ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΦΥΣΙΚΗΣ

Ανάλυση Μουσικών Σημάτων μέσω Wavelets

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της **Σιάννου Έλσας**

Επιβλέπων : Σταύρος Μαλτέζος,
ομότιμος καθηγητής Ε.Μ.Π.

Αθήνα, Δεκέμβριος 2022

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΦΥΣΙΚΗΣ

Ανάλυση Μουσικών Σημάτων μέσω Wavelets

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της **Σιάννου Έλσας**

Επιβλέπων : Σταύρος Μαλτέζος,
ομότιμος καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις 16 Ιανουαρίου 2023

…….	…….	…….
Σ. Μαλτέζος	Θ. Αλεξόπουλος	Ε. Ν. Γαζής
Ομ. Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Ομ. Καθηγητής Ε.Μ.Π.

Αθήνα, Ιανουάριος 2023

……

Σιάννου Έλσα

Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών

Ευχαριστίες

Πρωτίστως θα ήθελα να ευχαριστήσω τον κ. Σταύρο Μαλτέζο, επιβλέποντα καθηγητή μου, που ήταν πάντα παρών σε όλη την πορεία και ‘δημιουργία’ της διπλωματικής μου εργασίας. Με συμβούλεψε και μου έλυσε απορίες σε διάφορα κομμάτια της, ώστε η εργασία να ολοκληρωθεί με επιτυχία. Θα ήθελα επίσης να ευχαριστήσω για την ψυχολογική υποστήριξη, κατά την περίοδο αυτή, την οικογένεια μου και τους φίλους μου και ιδιαίτερα τον φίλο μου Φοίβο Αντουλινάκη.

Περίληψη

Σε αυτή την διπλωματική εργασία μελετάμε διάφορες μεθόδους ανάλυσης για την κατηγοριοποίηση μουσικών κομματιών και τραγουδιών στα αντίστοιχα μουσικά τους είδη, με κυρίαρχη την μέθοδο του διασκορπισμού των κυματιδίων (wavelet scattering). Αρχικά, στο πρώτο κεφάλαιο, μαθαίνουμε τι είναι η μουσική, και από μαθηματικής και φυσικής πλευράς, αλλά και τα συναισθήματα που αυτή δημιουργεί στον ακροατή και την ανάγκη που έχουμε για την ταξινόμηση της, τόσο για οργάνωση και ανακάλυψη καινούριας μουσικής, όσο και για την ψυχολογική ευχαρίστηση του κοινού. Στο δεύτερο κεφάλαιο εξηγούμε τι είναι τα wavelets (κυματίδια), τα οποία και χρησιμοποιούμε στις μεθόδους μας παρακάτω. Στο τρίτο κεφάλαιο καταλαβαίνουμε πως λειτουργεί ο wavelet scattering μετασχηματισμός. Στα επόμενα δύο κεφάλαια βλέπουμε δύο διαφορετικές μεθόδους ανάλυσης για κατηγοριοποίηση τραγουδιών της ίδιας βάσης δεδομένων στα μουσικά τους είδη. Στο τέταρτο κεφάλαιο, λοιπόν, έχουμε μια μέθοδο που χρησιμοποιεί διάφορα χαρακτηριστικά της μουσικής ώστε να συγκρίνει τα τραγούδια μεταξύ τους και να 'κρίνει' σε ποια κατηγορία αυτά ανήκουν. Τα αποτελέσματα της μεθόδου τα παίρνουμε έτοιμα από το αντίστοιχο άρθρο όπου αυτά αναφέρονται. Στο πέμπτο κεφάλαιο βλέπουμε και εκτελούμε, μέσω ενός κώδικα στη MATLAB, την μέθοδο wavelet scattering και προκύπτουν τα αποτελέσματά μας. Στο έκτο κεφάλαιο συγκρίνουμε αυτά τα αποτελέσματα μεταξύ τους, ώστε να δούμε πόσο υπερτερεί η δεύτερη μέθοδος, και στη συνέχεια τα συγκρίνουμε και με κάποιες άλλες μεθόδους που χρησιμοποιούν σαν βάση τα wavelets. Βλέπουμε επίσης την μέθοδο της ανθρώπινης ακοής συγκριτικά με τις αυτόματες μεθόδους ανάλυσης που εξηγήσαμε προηγουμένως. Τέλος, παραθέτουμε τα συμπεράσματα που προέκυψαν από αυτήν την μελέτη και κάποιες ιδέες για βελτίωση των μεθόδων αυτών. Στο παράρτημα παρέχουμε τον κώδικα της MATLAB.

Abstract

In this diploma thesis we examine different analysis methods for musical genre classifications of songs, with emphasis on the wavelet scattering method. For starters, in the first chapter, we explain what music is, from a mathematical and physical perspective, but also the feelings that it creates to the listeners and the need we have for classification, both for organizing and finding new music and also for the contentment of the audience. In the second chapter, we explain what wavelets are, which are used as our tools for the next described methods. In the third chapter we learn how does the wavelet scattering transform work. In the next two chapters we describe two different analysis methods that are used for musical genre classification of songs belonging to the same database. With that in mind, in the fourth chapter, we analyze a method that uses different features of music, so that it can compare the songs and find in which category each belongs to. The results of this method are mentioned in the relevant paper. In the fifth chapter we use a MATLAB code to train the data and test the wavelet scattering method and get our results for this musical classification. In the sixth chapter we compare the results of the two different methods, so that we can see a clear 'win' for the second method, and then we compare those results with some others that are obtained from different analysis methods that use wavelets as well. We also compare these automatic methods for musical genre classification with the human hearing method one. In the last chapter we provide the conclusions of this study and some ideas for improvement of the methods described above. In the appendix you can find the MATLAB code.

Περιεχόμενα

Κεφάλαιο 1

Τι είναι η μουσική;

1.1 Εισαγωγή

Η μουσική είναι πολύ σημαντικό στοιχείο της ανθρωπότητας και των διαφορετικών πολιτισμών. Είναι ένας συνδυασμός φωνητικών ή ορχηστρικών ήχων, συνήθως και των δύο, και δημιουργείται με βάση κάποια πολιτιστικά πρότυπα, όπως ρυθμό, μελωδία και αρμονία, για την μουσική του δυτικού κόσμου. Αποτελεί συνένωση της τέχνης, του συναισθήματος και της ψυχολογίας γενικότερα, της λογικής, ακόμα και της επιστήμης, αφού πέρα από την μαθηματική της και φυσική της σημασία, αφορά άμεσα και την φυσιολογία της ακοής [12]. Η μουσική μας συντροφεύει πολύ στην καθημερινότητα μας και έχει διαδοθεί σε όλη την ανθρωπότητα. Επιλέγουμε να την ακούσουμε για χαλάρωση, για ευχαρίστηση, ως υπόκρουση αλλά και για να μας δημιουργηθούν διάφορα συναισθήματα ή να αλλάξουμε συνειδητά την διάθεσή μας.

