ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΠΟΛΟΓΙΣΤΩΝ

Σημασιολογική κωδικοποίηση ηχητικών δεδομένων

Semantic audio & metadata

Υπεύθυνος καθηγητής
Ιωάννης Μουρτζόπουλος

Μάθημα
Ψηφιακή Τεχνολογία Ήχου (ECE_AK809)

Ονοματεπώνυμο:	Σταύρος Κάνιας
Αριθμός Μητρώου:	1066563
Τομέας:	Ηλεκτρονικής και Υπολογιστών
Έτος φοίτησης:	4ο

Περιεχόμενα

1. Εισαγωγή

Η παρούσα εργασία έχει ως στόχο την παρουσίαση γενικών αρχών της σημασιολογικής κωδικοποίησης ηχητικών σημάτων. Το θέμα αυτό έχει έρθει στο προσκήνιο τα τελευταία χρόνια λόγω της ανάγκης εμπλουτισμού και αυτόματης της παραγωγής των σύγχρονων μεταδεδομένων ώστε να εξυπηρετήσουν τις ανάγκες των τεχνολογικών εξελίξεων (Semantic Web, Metaverse, Streaming Services). Από άποψη δομής η εργασία ξεκινά παρουσιάζοντας την παρούσα κατάσταση στον τομέα των ηχητικών μεταδεδομένων αναλύοντας παράλληλα τον κομβικό ρόλο που κατέχει η σημασιολογική κωδικοποίηση του ήχου για την εξέλιξή τους. Έπειτα από μια σύντομη αναφορά σε θεμελιώδεις έννοιες της μουσικής μορφολογίας παρουσιάζεται η συνήθης αλγοριθμική διαδικασία (pipeline) που ακολουθείται για την εξαγωγή συμπερασμάτων και μεταδεδομένων για ένα μουσικό κομμάτι βάση του περιεχομένου του. Ακολουθεί εκτενής αναφορά στην εξαγωγή των βασικών μουσικών χαρακτηριστικών που χρησιμοποιούνται στη σύγχρονη σημασιολογική ανάλυση μουσικών σημάτων. Ιδιαίτερη έμφαση δίνεται στον πίνακα αυτο-ομοιότητας, ένα χαρακτηριστικό που αποτελεί το τελευταίο στάδιο της σημασιολογικής ανάλυσης και είναι είτε πηγή απευθείας εξαγωγής μουσικής πληροφορίας είτε η είσοδος σε κάποιο σύστημα μηχανικής μάθησης. Όλα όσα περιγράφονται στην εργασία έχουν υλοποιηθεί προγραμματιστικά σε γλώσσα Python και σε κάθε στάδιο της ανάλυσης παρουσιάζονται τα αποτελέσματα που προκύπτουν από την επεξεργασία ενός πραγματικού κομματιού που χρησιμοποιήθηκε για τις ανάγκες της εργασίας.

2. Μέθοδοι και Τεχνολογίες

2.1 Μεταδεδομένα

2.1.1 Κλασσική προσέγγιση και κατηγορίες μεταδεδομένων

Οι περισσότερες υπηρεσίες διανομής μουσικής ροής (streaming services) στη σύγχρονη βιομηχανία χρησιμοποιούν μεταδεδομένα (metadata) για να οργανώσουν τη μουσικής τους πληροφορία, να την προσαρμόσουν στο χρήστη με τη μορφή λιστών αναπαραγωγής (playlist) και για να βελτιώσουν την εμπειρία αναζήτησης της από το χρήστη [1]. Στην πραγματικότητα τα κλασσικά μεταδεδομένα δεν είναι κάτι άλλο από πληροφορία υπό τη μορφή κειμένου η οποία περιγράφει τα μουσικά, τεχνολογικά ή και νομικά χαρακτηριστικά ενός σήματος ήχου [2]. Τα μεταδεδομένα μπορεί να είναι είτε ενσωματωμένα σε επίπεδο φυσικού (CD) ή ψηφιακού (ID3 για αρχεία MP3) μέσου αποθήκευσης είτε να βρίσκονται εντός κάποιας βάσης δεδομένων και να συσχετίζονται με το αρχείο ήχου [3]. Επιπλέον ανάλογα με τον τύπο τους τα μεταδεδομένα χωρίζονται σε τρεις κατηγορίες:

1. *Μεταδεδομένα περιεχομένου*

Σε αυτή την κατηγορία μεταδεδομένων ανήκουν ο τίτλος, το όνομα του καλλιτέχνη, το όνομα του παραγωγού, η ημερομηνία κυκλοφορίας και το μουσικό είδος στο οποίο υπάγεται το κομμάτι.

2. *Μεταδεδομένα διαχείρισης*

Σε αυτή την κατηγορία μεταδεδομένων ανήκουν τα νομικά δικαιώματα διανομής και αναπαραγωγής, τα πνευματικά δικαιώματα του καλλιτέχνη και στατιστικά προτίμησης από τους ακροατές.

3. *Τεχνικά μεταδεδομένα*

Σε αυτή την κατηγορία μεταδεδομένων ανήκουν η κωδικοποίηση του αρχείου και η ψηφιακή του μορφή.

2.1.2 Προβλήματα της κλασσικής προσέγγισης και επίλυσή τους μέσω της σημασιολογικής κωδικοποίησης

Το κλασσικό μοντέλο των μουσικών μεταδεδομένων που περιγράψαμε παραπάνω έχει τρία βασικά προβλήματα [4]:

1. *Απαιτεί διαρκή ενημέρωση με μη αυτοματοποιημένο τρόπο*

Για κάθε νέο κομμάτι που εισάγεται στη βάση δεδομένων χρειάζεται ανθρώπινη παρέμβαση για να εισαχθούν οι πληροφορίες κειμένου ως μεταδεδομένα. Παράλληλα κάθε φορά που θέλουμε να προσθέσουμε μια νέα πληροφορία στα μεταδεδομένα θα πρέπει και πάλι να προστεθεί χειροκίνητα η νέα πληροφορία στα μεταδεδομένα του κάθε κομματιού.

2. *Δεν παρέχει πληροφορίες για τη βιωματική αντίληψη του κομματιού από το χρήστη.*

Είναι αδύνατο ο χρήστης να αναζητήσει πληροφορία βάση κάποιου μουσικού ή και συναισθηματικού χαρακτηριστικού το οποίο φέρει κάποιας μορφής υποκειμενικότητα. Για παράδειγμα είναι αδύνατο ο χρήστης με το κλασσικό μοντέλο μεταδεδομένων να αναζητήσει «χαρούμενη» ή «θλιμμένη», «ζωντανή» ή «χαλαρωτική» μουσική καθώς οι χαρακτηρισμοί αυτοί είναι γενικότερα υποκειμενικής φύσεως και απαιτούν κάποια μουσική γνώση για να εξαχθούν.

3. *Δίνει δυνατότητα αναζήτησης μόνο βάση κειμένου.*

Βάση του κλασσικού μοντέλου μεταδεδομένων είναι αδύνατο ο χρήστης να αναζητήσει ένα κομμάτι με χρήση κάποιου ηχητικού σήματος που είτε είναι χαρακτηριστικό του κομματιού από ρυθμικής ή μελωδικής άποψης, είτε περιέχει τμήμα των στίχων του.

Η σύγχρονη λύση στα παραπάνω προβλήματα είναι η χρήση μεταδεδομένων που παράγονται αυτόματα και βασίζονται αποκλειστικά στο ηχητικό σήμα χωρίς να απαιτούν κάποια εξωτερική πληροφορία. Αυτού του είδους τα μεταδεδομένα αποθηκεύονται σε ταμπέλες (tags), εντός του συνόλου των μεταδεδομένων, οι οποίες περιέχουν μουσική και συναισθηματική πληροφορία προερχόμενη από ισχυρούς αλγόριθμους εξαγωγής χαρακτηριστικών, σημασιολογικής κωδικοποίησης και μηχανικής μάθησης. Ως αποτέλεσμα δίνεται η δυνατότητα αναπαράστασης του ρυθμού, της χρονικής αγωγής, του μουσικού κλειδιού, της διαδοχής συγχορδιών, της μουσικής φόρμας, της συναισθηματικής διάθεσης και πολλών άλλων μουσικών και συναισθηματικών χαρακτηριστικών ενός κομματιού εντός των μεταδεδομένων του. Ταυτόχρονα είναι δυνατή η επεξεργασία ενός ηχητικού σήματος εισόδου και χρήση του ως πηγή αναζήτησης. Για να εκμεταλλευτούμε τις παραπάνω πληροφορίες μεταβαίνουμε από ένα μοντέλο ανάκτησης πληροφορίας βάση λεκτικής πληροφορίας σε ένα μοντέλο ανάκτησης με βάση το μουσικό περιεχόμενο (content-based retrieval). Οι βασικές ενέργειες αυτοματοποιημένης επεξεργασίας ενός μουσικού σήματος που χρησιμοποιούνται έντονα από τις σύγχρονες εφαρμογές είναι οι εξής:

1) Δομική ανάλυση
2) Εξαγωγή χαρακτηριστικού τμήματος (thumbnailing)
3) Αναγνώριση ρυθμού
4) Αναγνώριση χρονικής αγωγής
5) Αναγνώριση συγχορδιών
6) Εξαγωγή μουσικής ταυτότητας κομματιού

7) Εντοπισμός μέτρου ομοιότητας μεταξύ δύο κομματιών
8) Αναγνώριση εκτέλεσης
9) Αναγνώριση μουσικού κλειδιού
10) Εξαγωγή μελωδίας
11) Διαχωρισμός κρουστών και μελωδικών οργάνων
12) Εξαγωγής ανά νότα αναπαράστασης ενός κομματιού

Στην παρούσα εργασία θα ασχοληθούμε με τις εφαρμογές της σημασιολογικής κωδικοποίησης ηχητικών σημάτων στην αυτοματοποιημένη εξαγωγή δομικών μεταδεδομένων για ένα μουσικό κομμάτι. Συγκεκριμένα θα περιγράψουμε τα στάδια ανάλυσης για την αυτόματη εξαγωγή του χαρακτηριστικού μουσικού τμήματος (thumbnailing) ενός μουσικού κομματιού.

2.1 Μουσική μορφολογία

Το βασικό στοιχείο διαφοροποίησης της μουσικής από κάθε άλλου είδους θόρυβο είναι η ύπαρξη ιεραρχικής δομής. Από το επίπεδο μεμονωμένων φθόγγων, μεταβαίνουμε σε ένα επίπεδο απλών μουσικών φράσεων και φτάνουμε τελικά σε μια αφηρημένη μορφολογική προσέγγιση ενός μουσικού κομματιού που περιγράφει το λειτουργικό ρόλο κάθε συνόλου φράσεων στη χρονική εξέλιξή του (εισαγωγή, κουπλέ, ρεφρέν, οίκος, υπακοή κ.α). Είναι αρκετά εύκολο να αναγνωρίσουμε τα τμήματα αυτά μέσω μιας παρτιτούρας όπως, αυτή που φαίνεται στην Εικόνα 1, όπου συχνά ο συνθέτης έχει καταγράψει τη δομή της σύνθεσής του. Η αυτοματοποιημένη όμως τμηματοποίηση μιας ηχητικής κυματομορφής σε μουσικά τμήματα αποτελεί ένα σύνθετο πρόβλημα λόγω της χαοτικής μορφής των ηχητικών σημάτων όπως αυτού που φαίνεται στην Εικόνα 2.

Ευτυχώς για εμάς μπορούμε να βασιστούμε σε κάποια βασικά μουσικά χαρακτηριστικά ώστε να πραγματοποιήσουμε αυτόματη αναγνώριση της μορφολογικής δομής ενός μουσικού κομματιού και να το διαχωρίσουμε σε μουσικές ενότητες. Τα χαρακτηριστικά αυτά είναι:

1. Επανάληψη

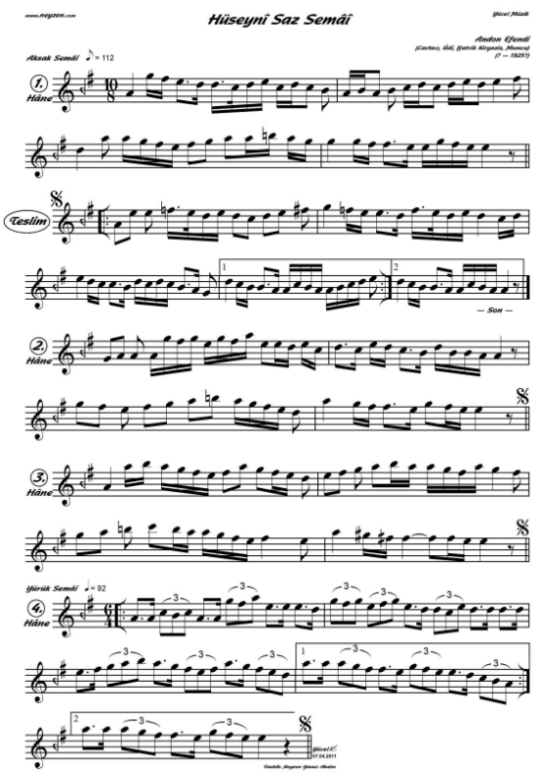
Συχνά στα μουσικά κομμάτια, η κεντρική μουσική ιδέα επαναλαμβάνεται, με μικρές διαφορές, αρκετές φορές κατά του κομματικού. Βασιζόμενοι σε αυτό μπορούμε να απομονώσουμε τμήματα όπου εμφανίζεται μια αλληλουχία φθόγγων, διαστημάτων ή και συγχορδιών.

2. Καινοτομία

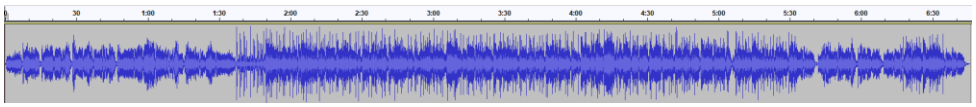
Η έννοια της καινοτομίας στην σημασιολογικής κωδικοποίησης της μουσικής αντιστοιχεί στην έννοια της χρονικής εξέλιξης και των αλλαγών που εμφανίζονται κατά τη διάρκειά της. Αποτελεί πολύ συχνό φαινόμενο σε μία σύνθεση η εμφάνιση έντονων αλλαγών στο τέμπο, το ρυθμό, το συχνοτικό περιεχόμενο και τη δυναμική. Σημεία μετάβαση από ‘γρήγορα’ τμήματα σε αργά ή από ‘απαλά’ τμήματα σε ‘έντονα’ είναι ιδανικοί υποψήφιοι για διαχωρισμό μουσικών τμημάτων.

3. Ομοιογένεια

Είναι αρκετά αξιόπιστη η υπόθεση πως τμήματα με κοινό συχνοτικό περιεχόμενο (και άρα παρόμοια ενορχήστρωση), παρόμοιο τέμπο ή και παρόμοιο αρμονικό ρυθμό κατέχουν είναι την ίδια είτε συμμετρική μορφολογική θέση στη χρονική εξέλιξη ενός μουσικού κομματιού. Ένας πίνακας που περιγράφει τα παραπάνω χαρακτηριστικά στην περίπτωση του κομματιού της Εικόνας 2 φαίνεται στην Εικόνα 3.



