Κάθε φορά που δημιουργείται μια νότα, τα στοιχεία που της αντιστοιχούν τίθενται ίσα με 0. Όταν έχουν χρησιμοποιηθεί όλα τα onsets, ελέγχεται πάλι η έξοδος Y_n για στοιχεία με τιμή μεγαλύτερη του κατωφλίου νότας. Θεωρώντας τα σημεία αυτά ως onsets, επαναλαμβάνεται η ίδια διαδικασία δημιουργίας νοτών, με μόνη διαφοροποίηση τον έλεγχο για νότες τόσο για χρονικές στιγμές μεγαλύτερες όσο και για μικρότερες από του κάθε onset. Τέλος, απορρίπτονται όσες νότες έχουν συνολική διάρκεια μικρότερη από την ελάχιστη διάρκεια νότας που έχει τεθεί. Οι 4 παράμετροι, γύρω από τις οποίες χτίζεται η τελική απεικόνιση των νοτών, τίθενται κάθε φορά ως υπερπαράμετροι, και χρίζουν εξερεύνησης. Η εξερεύνηση αυτών, όσο και η εκπαίδευση και αξιολόγηση του μοντέλου, αποτελούν αντικείμενο της επόμενης ενότητας.

4.2.4 Αποτελέσματα

4.2.4.1 Εκπαίδευση Μοντέλου

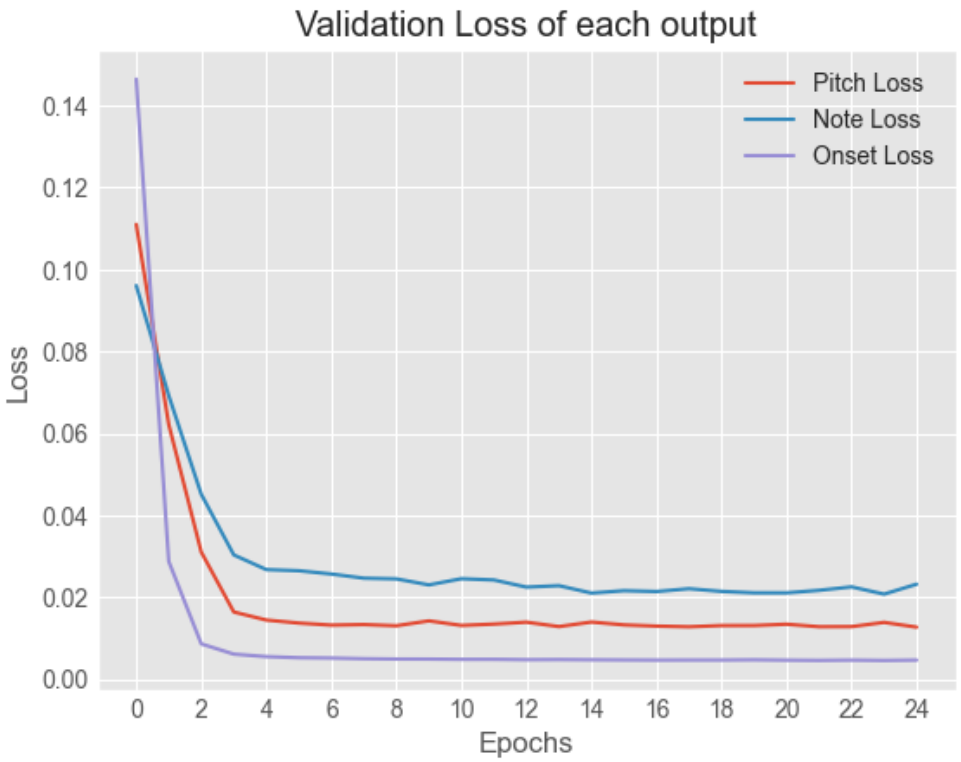
Για την εκπαίδευση του μοντέλου, χρησιμοποιήθηκαν, όπου ήταν δυνατό, οι ίδιες παράμετροι με αυτές που χρησιμοποιήθηκαν στην εκπαίδευση του Deep Salience μοντέλου. Μοναδική διαφορά, λόγω της ύπαρξης 3 διαφορετικών εξόδων, αποτελεί η επιλογή της συνάρτησης απώλειας, η οποία, όπως περιγράφηκε παραπάνω, επιλέχθηκε ως το άθροισμα των

3 επιμέρους απωλειών διασταυρούμενης επικύρωσης, οι οποίες αντιστοιχούν στις 3 εξόδους του μοντέλου. Οι παράμετροι της εκπαίδευσης παρουσιάζονται στον Πίνακα 4.4 παρακάτω:

Υπερπαράμετρος	Τιμή
Μέγεθος Δέσμης	64
Ρυθμός Μάθησης	10^{-3}
Εποχές	20
Βελτιστοποίηση	Adam [17]
Συνάρτηση Απώλειας	Άθροισμα τριών απωλειών διασταυρούμενης εντροπίας

Πίνακας 4.4: *Υπερπαράμετροι του μοντέλου Basic Pitch*

Παρατηρώντας την Εικόνα 4.7, φαίνεται ότι η απώλεια επικύρωσης μένει σταθερά μεγαλύτερη της αντίστοιχης της εκπαίδευσης μετά την δέκατη εποχή, ενώ στην αρχή της εκπαίδευσης κυμαίνεται γύρω από αυτό. Το συγκεκριμένο διάγραμμα, παρόλα αυτά, περιγράφει το συνολικό άθροισμα των επιμέρους απωλειών, και πιθανώς δεν παρέχει την συνολική εικόνα για τον τρόπο με τον οποίο εκπαιδεύεται το μοντέλο.



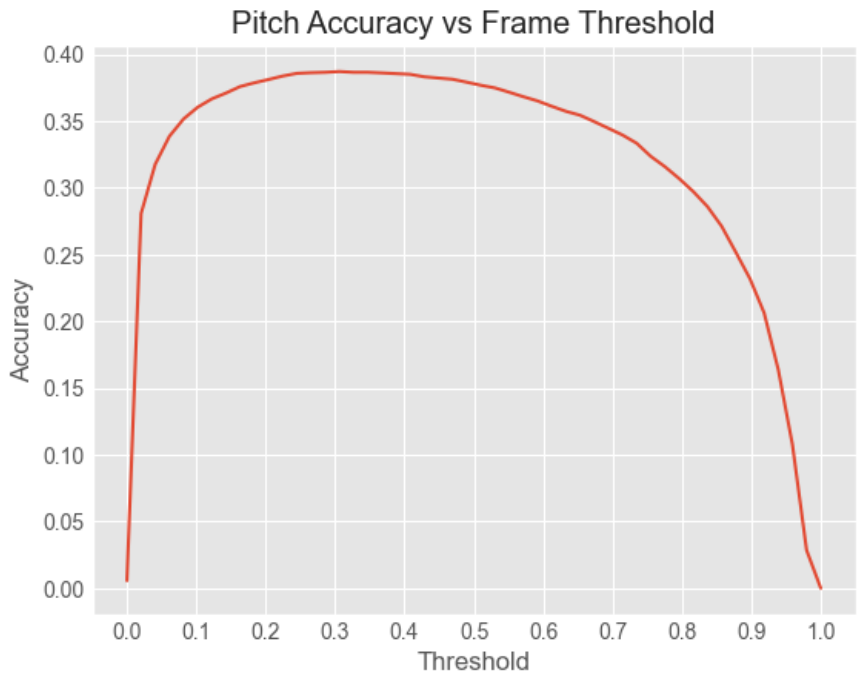
Εικόνα 4.8: *Pitch (κόκκινο), Note (μπλέ) και Onset (μώβ) απώλειες για το σετ επικύρωσης*

Για την περαιτέρω μελέτη, του τρόπου βελτιστοποίησης των επιμέρους απωλειών, παρέχεται η Εικόνα 4.8. Μελετώντας την, παρατηρείται ότι η απώλεια που μειώνεται ταχύτερα από τις 3, είναι αυτή που σχετίζεται με την απεικόνιση Y_o, δηλαδή τα onsets, ή στιγμές έναρξης των νοτών. Αυτό είναι αναμενόμενο, αφού τα συγκεκριμένα annotations, περιέχουν τα λιγότερα μη-μηδενικά στοιχεία, σε σχέση με τις απεικονίσεις Y_p, Y_n. Το ενδιαφέρον σημείο, είναι η σταθερή παραμονή της απώλειας τόνου σε χαμηλότερα επίπεδα από την απώλεια νότας, ενώ δεν διαφέρουν σημαντικά στον αριθμό των μη-μηδενικών στοιχείων που περι-

έχουν. Διαισθητικά, μπορεί να διατυπωθεί ο ισχυρισμός ότι τα 3 διαφορετικές εργασίες τις οποίες αναλαμβάνει τα επιλύσει το μοντέλο, είναι κατά σειρά αύξουσας δυσκολίας, η Onset, ακολουθούμενη από τον τόνο και τέλος τη νότα. Μετά την μελέτη της εκπαίδευσης και του τρόπου εκμάθησης, του δικτύου, μένει να επικυρωθεί το μοντέλο ώστε να επιλεχθούν οι παράμετροι που βελτιστοποιούν τις επιλεγμένες μετρικές.

4.2.4.2 Επικύρωση Μοντέλου

Στην περίπτωση του Deep Salience μοντέλου, η μόνη παράμετρος η οποία χρειαζόταν επιλογή, μετά την εκπαίδευση, ήταν το κατώφλι πλαισίου, το οποίο σχετιζόταν με τη μετρική της πιστότητας τόνου. Ξεκινώντας από το αντίστοιχο σημείο, για το παρόν μοντέλο, παρουσιάζεται η συγκεκριμένη μετρική ακρίβειας, συναρτήσει του του κατωφλίου πλαισίου στην Εικόνα 4.9. Παρατηρώντας την εικόνα φαίνεται ότι το μέγιστο της πιστότητας τόνου στο σύνολο επικύρωσης είναι ίσο με **38.71%**, και εντοπίζεται στην τιμή κατωφλίου **0.3061**. Η αναζήτηση κατωφλίου για το μέγιστο της πιστότητας, πραγματοποιήθηκε στο εύρος 0 έως 1, σε πλέγμα 50 ισαπέχουσων τιμών. Μια ενδιαφέρουσα παρατήρηση αποτελεί το γεγονός ότι η μετρική εμφανίζει τιμές κοντά στη βέλτιστη, για ένα μεγάλο εύρος ορίων γύρω από την τιμή του μέγιστου, κάτι που δεν συνέβη στην περίπτωση Deep Salience.