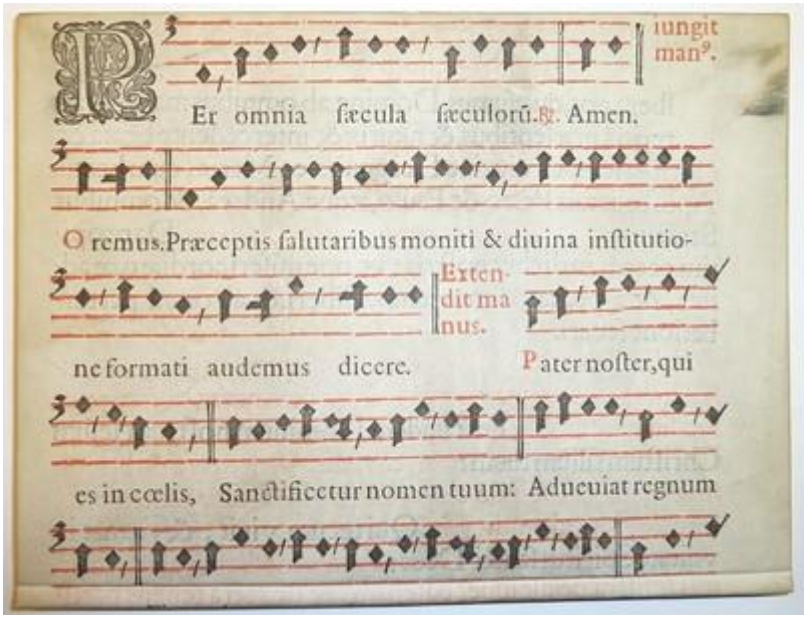
Σε αυτήν την διπλωματική εργασία μελετάμε την κατηγοριοποίηση διαφόρων μουσικών κομματιών και τραγουδιών σε συγκεκριμένα μουσικά είδη με διάφορες μεθόδους ανάλυσης με επίκεντρο την μέθοδο του μετασχηματισμού διασκορπισμού των κυματιδίων, όπου και εκτελούμε έναν κώδικα ώστε να προκύψουν τα αποτελέσματά μας και στη συνέχεια να τα συγκρίνουμε με αυτά των άλλων μεθόδων.

1.2 Ιστορική αναδρομή

Η μουσική είναι μια μορφή τέχνης που ξεκίνησε από την προϊστορική εποχή, όπως έχει φανεί από μουσικά όργανα, φτιαγμένα από κόκκαλα, πέτρες και ξύλα, που έχουν ανακαλυφθεί από εκείνη την εποχή. Γενικά ο κάθε πολιτισμός του κόσμου έχει την δική του ερμηνεία και ιστορία για την μουσική. [24] Στην Ινδία η μουσική χρησιμοποιούνταν για λατρευτικούς σκοπούς από τα πολύ παλιά χρόνια με τους ύμνους Vedic. Στην Κίνα, παρόμοια, αποτελούσε προσθήκη σε τελετές και αφηγήσεις. Ο Κομφούκιος πίστευε ότι μέσω της μουσικής μας φανερώνεται ο χαρακτήρας μας, αφού ακούγοντάς τη μας δημιουργούνται διάφορα συναισθήματα όπως η θλίψη, η χαρά, ο θυμός, η ικανοποίηση, η ευσέβεια και η αγάπη. Στην Αρχαία Ελλάδα η μουσική επίσης αποτελούσε σημαντικό ρόλο, μάλιστα η ίδια η λέξη θεωρείται ότι προέρχεται από τις Μούσες, που ήταν θεές της μουσικής και της ποίησης. Ήταν κυρίως φωνητική με τη βοήθεια της λύρας ή της κιθάρας. Σύμφωνα με τον Πλάτωνα και τον Αριστοτέλη η μουσική εξέφραζε συναισθήματα μιμούμενη τον λόγο και τις ανθρώπινες εκφράσεις και στη συνέχεια τα συναισθήματα του θεατή επηρεάζονταν από αυτήν [12]. Ο Πλάτωνας επίσης αναφέρει ότι ο ήχος είναι κίνηση του αέρα που μεταδίδεται μέσω της ακοής και του εγκεφάλου στην ψυχή και ότι επηρεάζει άμεσα την σωματική και ψυχική ευεξία. Για τον Πυθαγόρα η μουσική αποτελούσε τομέα των μαθηματικών, αυτός ήταν και ο πρώτος μουσικός αριθμολόγος (αριθμολογία: η ιδέα ότι όλα τα πράγματα μπορούν να εκφραστούν με αριθμητικούς όρους) και έθεσε την βάση για την ακουστική. Ανακάλυψε δηλαδή την αντιστοιχία μεταξύ της οξύτητας/τονικότητας (pitch) μιας νότας και του μήκους μιας χορδής, παρόλα αυτά δεν προχώρησε μέχρι τον υπολογισμό της οξύτητας μέσω των δονήσεων. Επιπλέον τόσο ο Πλάτωνας όσο και ο Πυθαγόρας θεώρησαν ότι η μουσική μπορεί να χρησιμοποιηθεί ως «όργανο ελέγχου», εφόσον μέσω αυτής επιβάλλεται πειθαρχία στους ανθρώπους καθώς παροτρύνονται ή αποτρέπονται κάποιες συμπεριφορές. Όσον αφορά τους Αρχαίους Αιγύπτιους η μουσική θεωρούνταν η φυσική της ψυχής. Οι Εβραίοι την χρησιμοποιούσαν για την ψυχική και σωματική θεραπεία και οι Ινδιάνοι χρησιμοποιούσαν την μουσική για τελετουργίες, θεραπεία και εξαγνισμό.

Για πολλούς αιώνες η μουσική πέρναγε από την μια γενιά στην επόμενη προφορικά, αλλά καταγραφή της ξεκίνησε από τους μεσαιωνικούς μοναχούς γύρω στα 500 μ. Χ. έως 1400 μ. Χ. . Αυτοί χρησιμοποιούσαν ένα σύστημα με αριθμούς βασισμένο στην νευματική σημειογραφία (neumes) που είναι

πρόγονος της σύγχρονης μουσικής σημειογραφίας μέσω του πενταγράμμου, όπως γνωρίζουμε [25].



Εικόνα 1.1: Η νευματική σημειογραφία (neumes)

Εκείνη την εποχή, επίσης, ξεκίνησε να αναπτύσσεται η πολυφωνία, δηλαδή πολλαπλοί ήχοι που δημιουργούν μια μελωδία με αρμονία. Ο Χριστιανισμός ανέκαθεν χρησιμοποίησε πάρα πολύ την μουσική, όπου δίνονταν έμφαση στα κείμενα με την βοήθεια της και γινόντουσαν αυτά κάποιες φορές άρθρα πίστης (article of faith). Από την άλλη όμως, θεωρούνταν επίσης ότι η μουσική μέσω της πνευματικής ανύψωσης που προσφέρει μπορεί να απομακρύνει από τον Θεό. Ο Άγιος Αυγουστίνος εξέφραζε τους φόβους του για την υπερίσχυση της μελωδίας στα λατρευτικά κείμενα. Ίδιες ανησυχίες είχε και ο Πλάτωνας στην εποχή του. Για αυτό και κατά την εδραίωση του Παπισμού τον 6ο αιώνα μ. Χ. περάστηκε νομοθεσία για την κάθαρση της λατρευτικής μουσικής από «επικίνδυνα» προς αυτούς ακούσματα [12].