Εικόνα 1: Huseyni Saz Semai, Αντώνης Κυριαζής, 19ος αιώνας.



Εικόνα 2: Αναπαράσταση του κομματιού της Εικόνας 1 στο πεδίο του χρόνου. [5]

Μουσικά τμήματα	Αυτοσχεδιαστική Εισαγωγή	1ος Οίκος	Υπακοή	2ος Οίκος	Υπακοή	3ος Οίκος	Υπακοή	4ος Οίκος	Υπακοή	Υπακοή
Επανάληψη	I	A1	B1	A2	B1	A3	B1	A4	B1	B1
Ρυθμός	Ελεύθερα	10/8	10/8	10/8	10/8	10/8	10/8	6/4	Ελεύθερα	10/8
Ενορχήστρωση	Σόλο	Σύνολο	Σύνολο	Σύνολο	Σύνολο	Σύνολο	Σύνολο	Σύνολο	Σόλο	Σύνολο

Εικόνα 3: Μορφολογική ανάλυση του κομματιού της Εικόνας 1.

2.2 Η αλγοριθμική προσέγγιση

Η πλέον σύγχρονη μέθοδος επίλυσης αυτού του προβλήματος αυτοματοποιημένης εξαγωγής μουσικών τμημάτων από ένα μουσικό κομμάτι βασίζεται σε δύο βασικές έννοιες, την εξαγωγή μουσικών χαρακτηριστικών και τους πίνακες αυτό-ομοιότητας. Η αλγοριθμική αναπαράσταση της διαδικασίας επίλυσης του προβλήματος παρουσιάζεται παρακάτω.

Βήμα 1ο

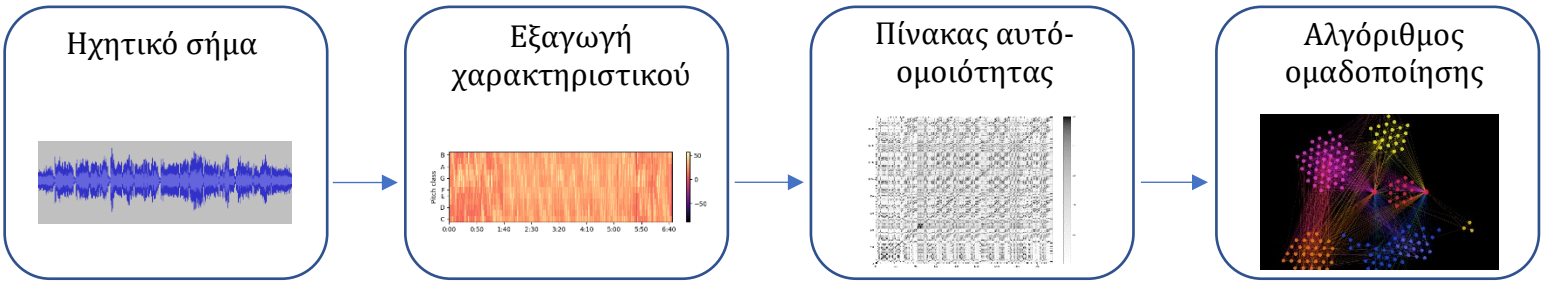
Το μουσικό σήμα μετασχηματίζεται σε διάφορες σημασιολογικές διαστάσεις-αναπαραστάσεις μουσικών χαρακτηριστικών οι οποίες μπορούν να περιγράψουν μουσικές έννοιες σε μια γλώσσα που είναι κατανοητή από τον υπολογιστή.

Βήμα 2ο

Για κάθε μια από τις σημασιολογικές αναπαραστάσεις που λάβαμε από το προηγούμενο βήμα υπολογίζεται ένα πίνακας που περιγράφει το βαθμό στον οποίο ένα τμήμα ταυτίζεται μουσικά με όλα τα υπόλοιπα τμήματα του κομματιού. Ο πίνακας αυτός ονομάζεται πίνακας αυτό-ομοιότητας. Στον κάθε πίνακα αυτό-ομοιότητας εμφανίζονται είτε δομές τύπου δέσμης είτε δομές τύπου μονοπατιού οι οποίες φανερώνουν ομοιότητα μουσικών τμημάτων είτε ως προς την ομοιομορφία είτε ως προς την επανάληψη αντίστοιχα.

Βήμα 3ο

Τροφοδοτούμε ένα μοντέλο μηχανικής μάθησης που εκτελεί έναν αλγόριθμο ομαδοποίησης με τα δεδομένα ομοιότητας που λάβαμε από τους πίνακες αυτό-ομοιότητας του προηγούμενου βήματος. Ο αλγόριθμος εντοπίζει της δομές ομοιότητας που αναφέραμε παραπάνω και ομαδοποιεί όλα τα τμήματα που τις εμφανίζουν σε μια κοινή συστάδα. Με τον τρόπο αυτό μπορούμε να συμπεράνουμε εάν ένα τμήμα επαναλαμβάνεται, με ποια άλλα τμήματα ταυτίζεται, το βαθμό στον οποίο ταυτίζεται με αυτά αλλά και το λειτουργικό του ρόλο μέσα στο κομμάτι ανάλογα με το σημείο στο οποίο το κάθε τμήμα εμφανίζεται. Το διάγραμμα ροής που περιγράφει την παραπάνω διαδικασία φαίνεται στην Εικόνα 4.



Εικόνα 4: Διάγραμμα ροής σύγχρονων αλγορίθμων μορφολογικής ανάλυσης.

2.3 Εξαγωγή χαρακτηριστικών και διαστάσεις ανάλυσης

Πολλές φορές είναι αναγκαίο να μην προσεγγίσουμε ένα πρόβλημα μέσω της πρώτης και διαισθητικής διάστασης που προκύπτει από την παρατήρηση αλλά να αναζητήσουμε κρυμμένα χαρακτηριστικά-διαστάσεις που μας δίνουν βαθύτερες πληροφορίες για τη φύση του προβλήματος που αναλύουμε. Στην περίπτωσή μας, το ηχητικό σήμα αποτελεί στην πραγματικότητα ένα σύνολο από μετρήσεις ηχητικής πίεσης στην πορεία του χρόνου πράγμα που το καθιστά ιδιαίτερα χαοτικό και δύσκολο στην επεξεργασία. Στην πορεία των χρόνων όμως έχουν ανακαλυφθεί νέα, πολύ χρήσιμα για την ανάλυση ενός ηχητικού σήματος, χαρακτηριστικά τα βασικότερα εκ των οποίων θα αναλύσουμε παρακάτω.

2.3.1 Τα χρωματικά χαρακτηριστικά (chroma features)

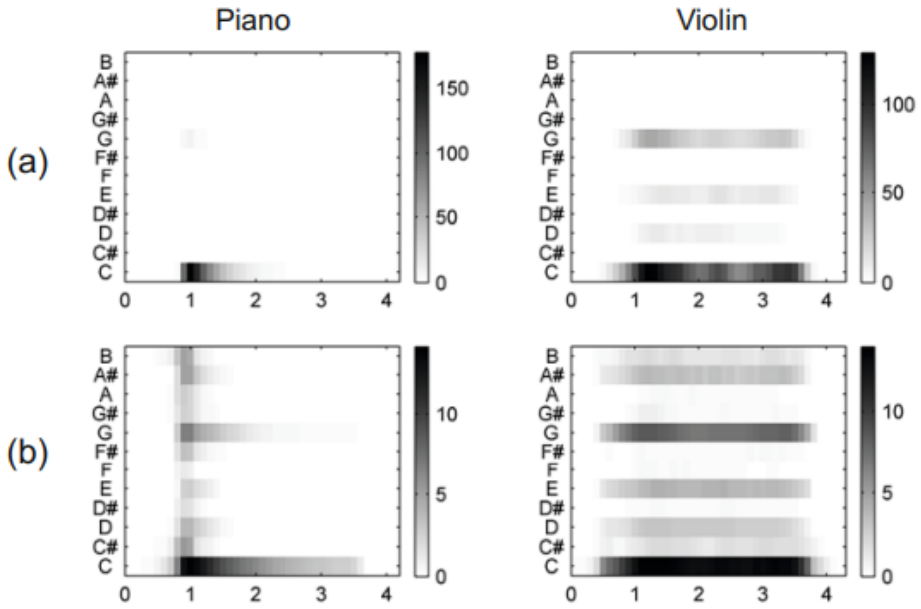
Βασιζόμενοι στο γεγονός ότι η ανθρώπινη ακοή επεξεργάζεται το συχνοτικό περιεχόμενο ενός ήχου με περιοδικό τρόπο, αναγνωρίζοντας τη διαφορά μεταξύ δύο φθόγγων που απέχουν κατά μια οκτάβα, μπορούμε να διαχωρίσουμε την έννοια της νότας από την έννοια του απόλυτου τονικού της ύψους. Με τον τρόπο αυτό μπορούμε να ονομάσουμε ως χρώμα ενός φθόγγου τη θέση του εντός του συνόλου των φθόγγων ενός είδους μουσικής και ως τονικό ύψος ενός φθόγγου την οκτάβα στη οποία ο φθόγγος αυτός ανήκει. Στην περίπτωση της ευρωπαϊκής κλασσικής μουσικής για παράδειγμα το χρώμα ενός φθόγγου θα ήταν η τιμή του εντός του συνόλου {C, C#, …B} ενώ το τονικό του ύψους θα ήταν η αρίθμηση της οκτάβας στην οποία ανήκει, για παράδειγμα C1 ή G3. Ως τονική οικογένεια ορίζουμε το σύνολο των φθόγγων που μοιράζονται το ίδιο χρώμα. Ένα τέτοιο σύνολο είναι το σύνολο των φθόγγων {C1, C2, C3,…}. Η παραπάνω ανάλυση μας οδηγεί στην ιδέα πώς η πληροφορία της ‘ποσότητας’ μιας τονικής οικογένειας σε ένα μουσικό κομμάτι είναι ιδιαίτερα χρήσιμη αλλά και αρκετά εύκολη στον υπολογισμό με χρήση φασματογραφήματος. Έχοντας στη διάθεσή μας ένα τονικό φασματογράφημα λογαριθμικής κλίμακας $y_{LF} : Z \cdot [0:127] \rightarrow R_{\geq 0}$ η αναπαράσταση των χρωματικών χαρακτηριστικών

ενός ηχητικού σήματος ή αλλιώς χρωματογραφήματος $Z \cdot [0:11] \rightarrow R_{\geq 0}$ μπορεί να παραχθεί από το άθροισμα της ενέργειας που εμφανίζεται σε κάθε τονική οικογένεια όπως φαίνεται στην παρακάτω σχέση.

$$C(n,c) := \sum_{\{p \in [0:127] : p \bmod 12 = c} y_{LF}(n,p) \qquad (2.3.1.1)$$

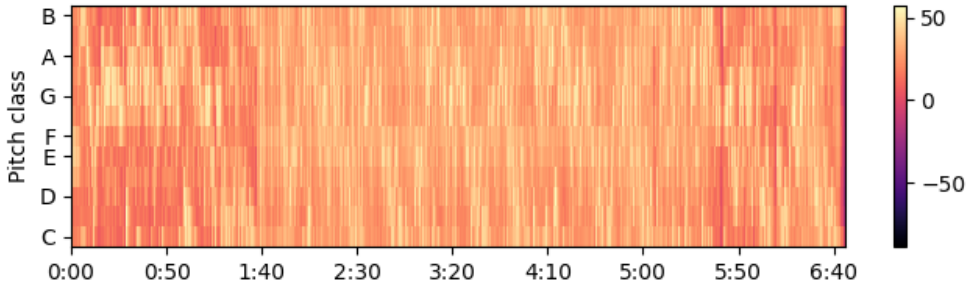
Όπου $c \in [0:11]$.

Ένα ενδιαφέρον παράδειγμα ανάλυσης χρωματικών χαρακτηριστικών φαίνεται παρακάτω.



Εικόνα 5: a) Αναπαράσταση χρωματικών χαρακτηριστικών της νότας C4 από πιάνο στα αριστερά και από βιολί στα δεξιά. b) Οι ίδιες χρωματικές αναπαραστάσεις μετά από λογαριθμική συμπίεση.

Από την παραπάνω εικόνα γίνεται φανερή η αξία αναπαράστασης των χρωματικών χαρακτηριστικών. Παρατηρούμε ότι η ίδια νότα παιγμένη από διαφορετικό όργανο παρουσιάζει διαφορετική χρωματική ταυτότητα λόγω των αρμονικών συχνοτήτων που δημιουργούν τη χροιά του οργάνου. Με τον τρόπο αυτό μπορούμε να αναγνωρίσουμε σημεία ενός κομματιού όπου ακούγεται ένα συγκεκριμένο όργανο ή μια συγκεκριμένη ενορχήστρωση και να το χαρακτηρίσουμε ως ξεχωριστό τμήμα. Παρακάτω φαίνεται το χρωματογράφημα του κομματιού της Εικόνας 6.



Εικόνα 6: Χρωματογράφημα του κομματιού της Εικόνας 2.

2.3.2 Τα προσαρμοσμένα φασματικά χαρακτηριστικά (cepstral features)

Παρότι τα χρωμογραφήματα είναι πολύ χρήσιμα όταν η ανάλυση γίνεται από μια μηχανή, πολύ συχνά χρειαζόμαστε να προσομοιώσουμε την ανθρώπινη ακοή ώστε να μπορέσουμε να προσεγγίσουμε την πραγματική αντίληψη του ακροατή ως προς το συχνοτικό περιεχόμενο ενός κομματιού. Για παράδειγμα, στην προσπάθεια εντοπισμού του τμήματος με το μεγαλύτερο αντίκτυπο στον ακροατή θα πρέπει να λάβουμε υπόψιν μας κατά πόσο το κάθε τμήμα ενός μουσικού κομματιού είναι αντιληπτό από τον ακροατή. Παρομοίως, δεν θα έχει νόημα να χαρακτηρίσουμε ένα τμήμα Α διαφορετικό από το τμήμα Β στην περίπτωση που εμφανίζουν μεν διαφορετικά συχνοτικά χαρακτηριστικά τα οποία όμως δε γίνονται αντιληπτά στην πράξη από τον ακροατή ο οποίος θα εκλάβει τα δύο τμήματα ως πανομοιότυπα. Για το σκοπό αυτό λοιπόν χρησιμοποιούμε την τράπεζα φίλτρων Mel τα οποία προσαρμόζουν το συχνοτικό περιεχόμενο ενός ηχητικού σήματος στο ανθρώπινο αυτί. Παρακάτω περιγράφονται σύντομα τα βήματα τις διαδικασίας εξόρυξης χαρακτηριστικών μέσω των φίλτρων Mel [6]. Στο τέλος της διαδικασίας καταλήγουμε σε κάποιες σταθερές που περιγράφουν ένα ηχητικό σήμα βάση της ανθρώπινης ακοής και οι οποίες ονομάζονται σταθερές cepstral.