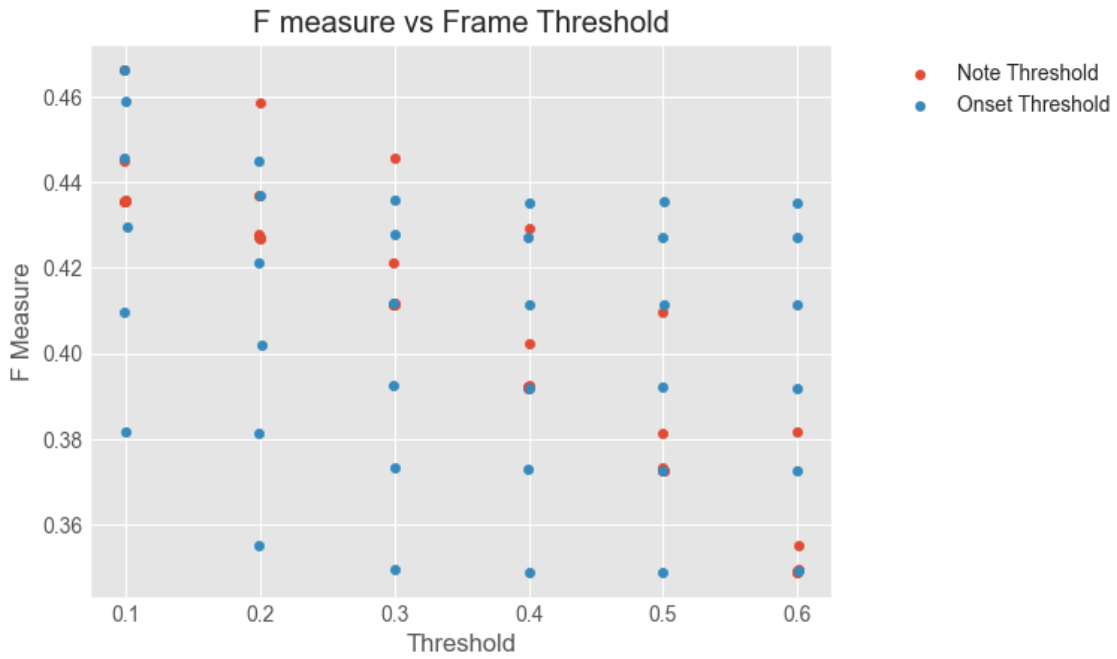
Εικόνα 4.9: *Πιστότητα τόνου μοντέλου συναρτήσει του κατωφλίου*

Εκτός από τη μετρική πιστότητας τόνου, το συγκεκριμένο μοντέλο παρέχει τη δυνατότητα υπολογισμού των δύο μετρικών F, οι οποίες έχουν αναφερθεί. Αυτό μπορεί να πραγματοποιηθεί λόγω της ύπαρξης προβλέψεων για τις νότες και τα onsets από το μοντέλο, πάνω στα οποία στηρίζεται ο υπολογισμών των μετρικών F. Για τον υπολογισμό αυτών, χρειάζεται η επιλογή 4 παραμέτρων. Αυτές είναι οι όριο έναρξης νότας, όριο νότας, ελάχιστη διάρκεια νότας και ανοχή. Για την εύρεση της βέλτιστης τετράδας, των παραπάνω παραμέτρων, πραγματοποιήθηκε αναζήτηση πλέγματος στις τιμές (Πίνακας 4.5):

Υπερπαράμετρος	Τιμή
Όριο Έναρξης Νότας	[0.1,0.7], με βηματισμό = 0.1
Όριο Νότας	[0.1,0.7], με βηματισμό = 0.1
Ελάχιστη Διάρκεια Νότας	5
Ανοχή	11

Πίνακας 4.5: *Παράμετροι για την επικύρωση του μοντέλου Basic Pitch*

Οι δύο από τις παραπάνω παραμέτρους με τη λιγότερη επιρροή στην απόδοση του μοντέλου, είναι η ανοχή και η ελάχιστη διάρκεια νότας. Τα δύο κατώφλια είναι αυτά τα οποία, κατά κύριο λόγο, διαμορφώνουν την απόδοση του μοντέλου, όσον αφορά τις μετρικές F. Χρησιμοποιώντας τη μετρική F, αφού η Fo είναι μια πιο ελαστική μορφή της, παρατηρείται η κατανομή των τιμών της στην Εικόνα 4.10. Όσο αφορά την ίδια τη μετρική, φαίνεται να μειώνεται με την αύξηση του ορίου νότας, καθώς και, σε μικρότερο βαθμό, και με την αύξηση του ορίου έναρξης νότας. Πραγματοποιώντας έλεγχο πλέγματος, λοιπόν, η τετράδα (όριο έναρξης νότας, όριο νότας, ελάχιστη διάρκεια νότας, ανοχή) με τη βέλτιστη απόδοση είναι η **(0.1, 0.1, 5, 11)**, με αποτελέσματα μετρικών **F = 46.64%** και **Fo = 65.79%**.



Εικόνα 4.10: *Μετρική F μοντέλου Basic Pitch συναρτήσει των κατωφλιών νότας και onset*

4.2.4.3 Αξιολόγηση Μοντέλου

Μετά την επικύρωση του μοντέλου και την επιλογή των κατάλληλων παραμέτρων και για τις 3 μετρικές, το μόνο που μένει είναι η αξιολόγησή του στο σύνολο ελέγχου. Επιπλέον, υπολογίστηκε και η πιστότητα νότας (με παρόμοιο τρόπο με την πιστότητα τόνου), αλλά στην πρόβλεψη για την Y_n αναπαράσταση αντί της Y_p. Για αυτό τον σκοπό χρησιμοποιήθηκαν, όπως και στο Deep Salience, όλοι οι δυνατοί τύποι συνόλου ελέγχου, οι οποίοι παρέχονται στο Guitarset. Όπως φαίνεται στον Πίνακα 4.6, οι μετρικές δεν εμφανίζουν ιδιαίτερη δια-

κύμανση μεταξύ των διαφορετικών συνόλων ελέγχου, και αντανακλούν, σε ικανοποιητικό βαθμό, την απόδοση στο σύνολο επικύρωσης.

Σύνολο Ελέγχου	Πιστότητα τόνου	Πιστότητα νότας	Μετρική F	Μετρική Fo
Εξαφωνικό	37.96%	38.24%	47.10%	66.52%
Εξαφωνικό Debleeded	37.94%	38.54%	48.51%	67.09%
Εξαφωνικό και Εξαφωνικό Debledeed	37.95%	38.39%	47.80%	66.80%
Μονοκάναλο	36.57%	36.16%	44.36%	61.44%
Μείξη μονοκάναλου και εξαφωνικού	37.72%	38.62%	48.67%	67.46%

Πίνακας 4.6: *Σύγκριση επίδοσης της πιστότητας τόνου και νότας και των μετρικών F και Fo σε διαφορετικά σύνολα ελέγχου για το Basic Pitch μοντέλο*

Σε αυτό το σημείο αξίζει να γίνει μια σύγκριση της απόδοσης του παρόντος μοντέλου, με την απόδοση του αντίστοιχου μοντέλου το οποίο προτάθηκε στην αρχική δημοσίευση [3]. Όσο αφορά τις μετρικές F, οι οποίες αποτελούν και πιο αντιπροσωπευτικές με παρτιτούρα μετρικές, η απόδοση του μοντέλου της δημοσίευσης ήταν **F = 56%** και **Fo = 79%**. Η βελτιωμένη απόδοση της αρχικής δημοσίευσης, μπορεί να αποδοθεί, κυρίως, στη χρήση πολλών διαφορετικών συνόλων δεδομένων πέραν του Guitarset. Συγκεκριμένα, και μη λαμβάνοντας υπόψη τις διαφορετικές διάρκειες των ηχογραφήσεων κάθε συνόλου δεδομένων, χρησιμοποιήθηκε περίπου το 10% του συνολικού όγκου δεδομένων για την εκπαίδευση του μοντέλου της συγκεκριμένης διπλωματικής.

4.3 Συνελικτικά LSTM

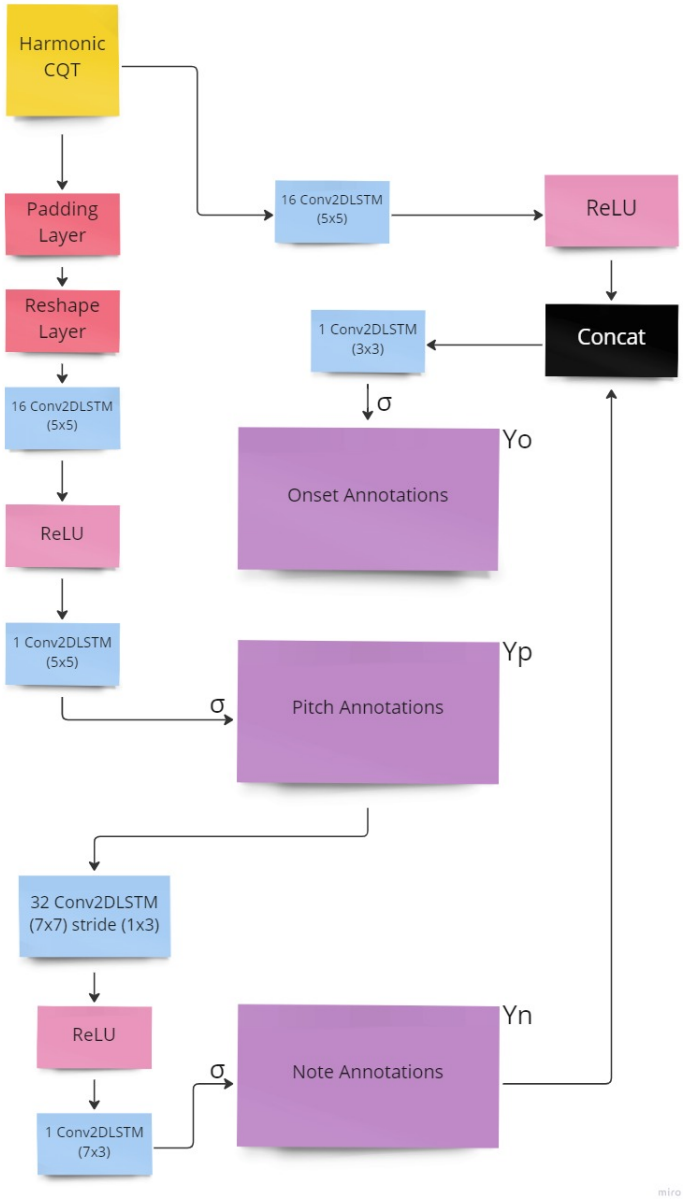
4.3.1 Περιγραφή Μοντέλου

Αφού εξετάστηκαν τα δύο προηγούμενα βιβλιογραφικά μοντέλα, συμπεριλαμβανομένης της τρέχουσας στάθμης της τεχνικής, κρίθηκε απαραίτητος ο πειραματισμός και η επέκταση του Basic Pitch μοντέλου. Κυριότερο στόχο, δεν αποτελεί, κατά ανάγκη, η βελτίωση της στάθμης της τεχνικής, αλλά η διατύπωση πρωτότυπων ιδεών και μεθόδων.

Η πρώτη, εξ αυτών, είναι η χρήση κάποιας μορφής αναδρομικού μοντέλου, όσον αφορά τον χρόνο. Κάτι τέτοιο, βασίζεται στην χρονική συσχέτιση των διαφορετικών νοτών ενός δεδομένου έργου, αφού, συνήθως, θα βασίζονται όλες στην ίδια μουσική κλίμακα. Επιπλέον, οι νότες που επιλέχθηκαν να εκτελεστούν μια δεδομένη χρονική στιγμή, επηρεάζουν τις επερχόμενες, χρονικά, νότες βάσει των προτιμήσεων και του μουσικού ο οποίος τις εκτελεί.

Αφού επικυρώθηκε η υψηλή απόδοση του μοντέλου της στάθμης της τεχνικής, όπως φαίνεται στην προηγούμενη ενότητα, υλοποιήθηκε μια αναδρομική παραλλαγή του Basic Pitch μοντέλου, η οποία παρουσιάζεται στην Εικόνα 4.11. Ο επαναλαμβανόμενος χαρακτήρας, προσδίδεται στο Basic Pitch μέσω χρήσης ConvLSTM [4] στρωμάτων δύο διαστάσεων. Η υλοποίηση αυτή εμπεριέχει CNN στρώματα εντός της δομής ενός LSTM, όπως περιγράφηκε λεπτομερώς στα εισαγωγικά Κεφάλαια.

Παρόλα αυτά, για να υφίσταται ένα επαναλαμβανόμενο μοντέλο, χρειάζεται και τα δεδομένα εισόδου να είναι δομημένα, διαστατικά, ως $(n_{timestep}, f, t)$, αντί (f, t) όπως προκύπτουν



Εικόνα 4.11: *Αρχιτεκτονική Basic Pitch ConvLSTM Μοντέλου*

μετά τον μετασχηματισμό HCQT. Η νέα διάσταση $n_{timestep}$ ουσιαστικά προκύπτει χωρίζοντας τα δεδομένα σε ίσα, στη χρονική διάσταση, τμήματα, και στοιβάζοντάς τα διαδοχικά.

Εφόσον και σε αυτήν την περίπτωση, χωρίζεται κάθε αρχικό αρχείο ήχου σε τμήματα 2 δευτερολέπτων, όπως θα περιγραφεί στην ενότητα της προ-επεξεργασίας, θα είχε σε πρώτη προσέγγιση νόημα, να στοιβαχτούν όλα τα τμήματα αυτά για κάθε αρχικό αρχείο, οπότε να οριστεί αναλόγως το κάθε δείγμα εισόδου. Το πρόβλημα με αυτήν την προσέγγιση έγκειται στο γεγονός ότι το μοντέλο γίνεται χρονικά μη αποδοτικό, αφού ο χρόνος εκπαίδευσης είναι πολύ αυξημένος. Αντί αυτού, το κάθε τμήμα 2 δευτερολέπτων, αφού εισέλθει στο δίκτυο, χωρίζεται ξανά, σε επιλεγμένο αριθμό χρονικών βημάτων, ο οποίος έχει δοθεί ως παράμετρος στον ορισμό του δικτύου.

