ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΕΥΦΥΩΝ ΣΥΣΤΗΜΑΤΩΝ, ΠΕΡΙΕΧΟΜΕΝΟΥ ΚΑΙ ΑΛΛΗΛΕΠΙΔΡΑΣΗΣ

Αυτόματη Αναγνώριση Συγχορδίας με Μεθόδους Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Έντμοντ-Γρηγόρη Γ. Ντερβάκου

Επιβλέπων: Γ. Στάμου
Αν. Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2018

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΕΥΦΥΩΝ ΣΥΣΤΗΜΑΤΩΝ, ΠΕΡΙΕΧΟΜΕΝΟΥ ΚΑΙ ΑΛΛΗΛΕΠΙΔΡΑΣΗΣ

Αυτόματη Αναγνώριση Συγχορδίας με Μεθόδους Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Έντμοντ-Γρηγόρη Γ. Ντερβάκου

Επιβλέπων: Γ. Στάμου
Αν. Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την ...η Ιουλίου 2018.

.....	.....	.....
Γ. Στάμου	Α-Γ Σταφυλοπάτης	Ν. Παπασπύρου
Αν. Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αν. Καθηγητής Ε.Μ.Π.

.....

Έντμοντ - Γρηγόρης Γ. Ντερβάκος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η Αυτόματη Αναγνώριση Συγχορδίας είναι κομβικό έργο για τον τομέα της ανάκτησης πληροφορίας από μουσική. Οι συγχορδίες κωδικοποιούν την αρμονική πληροφορία ενός κομματιού και καθορίζουν σε μεγάλο βαθμό το μεταδιδόμενο συναίσθημα. Για μουσικούς είναι ίσως το αποδοτικότερο μέσο επικοινωνίας όταν αυτοί παίζουν σε σύνολο, ενώ σε κάποιες περιπτώσεις οι αλληλουχίες συγχορδιών αντικαθιστούν την αναλυτική παρτιτούρα.

Όπως με τα περισσότερα έργα του τομέα ανάκτησης πληροφορίας από μουσική, τα συστήματα αυτόματης αναγνώρισης συγχορδίας ακολουθούν την τάση να αντικαθιστούν στάδια επεξεργασίας σήματος, εξαγωγής χαρακτηριστικών και στατιστικών μοντέλων με αρχιτεκτονικές βαθιάς μηχανικής μάθησης. Στην παρούσα εργασία ακολουθήθηκε ο πιο παραδοσιακός δρόμος, και για την πρόβλεψη χρησιμοποιήθηκαν υψηλού επιπέδου χαρακτηριστικά - αυτά που παρέχει το Spotify μέσω του API του. Αυτά περιέχουν τμηματοποιήσεις του κομματιού μουσικής, με επιπλέον χαρακτηριστικά να αντιστοιχούν σε κάθε τμήμα.

Για το πειραματικό μέρος της εργασίας εκπαιδεύτηκαν διαφορετικά νευρωνικά δίκτυα με σκοπό να αναγνωρίζουν τη συγχορδία που ακούγεται σε κάθε τμήμα από την τμηματοποίηση που παρέχει το Spotify. Συγκεκριμένα χρησιμοποιήθηκαν: Απλό Multi Layer Perceptron ή Feedforward Νευρωνικά Δίκτυα ή FNN, Συνελικτικά Νευρωνικά Δίκτυα ή CNN, η παραλλαγή του Αναδρομικού Νευρωνικού δικτύου Μακράς Βραχυπρόθεσμης Μνήμης ή LSTM καθώς και μια πιο πολύπλοκη αρχιτεκτονική, τον Κωδικοποιητή - Αποκωδικοποιητή LSTM - (LSTM Encoder Decoder).

Το σύνολο εκπαίδευσης αποτελούνταν από συνδυασμό συνόλων δεδομένων επισημειωμένων με συγχορδίες ανά χρονική στιγμή, που χρησιμοποιούνται ευρέως στη βιβλιογραφία. Αξιολογήθηκε η συμπεριφορά των μοντέλων με διαφορετικές παραμέτρους, διαφορετικά χαρακτηριστικά καθώς και η αποτελεσματικότητα τεχνικών προ-επεξεργασίας δεδομένων όπως η επαύξηση και το φιλτράρισμα των δεδομένων.

Abstract

Audio Chord Recognition (ACR) is perhaps one of the most important tasks in the area of Music Information Retrieval (MIR). Chords are a representation of harmony in music, and harmony largely determines the effects of music on humans, such as feelings conveyed. Chord transcriptions are the most efficient way to communicate musical ideas in an ensemble of musicians and in some cases they replace actual sheet music.

As with many MIR tasks, there is a shift occurring in the ACR state-of-the-art, moving from statistical models and high level feature extraction, to Deep Learning, thus alleviating most of the need for signal processing [4]. For this thesis, ACR was attempted on high level features obtained through Spotify's open API. These features include segmentations of each track, with corresponding features assigned to each segment.

For the experimental part of the thesis, different architectures of Artificial Neural Networks were trained to classify each segment of a song to a chord. Models used include: Feedforward Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks in the form of LSTM and a more complex Encoder-Decoder architecture with LSTM layers.

The dataset used consisted of a combination of chord-labeled songs which are used every year for the MIREX chord estimation task [8]. For all models, we explored the effects of using different data preprocessing methods, such as data augmentation and data filtering, different features, and evaluated all models on a separate test set.

Πίνακας Περιεχομένων

Κεφάλαιο 1 - Εισαγωγή - Ανασκόπηση

Η έννοια της συγχορδίας καθώς και η κατανόηση της λειτουργίας και της χρησιμότητάς της, στηρίζεται σε κάποιο βαθμό στη γνώση μουσικής θεωρίας. Για το λόγο αυτό στο δεύτερο κεφάλαιο της εργασίας θα γίνει μια σύντομη αναφορά στις συγχορδίες στο πλαίσιο της θεωρίας μουσικής.

Στο τρίτο κεφάλαιο θα αναφερθούν τρόποι αναγνώρισης συγχορδίας από άνθρωπο και από μηχανή καθώς και οι δυνατότητες της παρούσας τεχνολογίας για το έργο.

Στα κεφάλαια 4 και 5 περιγράφεται η διαδικασία προετοιμασίας των πειραμάτων της εργασίας, με το πρώτο να περιέχει τη συλλογή και την προ-επεξεργασία των δεδομένων και το τελευταίο τις διάφορες παραμέτρους και υπερπαραμέτρους των μοντέλων και πως αυτές επιλέχθηκαν για τα πειράματα.

Στη συνέχεια στο έκτο κεφάλαιο παρουσιάζονται όλα τα πειραματικά αποτελέσματα για τα διαφορετικά μοντέλα που χρησιμοποιήθηκαν, μια σύντομη περιγραφή του κάθε μοντέλου καθώς και η αξιολόγηση κάθε πειράματος.

Στο έβδομο κεφάλαιο γίνεται μια συνολική αξιολόγηση και ανάλυση σφάλματος βασισμένη στα δεδομένα και όχι στα μοντέλα.

Τέλος στο όγδοο κεφάλαιο περιγράφεται η διαδικασία της υλοποίησης όλων των παραπάνω. Ακολουθούν τα συμπεράσματα και η βιβλιογραφία.

Κεφάλαιο 2 - Θεωρία Μουσικής

2.1 Ορισμός Μουσικής

Η μουσική είναι μια “δεξιότητα” που ο ανθρώπινος εγκέφαλος έχει αναπτύξει παράλληλα με τη γλώσσα - εν γένει να δίνει αφηρημένο νόημα σε ήχους [1]. Ο ήχος είναι ένα κύμα πίεσης που διαδίδεται σε κάποιο μέσο. Γενικά ο άνθρωπος αντιλαμβάνεται, με αισθητήριο όργανο το αυτί, τέτοια κύματα που φέρουν συχνότητες μεταξύ (~20Hz και 20kHz). Για κάποιο εύρος συχνοτήτων (μεταξύ ~50Hz και 1kHz), το ανθρώπινο αυτί έχει πολύ μεγάλη ευαισθησία σε μικρές αλλαγές στη συχνότητα και αντιλαμβάνεται ήχους με πολύ μεγάλη ακρίβεια. Σε αυτές τις συχνότητες κυμαίνεται και η ομιλία. Αν σε έναν ήχο ξεχωρίζει (έχει μεγαλύτερη ένταση) μία συχνότητα μέσα σε αυτό το εύρος, τότε τον εκλαμβάνουμε ως τόνο [2]. Αν από την άλλη ένας ήχος επαναλαμβάνεται με συχνότητα ~(0.5Hz-3Hz) εκλαμβάνουμε τον ήχο ως ρυθμό (όχι με μοναδικό αισθητήριο όργανο το αυτί). Οι τόνοι και ο ρυθμός είναι τα στοιχειώδη συστατικά της μουσικής.

Δεν αρκούν όμως τόνοι και ρυθμός για να εκλάβει ο εγκέφαλος έναν ήχο ως μουσική. Σε ένα κομμάτι μουσικής διαφορετικοί τόνοι αλληλεπιδρούν μεταξύ τους με πολύπλοκους τρόπους. Δομούνται σε αλληλουχίες ή ακούγονται ταυτόχρονα, επαναλαμβάνονται ή όχι, ρωτάνε - απαντάνε κ.α. Όπως ακριβώς και στην ομιλία συνδυάζονται διαφορετικές συχνότητες για δημιουργία διαφορετικών φωνημάτων, ή χρησιμοποιούνται διαφορετικοί τόνοι για μετάδοση συναισθήματος ή και νοήματος (πχ. Ερώτηση και κατάφαση). Στη μουσική, μια ακολουθία από τόνους ορίζεται ως μελωδία, ενώ τον τρόπο αλληλεπίδρασης μεταξύ διαφορετικών τόνων επιχειρεί να εξηγήσει η αρμονία. Μπορούμε να συγκεκριμενοποιήσουμε λοιπόν τα συστατικά της μουσικής σε: ρυθμό, μελωδία, αρμονία.

Η μελωδία και ο ρυθμός είναι γενικά κατανοητές έννοιες και μπορεί κανείς να δει τη σύνδεσή τους με τη γλώσσα, καθώς και το βιολογικό λόγο που ο άνθρωπος ανέπτυξε αυτές τις δεξιότητες (τουλάχιστον σε κάποιο βαθμό). Η αρμονία από την άλλη κρύβει ακόμα πολλά ερωτηματικά ως προς το ρόλο της, το

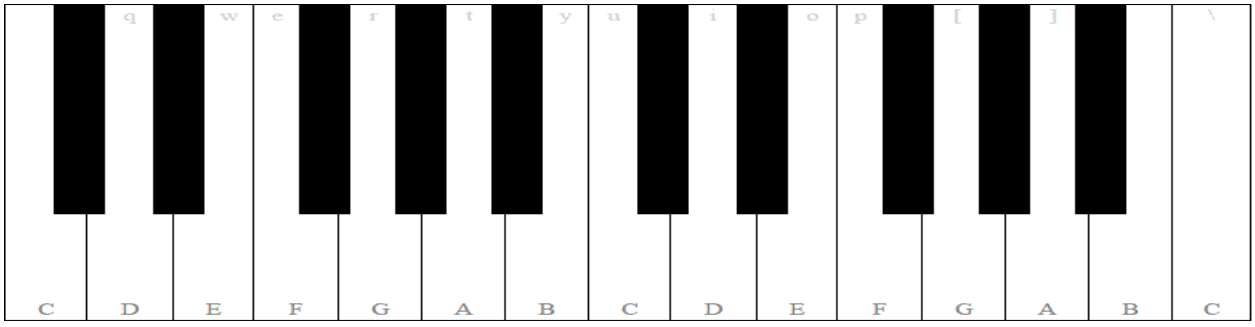
λόγο που προκαλεί έντονα συναισθήματα στους ανθρώπους, ενώ αποτελεί έννοια πιο δυσνόητη σε σχέση με τη μελωδία και το ρυθμό.

2.2 Αρμονία στη Μουσική - Αρμονική Σειρά

Η φυσική βάση της αρμονίας είναι η επαλληλία κυμάτων. Αρχικά κάθε κύμα ισοδυναμεί με (πεπερασμένο ή άπειρο) άθροισμα απλών αρμονικών κυμάτων (μία συχνότητα το καθένα). Όταν εκλαμβάνουμε το κύμα ως τόνο, τότε η "νότα" που ακούμε είναι μία από τις συχνότητες που συνιστούν το κύμα (η επικρατούσα ή θεμελιώδης). Όλες οι άλλες συχνότητες εκλαμβάνονται από τον εγκέφαλο σχετικά με την επικρατούσα και καθορίζουν τις υπόλοιπες ποιότητες του ήχου εκτός από τον τόνο.

Η "σχετικότητα" μεταξύ δύο συχνοτήτων εκφράζεται με βάση την αναλογία τους και προκύπτει από την αρμονική σειρά. Συγκεκριμένα, η πιο "όμοια" συχνότητα σε μια άλλη θα είναι η διπλάσιά της, η επόμενη η τριπλάσια κ.ο.κ. Η διπλάσια συχνότητα μιας άλλης λέγεται πως απέχει μια οκτάβα από αυτήν.

Στη σύγχρονη δυτική μουσική έχουν καθοριστεί 12 θεμελιώδεις συχνότητες, οι 12 νότες, οι οποίες επαναλαμβάνονται ανά οκτάβα. Αυτό φαίνεται εύκολα στα πλήκτρα ενός πιάνο.



Τα γράμματα στο κάτω μέρος των λευκών πλήκτρων είναι ονομασίες νοτών (C - Ντο, D - Ρε κλπ). Τα μαύρα πλήκτρα είναι επίσης νότες. Παρατηρούμε πως κάθε 12 πλήκτρα επαναλαμβάνονται.

Η πιο όμοια με μια νότα είναι αυτή με τη διπλάσια συχνότητα, και η ομοιότητα είναι τόσο ισχυρή που αναθέτουμε την ίδια ονομασία (πχ μια οκτάβα πάνω από την Ντο είναι επίσης Ντο). Αν συνεχίσουμε να πολλαπλασιάζουμε με ακέραιους αριθμούς τις συχνότητες προκύπτουν επίσης νότες. Ας πάρουμε για παράδειγμα τη νότα Λα (διεθνώς γράφεται "A"). Η μεσαία Λα σε ένα πιάνο είναι στα 110 Hz.

110 Hz (110.00 Hz): **Λα (A)**

220 Hz (220.00 Hz): **Λα (A)**

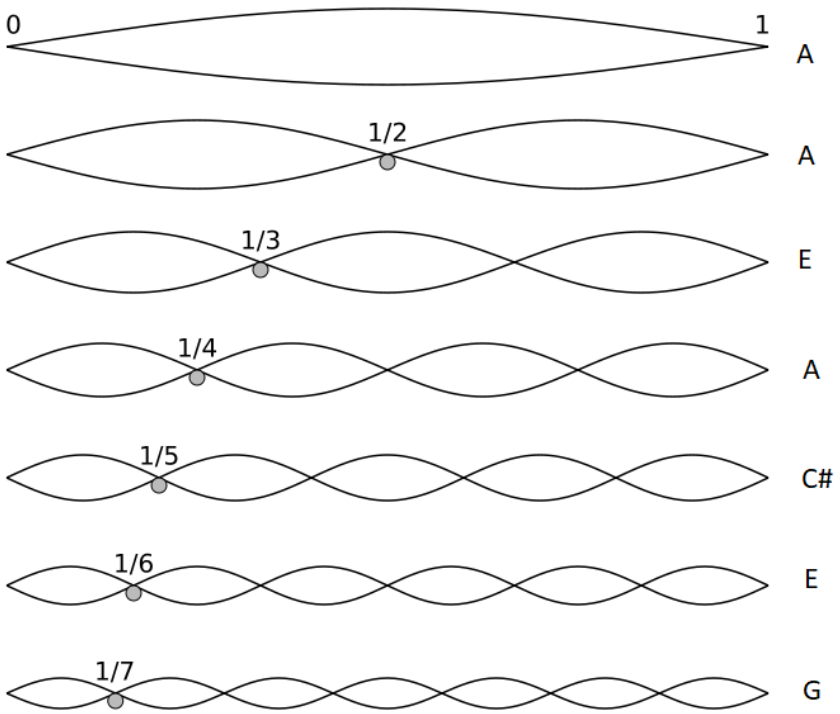
330 Hz (329.63 Hz): **Μι (E)**

440 Hz (440 Hz): **Λα (A)**

550Hz (554.57 Hz): **Ντο-δίεση (C#)**

660Hz (659.25 Hz): **Μι (E)**

Στις παρενθέσεις αναγράφονται οι πραγματικές τιμές της θεμελιώδους συχνότητας που θα μετρούσε κανείς αν άκουγε κάποιο μουσικό όργανο. Αυτές απέχουν από τις αρμονικές λόγω μοντέρνου κουρδίσματος (equal temperament). Σύμφωνα με το κούρδισμα αυτό, η αναλογία των συχνοτήτων δύο διαδοχικών νοτών είναι $\sqrt[12]{2} \approx 1.05946$ και έχει επιλεγεί ώστε να υπάρχει συμμετρία στα διαστήματα - ανεξαρτήτως τοποθεσίας στα πλήκτρα. Δηλαδή αν ηχούν δυο νότες που απέχουν Ν πλήκτρα στο πιάνο, τότε πρέπει η αναλογία των συχνοτήτων τους να είναι ανεξάρτητη των πλήκτρων καθ' αυτών και να εξαρτάται μόνο από την μεταξύ τους απόστασή.



Αρμονική σειρά με τονική νότα τη Λα (A)

Η θεωρία της αρμονίας βασίζεται στο γεγονός πως όταν ηχούν μαζί δυο "όμοιοι" τόνοι, τότε η προκύπτουσα κυματομορφή είναι απλή και την εκλαμβάνουμε ως εύηχη (συμφωνία - συνήχηση). Αντίθετα αν δυο "ανόμοιοι" τόνοι ηχούν μαζί τους εκλαμβάνουμε ως παραφωνία - ενώ η κυματομορφή είναι πιο πολύπλοκη.

2.3 Διαστήματα - Συγχορδίες

Δύο τόνοι χαρακτηρίζονται από το διάστημά τους (την απόστασή τους σε ημιτόνια - πόσα πλήκτρα απέχουν σε ένα πιάνο). Παρατηρώντας το παραπάνω παράδειγμα, μπορούμε να εξάγουμε τα 3 απλούστερα και πιο εύηχα διαστήματα:

Οκτάβα $(\frac{2}{1})$: Δυο Λα ακούγονται μαζί - δεν προκαλούν διαφορετικό συναίσθημα από μια Λα να ακούγεται μόνη της.

Πέμπτη $(\frac{3}{2})$: Μια Λα και μια Μι ακούγονται μαζί. Περιγράφεται ως εύηχο αλλά ουδέτερο συναισθηματικά διάστημα [3].

(ματζόρε) Τρίτη $(\frac{5}{4})$: Μια Λα και μια Ντο δίεση ακούγονται μαζί. Προκαλεί συναισθήματα ευτυχίας και μεγαλειότητας, ευστάθειας, φωτεινότητας.

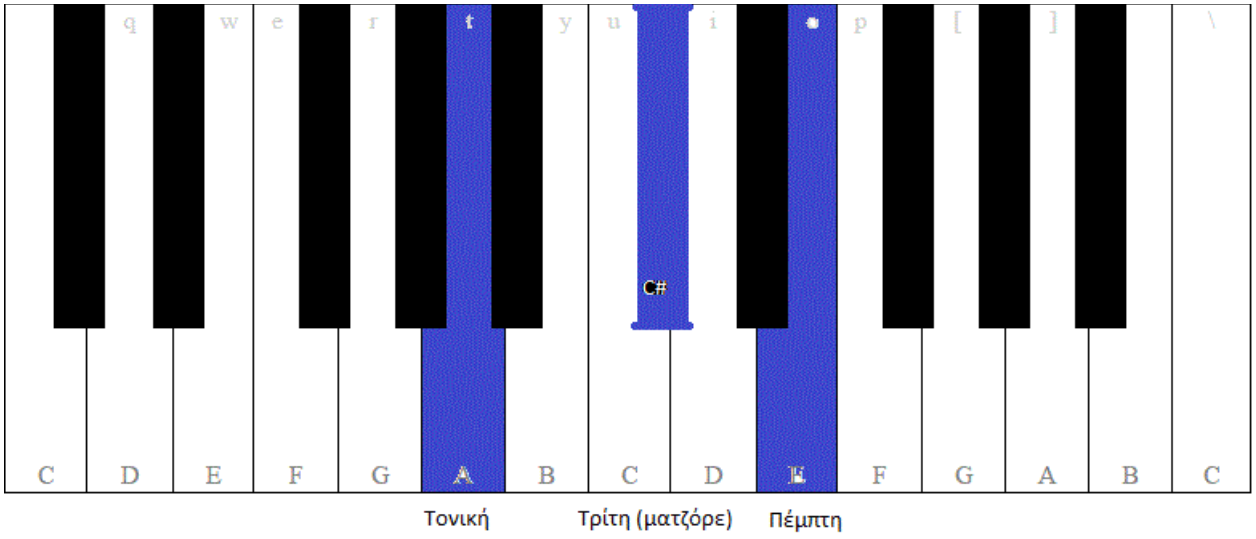
Η ομαδοποίηση δύο ή περισσότερων τόνων που ηχούν μαζί, ή κοντά χρονικά, ονομάζεται **συγχορδία**.

Όπως είναι προφανές υπάρχουν δυνητικά άπειρες συγχορδίες, αφού μπορεί να συνδυαστεί οποιοσδήποτε αριθμός τόνων για να συντεθεί. Σε κάθε περίπτωση όμως ο ανθρώπινος εγκέφαλος θα επιχειρεί να τις αποκωδικοποιήσει σε σχέση με έναν τόνο. Αυτός ο τόνος μπορεί να είναι είτε αυτός που ακούγεται πιο δυνατά, είτε ο πιο μπάσος, είτε ο αμέσως προηγούμενος, είτε οτιδήποτε άλλο τον έκανε να ξεχωρίσει . Ονομάζεται τονική νότα της συγχορδίας.

Έτσι κάθε συγχορδία μπορεί να χαρακτηριστεί αρχικά από την τονική της νότα. Στη μουσική σημειογραφία (πχ σε μια παρτιτούρα) κάθε συγχορδία θα αναγράφεται ξεκινώντας με την τονική της νότα. Για παράδειγμα όλες οι παρακάτω συγχορδίες έχουν τονική νότα τη Λα (A): *Amin, A7, Amaj7, A7(b9)...*

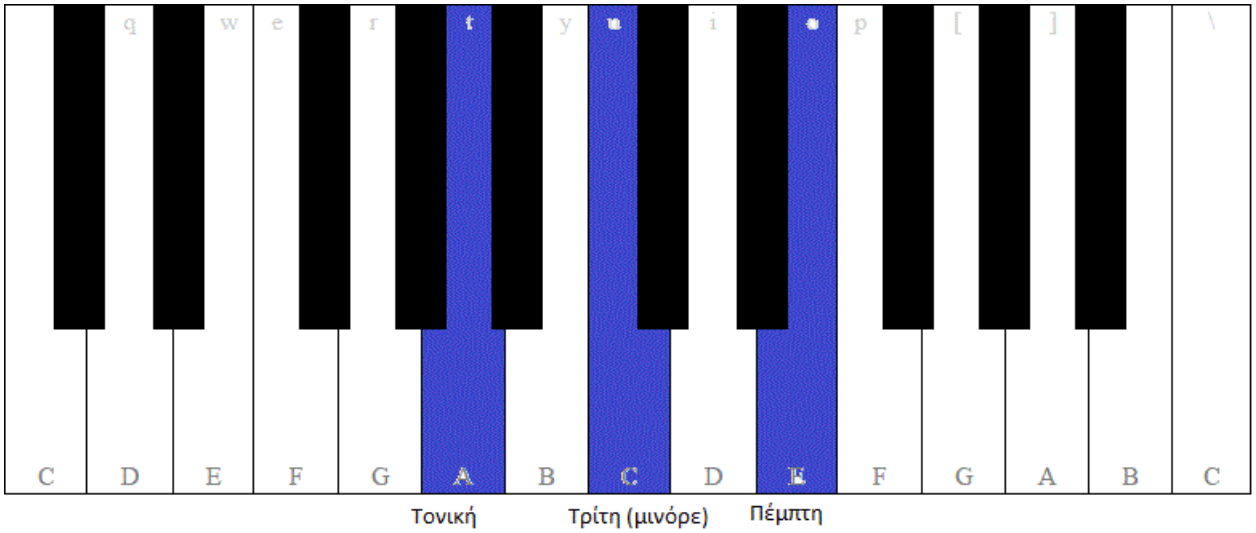
Το πρώτο έργο για ένα σύστημα αυτόματης αναγνώρισης συγχορδιών θα είναι άρα η αναγνώριση της τονικής νότας της συγχορδίας. Προκύπτει πρόβλημα κατηγοριοποίησης με 13 κατηγορίες (12 νότες + 1 κατηγορία για “όχι συγχορδία”).

Η επόμενη απλή κλάση συγχορδιών στηρίζεται στα παραπάνω διαστήματα (τα πρώτα της αρμονικής σειράς - τα πιο εύηχα). Ας ξεκινήσουμε με τονική τη Λα (A), την πέμπτη της: Μι (E) (3η στην αρμονική σειρά) και τη ματζόρε τρίτη της: Ντο-δίεση (C#) (5η στην αρμονική σειρά). Στα πλήκτρα ενός πιάνο:



Αυτό το σύνολο νοτών (με τονική τη Λα) συνθέτουν τη συγχορδία Λα ματζόρε ή σκέτο Λα (Amaj - A). Παρατηρούμε πως πέραν των διαστημάτων της ματζόρε τρίτης (A - C#) και της πέμπτης (A - E) που παρουσιάστηκαν παραπάνω, εμφανίζεται ένα ακόμα διάστημα (C# - E). Το διάστημα αυτό είναι ένα ημιτόνιο (ένα πλήκτρο) μικρότερο από τη ματζόρε τρίτη και ονομάζεται μινόρε τρίτη και είναι η 19η νότα της αρμονικής σειράς (με τονική C# στο παράδειγμα). Χαρακτηριστικό της μινόρε τρίτης είναι πως από μόνη της προκαλεί συναισθήματα μελαγχολίας, λύπης και στοϊκότητας. Επίσης είναι προφανές πως η μινόρε τρίτη της ματζόρε τρίτης μιας νότας είναι η πέμπτη της, αφού το διάστημα μινόρε τρίτης είναι 3 ημιτόνια, της ματζόρε 4 και της πέμπτης 7.

Χρησιμοποιώντας τα ίδια διαστήματα (τρίτες) μπορούμε να συνθέσουμε μία ακόμα συγχορδία αποτελούμενη από τρεις νότες, εναλλάσσοντας συμμετρικά τα δύο διαστήματα (τη μινόρε με τη ματζόρε τρίτη). Στα πλήκτρα του πιάνο:



Αυτή η συγχορδία ονομάζεται Λα μινόρε (Amin). Έχει την ίδια τονική με τη Λα ματζόρε (παραπάνω), το ίδιο πλήθος νοτών (3) και τα ίδια διαστήματα (μινόρε τρίτη, ματζόρε τρίτη, πέμπτη). Η διαφορά τους είναι η σειρά των διαστημάτων (πρώτα η μινόρε τρίτη, ύστερα η ματζόρε τρίτη). Παρά τις ομοιότητές τους όμως, ακούγονται εντελώς διαφορετικά και προκαλούν αντίθετα συναισθήματα (απλουστευμένα: λύπη - χαρά), και αυτό οφείλεται στο γεγονός ότι αντιλαμβανόμαστε διαστήματα με βάση την τονική.

Με αυτόν τον τρόπο μπορούμε να συνθέσουμε τις βασικότερες και τις πιο διαδεδομένες συγχορδίες της σύγχρονης δυτικής μουσικής, τις ματζόρε και μινόρε τριάδες (major and minor triads). Αυτές είναι 24 διαφορετικές συγχορδίες που χαρακτηρίζονται από την τονική τους (12) και το είδος της τρίτης σε σχέση με την τονική (μινόρε - ματζόρε).

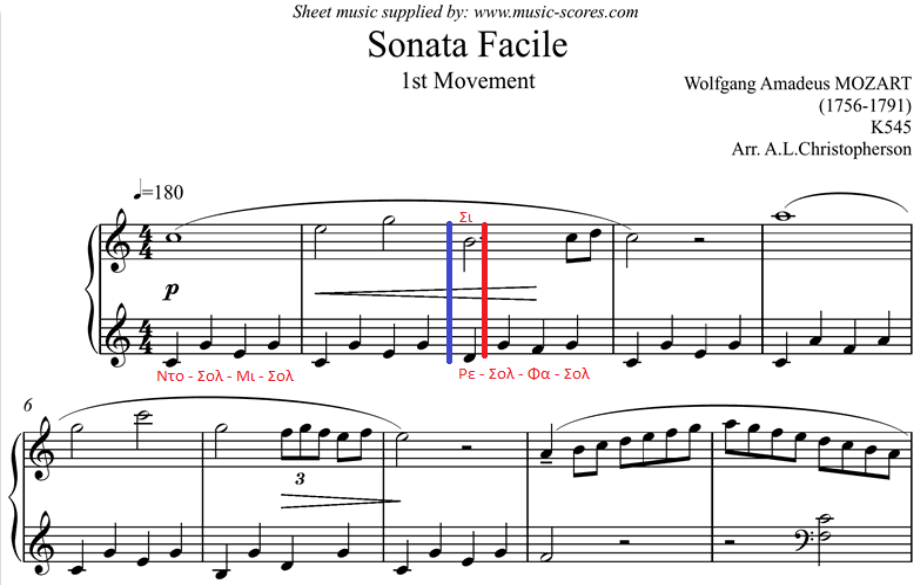
Ένα έργο λοιπόν για σύστημα αυτόματης αναγνώρισης συγχορδίας είναι να αναγνωρίζει την τρίτη της τονικής (καθώς και την τονική) και να κατηγοριοποιεί σε 25 κατηγορίες κάθε (πιθανή) συγχορδία.

Οι ιδιότητες του διαστήματος της τρίτης έχουν οδηγήσει στο μοντέλο της “Τριτογενούς Αρμονίας” (Tertiary Harmony), το οποίο στηρίζεται στη στοίβαξη ματζόρε και μινόρε τρίτων για τη δημιουργία συγχορδιών που έχουν νόημα. Στα

πλαίσια της εργασίας δεν θα ασχοληθούμε με αυτές παρά μόνο με την τονική και τις ματζόρε μινόρε τριάδες.

2.4 Δυσκολίες - Ασάφεια

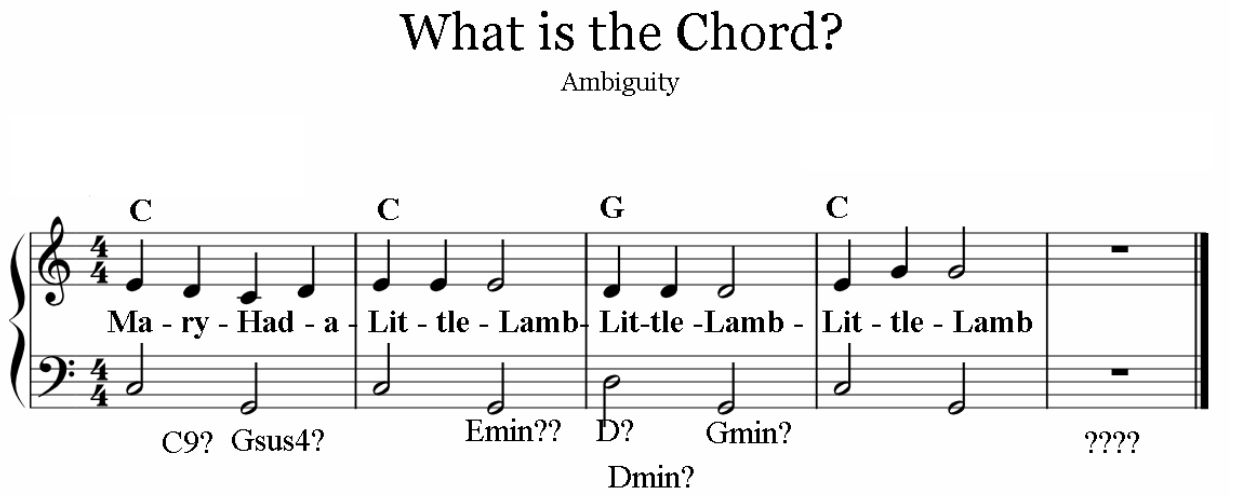
Τέλος, όταν λέμε πως σε κάποιο σημείο ενός κομματιού μουσικής η συγχορδία είναι Χ, αυτό δεν σημαίνει πως ηχούν όλες οι νότες της συγχορδίας. Μπορεί κάποιες νότες να εννοούνται από τα συμφραζόμενα, και η σχέση αυτή (σημείου με συμφραζόμενα) ενδέχεται να είναι πολύπλοκη. Η θεωρία της μουσικής επιχειρεί να εξηγήσει τις σχέσεις, αλλά είναι εκτός πεδίου εφαρμογής για την παρούσα εργασία. Σαν παράδειγμα μπορούμε να μελετήσουμε μια απλή σονάτα του Mozart, την K545 - Sonata Facile.