1. Προέμφαση

Σε αυτό το στάδιο εφαρμόζουμε ένα φίλτρο προέμφασης το οποίο αφαιρεί συχνότητες όπως αυτές που εισάγονται από την ταλάντωση των χειλιών και της γλώσσας κατά την εξαγωγή του ήχου από το στόμα, ενισχύοντας παράλληλα τις υψηλές φωνητικές συχνότητες αντισταθμίζοντας την απότομη εκ φύσεως απόσβεσή τους. Το στάδιο αυτό πραγματοποιείται κυρίως σε σήματα φωνής. Το πιο συχνά χρησιμοποιούμενο φίλτρο προέμφασης είναι το:

$$H(z) = 1 - b \cdot z^{-1} \qquad (2.3.2.1)$$

Με το b να ορίζει την κλίση του φίλτρου στις υψηλές συχνότητες και να λαμβάνει τιμές συνήθως μεταξύ 0.4 και 1.0.

2. Παραθυροποίηση

Λόγω της εγγενούς αστάθειας του ηχητικού σήματος επιθυμούμε την ανάλυσή του σε μικρή περίοδο χρόνου. Για το λόγο αυτό εφαρμόζουμε παράθυρα της τάξης των $20ms$ τα οποία προχωράνε ανά $10ms$ σε κάθε καρέ. Ο λόγος που αφήνουμε τα παράθυρα να συμπίπτουν είναι πως έτσι αυξάνεται σημαντικά η πιθανότητα ένα ηχητικό φαινόμενο να βρεθεί στο κέντρο του παραθύρου και έτσι να αναλυθεί πλήρως εμφανίζοντας μεγαλύτερη ακρίβεια στο φάσμα του. Στο βήμα αυτό χρησιμοποιούνται συνήθως

παράθυρα Hanning ή Hamming. Με τον τρόπο αυτό τρόπο επίσης βοηθάμε την ενίσχυση των αρμονικών και την εξομάλυνση των ηχητικών ακμών που εμφανίζονται στο επόμενο βήμα.

3. Εφαρμογή Γρήγορου Μετασχηματισμού Fourier.

Σε αυτό το βήμα λαμβάνουμε το διακριτό φάσμα του σήματος που λάβαμε από το προηγούμενο βήμα χρησιμοποιώντας το Γρήγορο Μετασχηματισμό Fourier όπως φαίνεται παρακάτω:

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{\frac{-j \cdot 2\pi \cdot n \cdot k}{N}} \tag{2.3.2.2}$$

Όπου $0 \leq k \leq N-1$ και N ο αριθμός των σημείων του μετασχηματισμού.

4. Εξαγωγή Mel φάσματος

Στο σημείο αυτό ήρθε η ώρα να προσαρμόσουμε το πραγματικό φάσμα του σήματος στο ανθρώπινο αυτί εφαρμόζοντας μια σειρά από ζωνοπερατά, μη γραμμικά και συνήθως τριγωνικά φίλτρα τα οποία προέρχονται από την τράπεζα φίλτρων Mel και εξάγονται από την παρακάτω σχέση:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \dfrac{2 \cdot (k - f(m-1))}{f(m) - f(m-1)}, & f(m-1) \leq k \leq f(m) \\ \dfrac{2 \cdot (f(m+1) - k)}{f(m+1) - f(m)}, & f(m) \leq k \leq f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{2.3.2.3}$$

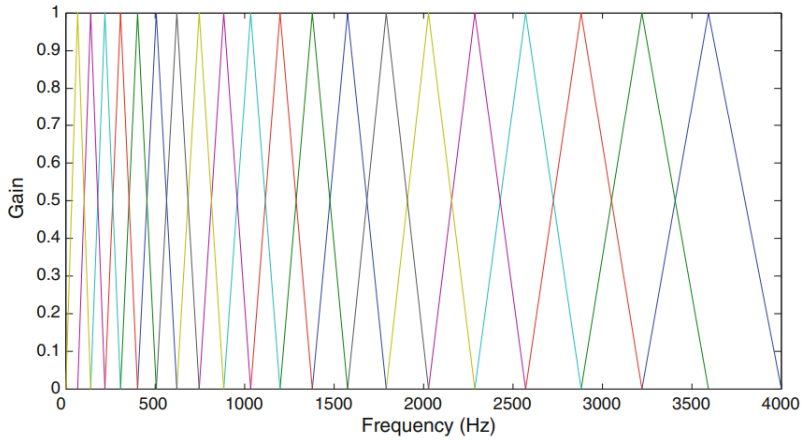
Πιο συγκεκριμένα στην παραπάνω σχέση M είναι το πλήθος των φίλτρων που αποφασίζουμε να χρησιμοποιούμε και $H_m(k)$ το βάρος που δίνεται στην φασματική ενεργειακή ζώνη k κατά τη συνεισφορά της στη ζώνη εξόδου m, με το m να λαμβάνει τιμές από 0 έως $M-1$.

Έχοντας παράγει τα παραπάνω φίλτρα μπορούμε να εξάγουμε το φάσμα Mel του φάσματος $X(k)$ πολλαπλασιάζοντας το μέτρο του με κάθε ένα από τα φίλτρα $H_m(k)$ και λαμβάνοντας το συνολικό άθροισμα. Αυτή η διαδικασία φαίνεται στην παρακάτω σχέση:

$$s(m) = \sum_{k=0}^{N-1} |X(k)^2| \cdot H_m(k) \tag{2.3.2.4}$$

Στην πράξη από την παραπάνω σχέση προκύπτει ότι η κλίμακα Mel είναι προσεγγιστικά γραμμική κάτω από το $1\ kHz$ και μη γραμμική πάνω από το $1\ kHz$. Ένας προσεγγιστικός τύπος που συχνά χρησιμοποιείται για την εξαγωγή της αντιληπτής συχνότητας κλίμακας f_{Mel} που προκύπτει από την ακοή της απόλυτης συχνότητας f είναι ο ακόλουθος:

$$f_{Mel} = 2595 \cdot \log_{10}(1 + \frac{f}{700}) \tag{2.3.2.5}$$



Εικόνα 7: Τράπεζα των πρώτων 20 φίλτρων Mel.

5. Εφαρμογή διακριτού μετασχηματισμού συνημιτόνου

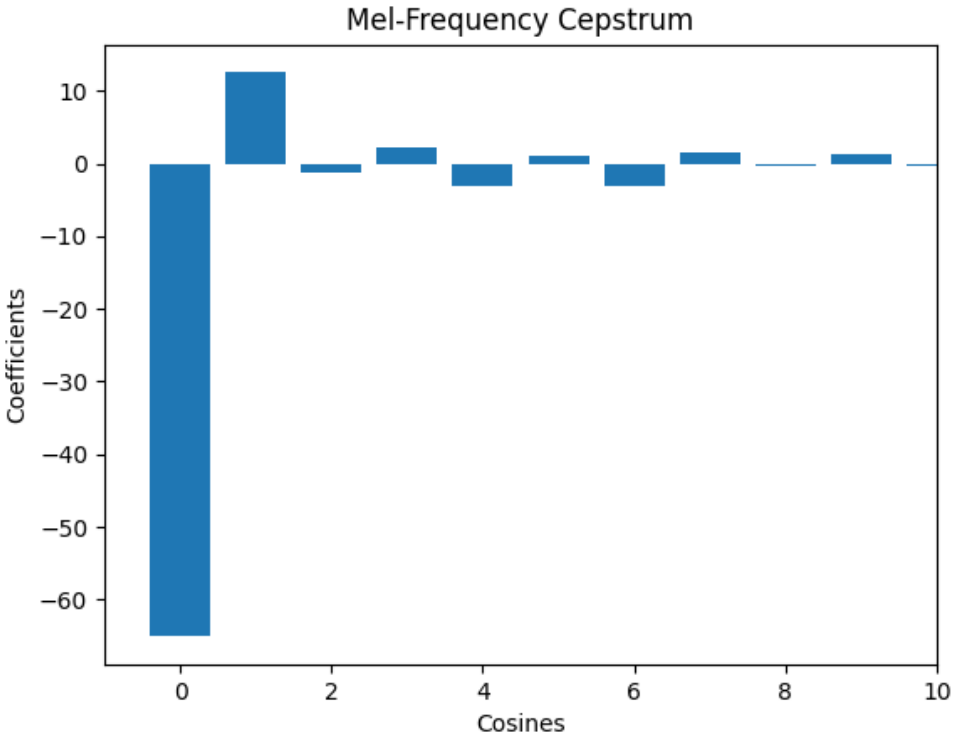
Στο τελευταίο βήμα της διαδικασίας χρειάζεται να αποσυσχετίσουμε τις ενεργειακές στάθμες γειτονικών περιοχών οι οποίες συσχετίστηκαν με την εφαρμογή των φίλτρων $\boldsymbol{Mel}$ στο προηγούμενο βήμα. Αυτό επιτυγχάνεται με τη χρήση ενός διακριτού μετασχηματισμού συνημιτόνου. Εδώ αξίζει να παρατηρήσουμε ότι μετασχηματίζουμε ουσιαστικά το φάσμα ενός σήματος δηλαδή εφαρμόζουμε μετασχηματισμό επί του μετασχηματισμού Fourier. Ως αποτέλεσμα φεύγουμε από το πεδίο συχνότητας και περνάμε στο πεδίο 'cepstral' (αναγραμματισμός του spectral). Προτού δε εφαρμόσουμε το διακριτό μετασχηματισμό συνημιτόνου, για να έρθουμε ακόμη πλησιέστερα στην ανθρώπινη ακοή, λογαριθμίζουμε το σήμα που προέκυψε από το προηγούμενο στάδιο. Τα παραπάνω περιγράφονται μαθηματικά από τον παρακάτω τύπο:

$$c(n) = \sum_{m=0}^{M-1} \log_{10} s(m) \cdot \cos\left(\frac{\pi \cdot n \cdot (m - 0.5)}{M}\right) \qquad (2.3.2.6)$$

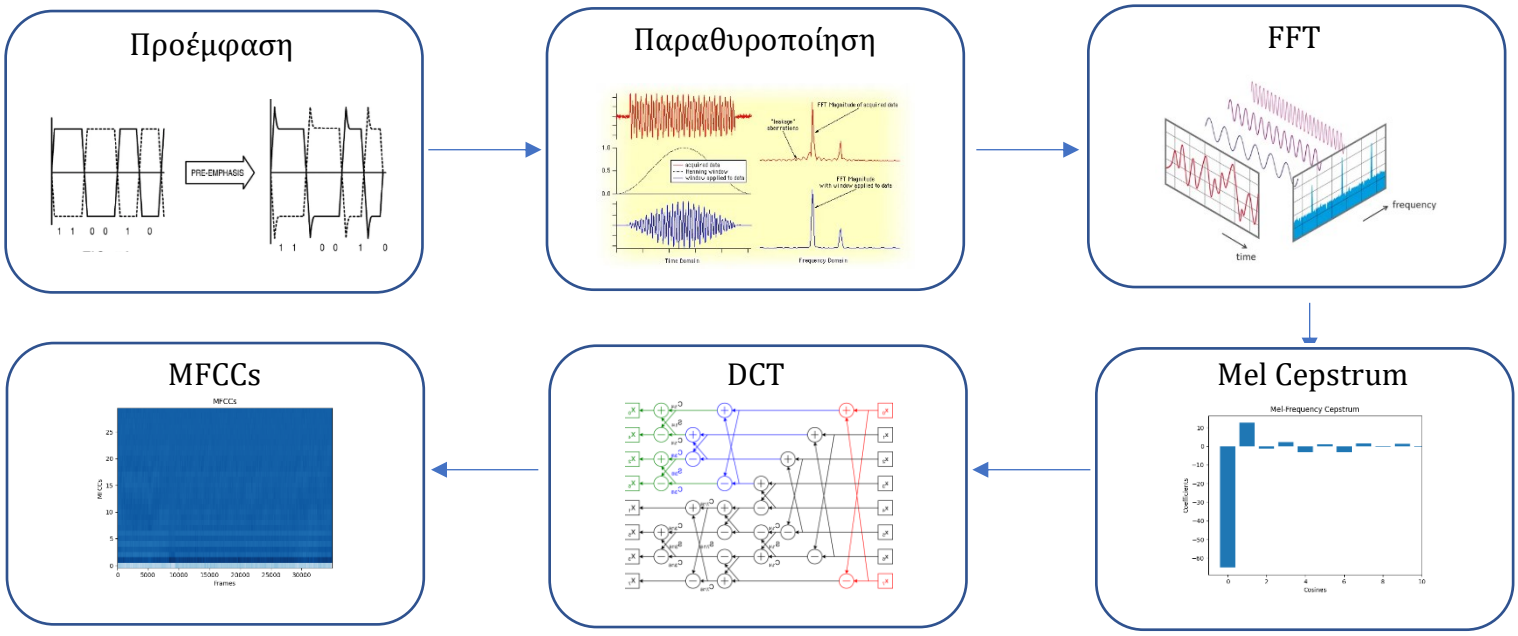
Όπου $c(n)$ η ν-οστή σταθερά MFC με το n να λαμβάνει ακέραιες τιμές από 0 έως $C - 1$, όπου C ο αριθμός των σταθερών που επιθυμούμε να εξάγουμε. Επειδή το μεγαλύτερο ποσοστό της πληροφορίας βρίσκεται εντός των αρχικών σταθερών MFC συνήθως επιλέγουμε να κρατήσουμε από 8 μέχρι 13 σταθερές αγνοώντας τη σταθερά που προκύπτει για $n = 0$ λόγω του χαμηλού ενεργειακού της περιεχομένου.