Κάθε ένα από τα δείγματα εισόδου, 2 δευτερολέπτων, μετατρέπεται, μετά τον μετασχη-

ματισμό HCQT, σε πίνακα διάστασης 173 στον χρονικό άξονα. Ο αριθμός 173, δεν είναι ιδιαίτερα βολικός ώστε να δημιουργηθούν ίσα τμήματα, οπότε πραγματοποιείται "γέμισμα" εντός του δικτύου πριν μετασχηματιστεί σε επαναλαμβανόμενο δείγμα μέσω στρώματος μετασχηματισμού. Στη συνέχεια, τα δεδομένα περνούν από ένα δισδιάστατο ConvLSTM στρώμα βάθους 16, με πυρήνα 5x5, προτού περάσουν από συνάρτηση ενεργοποίησης ReLU. Τέλος, για την έξοδο του Y_p, τα δεδομένα περνούν από ένα συνελικτικό LSTM στρώμα βάθους 1 και πυρήνα 5x5, ακολουθούμενο από σιγμοειδή συνάρτηση ενεργοποίηση.

Μετά την πρόβλεψη σχετικά με την αναπαράσταση Y_p, η έξοδος περνά από δισδιάστατο ConvLSTM στρώμα βάθους 32, πυρήνα 7x7, με βηματισμό 1x3. Η επιλογή του συγκεκριμένου βηματισμού, όπως και στο Basic Pitch μοντέλο, πραγματοποιείται ώστε να αντανακλάται το γεγονός ότι η αναπαράσταση Y_p, έχει τριπλάσιο αριθμό "κάδων" από τις υπόλοιπες. Στη συνέχεια, υπάρχει συνάρτηση ενεργοποίησης ReLU, και στρώμα βάθους 1 και πυρήνας 7x3. Αφού τα δεδομένα διέλθουν από σιγμοειδή συνάρτηση ενεργοποίησης, προκύπτει η πρόβλεψη Y_n, για τα annotations νότας.

Στο τελευταίο κλάδο του δικτύου, τα αρχικά δεδομένα εισόδου περνούν από συνελικτικό στρώμα LSTM βάθους 16 και πυρήνα 5x5, προτού περάσουν από συνάρτηση ενεργοποίησης ReLU. Στη συνέχεια, στοιβάζονται με την πρόβλεψη για το Y_n, που έχει ήδη εξαχθεί, και αφού περάσουν από δισδιάστατο στρώμα ConvLSTM βάθους 1, πυρήνα 3x3 και σιγμοειδούς συνάρτησης ενεργοποίησης, προκύπτει η πρόβλεψη του μοντέλου για το Y_o.

Τέλος, για την εκπαίδευση του δικτύου, χρησιμοποιείται το άθροισμα των τριών επιμερους απωλειών διασταυρούμενης εντροπίας, οι οποίες αντιστοιχούν στις εξόδους για τόνο, νότα και onsets αντίστοιχα, όπως είχαν περιγραφεί για το Basic Pitch μοντέλο στην Εξίσωση 4.1.

4.3.2 Προ-επεξεργασία

Στο στάδιο της προ-επεξεργασίας πραγματοποιούνται πολλά από τα βήματα της περίπτωσης του Basic Pitch μοντέλου, συμπεριλαμβανομένων της ίδιας επαύξησης δεδομένων και μετασχηματισμού HCQT, καθώς και του διαχωρισμού σε "κάδους" που πραγματοποιείται. Πρόσθετα βήματα αποτελούν τα τμήματα γεμίσματος και αλλαγής σχήματος, τα οποία μετασχηματίζουν τα δεδομένα σε μορφή κατάλληλη για αναδρομικά μοντέλα. Στη συγκεκριμένη περίπτωση, πρέπει ο αριθμός των χρονικών να διαιρεί ακριβώς τον αριθμό 180, δηλαδή τη διάσταση του χρόνου μετά το γέμισμα. Από όλους τους αριθμούς που πληρούν το παραπάνω κριτήριο, η βέλτιστη τιμή, τουλάχιστον για το Guitarset, βρέθηκε να είναι $n_{timestep} = 6$. Στη γενική περίπτωση, τα annotations έχουν τις παρακάτω διαστάσεις (Πίνακας 4.7)

Annotation	Συμβολισμός	Διάσταση
Τόνος	Y_p	$(n_{timestep}, 264, \frac{180}{n_{timestep}})$
Νότα	Y_n	$(n_{timestep}, 88, \frac{180}{n_{timestep}})$
Onset	Y_o	$(n_{timestep}, 88, \frac{180}{n_{timestep}})$

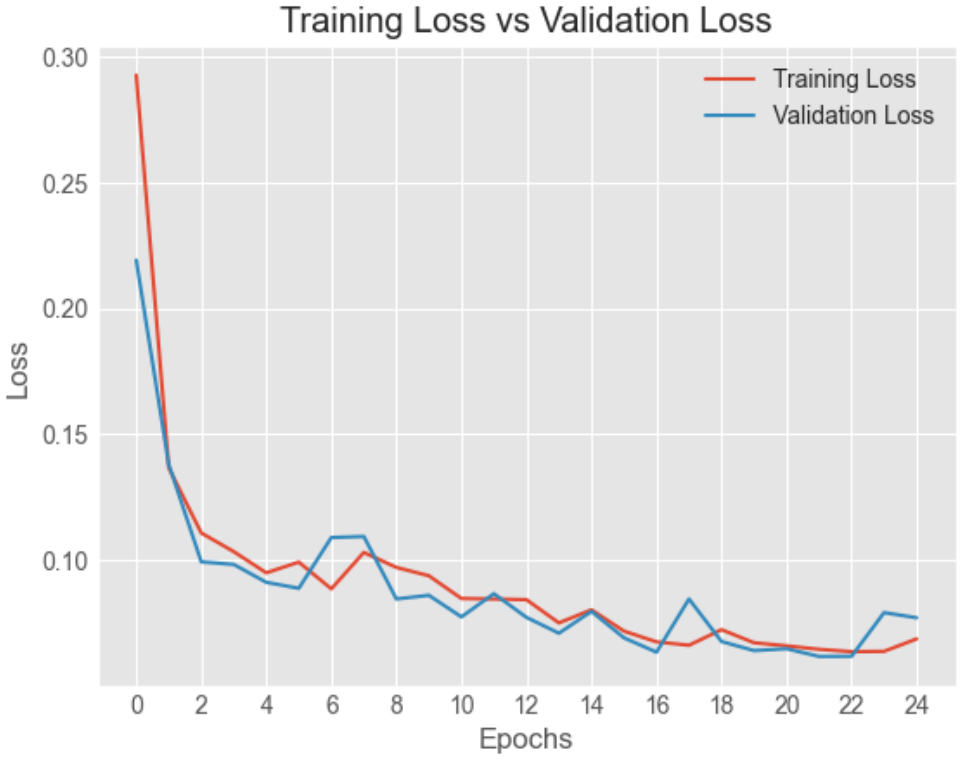
Πίνακας 4.7: *Διαστάσεις annotations για το Basic Pitch ConvLSTM μοντέλο*

Ο χρονικός τεμαχισμός, γίνεται με 173 "κάδους" ανά 2 δευτερόλεπτα για όλα τα annotations. Κατά τη διέλευση των δεδομένων εντός του μοντέλου, εκτός από το αρχικό γέμισμα,

εφαρμόζεται, στα δισδιάστατα ConvLSTM στρώματα, κατάλληλο γέμισμα, ώστε να μην αλλάζουν οι διαστάσεις των δεδομένων, εξαιρώντας τις 2 περιπτώσεις όπου γίνεται επιτηδευμένα εφαρμόζοντας τον βηματισμό 1x3, ώστε να να αντανακλάται ο υποτριπλασιασμός των συχνοτικών "κάδων". Τέλος, εφαρμόζεται ο ίδιος διαχωρισμός σε σύνολα δεδομένων εκπαίδευσης, επικύρωσης και ελέγχου, με αναλογία 80:10:10, διαχωρίζοντας τα δεδομένα πάντα με τέτοιον τρόπο, ώστε αν υπάρχει το κομμάτι μελωδίας μια σύνθεσης σε ένα σύνολο, να υπάρχει και η αντίστοιχη εκτέλεση συνοδείας.

4.3.3 Μέτα-επεξεργασία

Στην περίπτωση του παρόντος μοντέλου, χρησιμοποιείται, βασικά, ο ίδιος αλγόριθμος μετα-επεξεργασίας, με αυτόν που χρησιμοποιήθηκε στο Basic Pitch, και βασίζεται στον αλγόριθμο από το Onsets and Frames [10], ο οποίος εισάγει τις 4 παραμέτρους που έχουν ήδη εξηγηθεί (όριο έναρξης νότας, όριο νότας, ελάχιστη διάρκεια νότας και ανοχή).



Εικόνα 4.12: *Απώλεια εκπαίδευσης (κόκκινη) και επικύρωσης (μπλε) για το Basic Pitch ConvLSTM μοντέλο*

Παρόλα αυτά, τόσο οι έξοδοι όσο και οι ετικέτες annotations του δικτύου, είναι της μορφής ($n_{timestep}$ = 6, 88 ή 264, 30), με την κύρια διαφοροποίηση να βρίσκεται στον αριθμό των συχνοτικών "κάδων", ο οποίος είναι τριπλάσιος στην Y_p αναπαράσταση. Για την εκμετάλλευση του προαναφερθέντα αλγόριθμου, η κατάλληλη μορφή των απεικονίσεων, καθώς και των προβλέψεων, θα έπρεπε να είναι (88 ή 264, 173). Για την εκπλήρωση του παραπάνω κριτηρίου, πραγματοποιήθηκε, πριν την εκτέλεση του αλγορίθμου, διαδικασία αναστροφής των γεμισμάτων και των διαδικασιών αλλαγής σχήματος των δεδομένων που πραγματοποιήθηκαν αρχικά. Έτσι, οι προβλέψεις και τα annotations μετασχηματίζονται ως

$(n_{timestep} = 6, f_{bins}, 30) \rightarrow (f_{bins}, 180)$, και στη συνέχεια αφαιρούνται οι τελευταίοι 7 χρονικά "κάδοι", οι οποίοι εξάλλου αποτελούσαν το γέμισμα που προστέθηκε, ώστε να μετασχηματιστούν ως $(f_{bins}, 180) \rightarrow (f_{bins}, 173)$. Κατά αυτόν τον τρόπο, έχει γίνει πρόβλεψη μετατρέποντας τα δεδομένα σε επαναληπτικά, και τελικά μέσω της μέτα-επεξεργασίας, μετατρέπονται πάλι στην αρχική τους μορφή.