Στην παραπάνω παρτιτούρα αναγράφονται οι νότες που ακούγονται με τη σειρά που ακούγονται. Από την αρχή μέχρι και τη μπλε γραμμή (τα δύο πρώτα μέτρα) οι μόνες νότες που ακούγονται είναι Ντο, Μι, Σολ. Αν ελέγξουμε τις αποστάσεις των νοτών αυτών, με βάση τα παραπάνω παρατηρούμε πως η Μι απέχει από την Ντο 4 ημιτόνια (ματζόρε τρίτη) και η Σολ 7 (πέμπτη), άρα με βάση τον ορισμό μας οι νότες αυτές συνθέτουν την Ντο ματζόρε συγχορδία. Πράγματι, αυτή είναι η συγχορδία που ακούγεται. Τι συμβαίνει όμως από τη μπλε γραμμή και

ύστερα; Οι πρώτες νότες που ακούγονται (μέχρι την κόκκινη γραμμή) είναι η Ρε και η Σι. Οι νότες αυτές συνυπάρχουν σε πολλές συγχορδίες. Ακόμα και να μην υπήρχαν οι υπόλοιπες νότες (μετά την κόκκινη γραμμή), ένας μουσικός θα έλεγε πως η συγχορδία στο σημείο που ηχούν η Ρε και η Σι είναι η Σολ Ματζόρε, ενώ όμως δεν ηχεί η τονική της συγχορδίας (Σολ). Αυτό συμβαίνει για πολλούς λόγους, όπως: Η προηγούμενη νότα (που είναι Σολ, αλλά ηχεί σαν πέμπτη της Ντο ματζόρε αντί για τονική της Σολ ματζόρε), η κλίμακα του κομματιού, οι επόμενες νότες, συνηθέστερη αλληλουχία (η αλλαγή συγχορδίας από Ντο ματζόρε σε Σολ ματζόρε είναι από τις πιο διαδεδομένες) κ.α.

Από την παραπάνω απλουστευμένη ανάλυση ενός σημείου ενός απλού κομματιού, μπορεί να καταλάβει κανείς τον βαθμό της πολυπλοκότητας αλλά και ενδεχομένως της ασάφειας που θα κληθεί να επιλύσει ένα σύστημα αυτόματης αναγνώρισης συγχορδίας.



Επίσης, στην παραπάνω παρτιτούρα, οι σωστές συγχορδίες αναγράφονται στο πάνω μέρος, ενώ από κάτω αναγράφονται πιθανές συγχορδίες που θα προέκυπταν αν κάποιος λάμβανε υπ’ όψη μονάχα τις νότες που ακούγονταν στιγμιαία και όχι ολόκληρο το πλαίσιο.

Κεφάλαιο 3 - Αναγνώριση Συγχορδίας

3.1 Αναγνώριση Συγχορδίας - Άνθρωπος

Λίγοι άνθρωποι έχουν τη δυνατότητα να αναγνωρίζουν ακριβώς μια συγχορδία που ακούγεται. Συγκεκριμένα ένα μικρό ποσοστό του πληθυσμού έχει τη δυνατότητα να αναγνωρίζει την τονική νότα απλώς ακούγοντάς την (perfect pitch). Δεδομένης όμως της τονικής νότας, γίνεται εύκολο για τους περισσότερους να αναγνωρίζουν τους χρωματισμούς της συγχορδίας (πχ τρίτες, έβδομες) χωρίς μάλιστα να απαιτείται κάποια ιδιαίτερη μουσική παιδεία. Η αναγνώριση βασίζεται στη συνάφεια των διαφόρων διαστημάτων με διαφορετικά συναισθήματα.

Επιπρόσθετα, ένας ειδικός στον τομέα, μπορεί να εξάγει συμπεράσματα για την εκάστοτε συγχορδία διαβάζοντας μία παρτιτούρα ή ακούγοντας την συγχορδία σε κάποιο πλαίσιο (context). Με αυτόν τον τρόπο μπορεί να είναι πιο ακριβής στον προσδιορισμό των χρωματισμών, ειδικά σε πιο πολύπλοκες συγχορδίες όπου συγχέονται πολλά διαφορετικά διαστήματα.

3.2 Αναγνώριση Συγχορδίας από Μηχανή - Παρούσα Τεχνολογία (State-of-the-art)

Γενικά στον τομέα της ανάκτησης πληροφορίας από μουσική (Music Information Retrieval), μεγάλο μέρος της δουλειάς που απαιτείται έγκειται στην εξαγωγή χαρακτηριστικών που χρησιμοποιούνται κατά την πρόβλεψη. Η διαδικασία αυτή φαίνεται να έχει φτάσει στα όριά της χρησιμοποιώντας παραδοσιακές μεθόδους επεξεργασίας σήματος. Ερευνητές επομένως ψάχνουν για εναλλακτικές στην παραπάνω διαδικασία. Η καλύτερη εναλλακτική με την παρούσα τεχνολογία φαίνεται να είναι η βαθιά μηχανική μάθηση, αλλά υπάρχουν ακόμη προβλήματα που πρέπει να αντιμετωπιστούν [4].

Η αναγνώριση συγχορδίας από μηχανή στηρίζεται αρχικά στην εξαγωγή πληροφορίας για τις νότες που ακούγονται ανά πάσα στιγμή, και στη συνέχεια αξιοποίηση της πληροφορίας αυτής για τον καθορισμό της συγχορδίας. Κατά τους

καιρούς έχουν χρησιμοποιηθεί διάφορες μέθοδοι επεξεργασίας σήματος αλλά και μηχανικής μάθησης για τα παραπάνω στάδια [28,29,30]. Ενώ παλαιότερα επικρατούσε η χρήση στατιστικών μοντέλων στο δεύτερο στάδιο, όπως τα Hidden Markov Models (HMM) και τα Μπεϋζιανά Δίκτυα, πλέον χρησιμοποιούνται όλο και περισσότερο αρχιτεκτονικές νευρωνικών δικτύων, οι οποίες έχουν πλεονεκτήματα αλλά και μειονεκτήματα έναντι των πρώτων [32].

Μάλιστα, αξιοποιώντας την πρόοδο του τομέα της Βαθιάς Μηχανικής Μάθησης (Deep Learning) τα τελευταία χρόνια, αντικαθίστανται και τμήματα του πρώτου σταδίου (επεξεργασία σήματος - εξαγωγή χαρακτηριστικών) από νευρωνικά δίκτυα. Έτσι, τα αποδοτικότερα συστήματα αναγνώρισης συγχορδίας τείνουν να γίνουν end-to-end ή άκρο-σε-άκρο, όπου δεν υπάρχουν ενδιάμεσα στάδια επεξεργασίας από το σήμα μέχρι την πρόβλεψη. Παρ' όλα αυτά, σε αντίθεση με άλλους τομείς όπως η όραση υπολογιστών, δεν έχει επικρατήσει ακόμα η βαθιά μηχανική μάθηση στον τομέα της ανάκτησης πληροφορίας από μουσική.

Το πλεονέκτημα των βαθιών αρχιτεκτονικών μηχανικής μάθησης που χρησιμοποιούνται, όπως βαθιά συνελικτικά και αναδρομικά νευρωνικά δίκτυα, είναι πως έχουν τη δυνατότητα να μάθουν σε χαρακτηριστικά χαμηλότερου επιπέδου σε σχέση με παλαιότερες τεχνολογίες και έτσι μειώνεται σε μεγάλο βαθμό η προεπεξεργασία που απαιτούν τα δεδομένα [5] ενώ παράλληλα αυξάνεται η απόδοση.

Ενδεικτικά είναι τα αποτελέσματα του MIREX2017, στο οποίο την καλύτερη απόδοση σε όλες τις κατηγορίες εκτός από μία είχε το KBK2 [6], το οποίο χρησιμοποιεί συνελικτικό νευρωνικό δίκτυο για να μάθει τα χαρακτηριστικά που χρησιμοποιούνται κατά την πρόβλεψη. Στον ίδιο διαγωνισμό, το CM2 [7] το οποίο εξάγει χαρακτηριστικά με πιο παραδοσιακό τρόπο (chordino - nnls chroma), έχει τη χειρότερη απόδοση. Βέβαια εξακολουθεί να είναι χρήσιμη η "χειροκίνητη" εξαγωγή χαρακτηριστικών, αφού σε αντίθεση με τις βαθιές αρχιτεκτονικές νευρωνικών δικτύων, μας προσφέρουν πολύ καλύτερη επίγνωση του τρόπου λειτουργίας ενός συστήματος καθώς και κατανόηση των χαρακτηριστικών και της χρήσιμης πληροφορίας τους.

3.3 Διαφοροποίηση, Περιορισμοί και Πλεονεκτήματα Παρούσας Εργασίας

Στην παρούσα εργασία δεν ασχοληθήκαμε με εξαγωγή χαρακτηριστικών, αλλά αξιολογήσαμε διάφορα μοντέλα ως προς την απόδοσή τους στην πρόβλεψη συγχορδίας χρησιμοποιώντας τα χαρακτηριστικά που παρέχει το Spotify μέσω του API του.

Τα χαρακτηριστικά που παρέχει το Spotify, δεν προκύπτουν από μοντέλα μηχανικής μάθησης, αλλά με διαφορετικές μεθόδους επεξεργασίας σήματος (βλ. κεφάλαιο: Ανάλυση Ήχου - Spotify API). Περιοριζόμαστε άρα από την "ποιότητα" των χαρακτηριστικών αυτών και από την πληροφορία που ενδεχομένως να χάθηκε κατά την εξαγωγή τους.

Αυτή η προσέγγιση βέβαια έχει και πλεονεκτήματα. Από τη μία μπορούμε να εξάγουμε συμπεράσματα για τη συμβατότητα ηχητικών χαρακτηριστικών υψηλού επιπέδου με διαφορετικές αρχιτεκτονικές Νευρωνικών Δικτύων καθώς και να αξιολογήσουμε και να κατανοήσουμε καλύτερα τη λειτουργία τους. Τα χαρακτηριστικά που παρέχει το Spotify είναι πολύ πιο κατανοητά στον άνθρωπο από ότι ένα σήμα ή ένα φασματογράφημα.

Από την άλλη, η μορφή των δεδομένων είναι πιο διαχειρίσιμη και ευκόλως διαδιδόμενη, αφού αντί για ασυμπίεστα αρχεία .wav (~50 MB ανά κομμάτι) τα σύνολα δεδομένων αποτελούνται από αρχεία .csv (~200 KB ανά κομμάτι).

3.4 Αποτελέσματα

Όπως αναμενόταν, δεν καταφέραμε να φτάσουμε τα επίπεδα ακρίβειας της παρούσας τεχνολογίας. Αντ' αυτού, τα αποτελέσματα είναι χρήσιμα για την αξιολόγηση της συμπεριφοράς των μοντέλων, όταν η ποιότητα των δεδομένων είναι περιορισμένη. Αυτό μπορεί να συμβαίνει συχνά σε πραγματικές εφαρμογές, όπως εξάλλου θα ήταν η επιτυχής αναγνώριση συγχορδιών μέσω των χαρακτηριστικών του Spotify, καθώς η πλατφόρμα έχει εκατομμύρια χρήστες κάθε μέρα.

Κεφάλαιο 4 - Δεδομένα

Για την εκπαίδευση μοντέλων επιβλεπόμενης μάθησης (supervised learning), χρησιμοποιούνται *διανύσματα χαρακτηριστικών (features)* ως είσοδος, για να προβλεφθεί η επιθυμητή έξοδος. Η επιθυμητή έξοδος κωδικοποιείται επίσης σε διάνυσμα με κάθε θέση του να αντιστοιχεί σε μία κατηγορία (συγχορδία σε αυτήν την περίπτωση).

Για την εργασία, οι πίνακες χαρακτηριστικών κατασκευάστηκαν από τα δεδομένα που παρέχει το Spotify API, ενώ τα διανύσματα εξόδου από διάφορες πηγές επισημειωμένων με συγχορδίες δεδομένων.

4.1 Συλλογή Δεδομένων

4.1.1 Επισημειωμένα Δεδομένα

Υπάρχουν σύνολα δεδομένων επισημειωμένα με συγχορδίες που χρησιμοποιούνται ευρέως στην περιοχή της αυτόματης αναγνώρισης συγχορδίας [8,9]. Αποτελούνται από αρχεία της μορφής:

Roxanne - The Police, *from RWC uspoplabels*

Start Time	End Time	Chord
0.000	0.245	G:min
0.245	8.893	G:min
8.893	10.595	F:maj
...	...	...
341.914	370.000	N

Δηλαδή στο συγκεκριμένο κομμάτι, για κάθε τμήμα από <start time,end time> αντιστοιχεί η συγχορδία <Chord>
Χρησιμοποιήθηκαν τελικά τα:

1) McGill Billboard (http://ddmal.music.mcgill.ca/research/billboard)
2) Isophonics (http://www.isophonics.net/content/reference-annotations)
3) uspopLabels (https://github.com/tmc323/Chord-Annotations/tree/master/uspopLabels)
4) Robbie Williams Annotations (https://www.researchgate.net/publication/260399240_Chord_and_Harmony_annotations_of_the_first_five_albums_by_Robbie_Williams)

Σε αυτά περιέχονται επισημειώσεις για 1102 κομμάτια (ενδεχομένως να υπάρχουν επικαλύψεις).

4.1.2 Ανάλυση Ήχου - Spotify API

Οι πίνακες χαρακτηριστικών κατασκευάστηκαν από τα δεδομένα που παρέχει το Spotify API με την κλήση του Get Audio Analysis for Track.
Αυτή επιστρέφει ένα αντικείμενο της μορφής:

```
{
    "bars": "[...]",
    "beats": "[...]",
    "meta": "{...}",
    "sections": "[...]",
    "segments": "[...]",
    "tatums": "[...]",
    "track": "{...}"
}
```

Τα αντικείμενα που ορίζονται ως “[...]” αντιστοιχούν σε τμηματοποιήσεις του κομματιού μουσικής, με διαφορετικά κριτήρια. Τα “sections” επιχειρούν να χωρίσουν το κομμάτι με κριτήριο τη δομή (πχ κουπλέ, ρεφρέν), τα “tatums” και “beats” με κριτήριο το ρυθμό. Λεπτομερέστερα για το κάθε αντικείμενο στο documentation της εταιρίας που παρέχει τις υπηρεσίες της στο Spotify [10].

Η πιο αναλυτική από τις παραπάνω τμηματοποιήσεις είναι η τμηματοποίηση σε segments. Αυτή χρησιμοποιήθηκε για τη δημιουργία των πινάκων χαρακτηριστικών.

Συγκεκριμένα ένα αντικείμενο *segments* είναι μια λίστα από αντικείμενα της μορφής:

```
"segments": [
  {
    "start": "<float>",
    "duration": "<float>",
    "confidence": "<float>",
    "loudness_start": "<float>",
    "loudness_max_time": "<float>",
    "loudness_max": "<float>",
    "loudness_end": "<float>",
    "pitches": "<float[12]>",
    "timbre": "<float[12]>",
  }
 ],
```

Όπου:

- **"start"**: Χρόνος έναρξης του segment σε σχέση με την αρχή του κομματιού σε δευτερόλεπτα. (float >= 0.0)
- **"duration"**: Διάρκεια του segment σε δευτερόλεπτα. (float usually <1.0)
- **"confidence"**: Αξιοπιστία της τμηματοποίησης (float < 1.0)
- **"Loudness_start"**: Ένταση του ήχου στην αρχή του segment (float)
- **"Loudness_max_time"**: Σημείο μέγιστης έντασης στο segment σε σχέση με την αρχή του. (float < "duration")
- **"loudness_max"**: Η μέγιστη ένταση του segment. (float)
- **"loudness_end"**: Η ένταση στο τέλος του κομματιού (μόνο για το τελευταίο segment). (float)
- **"pitches"**: Chroma vector [11] - Σχετική επικράτηση της κάθε νότας (12 νότες) σε σχέση με την επικρατούσα. Στην επικρατούσα ανατίθεται η τιμή 1.0. Ο πίνακας pitches περιέχει 12 θέσεις (μία για κάθε νότα) ξεκινόντας από την Ντο στη θέση 0, Ντο δίεση στη θέση 1 κλπ.

- **"timbre"**: Διάνυσμα που περιέχει 12 υψηλού επιπέδου χαρακτηριστικά της φασματικής επιφάνειας - περιγράφουν ιδιότητες του ήχου όπως μπάσο-πρίμο, ατάκα, ένταση κ.α. Προκύπτουν από εφαρμογή φίλτρων στο φασματογράφημα του segment. (βλ. Documentation [10]).

Συνολικά για κάθε segment προκύπτουν 29 διαφορετικά χαρακτηριστικά. Μπορούμε να προσθέσουμε ως χαρακτηριστικά πληροφορία από τις υπόλοιπες τμηματοποιήσεις (όπως π.χ. Σε ποιο section ανήκει το segment). Με αυτόν τον τρόπο προκύπτουν 47 χαρακτηριστικά. Συγκεκριμένα:

- 29 segment features
- New Bar (1)
- New beat (1)
- New Section (1)
- Section Key (12)
- Key Confidence (1)
- Section Mode (1)
- Mode Confidence (1)

Συνδυάζοντας τα επισημειωμένα δεδομένα με τα χαρακτηριστικά ανατίθεται σε κάθε segment μία συγχορδία. Ο τρόπος που συνδυάστηκαν είναι ο εξής:

Εάν το segment είναι εξ'ολοκλήρου μέσα σε κάποιο διάστημα επισημειωμένο με μία συγχορδία, τότε ανατίθεται σε αυτό αυτή η συγχορδία. Εάν ένα segment περιέχεται σε παραπάνω από ένα διαστήματα επισημειωμένα με συγχορδίες, τότε ανατίθεται σε αυτό η συγχορδία του διαστήματος με τη μεγαλύτερη επικάλυψη.

Το κάθε σύστημα εκπαιδεύτηκε ώστε να προβλέπει τη συγχορδία που αντιστοιχεί στο segment δεδομένων των χαρακτηριστικών του.

4.2 Προεπεξεργασία Δεδομένων

4.2.1 Επιλογή Χαρακτηριστικών

Τα περισσότερα πειράματα έγιναν με διάφορους συνδυασμούς των 29 χαρακτηριστικών του segment. Έγιναν και κάποια με τα 47 συνολικά χαρακτηριστικά όλων των τμηματοποιήσεων. Για την εκτέλεση, τα δεδομένα οργανώθηκαν διαφορετικά για κάθε είδος Νευρωνικού Δικτύου που εκπαιδεύτηκε, συγκεκριμένα:

4.2.1α) Feedforward - Multi Layer Perceptron

Επειδή η συγχορδία εξαρτάται από τα "μουσικά συμφραζόμενα" (βλ. Θεωρία Μουσικής), και η αρχιτεκτονική ενός feedforward νευρωνικού δικτύου "βλέπει" ανεξάρτητα κάθε παράδειγμα εισόδου, σε κάθε segment προστέθηκαν τα χαρακτηριστικά των N προηγούμενων και N επόμενων segments. Εκτελέστηκαν πειράματα με διάφορες τιμές του N.

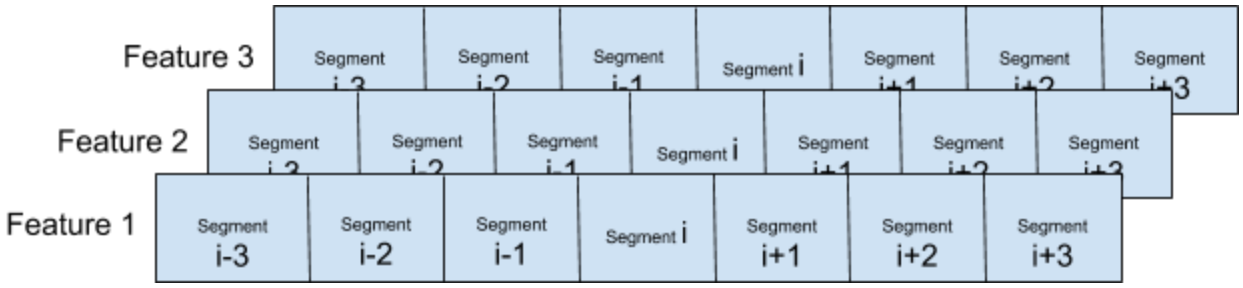
Για παράδειγμα αν είχαν επιλεγεί 24 από τα 29 χαρακτηριστικά και N=3, τότε ο πίνακας χαρακτηριστικών θα είχε $24 + 2 * 3 * 24 = 162$ στήλες /χαρακτηριστικά για κάθε segment.

4.2.1β) Convolutional - Συνελικτικό

a) Συνέλιξη στο πεδίο του χρόνου

Τα δεδομένα οργανώνονται ως όγκοι (3-διάστατοι πίνακες) με ύψος 1, πλάτος N (από πόσα segments εξάγεται η συγχορδία) και βάθος τον αριθμό των χαρακτηριστικών. Αυτά ύστερα συνελίσσονται με οριζόντια φίλτρα, τα οποία

εξάγουν ουσιαστικά πληροφορία για την εξέλιξη στο χρόνο του κάθε χαρακτηριστικού (τα segments αποτελούν χρονική ακολουθία)



b) Συνέλιξη στα πεδία του χρόνου και των *pitches*

Τα δεδομένα οργανώνονται ως επιφάνειες (διδιάστατοι πίνακες) με ύψος N (segments) και πλάτος 12 (pitches). Με αυτόν τον τρόπο, τα φίλτρα ενός συνελικτικού επιπέδου (convolutional layer) λαμβάνουν πληροφορία και για τις νότες που ακούγονται (pitches) αλλά και την εξέλιξή τους στο χρόνο.

Μία από αυτές τις επιφάνειες (ένα παράδειγμα εισόδου) θα είχε τη μορφή (N=3):

Segment i-3	C	C#	D	D#	E	F	F#	G	G#	A	A#	B
Segment i-2	C	C#	D	D#	E	F	F#	G	G#	A	A#	B
Segment i-1	C	C#	D	D#	E	F	F#	G	G#	A	A#	B
Segment i	C	C#	D	D#	E	F	F#	G	G#	A	A#	B
Segment i+1	C	C#	D	D#	E	F	F#	G	G#	A	A#	B
Segment i+2	C	C#	D	D#	E	F	F#	G	G#	A	A#	B
Segment i+3	C	C#	D	D#	E	F	F#	G	G#	A	A#	B

Τα υπόλοιπα χαρακτηριστικά πέραν των pitches μπορούν να κρατηθούν και να ενωθούν με τα αποτελέσματα της συνέλιξης σε κάποιο από τα *fully connected layers* ενός συνελικτικού νευρωνικού δικτύου.

4.2.1γ) LSTM - bLSTM - Encoder - Decoder LSTM

Ένα αναδρομικό νευρωνικό δίκτυο, όπως η παραλλαγή τού: LSTM, παίρνει σαν είσοδο μια ταξινομημένη αλληλουχία δεδομένων. Χρησιμοποιήθηκαν δύο διαφορετικές μεθοδολογίες.

a) Κατηγοριοποίηση **μιας αλληλουχίας** από N segments σε **μία συγχορδία**.

Σε αυτή την περίπτωση τα δεδομένα χωρίστηκαν σε αλληλουχίες στις οποίες αντιστοιχεί μία συγχορδία. Επιλέγεται κάποιο N και στις αλληλουχίες με περισσότερα των N segments απορρίπτονται τα επιπλέον των N segments, ενώ στις αλληλουχίες με λιγότερα των N προστίθενται στο τέλος της segments με τιμή 0 σε κάθε χαρακτηριστικό (zero padding), έτσι προκύπτει μια λίστα από ακολουθίες μήκους N στην κάθε μία από τις οποίες αντιστοιχεί μια συγχορδία.

b) Κατηγοριοποίηση **κάθε segment** από μια αλληλουχία σε **μία συγχορδία.**

Σε αυτή την περίπτωση τα δεδομένα χωρίζονται σε αλληλουχίες ίδιου μήκους. Το κάθε αναδρομικό δίκτυο παράγει πρόβλεψη για κάθε βήμα της ακολουθίας.

4.2.2 Επιλογή Κατηγοριών

Επειδή το πλήθος των πιθανών συγχορδιών είναι ουσιαστικά άπειρο (βλ. Θεωρία Μουσικής), περιοριστήκαμε στις δύο παρακάτω ομαδοποιήσεις.

a) **Τονική:** Οι συγχορδίες ομαδοποιήθηκαν με βάση την τονική τους νότα. Προκύπτουν με αυτόν τον τρόπο 13 διαφορετικές κατηγορίες (μία για κάθε νότα (12) + μία “no chord” κατηγορία).

b) **Majmin:** Οι συγχορδίες ομαδοποιήθηκαν με βάση την τονική τους νότα και το είδος της τρίτης τους (ματζόρε/μινόρε). Προκύπτουν 25 κατηγορίες (δύο για κάθε νότα (24) + μία “no chord” κατηγορία).

Περισσότερη βάση δόθηκε στη Majmin κατηγοριοποίηση καθώς από τη μία περιέχει την κατηγοριοποίηση στην Τονική, και από την άλλη είναι η πιο διαδεδομένη από τις κατηγοριοποιήσεις συγχορδιών που χρησιμοποιούνται στη σύγχρονη δυτική μουσική.

Πέραν των παραπάνω ομαδοποιήσεων, χρησιμοποιούνται συχνά και η ομαδοποίηση ανάλογα με την τρίτη και την έβδομη (49 κατηγορίες) καθώς και την πέμπτη (145 κατηγορίες). (MIREX audio chord estimation, http://www.music-ir.org/mirex/wiki/2017:Audio_Chord_Estimation_Results).

Λεπτομερέστερα για τις συγχορδίες και τα διαστήματα (τρίτη, πέμπτη, έβδομη) και ορολογία στο κεφάλαιο Θεωρία Μουσικής.

4.2.3 Κανονικοποίηση

Στις περισσότερες περιπτώσεις, από κάθε στοιχείο κάθε πίνακα αφαιρέθηκε η μέση τιμή της στήλης και κάθε στοιχείο διαιρέθηκε με την τυπική απόκλιση της στήλης (standardization). Έγιναν και πειράματα όπου οι τιμές των pitches κανονικοποιήθηκαν μόνο ως προς τις τιμές των pitches του ίδιου segment. Όπως και οι τιμες των timbre μόνο ως προς τις τιμές των timbre του ίδιου segment. Με αυτόν τον τρόπο όλα τα χαρακτηριστικά καταλήγουν να έχουν μέση τιμή 0 και τυπική απόκλιση 1. Αυτό βοηθά την εκπαίδευση ενός συστήματος μηχανικής μάθησης, καθώς οι συναρτήσεις ενεργοποίησης των νευρώνων είναι ευαίσθητες στο μέτρο της εισόδου καθώς και στην κατανομή της. Οι μέσες τιμές και οι αποκλίσεις των χαρακτηριστικών κρατώνται ως παράμετροι και χρησιμοποιούνται κατά την πρόβλεψη.

4.2.4 Επαύξηση (augmentation)

Αξιοποιώντας τη συμμετρία του ορισμού των νοτών και των συγχορδιών (βλ. Θεωρία Μουσικής) δίνεται η δυνατότητα να δημιουργήσουμε παραπάνω δεδομένα με τον εξής τρόπο:

Μεταφέρουμε κάθε τιμή του πίνακα **"pitches"** μια θέση δεξιά (και την τιμή της τελευταίας θέσης στην πρώτη) και μετατοπίζουμε τη συγχορδία προς τα πάνω κατά ένα ημιτόνιο.

Πχ. Από το segment:

p[0]	p[1]	p[2]	p[3]	p[4]	p[5]	p[6]	p[7]	p[8]	p[9]	p[10]	p[11]	f[..]	chord
0.624	0.758	0.825	0.548	0.699	0.661	0.693	**1.0**	0.688	0.71	0.656	0.585	...	**Gmin**

Προκύπτει το:

p[0]	p[1]	p[2]	p[3]	p[4]	p[5]	p[6]	p[7]	p[8]	p[9]	p[10]	p[11]	f[..]	chord
0.585	0.624	0.758	0.825	0.548	0.699	0.661	0.693	**1.0**	0.688	0.71	0.656	...	**G#min**

Όπου *p* είναι το διάνυσμα *"pitches", f[...]* είναι τα υπόλοιπα χαρακτηριστικά και *chord* είναι η συγχορδία που αντιστοιχεί στο segment.

Με αυτόν τον τρόπο προκύπτουν έγκυρα δεδομένα και το συνολικό πλήθος τους πολλαπλασιάζεται επί 12 (επαναλαμβάνοντας την παραπάνω διαδικάσια 11 φορές).

4.2.5 Φίλτρα

Επειδή στο σύνολο δεδομένων κάποιες συγχορδίες εμφανίζονται πολύ σπάνια (πχ. *G#min* εμφανίζεται σε *7,918* segments), ενώ άλλες πολύ συχνά (πχ. *Gmaj* σε *88,225* segments), δίνεται η δυνατότητα να επιλεγούν τυχαία ίδιο πλήθος παραδειγμάτων από κάθε συγχορδία ώστε να αποφευχθεί η πόλωση. Στην περίπτωση της *majmin* κατηγοριοποίησης (25 κατηγορίες) μπορούν να επιλεγούν

ίδιο πλήθος σε όλες τις maj συγχορδίες και ίδιο σε όλες τις min (διαφορετικά μεταξύ τους), ή ίδιο πλήθος για όλες τις κατηγορίες (ίδιο για maj kai min)

Ένα τυπικό train set (60% του συνόλου δεδομένων) έχει την εξής κατανομή σε κατηγορίες:

No chord	17062
Amin	16797
A	47797
A#min	3431
A#	23097
Bmin	12417
B	19454
Cmin	6778
C	45707
C#min	7853
C#	13252
Dmin	13660
D	48985
D#min	3658
D#	18299
Emin	14205
E	38713
Fmin	5466
F	30886
F#min	8021
F#	15522
Gmin	9218
G	50228
G#min	3553
G#	18134

Για 462193 τμήματα (segments) (60% αρχικού συνόλου).

Αν φιλτραριστούν, με την κατηγορία με τα λιγότερα τμήματα (3,431), προκύπτουν $3431 * 25 =$ 85,775 τμήματα. Εάν στη συνέχεια γίνει η επαύξηση,

προκύπτουν $85775 * 12 = 1,029,300$ τμήματα. Δηλαδή σχεδόν διπλάσιο μέγεθος από το αρχικό σύνολο εκπαίδευσης, στο οποίο τα παραδείγματα είναι ισοκατανεμημένα στις κατηγορίες.

Κεφάλαιο 5 - Προετοιμασία Πειραμάτων - Υπερπαράμετροι

Για την εκπαίδευση όλων των μοντέλων, τα δεδομένα χωρίστηκαν σε 3 σύνολα.

a) Το σύνολο εκπαίδευσης (train set), το οποίο χρησιμοποιείται για τον καθορισμό των παραμέτρων του εκάστοτε μοντέλου. Περιέχει από 60% εώς 70% του συνόλου όλων των δεδομένων.
b) Το σύνολο ανάπτυξης (dev set), το οποίο περιέχει δεδομένα **από την ίδια κατανομή** με το σύνολο εκπαίδευσης (μπορεί να περιέχει segments από ίδια κομμάτια, κομμάτια από ίδιους δίσκους, δίσκους από ίδια σύνολα δεδομένων). Χρησιμοποιείται για αξιολόγηση του μοντέλου καθώς και της διασποράς των δεδομένων.
c) Το σύνολο δοκιμής (test set), το οποίο περιέχει δεδομένα **από διαφορετική κατανομή** με το σύνολο εκπαίδευσης (περιέχει segments από διαφορετικά κομμάτια, κομμάτια από διαφορετικούς δίσκους, δίσκους από διαφορετικά σύνολα δεδομένων). Χρησιμοποιείται για αξιολόγηση των μοντέλων καθώς και της πόλωσης των δεδομένων.

Ο χωρισμός του συνόλου δεδομένων στα δύο πρώτα υποσύνολα (train - dev) γίνεται τυχαία πριν από κάθε εκτέλεση. Το Test Set χρησιμοποιήθηκε το ίδιο για όλα τα πειράματα ώστε τα μοντέλα να είναι συγκρίσιμα σε κάποιο βαθμό και να επιχειρηθεί βελτιστοποίηση. Το Test Set αποτελούταν από 331 κομμάτια από το αρχικό σύνολο.

Για κάθε μοντέλο επιλέχθηκαν υπερ-παράμετροι της εκτέλεσης με διάφορα κριτήρια. Για τις γενικότερες υπερ-παραμέτρους, που εφαρμόζονται σε όλα τα μοντέλα αναφέρονται παρακάτω οι επιλογές που έγιναν.