Μέσα από όλη αυτή τη διαδικασία έχουμε καταφέρει να εξάγουμε τις σταθερές που περιγράφουν το σήμα μας στο πεδίο 'cepstral'. Σε αυτό το πεδίο μπορούμε να λάβουμε την απάντηση στο εξής ερώτημα. Ποια θα είναι εκείνη η συχνότητα που θα θεωρήσει το ανθρώπινο αυτό ως τονικό κέντρο; Αυτό από τεχνικής άποψης δηλαδή το πλάτος μιας ζώνης συχνοτήτων στο πεδίο 'cepstral' μας δείχνει κατά πόσο οι συχνότητες εντός της ζώνης διήρκησαν κατά τη διάρκεια του κομματιού. Αυτό που στην πραγματικότητα μετράμε είναι η συχνότητα εμφάνισης ενός περιοδικού σήματος, δηλαδή μια μουσικής νότας, εντός του κομματιού. Το κομμάτι της Εικόνας 2 έχει ως κεντρική νότα, τη νότα Σι και κινείται εντός του πενταχόρδου Σι4 - Φα#5 όντας στο maqam Σι Huseyni. Στην παρακάτω εικόνα φαίνεται η αναπαράσταση του σήματος της Εικόνας 2 στο πεδίο 'cepstral'. Από αυτή την αναπαράσταση μπορούμε να εξάγουμε δύο πολύ σημαντικές πληροφορίες. Η πρώτη μας δίνεται από τη μέγιστη τιμή του διαγράμματος η οποία βρίσκεται στην περιοχή $88\ Hz - 166\ Hz$. Αυτή η ζώνη συχνοτήτων ανήκει στο κρουστό όργανο του κομματιού το οποίο είναι απολύτως λογικό να εμφανίζει τη μέγιστη συγκέντρωση συχνοτήτων λόγω των

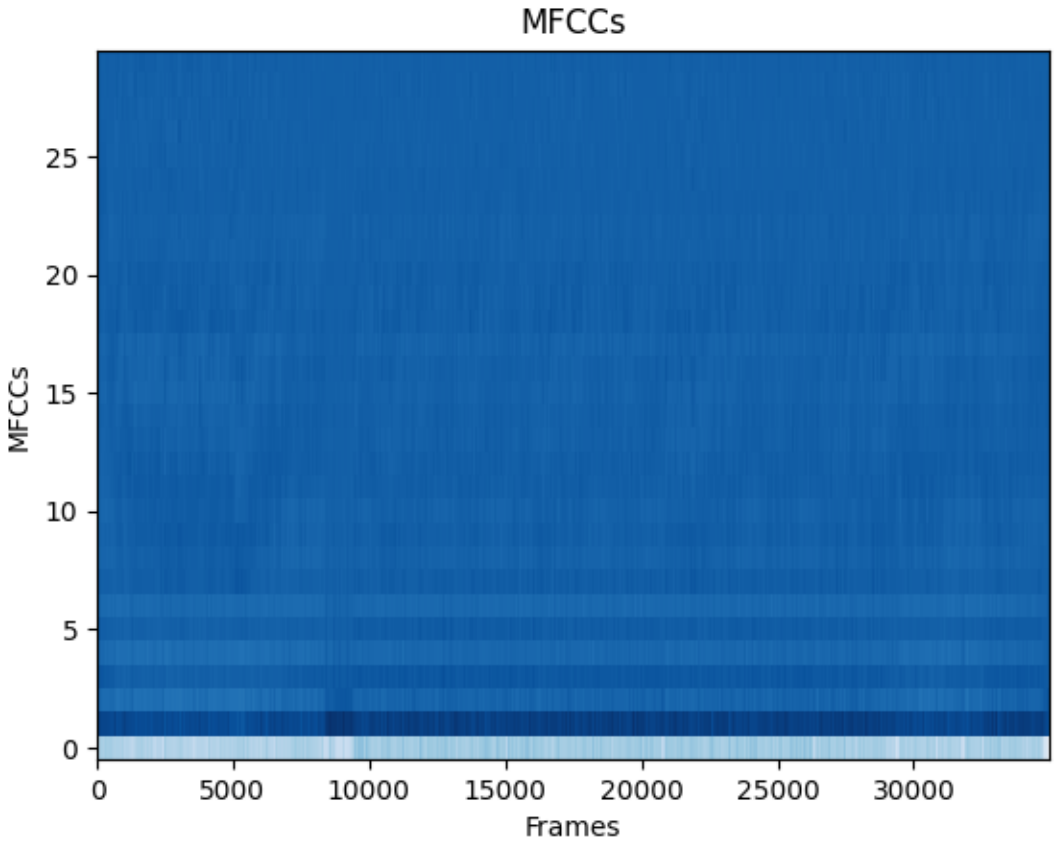
περιορισμένων αρμονικών του δυνατοτήτων. Για το νέυ, το μελωδικό όργανο του κομματιού βλέπουμε ότι το πλάτος των συχνοτήτων του μοιράζονται σε μεγάλο πεδίο του φάσματος καθώς είναι ένα όργανο με πολλές αρμονικές και τεράστιες αρμονικές δυνατότητες. Αυτό όμως που αξίζει να παρατηρήσουμε είναι το μεγάλο πλάτος που εμφανίζεται γύρω από τις βασικές νότες της μελωδίας Σι, Ρε και Φα#. Βάση αυτής της παρατήρησης μπορεί κάποιος να εξάγει την πιθανή πορεία του μέλους στο συγκεκριμένο κομμάτι.



Εικόνα 8: Αναπαράσταση των πρώτων 10 ζωνών του κομματιού της Εικόνας 2 σε 250 ζώνες του πεδίου cepstral όπου η κάθε ζώνη n καλύπτει την περιοχή συχνοτήτων $(\frac{n \cdot F_s}{250}, \frac{(n+1) \cdot F_s}{250})$.



Εικόνα 9: Διάγραμμα ροής του αλγορίθμου εξαγωγής των των σταθερών MFC.



Εικόνα 10: Αναπαράσταση των σταθερών MFC για ανάλυση 250 ζωνών στο πεδίο cepstral.

2.3.3 Τα ρυθμικά χαρακτηριστικά (tempo features)

Μια ακόμα πολύ ενδιαφέρουσα διάσταση από την οποία μπορεί κάποιος να εξάγει μουσική πληροφορία από ένα ηχητικό σήμα είναι αυτή του ρυθμού. Συγκεκριμένα στην εργασία αυτή θα ασχοληθούμε με την έννοια της ρυθμικής συνέπειας, δηλαδή αυτό που ονομάζουμε στη μουσική tempo ή χρονική αγωγή. Η διαδικασία αυτή επιτυγχάνεται με δύο στάδια τα οποία περιγράφουμε παρακάτω.

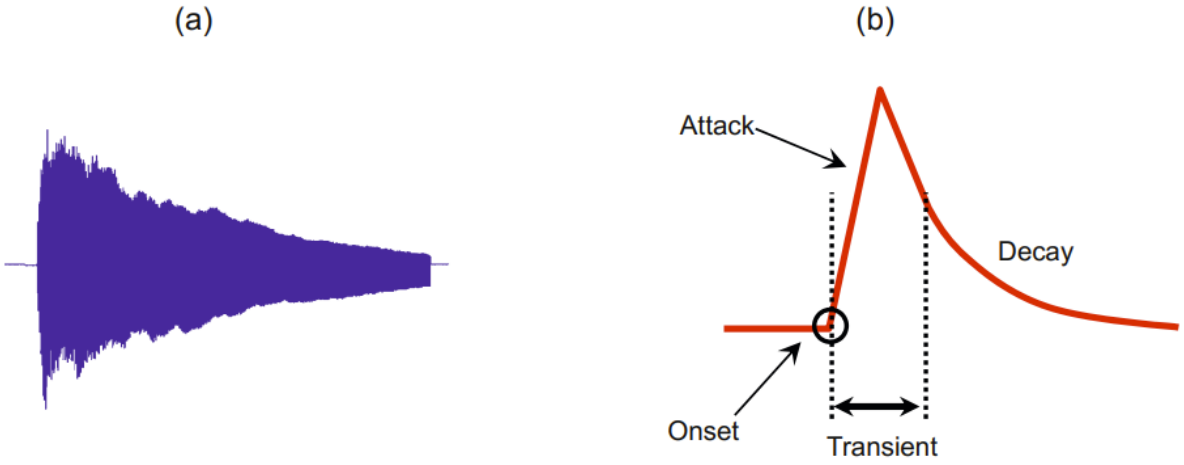
1. *Εντοπισμός έναρξης φθόγγων*

Σε αυτό το στάδιο ανάλυσης προσπαθούμε να εντοπίσουμε το σημείο εκκίνησης μια νότας, δηλαδή στην πράξη τη στιγμή που μια νότα παράγεται από τον οργανοπαίκτη ή τον τραγουδιστή. Για να κατανοήσουμε ποιο σημείο αναζητούμε στην πραγματικότητα αναλύουμε την κάθε νότα σε τρεις φάσεις οι οποίες φαίνονται και σχηματικά στην Εικόνα 11 :

i) Η φάση της επίθεσης (attack phase) κατά την οποία το μέσο παραγωγής της νότας πάλλεται έντονα δημιουργώντας ραγδαία ενεργειακή αύξηση στο σήμα, μέχρι να φτάσει κάποια στιγμή σε μια μέγιστη τιμή. Σε αυτή τη φάση βρίσκεται μια νότα που παράγεται για παράδειγμα από ένα νυκτό όργανο μόλις χτυπηθεί η χορδή από την πένα.

ii) Η φάση της απόσβεσης (decay phase) στην οποία εισέρχεται η νότα μόλις φτάσει στο μέγιστο πλάτος της. Από το μέγιστο και μετά η νότα χάνει σταδιακά την ενέργειά της καθώς αυτή διαχέεται ομοιόμορφα στο ακουστικό μέσο (π.χ. αέρας). Σε ένα τμήμα της φάσης απόσβεσης η νότα παραμένει αντιληπτή από τον ακροατή μέχρι να εξασθενίσει τόσο ώστε να μην μπορεί πλέον να εντοπιστεί από το ανθρώπινο αυτί.

iii) Η μεταβατική φάση (transient phase) η οποία περιγράφει τη διάρκεια ζωής μιας νότας. Θεωρούμε ότι η φάση αυτή ξεκινάει από την φάση της επίθεσης και τελειώνει μόλις η νότα πάψει να είναι αντιληπτή από τον ακροατή.



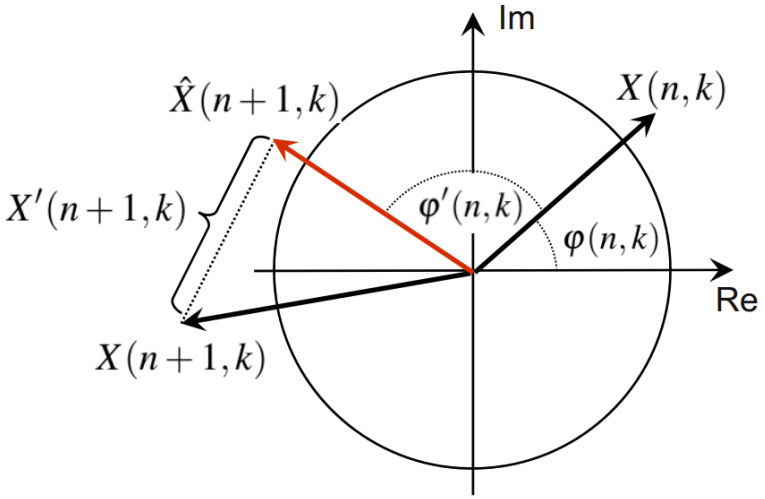
Εικόνα 11: α) Νότα στο πεδίο του χρόνου. β) Εξιδανικευμένη περιγραφή χρονικής εξέλιξης νότας στο πεδίο του χρόνου.

Από την παραπάνω εικόνα γίνεται αντιληπτό ότι το σημείο που αναζητούμε είναι στην πραγματικότητα η αρχή της μεταβατικής περιόδου, δηλαδή η αρχή της μεταβολής ενέργειας στη συγκεκριμένη συχνότητα. Με τον τρόπο αυτό μπορούμε να αναγνωρίσουμε τη χρονική στιγμή που 'παίχτηκε' μια νότα από τον οργανοπαίχτη ή αρθρώθηκε μια συλλαβή από τον τραγουδιστή με μεγάλη ακρίβεια.

Υπάρχουν πολλοί τρόποι για τον υπολογισμό του σημείου έναρξης, άλλοι βασίζονται στην ενεργειακή μεταβολή και άλλοι στη φασματική μεταβολή είτε αναλύοντας τη μεταβολή του πλάτους είτε της συχνότητας. Όλοι τους όμως προσπαθούν να εντοπίσουν το χαρακτηριστικό που ονομάζεται καινοτομία, το οποίο στη συγκεκριμένη περίπτωση είναι ο βαθμός μεταβολής της ηχητικής πίεσης κατά την εξέλιξή του σήματος στο χρόνο. Ο πιο σύγχρονος τρόπος για τον εντοπισμό της έναρξης μιας νότας αναλύει τη φασματική μεταβολή και ως προς το μέτρο αλλά και ως προς τη φάση μεταφέροντας έτσι το πρόβλημα στο μιγαδικό επίπεδο. Η κεντρική ιδέα είναι η στάθμιση μεταξύ της φασικής πληροφορίας και των σταθερών μέτρου. Η μέθοδος αυτή ονομάζεται 'Complex-Domain Novelty' (CD) και βασίζεται μαθηματικά στην εξαγωγή της συνάρτησης καινοτομίας (novelty function).

Έστω η σταθερά μετασχηματισμού Fourier $X(n,k)$ που αντιστοιχεί στο καρέ n. Από το γράφημα της Εικόνας 12 μπορούμε να εξάγουμε μια προσέγγιση για τη σταθερά μετασχηματισμού Fourier για το επόμενο καρέ την οποία και συμβολίζουμε με $\hat{X}(n+1,k)$.

$$\hat{X}(n+1,k) = |X(n,k)| \cdot \exp\left(2 \cdot \pi \cdot i\big(\varphi(n,k) + \varphi'(n,k)\big)\right) \qquad (2.3.3.1)$$



Εικόνα 12: Διανυσματικό διάγραμμα των μεγεθών της εξίσωσης 2.3.3.2 στο μιγαδικό επίπεδο.

Με βάση την παραπάνω σχέση μπορούμε να λάβουμε την τιμή της συνολικής καινοτομίας $X'(n,k+1)$ όπως φαίνεται παρακάτω.

$$X'(n+1,k) = |X'(n,k+1) - X(n,k+1)| \qquad (2.3.3.2)$$

Η τιμή της καινοτομίας εκφράζει το βαθμό της αστάθειας-μεταβολής για τη σταθερά k του καρέ n. Η τιμή της όμως υπό τη παρούσα μορφή δεν λαμβάνει υπόψιν της την αύξουσα η τη φθίνουσα πορεία της μεταβολής. Για το λόγο αυτό χωρίζουμε την καινοτομία σε ένα αύξον τμήμα $X^+(n,k)$ και σε ένα φθίνον τμήμα $X^-(n,k)$ όπως φαίνεται παρακάτω.

$$X^+(n,k)\begin{cases} X'(n,k), & \text{όπου } |X(n,k)| > |X(n-1,k)| \\ 0, & \text{αλλού} \end{cases} \qquad (2.3.3.3)$$

$$X^-(n,k)\begin{cases} X'(n,k), & \text{όπου } |X(n,k)| \leq |X(n-1,k)| \\ 0, & \text{αλλού} \end{cases} \qquad (2.3.3.4)$$

Αφού εμείς ενδιαφερόμαστε κατά την αναζήτηση του σημείου έναρξης φθόγγου για την αυξητική μεταβολή του σήματος, για να υπολογίσουμε τη συνολική τιμή της αύξουσας μεταβολής, ορίζουμε ως μιγαδικής συνάρτηση καινοτομίας $\Delta_{Complex}$ το παρακάτω άθροισμα:

$$\Delta_{Complex}(n,k) = \sum_{k=0}^{K} X^{+}(n,k) = CD(n) \quad (2.3.3.5)$$

Όπου K το σύνολο των σταθερών που προκύπτουν από το μετασχηματισμό Fourier του σήματος για το καρέ n. Εάν συνεπώς η τιμή της $CD(n)$ για κάποιο καρέ n υπερβεί ένα προκαθορισμένο κατώφλι t τότε θεωρούμε πως σε εκείνο το καρέ ξεκινά ένας φθόγγος.