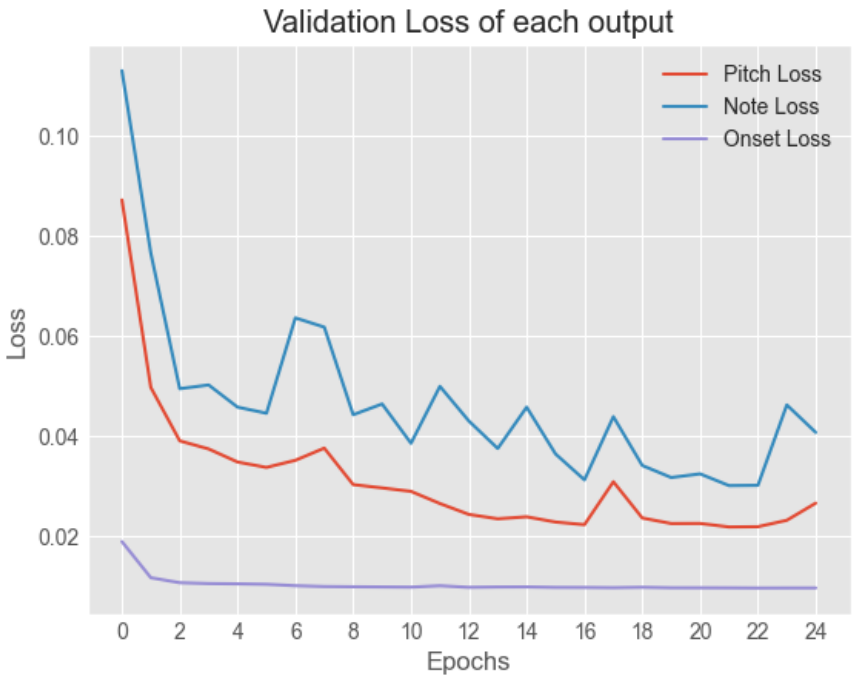
4.3.4 Αποτελέσματα

4.3.4.1 Εκπαίδευση Μοντέλου

Το συγκεκριμένο μοντέλο, αποτελώντας παραλλαγή του Basic Pitch, χρησιμοποιεί ταυτόσημες παραμέτρους εκπαίδευσης, με διαφορά τον αριθμό των χρονικών βημάτων, παράμετρος η οποία δεν υπάρχει στο αρχικό Basic Pitch μοντέλο, αφού δεν είναι αναδρομικό. Οι παράμετροι αυτοί είναι οι παρακάτω (Πίνακας 4.8):

Υπερπαράμετρος	Τιμή
Μέγεθος Δέσμης	64
$n_{timestep}$	6
Ρυθμός Μάθησης	10^{-3}
Εποχές	20
Βελτιστοποίηση	Adam [17]
Συνάρτηση Απώλειας	Άθροισμα τριών απωλειών διασταυρούμενης εντροπίας

Πίνακας 4.8: *Υπερπαράμετροι του Basic Pitch ConvLSTM μοντέλου*



Εικόνα 4.13: *Απώλεια τόνου (κόκκινο), νότας (μπλέ) και onset (μώβ) για την εκπαίδευση του Basic Pitch ConvLSTM μοντέλου*

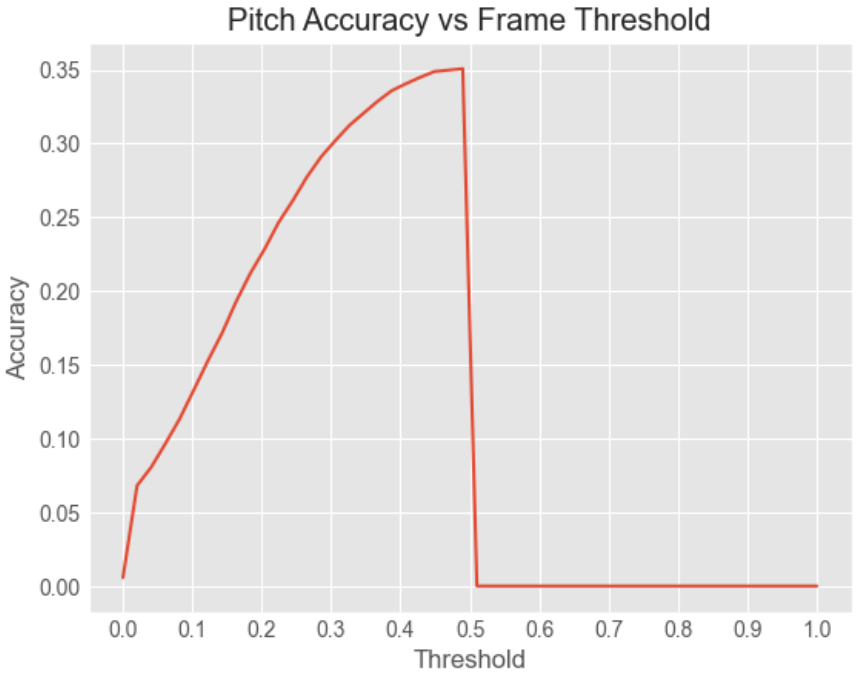
Κατά την εκπαίδευση, τόσο η απώλεια εκπαίδευσης, όσο και αυτή της επικύρωσης, πα-

ραμένουν σε παρόμοια επίπεδα, και εναλλάσσονται μεταξύ τους με το πέρασμα των εποχών, όπως φαίνεται στην Εικόνα 4.12. Τελικά επιστρέφονται τα βάρη του μοντέλου από την εποχή με το βέλτιστο σφάλμα επικύρωσης, χρησιμοποιώντας πρόωρο τερματισμό.

Το διάγραμμα αυτό, όπως και στην περίπτωση του Basic Pitch, περιγράφει τη συνολική απώλεια, η οποία είναι το άθροισμα των 3 επιμέρους που αντιστοιχούν στα Y_p, Y_n & Y_o, και παρουσιάζονται αναλυτικότερα στην Εικόνα 4.13.Εδώ φαίνεται να υπάρχει αντίστοιχη συμπεριφορά των επιμέρους απωλειών, με αυτή του Basic Pitch, με το απώλεια νότας να κυμαίνεται σε υψηλότερα επίπεδα από την απώλεια τόνου, η οποία με τη σειρά της κυμαίνεται σε υψηλότερα επίπεδα από την απώλεια onset. Μετά την εκπαίδευση, μένει να πραγματοποιηθεί η επικύρωση του μοντέλου, για την επιλογή των παραμέτρων του αλγορίθμου μετα-επεξεργασίας.

4.3.4.2 Επικύρωση Μοντέλου

Έχοντας πραγματοποιήσει την εκπαίδευση του δικτύου, χρειάζεται να πραγματοποιηθεί επικύρωση, ώστε να βρεθεί, αρχικά, το κατώφλι που θα χρησιμοποιηθεί για τον υπολογισμό της FLA. Για τον υπολογισμό της μετρικής αυτής, απλώς χρησιμοποιείται μια τιμή κατωφλίου, βάσει της οποίας ορίζονται 0 ή 1 τα στοιχεία της πρόβλεψης αν έχουν χαμηλότερη ή υψηλότερη τιμή από αυτό, αντίστοιχα.



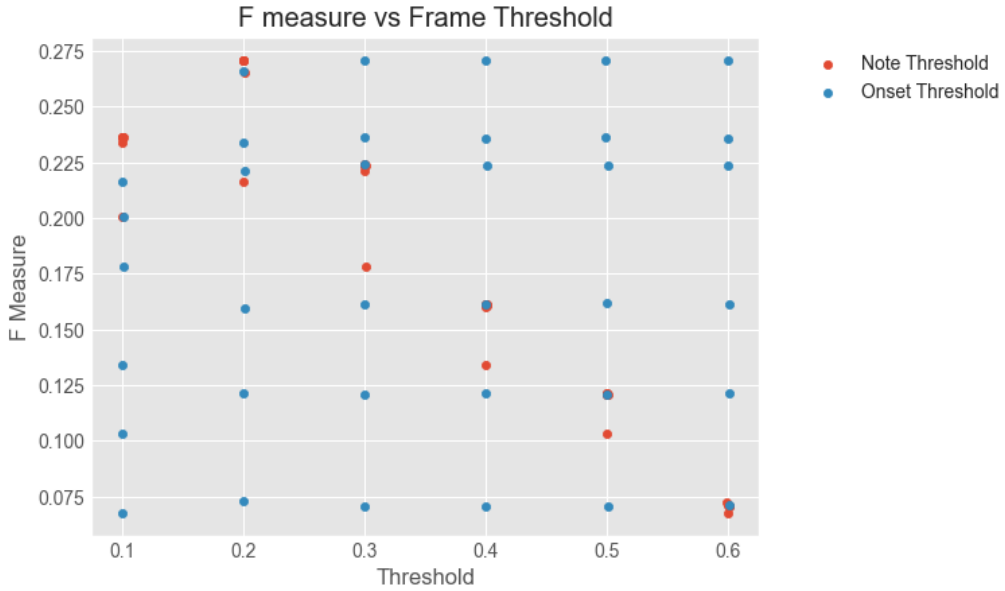
Εικόνα 4.14: *FLA συναρτήσει της τιμή κατωφλίου πλαισίου κατά της επικύρωση του Basic Pitch ConvLSTM μοντέλου*

Το πιο αξιοσημείωτο στην Εικόνα 4.14, είναι η απότομη μείωση, σχεδόν στο 0, μετά την τιμή ορίου 0.5. Κάτι τέτοιο, θα μπορούσε να οφείλεται στην απουσία στρωμάτων κανονικοποίησης δέσμης εντός του δικτύου. Η αναζήτηση του βέλτιστου κατωφλίου, πραγματοποιήθηκε σε πλέγμα 50 ισαπεχουσών τιμών από το 0 έως το 1. Ακολουθώντας τη διαδικασία αυτή, επιλέχθηκε όριο ίσο με **0.4898**, με τιμή πιστότητας τόνο ίση με **35.08%**. Φαίνεται, λοιπόν,

ότι παρά το απότομο όριο στο διάγραμμα, σε σχέση με τα προηγούμενα μοντέλα, η ακρίβεια κυμαίνεται στα ίδια επίπεδα.

Στο επόμενο στάδιο, πραγματοποιείται διαδικασία επικύρωσης του μοντέλου, βάσει των μετρικών F, οι οποίες υπολογίζονται σε επίπεδο νοτών. Οι παράμετροι οι οποίες εξερευνήθηκαν, είναι το όριο έναρξης νότας, το όριο νότας, η ελάχιστη διάρκεια νότας και η ανοχή, σχηματίζοντας ένα 4-διάστατο πλέγμα αναζήτησης παρόμοιο με του Πίνακα 4.5. Παρατηρήθηκε, παρόλα αυτά, ότι τα 2 κατώφλια είναι οι κύριοι παράγοντες διαφοροποίησης των μετρικών, επομένως η εξερεύνηση εστιάστηκε σε αυτά.

Στην Εικόνα 4.15 παρουσιάζεται η μετρική F, για διαφορετικές τιμές των 2 ξεχωριστών τιμών κατωφλίου. Για την επιλογή των βέλτιστων παραμέτρων, επιλέχθηκε, και πάλι, η μετρική F, αφού μέσω αυτής της διαδικασίας επιλέγεται και η βέλτιστη Fno. Ακολουθώντας την παραπάνω λογική οι βέλτιστες παράμετροι **(όριο έναρξης νότας, όριο νότας, ελάχιστη διάρκεια νότας, ανοχή)** προκύπτουν ίσες με **(0.3, 0.2, 11, 5)**, με βέλτιστες τιμές **F = 27.07%** και **Fno = 48.24%** για τις μετρικές. Παρατηρείται, όπως και στην περίπτωση του Basic Pitch, ότι η τιμή του κατωφλίου νότας έχει μεγαλύτερη συσχέτιση με την απόδοση του μοντέλου, από ότι έχει η τιμή του κατωφλίου onset.



Εικόνα 4.15: *Μετρικές F συναρτήσει των κατωφλίων νότας και onsets κατά της επικύρωση του Basic Pitch ConvLSTM μοντέλου*

4.3.4.3 Αξιολόγηση Μοντέλου

Μετά την εκπαίδευση και αξιολόγηση του μοντέλου, χρειάζεται να πραγματοποιηθεί ο έλεγχος της απόδοσής του, στο σύνολο ελέγχου, το οποίο αποτελεί το 10% του αρχικού συνόλου δεδομένων και έχει κρατηθεί για αυτόν τον σκοπό, χωρίς να έχει αλληλεπιδράσει με οποιονδήποτε τρόπο με το μοντέλο. Τα δείγματα εντός του συνόλου ελέγχου, παρέχονται σε 5 διαφορετικούς τύπους, όπως και στο προηγούμενο μοντέλο, ανάλογα με τον τρόπο ηχογράφησης.