Η μουσική όπως την γνωρίζουμε σήμερα ξεκίνησε να δημιουργείται κατά την περίοδο της Αναγέννησης. Ο J. Kepler (1571-1630) είχε σαν ιδέα ότι η μουσική συνδέεται με την κίνηση των πλανητών. Ο R. Descartes (1596-1650) αναγνώριζε την μαθηματική βάση της μουσικής, αλλά είχε και αυτός τις «πλατωνικές αντιλήψεις», δηλαδή να ακολουθείται συγκεκριμένος ρυθμός και απλές μελωδίες ώστε να μην έχει δημιουργικές και ανήθικες επιδράσεις στους ανθρώπους. Ο γερμανός φιλόσοφος A. Schopenhauer (1788–1860) αναγνώριζε

την σύνδεση μεταξύ του ανθρώπινου συναισθήματος και της μουσικής και την θεωρούσε ως την πιο σημαντική μεταξύ των τεχνών διότι έχει την ικανότητα να αντικατοπτρίσει την βούληση του σύμπαντος. Ο R. Scruton (1944-2020) υποστήριζε την ηθική εκπαίδευση της μουσικής και την δύναμή της να δομήσει τον χαρακτήρα του ανθρώπου.

Κάποιοι μαρξιστές ιστορικοί υποστήριζαν την άποψη ότι η μουσική και το τραγούδι, πιο συγκεκριμένα, είναι επακόλουθο της εργασιακής δραστηριότητας. Η ανάγκη, λοιπόν, για τον συντονισμό των εργατών ως σύνολο της ομάδας οδήγησε στην δημιουργία ρυθμικών και επαναλαμβανόμενων μοτίβων ώστε να εμψυχώνεται το ανθρώπινο δυναμικό. Τέτοιου είδους μουσική είχαν οι αγρότες και μετανάστες στην Αμερική αλλά και οι ανθρακωρύχοι που το τραγούδι τους βοήθαγε να εξομαλύνουν τις πολύ δύσκολες συνθήκες εργασίας τους. Οι ίδιοι επίσης τραγουδάγανε σε διαμαρτυρίες και γράφτηκαν και διάφορα τραγούδια από καλλιτέχνες για τις κακές συνθήκες εργασίας των ανθρακωρύχων. Επίσης υπάρχει και το καθαρά πολιτικό τραγούδι των συνδικάτων που εξελίχθηκε και αυτό σε τραγούδι διαμαρτυρίας και μέχρι και σήμερα στις πορείες διαμαρτυρίας που γίνονται έναντι στο σύστημα ακούγονται τραγουδιστά συνθήματα. Παρόμοιου είδους τραγούδια επικρατούν σε όλο τον κόσμο.

Η μουσική επίσης ήταν για πολλούς τρόπος να ξεφύγεις από την πραγματικότητα η οποία μπορεί να ήταν πολύ σκληρή. Έτσι ίσχυε για τους Αφροαμερικανούς σκλάβους που δούλευαν στα χωράφια και στους οικισμούς γενικότερα των πλούσιων Αμερικανών τον 19ο αιώνα. Με επιρροές από την αφρικανική κουλτούρα τους με τις ιδιαίτερες μουσικές παραδόσεις άρχισαν να δημιουργούνται η τζαζ και η μπλουζ μουσικές. Εξελίχθηκαν λοιπόν κυρίως στις Ηνωμένες Πολιτείες διότι τους απέτρεπαν να κρατήσουν τις μουσικές παραδόσεις της πατρίδας τους και οι ίδιοι ένιωσαν την ανάγκη να αντικαταστήσουν αυτήν την απώλεια με μια άλλη μορφή μουσικής έκφρασης. Υπήρξε επιρροή, πέρα του αφρικανικού ρυθμού, και από τις ευρωπαϊκές αρμονικές δομές [26]. Και οι 2 μουσικές μετεξελίχθηκαν και σε άλλα μουσικά είδη όπως η swing και η boogie-woogie αντίστοιχα στις αρχές του 20ου αιώνα, που αποτελούν και γνωστούς χορούς που έχουν πολλά κοινά στοιχεία μεταξύ τους. Άλλο ένα γνώρισμα, λοιπόν, της μουσικής γενικότερα είναι η συνοδεία της στο χορό. Οι εργάτες σύχναζαν στα μέρη όπου ακουγόταν αυτή η μουσική ώστε να χαλαρώσουν μετά το τέλος μιας κουραστικής μέρας χορεύοντας ή απλά ακούγοντας τους οργανοπαίχτες. Στις αρχές είχε και κοινωνικό χαρακτήρα εφόσον κυρίως Αφροαμερικανοί συναντιόντουσαν για να καταπολεμήσουν με αυτόν τον τρόπο τον ρατσισμό που βίωναν από τους λευκούς. Με το πέρασμα

των χρόνων αυτά τα είδη και οι αντίστοιχοι χοροί τους διαδόθηκαν και στον υπόλοιπο πληθυσμό ή κυρίως στους πιο «ανοιχτόμυαλους» λευκούς. Πλέον η χαλάρωση και η καλοπέραση με το άκουσμα μουσικής, είτε σε μεμονωμένο επίπεδο είτε στο κοινωνικό σύνολο, όπως για παράδειγμα σε κάποιο μαγαζί όπου παίζει ζωντανά ένα μουσικό σχήμα, έχει επικρατήσει έως και σήμερα.
Παρόλα αυτά, όσον αφορά στους διάφορους πολιτισμούς ο καθένας έχει ιδιαίτερα χαρακτηριστικά και κανόνες αντίληψης της μουσικής και επηρεάζουν με αυτά την συναισθηματική αντίδραση του κάθε ανθρώπου στα ακούσματά του. Για παράδειγμα σε κάποια μέρη της Αφρικής, η μουσική που δεν έχει σταθερό ρυθμό ώστε να χρησιμοποιείται στον χορό, θεωρείται μορφή θρήνου [12].

1.3 Χαρακτηριστικά της μουσικής

Η μουσική αποτελείται από πολλά συστατικά ή χαρακτηριστικά που οργανώνοντάς τα σε κατηγορίες μπορούμε να τα κατανοήσουμε καλύτερα. Αυτά είναι τα εξής [27]:

1.3.1 Ήχος

Ο ήχος είναι η δυνατότητα κάποιου να ακούει ή να αισθάνεται τις δονήσεις κάποιας κίνησης των σωματιδίων του αέρα που προκαλείται από μια φωνή ή ένα μουσικό όργανο. Αποτελείται από τα παρακάτω χαρακτηριστικά.