5.1 Συναρτήσεις Ενεργοποίησης

Μία από τις υπερ-παραμέτρους που πρέπει να τεθούν για τα διάφορα μοντέλα είναι οι συναρτήσεις ενεργοποίησης. Αυτές συνήθως επηρεάζουν την

ταχύτητα σύγκλισης του μοντέλου, ενώ κάποιες φορές επηρεάζουν και την απόδοση. Οι συναρτήσεις ενεργοποίησης καθορίζουν την έξοδο του νευρώνα, πρέπει να είναι μη γραμμικές και να έχουν εύκολα υπολογίσιμες παραγώγους. Αυτό είναι σημαντικό για τον αλγόριθμο Backpropagation που αναλύεται σε επόμενο κεφάλαιο.

Επιλέχθηκαν συναρτήσεις κυρίως με βάση τη βιβλιογραφία, αλλά έγιναν και πειράματα εναλλάσσοντάς τες. Συγκεκριμένα, στην περίπτωση πλήρως συνδεμένων επιπέδων (όπως σε ένα Feedforward) ως συνάρτηση ενεργοποίησης χρησιμοποιήθηκε η κυρίως η ReLU [12]. Στο τελευταίο επίπεδο χρησιμοποιείται Softmax, που ενδείκνυται για κατηγοριοποίηση σε πάνω από 2 κατηγορίες, καθώς οι έξοδοι αντιπροσωπεύουν πιθανότητες (τιμές<1, άθροισμα 1). Σε άλλου τύπου επίπεδα όπως το LSTM χρησιμοποιήθηκαν και η sigmoid και tanh.

5.2 Συναρτήσεις Κόστους

Η συνάρτηση κόστους στο πλαίσιο των νευρωνικών δικτύων είναι η συνάρτηση που επιχειρούμε να ελαχιστοποιήσουμε. Εκφράζει μια μορφή απόστασης των προβλέψεων του συστήματος από τις πραγματικές τιμές. Χρησιμοποιήθηκε η κατηγορηματική διασταυρούμενη εντροπία (categorical cross - entropy) ή λογαριθμική απώλεια, η οποία εφαρμόζεται στις εξόδους του τελευταίου επιπέδου κάθε μοντέλου που έχει συνάρτηση ενεργοποίησης Softmax. Η τιμή της διασταυρούμενης εντροπίας δίνεται από τον τύπο:

$$Cost(y,\hat{y}) = -\frac{1}{N}\sum_{n=1}^{N}[y_n log(\hat{y}_n) + (1-y_n) log(1-\hat{y}_n)]$$

Όπου N είναι το σύνολο των παραδειγμάτων εισόδου, $\hat{y}$ είναι η πρόβλεψη και y είναι η πραγματική τιμή. Είναι σημαντικό οι είσοδοι να αντιπροσωπεύουν πιθανότητες (έξοδος Softmax), καθώς η έννοια της εντροπίας και η λειτουργία της συνάρτησης κόστους στηρίζονται στη θεωρία πιθανοτήτων.

5.3 Βελτιστοποιητές

Ένα σημαντικό κομμάτι των τεχνητών νευρωνικών δικτύων είναι ο αλγόριθμος βελτιστοποίησης. Αυτός είναι υπεύθυνος για την εξέλιξη των παραμέτρων του μοντέλου κατά τη διάρκεια της εκπαίδευσης. Στο πλαίσιο των νευρωνικών δικτύων και του αλγορίθμου backpropagation, ελαχιστοποιούν τη συνάρτηση κόστους. Για τη βελτιστοποίηση εφαρμόστηκαν διάφορες παραλλαγές του αλγορίθμου κατάβασης βαθμίδας (Stochastic/Minibatch Gradient Descent με ορμή, ή με απόσβεση ρυθμού μάθησης, ή η παραλλαγή Adam [13]). Ο αλγόριθμος κατάβασης βαθμίδας στηρίζεται στον υπολογισμό των παραγώγων της συνάρτησης κόστους ως προς τις παραμέτρους του μοντέλου και την αλλαγή των τιμών τους προς την κατεύθυνση αρνητικής κλίσης (ελαχιστοποίηση). Για αυτό ονομάζεται αλγόριθμος “κατάβασης”.

Οι υπερπαράμετροι του βελτιστοποιητή θέτονταν κάθε φορά ανά περίσταση: το minibatch size ανάλογα με το μέγεθος του μοντέλου και των δεδομένων για αποδοτικότερη εκπαίδευση, και για τις υπόλοιπες υπερπαραμέτρους (ρυθμός μάθησης, *β κλπ.)* επιχειρήθηκε βελτιστοποίηση όπου ενδεικνυόταν.

5.4 Συστηματοποίηση - Regularization

Σε σύνθετα μοντέλα μηχανικής μάθησης ένας κίνδυνος είναι το *overfitting*. Ορίζεται ως η κατάσταση κατά την οποία το μοντέλο έχει συνεχώς αυξανόμενη απόδοση στο σύνολο εκπαίδευσης (train set) αλλά μειούμενη στα άλλα σύνολα δεδομένων. Δηλαδή το μοντέλο χάνει την ικανότητα γενίκευσης. Συμβαίνει συνήθως εξ’αιτίας της πολυπλοκότητας του συστήματος, η οποία οδηγεί στην εκμάθηση χαρακτηριστικών-θορύβου που είναι αποκλειστικά στο σύνολο εκπαίδευσης. Το σύστημα μετατρέπεται έτσι σε μια μορφή μνήμης που “θυμάται” παραδείγματα του συνόλου εκπαίδευσης, χωρίς να έχει μάθει να γενικεύει και να προβλέπει την έξοδο παραδειγμάτων που δεν έχει δει στο παρελθόν.

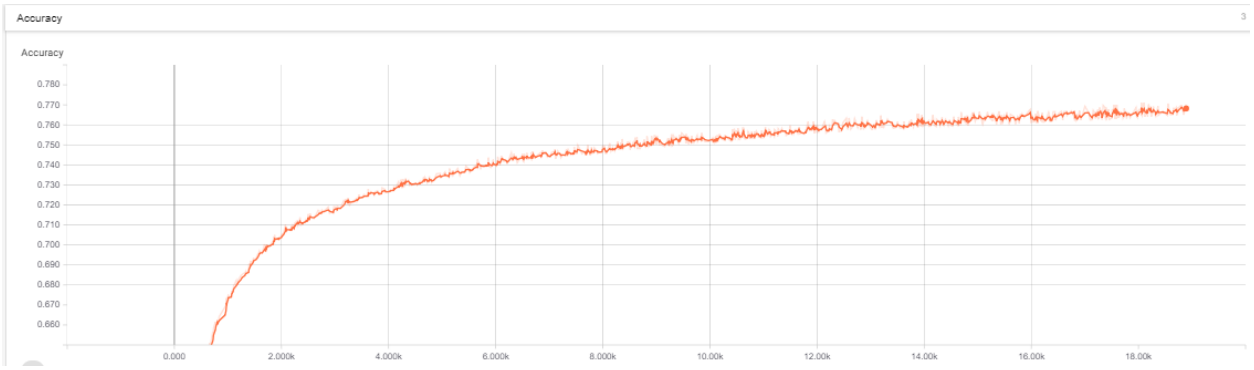
Ο τρόπος που αντιμετωπίζεται το overfitting, (πέραν της χρήσης απλούστερων μοντέλων - περισσότερων δεδομένων) είναι το Regularization ή

συστηματοποίηση. Υπάρχουν πολλές μέθοδοι συστηματοποίησης. Δύο διαδεδομένες που χρησιμοποιήθηκαν στην παρούσα εργασία είναι η L2 regularization και η Dropout.

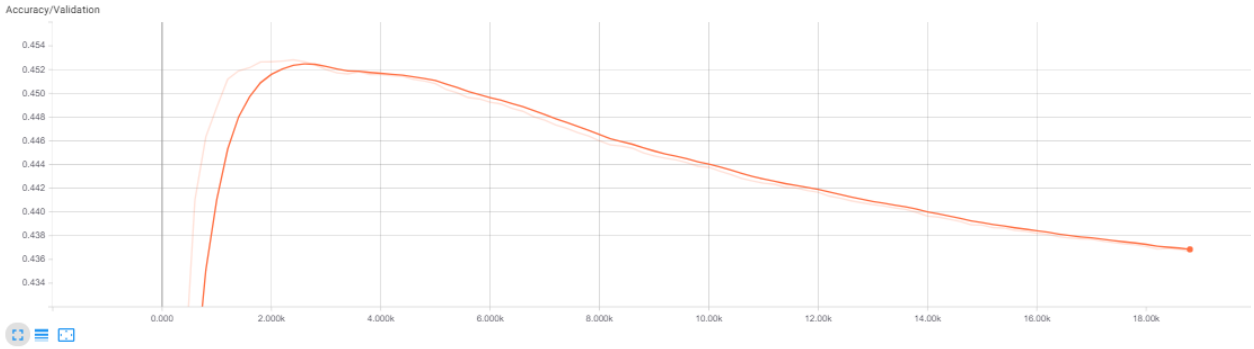
Η L2 συστηματοποίηση ή Lasso Regularization υλοποιείται ως πρόσθεση ενός όρου στη συνάρτηση κόστους. Ο όρος αυτός είναι ανάλογος του τετραγώνου του μέτρου των παραμέτρων. Έτσι αν μια παράμετρος (πχ το βάρος W ενός νευρώνα) τείνει να αυξηθεί πολύ σε σχέση με άλλες παραμέτρους, θα αυξηθεί το κόστος χάρη στον όρο που προστέθηκε και θα αντισταθμιστεί η αύξηση της τιμής της παραμέτρου. Έτσι δεν προκύπτουν νευρώνες που είναι πιο "δυνατοί" από άλλους, και έτσι αντιμετωπίζεται το overfitting, καθώς είναι δυσκολότερο για πλήθος νευρώνων να συν-εξελιχθούν σε χαρακτηριστικά - θόρυβο του συνόλου εκπαίδευσης.

Η Dropout [14] είναι μια μέθοδος που χρησιμοποιείται κυρίως σε βαθύτερα νευρωνικά δίκτυα (με πολλά κρυφά επίπεδα). Ο τρόπος που λειτουργεί είναι σε κάθε βήμα εκπαίδευσης, απενεργοποιούνται τυχαία διαφορετικοί νευρώνες (μαζί με τις συνδέσεις τους) ενώ ταυτόχρονα κλιμακώνονται οι έξοδοι τους (ώστε να διατηρείται η κατανομή των εξόδων. Για παράδειγμα, αν απενεργοποιούνταν οι μισοί νευρώνες (Dropout = 0.5), τότε το μέτρο των εξόδων των υπόλοιπων νευρώνων διπλασιάζονται. Με αυτόν τον τρόπο ουσιαστικά εκπαιδεύονται ταυτόχρονα πολλά διαφορετικά δίκτυα (το αρχικό δίκτυο με τυχαίους νευρώνες απενεργοποιημένους - διαφορετικούς σε κάθε βήμα) και κατά την πρόβλεψη που όλοι οι νευρώνες είναι ενεργοποιημένοι λαμβάνεται ο μέσος όρος των προβλέψεων των πολλών τυχαίων υποδικτύων.

Ένα από τα από τα απλούστερα feedforward μοντέλα που εκπαιδεύτηκε σε μικρό σύνολο δεδομένων (filtered - 5000 segments ανά κατηγορία) μπορεί να αποτελέσει παράδειγμα του φαινομένου του overfitting.



Από πάνω φαίνεται η εξέλιξη της ακρίβειας στο σύνολο εκπαίδευσης του μοντέλου. Εκ πρώτης όψεως φαίνεται να πηγαίνει πολύ καλά και να έχει φτάσει υψηλά επίπεδα ακρίβειας. Αν κοιτάξει κανείς όμως την ακρίβεια στο σύνολο δοκιμής - test set, παρατηρεί εύκολα το overfitting.



Από πάνω παρουσιάζεται η εξέλιξη της ακρίβειας του ίδιου μοντέλου πάνω στο test set. Η τιμή της ακρίβειας από τη μία είναι πάντοτε αρκετά μικρότερη απ' ότι στο σύνολο εκπαίδευσης. Από την άλλη ξεκινάει να φθήνει από ένα σημείο και ύστερα - ενώ στο train set αυξάνει συνεχώς. Αυτό ακριβώς το φαινόμενο αντιμετωπίζεται με το regularization.

5.5 Μετρικές

Ως κύρια μετρική αξιολόγησης χρησιμοποιήθηκε το *accuracy* ή "ακρίβεια" της πρόβλεψης. Ορίζεται ως $\frac{\#\text{Σωστών προβέψεων}}{\#\text{Συνολου Δεδομένων}}$. Κατασκευάστηκαν επίσης πίνακες σύγχυσης - confusion matrices που χρησιμοποιούνται συχνά σε συστήματα κατηγοριοποίησης σε παραπάνω από μία κατηγορίες. Σε αυτούς οι στήλες αντιπροσωπεύουν τις προβλεπόμενες κατηγορίες, ενώ οι γραμμές τις πραγματικές. Μας δίνουν έτσι πολύ καλή εικόνα για την απόδοση του συστήματος σε διαφορετικά δεδομένα.

5.6 Early Stopping

Στο τέλος της εκπαίδευσης, κρατώνται τα βάρη και οι υπερπαράμετροι του βήματος με την καλύτερη επίδοση στο σύνολο δοκιμής (ή στο σύνολο ανάπτυξης σε κάποιες περιπτώσεις). Έτσι κάθε φορά αξιολογείται η βέλτιστη κατάσταση του κάθε μοντέλου. Δίνεται ένα παράθυρο εποχών στο οποίο εάν δεν αυξηθεί η απόδοση, η εκπαίδευση σταματά.

Κεφάλαιο 6 - Εκτέλεση Πειραμάτων

6.1 Naive Bayes (baseline)

Ο Naive Bayes είναι ένας απλός ταξινομητής που χρησιμοποιείται στη Μηχανική Μάθηση και στηρίζεται στο θεώρημα του Bayes για τη δεσμευμένη πιθανότητα. [15]

Με την παραδοχή ανεξαρτησίας των χαρακτηριστικών δεδομένης της συγχορδίας, μπορούμε να γράψουμε:

$P(x_0|x_1,..x_n,C) = P(x_0|C)$, όπου x_i είναι τα χαρακτηριστικά και C η συγχορδία. Με την παραδοχή αυτή και με το θεώρημα του Bayes, προκύπτει ο κλειστός τύπος για την πρόβλεψη του μοντέλου y:

$$y = argmax_{k \in chords} P(C_k) \prod_{i}^{features} P(x_i|C_k)$$

Οι πιθανότητες $P(x_i|C_i)$ και $P(C_k)$ υπολογίζονται μετρώντας παραδείγματα στο σύνολο εκπαίδευσης. Επειδή τα χαρακτηριστικά είναι συνεχή, για τον υπολογισμό των πιθανοτήτων πρέπει να υποθέσουμε πως ακολουθούν κάποια κατανομή.

Για την εκπαίδευση, κατασκευάστηκαν πίνακες χαρακτηριστικών όπως αναφέρεται στο κεφάλαιο “Επιλογή Χαρακτηριστικών: Feedforward” και εκτελέστηκαν πειράματα με διάφορους συνδυασμούς χαρακτηριστικών, με διάφορες τιμές του N και για τις δύο κατηγοριοποιήσεις (**τονική** ή **majmin**).

Για τον Naive Bayes δεν χρησιμοποιήθηκε σύνολο ανάπτυξης (dev set), ενώ από τις διαφορετικές κατανομές που υποτέθηκαν για τα χαρακτηριστικά, μόνο η κανονική κατανομή κατέληγε σε σχετικά καλές προβλέψεις.

Ενδεικτικά αποτελέσματα (train size=741,585 segments, test size=36,599 segments:

Naive bayes accuracy (Gaussian Distribution assumption)

Accuracies are averages of 2 runs

Chord Classes	N[1]	Features	#Features	Train Accuracy	Test Accuracy
tonic	0	all	29	42.74%	40.74%
majmin	0	all	29	34.16%	31.66%
tonic	0	pitches	12	42.8%	42.31%
majmin	0	pitches	12	34.36%	33.11%
tonic	1	all	87	52.43%	50.52%
majmin	1	all	87	41.94%	37.22%
tonic	1	pitches	36	52.29%	50.86%
majmin	1	pitches	36	43.9%	42.3%
tonic	2	all	145	55.22%	52.63%
majmin	2	all	145	44.63%	38.38%
tonic	2	pitches	60	54.83%	52.59%
majmin	2	pitches	60	45.55%	43.54%
tonic	5	all	319	55.42%	53.28%
majmin	5	all	319	44.45%	37.81%
tonic	5	pitches	132	55.31%	56.84%

[1] Τα χαρακτηριστικά των N προηγούμενων και N επόμενων segments προστίθενται ως χαρακτηριστικά του κάθε segment

majmin	5	pitches	132	43.64%	41.09%
tonic	10	all	609	51.42%	53.52%
majmin	10	all	609	40.95%	38.73%
tonic	10	pitches	252	51.39%	50.11%
majmin	10	pitches	252	44.54%	39.44%

Υπενθυμίζεται πως για την πρόβλεψη της συγχορδίας ενός segment χρησιμοποιούνται τα χαρακτηριστικά των N προηγούμενων και N επόμενων segments (καθώς και το ίδιο το segment).

Από τον παραπάνω πίνακα μπορούμε να εξάγουμε τα εξής συμπεράσματα:

1) Τα Test Accuracy και Train Accuracy απέχουν περισσότερο μεταξύ τους όταν χρησιμοποιούνται όλα τα χαρακτηριστικά (σε αντίθεση με όταν χρησιμοποιούνται μόνο τα “pitches”). Αυτό σημαίνει πως τα περαιτέρω χαρακτηριστικά **στην περίπτωση του Naive Bayes** προσθέτουν θόρυβο στα δεδομένα και δεν συνεισφέρουν πληροφορία για την πρόβλεψη. Αυτό μπορεί να οφείλεται στις προϋποθέσεις του Naive Bayes για την ανεξαρτησία των χαρακτηριστικών και άρα σε άλλα μοντέλα μπορεί να είναι χρήσιμα.
2) Με N=0 segments η απόδοση είναι πολύ χειρότερη από ότι με άλλες τιμές του N. Αυτό αναμενόταν λόγω της χρονικής εξάρτησης της συγχορδίας. Από την άλλη για μεγαλύτερες τιμές του N αυξάνει η υπολογιστική πολυπλοκότητα, ενώ για πολύ μεγάλες (πχ N=10) πέφτει η απόδοση.

6.2 Feedforward Neural Network - Multi Layer Perceptron

Το πρώτο μοντέλο που χρησιμοποιήθηκε επιχειρώντας να αυξηθεί η ακρίβεια των προβλέψεων είναι η απλούστερη δομή νευρωνικού δικτύου, το Feedforward Νευρωνικό Δίκτυο ή Multi Layer Perceptron.

Αυτό αποτελείται από επίπεδα (layers) που περιέχουν πλήθος νευρώνων το κάθε ένα. Ένας νευρώνας λαμβάνει είσοδο από όλους τους νευρώνες του προηγούμενου επιπέδου, και εφαρμόζει κάποια μη-γραμμική συνάρτηση (activation function) σε κάποιο γραμμικό συνδυασμό των εισόδων. Αν οι είσοδοι ενός επιπέδου l γραφούν σε μορφή διανύσματος $x^{[l]}$ τότε ο γραμμικός συνδυασμός των εισόδων θα είναι της μορφής $z^{[l]} = W^{[l]T} * x^{[l]} + b^{[l]}$ (εσωτερικό γινόμενο διανυσμάτων W και x, + b). Ύστερα εφαρμόζεται η μη-γραμμική συνάρτηση, έστω g και προκύπτει η είσοδος του επόμενου επιπέδου, δηλαδή $x^{[l+1]} = g(z^{[l]})$.

Η είσοδος του πρώτου επιπέδου είναι το διάνυσμα χαρακτηριστικών. Η έξοδος του τελευταίου επιπέδου είναι διάνυσμα των κατηγοριών που επιχειρούμε να προβλέψουμε. Ο τρόπος που μαθαίνει να κάνει σωστές προβλέψεις ένα τέτοιο σύστημα είναι ο αλγόριθμος backpropagation [16].

Σύμφωνα με αυτόν: Αρχικά τίθεται μια συνάρτηση κόστους η οποία εξαρτάται από την έξοδο του τελευταίου επιπέδου. Συνήθως είναι κάποια μορφή απόστασης των προβλέψεων του δικτύου με τις πραγματικές τιμές, και είναι η συνάρτηση που επιχειρούμε να ελαχιστοποιήσουμε [17].

Με μεθόδους ανάλυσης και με προσεκτική επιλογή των μη γραμμικών συναρτήσεων, ο αλγόριθμος backpropagation καταλήγει στην εξάρτηση του διανύσματος εξόδου από το διάνυσμα εισόδου, διανύοντας ανάποδα το δίκτυο και υπολογίζοντας σε κάθε βήμα την παράγωγο της συνάρτησης κόστους σε σχέση με την παράμετρο που διανύεται. Στη συνέχεια ακολουθώντας την κλίση του κόστους στον χώρο των παραμέτρων (gradient descent) ή με παραλλαγές, μεταβάλλονται οι τιμές όλων των παραμέτρων (W kai b στο παράδειγμα) με τέτοιο τρόπο ώστε να μειωθεί η τιμή της συνάρτησης κόστους (ανάλογα με το πρόσημο και το μέτρο της παραγώγου σε κάθε βήμα). Στη συνέχεια επιχειρείται πρόβλεψη με τις νέες παραμέτρους και η διαδικασία επαναλαμβάνεται. Τελικά η συνάρτηση κόστους θα συγκλίνει σε μια ελάχιστη τιμή και το σύστημα θα προβλέπει συχνότερα την έξοδο με επιτυχία.

Στους παρακάτω πίνακες φαίνονται αποτελέσματα για διάφορα Feedforward νευρωνικά δίκτυα, με διάφορους συνδυασμούς των χαρακτηριστικών (και για τις δύο ομαδοποιήσεις (tonic, majmin)). Οι αναγραφόμενες τιμές για τα accuracies αντιστοιχούν στο βήμα της εκπαίδευσης που είχε το καλύτερο *Test Accuracy*. Κάτω από κάθε πίνακα παρουσιάζονται ενδεικτικά διαγράμματα της εξέλιξης των *accuracies* και της *συνάρτησης απώλειας (loss function)*.

6.2.1. Ένα Κρυφό Επίπεδο - 12 νευρώνες

6.2.1α. Τονική (13 κατηγορίες) - Initial Train Set (487k Segments)

Best test accuracy

N[2]	Features	#features	Train Accuracy	Dev Accuracy	Test Accuracy
0	pitches	12	48.97%	46.75%	48.51%
0	all	29	49.40%	48.08%	49.67%
0	*pt*[3]	24	49.56%	49.72%	43.59%
1	pitches	36	58.27%	**5**8.63%	59.88%
1	all	87	59.43%	62.19%	54.26%
1	*pt*	72	59.08%	59.99%	58.41%
2	pitches	60	61.56%	61.91%	62.05%
2	all	145	62.42%	62.48%	62.43%
2	*pt*	120	62.67%	61.13%	58.61%
4	pitches	108	62.13%	62.30%	63.17%
5	pitches	132	63.62%	66.08%	59.38%
5	all	319	64.30%	68.92%	64.02%

[2] N: Σε κάθε segment προστίθενται ως χαρακτηριστικά, τα χαρακτηριστικά των N προηγούμενων και N επόμενων segments.
[3] *pt : pitches and timbre*

5	*pt*	264	63.39%	64.59%	64.18%
10	pitches	252	63.72%	63.55%	60.47%
10	all	609	65.22%	62.36%	58.19%

6.2.1β. Ματζόρε/Μινόρε (25 κατηγορίες)

Best test accuracy

N[4]	Features	#features	Train Accuracy	Dev Accuracy	Test Accuracy
0	pitches	12	40.90%	42.08%	38.10%
0	all	29	40.91%	38.08%	36.35%
0	*pt*[5]	24	41.04%	46.93%	39.35%
1	pitches	36	50.04%	49.97%	49.20%
1	all	87	50.91%	47.69%	51.09%
1	*pt*	72	50.53%	51.99%	48.75%
2	pitches	60	53.71%	50.64%	54.35%
2	all	145	53.65%	51.15%	52.77%
2	*pt*	120	54.96%	48.18%	51.23%
4	pitches	108	54.70%	54.27%	53.30%
5	pitches	132	55.93%	50.99%	53.43%
5	all	319	56.11%	55.46%	53.66%
5	*pt*	264	56.10%	56.79%	51.47%

[4] N: Σε κάθε segment προστίθενται ως χαρακτηριστικά, τα χαρακτηριστικά των N προηγούμενων και N επόμενων segments.
[5] *pt : pitches and timbre*

6.2.2. Ένα Κρυφό Επίπεδο - 60 Νευρώνες - Initial Train Set (487k Segments)

6.2.2α. Τονική

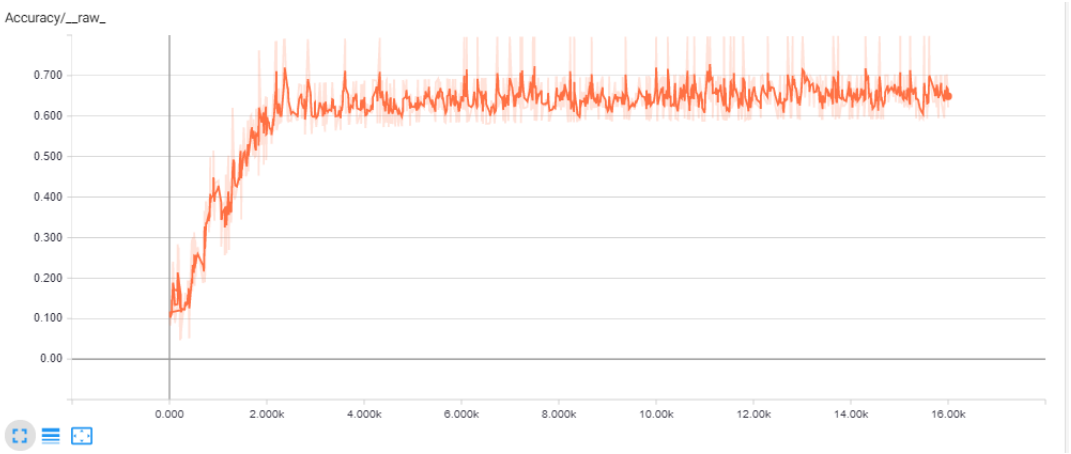
Best test accuracy

N[6]	Features	#features	Train Accuracy	Dev Accuracy	Test Accuracy
0	pitches	12	51.35%	52.89%	50.08%
0	all	29	52.95%	46.42%	48.48%
0	*pt*[7]	24	52.47%	49.73%	51.41%
1	pitches	36	61.64%	58.31.%	53.51%
1	all	87	62.41%	62.80%	56.07%
1	*pt*	72	62.53%	58.42%	58.03%
2	pitches	60	63.95%	64.93%	57.00%
2	all	145	65.32%	67.54%	61.71%
2	*pt*	120	65.20%	64.15%	61.17%
5	pitches	132	65.71%	65.36%	58.45%
5	all	319	68.00%	63.38%	58.76%
5	*pt*	264	66.47%	65.38%	58.19%
10	all	609	69.27%	65.38%	60.71%

[6] N: Σε κάθε segment προστίθενται ως χαρακτηριστικά, τα χαρακτηριστικά των N προηγούμενων και N επόμενων segments.
[7] *pt : pitches and timbre*

Οπτικοποίηση - Διαγράμματα:

i) N=2, 60 νευρώνες, pitches - 500 epochs - Αναγνώριση Τονικής

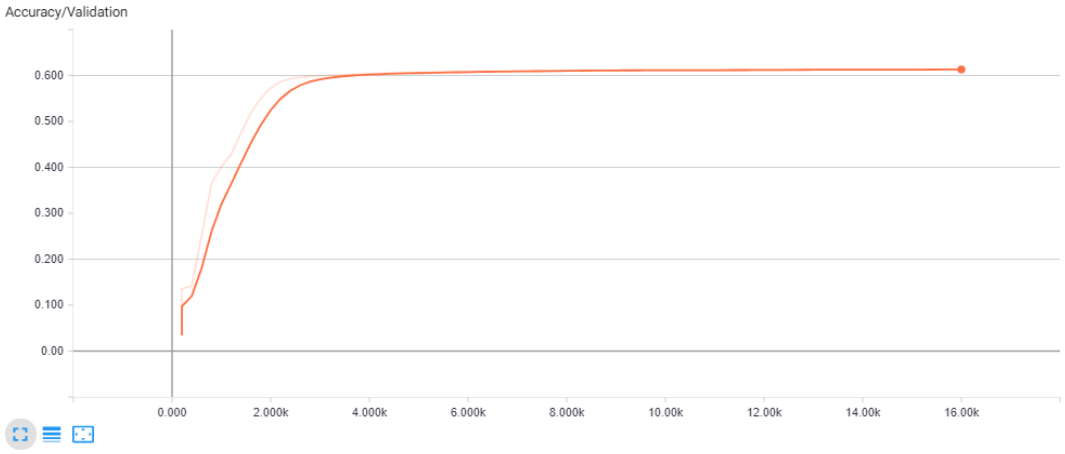


Στο παραπάνω διάγραμμα, ο οριζόντιος άξονας αντιπροσωπεύει τα training steps. Ένα training step ισοδυναμεί με ένα πέρασμα από 16384 segments (στο συγκεκριμένο μοντέλο). Έτσι μία εποχή (epoch - πέρασμα από όλα τα δεδομένα εκπαίδευσης) αντιστοιχεί σε περίπου 45 training steps. Το πείραμα έτρεξε για 500 epochs. Ο κάθετος άξονας αντιπροσωπεύει την ακρίβεια (accuracy).

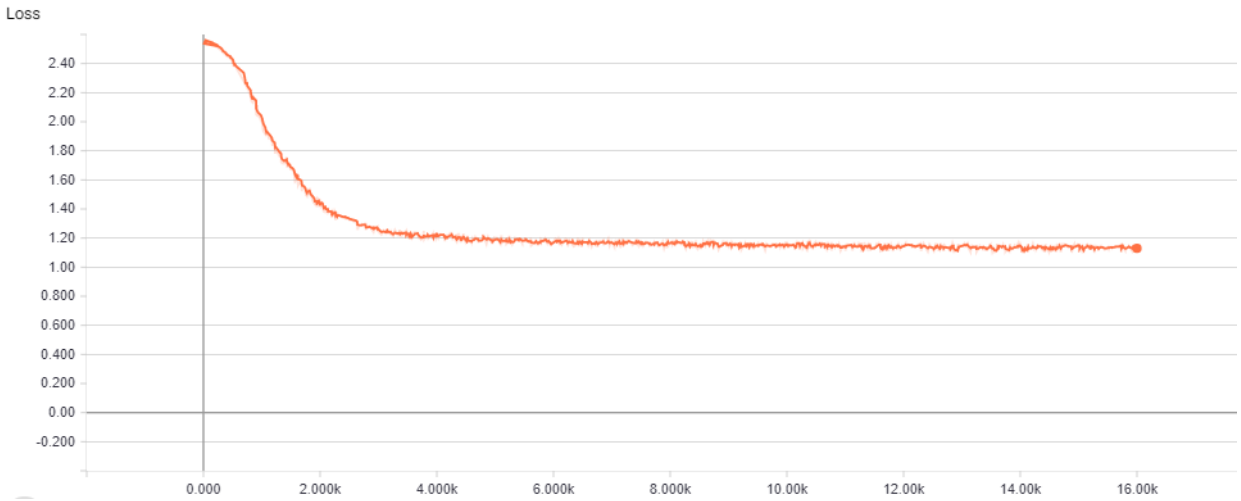
Η καμπύλη με το έντονο χρώμα είναι ο κινητός μέσος όρος (με ολισθαίνων παράθυρο) της ακρίβεια στο train set, ενώ με το λιγότερο έντονο είναι η τιμή της ακρίβειας στο τρέχον minibatch.

Με αύξηση του παραθύρου πετυχαίνεται λείανση των γραφημάτων και φαίνεται καθαρότερα πως δεν αυξάνει άλλο η ακρίβεια.