Όπως φαίνεται στον Πίνακα 4.9, η βέλτιστη απόδοση παρουσιάζεται στο **Εξαφωνικό**

Σύνολο Ελέγχου	Πιστότητα τόνου	Πιστότητα νότας	Μετρική F	Μετρική Fno
Εξαφωνικό	34.37%	31.35%	28.58%	51.33%
Εξαφωνικό Debleeded	34.39%	31.24%	29.29%	51.85%
Εξαφωνικό και Εξαφωνικό Debledeed	34.38%	31.29%	28.94%	51.59%
Μονοκάναλο	33.43%	29.42%	25.63%	45.68%
Μείξη μονοκάναλου και εξαφωνικού	34.02%	30.56%	27.65%	49.93%

Πίνακας 4.9: *Σύγκριση επίδοσης της πιστότητας τόνου και νότας και των μετρικών F και Fno σε διαφορετικά σύνολα ελέγχου για το Basic Pitch ConvLSTM μοντέλο*

Debleeded σύνολο, όπου οι μετρικές F είναι ίσες με **F = 29.29%** και **Fno = 51.85%**. Επιπλέον, υπολογίστηκαν οι πιστότητες τόνου και νότας, οι οποίες εκφράζουν τον υπολογισμό της ακρίβειας σε επίπεδο χρονικού πλαισίου (FLA) για τις απεικονίσεις Y_p και Y_n αντίστοιχα. Στο παραπάνω σύνολο ισούνται με **Pitch Acc = 34.39%** και **Note Acc = 31.24%**, με την πρώτη να είναι η βέλτιστη από όλα τα σύνολα. Την καλύτερη απόδοση βάσει της ακρίβειας νοτών παρουσιάζει το Εξαφωνικό σύνολο, με **Note Acc = 31.35%**.

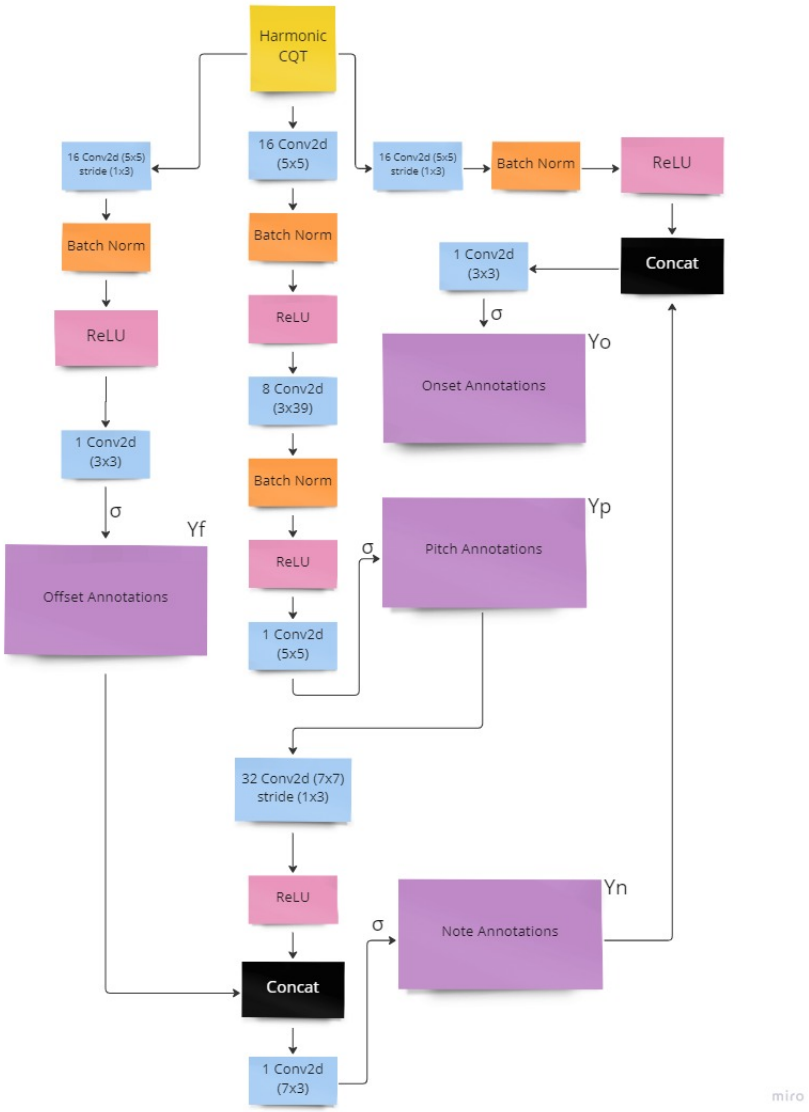
4.4 Basic Pitch with Offsets

4.4.1 Περιγραφή Μοντέλου

Αφού πραγματοποιήθηκε η εξερεύνηση των χρονικά επαναλαμβανόμενων μοντέλων, ήταν προφανές ότι, τουλάχιστον μέσω της συνελικτικής LSTM αρχιτεκτονικής, χρειαζόταν μια διαφορετική προσέγγιση ώστε να υπάρχουν συγκρίσιμα αποτελέσματα με το Basic Pitch μοντέλο. Η ιδέα για την συμπερίληψη των offsets στις προβλέψεις του μοντέλου προήλθε από το γεγονός ότι ήταν η μοναδική αναπαράσταση, η οποία παρότι μπορούσε να εξαχθεί το ίδιο εύκολα με τα onsets, δεν είχε συμπεριληφθεί στο αρχικό μοντέλο.

Αξίζει να σημειωθεί, ότι η αρχική δημοσίευση του Basic Pitch [3], επισημαίνει τη δυσκολία προσδιορισμού των offsets, αφού μπορούν να γίνουν ιδιαίτερα ασαφή τόσο λόγω των διάφορων εφέ των οργάνων και του χώρου, όσο λόγω του γεγονότος ότι ο μουσικός συνήθως θα τονίσει την έναρξη και όχι το τέλος μιας νότας. Παρόλα αυτά, ειδικά στην περίπτωση της παρτιτούρας, σημειώνονται τόσο οι νότες όσο και η διάρκειά τους. Επιπλέον υπάρχει ήδη μια μετρική, η F, η οποία λαμβάνει υπόψη τόσο τα onsets όσο και τα offsets, πάντα παρέχοντας μια χρονική ανοχή. Για τους παραπάνω λόγους, κρίθηκε ενδιαφέρον να εξερευνηθεί το ενδεχόμενο προσθήκης των offsets στο μοντέλο, για την πιθανή προσφορά τους στην πρόβλεψη των υπόλοιπων αναπαραστάσεων.

Όπως φαίνεται στην Εικόνα 4.16, η βασική διαφορά του τελευταίου μοντέλου, με το Basic Pitch, είναι ο νέος κλάδος που σχετίζεται με τα offsets, και το νέο τμήμα όπου στοιβάζεται η έξοδός του με τα δεδομένα ενός άλλου τμήματος του δικτύου, ώστε τελικά να εξαχθεί η πρόβλεψη για το Y_n. Συγκεκριμένα, για την κατασκευή του offset κλάδου, υπόκεινται τα δεδομένα, μετά τον μετασχηματισμό HCQT, σε συνελικτικό στρώμα βάθους 16, πυρήνα 5x5 και βηματισμού 1x3. Ο συγκεκριμένος βηματισμός, υποτριπλασιάζει τον αριθμό των συχνοτικών "κάδων", αφού η αναπαράσταση Y_f έχει υποστεί την ίδια προεπεξεργασία με τις



Εικόνα 4.16: *Αρχιτεκτονική Basic Pitch Offset Μοντέλου*

Y_o, Y_n, όπως θα αναλυθεί στην επόμενη ενότητα.

Συνεχίζοντας, εφαρμόζεται κανονικοποίηση δέσμης, καθώς και ενεργοποίηση ReLU, πριν εφαρμοστούν συνελίξεις πυρήνα 3x3, με βάθος 1. Τέλος, για να εξαχθεί η πρόβλεψη για το Y_f, εφαρμόζεται σιγμοειδής συνάρτηση ενεργοποίησης, ώστε να ερμηνευτούν τα στοιχεία κάθε "κάδου" ως πιθανότητες μεταξύ 0 και 1. Αφού παραχθεί η πρόβλεψη, διοχετεύεται στη δεύτερη περίπτωση του δικτύου όπου γίνεται στοίβαγμα, με την πρώτη να είναι αυτή που πραγματοποιείται με την έξοδο Y_p για την πρόβλεψη των onsets. Το σημείο όπου πραγματοποιείται, σε αυτήν την περίπτωση, είναι πριν το τελευταίο συνελικτικό στρώμα, το οποίο θα οδηγήσει τελικά στην πρόβλεψη των νοτών. Διαισθητικά, χρησιμοποιείται η πληροφορία του Y_p μαζί με αυτή του Y_f, έτσι ώστε να μπορεί το δίκτυο, πιθανώς, να διακρίνει καλύτερα το σημείο όπου τερματίζεται κάθε νότα, προσδίδοντας μεγαλύτερη ευκρίνεια σε σχέση με το Basic Pitch μοντέλο.

4.4.2 Προ-επεξεργασία

Προτού πραγματοποιηθεί η εκπαίδευση του δικτύου, είναι σημαντικό να πραγματοποιηθεί η κατάλληλη προ-επεξεργασία, ώστε να μετατραπούν τα annotations σε αναπαραστάσεις "κάδων" συχνότητας και χρόνου. Για τη νέα αναπαράσταση των offsets, χρησιμοποιείται η ίδια διαδικασία τεμαχισμού όπως και στην περίπτωση των onsets, αφού έχουν πολλές εννοιολογικές ομοιότητες. Έτσι, αφού τα αρχικά αρχεία ήχου χωριστούν, με κατάλληλο γέμισμα, σε ίσα τμήματα 2 δευτερολέπτων, παράγονται οι 3 αναπαραστάσεις που έχουν ήδη παρουσιαστεί, μαζί με τη νέα που αφορά τα offsets, και έχει διαστάσεις (88,173). Έχοντας αυτά κατά νου, οι αναπαραστάσεις, μετά την επεξεργασία, σχηματίζονται ως εξής (Πίνακας 4.10):

Annotation	Συμβολισμός	Διάσταση
Τόνος	Y_p	(264, 173)
Νότα	Y_n	(88, 173)
Onset	Y_o	(88, 173)
Offset	Y_o	(88, 173)

Πίνακας 4.10: *Διαστάσεις annotations για το Basic Pitch with offsets μοντέλο*

Η ανάλυση, είναι 173 "κάδοι" ανά 2 δευτερόλεπτα, στον χρονικό άξονα, και 1 ή 3 "κάδοι" ανά ημιτόνια στον συχνοτικό, ανάλογα με το είδος της αναπαράστασης. Ο διαχωρισμός των δεδομένων είναι, όπως σε κάθε μοντέλο 80:10:10, ενώ στα δεδομένα εισόδου, πριν εισέλθουν στα εκπαιδεύσιμα τμήματα του δικτύου, πραγματοποιείται η ίδια επαύξηση δεδομένων που έχει περιγραφεί ήδη από το Deep Salience μοντέλο, και ο μετασχηματισμός HCQT με 8 αρμονικές.

4.4.3 Μέτα-επεξεργασία

Το στάδιο της μέτα-επεξεργασίας ακολουθεί τον ίδιο αλγόριθμο με αυτό των προηγούμενων δύο μοντέλων, ο οποίος είναι έχει προκύψει ως επέκταση του αλγορίθμου του Onsets and Frames [10], και διατυπώθηκε στην αρχική δημοσίευση του Basic Pitch [3]. Ο συγκεκριμένος αλγόριθμος, χρησιμοποιεί τις προβλέψεις των αναπαραστάσεων Y_o και Y_n, και, ορίζοντας 4 παραμέτρους προς βελτιστοποίηση, επιχειρεί να κατασκευάσει μια τελική αναπαράσταση με τις νότες που εκτελέστηκαν στο αρχικό μουσικό έργο. Οι 4 παράμετροι, όπως έχουν ήδη παρουσιαστεί, είναι το όριο έναρξης νότας, το όριο νότας, η ελάχιστη διάρκεια νότας και η ανοχή.