Τα τελευταία χρόνια έχει προστεθεί στους αλγορίθμους εντοπισμού έναρξης νότας και μια νέα τεχνική η οποία αναπτύχθηκε για την αντιμετώπιση του vibrato καθώς ενώ κατά την εκτέλεσή του γίνεται εκπομπή ενέργειας σε κάθε νότα εμείς θα θέλαμε να το εκλαμβάνουμε ως μια νότα. Η τεχνική αυτή ονομάζεται SuperFlux (SF) [7] με τη μαθηματική της ανάλυση όμως να ξεφεύγει από τα πλαίσια της εργασίας αυτής. Χρησιμοποιώντας λοιπόν τις δύο τεχνικές CD και SF καταλήγουμε στην ακόλουθη συνάρτηση εντοπισμού του σημείου εκκίνησης ενός φθόγγου.

$$ODF(n) = a \cdot CD(n) + (1-a) \cdot SF(n) \quad (2.3.3.6)$$

Όπου n το καρέ υπό ανάλυση και όπου α η σταθερά γραμμικού συνδυασμού η οποία συνήθως λαμβάνει την τιμή 0.3.

2. Εξαγωγή τεμπογραφήματος

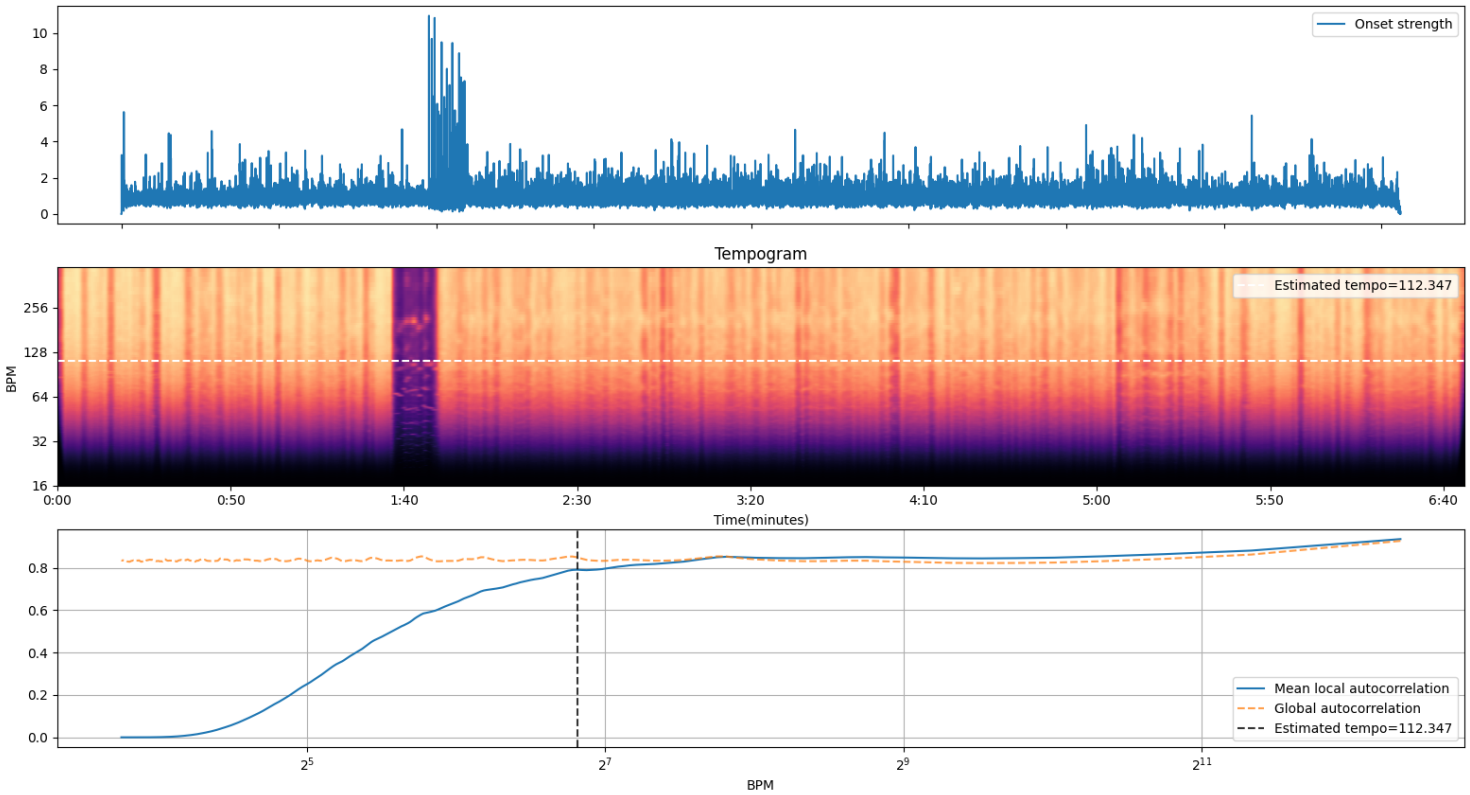
Το τεμπογράφημα είναι μια αναπαράσταση ενός ηχητικού σήματος βασιζόμενη σε χρονικούς παλμούς (καρέ) έτσι ώστε να καθίστανται φανερές οι μεταβολές στης χρονικής συνέπειας ως προς ένα τέμπο τ εκπεφρασμένο σε BPM. Η διαδικασία εξαγωγής του τεμπογραφήματος μπορεί να χωριστεί σε δύο τμήματα [7]. Το πρώτο είναι το στάδιο αυτό του εντοπισμού του χρονικού σημείου εκκίνησης του κάθε φθόγγου το οποίο αναλύσαμε λεπτομερώς στην προηγούμενη ενότητα. Στο δεύτερο στάδιο πραγματοποιείται προσέγγιση ενός μεγέθους που ονομάζουμε 'τοπικό τέμπο' το οποίο εκφράζει το τέμπο ως πυκνότητα φθόγγων για ένα συγκεκριμένο και αρκετά μικρό χρονικό παράθυρο χρησιμοποιώντας την αυτοσυσχέτηση της συνάρτηση εντοπισμού εκκίνησης φθόγγου (Onset Detection Function – ODF). Η διαδικασία αυτή βασίζεται στο γεγονός ότι ένα μουσικό σήμα εμφανίζει συμπαγή και τοπικώς περιοδικά χρονικά μοτίβα (αυτό που μουσικά ονομάζουμε μέτρα). Αυτά τα μοτίβα θα εμφανίζουν κορυφές στην συνάρτηση αυτοσυσχέτισης με χρονική καθυστέρηση l. Η μαθηματική έκφραση της τοπικής αυτοσυσχέτισης ενός μουσικού σήματος με τη χρήση ορθογωνικών παραθύρων W φαίνεται παρακάτω.

$$A(t,l) = \sum_{n \in Z} \frac{ODF(n) \cdot ODF(n+l) \cdot W(n-t)}{2 \cdot N + 1 - l} \quad (2.3.3.7)$$

Όπου ο χρόνος $t \in Z$, η χρονοκαθυστέρηση $l \in [0: N]$ και το τοπικό τέμπο τ δίνεται ως $\tau = \frac{60}{r \cdot l}$. Η σταθερά r εισάγεται για να βελτιώσει την ανάλυση στον άξονα του χρόνου έπειτα από επαναδειγματοληψία.

Το τεμπογράφημα για το μουσικό σήμα της Εικόνας 2 φαίνεται στην Εικόνα 13. Ιδιαίτερο ενδιαφέρον αποτελεί η σκοτεινή περιοχή στο δεύτερο υποδιάγραμμα που ξεκινάει λίγο πριν τη χρονική στιγμή 1:40. Αυτό το διάστημα ανήκει στα μέτρα που δίνει το κρουστό και μας δείχνει ότι υπάρχει σχεδόν μηδαμινή ρυθμική απόκλιση. Επίσης στο τρίτο υποδιάγραμμα βλέπουμε πως στην αρχή του κομματιού όπου πραγματοποιείται αυτοσχεδιαστική εισαγωγή δεν υπάρχει σταθερό τέμπο με την πυκνότητα των

φθόγγων όμως να αυξάνει όσο προχωράει ο αυτοσχεδιασμός. Τέλος από το ίδιο υποδιάγραμμα βλέπουμε πως η μέση αυτοσυσχέτιση ως προς το μέσο τέμπο του κομματιού είναι λίγο πάνω από 0.8 , συνεπώς μπορούμε να συμπεράνουμε πως το σύνολο ήταν πάνω από 80% συνεπές ως προς το μέσο τέμπο που είχε στο κομμάτι.



Εικόνα 13: α) Διάγραμμα καθαρότητας εκκίνησης φθόγγων του κομματιού της Εικόνας 2. β) Τεμπογράφημα του κομματιού της Εικόνας 2. γ) Τιμή συνάρτησης αυτοσυσχέτισης κατά τη διάρκεια του κομματιού της Εικόνας 2.

2.4 Πίνακας αυτο-ομοιότητας (Self-Similarity Matrix)

2.4.1 Βασικές έννοιες

Έστω ένας χώρος εξαχθέντος χαρακτηριστικού F που περιέχει όλα τα στοιχεία μιας ακολουθίας $X = (x1, x2, .., x_n)$ με τα x_i να είναι οι τιμές που λαμβάνουμε από έναν αλγόριθμο εξαγωγής μουσικού χαρακτηριστικού, όπως αυτοί που εξετάστηκαν παραπάνω. Το μέτρο ομοιότητας που δίνει τη δυνατότητα να συγκρίνουμε μεταξύ τους τα στοιχεία του χώρου F ορίζεται ως:

$$s : F \times F \rightarrow R \qquad (2.4.1.1)$$

Τυπικά η τιμή του μέτρου ομοιότητας για δύο στοιχεία $x, y \in F$ το οποίο συμβολίζεται ως $s(x, y)$ είναι μεγάλη εάν τα στοιχεία εμφανίζουν ομοιότητα και μικρή εάν δεν εμφανίζουν. Η βασική ιδέα είναι να συγκρίνουμε κάθε στοιχείο της ακολουθίας X με όλα τα υπόλοιπα για να εντοπίσουμε όμοια τμήματα

εντός του μουσικού κομματιού. Το αποτέλεσμα αυτής της διαδικασίας θα είναι ένας πίνακας αυτο-ομοιότητας $S \in R^{N*N}$ ορισμένος από τη σχέση:

$$S(n,m) \coloneqq s(x_n, x_m) \tag{2.4.1.2}$$

Όπου $x_n, x_m \in F$, $n,m \in [1:N]$ και $(n,m) \in [1:N] \times [1:N]$. Κάθε πλειάδα (n,m) ονομάζεται κύτταρο (cell) τιμή $S(n,m)$ ονομάζεται τιμή (score) του πίνακα S.

Υπάρχουν αρκετοί τρόποι ορισμού του μέτρου ομοιότητας με πιο κοινό τον ορισμό του ως το μέτρο του εσωτερικού γινομένου μεταξύ δύο διανυσμάτων. Υποθέτοντας λοιπόν Ευκλείδειο χώρο $F = R^D$, $D \in N$ η τιμή του μέτρου ομοιότητας για το κύτταρο (n,m) του πίνακα αυτό-ομοιότητας θα δίνεται από τη σχέση:

$$s(x,y) \coloneqq |\langle x|y\rangle| \tag{2.4.1.3}$$

Η παραπάνω σχέση πληροί την ιδιότητα της μέγιστης ταύτισης του κάθε στοιχείου ως προς τον εαυτό του εάν τα διανύσματα x και y είναι κανονικοποιημένα. Ως αποτέλεσμα η ομοιότητα κρύβεται στη γωνία μεταξύ των δύο διανυσμάτων x και y εντός του διανυσματικού χώρου F. Εάν η γωνία είναι ορθή τότε η ταύτιση είναι μηδενική και το μέτρο ομοιότητας θα έχει τιμή 0. Εάν τα διανύσματα x και y είναι κανονικοποιημένα και η γωνία μεταξύ τους είναι μηδενική τότε το μέτρο ομοιότητας θα λάβει τη μέγιστη τιμή του η οποία είναι 1. Ως αποτέλεσμα πάντοτε οι πίνακες αυτό-ομοιότητας εμφανίζουν μια διαγώνιο τιμής 1 η οποία οφείλεται στην ομοιότητα του κάθε στοιχείου ως προς τον εαυτό του.

Για μαθηματική περιγραφή της μουσικής πληροφορίας του πίνακα αυτό-ομοιότητας ακολουθούμε την εξής λογική. Έστω X η ακολουθία τιμών ενός μουσικού χαρακτηριστικού και S ο πίνακας αυτό-ομοιότητας που προκύπτει από αυτό. Ένα τμήμα a του πίνακα ορίζεται ως ένα σύνολο $a = [s:t] \subseteq [1:N]$ με τα s και t να ορίζουν την χρονική στιγμή έναρξης και λήξης του τμήματος. Ένα τέτοιο διάστημα αναπαριστά ένα χρονικό διάστημα επί του άξονα x στη γραφική παράσταση του πίνακα αυτό-ομοιότητας. Βάση του παραπάνω ορισμού το μήκος αυτού τμήματος a θα είναι $|\alpha| \coloneqq t - s + 1$.

Έπειτα επί της γραφικής παράστασης του πίνακα αυτό-ομοιότητας φέρνουμε δύο κατακόρυφες και παράλληλες ευθείες οι οποίες διέρχονται από τα άκρα του τμήματος a. Η περιοχή που περικλείουν οι δύο παράλληλες είναι η περιοχή συσχέτισης του a με το υπόλοιπο κομμάτι. Για κάθε επαγόμενο τμήμα $a' = [s':t']$ που μπορεί να οριστεί επί του άξονα y δημιουργείται μια παραλληλόγραμμη περιοχή με επιφάνεια $|\alpha| \times |\alpha'|$ εντός της οποίας φαίνεται η συσχέτιση των τμημάτων a και a'. Ανάλογα λοιπόν με τον τύπο συσχέτισης των δύο διαστημάτων εμφανίζονται διάφορες γεωμετρικές δομές εντός του παραλληλογράμμου. Οι δύο σημαντικότερες εξ αυτών, τις οποίες και θα αναλύσουμε, είναι οι συμπαγείς δομές (blocks) και οι δομές μονοπατιού (paths).

2.4.2 Συμπαγείς (block) δομές

Εάν το χαρακτηριστικό που εξαγάγαμε λαμβάνει σταθερή τιμή για ένα χρονικό διάστημα όλα τα διανύσματα x και y του διαστήματος αυτού θα εμφανίζουν μεγάλο μέτρο ομοιότητας μεταξύ τους. Ως αποτέλεσμα στον πίνακα αυτό-ομοιότητας εμφανίζονται συμπαγείς παραλληλόγραμμες δομές με υψηλή τιμή ομοιότητας. Μέσω αυτών των συμπαγών δομών μπορούμε λοιπόν να εντοπίσουμε την ύπαρξη μουσικής ομοιογένειας σε ένα μουσικό τμήμα.