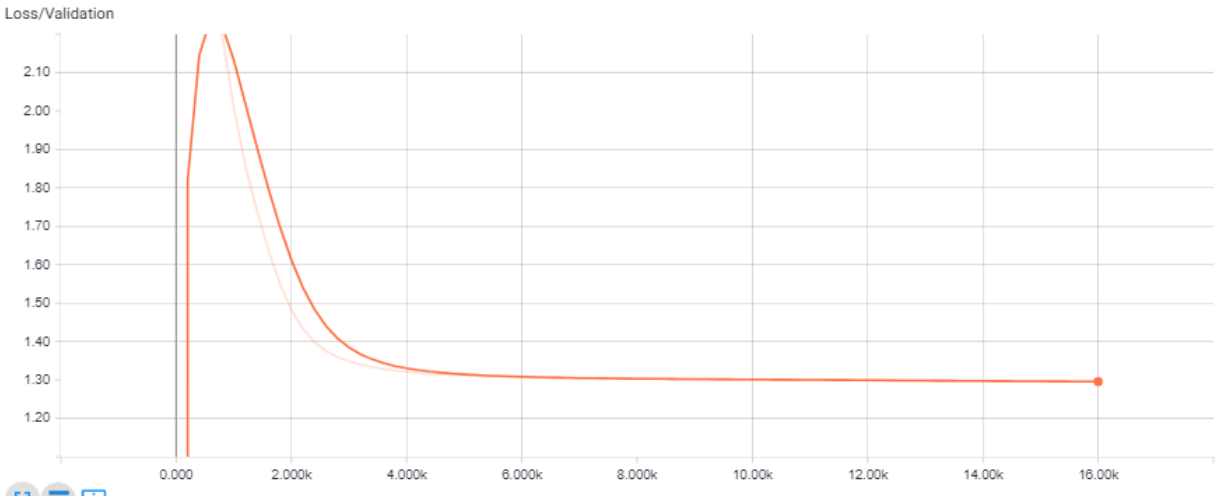
(Κάτω) Ακρίβεια Test Set



(Κάτω) Τιμή Συνάρτησης Κόστους - Train Set



(Κάτω) Τιμή Συνάρτησης Κόστους - Test Set



Στα διαγράμματα απώλειας φαίνεται επίσης ο "τοίχος" που συναντά κοντά στην τιμή ακρίβειας 60% για αναγνώρισης τονικής

6.2.2β) Ματζόρε- Μινόρε (25 κατηγορίες)

Best test accuracy

N[8]	Features	#features	Train Accuracy	Dev Accuracy	Test Accuracy
0	pitches	12	41.97%	43.60%	44.48%
0	all	29	43.29%	43.06%	44.59%
0	*pt*[9]	24	45.20%	45.07%	41.18%
1	pitches	36	52.41%	54.41%	44.74%
1	all	87	53.35%	53.32%	47.27%
1	*pt*	72	53.26%	53.34%	45.22%
2	pitches	60	56.54%	56.03%	50.39%

[8] N: Σε κάθε segment προστίθενται ως χαρακτηριστικά, τα χαρακτηριστικά των N προηγούμενων και N επόμενων segments.
[9] *pt : pitches and timbre*

2	all	145	58.14%	56.98%	53.27%
2	*pt*	120	58.06%	55.43%	52.05%
4	pitches	108	58.91%	57.14%	50.10%
5	pitches	132	59.33%	54.37%	53.29%
5	all	319	61.81%	56.03%	49.76%
5	*pt*	264	61.00%	56.24%	50.74%
10	pitches	252	60.09%	54.92%	52.15%
10	all	609	63.07%	55.28%	49.94%

6.2.3 Ένα Κρυφό Επίπεδο - 600 Νευρώνες

6.2.3α) Αρχικό Σύνολο (487k Segments) - Ματζόρε Μινόρε (25 κατηγορίες)

Best test accuracy

N[10]	Features	#features	Train Accuracy	Dev Accuracy	Test Accuracy
5	pitches	132	59.34%	58.04%	54.79%
5	pt	264	62.16%	58.86%	54.50%

6.2.3β) Φιλτραρισμένο Σύνολο (85k segments) - Ματζόρε Μινόρε (25 κατηγορίες)

Best test accuracy

Num Neurons	Regularizer	N	Features	#features	Train Accuracy	Test Accuracy
12	None	5	pitches	132	52.46%	50.89%
600	None	5	pitches	132	58.15%	52.63%

[10] N: Σε κάθε segment προστίθενται ως χαρακτηριστικά, τα χαρακτηριστικά των N προηγούμενων και N επόμενων segments.

600	L2	5	pitches	132	56.23%	53.18%
600	None	11	all	667	70.11%	49.66%
600	L2	11	all	667	61.19%	50.83

6.2.3γ) Επαυξημένο Φιλτραρισμένο Σύνολο (877k segments)

Num Neurons	Regularizer	N	Features	#features	Train Accuracy	Test Accuracy
60	L2	5	pitches	132	51.28%	53.45%

6.2.4. Αξιολόγηση Μοντέλων Ενός Κρυφού Επιπέδου - Confusion Matrix

Για την αξιολόγηση μοντέλων κατηγοριοποίησης (classification) με παραπάνω από 2 κατηγορίες, χρησιμοποιούνται συχνά confusion matrices ή "πίνακες σύγχυσης" [18]. Αυτοί αναγράφουν το ποσοστό των προβλέψεων που έκανε σε κάθε κατηγορία το σύστημα, χωρισμένες με βάση την αληθινή κατηγορία. Παρακάτω, οι στήλες του πίνακα αναφέρονται στη συγχορδία που προβλεφθηκε, ενώ οι γραμμές στην πραγματική.

Συγκρίνουμε την απόδοση του ίδιου μοντέλου εκπαιδευμένου α) στο αρχικό Train Set, β) στο Φιλτραρισμένο Train Set και γ) στο Επαυξημένο Φιλτραρισμένο Train Set, παρατηρόντας τους πίνακες σύγχυσης που προκύπτουν από τις προβλέψεις του μοντέλου στο Test Set (ίδιο σε όλες τις περιπτώσεις)

α) N=5, pitches, 600 Neurons, L2, Trained on Initial Train Set - Confusion Matrix

	N	Amin	A	A#mir	A#	Bmin	B	Cmin	C	C#mir	C#	Dmin	D	D#mir	D#	Emin	E	Fmin	F	F#min	F#	Gmin	G	G#mir	G#	Popula
N	0.1090	0.0257	0.0872	0.0005	0.0263	0.0141	0.0486	0.0031	0.0941	0.0063	0.0317	0.0091	0.1063	0.0035	0.0510	0.0295	0.1301	0.0010	0.0363	0.0034	0.0445	0.0042	0.0953	0.0005	0.0367	7538
Amin	0.0047	0.3632	0.3030	0.0005	0.0103	0.0015	0.0006	0	0.0717	0.0003	0.0003	0.0252	0.0582	0	0.0021	0.0111	0.0116	0.0010	0.0715	0.0030	0.0001	0.0005	0.0537	0.0011	0.0034	9302
A	0.0075	0.0461	0.6974	0.0001	0.0093	0.0066	0.0106	0.0000	0.0077	0.0091	0.0039	0.0081	0.0629	0.0002	0.0031	0.0288	0.0451	0	0.0100	0.0108	0.0041	0.0002	0.0219	0.0009	0.0049	20240
A#mir	0.0028	0.0003	0.0047	0.1584	0.3000	0	0.0088	0.0009	0.0066	0.0031	0.1568	0.0006	0.0025	0.1032	0.0861	0	0.0095	0.0307	0.0183	0	0.0240	0.0006	0	0.0110	0.0700	3156
A#	0.0107	0.0004	0.0044	0.0085	0.6504	0.0009	0.0074	0.0108	0.0327	0	0.0133	0.0321	0.0059	0.0045	0.0930	0.0002	0.0005	0.0095	0.0540	0	0.0057	0.0206	0.0158	0	0.0168	11013
Bmin	0.0103	0.0044	0.0687	0.0067	0.0044	0.3601	0.1974	0.0065	0.0314	0.0014	0.0029	0.0016	0.1251	0.0008	0.0016	0.0213	0.0565	0.0002	0.0046	0.0185	0.0170	0	0.0552	0.0014	0.0004	4740
B	0.0049	0.0004	0.0618	0.0002	0.0296	0.0506	0.5308	0.0008	0.0248	0.0106	0.0153	0.0006	0.0183	0.0095	0.0108	0.0109	0.0916	0.0002	0.0031	0.0106	0.0685	0.0004	0.0197	0.0101	0.0146	9702
Cmin	0.0081	0.0050	0	0.0025	0.0495	0	0.0010	0.3274	0.3407	0	0.0106	0.0018	0.0056	0	0.0790	0	0	0.0110	0.0623	0	0	0.0119	0.0294	0	0.0533	4780
C	0.0030	0.0178	0.0082	0.0004	0.0146	0.0015	0.0275	0.0088	0.6924	0.0019	0.0109	0.0168	0.0379	0.0013	0.0037	0.0158	0.0145	0.0008	0.0458	0.0009	0.0033	0.0038	0.0634	0.0002	0.0036	22651
C#mir	0.0006	0.0008	0.1388	0.0002	0.0013	0.0006	0.0512	0	0.0019	0.4150	0.1421	0.0008	0.0193	0.0028	0.0084	0.0002	0.1119	0	0.0004	0.0052	0.0724	0	0.0028	0.0030	0.0195	4609
C#	0.0068	0	0.0055	0.0038	0.0124	0.0014	0.0252	0.0011	0.0057	0.0317	0.6312	0.0012	0.0087	0.0211	0.0281	0.0024	0.0126	0.0098	0.0161	0.0026	0.0866	0.0006	0.0063	0.0045	0.0734	11528
Dmin	0.0016	0.0152	0.0462	0	0.0332	0.0013	0	0.0006	0.0432	0.0033	0.0067	0.4542	0.2577	0.0058	0.0128	0.0020	0.0078	0	0.0561	0.0001	0.0027	0.0041	0.0422	0.0001	0.0018	7216
D	0.0058	0.0079	0.0763	0	0.0041	0.0107	0.0039	0.0001	0.0166	0.0015	0.0041	0.0290	0.6868	0.0000	0.0033	0.0062	0.0180	0.0000	0.0074	0.0077	0.0066	0.0018	0.0956	0.0000	0.0053	25278
D#mir	0.0006	0	0.0024	0	0.0092	0	0.0141	0.0003	0	0.0055	0.0457	0	0.0017	0.3262	0.4150	0	0.0127	0.0103	0.0010	0	0.0812	0	0	0.0116	0.0619	2906
D#	0.0060	0.0019	0.0019	0.0028	0.0547	0	0.0027	0.0160	0.0033	0.0022	0.0256	0.0006	0.0010	0.0131	0.7108	0	0.0034	0.0125	0.0105	0.0004	0.0097	0.0113	0.0217	0.0022	0.0843	10117
Emin	0.0136	0.0188	0.0822	0	0.0043	0.0055	0.0073	0	0.0808	0.0005	0.0006	0.0015	0.0473	0.0014	0.0058	0.3293	0.2518	0.0092	0.0478	0.0015	0.0032	0.0002	0.0855	0	0.0007	8999
E	0.0085	0.0081	0.0903	0.0001	0.0056	0.0032	0.0328	0	0.0151	0.0157	0.0028	0.0007	0.0230	0.0016	0.0093	0.0329	0.6993	0.0006	0.0111	0.0023	0.0135	0.0000	0.0080	0.0063	0.0075	17979
Fmin	0.0054	0.0029	0.0011	0.0029	0.0690	0	0.0041	0.0226	0.0644	0.0002	0.0681	0.0082	0.0029	0.0002	0.0265	0.0002	0.0011	0.3111	0.2443	0	0.0059	0.0050	0.0198	0	0.1330	4374
F	0.0057	0.0200	0.0082	0.0004	0.0427	0.0002	0.0029	0.0039	0.0809	0.0007	0.0114	0.0275	0.0059	0	0.0097	0.0070	0.0165	0.0216	0.6595	0.0023	0.0095	0.0047	0.0491	0.0004	0.0080	14098
F#min	0.0045	0.0018	0.1565	0	0.0004	0.0109	0.0508	0	0.0027	0.0214	0.0234	0.0002	0.1098	0	0.0013	0.0027	0.0344	0	0.0038	0.4054	0.1453	0	0.0182	0.0022	0.0034	4388
F#	0.0051	0.0016	0.0086	0.0017	0.0210	0.0093	0.0681	0.0020	0.0061	0.0181	0.0710	0	0.0115	0.0192	0.0085	0.0015	0.0385	0.0012	0.0080	0.0354	0.6045	0.0017	0.0202	0.0069	0.0290	8580
Gmin	0.0103	0.0034	0.0101	0	0.0727	0.0002	0	0.0138	0.0643	0	0.0002	0.0189	0.0616	0	0.0204	0.0024	0	0.0019	0.0441	0.0017	0.0002	0.3151	0.3457	0	0.0120	4055
G	0.0046	0.0122	0.0274	0	0.0040	0.0077	0.0049	0.0041	0.0874	0.0012	0.0012	0.0078	0.0678	0.0001	0.0034	0.0219	0.0051	0.0002	0.0200	0.0009	0.0033	0.0141	0.6917	0	0.0080	24060
G#mir	0.0026	0.0023	0.0066	0.0306	0.0049	0	0.0736	0.0014	0.0002	0.0335	0.1187	0	0.0046	0.0497	0.0161	0	0.0523	0.0069	0.0011	0.0005	0.0771	0	0.0005	0.1800	0.3358	3460
G#	0.0113	0.0031	0.0058	0.0078	0.0330	0	0.0075	0.0235	0.0202	0.0044	0.1098	0.0001	0.0031	0.0035	0.0862	0	0.0045	0.0232	0.0092	0	0.0254	0.0013	0.0098	0.0048	0.6007	9897
sums	0.2556	0.5644	1.9045	0.2292	1.4680	0.4872	1.1828	0.4486	1.7961	0.5886	1.5095	0.6477	1.7365	0.5694	1.6975	0.5270	1.6314	0.4644	1.4478	0.5135	1.3155	0.4030	1.7725	0.2483	1.5891	25466

Παρατηρούμε πως τα περισσότερα λάθη στην περίπτωση των 25 κατηγοριών (ματζόρε μινόρε) όταν το μοντέλο είναι εκπαιδευμένο στο αρχικό σύνολο δεδομένων προέρχονται από τη σύγχυση της μινόρε συγχορδίας με ματζόρε.

β) N=5, pitches, 600 Neurons, L2, Trained on Filtered Dataset - Confusion Matrix

	N	Amin	A	A#mir	A#	Bmin	B	Cmin	C	C#mir	C#	Dmin	D	D#mir	D#	Emin	E	Fmin	F	F#min	F#	Gmin	G	G#mir	G#	Popula
N	0.1525	0.0204	0.0279	0.0527	0.0214	0.0380	0.0626	0.0302	0.0212	0.0202	0.0350	0.0181	0.0451	0.0390	0.0520	0.0367	0.0780	0.0201	0.0202	0.0147	0.0553	0.0321	0.0225	0.0376	0.0449	7538
Amin	0.0271	0.4925	0.1409	0.0170	0.0074	0.0199	0.0092	0.0054	0.0304	0.0045	0.0013	0.0457	0.0320	0.0003	0.0031	0.0308	0.0091	0.0111	0.0590	0.0113	0.0015	0.0077	0.0228	0.0035	0.0048	9302
A	0.0321	0.0630	0.4759	0.0145	0.0074	0.0304	0.0253	0.0019	0.0028	0.0463	0.0134	0.0197	0.0349	0.0035	0.0048	0.0495	0.0359	0.0015	0.0086	0.0718	0.0124	0.0025	0.0101	0.0200	0.0101	20240
A#mir	0.0019	0	0.0015	0.5592	0.0614	0	0.0057	0.0053	0.0003	0.0063	0.0744	0.0006	0	0.1025	0.0396	0	0.0012	0.0605	0.0022	0	0.0114	0.0006	0	0.0313	0.0325	3156
A#	0.0069	0.0007	0.0009	0.1375	0.5023	0.0019	0.0052	0.0408	0.0065	0.0000	0.0085	0.0288	0.0002	0.0427	0.0882	0.0005	0.0003	0.0516	0.0140	0.0005	0.0029	0.0403	0.0014	0.0032	0.0123	11013
Bmin	0.0200	0.0052	0.0128	0.0352	0.0027	0.4685	0.2191	0.0149	0.0025	0.0086	0.0056	0.0040	0.0280	0.0143	0.0014	0.0242	0.0221	0.0023	0.0014	0.0580	0.0255	0.0006	0.0107	0.0092	0.0014	4740
B	0.0086	0.0004	0.0101	0.0303	0.0117	0.0660	0.5195	0.0045	0.0030	0.0315	0.0159	0.0017	0.0066	0.0553	0.0077	0.0110	0.0368	0.0065	0.0001	0.0289	0.0688	0.0025	0.0047	0.0606	0.0062	9702
Cmin	0.0050	0.0046	0	0.0355	0.0242	0.0048	0.0014	0.6525	0.0410	0	0.0050	0.0033	0.0002	0.0037	0.0644	0	0	0.0696	0.0081	0	0	0.0253	0.0035	0.0008	0.0464	4780
C	0.0190	0.0335	0.0033	0.0096	0.0188	0.0116	0.0257	0.1521	0.4456	0.0070	0.0122	0.0437	0.0141	0.0080	0.0054	0.0371	0.0079	0.0254	0.0425	0.0026	0.0067	0.0237	0.0357	0.0025	0.0054	22651
C#mir	0.0017	0.0008	0.0548	0.0095	0	0.0010	0.0357	0.0013	0	0.5693	0.1286	0.0004	0.0013	0.0431	0.0039	0.0002	0.0175	0.0006	0	0.0182	0.0540	0	0	0.0477	0.0095	4609
C#	0.0041	0.0005	0.0009	0.0677	0.0021	0.0015	0.0161	0.0040	0.0010	0.0423	0.6074	0.0007	0.0009	0.0728	0.0104	0.0005	0.0028	0.0324	0.0002	0.0078	0.0533	0.0013	0.0022	0.0289	0.0369	11528
Dmin	0.0160	0.0191	0.0213	0.0098	0.0429	0.0124	0.0011	0.0153	0.0139	0.0085	0.0123	0.5745	0.1043	0.0173	0.0094	0.0094	0.0049	0.0022	0.0478	0.0020	0.0024	0.0300	0.0177	0.0012	0.0026	7216
D	0.0341	0.0193	0.0458	0.0041	0.0066	0.0479	0.0066	0.0043	0.0077	0.0098	0.0185	0.1067	0.4831	0.0035	0.0068	0.0183	0.0138	0.0021	0.0056	0.0511	0.0186	0.0115	0.0572	0.0058	0.0103	25278
D#mir	0.0003	0	0.0003	0.0227	0.0030	0	0.0048	0.0003	0	0.0089	0.0203	0	0	0.7209	0.1355	0	0.0013	0.0158	0.0003	0.0003	0.0240	0	0	0.0302	0.0103	2906
D#	0.0026	0.0011	0.0003	0.0458	0.0234	0.0000	0.0025	0.0340	0.0000	0.0035	0.0206	0.0009	0.0000	0.1416	0.5802	0.0008	0.0003	0.0380	0.0012	0.0001	0.0044	0.0193	0.0032	0.0144	0.0600	10117
Emin	0.0373	0.0202	0.0345	0.0077	0.0041	0.0242	0.0148	0.0018	0.0265	0.0138	0.0020	0.0074	0.0256	0.0072	0.0074	0.4506	0.1861	0.0486	0.0137	0.0128	0.0050	0.0043	0.0353	0.0022	0.0057	8999
E	0.0234	0.0078	0.0342	0.0140	0.0041	0.0126	0.0507	0.0008	0.0035	0.1228	0.0041	0.0017	0.0082	0.0138	0.0159	0.0546	0.5089	0.0072	0.0044	0.0132	0.0282	0.0009	0.0026	0.0533	0.0078	17979
Fmin	0.0073	0.0020	0.0004	0.0428	0.0251	0.0002	0.0011	0.0523	0.0100	0.0016	0.0582	0.0086	0.0002	0.0080	0.0201	0.0004	0.0002	0.6177	0.0155	0.0006	0.0013	0.0160	0.0045	0.0084	0.0962	4374
F	0.0216	0.0235	0.0026	0.0207	0.0440	0.0020	0.0037	0.0335	0.0336	0.0055	0.0137	0.0414	0.0013	0.0024	0.0117	0.0103	0.0080	0.1848	0.4592	0.0067	0.0100	0.0365	0.0125	0.0016	0.0083	14098
F#min	0.0045	0.0011	0.0417	0.0041	0.0004	0.0091	0.0410	0.0009	0.0004	0.0360	0.0268	0.0009	0.0296	0.0105	0.0013	0.0075	0.0066	0.0022	0.0008	0.5656	0.1636	0.0002	0.0072	0.0328	0.0036	4388
F#	0.0057	0.0011	0.0029	0.0355	0.0047	0.0085	0.0558	0.0064	0.0011	0.0284	0.0826	0	0.0016	0.0868	0.0048	0.0018	0.0079	0.0106	0.0019	0.0561	0.5487	0.0032	0.0029	0.0231	0.0168	8580
Gmin	0.0219	0.0061	0.0029	0.0049	0.0742	0.0024	0	0.0670	0.0150	0.0004	0.0036	0.0182	0.0295	0.0004	0.0221	0.0041	0	0.0236	0.0207	0.0083	0.0066	0.5617	0.0872	0.0002	0.0175	4055
G	0.0261	0.0235	0.0147	0.0041	0.0057	0.0459	0.0108	0.0361	0.0462	0.0045	0.0058	0.0222	0.0438	0.0016	0.0073	0.0460	0.0025	0.0061	0.0181	0.0073	0.0139	0.1077	0.4832	0.0034	0.0122	24060
G#mir	0.0017	0	0.0008	0.1225	0.0005	0.0011	0.0289	0.0014	0	0.0367	0.0716	0	0	0.0702	0.0066	0	0.0170	0.0112	0	0.0020	0.0439	0.0002	0	0.4294	0.1534	3460
G#	0.0039	0.0024	0.0015	0.0721	0.0084	0.0003	0.0039	0.0398	0.0012	0.0075	0.0980	0.0002	0.0003	0.0385	0.0640	0	0.0005	0.0610	0.0009	0.0006	0.0131	0.0039	0.0004	0.0592	0.5177	9897
sums	0.4867	0.7501	0.9342	1.3808	0.9072	0.8113	1.1528	1.2080	0.7148	1.0250	1.3472	0.9508	0.8912	1.5105	1.1750	0.7956	0.9707	1.3140	0.7473	0.9417	1.1764	0.9329	0.8284	0.9117	1.1345	25466

Αν κρατηθούν ίδιος αριθμός segments για κάθε κατηγορία (filtered dataset), τότε στους πίνακες σύγχυσης φαίνεται πως αντιμετωπίστηκε το πρόβλημα με τη σύγχυση ματζόρε και μινόρε συγχορδιών. Παρ' όλα αυτά η συνολική ακρίβεια στα δύο παραπάνω μοντέλα (στο σύνολο δοκιμής - test set) είναι σχεδόν ίδια. Μάλιστα είναι 3% καλύτερη όταν υπήρχε μεγάλη σύγχυση εξ' αιτίας της μη ισορροπημένης κατανομής των συγχορδιών (στα σύνολα δεδομένων, αλλά και στην πραγματικότητα).

γ) N=5, pitches, 600 Neurons, L2, Trained on Filtered Augmented Dataset - Confusion Matrix

	N	Amin	A	A#mir	A#	Bmin	B	Cmin	C	C#mir	C#	Dmin	D	D#mir	D#	Emin	E	Fmin	F	F#min	F#	Gmin	G	G#mir	G#	Popula
N	0.160	0.047	0.037	0.017	0.012	0.035	0.021	0.045	0.058	0.065	0.038	0.032	0.072	0.022	0.020	0.047	0.061	0.009	0.021	0.047	0.035	0.020	0.036	0.012	0.022	7538
Amin	0.008	0.639	0.086	0.007	0.002	0.006	0.000	0.002	0.043	0.004	0.002	0.040	0.038	0.000	0.000	0.017	0.003	0.006	0.045	0.012	0	0.003	0.022	0.002	0.002	9302
A	0.019	0.133	0.520	0.006	0.003	0.022	0.008	0.000	0.007	0.051	0.003	0.013	0.042	0.001	0.000	0.037	0.017	0.000	0.006	0.078	0.001	0.001	0.008	0.003	0.006	20240
A#mir	0.001	0.001	0.002	0.596	0.025	0	0.004	0.011	0.003	0.008	0.104	0.000	0.000	0.061	0.025	0	0.001	0.052	0.007	0.000	0.015	0.001	0	0.045	0.027	3156
A#	0.017	0.003	0.001	0.165	0.423	0.002	0.004	0.062	0.019	0.001	0.013	0.044	0.001	0.017	0.051	0.000	0.000	0.054	0.038	0.000	0.004	0.059	0.000	0.000	0.009	11013
Bmin	0.028	0.018	0.017	0.012	0.000	0.542	0.083	0.016	0.017	0.021	0.016	0.006	0.046	0.003	0	0.027	0.018	0.001	0.003	0.081	0.011	0.000	0.017	0.004	0.000	4740
B	0.011	0.004	0.021	0.017	0.008	0.096	0.439	0.012	0.014	0.058	0.022	0.002	0.008	0.024	0.003	0.018	0.049	0.002	0.000	0.081	0.055	0.001	0.007	0.027	0.007	9702
Cmin	0.009	0.005	0	0.013	0.012	0.000	0.000	0.680	0.095	0.004	0.017	0.013	0.000	0.002	0.037	0	0	0.030	0.021	0	0.000	0.022	0.005	0.000	0.024	4780
C	0.010	0.044	0.001	0.003	0.005	0.004	0.022	0.073	0.608	0.011	0.013	0.051	0.014	0.002	0.000	0.024	0.003	0.008	0.036	0.006	0.002	0.013	0.030	0.001	0.001	22651
C#mir	0.001	0.003	0.071	0.001	0	0.000	0.032	0.001	0.001	0.649	0.097	0	0.002	0.013	0.001	0.001	0.019	0.000	0	0.048	0.037	0	0	0.007	0.007	4609
C#	0.003	0.000	0.001	0.043	0.000	0.003	0.015	0.007	0.002	0.066	0.636	0.000	0.001	0.035	0.006	0.001	0.003	0.030	0.000	0.020	0.069	0.001	0.002	0.011	0.031	11528
Dmin	0.011	0.039	0.018	0.006	0.013	0.010	0	0.007	0.029	0.016	0.008	0.620	0.110	0.013	0.003	0.005	0.002	0.000	0.033	0.002	0.001	0.024	0.015	0.000	0.000	7216
D	0.015	0.026	0.051	0.000	0.001	0.030	0.001	0.002	0.014	0.012	0.007	0.107	0.542	0.001	0.002	0.020	0.009	0.001	0.005	0.066	0.002	0.013	0.060	0.001	0.002	25278
D#mir	0.000	0	0.000	0.012	0.003	0	0.009	0.001	0	0.012	0.033	0	0.001	0.739	0.086	0.001	0.002	0.012	0.001	0.002	0.042	0.001	0	0.022	0.014	2906
D#	0.016	0.002	0.000	0.037	0.024	0	0.002	0.067	0.000	0.005	0.034	0.003	0.000	0.113	0.551	0.003	0.001	0.027	0.005	0.001	0.004	0.022	0.005	0.010	0.057	10117
Emin	0.025	0.058	0.046	0.002	0.001	0.019	0.004	0.001	0.059	0.019	0.000	0.008	0.026	0.005	0.002	0.496	0.096	0.042	0.016	0.020	0.001	0.001	0.043	0.000	0.000	8999
E	0.025	0.025	0.058	0.003	0.002	0.014	0.024	0.000	0.011	0.148	0.003	0.003	0.012	0.005	0.003	0.096	0.475	0.004	0.004	0.029	0.011	0.000	0.002	0.023	0.004	17979
Fmin	0.003	0.005	0	0.030	0.009	0	0.001	0.072	0.022	0.003	0.058	0.008	0.000	0.007	0.010	0.000	0.000	0.604	0.059	0.005	0.004	0.016	0.008	0.005	0.059	4374
F	0.011	0.041	0.002	0.016	0.019	0.000	0.000	0.024	0.073	0.005	0.013	0.066	0.001	0.001	0.005	0.006	0.008	0.073	0.559	0.014	0.009	0.022	0.018	0.000	0.001	14098
F#min	0.005	0.003	0.063	0.001	0	0.014	0.026	0.000	0.001	0.036	0.025	0.000	0.028	0.005	0	0.005	0.007	0.000	0.001	0.686	0.065	0.000	0.008	0.007	0.000	4388
F#	0.009	0.001	0.003	0.034	0.003	0.013	0.049	0.006	0.003	0.047	0.068	0.001	0.002	0.035	0.000	0.001	0.011	0.006	0.003	0.165	0.490	0.004	0.004	0.015	0.011	8580
Gmin	0.012	0.021	0.004	0.000	0.024	0.001	0	0.068	0.046	0.000	0.000	0.032	0.041	0.000	0.014	0.002	0	0.010	0.033	0.014	0.001	0.549	0.112	0	0.005	4055
G	0.014	0.044	0.013	0.001	0.000	0.025	0.002	0.024	0.104	0.006	0.003	0.032	0.049	0.000	0.002	0.051	0.000	0.001	0.013	0.017	0.003	0.060	0.519	0.000	0.004	24060
G#mir	0.005	0.001	0.002	0.088	0.001	0.000	0.033	0.003	0	0.087	0.126	0	0	0.021	0.006	0	0.022	0.016	0	0.017	0.045	0.000	5	0.356	0.162	3460
G#	0.011	0.003	0.001	0.038	0.007	0.000	0.001	0.065	0.004	0.021	0.132	0.000	0.000	0.024	0.040	0	0.001	0.073	0.001	0.004	0.017	0.005	0.001	0.030	0.507	9897
sums	0.443	1.178	1.030	1.160	0.609	0.847	0.792	1.261	1.244	1.367	1.485	1.094	1.047	1.160	0.878	0.869	0.820	1.072	0.921	1.425	0.937	0.849	0.933	0.591	0.973	25466

Στο επαυξημένο φιλτραρισμένο πετυχαίνεται ακρίβεια ίδια με του αρχικού, ενώ ταυτόχρονα αποφεύγεται η σύγχυση.

6.2.5) Συμπεράσματα Μοντέλων Ενός Κρυφού Επιπέδου

α) Ακόμα και με ένα κρυφό επίπεδο με λίγους νευρώνες (12) η απόδοση είναι καλύτερη από τον Naive Bayes. Αυτό αναμενόταν γιατί το νευρωνικό δεν έχει προϋποθέσεις ανεξαρτησίας όπως ο Naive Bayes και προσεγγίζει πιο πολύπλοκες σχέσεις εισόδου - εξόδου.

β) Τα παραπάνω χαρακτηριστικά πέραν των pitches δεν φαίνεται να επηρεάζουν σημαντικά την επίδοση των μοντέλων, ενώ όταν συμπεριλαμβάνονται εντείνεται το φαινόμενο του overfitting.

γ) Τα αποτελέσματα δεν διαφέρουν σημαντικά σε σχέση με τον αριθμό των νευρώνων (12 vs 60 vs 600) στην περίπτωση του ενός κρυφού επιπέδου. Αυτό μπορεί να οφείλεται στο γεγονός πως με μόνο ένα κρυφό επίπεδο η πολυπλοκότητα της συνάρτησης εισόδου-εξόδου που προσεγγίζει ένα νευρωνικό είναι περιορισμένη. Ενδέχεται ο περιορισμός να προκύπτει από τα ίδια τα δεδομένα.

δ) Συγκρίνοντας τα Train, Dev, Test Accuracies για κάθε περίπτωση παρατηρούμε πως η ικανότητα γενίκευσης του δικτύου είναι χειρότερη στην περίπτωση των περισσότερων νευρώνων καθώς και των περισσότερων χαρακτηριστικών πέραν των pitches (απέχουν περισσότερο μεταξύ τους τα accuracies τη στιγμή του μέγιστου test accuracy).

Καταλήγουμε πως το feedforward νευρωνικό δίκτυο με 1 κρυφό επίπεδο, δεν είναι καλό μοντέλο για το παρόν έργο, όπως αναμενόταν.

6.2.6. Δύο Κρυφά Επίπεδα - Τυχαία Αναζήτηση

Για τα πειράματα με πολλαπλά κρυφά επίπεδα, επειδή ο χώρος των υπερπαραμέτρων είναι πολύ μεγάλος και είναι αδύνατο να γίνουν πειράματα με όλους τους δυνατούς συνδυασμούς, αυτές επιλέχθηκαν τυχαία για κάθε μοντέλο. Ύστερα θα επιχειρήσουμε να εξάγουμε συμπεράσματα από τα τυχαία τρεξίματα.