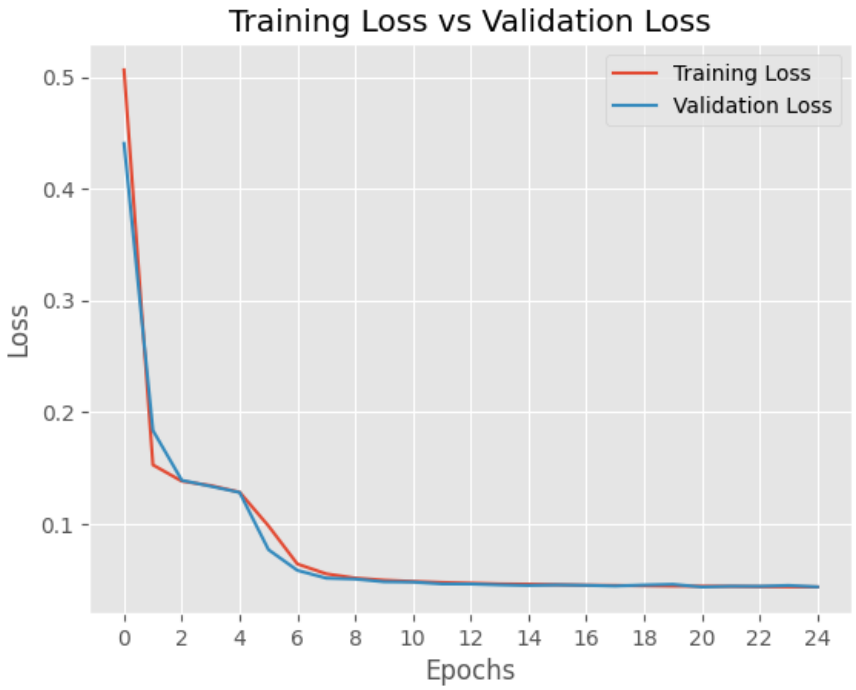
Το στάδιο της εφαρμογής του αλγορίθμου, παρότι είναι ακριβώς ίδιο με του Basic Pitch μοντέλου. Πιθανώς θα μπορούσε να συμπεριλαμβάνει και την Y_f αναπαράσταση, κάτι που θα μπορούσε να βελτιώσει τα αποτελέσματα της μετρικής F, η οποία τα λαμβάνει υπόψη. Αξίζει να σημειωθεί, ότι ο αλγόριθμος, όπως έχει εξηγηθεί ήδη στην ενότητα για το Basic Pitch, εξάγει κάποια υποψήφια onsets, για τα οποία, στα τελικά στάδια του αλγορίθμου, πραγματοποιείται έρευνα για ύπαρξη νοτών τόσο μπροστά όσο και πίσω στον χρόνο. Θα μπορούσε να γίνει χρήση της Y_f, ιδιαίτερα για την έρευνα πίσω στο χρόνο, αφού τα offsets εξ' ορισμού τοποθετούνται στο τέλος της νότας.

4.4.4 Αποτελέσματα

4.4.4.1 Εκπαίδευση Μοντέλου

Κατά την εκπαίδευση του μοντέλου, ακολουθείται ταυτόσημη διαδικασία με αυτή του Basic Pitch, με μια μικρή διαφορά στον ορισμό της συνάρτησης απώλειας την οποία το δίκτυο επιχειρεί να ελαχιστοποιήσει. Η διαφοροποίηση, η οποία προέρχεται από την επιπλέον πρόβλεψη Y_f, οδηγεί στη χρήση της συνάρτησης σφάλματος της Εξίσωσης 4.2, η οποία ισούται, πλέον, με το άθροισμα 4 επιμέρους απωλειών διασταυρούμενης εντροπίας, αντί 3.

$$L = L_p + L_n + L_o + L_f \tag{4.2}$$

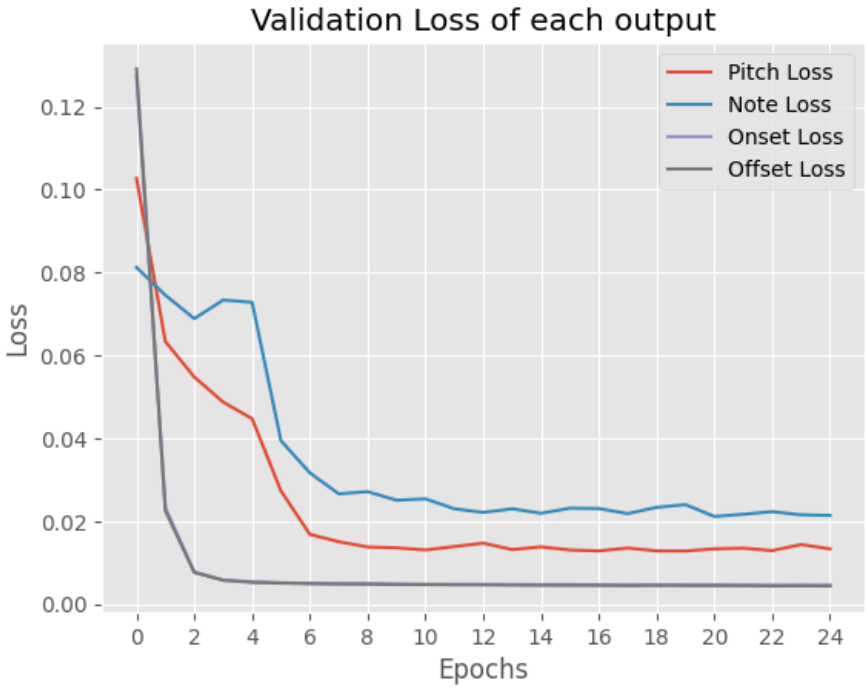


Εικόνα 4.17: *Απώλεια εκπαίδευσης (κόκκινο) και επικύρωσης (μπλε) κατά την εκπαίδευση του Basic Pitch with offsets μοντέλου*

Με τη συνάρτηση απώλειας να αποτελεί τη μοναδική, ουσιαστικά, διαφορά, οι παράμετροι της εκπαίδευσης του δικτύου διατηρούνται είναι όπως στον Πίνακα 4.1, με τη διαφορά ότι η συνάρτηση απώλειας απώλειας είναι το άθροισμα 4 επιμέρους απωλειών διασταυρούμενης επικύρωσης.

Στην Εικόνα 4.17 φαίνονται οι απώλειες εκπαίδευσης και επικύρωσης και παρατηρείται η διακύμανσή τους σε παραπλήσιες τιμές. Επιπλέον, μετά την εκπαίδευση, όπως και στα προηγούμενα μοντέλα, επιστρέφονται τα βάρη του δικτύου για τα οποία η απώλεια επικύρωσης ήταν ελάχιστη.

Για την περαιτέρω ανάλυση της πορείας της εκπαίδευσης, παρουσιάζεται η πορεία των 4 επιμέρους απωλειών επικύρωσης συναρτήσει των εποχών, στην Εικόνα 4.18. Αξίζει να σημειωθεί, ότι οι απώλειες των onsets και offsets, κυμαίνονται σε τόσο κοντινές τιμές, ώστε η γραφική παράσταση της μιας να υπερκαλύπτει την άλλη στο διάγραμμα. Με τις απώλειες



Εικόνα 4.18: *Απώλεια τόνου (κόκκινο), νότας (μπλέ), onset (μώβ) και offset (γκρι) κατά την εκπαίδευση του Basic Pitch with offsets μοντέλου*

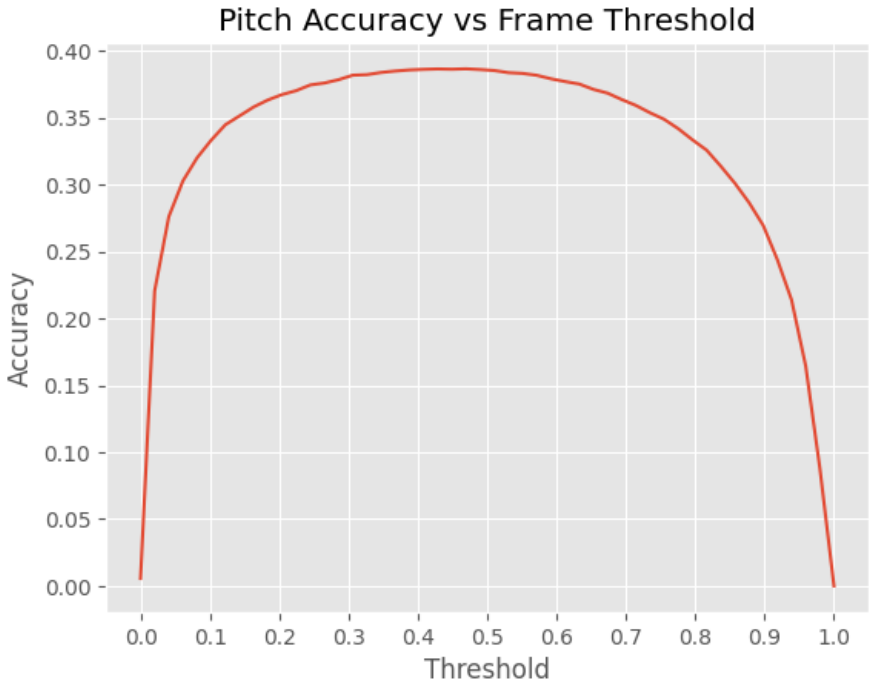
αυτές να κυμαίνονται στα χαμηλότερα επίπεδα, ακολουθεί η απώλεια επικύρωσης του τόνου, και τέλος αυτή των νοτών, η οποία είναι σταθερά μεγαλύτερη από τις υπόλοιπες. Με την πραγματοποίηση της εκπαίδευσης, σειρά έχει η επικύρωση του μοντέλου.

4.4.4.2 Επικύρωση Μοντέλου

Ξεκινώντας την επικύρωση του μοντέλου, μελετάται, αρχικά, η μετρική της πιστότητας τόνου, η οποία υπολογίζεται σε επίπεδο χρονικών "κάδων", χρησιμοποιώντας την πρόβλεψη της Y_p αναπαράστασης. Για τον υπολογισμό αυτής, ορίζεται ένα κατώφλι, το οποίο χρησιμοποιείται στην έξοδο του δικτύου, ώστε να τεθούν τα στοιχεία της είτε ως 0 είτε ως 1. Για την εύρεση της βέλτιστης τιμής του κατωφλίου, ελέγχθηκαν 50 ισαπέχουσες τιμές από το 0 ως το 1, όπως φαίνεται στην Εικόνα 4.19. Τα αποτελέσματα της παραπάνω έρευνας, επέστρεψαν βέλτιστη τιμή κατωφλίου ίση με **0.4694**, με πιστότητα στο σύνολο επικύρωσης ίση με **38.65%**.

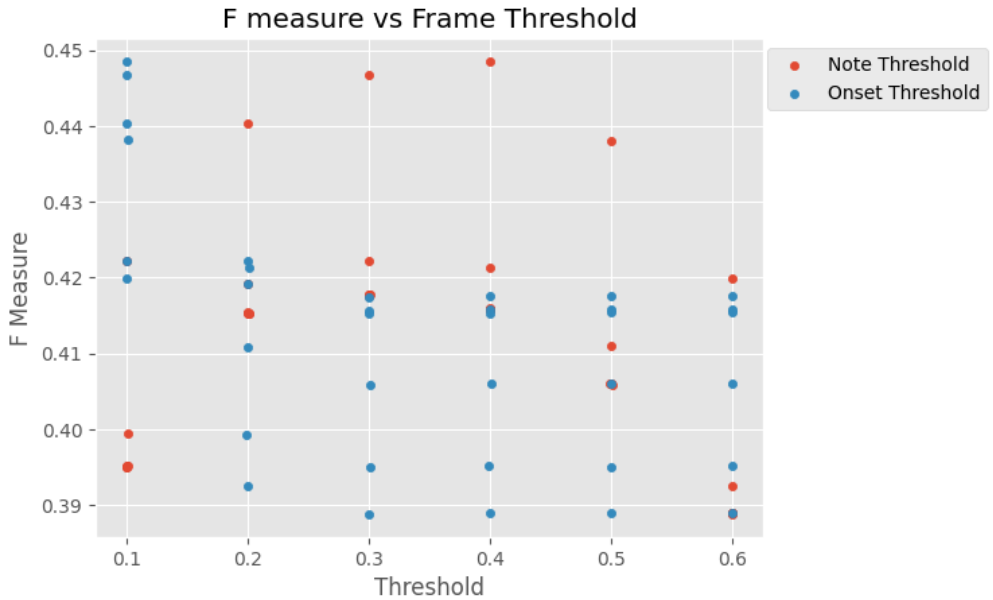
Έχοντας καθορίσει την ιδανική τιμή για τον υπολογισμό της πιστότητας τόνου, σειρά έχουν οι μετρικές σε επίπεδο νοτών, οι οποίες είναι οι F και Fno. Για τον υπολογισμό τους, εφαρμόζεται ο αλγόριθμος μετα-επεξεργασίας ο οποίος έχει ήδη περιγραφεί και καθορίζεται από την αρχικοποίηση 4 διαφορετικών παραμέτρων. Η έρευνα για τις παραμέτρους αυτές, πραγματοποιείται εντός του πλέγματος που περιγράφεται στον Πίνακα 4.5.