- **Απόηχος ή αρμονική (overtone):** Είναι μια οξύτητα (pitch) που αποτελεί κομμάτι της αρμονικής σειράς μιας θεμελιώδους οξύτητας/τονικότητας και μπορεί να ακουστεί συγχρόνως με αυτήν. Πιο συγκεκριμένα αρμονική ονομάζουμε ένα κύμα με συχνότητα θετικό ακέραιο πολλαπλάσιο της θεμελιώδους συχνότητας, δηλαδή την συχνότητα του αρχικού περιοδικού σήματος.
- **Ηχόχρωμα (timbre):** Είναι η χροιά ενός ήχου και μέσω αυτού μπορούμε να ξεχωρίσουμε τα διαφορετικά μουσικά όργανα ή τις ιδιαίτερες φωνές των ανθρώπων μεταξύ τους. Είναι το «χρώμα», ο χαρακτηρισμός που αποδίδουμε σε έναν ήχο, για παράδειγμα για μια ανθρώπινη φωνή μπορούμε να πούμε ότι είναι ‘ρινική’, ‘ηχηρή’, ‘έντονη’, ‘θερμή’, ‘ψυχρή’, ‘υψηλή’, ‘χαμηλή’, ‘διαπεραστική’, ‘βαθιά’ κ.α. ώστε να την ξεχωρίσουμε

από κάποια άλλη. Τα μουσικά όργανα ταξινομούνται σε οικογένειες ανάλογα με τον τρόπο κατασκευής τους (υλικό, σχήμα) και τις φυσικές τους ιδιότητες. Έτσι έχουμε 4 οικογένειες: τα έγχορδα (όπως το βιολοντσέλο και η κιθάρα), τα ξύλινα πνευστά (κλαρινέτο, σαξόφωνο και άλλα), τα χάλκινα πνευστά (για παράδειγμα η τρομπέτα και το κόρνο) και τα κρουστά (π.χ. το πιάνο και τα ντραμς).

- **Οξύτητα/τονικότητα (pitch):** Είναι η συχνότητα που αντιστοιχεί σε συγκεκριμένη νότα. Η νότα λα (A) έχει συχνότητα 440Hz και με βάση αυτήν θεωρούμε αν μια οξύτητα είναι υψηλή, αν δηλαδή ξεπερνάει την συχνότητα της λα, ή αντίστοιχα αν είναι χαμηλή.
- **Ένταση:** Πόσο δυνατός/ισχυρός ή απαλός είναι ένας ήχος.
- **Διάρκεια:** Είναι η αξία ενός ήχου, αν τον ακούμε για μεγάλο ή μικρό χρονικό διάστημα.

1.3.2 Μελωδία

Η μελωδία είναι η αλληλουχία μουσικών νοτών ή αλλιώς μια σειρά από οξύτητες που οργανώνονται σε φράσεις. Είναι το κεντρικό θέμα που ακούγεται σε ένα μουσικό κομμάτι και επαναλαμβάνεται καθόλη τη διάρκεια του τραγουδιού.

1.3.3 Αρμονία

Η αρμονία είναι ο ταυτόχρονος συνδυασμός διαφόρων νοτών που σχηματίζουν τις συγχορδίες, οι οποίες είναι συνοδευτικές στην μελωδία. Αυτή ενισχύει το μουσικό κομμάτι και του προσφέρει «χρώμα».

1.3.4 Ρυθμός

Ο ρυθμός είναι η οργάνωση της μουσικής στο χρόνο. Είναι η επανάληψη των νοτών και των παύσεων (σιωπές) κατά την διάρκεια ενός μουσικού κομματιού. Ή αλλιώς είναι ο παλμός, ο σταθερός χτύπος δηλαδή, που εμφανίζεται ανά τακτά χρονικά διαστήματα σε ένα κομμάτι. Τα χαρακτηριστικά του είναι τα ακόλουθα:

- **Αξίες των νοτών:** Είναι η χρονική διάρκεια που πρέπει ο μουσικός να παίξει την νότα ή να «κρατήσει» την παύση αντίστοιχα που διαβάζει στο πεντάγραμμο [28].

Term	Symbol	Value
whole note		4 beats
half note		2 beats
quarter note		1 beat
eighth note		1/2 beat
joined eighth notes		1/2 + 1/2=1
sixteenth note		1/4 beat
joined sixteenth notes		1/4 + 1/4 + 1/4 + 1/4=1

Term	Symbol	Value
whole note rest		4 beats of silence
half note rest		2 beats of silence
quarter note rest		1 beat of silence
eighth note rest		1/2 beat of silence
sixteenth note rest		1/4 beat of silence

Πίνακας 1.1: Στον αριστερό πίνακα έχουμε την αξία των νοτών και τον συμβολισμό τους και στον δεξιό πίνακα έχουμε την αξία και τον συμβολισμό των παύσεων. Οι ονομασίες των νοτών στα ελληνικά είναι: ολόκληρο που διαρκεί 4 χρόνους (ή μπιτ), μισό που διαρκεί 2 χρόνους, τέταρτο που διαρκεί 1 χρόνο, όγδοο που διαρκεί μισό χρόνο, δέκατο έκτο που είναι το μισό του ογδόου, τριακοστό δεύτερο που είναι το μισό του δέκατου έκτου και εξηκοστό τέταρτο που είναι το μισό του τριακοστού δευτέρου. Τα δύο τελευταία έχουν ως σύμβολο ένα τέταρτο με 3 και 4 γραμμούλες αντίστοιχα προς τα δεξιά.

- **Μέτρο:** Είναι ενότητα συγκεκριμένης χρονικής αξίας στο πεντάγραμμο. Διευκολύνει τον μουσικό να καταλαβαίνει την χρονική του θέση στο κομμάτι που διαβάζει. Τα μέτρα μεταξύ τους χωρίζονται με την υποδιαστολή.
- **Μέγεθος του μέτρου/χρόνος:** Βρίσκεται στην αρχή του πενταγράμμου και καθορίζει την χρονική αξία του μέτρου. Έχει την μορφή κλάσματος συνήθως, όπου ο αριθμητής συμβολίζει την αξία του κάθε μέτρου και ο παρανομαστής την βασική χρονική μονάδα μέτρησης. Τα 4/4 (τέσσερα τέταρτα) για παράδειγμα, που είναι και το πιο κοινό μέγεθος μέτρου,

σημαίνει ότι σε κάθε μέτρο θέλουμε να έχουμε 4 φορές από την αξία του τετάρτου. Αυτό πέρα από το κλάσμα συμβολίζεται και ως $\mathbf{C}$.