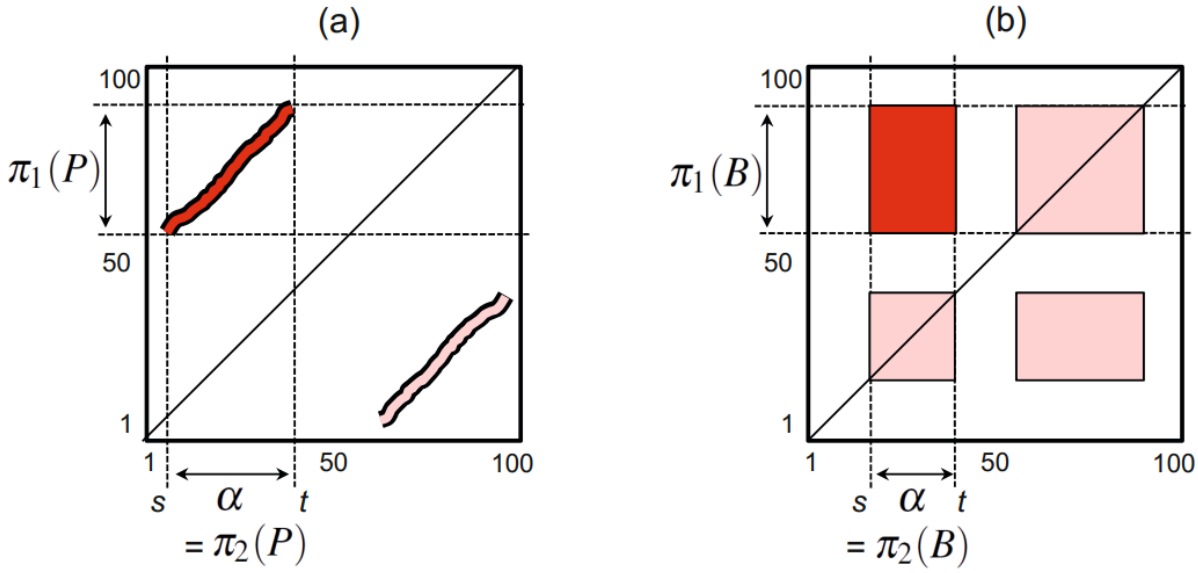
Ως συμπαγή δομή επί ενός μουσικού τμήματος $a = [s:t]$ του άξονα x ορίζουμε τη γεωμετρική δομή:

$$B = \alpha' \times \alpha \subseteq [1:N] \times [1:N] \tag{2.4.2.1}$$

όπου το $\alpha' = [s':t']$ είναι το επαγόμενο μουσικό τμήμα με το οποίο συσχετίζουμε το τμήμα α.

Τέλος ορίζουμε την τιμή (score) της συμπαγούς δομής μέσω της σχέσης:

$$\sigma(B) = \sum_{(n,m)\in B} S(n,m) \tag{2.4.2.2}$$



Εικόνα 14: (α) Εξιδανικευμένη αναπαράσταση δομής μονοπατιού. (β) Εξιδανικευμένη αναπαράσταση συμπαγούς δομής.

2.4.3 Δομές μονοπατιού (path)

Εάν δύο διαφορετικά τμήματα α και α'ενός μουσικού κομματιού είναι όμοια ως προς τη μουσική τους εξέλιξη με το δεύτερο να αποτελεί επανάληψη του πρώτου αυτό σημαίνει πως τα δύο μουσικά αυτά τμήματα θα φαίνεται να εξελίσσονται παράλληλα εντός του παραλληλογράμμου συσχέτισής τους. Αυτό οδηγεί στην εμφάνιση μιας διαγωνίου παράλληλης με την κύρια διαγώνιο εντός του παραλληλογράμμου συσχέτισης η οποία εμφανίζει υψηλές τιμές ομοιότητας. Μέσω αυτών των διαγώνιων οι οποίες ονομάζονται και δομές μονοπατιού μπορούμε λοιπόν να εντοπίσουμε τα χαρακτηριστικά επανάληψης ενός μουσικού κομματιού.

Ως δομή μονοπατιού μήκους L επί ενός μουσικού τμήματος $\alpha = [s:t]$ του άξονα x ορίζουμε τη γεωμετρική ακολουθία:

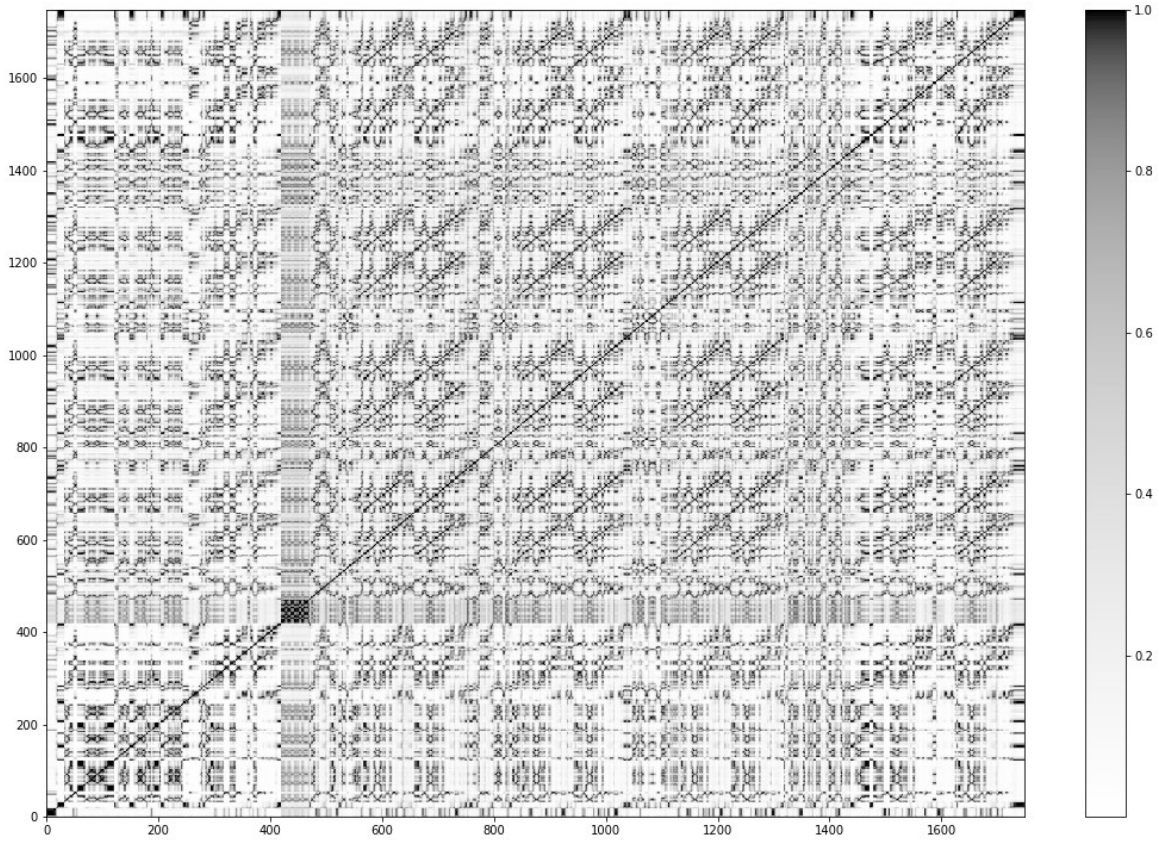
$$P = ((n_1, m_1), \dots, (n_L, m_L)) \tag{2.4.3.1}$$

Όπου τα κύτταρα $(n_l, m_l) \in [1: N]^2$και $l \in [1: L]$. Επιπλέον λόγω οριακών συνθηκών ισχύει $m_1 = s$, $m_L = t$ ενώ επίσης ισχύει ότι $(n_{l+1}, m_{l+1}) - (n_l, m_l) \in \Sigma$, όπου Σ είναι το σύνολο των αποδεκτών βημάτων επί του πίνακα αυτό-ομοιότητας. Συνήθως χρησιμοποιούμε ως Σ το σύνολο $\{(2,1), (1,2), (1,1)\}$.

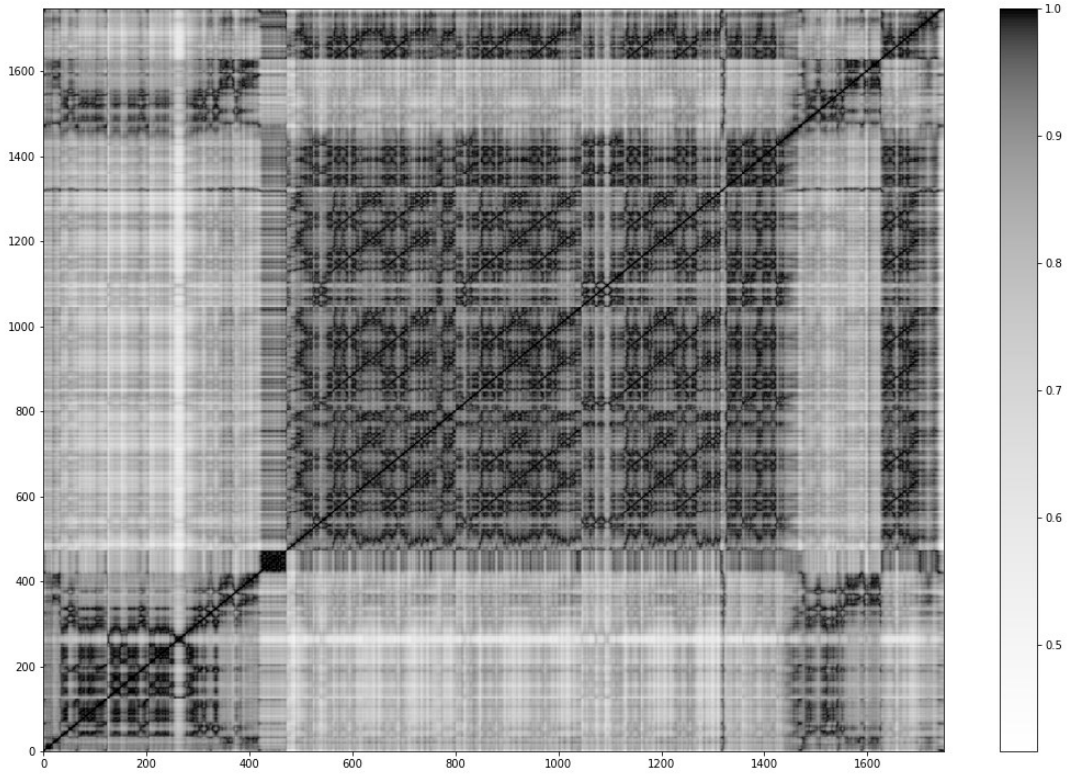
Τέλος ορίζουμε την τιμή (score) της δομής μονοπατιού μέσω της σχέσης:

$$\sigma(B) = \sum_{l=1}^{L} S(n_l, m_l) \tag{2.4.3.2}$$

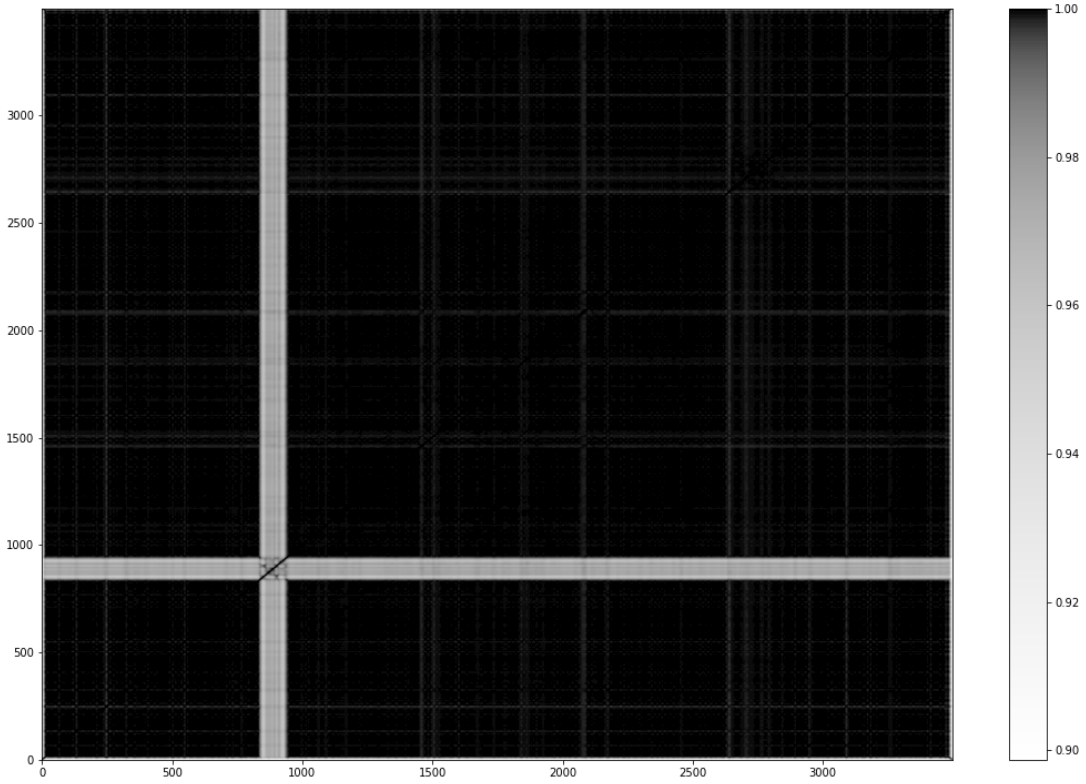
Παρακάτω φαίνονται οι τρεις πίνακες αυτό-ομοιότητας που προέκυψαν βάση των τριών χαρακτηριστικών που εξαγάγαμε στην προηγούμενη ενότητα. Ο πίνακας που προέκυψε από τα χρωματικά χαρακτηριστικά έχει την περισσότερη ποικιλία καθώς λαμβάνει υπόψιν του το μεγαλύτερο εύρος μεταβολών. Τη μεγαλύτερη ακρίβεια εμφανίζει ο πίνακας που προέκυψε από τις MFCCs στον οποίο και φαίνονται πολύ καθαρά οι επαναληπτικές δομές αλλά και οι ομοιογενείς ομάδες. Τέλος, ο πίνακας που προέκυψε από το τεμπογράφημα παρουσιάζει στην πράξη μόνο συμπαγείς δομές. Αυτό που μπορεί εύκολα να διακριθεί είναι ότι το τοπικό τέμπο 'μονοπωλείται' από το νέυ όποτε αυτό είναι παρόν και πως τα μέτρα που δίνει το κρουστό ταυτίζονται στην πράξη μόνο με τα σημεία όπου εμφανίζονται ανοιχτόχρωμες κατακόρυφες γραμμές. Βλέπουμε όμως ότι η διαφορά είναι της τάξης του 0,1 και άρα συνεπώς η ηχογράφηση έχει αρκετά σταθερή χρονική αγωγή ως προς τα μέτρα του κρουστού και άρα ως προς την απόλυτη ρυθμική αγωγή του κομματιού.