Μετά την πραγματοποίηση της παραπάνω έρευνας, προέκυψαν τα αποτελέσματα της Εικόνας 4.20, όπου φαίνεται, για άλλη μια φορά, η μεγαλύτερη συσχέτιση της απόδοσης του μοντέλου με το κατώφλι νότας. Το διάγραμμα, περιγράφει τη μετρική F συναρτήσει των 2 διαφορετικών τύπων κατωφλίου, για τις τιμές του πλέγματος το οποίο ερευνήθηκε. Παρουσι-



Εικόνα 4.19: *FLA συναρτήσει της τιμή κατωφλίου πλαισίου κατά την επικύρωση του Basic Pitch with offsets μοντέλου*

άζεται η συγκεκριμένη μετρική, αφού χρησιμοποιώντας τη αποφασίστηκε η βέλτιστη τετράδα παραμέτρων **(όριο έναρξης νότας, όριο νότας, ελάχιστη διάρκεια νότας, ανοχή)**, η οποία είναι ίση με **(0.1, 0.4, 11, 5)**. Βάσει αυτής, προκύπτουν για τις μετρικές, τιμές **F = 44.84%** και **Fno = 63.73%** στο σύνολο επικύρωσης.



Εικόνα 4.20: *Μετρική F συναρτήσει των κατωφλίων νότας και onsets κατά την επικύρωση του Basic Pitch with offsets μοντέλου*

4.4.4.3 Αξιολόγηση Μοντέλου

Για την αξιολόγηση του μοντέλου, πραγματοποιείται υπολογισμός 4 διαφορετικών μετρικών στο σύνολο δεδομένων ελέγχου, το οποίο έχει κρατηθεί εξ αρχής, βάσει των βέλτιστων παραμέτρων οι οποίες αποφασίστηκαν στο στάδιο επικύρωσης. Για τον έλεγχο της απόδοσης, χρησιμοποιούνται οι F & Fno ως μετρικές επιπέδου νότας, και οι πιστότητες νότας και τόνου ως μετρικές επιπέδου χρονικού πλαισίου. Η μετρική πιστότητας νότας υπολογίζεται με τον ίδιο τρόπο με την πιστότητα τόνου, αλλά στην έξοδο που προβλέπει την Y_n αναπαράσταση αντί της Y_p.

Σύνολο Ελέγχου	Πιστότητα τόνου	Πιστότητα νότας	Μετρική F	Μετρική Fno
Εξαφωνικό	37.78%	39.11%	45.61%	66.56%
Εξαφωνικό Debleeded	37.77%	38.98%	46.86%	67.15%
Εξαφωνικό και Εξαφωνικό Debledeed	37.77%	39.04%	46.23%	66.86%
Μονοκάναλο	36.79%	37.43%	43.34%	62.03%
Μείξη μονοκάναλου και εξαφωνικού	37.27%	38.71%	46.52%	66.43%

Πίνακας 4.11: *Σύγκριση επίδοσης της πιστότητας τόνου και νότας και των μετρικών F και Fno σε διαφορετικά σύνολα ελέγχου για το Basic Pitch with offsets μοντέλο*

Η σύγκριση των μετρικών, που παρουσιάζεται στον Πίνακα 4.11, πραγματοποιείται μεταξύ των 5 διαφορετικών τύπων του συνόλου ελέγχου, ανάλογα με τον τρόπο ηχογράφησης ή την μίξη. Οι βέλτιστες τιμές για τις μετρικές F & Fno, εντοπίζονται στο Εξαφωνικό Debleeded σύνολο ελέγχου, ενώ οι βέλτιστες πιστότητες, στο Εξαφωνικό σύνολο ελέγχου, όπου δεν έχει εφαρμοστεί Debleeding.

Κεφάλαιο 5

Σύγκριση Αποτελεσμάτων

5.1 Μεθοδολογία Σύγκρισης

Μετά την εκπαίδευση, επικύρωση και έλεγχο του κάθε μοντέλου ξεχωριστά, καθίσταται χρήσιμη η σύγκρισή τους, η οποία βέβαια δύνανται να πραγματοποιηθεί σε έναν αριθμό διαφορετικών επιπέδων. Σε κάθε περίπτωση, θα χρησιμοποιηθούν οι τιμές των μετρικών όπως προέκυψαν για το σύνολο δεδομένων ελέγχου, το οποίο αποτελεί την αντικειμενικότερη παρουσίαση της απόδοσης του κάθε μοντέλου σε νέα δεδομένα. Έχοντας αυτά κατά νου, οι άξονες της σύγκρισης είναι οι εξής:

1. Σύγκριση σε επίπεδο υπολογιστικού κόστους του κάθε μοντέλου.
2. Σύγκριση των μετρικών οι οποίες υπολογίζονται σε επίπεδο χρονικού πλασίου, δηλαδή τις πιστότες τόνου και νότας, όπου η τελευταία υπολογίστηκε.
3. Σύγκριση των μετρικών οι οποίες υπολογίζονται σε επίπεδο νότας, δηλαδή τις F και Fno.
4. Σύγκριση της απόδοσης στο μονοκαναλικό σύνολο ελέγχου, το οποίο αποτελεί τον απλούστερο τρόπο ηχογράφησης.

5.2 Αναλυτική Παρουσίαση & Συμπεράσματα

Ξεκινώντας με τη σύγκριση του υπολογιστικού κόστους, παρουσιάζονται στον Πίνακα 5.1, τόσο ο μέσος χρόνος εκπαίδευσης ανά εποχή, όσο και ο αριθμός των συνολικών εκπαιδεύσιμων παραμέτρων του κάθε δικτύου.

Μοντέλο	Πλήθος εκπαιδεύσιμων παραμέτρων	Χρόνος ανά εποχή
Deep Salience	143.635	411
Basic Pitch	45.742	139
Basic Pitch ConvLSTM	470.670	521
Basic Pitch with offsets	52.548	150

Πίνακας 5.1: *Σύγκριση του υπολογιστικού κόστους μεταξύ των μοντέλων*

Μια αρχική παρατήρηση, αποτελεί το γεγονός ότι, όπως ήταν αναμενόμενο, ο αριθμός των εκπαιδεύσιμων παραμέτρων είναι αναλογικός του μέσου χρόνου εκπαίδευσης ανά επο-

χή. Το μοντέλο με τις περισσότερες παραμέτρους είναι το Basic Pitch ConvLSTM, το οποίο περιέχει αναδρομική δομή, μέσω του LSTM, οπότε είναι πιο περίπλοκο από τα υπόλοιπα δύο Basic Pitch μοντέλα. Τη δεύτερη πιο περίπλοκη, από υπολογιστική άποψη, δομή, κατέχει το Deep Salience μοντέλο, κυρίως λόγω του μεγαλύτερου βάθους των φίλτρων που χρησιμοποιεί σε κάθε συνελικτικό στρώμα. Τέλος, τα δύο εναπομείναντα, μοντέλα παρουσιάζουν παρόμοια χαρακτηριστικά όσο αφορά το υπολογιστικό κόστος τους. Η σύγκριση του κόστους εκπαίδευσης του κάθε δικτύου, είναι ένα σημαντικό σημείο, ιδιαιτέρως, στην πιθανή περίπτωση παροχής νέων δεδομένων, όπου θα ήταν χρήσιμη η επανεκπαίδευση του δικτύου για τη βελτιστοποίηση της απόδοσής του.

Μοντέλο	Πιστότητα Τόνου	Πιστότητα Νότας
Deep Salience	37.39%	-
Basic Pitch	37.96%	38.62%
Basic Pitch ConvLSTM	34.39%	31.35%
Basic Pitch with ffsets	37.78%	39.11%

Πίνακας 5.2: *Σύγκριση επίδοσης της πιστότητας τόνου και νότας μεταξύ των μοντέλων*

Συνεχίζοντας στο επόμενο στάδιο σύγκρισης, σειρά έχουν οι μετρικές πιστότητας τόνου και νότας, οι οποίες παρουσιάζονται στον Πίνακα 5.2. Και τα 4 μοντέλα φαίνεται να αποδίδουν σχεδόν στο ίδιο επίπεδο, με το Basic Pitch ConvLSTM να έχει τη χειρότερη απόδοση στις συγκεκριμένες μετρικές. Στην περίπτωση του Deep Salience, δεν υπολογίστηκε πιστότητα νότας, αφού το συγκεκριμένο μοντέλο επιχειρεί να προβλέψει μόνο την Y_p αναπαράσταση. Την βέλτιστη απόδοση, με πολύ μικρή διαφορά, παρουσιάζει το Basic Pitch μοντέλο στην περίπτωση της πιστότητας τόνου, και το Basic Pitch with offsets στην περίπτωση της πιστότητας νοτών. Η βελτιωμένη απόδοση στη μετρική των νοτών, θα μπορούσε, πιθανώς, να αποδοθεί στη χρήση των offsets, αλλά η διαφορά με το μοντέλο Basic Pitch θεωρείται πολύ μικρή ώστε να διατυπωθεί τέτοιο επιχείρημα.

Μοντέλο	Μετρική F	Μετρική Fno
Deep Salience	-	-
Basic Pitch	48.67%	67.46%
Basic Pitch ConvLSTM	29.29%	51.85%
Basic Pitch Offsets	46.86%	67.15%

Πίνακας 5.3: *Σύγκριση επίδοσης των μετρικών F & Fno μεταξύ των μοντέλων*

Προχωρώντας στις μετρικές σε επίπεδο νοτών, παρουσιάζονται τα συγκριτικά αποτελέσματα στον Πίνακα 5.3. Αρχικά, το μοντέλο Deep Salience δεν συμμετέχει στην συγκεκριμένη σύγκριση, αφού για τον υπολογισμό των F και Fno χρειάζεται δυνατότητα πρόβλεψης των Y_o και Y_n, κάτι το οποίο δεν είναι διαθέσιμο για το συγκεκριμένο μοντέλο. Την βέλτιστη απόδοση, και στις δύο μετρικές, παρουσιάζει το Basic Pitch, με το Basic Pitch Offsets να έρχεται δεύτερο, και το Basic Pitch ConvLSTM να έχει τη χειρότερη απόδοση εκ των τριών. Φαίνεται, μέσω της σύγκρισης, ότι η προσθήκη του νέου κλάδου πρόβλεψης των Offsets στο Basic Pitch μοντέλο, δεν βελτίωσε την απόδοση των μετρικών, τουλάχιστον διατηρώντας τον ίδιο αλγόριθμο μετα-επεξεργασίας.

Στο τελευταίο στάδιο της σύγκρισης, έχει ενδιαφέρον η παρουσίαση των προαναφερθέντων μετρικών στο μονοκαναλικό σύνολο ελέγχου, κάτι που πραγματοποιείται με τη βοήθεια του Πίνακα 5.4. Το σύνολο αυτό αντιπροσωπεύει τον ευκολότερο τρόπο ηχογράφησης μιας μουσικής εκτέλεσης, επομένως σε περίπτωση χρήσης ενός εκ των μοντέλων σε πραγματικές συνθήκες, για παράδειγμα εντός κάποιας εφαρμογής κινητού, θα ήταν ο πιο πιθανός τρόπος χρήσης της. Τα υπόλοιπα σύνολα, τα οποία περιλαμβάνουν τις εξαφωνικές ηχογραφήσεις, σε οποιαδήποτε μορφή τους, αποτελούν έναν αρκετά δυσκολότερο τρόπο ηχογράφησης, αφού απαιτείται εξειδικευμένος εξοπλισμός, όπως είναι ο εξαφωνικός μαγνήτης που παρουσιάστηκε στα αρχικά κεφάλαια. Τη βέλτιστη απόδοση, συγκριτικά, κατέχουν τα μοντέλα Basic Pitch και Basic Pitch Offsets, με τις διαφορές στις αντίστοιχες μετρικές τους να κυμαίνονται περίπου στο 1%. Τη χείροτερη απόδοση, για άλλη μια φορά, παρουσιάζει το Basic Pitch CovnLSTM μοντέλο, εξαιρώντας, φυσικά, το Deep Salience, το οποίο δύνανται να συμμετέχει στον υπολογισμό μιας μόνο εκ των 4 διαθέσιμων μετρικών.