- **Μπιτ/ρυθμός/παλμός (beat):** Μπορεί να είναι σαφής ή ασαφής, αναλόγως αν υπάρχει όργανο που να παίζει πάνω στο μπιτ ή στον ρυθμό της μουσικής ή αντίστοιχα αν δεν υπάρχει τέτοιο όργανο. Συνδέεται με το μέγεθος του μέτρου που είδαμε πριν και κάθε χρόνος του μέτρου έχει την δική του διαρρύθμιση των ισχυρών και των αδύναμων μπιτ σε κάθε μέτρο. Για παράδειγμα στα 4/4 που είδαμε και πριν, όσον αφορά στην κλασσική μουσική, η διαρρύθμιση των μπιτ σε κάθε μέτρο είναι ως εξής: **1**, 2, **3**, 4. Άρα είναι πιο ισχυρά το πρώτο και το τρίτο μπιτ, με το πρώτο ακόμα πιο ισχυρό από το τρίτο. Δίνεται έμφαση, δηλαδή, σε αυτά τα 2 περισσότερο.

1.3.5 Δομή ή μορφή (Structure)

Η δομή στην μουσική είναι η σειρά αλληλουχίας του κάθε μέρους ή τμήματος του μουσικού κομματιού. Μέσω αυτής ξέρουμε ποιο τμήμα πρέπει να παιχτεί και πότε, πόσες φορές πρέπει να παιχτεί και που επαναλαμβάνεται η μουσική. Για παράδειγμα το κουπλέ, το ρεφρέν, η εισαγωγή, το κλείσιμο κλπ.

1.3.6 Υφή (Texture)

Η υφή στην μουσική είναι ο συνδυασμός της μελωδίας, της αρμονίας και του ρυθμού και αναδεικνύει την ποιότητα και τον ήχο του κάθε τραγουδιού. Έχουμε 3 βασικά είδη υφής:

- **Μονοφωνία (Monophony):** Μια μελωδία χωρίς κάποια άλλη παρεμβολή, όπως για παράδειγμα μια φωνή που τραγουδάει μόνη της, σόλο. Μπορεί η μελωδία να περιέχει περισσότερες φωνές ή μουσικά όργανα, αρκεί όμως όλα να τραγουδάνε/παίζουνε τις ίδιες νότες ταυτόχρονα, είτε στην ίδια είτε σε διαφορετικές οκτάβες.
- **Ομοφωνία (Homophony):** Μια μελωδία με την συνοδευτική της αρμονία. Για παράδειγμα ένας πιανίστας που ταυτόχρονα τραγουδάει ή ένας τραγουδιστής και η μπάντα του.

- **Πολυφωνία (Polyphony):** Περισσότερες από μια μελωδίες που ακούγονται ταυτόχρονα και ίσως και τις αρμονίες τους. Για παράδειγμα δύο τραγουδιστές με την μπάντα τους που τραγουδάνε διαφορετικές μελωδίες ή να τραγουδάνε την ίδια μελωδία αλλά ο ένας με καθυστέρηση σε σχέση με τον άλλο.

1.3.7 Έκφραση (Expression)

Με τον όρο έκφραση εννοούμε τον τρόπο με τον οποίο παίζεται ένα συγκεκριμένο μουσικό κομμάτι. Έχει χαρακτηριστικά και από τον ρυθμό και από την μελωδία:

- **Δυναμικές (Dynamics):** Είναι η ένταση, ξανά, αλλά όχι μόνο για το 'δυνατά' και 'σιγανά'. Περιέχει και τα crescendo, diminuendo όπου υπάρχει μια αλληλουχία από το 'σιγανά' στο 'δυνατά' και αντίστροφα και δεν έρχεται ξαφνικά.
- **Τέμπο (Tempo):** Είναι χαρακτηριστικό του ρυθμού και είναι το πόσο γρήγορα ή αργά παίζεται ή τραγουδιέται ένα μουσικό κομμάτι. Μετριέται σε μπιτ/παλμούς ανά λεπτό (bpm: beats per minute). Βρίσκοντας τα beat του κομματιού μπορούμε στη συνέχεια να καταλάβουμε την ταχύτητα αυτού. Υπάρχουν κάποιοι ιταλικοί όροι (ιταλικοί διότι οι Ιταλοί συνθέτες ήταν οι πρώτοι που άρχιζαν να καταγράφουν όρους έκφρασης στις συνθέσεις τους και στη συνέχεια διαδόθηκαν και στον υπόλοιπο κόσμο) που σημειώνονται στην αρχή του κομματιού και μας εξηγούν τον τρόπο που θα πρέπει να παιχτεί το μουσικό αυτό κομμάτι. Κάποιοι από αυτούς είναι οι εξής:
    - *Largo:* Αργά και πλατιά (οι νότες διαρκούν μεγάλο χρονικό διάστημα)
    - *Adagio:* Αργά
    - *Andante:* Μεσαίου ρυθμού, σαν το περπάτημα
    - *Moderato:* Μέτριας ταχύτητας
    - *Allegro:* Γρήγορα
    - *Presto:* Πολύ γρήγορα
    - *Prestissimo:* Πάρα πολύ γρήγορα

Επίσης το tempo μπορεί να αλλάξει κατά την διάρκεια του κομματιού, δεν παραμένει σταθερό συνέχεια, έτσι έχουμε:

- *Accelerando:* Σταδιακά πιο γρήγορα
- *Ritenuto:* Σταδιακά πιο αργά
- *A tempo:* Επιστροφή στο αρχικό tempo

- **Άρθρωση ή έμφαση (Articulation or accent):** Είναι ο τρόπος με τον οποίο παίζονται συγκεκριμένες νότες ή προφέρονται κάποιες λέξεις (σε τραγούδι αντίστοιχα). Μπορεί να γράφονται σαν λέξη στα σημεία όπου επιθυμούμε ή να έχουν κάποιο συμβολισμό:
  - *Staccato:* Κοφτά (Συμβολίζεται σαν τελεία πάνω από την νότα που θα παιχτεί κοφτά)
  - *Legato:* Ομαλά και ενωμένες οι νότες, δεν υπάρχει κενό ή παύση ανάμεσα τους
  - *Marcato:* Νότα, συγχορδία ή τμήμα που παίζεται πολύ πιο έντονα και δυνατά
  - *Sforzando:* Ξαφνική ένταση/έμφαση
  - *Subito:* Ξαφνικά ή απευθείας παίξιμο νότας

1.4 Είδη μουσικής και ανάγκη ταξινόμησής της

Σε αυτήν την εργασία, όπως αναφέραμε και προηγουμένως, θα επικεντρωθούμε στην κατηγοριοποίηση διαφόρων τραγουδιών σε μουσικά είδη. Με τον όρο ‘μουσικό είδος’ εννοούμε έναν τρόπο περιγραφής ενός μουσικού κομματιού τόσο για τους χρήστες όσο και για την μουσική βιομηχανία [12]. Βοηθάει στην οργάνωση βάσεων μουσικών δεδομένων και μουσικών βιβλιοθηκών και γενικότερα στην διευκόλυνση των χρηστών στην εύρεση των μουσικών κομματιών που επιθυμούν ή παρόμοιων με αυτά. Τα όρια μεταξύ των διαφορετικών ειδών μουσικής δεν είναι σαφή, μάλιστα μπορούν κάποια τραγούδια να ταιριάζουν σε περισσότερα από ένα μουσικά είδη, όμως είναι η καλύτερη επιλογή περιγραφικού προσδιορισμού ενός κομματιού που έχουμε, καθώς διευκολύνεται η εύρεση ομοιοτήτων και διαφοροποιήσεων μεταξύ των διαφορετικών ειδών.