6.2.6α) Τονική (13 κατηγορίες)

#neurons per layer	N	features	#features	Train Accuracy	Dev Accuracy	Test Accuracy
12-24	2	pitches	60	61.07%	64.87%	58.80%
96-48	2	pitches	60	64.10%	65.30%	58.68%
30-30	5	pitches	132	64.77%	64.93%	61.84%
100-100	7	pitches	180	65.54%	64.60%	59.77%
50-100	6	pitches	156	65.14%	64.97%	61.50%

50-50	6	all	377	65.75%	64.79%	61.17%
100-50	6	all	377	64.58%	64.63%	61.06%
25-50	7	all	435	65.62%	65.12%	60.99%
100-50	4	all	261	65.37%	65.43%	64.63%
100-50	3	all	203	67.48%	65.42%	59.99%

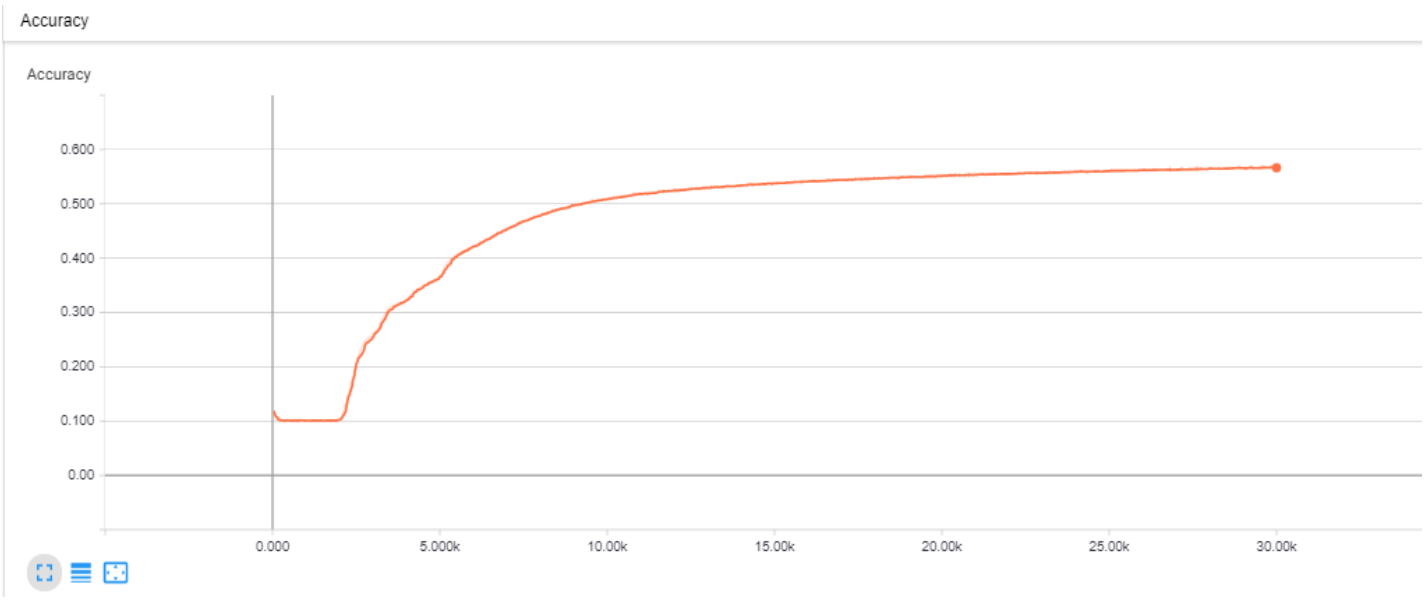
6.2.6β) Ματζόρε-μινόρε (25 κατηγορίες):

#neurons per layer	N	features	#features	Train Accuracy	Dev Accuracy	Test Accuracy
12-24	2	pitches	60	53.64%	56.05%	50.85%
96-48	2	pitches	60	57.27%	56.68%	53.68%
30-30	5	pitches	132	54.30%	56.07%	52.78%
100-100	5	pitches	180	56.65%	58.28%	54.54%
200-200	6	pitches	156	57.68%	57.19%	55.01%
512-512	4	pitches	108	59.36%	58.99%	55.15%
512-512	5	*pdc*[11]	154	60.60%	63.48%	55.54%
512-512	5	all	319	62.28%	61.12%	54.53%
50-50	6	all	377	58.50%	59.79%	53.53%

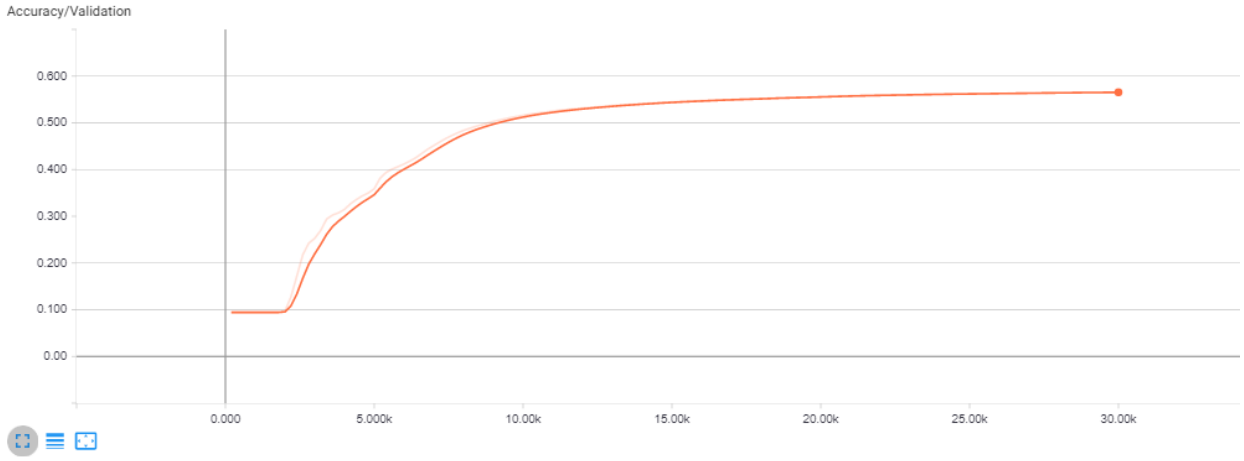
[11] *Pdc: pitches, duration, confidence*

Διαγράμματα - Οπτικοποιήσεις

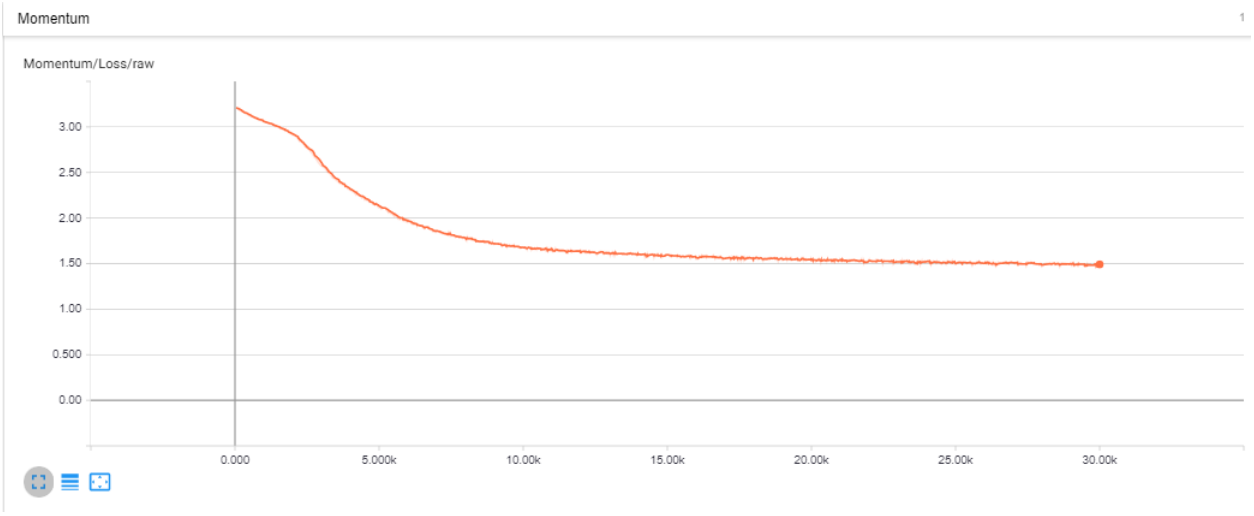
2 Layers, 100 neurons per layer, N=5, pitches,1000 epochs, 30000 training steps:



(πάνω) Ακρίβεια στο Train Set



(πάνω) Ακρίβεια στο Test Set



(πάνω) Τιμή Συνάρτησης Απώλειας στο Train Set

6.2.7 Πολλά Κρυφά Επίπεδα - Ματζόρε Μινόρε (25 κατηγορίες)

6.2.7α) Αρχικό Σύνολο Εκπαίδευσης

#neurons per layer	N	features	#features	Train Accuracy	Dev Accuracy	Test Accuracy
60-60-60	5	pitches	132	58.30%	58.93%	55.67%
60-60-60	5	All (-)timbre	187	60.67%	59.32%	55.16%
512-512--512	2	pitches	60	55.41%	55.64%	52.56%
512-512--512	5	pitches	132	59.58%	58.95%	54.93%
512-512--512	15	pitches	372	64.40%	61.83%	53.61%

6.2.7β) Επαυξημένο - Φιλτραρισμένο Σύνολο Εκπαίδευσης

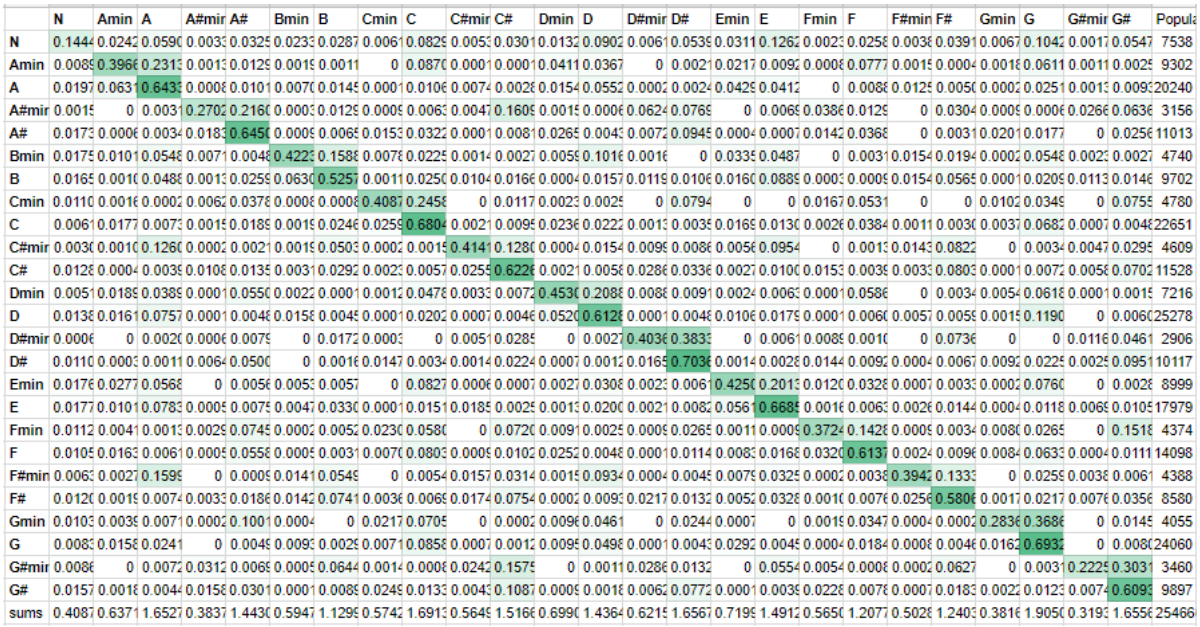
#neurons per layer	N	features	#features	Train Accuracy	Dev Accuracy	Test Accuracy
60-60-60	2	pitches	60	52.47%	49.91%	51.04%
60-60-60	5	pitches	60	54.07%	55.13%	54.46%
120-120	5	pitches	60	54.4%	53.34%	52.31%

Ενώ η απόδοση στο Test Set δεν είναι καλύτερη αν για σύνολο εκπαίδευσης χρησιμοποιηθεί το φιλτραρισμένο - επαυξημένο σύνολο ως Train Set, το μοντέλο συμπεριφέρεται αρκετά διαφορετικά από ένα εκπαιδευμένο στο αρχικό σύνολο δεδομένων όπως και με ένα κρυφό επίπεδο. Αυτό φαίνεται εύκολα από τους προκύπτοντες πίνακες σύγχυσης ενός από τα μοντέλα εκπαιδευμένο α) στο αρχικό σύνολο δεδομένων, β) στο φιλτραρισμένο-επαυξημένο.

Το μοντέλο που χρησιμοποιήθηκε ήταν με τρία κρυφά επίπεδα, 60 νευρώνες ανά επίπεδο, pitches ως χαρακτηριστικά, N=5 (συνολικό πλήθος χαρακτηριστικών=132), για τη ματζόρε μινόρε κατηγοριοποίηση.

6.2.8 Αξιολόγηση Πολλών Κρυφών Επιπέδων

α) Εκπαιδευμένο στο αρχικό σύνολο εκπαίδευσης - προβλέψεις στο Test Set. (55.16% Test Accuracy)



	N	Amin	A	A#mir	A#	Bmin	B	Cmin	C	C#mir	C#	Dmin	D	D#mir	D#	Emin	E	Fmin	F	F#min	F#	Gmin	G	G#mir	G#	Popula
N	0.1444	0.0242	0.0590	0.0033	0.0325	0.0233	0.0287	0.0061	0.0825	0.0055	0.0301	0.0132	0.0902	0.0061	0.0536	0.0311	0.1262	0.0023	0.0256	0.0038	0.0391	0.0067	0.1042	0.0017	0.0547	7538
Amin	0.0085	0.3966	0.2313	0.0013	0.0129	0.0019	0.0011	0	0.0870	0.0001	0.0001	0.0411	0.0367	0	0.0021	0.0217	0.0092	0.0008	0.0777	0.0015	0.0004	0.0018	0.0611	0.0011	0.0025	9302
A	0.0197	0.0631	0.6433	0.0008	0.0101	0.0070	0.0145	0.0001	0.0106	0.0074	0.0028	0.0154	0.0552	0.0002	0.0024	0.0429	0.0412	0	0.0088	0.0125	0.0050	0.0002	0.0251	0.0013	0.0093	20240
A#mir	0.0015	0	0.0031	0.2702	0.2160	0.0003	0.0129	0.0009	0.0063	0.0047	0.1609	0.0015	0.0006	0.0624	0.0769	0	0.0069	0.0386	0.0129	0	0.0304	0.0009	0.0006	0.0266	0.0636	3156
A#	0.0173	0.0006	0.0034	0.0183	0.6450	0.0009	0.0065	0.0153	0.0322	0.0001	0.0081	0.0265	0.0043	0.0072	0.0945	0.0004	0.0007	0.0142	0.0368	0	0.0031	0.0201	0.0177	0	0.0256	11013
Bmin	0.0175	0.0101	0.0548	0.0071	0.0048	0.4223	0.1588	0.0078	0.0225	0.0014	0.0027	0.0059	0.1016	0.0016	0	0.0335	0.0487	0	0.0031	0.0154	0.0194	0.0002	0.0548	0.0023	0.0027	4740
B	0.0165	0.0010	0.0488	0.0013	0.0259	0.0630	0.5257	0.0011	0.0250	0.0104	0.0166	0.0004	0.0157	0.0119	0.0106	0.0160	0.0889	0.0003	0.0009	0.0154	0.0565	0.0001	0.0205	0.0113	0.0146	9702
Cmin	0.0110	0.0016	0.0002	0.0062	0.0378	0.0008	0.0008	0.4087	0.2458	0	0.0117	0.0023	0.0025	0	0.0794	0	0	0.0167	0.0531	0	0	0.0102	0.0349	0	0.0755	4780
C	0.0061	0.0177	0.0073	0.0015	0.0189	0.0019	0.0246	0.0255	0.6804	0.0021	0.0095	0.0236	0.0222	0.0013	0.0035	0.0165	0.0130	0.0026	0.0384	0.0011	0.0030	0.0037	0.0682	0.0007	0.0048	22651
C#mir	0.0030	0.0010	0.1260	0.0002	0.0021	0.0019	0.0503	0.0002	0.0015	0.4141	0.1280	0.0004	0.0154	0.0099	0.0086	0.0056	0.0954	0	0.0013	0.0143	0.0822	0	0.0034	0.0047	0.0295	4609
C#	0.0128	0.0004	0.0035	0.0108	0.0135	0.0031	0.0292	0.0023	0.0057	0.0255	0.6226	0.0021	0.0058	0.0286	0.0336	0.0027	0.0100	0.0153	0.0035	0.0033	0.0803	0.0001	0.0072	0.0058	0.0702	11528
Dmin	0.0051	0.0185	0.0385	0.0001	0.0550	0.0022	0.0001	0.0012	0.0478	0.0033	0.0072	0.4530	0.2088	0.0088	0.0091	0.0024	0.0063	0.0001	0.0586	0	0.0034	0.0054	0.0618	0.0001	0.0015	7216
D	0.0138	0.0161	0.0757	0.0001	0.0048	0.0158	0.0045	0.0001	0.0202	0.0007	0.0046	0.0520	0.6128	0.0001	0.0048	0.0106	0.0175	0.0001	0.0060	0.0057	0.0055	0.0015	0.1190	0	0.0060	25278
D#mir	0.0006	0	0.0020	0.0006	0.0079	0	0.0172	0.0003	0	0.0051	0.0285	0	0.0027	0.4036	0.3833	0	0.0061	0.0085	0.0010	0	0.0736	0	0	0.0116	0.0461	2906
D#	0.0110	0.0003	0.0011	0.0064	0.0500	0	0.0016	0.0147	0.0034	0.0014	0.0224	0.0007	0.0012	0.0165	0.7036	0.0014	0.0028	0.0144	0.0092	0.0004	0.0067	0.0092	0.0225	0.0025	0.0951	10117
Emin	0.0176	0.0277	0.0568	0	0.0056	0.0053	0.0057	0	0.0827	0.0006	0.0007	0.0027	0.0308	0.0023	0.0061	0.4250	0.2013	0.0120	0.0328	0.0007	0.0033	0.0002	0.0760	0	0.0028	8999
E	0.0177	0.0101	0.0783	0.0005	0.0075	0.0047	0.0330	0.0001	0.0151	0.0185	0.0025	0.0013	0.0200	0.0021	0.0082	0.0561	0.6685	0.0016	0.0063	0.0026	0.0144	0.0004	0.0118	0.0069	0.0105	17979
Fmin	0.0112	0.0041	0.0013	0.0029	0.0745	0.0002	0.0052	0.0230	0.0580	0	0.0720	0.0091	0.0025	0.0005	0.0265	0.0011	0.0009	0.3724	0.1428	0.0009	0.0034	0.0080	0.0265	0	0.1518	4374
F	0.0105	0.0163	0.0061	0.0005	0.0558	0.0005	0.0031	0.0070	0.0803	0.0009	0.0102	0.0252	0.0048	0.0001	0.0114	0.0083	0.0168	0.0320	0.6137	0.0024	0.0096	0.0084	0.0633	0.0004	0.0111	14098
F#min	0.0063	0.0027	0.1599	0	0.0009	0.0141	0.0549	0	0.0054	0.0157	0.0314	0.0015	0.0934	0.0004	0.0045	0.0075	0.0325	0.0002	0.0038	0.3942	0.1333	0	0.0259	0.0038	0.0061	4388
F#	0.0120	0.0015	0.0074	0.0033	0.0186	0.0142	0.0741	0.0036	0.0065	0.0174	0.0754	0.0002	0.0093	0.0217	0.0132	0.0052	0.0328	0.0010	0.0076	0.0256	0.5806	0.0017	0.0217	0.0076	0.0356	8580
Gmin	0.0103	0.0036	0.0071	0.0002	0.1001	0.0004	0	0.0217	0.0705	0	0.0002	0.0096	0.0461	0	0.0244	0.0007	0	0.0019	0.0347	0.0004	0.0002	0.2836	0.3686	0	0.0145	4055
G	0.0083	0.0158	0.0241	0	0.0049	0.0093	0.0025	0.0071	0.0858	0.0007	0.0012	0.0095	0.0498	0.0001	0.0043	0.0292	0.0045	0.0004	0.0184	0.0008	0.0046	0.0162	0.6932	0	0.0080	24060
G#mir	0.0086	0	0.0072	0.0312	0.0069	0.0005	0.0644	0.0014	0.0008	0.0242	0.1575	0	0.0011	0.0286	0.0132	0	0.0554	0.0054	0.0008	0.0002	0.0627	0	0.0031	0.2225	0.3031	3460
G#	0.0157	0.0018	0.0044	0.0158	0.0301	0.0001	0.0089	0.0245	0.0133	0.0043	0.1087	0.0009	0.0018	0.0062	0.0772	0.0001	0.0039	0.0228	0.0078	0.0007	0.0183	0.0022	0.0123	0.0074	0.6093	9897
sums	0.4087	0.6371	1.6527	0.3837	1.4430	0.5947	1.1295	0.5742	1.6913	0.5645	1.5166	0.6990	1.4364	0.6215	1.6567	0.7195	1.4912	0.5650	1.2077	0.5028	1.2403	0.3816	1.9050	0.3193	1.6556	25466

β) Εκπαιδευμένο στο Επαυξημένο Φιλτραρισμένο του αρχικού συνόλου εκπαίδευσης - προβλέψεις στο Test Set (54.46% Accuracy).

	N	Amin	A	A#mir	A#	Bmin	B	Cmin	C	C#mir	C#	Dmin	D	D#mir	D#	Emin	E	Fmin	F	F#min	F#	Gmin	G	G#mir	G#	Popula
N	0.1594	0.0474	0.0457	0.0168	0.0092	0.0359	0.0180	0.0486	0.0695	0.0585	0.0359	0.0279	0.0664	0.0208	0.0250	0.0525	0.0648	0.0107	0.0205	0.0325	0.0408	0.0236	0.0332	0.0165	0.0177	7538
Amin	0.0092	0.6453	0.1147	0.0067	0.0019	0.0068	0.0006	0.0033	0.0482	0.0031	0.0015	0.0239	0.0327	0.0005	0.0006	0.0138	0.0045	0.0062	0.0391	0.0080	0.0002	0.0041	0.0193	0.0015	0.0032	9302
A	0.0183	0.1232	0.5726	0.0051	0.0028	0.0203	0.0078	0.0011	0.0080	0.0414	0.0048	0.0130	0.0314	0.0013	0.0009	0.0406	0.0218	0.0006	0.0063	0.0527	0.0036	0.0027	0.0086	0.0036	0.0055	20240
A#mir	0.0015	0.0015	0.0038	0.6166	0.0294	0	0.0038	0.0066	0.0050	0.0063	0.0868	0.0006	0.0003	0.0535	0.0193	0	0.0012	0.0427	0.0063	0	0.0155	0.0025	0	0.0643	0.0313	3156
A#	0.0130	0.0062	0.0016	0.1733	0.4295	0.0009	0.0027	0.0527	0.0269	0.0016	0.0127	0.0292	0.0009	0.0175	0.0461	0.0006	0.0003	0.0614	0.0437	0.0003	0.0037	0.0631	0.0015	0.0008	0.0088	11013
Bmin	0.0320	0.0168	0.0244	0.0196	0	0.5413	0.0909	0.0145	0.0111	0.0238	0.0149	0.0086	0.0386	0.0027	0	0.0293	0.0234	0.0012	0.0027	0.0637	0.0149	0.0004	0.0196	0.0046	0	4740
B	0.0139	0.0025	0.0269	0.0174	0.0096	0.0984	0.4296	0.0073	0.0112	0.0459	0.0216	0.0016	0.0102	0.0232	0.0035	0.0161	0.0611	0.0034	0.0013	0.0659	0.0778	0.0016	0.0063	0.0343	0.0080	9702
Cmin	0.0069	0.0033	0	0.0175	0.0104	0.0039	0	0.6855	0.0960	0.0046	0.0150	0.0108	0.0002	0.0039	0.0349	0	0	0.0294	0.0194	0.0002	0.0006	0.0248	0.0048	0.0002	0.0263	4780
C	0.0114	0.0403	0.0032	0.0044	0.0063	0.0048	0.0211	0.0620	0.6484	0.0066	0.0120	0.0335	0.0132	0.0026	0.0010	0.0225	0.0045	0.0075	0.0334	0.0044	0.0037	0.0164	0.0316	0.0026	0.0015	22651
C#mir	0.0032	0.0015	0.0822	0.0015	0.0002	0.0004	0.0338	0.0019	0.0017	0.6161	0.1019	0.0006	0.0017	0.0158	0.0025	0.0013	0.0286	0	0.0004	0.0316	0.0548	0	0	0.0069	0.0101	4609
C#	0.0057	0.0002	0.0025	0.0421	0.0005	0.0040	0.0115	0.0078	0.0037	0.0628	0.6266	0.0007	0.0014	0.0397	0.0098	0.0021	0.0053	0.0346	0.0017	0.0084	0.0773	0.0013	0.0023	0.0130	0.0340	11528
Dmin	0.0063	0.0475	0.0271	0.0063	0.0159	0.0056	0.0002	0.0101	0.0461	0.0163	0.0099	0.5697	0.1083	0.0130	0.0036	0.0070	0.0045	0.0022	0.0445	0.0034	0.0023	0.0307	0.0169	0.0004	0.0006	7216
D	0.0162	0.0289	0.0687	0.0007	0.0013	0.0312	0.0020	0.0034	0.0164	0.0141	0.0069	0.0981	0.5320	0.0014	0.0030	0.0209	0.0104	0.0016	0.0062	0.0564	0.0046	0.0142	0.0583	0.0006	0.0013	25278
D#mir	0.0006	0	0.0010	0.0113	0.0041	0	0.0072	0.0010	0	0.0103	0.0227	0	0.0013	0.6896	0.1424	0	0.0034	0.0175	0.0013	0.0024	0.0509	0.0006	0	0.0203	0.0113	2906
D#	0.0196	0.0025	0.0017	0.0466	0.0244	0	0.0010	0.0610	0.0003	0.0031	0.0256	0.0025	0.0011	0.0881	0.5775	0.0036	0.0012	0.0352	0.0052	0.0017	0.0065	0.0266	0.0028	0.0095	0.0504	10117
Emin	0.0346	0.0537	0.0530	0.0021	0.0015	0.0141	0.0047	0.0024	0.0610	0.0095	0.0035	0.0085	0.0238	0.0058	0.0022	0.4934	0.1146	0.0341	0.0165	0.0136	0.0024	0.0023	0.0407	0.0003	0.0004	8999
E	0.0308	0.0210	0.0643	0.0037	0.0017	0.0147	0.0235	0.0010	0.0110	0.1017	0.0043	0.0035	0.0134	0.0077	0.0045	0.0950	0.5163	0.0046	0.0068	0.0223	0.0152	0.0009	0.0030	0.0223	0.0053	17979
Fmin	0.0052	0.0057	0.0009	0.0333	0.0052	0	0.0011	0.0603	0.0358	0.0029	0.0534	0.0045	0	0.0086	0.0091	0.0004	0.0002	0.6266	0.0537	0.0020	0.0034	0.0189	0.0084	0.0054	0.0537	4374
F	0.0109	0.0475	0.0050	0.0123	0.0160	0.0005	0.0009	0.0242	0.0863	0.0021	0.0117	0.0430	0.0019	0.0008	0.0078	0.0087	0.0110	0.0822	0.5646	0.0070	0.0108	0.0213	0.0196	0.0009	0.0017	14098
F#min	0.0045	0.0015	0.0758	0.0015	0	0.0175	0.0227	0.0002	0.0004	0.0337	0.0280	0.0011	0.0278	0.0082	0.0002	0.0095	0.0088	0.0006	0.0027	0.6545	0.0763	0.0015	0.0097	0.0102	0.0013	4388
F#	0.0093	0.0030	0.0044	0.0285	0.0031	0.0120	0.0375	0.0074	0.0044	0.0336	0.0639	0.0013	0.0038	0.0320	0.0010	0.0030	0.0152	0.0075	0.0044	0.1136	0.5725	0.0043	0.0036	0.0158	0.0136	8580
Gmin	0.0108	0.0212	0.0064	0.0009	0.0286	0.0017	0	0.0495	0.0530	0.0022	0.0009	0.0231	0.0352	0	0.0165	0.0017	0	0.0128	0.0332	0.0133	0.0004	0.5807	0.1008	0	0.0061	4055
G	0.0163	0.0448	0.0217	0.0011	0.0010	0.0240	0.0022	0.0209	0.1063	0.0054	0.0029	0.0272	0.0410	0.0007	0.0024	0.0495	0.0010	0.0015	0.0160	0.0144	0.0046	0.0573	0.5316	0.0007	0.0036	24060
G#mir	0.0112	0.0005	0.0037	0.0887	0.0005	0.0008	0.0271	0.0037	0.0002	0.0572	0.0930	0	0.0002	0.0265	0.0078	0	0.0242	0.0170	0.0005	0.0104	0.0624	0.0023	0.0002	0.3817	0.1789	3460
G#	0.0090	0.0026	0.0036	0.0472	0.0098	0.0002	0.0014	0.0618	0.0044	0.0186	0.1175	0.0009	0.0005	0.0256	0.0447	0	0.0017	0.0709	0.0026	0.0027	0.0219	0.0056	0.0013	0.0275	0.5170	9897
sums	0.4614	1.1713	1.2160	1.2064	0.6142	0.8400	0.7522	1.1998	1.3564	1.1824	1.3785	0.9354	0.9884	1.0906	0.9673	0.8725	0.9293	1.1138	0.9349	1.1866	1.1280	0.9090	0.9251	0.6448	0.9937	25466

6.2.9. Συμπεράσματα Feedforward

Για το Feedforward φαίνεται η ακρίβεια να μην υπερβαίνει το 55% στο Test Set για τη ματζόρε μινόρε κατηγοριοποίηση, ανεξαρτήτως του μεγέθους του μοντέλου και συνόλου εκπαίδευσης. Αυτό μπορεί να οφείλεται σε κάποιο βαθμό στο γεγονός πως το N δεν μπορούσε να πάρει μεγάλες τιμές (segments, τα χαρακτηριστικά των οποίων προστίθενται στην αρχή και στο τέλος του διανύσματος χαρακτηριστικών), λόγω περιορισμού υπολογιστικών πόρων. Όσο αυξάνει το N, αυξάνει ο όγκος των δεδομένων καθώς και η απαιτούμενη πολυπλοκότητα του μοντέλου για να συνδυάσει τα χαρακτηριστικά και να παράξει πρόβλεψη. Τα υπόλοιπα μοντέλα (CNN, RNN) αναμένεται να έχουν καλύτερη απόδοση καθώς μπορούν αποδοτικότερα να αξιοποιήσουν την πληροφορία των γειτονικών segments.

Επιπρόσθετα, υπάρχει η πιθανότητα ο περιορισμός να προκύπτει σε μεγάλο βαθμό από τα ίδια τα δεδομένα και τα αναπόφευχτα σφάλματα. Θα μπορούμε να αποφανθούμε στο τέλος όλων των πειραμάτων.

6.3 Συνελικτικό Νευρωνικό Δίκτυο

Ο επόμενος τύπος νευρωνικού δικτύου που εκπαιδεύτηκε για αναγνώριση συγχορδιών είναι το συνελικτικό νευρωνικό δίκτυο (Convolutional Neural Network). Ο τρόπος που λειτουργεί βασίζεται στη λειτουργία του απλούστερου Feedforward νευρωνικού που εξετάσαμε στην προηγούμενη παράγραφο. Η διαφορά τους είναι πως αντικαθίσταται το στάδιο του γραμμικού συνδυασμού των εισόδων, με συνέλιξη φίλτρων με τις εισόδους. Η συνέλιξη [19] είναι μια πράξη μεταξύ δύο συναρτήσεων (σε μορφή πινάκων στην περίπτωσή μας) το αποτέλεσμα της οποίας είναι βαθμωτό και εκφράζει μια μορφή συσχέτισης της εισόδου με το φίλτρο. Τα συνελικτικά νευρωνικά δίκτυα χρησιμοποιούνται ευρέως για πληθώρα έργων και σε πολλά από αυτά έχουν την καλύτερη απόδοση σε σχέση με άλλα μοντέλα [20]. Οι παράμετροι που μαθαίνονται από το σύστημα είναι οι τιμές σε κάθε θέση του κάθε φίλτρου.

Πλεονέκτημα του συνελικτικού νευρωνικού έναντι του Feedforward είναι η καλύτερη αξιοποίηση υπολογιστικών πόρων, αφού σε κάθε στάδιο η συνέλιξη λαμβάνει υπόψη ένα μικρό μέρος της εισόδου αντί για ολόκληρη την είσοδο (όπως το feedforward). Προϋποθέτει βέβαια πως έχει νόημα η "κοντινότητα" των χαρακτηριστικών εισόδου. Για παράδειγμα σε ένα σύστημα αναγνώρισης εικόνας, έχει νόημα να αναφερθούμε σε "κοντινά pixels", ενώ σε ένα σύστημα που προβλέπει αξίες ακινήτων με βάση την περιοχή, τον αριθμό δωματίων κλπ, δεν έχει νόημα (ο αριθμός δωματίων είναι "κοντά" ή "μακριά" από την τοποθεσία του ακινήτου σαν χαρακτηριστικό;).