Εικόνα 15: Πίνακας αυτο-ομοιότητας βάση των χρωματικών χαρακτηριστικών.



Εικόνα 16: Πίνακας αυτο-ομοιότητας βάση των προσαρμοσμένων φασματικών χαρακτηριστικών.



Εικόνα 17: Πίνακας αυτο-ομοιότητας βάση των ρυθμικών χαρακτηριστικών.

2.5 Εφαρμογή στην εξαγωγή χαρακτηριστικού μουσικού τμήματος (thumbnailing)

Μια πολύ ενδιαφέρουσα εφαρμογή της εξαγωγής μουσικών χαρακτηριστικών και του πίνακα αυτο-ομοιότητας είναι η αυτόματη εξαγωγή του χαρακτηριστικού μουσικού τμήματος ενός κομματιού. Συνήθως το τμήμα αυτό είναι το ρεφρέν και είναι το μουσικό τμήμα που έχει το μεγαλύτερο αντίκτυπο στον ακροατή. Έστω λοιπόν ότι έχουμε μια τεράστια βάση μουσικών δεδομένων και μια εφαρμογή διανομής ροής (streaming) η οποία αποστέλλει στους χρήστες τη μουσική αυτή προς ακρόαση. Έστω επίσης πως θέλουμε να εισάγουμε μια καινούρια δυνατότητα στην εφαρμογή η οποία θα δίνει τη δυνατότητα στο χρήστη να ακούει το χαρακτηριστικό τμήμα ενός κομματιού πριν επιλέξει να το ακούσει ολόκληρο. Σε ένα τέτοιο σενάριο οι επιλογές είναι δύο. Στην πρώτη θα πρέπει να αναθέσουμε σε έναν άνθρωπο να ακούσει κάθε ένα από τα κομμάτια της βάσης δεδομένων, να αποφασίσει υποκειμενικά ποιο τμήμα του κομματιού το χαρακτηρίζει καλύτερα και να εντοπίσει το χρονικό διάστημα στο οποίο το τμήμα αυτό εμφανίζεται ώστε να μπορεί να αποκοπεί και να δοθεί στο χρήστη ως χαρακτηριστικό τμήμα. Είναι φανερό πως για μια βάση πολλών χιλιάδων κομματιών η ολοκλήρωση αυτής της διαδικασίας μπορεί να πάρει αρκετούς μήνες, αν όχι χρόνια. Στη δεύτερη επιλογή μπορούμε να εξάγουμε αυτόματα το χαρακτηριστικό τμήμα του κάθε κομματιού, θεωρώντας ότι είναι το τμήμα που εμφανίζεται με τη μεγαλύτερη συχνότητας στο κομμάτι, ακολουθώντας την αλγοριθμική διαδικασία που περιγράψαμε στην παράγραφο 2.3. Κατά τη διαδικασία αυτή εκμεταλλευόμαστε τις δομές μονοπατιού του πίνακα αυτό-ομοιότητας και εισάγουμε στο τέλος της αλγοριθμικής διαδικασίας ένα επιπλέον στάδιο ομαδοποίησης (clustering) κοινών μονοπατιών (οικογένειας μονοπατιών) [9] και υπολογισμού του μέτρου ταύτισής τους [10]. Όποια από τις δομές μονοπατιών μεγιστοποιήσει το μέτρο ταύτισης θεωρείται ως χαρακτηριστικό τμήμα του κομματιού ενώ η διαδικασία αυτή διαρκεί μόλις λίγα δευτερόλεπτα. Παρακάτω φαίνεται η έξοδος ενός προγράμματος υπολογισμού του χαρακτηριστικού τμήματος του κομματιού της Εικόνας 2. Το πρόγραμμα λειτουργεί με τη χρήση χρωματικών χαρακτηριστικών και δέχεται ως είσοδο την επιθυμητή διάρκεια του χαρακτηριστικού τμήματος. Με τον τρόπο αυτό μπορούμε να επιλέξουμε ακριβώς πόσο θα διαρκεί το χαρακτηριστικό τμήμα που επιθυμούμε να χρησιμοποιήσουμε στην εφαρμογή μας. Τυπικά τα χαρακτηριστικά μουσικά τμήματα έχουν διάρκεια λίγων δευτερολέπτων. Δίνοντας λοιπόν ως επιθυμητή διάρκεια τα 20 δευτερόλεπτα το πρόγραμμα επιτυχώς επέλεξε το τμήμα της Υπακοής το οποίο επαναλαμβάνεται 5 φορές εντός του κομματιού.

```
Thumbnail init: 153.23419329999615 with: 0.47457710937759956 of fitness value.
The best thumbnail for this song with length 19.97 starts at time: 153.23s
```

Εικόνα 18: Έξοδος αλγορίθμου εντοπισμού χαρακτηριστικού μουσικού τμήματος.

3. Συμπεράσματα

Στο σημείο αυτό έχει καταστεί πλέον φανερή η θεωρητική αλλά και πρακτική ισχύς της σημασιολογικής κωδικοποίησης των ηχητικών δεδομένων αλλά και των αλγορίθμων αξιοποίησής της. Σημειώνεται ότι αυτή η διαδικασία θα έπρεπε εναλλακτικά να πραγματοποιηθεί από ανθρώπινους ακροατές οι οποίοι θα έπρεπε να ακούσουν ένα μουσικό κομμάτι, πιθανώς αρκετές φορές, να αποφανθούν για τμήμα του κομματιού που θεωρούν ως χαρακτηριστικό και να εντοπίσουν τις χρονικές στιγμές έναρξης και λήξης του. Η διαδικασία αυτή θα ήταν και χρονοβόρα αλλά και αρκετά δύσκολη λόγω της ανθρώπινης υποκειμενικότητας (συναισθηματική κατάσταση του ακροατή, βαθμός μουσικής αντίληψης κ.α.). Έτσι

λοιπόν μέσω της αυτοματοποίησης μπορούμε να λάβουμε πολύ πιο σύντομα και χωρίς κάποια ανθρώπινη παρέμβαση ακριβή και αντικειμενικά μεταδεδομένα τα οποία απαιτούν μόνο το ίδιο το ηχητικό σήμα για τη δημιουργία τους. Η διαδικασία αυτή, όπως και κάθε διαδικασία υπολογιστικής αυτοματοποίησης, θα γίνεται επιπλέον ακόμη πιο ισχυρή με την αύξηση της διαθέσιμης υπολογιστικής ισχύος και με τη βελτίωση των αλγορίθμων που χρησιμοποιούνται. Αξίζει να σημειωθεί ότι και στην αλγοριθμική προσέγγιση αλλά και κατά τον υπολογισμό του πίνακα αυτο-ομοιότητας μπορεί να υπάρξει αρκετή ποικιλία με αρκετά βήματα βελτιστοποίησης να προστίθενται σε εφαρμογές που προορίζονται για την αγορά. Τέλος, αξίζει να παρατηρήσουμε πως από κάθε εξαγόμενο μουσικό χαρακτηριστικό μπορούμε να αντλήσουμε συγκεκριμένα και τις περισσότερες φορές πολύ εξειδικευμένα συμπεράσματα. Για το λόγο αυτό πρέπει να χρησιμοποιείται ένα σύνολο χαρακτηριστικών κατά τη διάρκεια της σημασιολογικής ανάλυσης των ηχητικών δεδομένων εάν θέλουμε τα συμπεράσματα στα οποία θα καταλήξουμε να ανταποκρίνονται στα δεδομένα ενός πραγματικού ακροατή. Εάν ο ανθρώπινος εγκέφαλος υπερτερεί κάπου σε σχέση με την αυτοματοποιημένη παραγωγή των ηχητικών δεδομένων από συστήματα σημασιολογικής κωδικοποίησης είναι στη δυνατότητά του να συνδυάζει αβίαστα ολόκληρο το σύνολο των μουσικών διαστάσεων (ρυθμού, μελωδίας, χροιάς, χρονικής αγωγής κλπ.) διαμορφώνοντας μια ολιστική αντίληψη για το άκουσμα. Αυτή ακριβώς η συνολική αντίληψη του ανθρώπου για τη μουσική αποτελεί την πρόκληση των αυτόματων συστημάτων σημασιολογικής κωδικοποίησης τα οποία εξετάζουν πάντα την κάθε διάσταση-χαρακτηριστικό μεμονωμένα. Η διαδικασία συνδυασμού της αυτόματα παραγόμενης μουσικής πληροφορίας αλλά και η ανακάλυψη νέων μουσικών διαστάσεων οι οποίες θα αναδεικνύουν κρυμμένες ιδιότητες των ηχητικών σημάτων που υποσυνείδητα επεξεργάζεται ο ανθρώπινος εγκέφαλος είναι τα επόμενα κομβικά σημεία στην εξέλιξη αυτής της τεχνολογίας

4. Βιβλιογραφία

[1] https://www.cambridgeaudio.com/usa/en/blog/metadata-digital-audio-files-%E2%80%93-what-it-where-it-and-how-tidy-it

[2] https://www.lifewire.com/importance-of-song-tags-in-music-files-2438758

[3]https://ecommons.luc.edu/cgi/viewcontent.cgi?article=1001&context=lib_facpubs

[4] Meinard Müller, Fundamentals of Music Processing Meinard Müller Audio, Analysis, Algorithms, Applications, Springer 2015.

[5] https://www.youtube.com/watch?v=7JIH-_2Q0Os

[6] K.S. Rao and Manjunath K.E., Speech Recognition Using Articulatory and Excitation Source Features, SpringerBriefs in Speech Technology, Appendix A, DOI 10.1007/978-3-319-49220-9

[7] Mi Tian, Gyorgy Fazekas, Dawn A. A. Black, Mark Sandler, On the use of the tempogram to describe audio content and its application to music structural segmentation, Centre for Digital Music, Queen Mary University of London E1 4NS, London, UK

[8] Sebastian Böck and Gerhard Widmer, Maximum filter vibrato suppression for onset detection, Department of Computational Perception Johannes Kepler University Linz, Austria.

[9] Meinard Müller, Fundamentals of Music Processing Meinard Müller Audio, Analysis, Algorithms, Applications, Springer 2015, pp. 197.

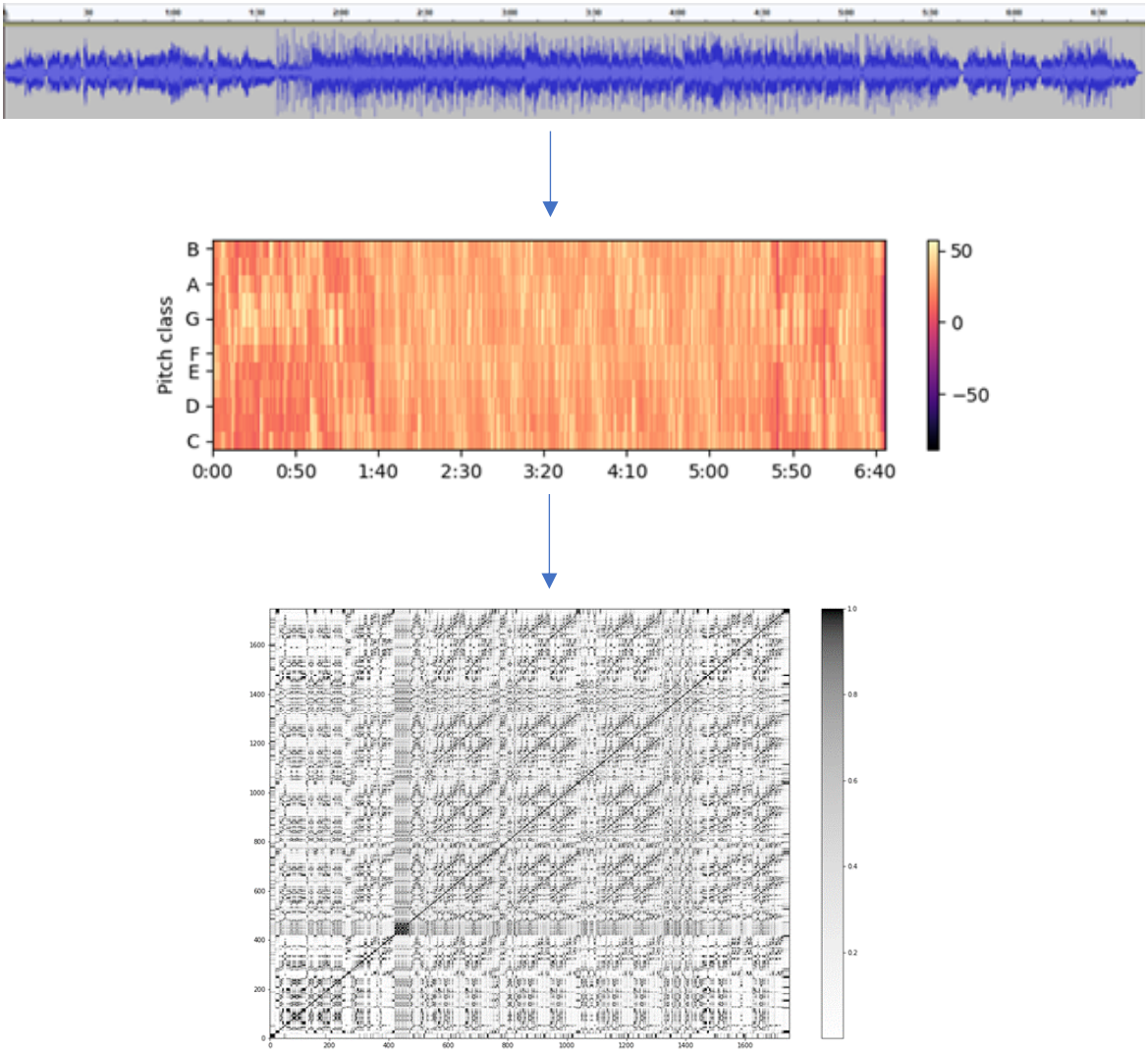
[10] Meinard Müller, Fundamentals of Music Processing Meinard Müller Audio, Analysis, Algorithms, Applications, Springer 2015, pp. 200-202.

5. Παραρτήματα

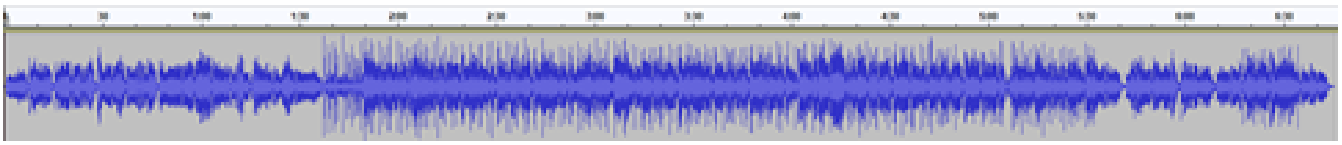
5.1 – Δομικά διαγράμματα εξαγωγής των πινάκων αυτό-ομοιότητας.