Μοντέλο	Πιστότητα τόνου	Πιστότητα νότας	Μετρική F	Μετρική Fno
Deep Salience	36.04%	-	-	-
Basic Pitch	36.57%	36.16%	44.36%	61.44%
Basic Pitch ConvLSTM	33.43%	29.42%	25.63%	45.48%
Basic Pitch Offsets	36.79%	37.43%	43.34%	62.03%

Πίνακας 5.4: *Σύγκριση επίδοσης των των μουτέλων στο μονοκαναλικό σύνολο ελέγχου*

Έχοντας εξετάσει το κάθε μοντέλο ξεχωριστά, και παρουσιάσει τη σύγκρισή τους σε 4 διαφορετικά επίπεδα, μένει μόνο, η παρουσίαση κάποιων συμπερασμάτων ως προϊόντα των παραπάνω εργασιών. Από τα 4 μοντέλα, φαίνεται να παρουσιάζουν τη βέλτιστη απόδοση τα Basic Pitch και Basic Pitch with offsets. Τόσο από άποψη υπολογιστικού κόστους, όσο και συνολικής απόδοσης στις 4 μετρικές που εξετάστηκαν, θα μπορούσε να επιλεχθεί είτε το ένα είτε το άλλο. Το Basic Pitch ConvLSTM μοντέλο, φαίνεται να αποδίδει σταθερά χειρότερα από τα υπόλοιπα δύο, απορρίπτοντας, κατά πάσα πιθανότητα, τα χρονικά αναδρομικά μοντέλα τα οποία είναι στο πνεύμα του Basic Pitch. Επιπλέον, αξίζει να σημειωθεί ότι πάντα τα αναδρομικά μοντέλα, θα είναι πιο σύνθετα υπολογιστικά από τα αντίστοιχα μη-αναδρομικά, κάτι το οποίο θα πρέπει να αντισταθμίζεται από τη βελτιωμένη απόδοσή τους.

Στην περίπτωση του Basic Pitch with offsets μοντέλου, αξίζει να σημειωθεί, ότι παρά την παραπλήσια απόδοση με το αρχικό Basic Pitch, ίσως προσφέρει περισσότερα μελλοντικά, αφού παρέχει μια επιπλέον πρόβλεψη. Η πρόβλεψη της αναπαράστασης των offsets Y_f, παρότι δεν φάνηκε να βελτιώνει την απόδοση του δικτύου μέσω της συνδρομής του κατά τη διάρκεια της εκπαίδευσης, θα μπορούσε, μελλοντικά, να ληφθεί υπόψη κατά τη διάρκεια του αλγορίθμου μετα-επεξεργασίας. Συγκεκριμένα, ο αλγόριθμος σε ένα από τα τελικά στάδιά του, επιχειρεί τον εντοπισμό υποψήφιων σημείων, τα οποία διερευνούνται τόσο μπροστά όσο και πίσω στον χρόνο. Ειδικά για την περίπτωση διερεύνησης πίσω στον χρόνο, θα μπορούσε να χρησιμοποιηθεί με κάποιο τρόπο η αναπαράσταση Y_f, αφού τα offsets, εν γένει, αποτελούν τα τελευταία χρονικά σημεία μιας νότας.

Μέρος III

Επίλογος

Κεφάλαιο 6

Επίλογος

6.1 Μελλοντικές Επεκτάσεις

Σαν τελευταίο, αλλά εξίσου σημαντικό τμήμα της παρούσας εργασίας, κρίθηκε απαραίτητο να αναφερθούν πιθανές επεκτάσεις και προτάσεις, οι οποίες πιθανώς να προσθέσουν στις ήδη υπάρχουσες προσεγγίσεις. Αρχικά, μια γενική παρατήρηση αποτελεί το γεγονός ότι χρησιμοποιήθηκε μόνο το Guitarset ως πηγή δεδομένων. Για τον έλεγχο των μοντέλων, καθώς και την εξαγωγή ασφαλέστερων συμπερασμάτων, θα ήταν ωφέλιμη η χρήση μεγαλύτερου συνόλου συνόλου διαφορετικών συνόλων δεδομένων για τη σύγκριση.

Έχοντας αναφέρει το παραπάνω, παρουσιάζονται παρακάτω μερικές μελλοντικές επεκτάσεις:

- Όπως φάνηκε, παρότι η διαδικασία εκπαίδευσης των δικτύων δεν ήταν ιδιαίτερα χρονοβόρα, το στάδιο της μετα-επεξεργασίας για την εύρεση των ιδανικών παραμέτρων προκαλούσε ιδιαίτερη επιμήκυνση στον χρόνο επικύρωσης, ο οποίος αυξανόταν εκθετικά με το εύρος πραγματοποιούμενης αναζήτησης πλέγματος. Ένας πιθανός τρόπος επίλυσης του συγκεκριμένου ζητήματος είναι η επιλογή μιας διαφορετικής συνάρτησης απώλειας κατά την εκπαίδευση, η οποία θα συσχετιζόταν περισσότερο με τις μετρικές F και Fno με τέτοιο τρόπο ώστε να μη χρειαζόταν ο αλγόριθμος μετά την εκπαίδευση.

- Θα παρουσίαζε ενδιαφέρον η λεπτομερής μελέτη της επίδρασης των διαφορετικών μεθόδων επαύξησης που μπορούν να πραγματοποιηθούν, ειδικά με τη μορφή ηχητικών εφέ, τα οποία χρησιμοποιούνται σε μεγάλο βαθμό και στην ίδια τη μουσική. Εκτός από τα εφέ Gain & Reverb, τα οποία χρησιμοποιήθηκαν στη συγκεκριμένη εργασία, υπάρχουν αρκετά πιο σύνθετα εφέ ήχου, τα οποία οδηγούν σε διάφορα στάδια παραμόρφωσης της αρχικής ηχογράφησης.

- Το Guitarset, γύρω από το οποίο δομήθηκε η παρούσα διπλωματική εργασία, περιέχει ηχογραφήσεις οι οποίες περιλαμβάνουν τη χρήση τόσο εξαφωνικού μαγνήτη όσο και μικροφώνου, σαν αυτό που θα χρησιμοποιούνταν σε κάποιο στούντιο ηχογράφησης, για την ηχογράφηση ακουστικής κιθάρας. Παρόλα αυτά, μια ακόμη πιο προσβάσιμη μέθοδος ηχογράφησης, αποτελεί η ηχογράφηση μέσω του μικροφώνου κινητού, τα οποία είναι ευρέως διαθέσιμα. Έχοντας αυτό κατά νου, θα είχε ενδιαφέρον η επέκταση του συγκεκριμένου συνόλου δεδομένων με ηχογραφήσεις μουσικών εκτελέσεων με χρήση μικροφώνου κινητού, ώστε να μελετηθεί και αυτή η προσέγγιση.

Βιβλιογραφία

[1] Q. Xi, R. Bittner, J. Pauwels, X. Ye και J. P. Bello. *Guitarset: A Dataset For Guitar Transcription*. *19th International Society for Music Information Retrieval Conference*, 2018.

[2] Rachel M. Bittner, Brian McFeel, Justin Salamon, Peter Li και Juan P. Bello. *Deep Salience Representation for F0 Estimation In Polyphonic Music*, 2017.

[3] Rachel M. Bittner, Juan José Bosch, David Rubinstein, Gabriel Meseguer-Brocal και Sebastian Ewert. *A Lightweight Instrument-Agnostic Model For Polyphonic Note Transcription And Multipitch Estimation*. 2022.

[4] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong και Wang-chun Woo. *Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting*. *CoRR*, abs/1506.04214, 2015.

[5] Marius Miron και Agustin Martorell. *Bach10 Sibelius Dataset*, 2017.

[6] Li Su και yi hsuan Yang. *Escaping from the Abyss of Manual Annotation: New Methodology of Building Polyphonic Datasets for Automatic Music Transcription*. τόμος 9617, σελίδες 309–321, 2016.

[7] Rachel Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam και Juan Pablo Bello. *MedleyDB Audio: A Dataset of Multitrack Audio for Music Research*, 2014.

[8] Emilio Molina, Ana M. Barbancho, Lorenzo J. Tardón και Isabel Barbancho. *Evaluation Framework for Automatic Singing Transcription*. *International Society for Music Information Retrieval Conference*, 2014.

[9] Tak Shing Chan, Tzu Chun Yeh, Zhe Cheng Fan, Hung Wei Chen, Li Su, Yi Hsuan Yang και Roger Jang. *Vocal activity informed singing voice separation with the i-Kala dataset*. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, σελίδες 718–722, 2015.

[10] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel και Douglas Eck. *Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset*. *International Conference on Learning Representations*, 2019.

[11] Ethan Manilow, Gordon Wichern, Prem Seetharaman και Jonathan Le Roux. *Cutting Music Source Separation Some Slakh: A Dataset to Study the Impact of Training Data Quality and Quantity*. *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019.

[12] Marius Miron, Joe Orti, Juan Bosch, Emilia Gómez και Jordi Janer. *Score-Informed Source Separation for Multichannel Orchestral Recordings*. *Journal of Electrical and Computer Engineering*, 2016:1–19, 2016.

[13] *Constant-Q Transform Toolbox for Music Processing*. Zenodo, 2010.

[14] Thomas Prätzlich, Rachel M. Bittner, Antoine Liutkus και Meinard Müller. *Kernel Additive Modeling for interference reduction in multi-channel music recordings*. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, σελίδες 584–588, 2015.

[15] Matthias Mauch και Simon Dixon. *PYIN: A fundamental frequency estimator using probabilistic threshold distributions*. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, σελίδες 659–663, 2014.

[16] Colin Raffel, Brian McFee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang και Daniel P. W. Ellis. *MIR_EVAL: A Transparent Implementation of Common MIR Metrics*. *ISMIR*, σελίδες 367–372, 2014.

[17] Diederik P Kingma και Jimmy Ba. *Adam: A method for stochastic optimization*. *arXiv preprint arXiv:1412.6980*, 2014.

Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια

CNN	Convolutional Neural Network
ConvLSTM	Convolutional LSTM
CQT	Constant-Q Transform
FLA	Frame Level Accuracy
GPU	Graphics Processing Unit
HCQT	Harmonic Constant Q Transform
LSTM	Long Short-Tem Memory
ML	Machine Learning
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Networks
Y_o	Onset Annotation
Y_f	Offset Annotation
Y_n	Note Annotation
Y_p	Pitch Annotation

Απόδοση ξενόγλωσσων όρων

Απόδοση	Ξενόγλωσσος όρος
Νευρωνικό	Neural
Δίκτυο	Network
Αναδρομικός	Recurrent
Συνέλιξη	Convolution
Αρμονικό	Harmonic
Στάθμη Τεχνικής	State of the Art
Στρώμα	Layer
Έναρξη	Onset
Λήξη	Offset
Νότα	Note
Τόνος (ηχητικός)	Pitch
Σύνολο Δεδομένων	Dataset
Ανάστροφη διάδοση	Backpropagation
Πυρήνας	Kernel
Σιγμοειδής	Sigmoid
Μακρύ	Long
Βραχυπρόθεσμο	Short-Term
Μνήμη	Memory
Εργασία	Task
Κρυφό	Hidden
Κελί	Cell
Κατάσταση	State
Λήθη	Forget
Είσοδος	Input
Έξοδος	Output
Πύλη	Gate
Εκπαίδευση	Training
Εναλλακτική	Workaraound
Σύρω	Slide
Σταθερός	Constant
Κάδος	Bin
Στοιβάζω	Stack
Δέσμη	Batch
Μέγεθος	Size