Στην περίπτωση της αναγνώρισης συγχορδίας με τα διαθέσιμα δεδομένα, η κοντινότητα έχει νόημα σε δύο άξονες: α) Στο χρόνο, προφανώς segments που ακούγονται διαδοχικά είναι κοντά μεταξύ τους. β) Στα pitches (κωδικοποιημένη συχνότητα). Έχει νόημα να πούμε πως η Ντο είναι κοντά (ή μακριά από) στην Ντο δίεση.

Επιχειρήθηκαν δύο διαφορετικές αρχιτεκτονικές του δικτύου, βασισμένες στους παραπάνω άξονες. Στην πρώτη περίπτωση τα δεδομένα οργανώνονται με τέτοιο τρόπο ώστε να εφαρμόζεται συνέλιξη στο πεδίο του χρόνου, ενώ στη δεύτερη στα πεδία του χρόνου και της συχνότητας (των pitches). (βλ. Προεπεξεργασία Δεδομένων - Convolutional).

Ένα συνελικτικό επίπεδο λαμβάνει είσοδο διαστάσεων:

$$input\ shape := [W_1,\ H_1\ , D_1]$$

Αποτελείται από F φίλτρα διαστάσεων

$$filter\ shape := [W_f,\ H_f]$$

Τα φίλτρα εφαρμόζονται σε ολόκληρη την είσοδο, διανύοντας την με βήμα S.

Στην είσοδο εφαρμόζεται "padding", με τη μορφή μηδενικών στα άκρα της, ώστε η έξοδος του συνελικτικού επιπέδου να έχει τις επιθυμητές διαστάσεις. Τοποθετούνται P μηδενικά σε κάθε άκρο της εισόδου

Τελικά η έξοδος του επιπέδου θα έχει διαστάσεις:

$output\ shape := [W_2, H_2, D_2]$

Όπου: $W_2 = \frac{W_1 - W_f + 2P}{S} + 1$

$H_2 = \frac{H_1 - H_f + 2P}{S} + 1$

$D_2 = F$

6.3.1 Συνέλιξη στο Πεδίο του Χρόνου

Είσοδος διαστάσεων [n, 1, 1+2*N, m]

Όπου n ο αριθμός των παραδειγμάτων εισόδου, N το πλήθος των προηγούμενων και επόμενων segments που συνελίσσονται και m το πλήθος των χαρακτηριστικών

Πρώτο συνελικτικό επίπεδο:

Φίλτρα διαστάσεων [1,N]

Έξοδος διαστάσεων [n, 1, N+2, f]

Όπου f ο αριθμός των φίλτρων

Δεύτερο συνελικτικό επίπεδο:

Φίλτρα διαστάσεων [1,N]

Έξοδος διαστάσεων [n,1,3,f]

Τρίτο συνελικτικό επίπεδο:

Φίλτρα διαστάσεων [1,3]

Έξοδος διαστάσεων [n, 1, 1, f]

Η έξοδος του τελευταίου συνελικτικού επιπέδου αποτελείται από f αριθμούς για κάθε παράδειγμα εισόδου. Αυτοί τροφοδοτούνται σε ένα πλήρως συνδεδεμένο επίπεδο (fully connected) με συνάρτηση ενεργοποίησης Softmax, η έξοδος του οποίου αποτελεί την πρόβλεψη.

A) N=3 F=16/32/64 m=12 (pitches) - Unfiltered Dataset

Input --> (?, 1, 7, 12)

Conv2D_1 --> (?, 1, 5, F)

Conv2D_2 --> (?, 1, 3, F)

Conv2D_3 --> (?, 1, 1, F)

Flatten --> (?, F)

FullyConnected --> (?, 25)

Num Filters	**Train Accuracy**	**Dev Accuracy**	**Test Accuracy**
16	52.12%	52.63%	54.01%
32	53.21%	51.66%	55.12%
64	55.29%	54.11%	55.20%

B) N=5 F=16/32/64 m=12 (pitches) - Unfiltered Dataset

Input --> (?, 1, 11, 12)

Conv2D_1 --> (?, 1, 7, F)

Conv2D_2 --> (?, 1, 3, F)

Conv2D_3 --> (?, 1, 1, F)

Flatten --> (?, F)

FullyConnected --> (?, 25)

Num Filters	**Train Accuracy**	**Dev Accuracy**	**Test Accuracy**
16	55.01%	55.23%	56.08%
32	54.60%	54.02%	56.35%
64	55.64%	54.82%	56.27%

Γ) N=33, F=128, m=12/24/, (all)

Input --> (505002, 1, 67, 12)

Conv2D_1 --> (?, 1, 35, 128)

Conv2D_2 --> (?, 1, 3, 128)

Conv2D_3 --> (?, 1, 1, 128)

Flatten --> (?, 128)

fc --> (?, 25)

Features	**Train Accuracy**	**Dev Accuracy**	**Test Accuracy**
all	55.61%	55.82%	55.30%

Ενώ τα παραπάνω μοντέλα εκπαιδεύτηκαν γρηγορότερα και το N μπορούσε να πάρει αρκετά μεγαλύτερες τιμές απ'ότι τα Feedforward, η απόδοση δεν είναι σημαντικά καλύτερη ανεξαρτήτως μεγέθους μοντέλου. Επίσης εξ'αιτίας της οργάνωσης των δεδομένων, κάποιο μοντέλο δύσκολα αξιοποιεί πληροφορία από τη συσχέτιση μεταξύ των διαφορετικών χαρακτηριστικών, αφού η συνέλιξη γίνεται στη διάσταση μόνο ενός χαρακτηριστικού κάθε φορά.

6.3.2. Συνέλιξη στα Πεδία του Χρόνου και των Pitches

Είσοδος διαστάσεων [n, 12, 2*N+1, 1]

6.3.2α) Ένα συνελικτικό Επίπεδο

Για διάφορες τιμές του μεγέθους των φίλτρων (Wf, Hf), του αριθμού των γειτονικών segments (N) και του πλήθους των φίλτρων προέκυψαν τα παρακάτω αποτελέσματα. Στη συνέχεια παρουσιάζονται αναλυτικότερα οι διαστάσεις των ενδιάμεσων επιπέδων του δικτύου.

Ματζόρε Μινόρε - Αρχικό Σύνολο Εκπαίδευσης

(Wf, Hf)	F	N	Train Accuracy	Dev Accuracy	Test Accuracy
(12,3)	128	3	50.44%	50.12%	52.64%
(12,7)	128	3	51.24%	51.44%	53.21%
(6,3)	64	3	51.92%	53.58%	53.87%
(12,3)	64	11	56.27%	55.78%	55.94%
(12,7)	64	11	56.08%	56.03%	55.96%
(12,7)	128	11	57.15%	57.93%	55.65%
(12,23)	32	11	56.76%	56.25%	55.45%
(12,23)	128	11	55.70%	54.67%	55.26%
(12,23)	32	21	57.30%	55.62%	56.39%

Ενδεικτικά παρουσιάζονται οι ενδιάμεσες διαστάσεις για κάποια από τα παραπάνω μοντέλα. Σε αυτές φαίνονται τα πλήθη των εκπαιδεύσιμων παραμέτρων ανάλογα με τις υπερπαραμέτρους που επιλέγονται (Wf,Hf,N,F)

a) Wf=12, Hf=3, N=3, F=128

Input --> (493435, 12, 7, 1)

Conv2D_1 (12,3) --> (493435, 1, 5, 128)

Flatten --> (493435, 640)

fc --> (493435, 25)

b) Wf=12,Hf=7, N=3,F=128

Input --> (493435, 12, 7, 1)

Conv2D_1 (12,7) --> (493435, 1, 1, 128)

Flatten --> (493435, 128)

fc --> (493435, 25)

c) Wf=6, Hf=3, N=3,F=64

Input --> (493435, 12, 7, 1)

Conv2D_1 (6,3)--> (493435, 7, 5, 64)

Flatten --> (493435, 2240)

fc --> (493435, 25)

d) Wf=12, Hf=3, N=11, F=64

Input --> (493435, 12, 23, 1)

Conv2D_1 --> (493435, 1, 21, 64)

Flatten --> (493435, 1344)

fc --> (493435, 25)

e) Wf=12, Hf=7, N=11, F=64

Input --> (493435, 12, 23, 1)

Conv2D_1 --> (493435, 1, 17, 64)

Flatten --> (493435, 1088)

fc --> (493435, 25)

f) Wf=12, Hf= 7, N=11, F=128

Input --> (493435, 12, 23, 1)

Conv2D_1 --> (493435, 1, 17, 128)

Flatten --> (493435, 2176)

fc --> (493435, 25)

e) Wf=12, Hf=23, N=11, F=32

Input --> (?, 12, 23, 1)

Conv2D_1 --> (?, 1, 1, 32)

Flatten --> (?, 32)

fc --> (?, 25)

6.3.2β) Πολλά Συνελικτικά Επίπεδα - Αρχικό Σύνολο Εκπαίδευσης

(W,H,F)_1	(W,H,F)_2	(W,H,F)_3	N	Train Accuracy	Dev Accuracy	Test Accuracy
(3,3,16)	(3,3,32)	(3,3,64)	5	56.08%	55.27%	56.89%
(3,3,16)	(3,3,32)	(3,3,64)	21	55.32%	54.94%	54.2%
(12,3,16)	(1,3,32)	(1,3,64)	5	55.16%	54.66%	57.00%

(12,3,16)	(1,3,32)	(1,3,64)	11	56.71%	55.96%	57.08%
(12,12,16)	(1,12,32)	(1,12,64)	21	55.38%	55.74%	55.98%

6.3.2γ) Πολλά Συνελικτικά Επίπεδα - Φιλτραρισμένο Σύνολο

Το Φιλτραρισμένο Σύνολο περιέχει ίδιο αριθμό παραδειγμάτων για κάθε κατηγορία (χωρίς επαύξηση). Εξαιτίας του μικρού μεγέθους του συνόλου δεδομένων, το μοντέλο αντιμετωπίζει με δυσκολία το overfitting (που δεν ήταν τόσο έντονο στο αντίστοιχο Feedforward).

Για το μοντέλο:

(W,H,F)_1	(W,H,F)_2	(W,H,F)_3	N
(12,12,16)	(1,12,32)	(1,12,64)	21

(αναγράφονται οι τιμές των accuracies στο βήμα με το καλύτερο Test Accuracy)

Regularizer	N	Train Accuracy	Test Accuracy
None	21	66.41%	35.72%
L2	21	67.04%	41.55%
L2,Dropout(0.5)	21	61.32%	45.87%
L2,Dropout(0.25)	21	49.06%	39.97%

Η μέθοδος Dropout εφαρμόστηκε στην έξοδο κάθε συνελικτικού επιπέδου

6.3.2δ) Πολλά συνελικτικά επίπεδα - Επαυξημένο Φιλτραρισμένο Σύνολο - (~1M segments)

Για το μοντέλο:

(W,H,F)_1	(W,H,F)_2	(W,H,F)_3	N

(12,12,16)	(1,12,32)	(1,12,64)	21

(αναγράφονται οι τιμές των accuracies στο βήμα με το καλύτερο Test Accuracy)

Regularizer	N	Train Accuracy	Test Accuracy
None	21	58.45%	51.60%
L2	21	49.10%	52.21%
L2,Dropout(0.5)	21	41.56%	49.05%
L2,Dropout(0.25)	21	40.53%	38.72%

Ενώ το μοντέλο όταν εκπαιδεύτηκε πάνω στο φιλτραρισμένο σύνολο δεδομένων δεν μπορούσε να αντιμετωπίσει το overfitting ακόμα και με πολύ ισχυρή συστηματοποίηση (regularization), όταν το σύνολο επαυξήθηκε η απόδοσή του αυξήθηκε σημαντικά και αντιμετωπίστηκε σε μέγαλο βαθμό το overfitting.

Για το μοντέλο:

(3,3,16)	(3,3,32)	(3,3,64)	21

Στο φιλτραρισμένο:

Regularizer	Train Accuracy	Test Accuracy
L2	72.09%	41.21%
L2+dropout(0.5)	66.52%	43.56%

Στο αρχικό:

max_pool*	Regularizer	Train Accuracy	Test Accuracy
No	L2+dropout(0.5)	54.73%	52.13%

Yes	L2+dropout(0.5)	53.66%	52.83%

Στα πειράματα με max_pool εισήχθει ένα max pool επίπεδο μετά από κάθε συνελικτικό επίπεδο. Το max pool επίπεδο εφαρμόζει ένα φίλτρο κάποιου μεγέθους [m,n] με κάποιο βήμα S, όπου η έξοδος του είναι το μέγιστο στοιχείο του πίνακα σε κάθε θέση του φίλτρου.

Στο επαυξημένο φιλτραρισμένο

max_pool	Regularizer	Train Accuracy	Test Accuracy
Yes	L2+dropout(0.5)	60.38%	50.45%

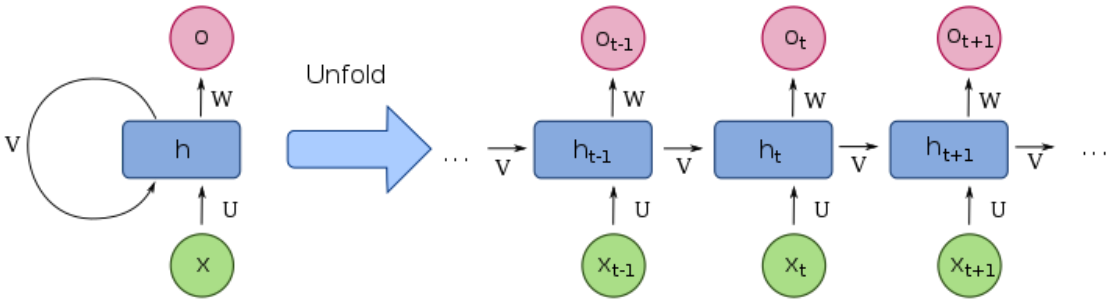
6.3.3 Συμπεράσματα Συνελικτικού Νευρωνικού

Τα συνελικτικά νευρωνικά δίκτυα που εκπαιδεύτηκαν, είχαν σε γενικές γραμμές ίδια επίδοση με τα αντίστοιχα Feedforward. Βέβαια, η εκπαίδευση στις περισσότερες περιπτώσεις γινόταν πολύ γρηγορότερα απ'ότι στο Feedforward. Προέκυψαν όμως διαφορετικά θέματα, όπως το Overfitting καθώς και η δυσκολία αξιοποίησης της συνέλιξης για όλα τα χαρακτηριστικά εισόδου.

6.4 Αναδρομικό Νευρωνικό Δίκτυο - LSTM (Σύστημα Μακράς Βραχυπρόθεσμης Μνήμης)

Το LSTM [21] είναι μια από τις πολλές παραλλαγές του Αναδρομικού Νευρωνικού Δικτύου (Recurrent Neural Network - RNN). Η διαφορά ενός RNN με ένα Feedforward νευρωνικό είναι πως το δίκτυο δεν αποτελείται από νευρώνες, αλλά από κελιά/κύτταρα (cells). Η δομή αυτών διαφέρει ανά τις παραλλαγές, αλλά σε όλα τα RNN, το κάθε κελί διατηρεί μια εσωτερική κατάσταση/μνήμη την οποία αξιοποιεί για την επεξεργασία αλληλουχιών δεδομένων. Αυτό κάνει τα RNN αποδοτικό μοντέλο για αναγνώριση συγχορδιών [37](από αλληλουχία ήχων σε

αλληλουχία συγχορδιών). Ένα απλό σχήμα από το άρθρο για τα RNN της wikipedia [36] δείχνει τη βασική λειτουργία του:

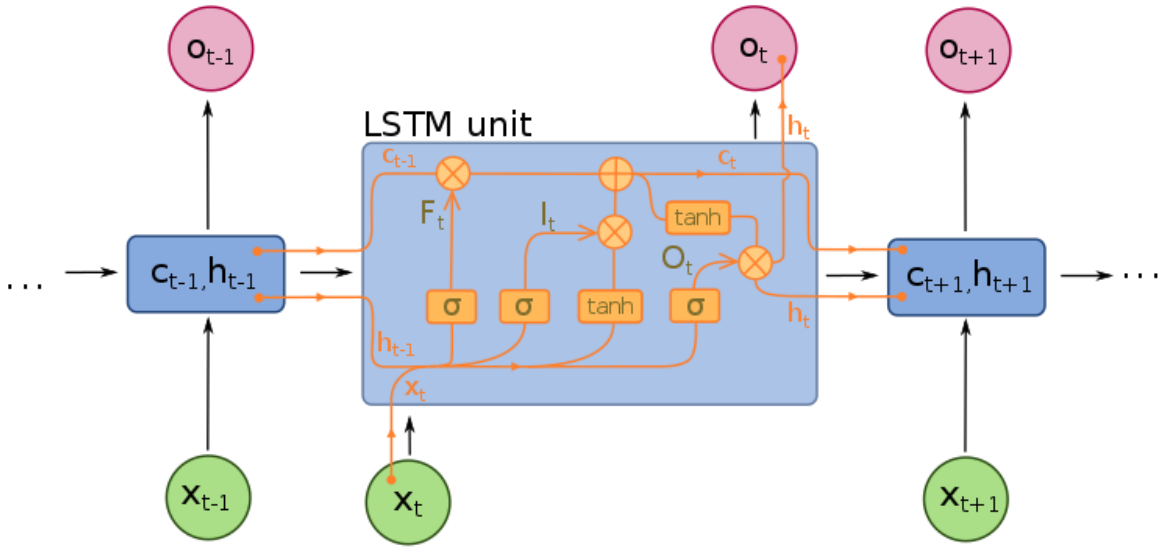


Στο παραπάνω σχήμα, το RNN λαμβάνει ως είσοδο μια ακολουθία X και παράγει έξοδο ακολουθία O. Στο δεξί μέρος του σχήματος φαίνεται η λειτουργία του σε κάθε βήμα της ακολουθίας (“ξεδιπλωμένο” στο χρόνο δίκτυο). Τα U,W,V είναι οι εκπαιδεύσιμες παράμετροι, και το h είναι η εσωτερική κατάσταση το κελιού. Η έξοδος καθορίζεται από την κρυφή κατάσταση και τα βάρη W. Η κρυφή κατάσταση καθορίζεται από την κρυφή κατάσταση του προυγούμενου βήματος της ακολουθίας και την είσοδο στο παρόν βήμα (και τα βάρη V, U).

Υπάρχουν δεκάδες παραλλαγές του RNN, όπου συνήθως εισάγεται πολυπλοκότητα στην εσωτερική λειτουργία του κελιού. Το LSTM είναι μία από τις πιο διαδεδομένες παραλλαγές που χρησιμοποιείται σε πληθώρα εφαρμογών. Αυτή χρησιμοποιήθηκε για τα πειράματα της παρούσας εργασίας.

Το LSTM αποτελείται από σχετικά πολύπλοκα κελιά, με τρεις πύλες το καθένα: την πύλη εισόδου, την πύλη εξόδου καθώς και την πύλη forget. Μια πύλη είναι ουσιαστικά ένας νευρώνας αποτελούμενος από παραμέτρους που μπορούν να βελτιστοποιηθούν για ελαχιστοποίηση μιας συνάρτησης κόστους, όπως στα υπόλοιπα νευρωνικά. Αυτό γίνεται συνήθως με παραλλαγές του αλγορίθμου Back Propagation, όπως και στα υπόλοιπα νευρωνικά. Κάθε πύλη έχει τις δικές της παραμέτρους, ξεχωριστές συναρτήσεις ενεργοποίησης και διαφορετικές εισόδους και εξόδους.

Χαρακτηριστικό του LSTM είναι πως διατηρεί δύο εσωτερικές καταστάσεις και η δομή του κυττάρου, κυρίως της πύλης forget, δίνει τη δυνατότητα στο δίκτυο να μαθαίνει τί πρέπει να "θυμάται" σε οποιοδήποτε στάδιο της ακολουθίας. Διαδεδομένο παράδειγμα είναι η μοντελοποίηση γλώσσας με LSTM [22]. Η δομή ενός κυττάρου φαίνεται στο παρακάτω διάγραμμα, από το ίδιο άρθρο της Wikipedia [36]:



Η είσοδος του κυττάρου x_t τη χρονική στιγμή t (για δεδομένα οργανωμένα σε χρονική ακολουθία) τροφοδοτείται στις διάφορες πύλες (F_t, I_t, O_t) και με βάση τις παραμέτρους των πυλών, και τις καταστάσεις c_t, h_t καθορίζεται η έξοδος o_t και οι καταστάσεις των κελιών στο επόμενο βήμα. Σε ένα LSTM επίπεδο (LSTM layer), οι εσωτερικές καταστάσεις είναι διανύσματα μήκους Λ, έτσι μετά από την επεξεργασία της εισόδου προκύπτει ως έξοδος του επιπέδου ένα διάνυσμα με Λ χαρακτηριστικά. Δίνεται επίσης η δυνατότητα να λάβουμε στην έξοδο ακολουθία ίδιου μήκους με την αρχική, που αποτελείται από τις εξόδους $O_t[\Lambda]$ για κάθε βήμα της ακολουθίας εισόδου. Τέλος δίνεται η δυνατότητα να λάβουμε στην έξοδο τις καταστάσεις C_t, H_t σε κάθε βήμα της ακολουθίας, ή μόνο την τελευταία κατάσταση (μετά από το πέρασμα όλης της ακολουθίας). Αυτές οι δυνατότητες θα αξιοποιηθούν στα παρακάτω μοντέλα.

Για το LSTM έγιναν δύο παραλλαγές της οργάνωσης των δεδομένων. Στην πρώτη τα δεδομένα χωρίζονται σε ακολουθίες σταθερού μήκους κάθε μία από τις οποίες αντιστοιχεί σε μία συγχορδία. Για να επιτευχθεί αυτό είτε προστίθεται zero-padding στις ακολουθίες με μικρότερο μήκος από το επιθυμητό, είτε απορρίπτονται κάποια segments από τις ακολουθίες με μεγαλύτερο μήκος από το επιθυμητό. Στη δεύτερη περίπτωση μια συγχορδία δεν αντιστοιχεί σε ολόκληρη την ακολουθία αλλά σε ένα από τα segments της. Στην περίπτωση αυτή χρησιμοποιούνται και οι ενδιάμεσες έξοδοι (O_t) για την πρόβλεψη, σε συνδυασμό με ένα πλήρως συνδεδεμένο επίπεδο με συνάρτηση ενεργοποίησης Softmax (για κατηγοριοποίηση).

Ως Test Set χρησιμοποιήθηκε το ίδιο (330 κομμάτια) με τα παραπάνω πειράματα.

6.4.1) Μία συγχορδία ανά ακολουθία

6.4.1α) Τονική (13 κατηγορίες) - Best Test Accuracy

#hidden	Sequence length	Features	#features	Train Accuracy	Dev Accuracy	Test Accuracy
10	10	pitches	12	64.45%	67.17%	69.20%
20	10	pitches	12	66.92%	62.68%	70.02%
20	20	all	29	67.11%	69.51%	61.24%
5	10	all	29	63.05%	59.90%	67.63%
15	10	all	29	67.74%	62.19%	70.14%

6.4.1β) Ματζόρε-Μινόρε (25 κατηγορίες) - Best Test Accuracy

#hidden	Sequence length	Features	#features	Train Accuracy	Dev Accuracy	Test Accuracy
10	10	pitches	12	61.26%	57.76%	59.67%

20	10	pitches	12	64.63%	59.01%	59.73%
20	20	pitches	12	63.40%	60.17%	60.13%
20	20	pitches,dur	13	65.34%	61.22%	59.85%
20	20	pitches,loud	20	64.42%	58.40%	60.15%
20	20	pitches,timbre	24	65.10%	61.76%	59.41%
20	20	all(-)timbre	17	64.48%	61.96%	60.40%
20	20	all	29	64.92%	55.98%	59.39%

6.4.1γ) Συμπεράσματα

Ενώ η απόδοση του συστήματος είναι αρκετά καλύτερη από τα υπόλοιπα μοντέλα που χρησιμοποιήθηκαν παραπάνω, αυτή δεν μπορεί τόσο εύκολα να αξιοποιηθεί κατά την πρόβλεψη, διότι συνοδεύεται από το πολύπλοκο πρόβλημα της βέλτιστης κατάτμησης ενός κομματιού σε αλληλουχίες, σε κάθε μία από τις οποίες αντιστοιχεί μια συγχορδία.

6.4.2 Μία συγχορδία ανά τμήμα (segment)

Ιδανικά θα θέλαμε το σύστημα να λαμβάνει σαν είσοδο ολόκληρο το κομμάτι ως αλληλουχία από segments. Πολλά κομμάτια όμως έχουν μεγάλη διάρκεια και θα προέκυπταν ακολουθίες από μερικές χιλιάδες Segments. Τόσο μεγάλα μήκη ακολουθίας δυσχεραίνουν τη διαδικασία της μάθησης στα αναδρομικά νευρωνικά δίκτυα. Αυτό συμβαίνει διότι κατά τον υπολογισμό της παραγώγου της συνάρτησης κόστους σε σχέση με τις διάφορες παραμέτρους στο πλαίσιο του αλγορίθμου Backpropagation αν εκτελείται κάποιος πολλαπλασιασμός σε κάθε βήμα της ακολουθίας, τότε η παράγωγος γίνεται εκθετική ως προς τα βήματα της ακολουθίας. Η αστάθεια αυτή οδηγεί σε παραγώγους με τεράστιες τιμές (*exploding gradients*) ή με πολύ μικρές τιμές (*vanishing gradients)* και το σύστημα δεν συγκλίνει.

Το LSTM αντιμετωπίζει σε κάποιο βαθμό το πρόβλημα αυτό, καθώς λόγω της δομής του κυττάρου LSTM εκτελούνται μόνο προσθέσεις μεταξύ των βημάτων της ακολουθίας. Η μάθηση όμως γίνεται πολύ πιο αργά με μεγάλα μήκη ακολουθίας και απαιτεί περισσότερους υπολογιστικούς πόρους ενώ αυξάνει ο κίνδυνος του *overfitting*, επομένως ο περιορισμός ως προς το μήκος της ακολουθίας εξακολουθεί να υπάρχει.

Για την εκπαίδευση απορρίφθηκαν οι ακολουθίες που περιείχαν μόνο μία συγχορδία (sequence balanced dataset).

α)Τονική (13 κατηγορίες) - Sequence Balanced Dataset

#hidden	Sequence length	#total sequences	Features	#features	Train Accuracy	Test Accuracy
10	10	58558	pitches	12	65.25%	62.77%
128	10	58558	pitches	12	66.5%	62.41%
20	10	58558	pitches	12	65.38%	64.60%
10	50	12922	pitches	12	51.45%	60.28%
128	50	12922	pitches	12	52.99%	61.39%

Τονική (13 κατηγορίες) - Augmented Sequence Balanced Dataset

#hidden	Sequence length	#total sequences	Features	#features	Train Accuracy	Test Accuracy
128	10	696,840	pitches	12	58.43	64.86%
128	50	153,883	pitches	12	52.40%	62.06%

Πίνακας Σύγχυσης - Αναγνώριση Τονικής (sequence length 10, num hidden 128 - Αρχικό Dataset

	A	A#	B	C	C#	D	D#	E	F	F#	G	G#	N	sums
A	0.68	0.01	0.02	0.03	0.01	0.08	0.00	0.09	0.02	0.02	0.03	0.01	0.01	29,117
A#	0.01	0.65	0.01	0.05	0.02	0.02	0.09	0.01	0.06	0.01	0.02	0.05	0.00	11,294
B	0.06	0.03	0.56	0.01	0.02	0.05	0.02	0.12	0.00	0.07	0.02	0.04	0.01	14,901
C	0.04	0.03	0.03	0.65	0.01	0.03	0.02	0.04	0.05	0.00	0.08	0.01	0.01	24,470
C#	0.05	0.03	0.03	0.01	0.51	0.01	0.05	0.08	0.02	0.09	0.00	0.10	0.01	14,840
D	0.10	0.02	0.03	0.03	0.00	0.66	0.00	0.04	0.02	0.02	0.08	0.00	0.01	28,880
D#	0.01	0.08	0.01	0.03	0.03	0.01	0.67	0.01	0.02	0.02	0.03	0.08	0.01	10,483
E	0.09	0.01	0.04	0.02	0.01	0.03	0.01	0.72	0.01	0.02	0.03	0.01	0.01	25,983
F	0.03	0.06	0.01	0.10	0.02	0.03	0.01	0.04	0.58	0.01	0.07	0.03	0.01	14,031
F#	0.05	0.02	0.08	0.01	0.07	0.04	0.02	0.06	0.01	0.58	0.01	0.04	0.01	13,383
G	0.05	0.02	0.02	0.09	0.00	0.10	0.01	0.04	0.02	0.01	0.62	0.00	0.01	26,689
G#	0.01	0.06	0.02	0.03	0.09	0.01	0.08	0.04	0.02	0.04	0.01	0.58	0.01	11,065
N	0.11	0.03	0.06	0.05	0.02	0.11	0.03	0.19	0.02	0.03	0.08	0.02	0.25	5,794.
														23093

Στο Επαυξημένο Σύνολο

	A	A#	B	C	C#	D	D#	E	F	F#	G	G#	N	sums
A	0.72	0.00	0.02	0.03	0.01	0.09	0.00	0.06	0.02	0.02	0.02	0.01	0.00	29,117
A#	0.01	0.61	0.00	0.07	0.04	0.05	0.05	0.00	0.07	0.02	0.02	0.07	0.00	11,294
B	0.04	0.02	0.58	0.01	0.03	0.07	0.01	0.09	0.00	0.09	0.02	0.03	0.00	14,901
C	0.03	0.01	0.03	0.76	0.01	0.05	0.01	0.01	0.04	0.01	0.04	0.02	0.00	24,470
C#	0.03	0.02	0.03	0.01	0.69	0.01	0.02	0.02	0.01	0.09	0.00	0.07	0.00	14,840
D	0.08	0.01	0.01	0.03	0.01	0.76	0.00	0.03	0.01	0.01	0.05	0.00	0.00	28,880
D#	0.00	0.05	0.01	0.05	0.04	0.02	0.63	0.01	0.02	0.04	0.01	0.11	0.00	10,483
E	0.09	0.00	0.03	0.03	0.03	0.05	0.01	0.70	0.01	0.02	0.02	0.02	0.00	25,983
F	0.02	0.03	0.00	0.12	0.04	0.05	0.01	0.02	0.58	0.02	0.05	0.04	0.00	14,031
F#	0.04	0.01	0.06	0.01	0.07	0.05	0.01	0.03	0.01	0.68	0.00	0.03	0.00	13,383
G	0.05	0.01	0.01	0.14	0.01	0.13	0.00	0.03	0.02	0.01	0.58	0.01	0.00	26,689
G#	0.01	0.04	0.02	0.02	0.13	0.01	0.03	0.01	0.01	0.04	0.00	0.68	0.00	11,065
N	0.09	0.03	0.05	0.14	0.07	0.28	0.02	0.14	0.02	0.06	0.06	0.05	0.00	5,794.
														23093

Δεν υπάρχει σημαντική διαφορά μεταξύ των δύο συνόλων. Στο επαυξημένο η συνολική ακρίβεια είναι καλύτερη. Αξιοσημείωτο είναι πως δεν έκανε καμία πρόβλεψη για την κατηγορία “όχι συγχορδία” όταν εκπαιδεύτηκε στο επαυξημένο.

β) Ματζόρε-μινόρε (25 κατηγορίες) - Sequence Balanced Dataset

#hidden	Sequence length	Features	#features	Train Accuracy	Test Accuracy
10	10	pitches	12	55.03%	52.93%
20	10	pitches	12	46.43%	48.8%
10	50	pitches	12	46.49%	54.24%
128	50	pitches	12	43.77%	51.17%
256	50	pitches	12	46.27%	54.40%

Ματζόρε - Μινόρε (25 κατηγορίες)- Augmented Sequence Balanced

#hidden	Sequence length	Features	#features	Train Accuracy	Test Accuracy
10	100	all	29	49.95%	45.06%
256	100	all	29	51.76%	48.29%

Τα δίκτυα LSTM που εκπαιδεύτηκαν για να προβλέπουν τη συγχορδία σε κάθε segment μιας ακολουθίας αποτελούνταν από τουλάχιστον δύο επίπεδα. Με ένα Dropout επίπεδο ενδιάμεσα για αποφυγή του overfitting.

Κάθε επίπεδο λαμβάνει σαν είσοδο μια ακολουθία μήκους L και επιστρέφει έξοδο μια ακολουθία με ίδιο μήκος. Η είσοδος του πρώτου επιπέδου είναι η αρχική ακολουθία των segments, όπου κάθε segment έχει M χαρακτηριστικά. Η έξοδος του πρώτου επιπέδου είναι μια ακολουθία με ίδιο μήκος αλλά αποτελούμενο από

H_1 χαρακτηριστικά (όπου το H_1 είναι η διάσταση του διανύσματος καταστάσεων, ή ο αριθμός των κρυφών μονάδων του πρώτου επιπέδου - υπερπαράμετρος του μοντέλου). Το επόμενο επίπεδο λαμβάνει την ακολουθία και παράγει μία με ίδιο μήκος αλλά το κάθε βήμα της αποτελείται από H_2 χαρακτηριστικά. Κάθε τμήμα της ακολουθίας εξόδου τροφοδοτείται σε ένα πλήρως συνδεδεμένο επίπεδο με συνάρτηση ενεργοποίησης Softmax και πλήθος νευρώνων ίσο με το πλήθος των κατηγοριών, η έξοδος του οποίου αποτελεί την πρόβλεψη.