Σημαντικό να επισημάνουμε ότι η κατηγοριοποίησης της μουσικής σε είδη βοηθάει για τραγούδια που δεν γνωρίζουμε. Για παράδειγμα, την δική μας συλλογή (playlist) την οργανώνουμε όπως εμείς επιθυμούμε, είτε αυτό είναι με βάση τους καλλιτέχνες, είτε αλφαβητικά, είτε χρονολογικά κλπ. . Για τραγούδια όμως που δεν γνωρίζουμε και θα θέλαμε να βρούμε παρόμοια με τις προσωπικές μας προτιμήσεις, η ταξινόμηση κατά μουσικά είδη είναι η καλύτερη

επιλογή. Στην εποχή μας, με την άνοδο της ψηφιακής μουσικής και την ύπαρξη των τεράστιων μουσικών πλατφόρμων στο διαδίκτυο, είναι πρακτικά αδύνατον να περιηγηθούμε ψάχνοντας ένα-ένα τα μουσικά κομμάτια της κάθε πλατφόρμας. Χρειαζόμαστε συνεπώς ισχυρά συστήματα διαχείρισης γνώσης (knowledge management). Η κατηγοριοποίηση της μουσικής αποτελείται από ένα τέτοιο σύστημα και με την ανάλυση της ακουστικής συμπεριφοράς (listening behavior) του κάθε ακροατή βελτιώνεται το σύστημα προσφέροντας στο κοινό προτάσεις για καινούρια τραγούδια, δημιουργία καινούριων λιστών μουσικής (playlist) και γενικότερα οργάνωση της μουσικής του καθενός.

Παρόλα αυτά, πολύ λίγα είδη είναι ξεκάθαρα καθορισμένα, όπως είπαμε και πριν, και συνήθως υπάρχει αλληλοεπικάλυψη με άλλα. Επίσης κάποια έχουν μεγάλο εύρος, όπως πχ η ροκ μουσική, η οποία έχει και πολλές υποκατηγορίες, ενώ άλλα είδη είναι πολύ πιο περιορισμένα [29]. Περιπλέκοντας περισσότερο τα πράγματα, το φαινόμενο εμφάνισης καινούριων μουσικών ειδών είναι αρκετά συχνό και επιπλέον η γνώση των ήδη υπαρχόντων ειδών μπορεί να αλλάξει με την πάροδο του χρόνου. Αυτό δημιουργεί περισσότερα προβλήματα, διότι θα πρέπει να «επανεκπαιδεύσουμε» (retraining) τα συστήματα των ταξινομητών (classifiers). Γενικά οι αλγόριθμοι που χρησιμοποιούμε ως ταξινομητές στην μηχανική μάθηση, όπως θα δούμε και στα επόμενα κεφάλαια με τον support vector machine (SVM) για παράδειγμα, χρειάζονται ένα μεγάλο training set, δηλαδή μια σειρά για εξάσκηση και ανάπτυξη του ταξινομητή. Έχει προταθεί επίσης η κατηγοριοποίηση με βάση την ομοιότητα των τραγουδιών, όμως εκεί δημιουργούνται άλλα προβλήματα ασάφειας και υποκειμενικότητας, οπότε δεν θεωρείται καλή εναλλακτική. Όμως, στην τελική, παρατηρείται ότι τα μουσικά είδη έχουν μεγάλη αξία πέρα από την χρησιμότητά τους στην οργάνωση και ανακάλυψη της μουσικής, πέρα δηλαδή από το καθαρά εμπορικό κομμάτι τους. Πάρα πολλά άτομα ταυτοποιούνται σε επίπεδο πολιτισμικό με τα αγαπημένα τους μουσικά είδη και για αυτό βλέπουμε ομάδες να ντύνονται ή να μιλάνε με τον ίδιο τρόπο, αναλόγως με την μουσική που ακούν, όπως για παράδειγμα με την ραπ. Μάλιστα το είδος μουσικής είναι τόσο σημαντικό για τους ακροατές, αφού κάποιες ψυχολογικές έρευνες έχουν δείξει ότι το μουσικό στυλ ενός κομματιού μπορεί να επηρεάσει την προτίμησή τους για αυτό περισσότερο κι από το ίδιο το κομμάτι [29]. Συνεπώς καταλήγουμε στο γεγονός ότι η αυτόματη κατηγοριοποίηση της μουσικής σε αντίστοιχα μουσικά είδη, παρότι δύσκολη και απαιτητική, είναι μεγάλης σημασίας και από εμπορικής και ερευνητικής πλευράς αλλά και από ψυχολογικής, για την ευχαρίστηση του ίδιου του κοινού.

1.5 Τελικά ποια είναι η επίδραση της μουσικής;

Οι λόγοι για τους οποίους ακούμε μουσική εν τέλει ποικίλλουν, μπορεί να είναι συναισθηματικοί, αισθητικοί, διανοητικοί, πρακτικοί ή και ηθικοί [12]. Η μουσική αγγίζει το σώμα και το μυαλό, μπορεί να μας κάνει να θυμηθούμε το παρελθόν μας ή το παρόν μας ακούγοντας τους στίχους των τραγουδιών που μας ταυτίζουν με τον καλλιτέχνη, ή απλώς μια μελωδία που κάνει πιο έντονα τα συναισθήματα μας. Επίσης πολλές φορές τη συνδέουμε με σημαντικά γεγονότα της ζωής μας, όπως ακριβώς συμβαίνει και στις ταινίες ή τα θέατρα. Δίχως μουσική αυτά θα ήταν ανιαρά, όσο καλοί και να ήταν οι ηθοποιοί ή η σκηνοθεσία. Γενικά η μουσική συνεισφέρει στην εξέλιξή μας και σε ολόκληρη την ζωή μας συνυπάρχει μαζί μας, αποτελώντας μέρος της κοινωνικής μας φύσης. Είναι πολύ βασικό μέρος της καθημερινότητάς μας, εφόσον υπάρχει παντού γύρω μας, την ακούμε, τη συζητάμε, τη θυμόμαστε αλλά και τη δημιουργούμε.