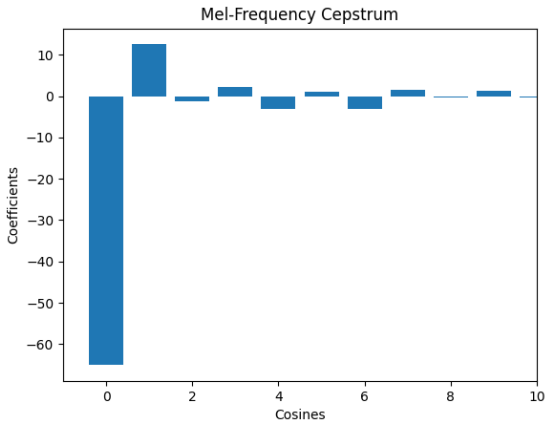
Στο σημείο αυτό παραθέτουμε σε αναλυτικό δομικό διάγραμμα τα βήματα εξαγωγής των πινάκων αυτό-ομοιότητας που φαίνονται στις εικόνες 15, 16 και 17.

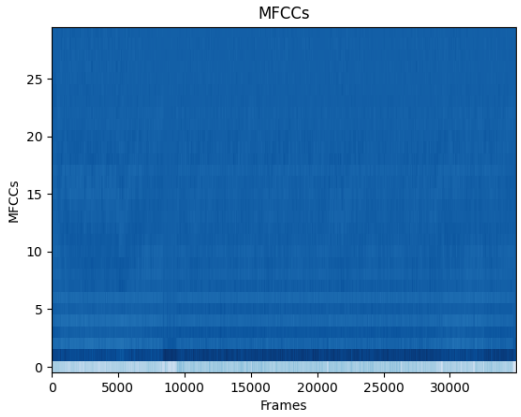
5.1.1 Εξαγωγή πίνακας αυτό-ομοιότητας χρωματικών χαρακτηριστικών

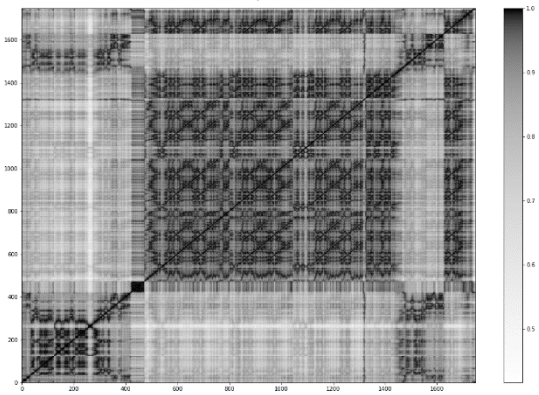


5.1.2 Εξαγωγή πίνακας αυτό-ομοιότητας φασματικών χαρακτηριστικών

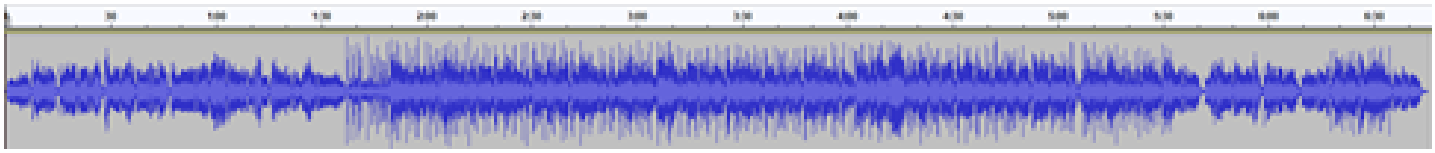


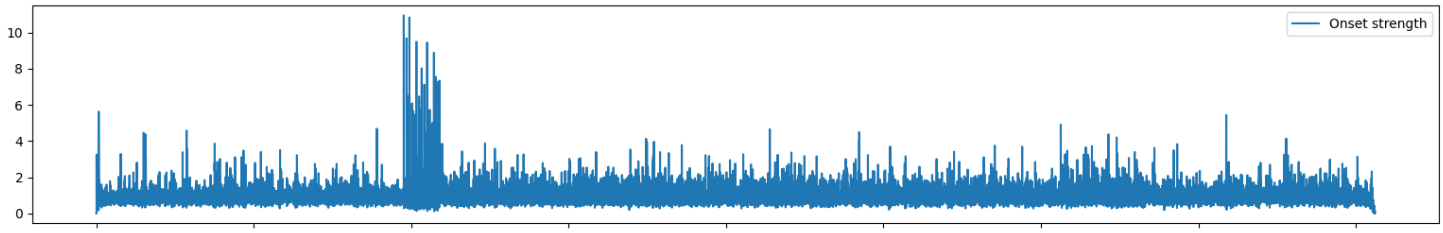


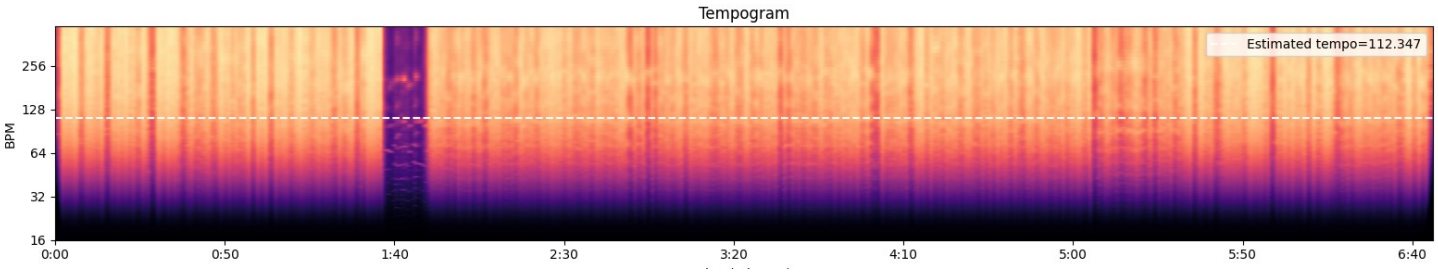


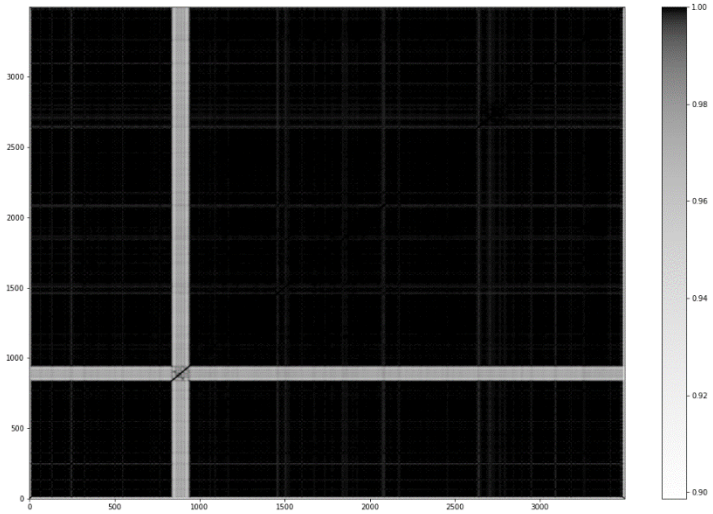


5.1.3 Εξαγωγή πίνακας αυτό-ομοιότητας ρυθμικών χαρακτηριστικών





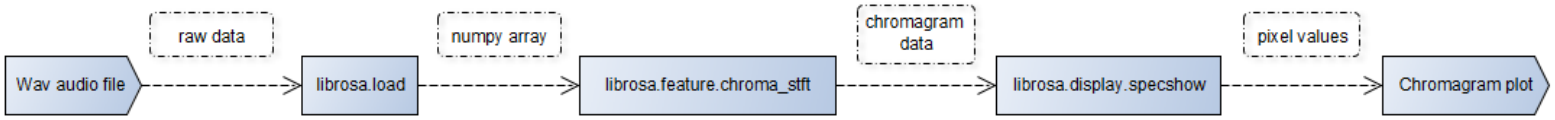




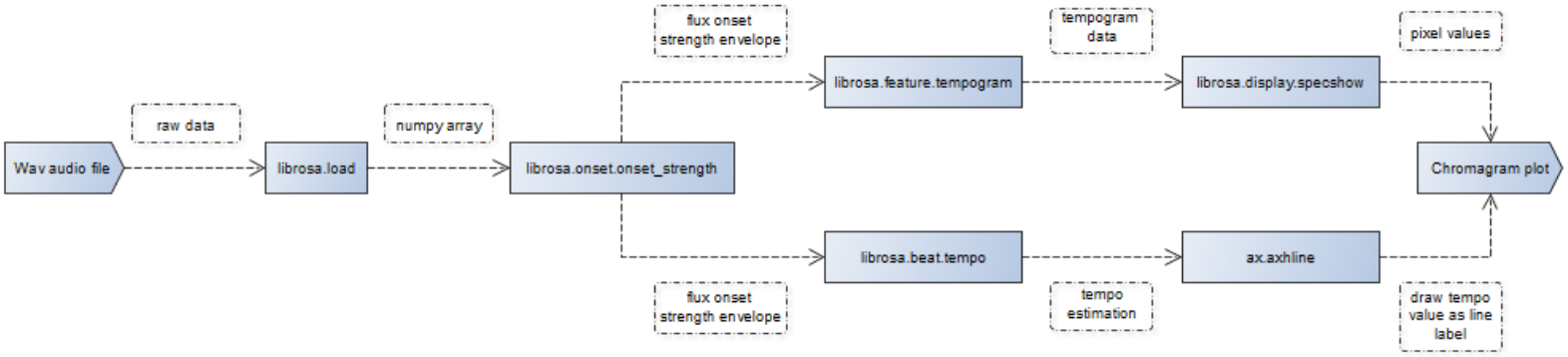
5.2 Διαγράμματα ροής για τους κώδικες εξαγωγής χαρακτηριστικών και δημιουργίας των πινάκων αυτο-ομοιότητας.

Στο σημείο αυτό παραθέτουμε τα διαγράμματα ροής για τους κώδικες εξαγωγής χαρακτηριστικών και υπολογισμού των πινάκων αυτό-ομοιότητας.

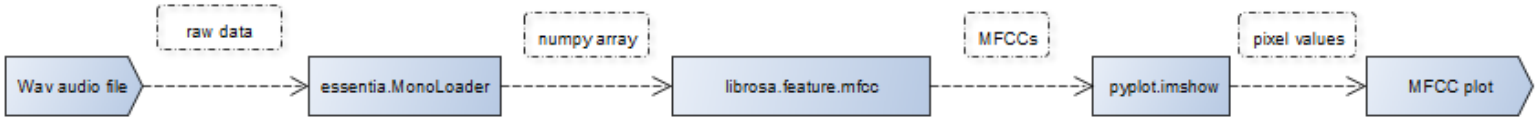
5.2.1 Κώδικας εξαγωγής χρωματογραφήματος



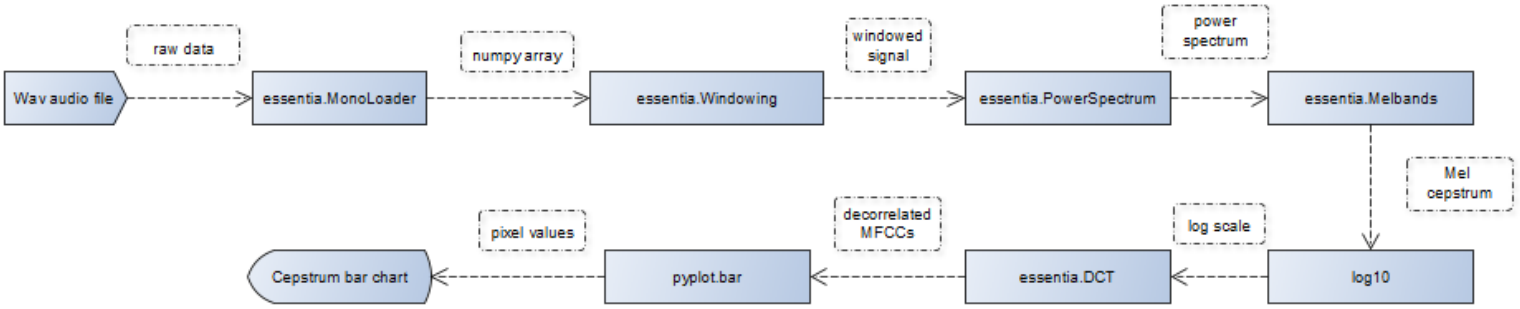
5.2.2 Κώδικας εξαγωγής τεμπογραφήματος και γραφημάτων εντοπισμού έναρξης φθόγγων



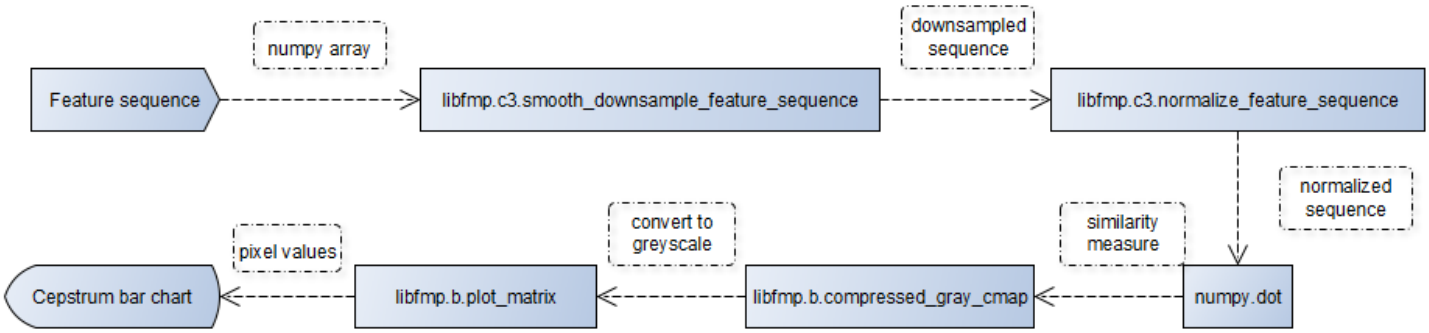
5.2.3 Κώδικας εξαγωγής σταθερών MFCC



5.2.4 Κώδικας εξαγωγής φάσματος cepstrum



5.2.5 Κώδικας εξαγωγή πίνακα αυτό-ομοιότητας από χαρακτηριστικό



Ο κώδικας που αναπτύχθηκε για τις ανάγκες της εργασίας βρίσκεται αναρτημένος στο GitHub, στο παρακάτω αποθετήριο:

https://github.com/StavrosKanias/Semantic-Audio-Project-2022.git