Αμφίδρομο LSTM (Bidirectional LSTM) [23]

Η συγχορδία του κάθε segment εξαρτάται εξίσου από τα προηγούμενα με τα επόμενα segments του. Για να αξιοποιηθεί αυτή η εξάρτηση, επεκτάθηκε το αναδρομικό νευρωνικό δίκτυο LSTM που παρουσιάστηκε παραπάνω, ώστε να είναι αμφίδρομο - Bidirectional. Ουσιαστικά αποτελείται από δύο LSTM: η είσοδος του ενός είναι η ακολουθία καθ'αυτή ενώ του άλλου είναι η αντίστροφη ακολουθία.

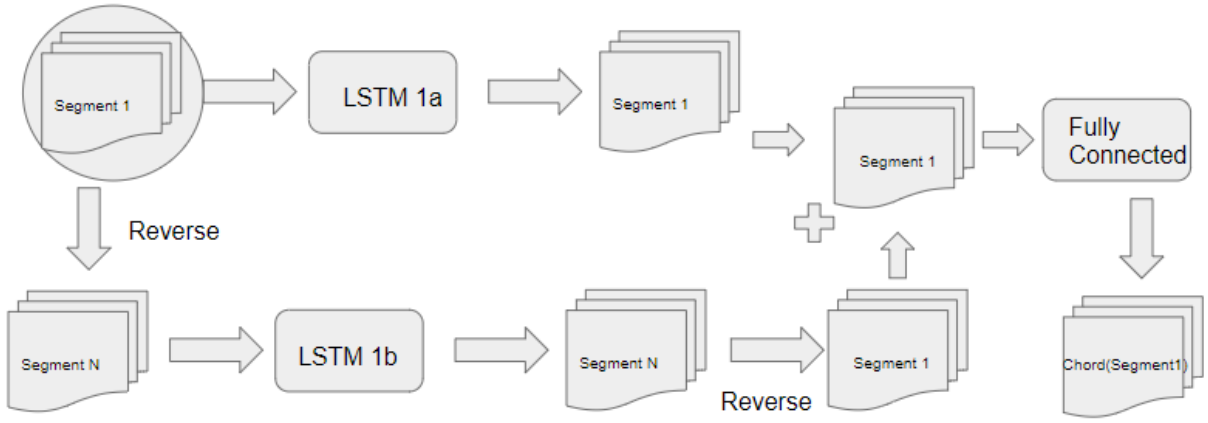
Η είσοδος οργανώνεται σε ακολουθίες μήκους L. Η κάθε ακολουθία περνά από το LSTM A, ενώ η αντίστροφή της από το LSTM B. Κάθε ένα από τα A και B αποτελείται από 128 κρυφές μονάδες (hidden units)/κελιά-κύτταρα. Συνδυάζοντας τις εξόδους προκύπτει μια ακολουθία μήκους L, όπου κάθε βήμα έχει 256 χαρακτηριστικά. Αυτή η ακολουθία τροφοδοτείται σε ένα πλήρως συνδεδεμένο Softmax επίπεδο, το οποίο αποτελείται από όσες κρυφές μονάδες όσες είναι οι κατηγορίες (25 για ματζόρε μινόρε, 13 για τονική). Η έξοδος αυτού αποτελεί τις προβλέψεις του συστήματος και τροφοδοτείται στη συνάρτηση απώλειας που βελτιστοποιείται όπως παραπάνω.

Για διάφορες τιμές του μήκους, εκπαιδευμένο στο *Sequence Balanced*[12] *Augmented*[13] *Dataset* (119.808 ακολουθίες αν L=50), στο *Sequence Balanced αρχικό Dataset* (10.035 ακολουθίες L=50) και στο αρχικό Dataset (14.000

[12] Sequence Balanced: Έχουν αφαιρεθεί οι ακολουθίες που περιέχουν μόνο μία συγχορδία
[13] Augmented: 12*αρχικό Dataset, όπου σε κάθε segment έχει μετατονιστεί το διάνυσμα των pitches και η συγχορδία

ακολουθίες αν L=50) για Ματζόρε Μινόρε, χρησιμοποιώντας μόνο τα pitches ή όλα τα χαρακτηριστικά.

Διαγραμματικά το δίκτυο:



Απόδοση (ματζόρε-μινόρε) L=50 (pitches):

Dataset	**Train Accuracy**	**Test Accuracy**
Filtered Augmented	44.60%	51.02%
Normal Train	56.33%	52.75%
Filtered Train	54.27%	47.74%

Απόδοση (ματζόρε-μινόρε) L=10:

Dataset	**Features**	**Train Accuracy**	**Test Accuracy**
Filtered Augmented	Pitches	55.77%	57.39%
Normal Train	Pitches	55.27%	52.81%
Filtered	All	57.74%	57.6%

Augmented			
Normal Train	All	55.43%	55.1%

Πίνακας Σύγχυσης (Filtered Augmented - All features)

	N	Amin	A	A#mir	A#	Bmin	B	Cmin	C	C#mir	C#	Dmin	D	D#mir	D#	Emin	E	Fmin	F	F#min	F#	Gmin	G	G#mir	G#	Population
N	0.0998	0.0126	0.0703	0.0022	0.0217	0.0084	0.0790	0.0101	0.0746	0.0276	0.0851	0.0046	0.0946	0.0088	0.0246	0.0056	0.1236	0.0013	0.0388	0.0145	0.0758	0.0034	0.0704	0.003	0.0386	5794
Amin	0.003	0.4636	0.2027	0.0048	0.0085	0.0026	0.0057	0.0005	0.0631	0.0018	0.0067	0.0206	0.0643	0	0.0007	0.0046	0.0151	0.0026	0.0680	0.0056	0.0012	0.0003	0.0438	0.0010	0.0077	7854
A	0.0063	0.0478	0.6648	0.0033	0.0064	0.0114	0.0220	0.0002	0.0096	0.0205	0.0088	0.0051	0.0447	0.0010	0.0011	0.0137	0.0430	0	0.0097	0.0303	0.0106	0.0004	0.0254	0.0011	0.0110	21263
A#min	0.0012	0	0.0044	0.4167	0.0744	0	0.0136	0.0022	0.0016	0.0056	0.2126	0	0	0.0582	0.0347	0	0.0028	0.0133	0.0092	0	0.0356	0.0006	0	0.0362	0.0770	2821
A#	0.0113	0.0014	0.0026	0.0887	0.5544	0.0004	0.0082	0.0202	0.0294	0.0006	0.0333	0.0104	0.0025	0.0141	0.0606	0	0.0006	0.0185	0.0728	0.0006	0.0097	0.0174	0.0122	0	0.0284	8473
Bmin	0.0068	0.0027	0.0471	0.0033	0.0012	0.3618	0.2474	0.0110	0.0101	0.0118	0.0160	0.0010	0.0882	0.0018	0.0002	0.0101	0.0628	0.0006	0.0035	0.0319	0.0286	0.0006	0.0442	0.0035	0.0025	4485
B	0.0022	0.0014	0.0366	0.0034	0.0144	0.0380	0.5653	0.0016	0.0126	0.0167	0.0353	0.0014	0.0117	0.0084	0.0041	0.0067	0.0952	0.0003	0.0016	0.0236	0.0793	0	0.0121	0.0127	0.0134	10416
Cmin	0.0044	0.0008	0.0008	0.0054	0.0348	0	0.0012	0.5297	0.1467	0.0031	0.0272	0.0037	0.0023	0.0002	0.0612	0	0.0002	0.0246	0.0584	0	0.0008	0.0044	0.0180	0	0.0716	4155
C	0.0032	0.0231	0.0086	0.0010	0.0146	0.0023	0.0314	0.0221	0.6484	0.0028	0.0222	0.0196	0.0325	0.0011	0.0022	0.0067	0.0136	0.0047	0.0592	0.0020	0.0067	0.0027	0.0618	0.0005	0.0048	20315
C#min	0.0002	0.0002	0.1171	0	0.0002	0.0006	0.0466	0.0010	0.0013	0.4370	0.1817	0	0.0052	0.0028	0.0013	0.0010	0.0621	0.0002	0.0015	0.0266	0.0900	0	0.0013	0.0045	0.0169	4697
C#	0.0006	0.0001	0.0034	0.0137	0.0028	0.0010	0.0261	0.0022	0.0040	0.0131	0.7178	0.0004	0.0017	0.0193	0.0127	0.0002	0.0086	0.0074	0.0053	0.0042	0.1064	0.0007	0.0026	0.0066	0.0383	10143
Dmin	0.0038	0.0208	0.0416	0.0015	0.0572	0.0043	0.0015	0.0016	0.0448	0.0096	0.0238	0.4468	0.1820	0.0116	0.0036	0.0012	0.0091	0.0007	0.0801	0.0016	0.0038	0.0046	0.0401	0	0.0023	5067
D	0.0075	0.0071	0.0893	0	0.0042	0.0160	0.0094	0.0003	0.0184	0.0028	0.0144	0.0347	0.6243	0.0008	0.0027	0.0045	0.0245	0.0004	0.0086	0.0246	0.0126	0.0013	0.0867	0.0006	0.0033	23813
D#min	0.0006	0	0.0010	0.0051	0.0051	0	0.0258	0.0003	0	0.0017	0.0836	0	0.0017	0.5732	0.1366	0	0.0076	0.0021	0.0020	0.0003	0.1046	0	0	0.0154	0.0313	2832
D#	0.0080	0.0010	0.0015	0.0168	0.0436	0.0000	0.0068	0.0227	0.0037	0.0010	0.0598	0.0004	0.0018	0.0684	0.6013	0.0010	0.0056	0.0107	0.0181	0.0010	0.0145	0.0071	0.0077	0.0038	0.0925	7651
Emin	0.0155	0.0193	0.1058	0.0007	0.0033	0.0055	0.0146	0.0004	0.0596	0.0055	0.0111	0.0017	0.0373	0.0016	0.0033	0.2606	0.3070	0.0210	0.0240	0.0051	0.0126	0.0003	0.0802	0	0.0026	6709
E	0.0096	0.0056	0.0741	0.0003	0.0036	0.0040	0.0372	0	0.0096	0.0397	0.0127	0.0006	0.0165	0.0026	0.0054	0.0156	0.6843	0.0004	0.0090	0.0124	0.0261	0.0006	0.0100	0.0086	0.0094	19274
Fmin	0.0033	0.0022	0	0.0277	0.0291	0	0.0056	0.0174	0.0446	0.0006	0.1316	0.0033	0.0025	0.0066	0.0236	0	0.0048	0.3948	0.1442	0.0036	0.0176	0.0011	0.0146	0.0011	0.1185	3441
F	0.0034	0.0201	0.0096	0.0033	0.0278	0.0002	0.0027	0.0103	0.0661	0.0024	0.0232	0.0236	0.0063	0.0006	0.0046	0.0014	0.0151	0.0213	0.6675	0.0051	0.0195	0.0027	0.0531	0	0.0083	10590
F#min	0.0006	0.0002	0.1001	0.0004	0.0004	0.0086	0.0566	0	0.0018	0.0156	0.0390	0	0.0556	0.0020	0	0.0036	0.0266	0	0.0020	0.5520	0.1128	0.0002	0.0143	0.0027	0.0020	4899
F#	0.0017	0.0003	0.0075	0.0114	0.0051	0.0074	0.0610	0.0018	0.0046	0.0103	0.0834	0.0012	0.0077	0.0143	0.0031	0.0006	0.0252	0.0006	0.0094	0.0441	0.6578	0.0005	0.0086	0.0078	0.0220	8484
Gmin	0.0091	0.0066	0.0121	0.0014	0.0798	0.0004	0	0.0177	0.0843	0.0007	0.0091	0.0150	0.0551	0	0.0345	0.0004	0.0002	0.0036	0.0657	0.0078	0.0066	0.3011	0.2714	0	0.0155	2658
G	0.0053	0.0167	0.0332	0.0003	0.0051	0.0121	0.0077	0.0088	0.0811	0.0018	0.0096	0.0108	0.0620	0.0004	0.0036	0.0140	0.0105	0.0008	0.0214	0.0063	0.0105	0.0088	0.6582	0.0007	0.0107	24031
G#min	0.0017	0	0.0081	0.0231	0.0017	0.0002	0.0494	0.0005	0	0.0335	0.2361	0	0.0008	0.0167	0.0083	0	0.0492	0.0075	0.0002	0.0052	0.0848	0.0002	0.0023	0.2636	0.2054	3097
G#	0.0048	0.0015	0.0052	0.0165	0.0166	0	0.0044	0.0200	0.0055	0.0086	0.1534	0.0004	0.0006	0.0111	0.0403	0	0.0022	0.0182	0.0086	0.0007	0.0305	0.0011	0.0037	0.0126	0.6336	7968
sums	0.2157	0.6565	1.6490	0.6515	1.0170	0.4864	1.3302	0.7036	1.4264	0.6744	2.2378	0.6066	1.4038	0.8273	1.0757	0.3528	1.6023	0.5572	1.3905	0.8118	1.5607	0.3606	1.5437	0.3878	1.4695	230930

Πρόβλεψη στο Fallout
(https://open.spotify.com/track/27oSbEifl4aip9TLOSywPu)

Το Fallout είναι ένα κομμάτι που έχει ποικιλία διαφορετικών ήχων (από πολύ καθαρό εώς πολύ θορυβώδη). Έχει επίσης σε πολλά σημεία Σολ ματζόρε (G) ακολουθούμενη από Σολ μινόρε (Gmin), το οποίο ενδέχεται να δυσκολέψει ένα σύστημα πρόβλεψης, καθώς οι συγχορδίες είναι πολύ όμοιες, ενώ η αλληλουχία δε συνηθίζεται.

Τις καλύτερες προβλέψεις στο κομμάτι έκανε το bLSTM που εκπαιδεύτηκε στο *Sequence Balanced Augmented Dataset.* Αυτές φαίνονται στον παρακάτω πίνακα.

Start Time	Predicted (FA)	Real
0.000	Dmin	Dmin
9.532	Gmin	Gmin
9.712	Dmin	Dmin - Amin(~12.00)
14.574	Gmin	G - Gmin(~15.50)
18.465	Dmin	Dmin - Amin (~20.00)
22.211	Gmin	G - Gmin(~24.00) - Dmin(~26.00)
28.439	Amin	Amin
30.215	G	G
33.205	Gmin	Gmin
34.702	Dmin	Dmin
37.943	Amin	Amin
39.173	Dmin	Dmin
43.207	G	Gmin
44.809	A#	A#
46.173	C	C
48.518	Dmin	Dmin
52.558	Gmin	Gmin - A# (~54.00)- Gmin(~56.00)
58.804	Dmin	Dmin
60.808	Amin	Amin
61.701	G	G - Gmin(~64.00)
66.775	Dmin	Dmin
68.394	Amin	Amin
69.933	C	Amin
70.791	G	G
72.214	Gmin	G

72.463	G	G
72.591	Gmin	Gmin
74.425	G	Dmin
76.168	Dmin	Dmin
77.682	Amin	Amin
78.164	F	Amin
78.438	G	G
79.174	Gmin	Gmin
83.122	Dmin	Dmin - Amin (~84.00)
86.425	G	G
89.17	Gmin	Gmin
91.214	Dmin	Dmin
93.692	Amin	Amin
94.587	Gmin	G
94.917	G	G
95.179	Gmin	Gmin
99.66	Dmin	Dmin - Amin(~102.00)-Dmin(~104.00)
107.236	Gmin	Gmin
107.468	G	Gmin
108.53	A#	A#
109.917	C	C
112.17	Dmin	Dmin
114.962	Gmin	Gmin
117.621	A#	A#
117.714	Gmin	A#
117.876	A#	A#

118.137	C	C
121.185	Gmin	Gmin
122.654	Dmin	Dmin - Amin (~123.50)
125.702	G	G - Gmin (~128.50)
130.962	Dmin	Dmin
133.213	Amin	Amin
134.81	G	G - Gmin(~136.00)
138.582	Dmin	Dmin
141.723	Amin	Amin
142.507	G	G
142.959	Gmin	Gmin
147.609	Dmin	Dmin-Amin(~146.00)
150.025	G	G
153.159	Gmin	Gmin
154.204	G	Gmin
154.941	Dmin	Dmin
157.124	Amin	Amin
158.876	G	G-Gmin(~160.00)
162.946	Dmin	Dmin
165.134	Amin	Amin
166.528	G	G
167.678	Gmin	Gmin
170.533	Dmin	Dmin
173.197	Amin	Amin
174.579	G	G - Gmin(~176.00)
178.788	Dmin	Dmin - Amin(~180.00)

183.432	G	G- Gmin (~184.00)
186.927	Dmin	Dmin
187.071	Gmin	Amin
189.77	Dmin	Amin
191.872	Amin	Amin

Παρατηρούμε πως μεγάλο ποσοστό των σφαλμάτων (πορτοκαλί) προκύπτουν όταν το σύστημα δεν προβλέπει σωστά αλλαγή συγχορδίας, κυρίως όταν δύο διαδοχικές συγχορδίες μοιάζουν (πχ G, Gmin).

Ενδεικτικά το ίδιο μοντέλο εκπαιδευμένο στο αρχικό Dataset παράγει τις εξής προβλέψεις στα πρώτα 30 δευτερόλεπτα του κομματιού.

0.000 : Dmin

0.447 : D

8.348 : G

9.532 : A#

9.898 : Dmin

13.806 : G

14.656 : Gmin

17.915 : Dmin

21.062 : A

22.082 : F

22.211 : G

23.219 : Gmin

27.203 : Dmin

28.439 : Amin

30.023 : Dmin

Παρά το γεγονός ότι το μοντέλο είχε καλύτερη ακρίβεια και στο Test Set (ίδιο και για τα δύο Train Sets) έχει πολύ χειρότερη απόδοση σε ένα τέτοιο κομμάτι. Η υψηλή ακρίβεια στο Test Set οφείλεται στην έλλειψη διακύμανσης στα δεδομένα (οι συχνότερες των συγχορδιών εμφανίζονται ~10 φορές περισσότερο από τις σπανιότερες). Αυτό είναι χαρακτηριστικό της δυτικής μουσικής γενικότερα, αλλά ένα σύστημα πρόβλεψης θα πρέπει να είναι αμερόληπτο.

Κωδικοποιητής - Αποκωδικοποιητής LSTM (LSTM Encoder - Decoder)

Η συγκεκριμένη αρχιτεκτονική έχει γνωρίσει μεγάλη επιτυχία σε προβλήματα μετάφρασης φυσικής γλώσσας [33]. Το μοντέλο αποτελείται από δύο LSTM δίκτυα, τον κωδικοποιητή (encoder) και τον αποκωδικοποιητή (decoder), με ίδιο αριθμό κρυφών μονάδων το καθένα, καθώς και ένα πλήρως συνδεδεμένο επίπεδο στο τέλος του αποκωδικοποιητή. Η λειτουργία τους είναι διαφορετική κατά την εκπαίδευση και κατά την πρόβλεψη.

LSTM Encoder - Decoder - Εκπαίδευση

Έστω κατά την εκπαίδευση, οι αρχικές ακολουθίες εισόδου X1, διαστάσεων [?, N, M] αποτελούνται από N segments η κάθε μία, με M χαρακτηριστικά το καθένα και οι ακολουθίες εξόδου Y[?, N, C] από N τμήματα η κάθε μία, με C πιθανές κατηγορίες το κάθε ένα.

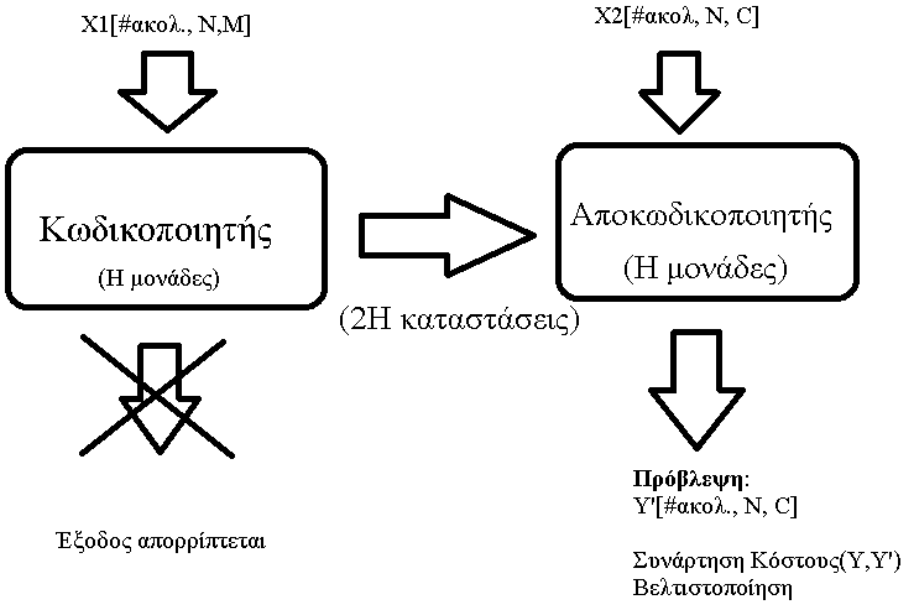
Θα κατασκευάσουμε ακολουθίες X2[?, N, C] μετατοπίζοντας κάθε συγχορδία από τις αλληλουχίες Y (εκτός από την τελευταία συγχορδία κάθε ακολουθίας) στην επόμενη θέση της ακολουθίας. Στην αρχή των μετατοπισμένων ακολουθιών προστίθενται διανύσματα μηδενικών, διάστασης [C].

Η εκπαίδευση γίνεται με είσοδο τους πίνακες [X1, X2] με σκοπό την πρόβλεψη του Y ως εξής:

Αρχικά το LSTM του κωδικοποιητή λαμβάνει την είσοδο X1 και επιστρέφει τις εσωτερικές καταστάσεις των κελιών/κυττάρων στο τέλος κάθε ακολουθίας.

Στη συνέχεια, οι καταστάσεις του αποκωδικοποιητή θέτονται ίδιες με του κωδικοποιητή (έξοδος κωδικοποιητή), και με είσοδο το [X2], το LSTM επιστρέφει ακολουθία (ίδιου μήκους, με χαρακτηριστικά όσα οι κρυφές του μονάδες), που μέσω ενός πλήρως συνδεδεμένου επιπέδου προβλέπουν την έξοδο [Y].

Δηλαδή ο αποκωδικοποιητής σε κάθε βήμα λαμβάνει είσοδο μια συγχορδία (από το [X2]) και μια κατάσταση (από τον κωδικοποιητή) και επιχειρεί με βάση αυτά να προβλέψει την επόμενη συγχορδία (από το [Y]) . Προκύπτει ένα μοντέλο της μορφής:



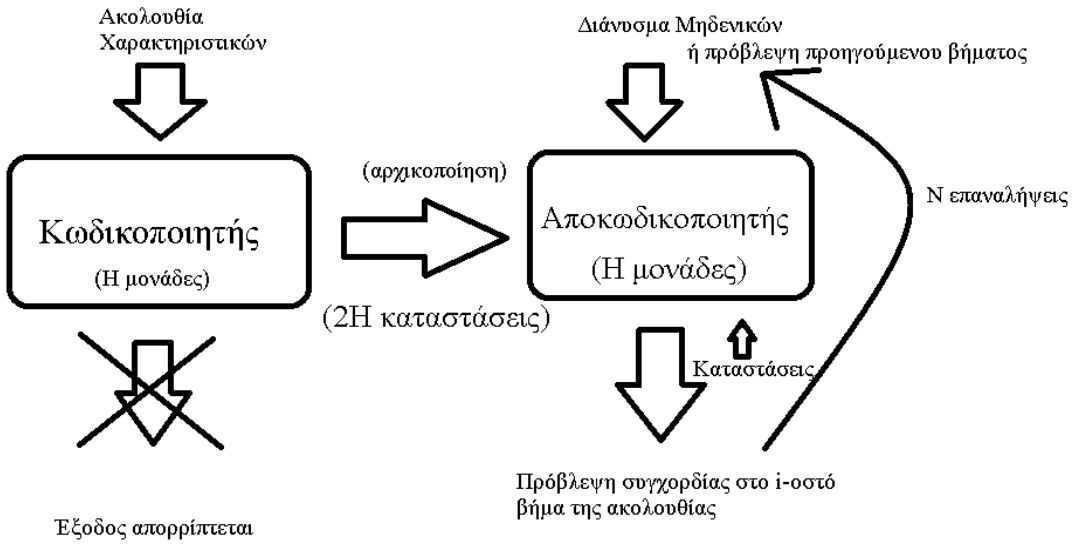
LSTM Encoder - Decoder Πρόβλεψη

Όταν χρησιμοποιείται το παραπάνω μοντέλο για πρόβλεψη συγχορδιών, τότε προφανώς δεν μπορούμε να κατασκευάσουμε την είσοδο του αποκωδικοποιητή [X2] , αφού δεν γνωρίζουμε εκ των προτέρων τις συγχορδίες.

Ο τρόπος που γίνεται η πρόβλεψη είναι ο εξής:

Αρχικά οι ακολουθίες [X1] τροφοδοτούνται στον κωδικοποιητή, και περνώνται οι καταστάσεις που επιστρέφει στον αποκωδικοποιητή, όπως και στην εκπαίδευση.

Ο αποκωδικοποιητής προβλέπει την πρώτη συγχορδία της ακολουθίας, με είσοδο διάνυσμα μηδενικών (έχοντας τις καταστάσεις του κωδικοποιητή). Θυμίζεται πως κατά την εκπαίδευση, το πρώτο τμήμα κάθε ακολουθίας X2 είναι επίσης διάνυσμα μηδενικών. Αφού προβλέψει την πρώτη συγχορδία, ανανεώνονται οι καταστάσεις (από το "πέρασμα" του διανύσματος μηδενικών μέσα από το δίκτυο) και χρησιμοποιείται η πρώτη συγχορδία (σε διάνυσμα) ώς είσοδος στο δεύτερο βήμα με τις νέες καταστάσεις. Η διαδικασία επαναλαμβάνεται μέχρι το τέλος της κάθε ακολουθίας. Σχηματικά:



Πειράματα

Εκτελέστηκαν πειράματα με διαφορετικά βάθη των LSTM δικτύων για τον κωδικοποιητή και αποκωδικοποιητή. Συγκεκριμένα το πρώτο μοντέλο είχε ένα LSTM επίπεδο σε κάθε υποδίκτυο, το δεύτερο είχε 3 LSTM επίπεδα στον κωδικοποιητή και 1 στον αποκωδικοποιητή και το τελευταίο 1 στον κωδικοποιητή και 3 στον αποκωδικοποιητή.

Α) 1 LSTM - 128 μονάδες

Στο πρώτο μοντέλο της παραπάνω αρχιτεκτονικής, οι κωδικοποιητής - αποκωδικοποιητής αποτελούνται από 1 LSTM ο κάθε ένας με 128 κρυφές μονάδες.

Αναγνώριση Τονικής

Στο Αρχικό Σύνολο Εκπαίδευσης:

Sequence Length	#Sequences	#Features	Train Accuracy	Test Accuracy
5	99688	Pitches(12)	82.5%	64.71%
10	49687	Pitches(12)	85.3%	64.09%
20	24724	Pitches(12)	86.27%	59.19%
5	99,341	Extended(46)	84.00%	64.64%
10	49,432	Extended(46)	85.70%	63.28%
20	24,852	Extended(46)	86.16%	53.77%

Αναγνώριση Ματζόρε-Μινόρε

στο Αρχικό Σύνολο Εκπαίδευσης:

Sequence Length	#Sequences	#Features	Train Accuracy	Test Accuracy
5	96,684	Pitches(12)	79.49%	55.52%
10	48,255	Pitches(12)	84.69%	54.60%
20	24,144	Pitches(12)	85.89%	48.13%
5	99,534	Extended(46)	86.98%	52.21%

10	49,767	Extended(46)	84.83%	53.06%
20	25024	Extended(46)	87.56%	45.83%

Στο Επαυξημένο Σύνολο Εκπαίδευσης:

Sequence Length	#Train Sequences	#Features	Train Accuracy	Test Accuracy
5	1,584,000	Pitches(12)	88.93%	59.68%
10	792,000	Pitches(12)	90%	61.96%

Πίνακας Σύγχυσης για μήκος ακολουθίας 10 στο αρχικό σύνολο εκπαίδευσης:

	N	Amin	A	A#mir	A#	Bmin	B	Cmin	C	C#mir	C#	Dmin	D	D#mir	D#	Emin	E	Fmin	F	F#mir	F#	Gmin	G	G#mir	G#	Population
N	0.3500	0.0340	0.0605	0	0.0302	0.0484	0.0136	0.0063	0.0407	0.0044	0.0184	0.0096	0.0743	0.0124	0.0132	0.0314	0.1168	0.0024	0.0108	0.0031	0.0281	0.0082	0.0531	0.0070	0.0219	5794
Amin	0.0155	0.4691	0.1623	0	0.0080	0.0058	0.0034	0.0001	0.0953	0.0001	0	0.0164	0.0417	0	0.0003	0.0207	0.0277	0.0002	0.0436	0.0050	0.0005	0.0008	0.0761	0.0019	0.0044	7854
A	0.0167	0.0564	0.6136	0	0.0047	0.0159	0.0105	0.0005	0.0105	0.0103	0.0022	0.0059	0.0593	0.0002	0.0013	0.0314	0.0926	0	0.0065	0.0162	0.0037	0.0003	0.0287	0.0070	0.0044	21263
A#mir	0.0007	0.0007	0.0053	0.1375	0.2828	0.0007	0.0067	0.0021	0.0031	0.0085	0.1109	0	0.0010	0.1261	0.0669	0	0.0049	0.0138	0.0120	0	0.0528	0	0	0.0471	0.1155	2821
A#	0.0105	0.0054	0.0055	0.0033	0.6236	0.0015	0.0089	0.0092	0.0651	0	0.0080	0.0230	0.0084	0.0057	0.0765	0.0017	0.0017	0.0131	0.0442	0	0.0093	0.0194	0.0122	0.0011	0.0416	8473
Bmin	0.0111	0.0140	0.0477	0.0049	0.0113	0.5157	0.0697	0	0.0089	0.0060	0.0037	0.0026	0.0726	0.0026	0.0035	0.0296	0.0838	0	0.0013	0.0211	0.0269	0	0.0573	0.0035	0.0011	4485
B	0.0122	0.0021	0.0519	0.0021	0.0255	0.1745	0.3303	0.0005	0.0117	0.0199	0.0208	0.0007	0.0120	0.0195	0.0045	0.0086	0.1191	0.0002	0.0014	0.0228	0.0949	0	0.0130	0.0344	0.0162	10416
Cmin	0.0199	0.0168	0.0002	0.0002	0.0880	0	0.0038	0.2880	0.2758	0	0.0199	0.0031	0.0002	0	0.0688	0.0004	0.0014	0.0125	0.0377	0	0	0.0163	0.0151	0.0009	0.1299	4155
C	0.0115	0.0323	0.0052	0.0000	0.0122	0.0115	0.0190	0.0071	0.6829	0.0022	0.0075	0.0136	0.0160	0.0023	0.0022	0.0200	0.0196	0.0020	0.0338	0.0007	0.0057	0.0058	0.0772	0.0024	0.0061	20315
C#mir	0.0051	0.0025	0.1234	0.0012	0.0014	0.0042	0.0404	0	0.0031	0.3736	0.1166	0.0008	0.0053	0.0085	0.0021	0.0019	0.1119	0	0.0017	0.0119	0.0657	0	0.0006	0.0806	0.0364	4697
C#	0.0161	0.0007	0.0071	0.0060	0.0111	0.0048	0.0122	0.0022	0.0136	0.0284	0.5507	0.0039	0.0024	0.0419	0.0139	0.0005	0.0145	0.0105	0.0114	0.0043	0.0859	0.0005	0.0028	0.0278	0.1256	10143
Dmin	0.0165	0.0517	0.0528	0	0.0882	0.0067	0	0.0003	0.0807	0.0124	0.0031	0.3291	0.1914	0	0.0067	0.0071	0.0177	0	0.0594	0.0007	0.0003	0.0092	0.0588	0.0023	0.0039	5067
D	0.0159	0.0185	0.1145	0	0.0078	0.0347	0.0023	0.0001	0.0340	0.0034	0.0018	0.0393	0.5208	0.0002	0.0009	0.0104	0.0437	0	0.0052	0.0136	0.0067	0.0017	0.1201	0.0001	0.0029	23813
D#mir	0.0049	0	0.0014	0.0010	0.0095	0	0.0141	0.0010	0.0010	0.0017	0.0444	0	0.0017	0.3894	0.3259	0	0.0091	0.0017	0.0010	0	0.0819	0.0007	0.0035	0.0346	0.0706	2832
D#	0.0104	0.0015	0.0009	0.0013	0.1095	0.0006	0.0010	0.0088	0.0291	0.0013	0.0260	0.0030	0.0014	0.0284	0.5494	0.0015	0.0069	0.0058	0.0103	0.0002	0.0130	0.0194	0.0142	0.0054	0.1491	7651
Emin	0.0251	0.0567	0.0424	0	0.0025	0.0219	0.0050	0	0.0606	0.0034	0.0013	0.0014	0.0310	0.0055	0.0011	0.3793	0.2507	0	0.0134	0.0038	0.0023	0.0002	0.0892	0.0019	0.0001	6709
E	0.0195	0.0069	0.0681	0.0002	0.0052	0.0111	0.0324	0	0.0085	0.0251	0.0029	0.0007	0.0178	0.0043	0.0047	0.0345	0.7005	0.0009	0.0032	0.0062	0.0152	0.0002	0.0109	0.0133	0.0064	19274
Fmin	0.0092	0.0072	0.0005	0.0023	0.0927	0.0002	0.0133	0.0235	0.0900	0.0008	0.0688	0.0046	0.0034	0.0058	0.0261	0.0002	0.0133	0.2813	0.1063	0	0.0133	0.0154	0.0220	0.0063	0.1915	3441
F	0.0130	0.0481	0.0087	0.0007	0.0541	0.0037	0.0017	0.0054	0.1525	0.0007	0.0150	0.0152	0.0067	0.0005	0.0058	0.0095	0.0273	0.0100	0.5235	0.0050	0.0114	0.0050	0.0618	0.0014	0.0121	10590
F#mir	0.0038	0.0006	0.1371	0	0.0010	0.0563	0.0424	0	0.0020	0.0379	0.0112	0.0004	0.0751	0.0010	0.0008	0.0177	0.0716	0	0.0040	0.3541	0.1365	0	0.0130	0.0226	0.0100	4899
F#	0.0080	0.0040	0.0119	0.0021	0.0153	0.0330	0.0473	0.0010	0.0086	0.0307	0.1016	0.0004	0.0068	0.0187	0.0051	0.0021	0.0327	0.0020	0.0070	0.0367	0.5383	0.0005	0.0091	0.0245	0.0516	8484
Gmin	0.0203	0.0056	0.0158	0	0.1177	0.0109	0.0022	0.0221	0.0639	0	0	0.0056	0.0891	0.0003	0.0410	0.0056	0.0026	0.0052	0.0323	0.0045	0.0052	0.3164	0.2189	0.0015	0.0124	2658
G	0.0138	0.0240	0.0236	0.0000	0.0073	0.0195	0.0049	0.0046	0.1184	0.0008	0.0005	0.0069	0.0551	0.0004	0.0043	0.0302	0.0138	0.0003	0.0168	0.0030	0.0055	0.0177	0.6205	0.0006	0.0064	24031
G#mir	0.0022	0	0.0158	0.0248	0.0022	0.0054	0.0574	0.0016	0.0019	0.0258	0.0422	0	0.0016	0.0920	0.0109	0	0.0988	0.0038	0	0.0029	0.0723	0.0009	0.0035	0.3099	0.2231	3097
G#	0.0036	0.0045	0.0038	0.0038	0.0341	0.0003	0.0076	0.0166	0.0188	0.0115	0.1135	0.0005	0.0013	0.0115	0.0630	0.0002	0.0060	0.0111	0.0055	0	0.0234	0.0032	0.0037	0.0268	0.6244	7968
sums	0.6367	0.8642	1.5811	0.1920	1.6469	0.9884	0.7513	0.4020	1.8818	0.6099	1.2921	0.4878	1.2978	0.7781	1.3001	0.6459	1.8910	0.3776	0.9933	0.5167	1.2999	0.4428	1.5866	0.6662	1.8685	230930