Κεφάλαιο 2

Η Θεωρία των Κυματιδίων

2.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα αναφερθούμε στον μετασχηματισμό των κυματιδίων (wavelets) ο οποίος χρησιμοποιείται για την ανάλυση δεδομένων (χρονοσειρών) σε διαφορετικές κλίμακες (scale) και αναλύσεις (resolutions) μέσω ειδικών συναρτήσεων. Με την επιλογή της κατάλληλης κλίμακας (ή χρονικού παραθύρου) μπορούμε να δούμε και τα προφανή χαρακτηριστικά ενός σήματος και τα πιο μικρά και ιδιαίτερα, με λίγα λόγια φαίνεται και το δάσος και τα δέντρα! Η βασική ιδέα του μετασχηματισμού των wavelets, βασίζεται στην ανάλυση ενός σήματος τόσο στον χρόνο όσο και στη συχνότητα και είναι εντοπισμένο σε αυτά συγχρόνως, σε αντίθεση με άλλους μετασχηματισμούς όπως αυτού του Fourier (FT). Επομένως παρακάτω θα δούμε τον μετασχηματισμό Fourier και τους περιορισμούς του που προσπαθούμε να λύσουμε με τον Short-Time Fourier Transform (STFT), τον wavelet μετασχηματισμό (WT) που μας λύνει και τους περιορισμούς του STFT και διάφορα παραδείγματα. Αυτό το κεφάλαιο βασίζεται στις αναφορές [18], [19], [20], [21], [22], [23].

2.2 Μετασχηματισμός Fourier

Ας ξεκινήσουμε αρχικά με την έννοια του μετασχηματισμού. Τους μαθηματικούς μετασχηματισμούς γενικότερα τους χρησιμοποιούμε για να

εξάγουμε όλες τις δυνατές πληροφορίες από τα σήματα που υπάρχουν σε ένα αρχικό σήμα (raw signal), για παράδειγμα ένα σήμα που είναι συνάρτηση του χρόνου (time-domain). Εν γένει υπάρχουν πληροφορίες οι οποίες σχετίζονται με την εμφάνιση των συχνοτήτων στο σήμα. Στην περίπτωση που έχουμε σήματα εξαρτώμενα από τον χρόνο, εφαρμόζοντας έναν μετασχηματισμό στο σήμα μας μπορούμε να περάσουμε από το χρονοεξαρτώμενο (time-domain) στο εξαρτώμενο από την συχνότητα (frequency spectrum) σήμα. Ο θεμελιώδης μετασχηματισμός είναι ο Fourier (FT). Ο μετασχηματισμός Fourier μιας συνάρτησης $f(t)$ ορίζεται ως το ολοκλήρωμα [21] :

$$F(\omega) = \int_{-\infty}^{+\infty} f(t)e^{-j\omega t}dt \qquad (2.1)$$

όπου ω είναι η κυκλική συχνότητα $\omega = 2\pi f$ όπου f: συχνότητα.

Συνεπώς ο μετασχηματισμός Fourier μπορεί να θεωρηθεί ως ένα σύστημα με είσοδο (input) σήματα $f(t)$ και έξοδο (output) τα αντίστοιχα μετασχηματισμένα σήματα $F(\omega)$. Είναι μια μιγαδική συνάρτηση, στην γενική περίπτωση, που έχει πραγματικό και φανταστικό μέρος.

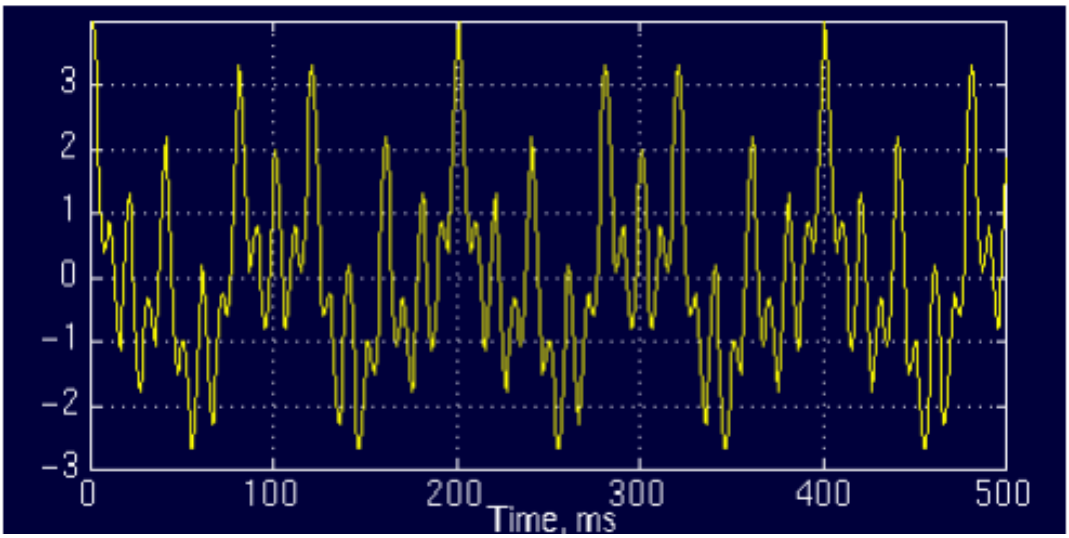
$$F(\omega) = A(\omega)e^{j\varphi(\omega)} \qquad (2.2)$$

όπου $A(\omega)$: το φάσμα και $\varphi(\omega)$: η γωνία φάσης Fourier της συνάρτησης $F(\omega)$ και ισχύει $j^2 = -1$. Επίσης το $A^2(\omega)$ είναι η ενέργεια ανά μονάδα συχνότητας του φάσματος.
Υπάρχει και ο αντίστροφος μετασχηματισμός Fourier, με τον οποίο γνωρίζοντας τον $F(\omega)$ μιας συνάρτησης $f(t)$ μπορούμε να βρούμε την αρχική συνάρτηση $f(t)$ ως εξής:

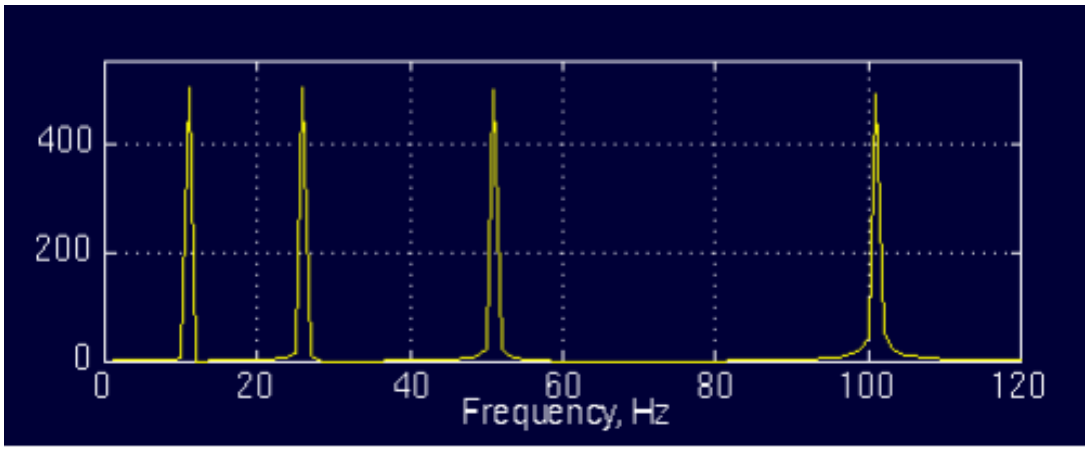
$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} F(\omega)e^{j\omega t}d\omega \qquad (2.3)$$