Πίνακας Σύγχυσης για μήκος ακολουθίας 10 στο επαυξημένο σύνολο:

	N	Amin	A	A#mir	A#	Bmin	B	Cmin	C	C#mir	C#	Dmin	D	D#mir	D#	Emin	E	Fmin	F	F#mir	F#	Gmin	G	G#mir	G#	Population
N	0.0833	0.0283	0.0809	0.0025	0.0182	0.0238	0.0612	0.0112	0.0743	0.0127	0.0519	0.0077	0.1416	0.0046	0.0336	0.0182	0.1042	0.0024	0.0315	0.0170	0.0398	0.0098	0.1073	0.0029	0.0296	5794
Amin	0.0053	0.5463	0.1541	0	0.0036	0.0036	0.0039	0.0007	0.0765	0.0002	0.0002	0.0160	0.0660	0	0.0028	0.0024	0.0131	0.0003	0.0477	0.0030	0.0003	0.0001	0.0481	0.0014	0.0033	7854
A	0.0035	0.0541	0.6987	0.0000	0.0016	0.0129	0.0162	0.0001	0.0076	0.0145	0.0048	0.0056	0.0622	0.0002	0.0019	0.0166	0.0378	0.0000	0.0046	0.0212	0.0044	0.0005	0.0192	0.0023	0.0081	21263
A#mir	0.0007	0.0003	0.0035	0.4732	0.0740	0	0.0095	0.0021	0.0070	0.0085	0.1035	0	0.0003	0.0411	0.0552	0	0.0014	0.0184	0.0088	0.0007	0.0397	0.0007	0.0003	0.0712	0.0790	2821
A#	0.0059	0.0027	0.0086	0.0218	0.6351	0.0014	0.0074	0.0224	0.0496	0.0010	0.0132	0.0153	0.0048	0.0050	0.0545	0	0.0007	0.0262	0.0574	0.0009	0.0092	0.0175	0.0149	0.0014	0.0221	8473
Bmin	0.0055	0.0057	0.0423	0.0044	0.0042	0.5701	0.0856	0	0.0066	0.0091	0.0167	0.0015	0.1043	0	0.0006	0.0149	0.0354	0	0.0046	0.0294	0.0107	0.0002	0.0443	0.0024	0.0004	4485
B	0.0012	0.0014	0.0458	0.0027	0.0119	0.0873	0.5495	0.0012	0.0074	0.0122	0.0228	0.0018	0.0192	0.0090	0.0035	0.0085	0.0618	0.0014	0.0027	0.0484	0.0609	0.0014	0.0094	0.0164	0.0110	10416
Cmin	0.0026	0.0072	0.0009	0.0021	0.0380	0.0004	0.0024	0.5018	0.1860	0.0024	0.0197	0.0043	0.0028	0	0.0779	0.0012	0	0.0108	0.0676	0	0.0007	0.0045	0.0158	0.0014	0.0486	4155
C	0.0029	0.0309	0.0087	0.0007	0.0076	0.0034	0.0313	0.0110	0.7032	0.0022	0.0109	0.0133	0.0248	0.0007	0.0026	0.0060	0.0098	0.0053	0.0393	0.0036	0.0036	0.0031	0.0709	0.0001	0.0029	20315
C#mir	0.0004	0.0004	0.0983	0	0.0008	0.0017	0.0542	0.0010	0.0023	0.5411	0.1153	0.0002	0.0093	0.0061	0.0053	0.0010	0.0410	0.0002	0.0023	0.0298	0.0689	0	0	0.0036	0.0157	4697
C#	0.0013	0	0.0057	0.0136	0.0036	0.0013	0.0182	0.0041	0.0066	0.0242	0.6827	0.0005	0.0036	0.0163	0.0101	0.0007	0.0082	0.0168	0.0057	0.0152	0.0877	0.0003	0.0054	0.0126	0.0542	10143
Dmin	0.0019	0.0408	0.0406	0.0005	0.0371	0.0051	0.0003	0.0019	0.0617	0.0136	0.0078	0.4406	0.2212	0	0.0025	0.0019	0.0047	0.0001	0.0582	0.0031	0.0011	0.0078	0.0442	0	0.0017	5067
D	0.0047	0.0146	0.0857	0	0.0014	0.0241	0.0036	0.0003	0.0162	0.0015	0.0065	0.0185	0.6979	0.0000	0.0005	0.0062	0.0165	0.0007	0.0045	0.0084	0.0046	0.0023	0.0778	0	0.0022	23813
D#mir	0.0010	0	0.0017	0.0060	0.0028	0	0.0215	0	0	0.0017	0.0625	0	0.0021	0.5307	0.2171	0	0.0049	0.0084	0.0031	0.0007	0.0971	0	0	0.0165	0.0215	2832
D#	0.0039	0.0006	0.0022	0.0060	0.0580	0	0.0032	0.0279	0.0180	0.0015	0.0298	0.0028	0.0079	0.0275	0.6548	0.0039	0.0028	0.0128	0.0145	0.0002	0.0105	0.0098	0.0098	0.0082	0.0824	7651
Emin	0.0155	0.0374	0.0909	0	0.0019	0.0165	0.0084	0.0002	0.0681	0.0046	0.0049	0.0034	0.0541	0.0041	0.0010	0.3605	0.2050	0	0.0165	0.0055	0.0050	0.0010	0.0927	0	0.0019	6709
E	0.0051	0.0088	0.0767	0.0001	0.0022	0.0125	0.0406	0.0004	0.0095	0.0327	0.0066	0.0006	0.0308	0.0016	0.0069	0.0446	0.6643	0.0022	0.0049	0.0110	0.0140	0.0004	0.0097	0.0062	0.0062	19274
Fmin	0.0017	0.0040	0.0020	0.0151	0.0435	0	0.0052	0.0154	0.0587	0.0017	0.0900	0.0066	0.0023	0.0014	0.0232	0.0002	0.0069	0.4722	0.0892	0.0058	0.0081	0.0061	0.0188	0.0078	0.1130	3441
F	0.0031	0.0373	0.0126	0.0021	0.0285	0.0016	0.0028	0.0063	0.0968	0.0013	0.0213	0.0231	0.0126	0.0005	0.0075	0.0020	0.0184	0.0135	0.6127	0.0088	0.0122	0.0031	0.0658	0.0001	0.0047	10590
F#mir	0.0014	0.0024	0.0913	0.0002	0	0.0185	0.0479	0.0002	0.0016	0.0328	0.0089	0.0004	0.0857	0.0014	0	0.0042	0.0277	0.0014	0.0020	0.5927	0.0620	0	0.0112	0.0028	0.0024	4899
F#	0.0004	0.0030	0.0064	0.0102	0.0033	0.0102	0.0588	0.0025	0.0049	0.0142	0.0687	0.0009	0.0105	0.0111	0.0051	0.0017	0.0235	0.0030	0.0100	0.1130	0.5901	0.0005	0.0104	0.0106	0.0252	8484
Gmin	0.0011	0.0026	0.0214	0.0011	0.0726	0.0041	0.0026	0.0376	0.0669	0.0011	0	0.0127	0.1060	0	0.0376	0.0033	0.0003	0.0022	0.0398	0.0056	0.0037	0.3875	0.1741	0.0030	0.0120	2658
G	0.0030	0.0214	0.0311	0	0.0037	0.0137	0.0047	0.0073	0.0824	0.0017	0.0042	0.0076	0.0737	0.0001	0.0026	0.0151	0.0046	0.0008	0.0185	0.0045	0.0063	0.0093	0.6774	0.0011	0.0042	24031
G#mir	0.0022	0.0022	0.0142	0.0413	0.0029	0.0029	0.0649	0.0016	0	0.0429	0.0997	0	0.0025	0.0125	0.0071	0	0.0526	0.0093	0.0006	0.0067	0.0791	0.0003	0.0029	0.4275	0.1233	3097
G#	0.0032	0.0022	0.0076	0.0092	0.0288	0.0037	0.0050	0.0273	0.0077	0.0100	0.1224	0.0011	0.0051	0.0122	0.0440	0	0.0033	0.0235	0.0056	0.0032	0.0247	0.0007	0.0028	0.0126	0.6326	7968
sums	0.1619	0.8556	1.6330	0.6137	1.0864	0.8199	1.1100	0.6854	1.6207	0.7908	1.5761	0.5856	1.7531	0.6872	1.2589	0.5142	1.3501	0.6330	1.1535	0.9396	1.2455	0.4678	1.5343	0.6130	1.3094	230930

Β) 3 και 1 LSTM - 128 μονάδες

Στο δεύτερο μοντέλο της παραπάνω αρχιτεκτονικής, ο κωδικοποιητής αποτελείται από τρία επίπεδα LSTM στη σειρά, με τις εξόδους των πρώτων να είναι ακολουθίες και την έξοδο του δεύτερου οι καταστάσεις για τον αποκωδικοποιητή. Ο αποκωδικοποιητής παραμένει ίδιος (1 επίπεδο LSTM, 1 Softmax)

Στο Αρχικό Σύνολο Εκπαίδευσης:

Sequence Length	#Sequences	#Features	Train Accuracy	Test Accuracy
5	97,859	Pitches(12)	79.77%	54.46%
10	48,304	Pitches(12)	83.57%	53.17%
20	24,903	Pitches(12)	86.35%	52.28%

Στο Επαυξημένο Σύνολο

Sequence Length	#Sequences	#Features	Train Accuracy	Test Accuracy
5	1,584,000	Pitches(12)	87.69%	61.02%
10	792,000	Pitches(12)	90.28%	62.71%

C) 1 και 3 LSTM:

Στο τρίτο μοντέλο, ο αποκωδικοποιητής αποτελείται από 3 LSTM στη σειρά, ενώ ο κωδικοποιητής από 1.

Στο Αρχικό Σύνολο Εκπαίδευσης:

Sequence Length	#Sequences	#Features	Train Accuracy	Test Accuracy
5	98,888	Pitches(12)	81.11%	54.71%
10	49,390	Pitches(12)	84.66%	50.90%
20	24,841	Pitches(12)	85.83%	43.34%

Στο Επαυξημένο Σύνολο

Sequence Length	#Sequences	#Features	Train Accuracy	Test Accuracy
5	1,584,000	Pitches(12)	87.00%	56.92%
10	792,000	Pitches(12)	89.61%	51.31%

Συμπεράσματα Κωδικοποιητή - Αποκωδικοποιητή

Η συγκεκριμένη αρχιτεκτονική είχε την καλύτερη επίδοση σε σχέση με τα μοντέλα που προηγήθηκαν. Επίσης η εκπαίδευση ήταν πολύ γρηγορότερη καθώς απαιτούνται πολύ λίγες εποχές (λιγότερες από 10) μέχρι να συγκλίνει. Βέβαια δυσχαιραίνεται η αξιολόγηση του μοντέλου κατά την εκπαίδευση, καθώς ο αποκωδικοποιητής λειτουργεί διαφορετικά απ' ότι κατά την πρόβλεψη. Μάλιστα στην υλοποίηση είναι διαφορετικό μοντέλο (ο αποκωδικοποιητής της εκπαίδευσης δεν επιστρέφει τις καταστάσεις του, ενώ της πρόβλεψης τις επιστρέφει). Συνέπεια είναι να αποφεύγεται δυσκολότερα το φαινόμενο του overfitting καθώς και η παρεμπόδιση μεθόδων όπως το Early Stopping (που μας επιτρέπει να κρατάμε τις παραμέτρους του μοντέλου στην εποχή με την καλύτερη επίδοση στο σύνολο δοκιμής).

Επιπρόσθετα, το μοντέλο έχει καλύτερη επίδοση όταν το δίκτυο του κωδικοποιητή είναι πιο πολύπλοκο, ενώ η πολυπλοκότητα του αποκωδικοποιητή δεν φαίνεται να επηρεάζει σε μεγάλο βαθμό. Αυτό μπορεί να οφείλεται στο γεγονός πως ο κωδικοποιητής επιτελεί δυσκολότερο έργο (ουσιαστικά να συμπυκνώσει την πληροφορία ολόκληρης της ακολουθίας στο διάνυσμα καταστάσεων). Ο αποκωδικοποιητής προβλέπει την επόμενη συγχορδία δεδομένης της παρούσας και των καταστάσεων, πρόβλημα με μικρότερη πολυπλοκότητα.

Κεφάλαιο 7 - Σφάλματα

7.1 Αναπόφευχτα Σφάλματα

Τα σύνολα δεδομένων που χρησιμοποιήθηκαν ενδέχεται να έχουν λάθη, όπως αναφέρεται στις σελίδες που τα προμηθεύουν. Επίσης μπορεί να προέκυψαν σφάλματα κατά τη συνένωση των επισημειομένων δεδομένων με τα χαρακτηριστικά του Spotify. Ενδεικτικά

Τα κομμάτια που προμηθεύονται, υπάρχουν σε πολλές εκδοχές στο Spotify (live, re-recording, remastered κλπ), και ενώ ελέγχθηκαν οι διάρκειες και οι πρώτες ή τελευταίες συγχορδίες των κομματιών προτού επιλεγεί η κάθε εκδοχή, εξαιτίας του τεράστιου όγκου δεδομένων είναι πιθανό να έγιναν λάθη.

Η ανάλυση ήχου που παρέχει το Spotify δεν είναι σε καμία περίπτωση τέλεια, όπως φαίνεται και από τις τιμές του confidence στα διάφορα segments. Επίσης φέρουν λιγότερη πληροφορία από το αρχικό αρχείο ήχου και επομένως είναι λογικό να μην μπορούμε να φτάσουμε πολύ υψηλά επίπεδα ακρίβειας.

7.2 Πόλωση Δεδομένων

Τα δεδομένα δεν αντιπροσωπεύουν το σύνολο της μουσικής, αλλά την κατανομή των συνόλων δεδομένων που χρησιμοποιήσαμε. Για παράδειγμα στα σύνολα δεδομένων δεν υπάρχουν έργα κλασικής μουσικής, όπως δεν υπάρχουν και από άλλα είδη μουσικής. Είναι αδύνατον να φτιαχτεί κάποιο σύνολο δεδομένων με κομμάτια μουσικής που να αντιπροσωπεύει το σύνολο της μουσικής που υπάρχει - αφού δεν γνωρίζει κανείς ποια είναι η πραγματική κατανομή.

7.3 Μουσικά Σφάλματα

Παρατηρώντας τους πίνακες σύγχυσης που κατασκευάστηκαν ύστερα από τα διαφορετικά πειράματα, συμπεράναμε πως μεγάλο ποσοστό των σφαλμάτων σε

όλες τις περιπτώσεις προέκυπτε από σύγχυση συγχορδιών που έχουν μεταξύ τους κοινές νότες. Το γεγονός αυτό ίσως να αναμενόταν από τη φύση των διαφορετικών μοντέλων (δεδομένου πως οι ίδιες οι νότες ήταν τα χαρακτηριστικά), αλλά με λίγη γνώση μουσικής θεωρίας θα αντιμετωπίζονταν πολλά από αυτά. Η αιτία είναι πως στη μουσική θεωρία υπάρχουν ορισμένοι κανόνες, οι οποίοι μεν δεν είναι αυστηροί αλλά ακολουθούνται στο μεγαλύτερο ποσοστό της δυτικής μουσικής. Πολλές από τις λάθος προβλέψεις συγχορδιών παραβιάζουν τέτοιου είδους κανόνες, γεγονός που θα αντιμετωπιζόταν εάν το σύστημα λάμβανε υπ' οψη τη θεωρία της μουσικής.

Κεφάλαιο 8 - Υλοποίηση

Όλα τα πειράματα και τα σενάρια (scripts) για τη συλλογή των δεδομένων υλοποιήθηκαν στη γλώσσα προγραμματισμού Python. Συγκεκριμένα, για την κατασκευή και εκπαίδευση των μοντέλων χρησιμοποιήθηκε το framework που έχει αναπτύξει η Google ειδικά για μηχανική μάθηση με νευρωνικά δίκτυα, το Tensorflow [24]. Τα πιο πολύπλοκα μοντέλα όπως το bLSTM υλοποιήθηκαν με την υψηλού επιπέδου αφαίρεση του tensorflow, το TFLearn [25]. Ο κωδικοποιητής - αποκωδικοποιητής υλοποιήθηκαν με τη βοήθεια της αντίστοιχης βιβλιοθήκης - του Keras [34]. Αυτά τα frameworks χρησιμοποιούνται ευρέως στον χώρο της μηχανικής μάθησης και παρέχουν βοηθητικές συναρτήσεις για εκπαίδευση, υπολογισμό μετρικών και επεξεργασία δεδομένων ενώ προσφέρουν τη δυνατότητα αξιοποίησης καρτών γραφικών (GPUs) για τους υπολογισμούς. Στις κάρτες γραφικών τα μοντέλα εκπαιδεύονται πολύ γρηγορότερα (~μία τάξη μεγέθους) καθώς είναι βελτιστοποιημένες για παράλληλες πράξεις με μεγάλους πίνακες (για να υπολογίζουν προβολές και άλλους μετασχηματισμούς που απαιτούνται για την εμφάνιση γραφικών).

Αξιοποιώντας την διεπαφή CUDA [31] που έχει αναπτύξει η κατασκευάστρια καρτών γραφικών NVIDIA, το Tensorflow έχει υλοποιημένες μεθόδους ώστε με μία γραμμή κώδικα, να μπορεί κανείς να τρέξει το μοντέλο του σε καρτά γραφικών (GPU) αντί για επεξεργαστή (CPU).

Κεφάλαιο 9 - Συμπεράσματα

Η προσέγγιση της εργασίας για το έργο της αυτόματης αναγνώρισης συγχορδίας ήταν αρκετά διαφορετική από άλλες. Υπήρχε σημαντικός περιορισμός στα δεδομένα και από άποψη πλήθους καθώς και ποιότητας. Συνήθως για αυτόματη αναγνώριση συγχορδίας, τα διανύσματα chroma (pitches) που χρησιμοποιούνται περιλαμβάνουν αναλυτικά τις νότες για περισσότερες από μία οκτάβες, ενώ η εξαγωγή των χαρακτηριστικών και η τμηματοποίηση του κομματιού γίνεται εξειδικευμένα για αναγνώριση συγχορδίας, είτε αυτή γίνεται αυτόματα με μηχανική μάθηση, είτε χειροκίνητα με επεξεργασία σήματος. Το Spotify παρέχει χαρακτηριστικά "γενικού σκοπού", τα οποία πρέπει να είναι φθηνά στον υπολογισμό, αφού υπολογίζονται για τα δεκάδες εκατομμύρια κομμάτια που υπάρχουν στην πλατφόρμα.

Επιπρόσθετα, δεν συνηθίζεται να χρησιμοποιούνται οι αρχιτεκτονικές που χρησιμοποιήθηκαν για την πρόβλεψη, αλλά για την εξαγωγή χαρακτηριστικών. Όταν επιχειρείται το έργο με χαρακτηριστικά υψηλού επιπέδου της μορφής της εργασίας, στη βιβλιογραφία χρησιμοποιούνται Κρυφά Μοντέλα Μαρκόβ (Hidden Markov Models), Βαθιά Δίκτυα Εμπιστοσύνης (Deep Belief Networks) και άλλα μοντέλα που δεν εξετάστηκαν στην παρούσα εργασία. Θα είχε ενδιαφέρον η σύγκριση της απόδοσης αυτών των μοντέλων στα χαρακτηριστικά του Spotify, με τα νευρωνικά δίκτυα που χρησιμοποιήθηκαν στην εργασία.

Ως προς την απόδοση των διαφορετικών μοντέλων, αν συλλέξουμε από κάθε αρχιτεκτονική την εκτέλεση με την καλύτερη επίδοση στο Test Set προκύπτει ο παρακάτω πίνακας:

Μοντέλο	**Σύνολο Εκπαίδευσης**	**Χαρακτηριστικά**	**Ακρίβεια στο Test Set**
Naive Bayes	Αρχικό	11*Pitches	43.54%
FNN	Αρχικό	11*Pitches	55.67%
CNN1	Αρχικό	11*Pitches	56.35%
CNN2	Αρχικό	23*Pitches	57.08%
LSTM-s2c	Αρχικό	20*(All-Timbre)	60.40%
LSTM2-s2s	Αρχικό	50*Pitches	54.40%

bLSTM-s2s	Επαυξημένο	10*All	57.60%
LSTMaec-s2s	Επαυξημένο	10*Pitches	62.71%

Παρατηρούμε καταρχάς πως οι απλές αρχιτεκτονικές νευρωνικών δικτύων είναι καλύτερες από μοντέλα όπως ο Naive Bayes. Επίσης, οι πιο πολύπλοκες όπως ο κωδικοποιητής-αποκωδικοποιητής LSTM ή το αμφίδρομο LSTM αποδίδουν καλύτερα από τις απλές.

Επίσης, τα απλούστερα μοντέλα φαίνεται να είναι πιο ευαίσθητα στην κατανομή των δεδομένων, αφού είχαν την καλύτερη επίδοση όταν εκπαιδεύονταν στο αρχικό σύνολο εκπαίδευσης. Αυτά τα μοντέλα ήταν πολωμένα προς τις συχνότερες κατηγορίες όπως φάνηκε στους πίνακες σύγχυσης. Αντίθετα, ο κωδικοποιητής - αποκωδικοποιητής και το αμφίδρομο LSTM ήταν αμερόληπτα ως προς τις κατηγορίες και είχαν καλύτερη επίδοση όταν εκπαιδεύονταν στο επαυξημένο σύνολο.

Τέλος, όταν χρησιμοποιούνταν παραπάνω χαρακτηριστικά πέραν των pitches, τα περισσότερα μοντέλα έτειναν να έχουν χειρότερη απόδοση. Προφανώς τα pitches είναι το σημαντικότερο χαρακτηριστικό για την πρόβλεψη, αλλά θα περίμενε κανείς να φέρουν χρήσιμη πληροφορία και τα υπόλοιπα χαρακτηριστικά. Όταν όμως το σύνολο δεδομένων ήταν το αρχικό (μη επαυξημένο), τα παραπάνω χαρακτηριστικά προκαλούσαν *overfitting*, καθώς πολλά από αυτά σχετίζονται με ολόκληρο το κομμάτι. Έτσι συσχέτισσαν άσκοπα τα ηχητικά χαρακτηριστικά ενός κομματιού (όπως το τί όργανα ακούγονται - timbre) με τη συγχορδία και προέκυπτε overfitting, αφού τα κομμάτια ήταν λίγα σε σχέση με τα segments (~1000 segments ανά κομμάτι). Στο επαυξημένο σύνολο φαίνεται να μην επηρεάζουν την απόδοση στις περισσότερες περιπτώσεις, αφού όταν δημιουργούνται νέα παραδείγματα εισόδου με τη διαδικασία της επαύξησης, μόνο ο πίνακας pitches αλλάζει. Τα υπόλοιπα χαρακτηριστικά αντιγράφονται ως έχουν, επομένως υπάρχουν τα ίδια ακριβώς χαρακτηριστικά συσχετισμένα με όλες τις κατηγορίες του προβλήματος.

Κεφάλαιο 10 - Βιβλιογραφία

[1] Stefan Koelsch κ.α. *Music, Language and Meaning: Brain Signatures of Semantic Processing*, in Nature Neuroscience, April 2004 https://www.nature.com/articles/nn1197

[2] Hermann Helmholtz. *On the Sensations of Tone*, *Chapter 1: On the Sensation of Sound in General*., Dover Publications, New York, 1954, ISBN: 0-486-60753-4, books.google (preview)

[3] Marco Costa, Pio Enrico Ricci Bitti and Luisa Bonfiglioili. *Psychological Connotations of Harmonic Musical Intervals*, Department of Psychology, University of Bologna, Viale Berti Pichat, 5, 1-40127 Bologna, Italy, link: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.577.9997&rep=rep1&type=pdf

[4] Eric J. Humphrey, Yann LeCun. *Moving Beyond Feature Design: Deep Architectures and Automatic Feature Learning in Music Informatics*, in ISMIR 2012, http://yann.lecun.com/exdb/publis/pdf/humphrey-ismir-12.pdf

[5] Filip Korzeniowski and Gerhard Widmer. *A Fully Convolutional Deep Auditory Model for Musical Chord Recognition "*, in arXiv:1612.05082 , 2016

[6] Filip Korzeniowski, Sebastian Bock, Florian Krebs and Gerhard Widmer. *MIREX Submissions for Chord Recognition and Key Estimation 2017,* from http://www.music-ir.org/mirex/abstracts/2017/KBK1.pdf

[7] Chris Cannam, Matthias Mauch. *MIREX 2017: VAMP Plugins from the Center for Digital Music*, from http://www.music-ir.org/mirex/abstracts/2017/CM2.pdf

[8] MIREX wiki. *2017 Audio Chord Estimation* , from http://www.music-ir.org/mirex/wiki/2017:Audio_Chord_Estimation#Data , visited December 2017

[9] Alexander Lerch. *Datasets*, from https://www.audiocontentanalysis.org/data-sets/,Visited December 2017

[10] Tristan Jehan, David DesRoches. *Analyzer Documentation*, January 7 2014, http://docs.echonest.com.s3-website-us-east-1.amazonaws.com/_static/AnalyzeDocumentation.pdf , visited December 2017

[11] Geoffroy Peeters. *Classifying Music Audio with Timbral and Chroma Features*, https://pdfs.semanticscholar.org/2c57/e9a1ec8ab8f850023a4e7ec0804be7248963.pdf

[12] V. Nair and G. E. Hinton. *Rectified linear units improve restricted boltzmann machines*. In Proc. 27th International Conference on Machine Learning, 2010

[13] Diederik P. Kingma, Jimmy Lei Ba. *ADAM: A Method for Stochastic Optimization*, 2017, arxiv:1412.6980

[14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov. *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*, 2014, link

[15] David D. Lewis. *Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval*, AT&T Labs - Research, NJ, USA, 1998 https://link.springer.com/content/pdf/10.1007%2FBFb0026666.pdf

[16] Y. Le Cun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, et al. *Handwritten digit recognition with a back-propagation network*. In Advances in neural information processing systems, 1990

[17] Katarzyna Janocha, Wojciech Marian Czarnecki, *On Loss Functions for Deep Neural Networks in Classification*, Deep Mind, London, UK, 2017

[18] Jesse Davis, Mark Goadrich. *The Relationship Between Precision-Recall and ROC Curves*, WI, USA, June 2006

[19] Dominguez-Torres, Alejandro. *Origin and history of convolution*, 2010 http://www.slideshare.net/Alexdfar/origin-adn-history-of-convolution

[20] Qiming Chen, Ren Wu . *CNN is all you Need*, NovuMind Inc, USA, 2017, arxiv:1712.09662

[21] Sepp Hochreiter, Jurgen Schmidhuber. *Long Short Term Memory*, Neural Computation, 1997, link: https://www.researchgate.net/publication/13853244_Long_Short-term_Memory

[22] Martin Sundermeyer, Ralf Schluter, and Hermann Ney. *LSTM Neural Networks for Language Modeling*, 2012, link: http://www.isca-speech.org/archive/archive_papers/interspeech_2012/i12_0194.pdf

[23] Alex Gravesa, Ju¨rgen Schmidhubera. *Framewise phoneme classification with bidirectional LSTM and other neural network architectures*, Munich, Germany, 2005

[24] Tensorlfow, An open source machine learning framework for everyone, https://www.tensorflow.org/

[25] TFlearn, Deep learning library featuring a higher-level API for TensorFlow., http://tflearn.org/

[26] Spotify Web Api, Spotify for Developers, https://developer.spotify.com/documentation/web-api/

[27] Ernst Levy. *A Theory of Harmony*, SUNI press, January 1, 1985

[28] Matt McVicar, Raúl Santos-Rodríguez, Yizhao Ni, and Tijl De Bie. *Automatic Chord Estimation from Audio: A Review of the State of the Art*, in IEEE, NJ, USA, February 2014

[29] Y. Ni, M. McVicar, R. Santos-Rodriguez, and T. De Bie. *An end-to-end machine learning system for harmonic analysis of music,* in IEEE, August 2012

[30] G. Zoia, Ruohua Zhou, D. Mlynek. *A multi-timbre chord/harmony analyzer based on signal processing and neural networks*, Multimedia Signal Processing 2004 IEEE 6th Workshop on, pp. 219-222, 2004

[31] CUDA compute platform, NVIDIA, https://developer.nvidia.com/about-cuda

[32] Maximilian Panzner and Philipp Cimiano. *Comparing Hidden Markov Models and Long Short Term Memory Neural Networks for Learning*, Action Representations Semantic Computing Group, CITEC, Bielefeld University, 2016

[33] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, Yoshua Bengio. *On the Properties of Neural Machine Translation: Encoder-Decoder Approaches*, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, **arXiv:1409.1259**

[34] Keras: The Python Deep Learning Library, https://keras.io/

[35] Long short-term memory. In Wikipedia. Retrieved March, 2018, from https://en.wikipedia.org/wiki/Long_short-term_memory

[36] Recurrent neural network. In Wikipedia. Retrieved March, 2018, from https://en.wikipedia.org/wiki/Recurrent_neural_network

[37] Nicolas Boulanger-Lewandowski, Yoshua Bengio and Pascal Vincent. *Audio Chord Recognition with Recurrent Neural Networks, in ISMIR 2013*

[38] Simon Haykin, Neural Networks and Learning Machines 3rd Edition, ISBN-13:978-0-13-147139-9, 2009