Συνεπώς, λοιπόν, ο μετασχηματισμός Fourier μας δίνει την πληροφορία για το ποιες συχνότητες υπάρχουν στο σήμα μας χωρίς όμως να γνωρίζουμε την χρονική στιγμή που εμφανίζεται η κάθε συχνότητα. Αν έχουμε στάσιμο σήμα (stationary signal), δηλαδή σήμα που η πληροφορία της συχνότητας να μην αλλάζει με την πάροδο του χρόνου, τότε γνωρίζουμε ότι το περιεχόμενο της συχνότητας υπάρχει για όλους τους χρόνους του σήματος, άρα δεν μας δημιουργεί κάποιο πρόβλημα. Ας δούμε ένα παράδειγμα [19]. Έστω ότι έχουμε το σύνθετο σήμα:

$x(t) = cos(2\pi 10t) + cos(2\pi 25t) + cos(2\pi 50t) + cos(2\pi 100t)$, το οποίο είναι στάσιμο, διότι οι συχνότητες 10, 25, 50 και 100 Hz υπάρχουν για οποιαδήποτε χρονική στιγμή. Βλέπουμε τα διαγράμματα που προκύπτουν για το συγκεκριμένο σήμα:



Σχήμα 2.1: Το διάγραμμα του πλάτους του στατικού σήματος στο χρόνο



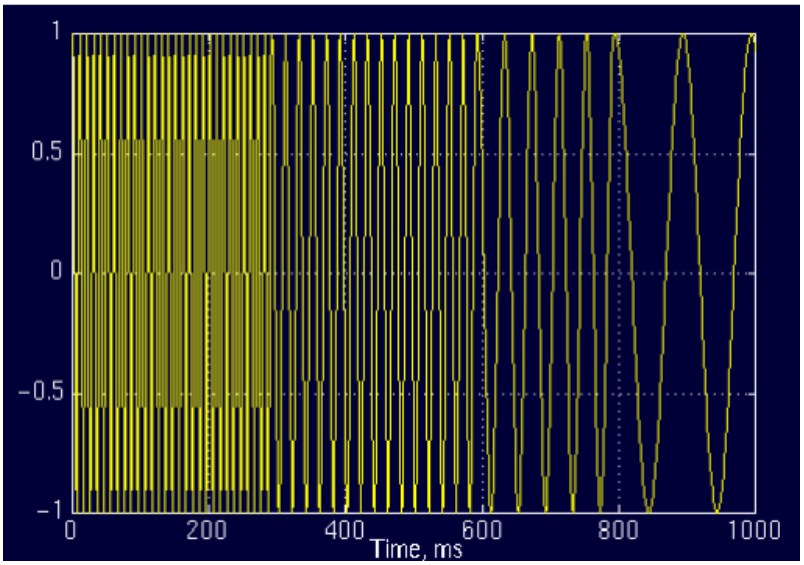
Σχήμα 2.2: Το διάγραμμα πλάτους-συχνότητας του ίδιο σήματος, μετασχηματισμός Fourier.

Στο *σχήμα 2.2* βλέπουμε το μετασχηματισμένο σήμα του *σχήματος 2.1*. Φαίνονται καθαρά οι συχνότητες που υπάρχουν στο σήμα, που όπως εξηγήσαμε και πριν, υπάρχουν για κάθε χρονική στιγμή.

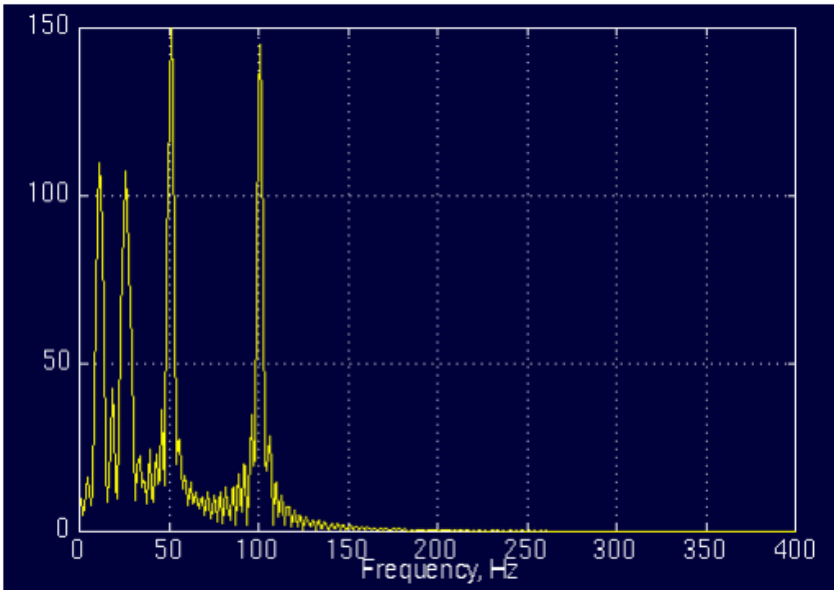
Ας δούμε ένα δεύτερο παράδειγμα σύνθετου σήματος:

$$x(t)\begin{cases} \sin(2\pi 100t), & t = 0 - 300\ ms \\ \sin(2\pi 50t), & t = 300 - 600\ ms \\ \sin(2\pi 25t), & t = 600 - 800\ ms \\ \sin(2\pi 10t), & t = 800 - 1000\ ms \end{cases}$$

Αυτό το σήμα, όπως φαίνεται είναι μη-στατικό (non-stationary signal).



Σχήμα 2.3: Το διάγραμμα του μη-στατικού σήματος στο χρόνο



Σχήμα 2.4: Ο μετασχηματισμός Fourier του ίδιου σήματος. Διάγραμμα συχνότητας

Στο *σχήμα 2.4* φαίνονται τέσσερις κορυφές που αντιστοιχούν και στις τέσσερις συχνότητες που έχουμε. Το μεγαλύτερο πλάτος των συχνοτήτων 50 και 100 Hz οφείλεται στην μεγαλύτερη χρονική διάρκεια που αυτές επικρατούν, 100 ms παραπάνω από τις δύο προηγούμενες συχνότητες. Αν συγκρίνουμε τώρα τους δύο FT στα *σχήματα 2.2* και *2.4* έχουμε ότι τα δύο αυτά διαγράμματα μοιάζουν πάρα πολύ εφόσον έχουν τέσσερις κορυφές που αντιστοιχούν στις ίδιες ακριβώς συχνότητες. Όμως στο πρώτο διάγραμμα οι συχνότητες αντιστοιχούν για κάθε χρόνο, ενώ στο δεύτερο διάγραμμα οι συχνότητες αντιστοιχούν για πολύ συγκεκριμένες χρονικές περιόδους. Συνεπώς παίρνουμε για δύο πολύ διαφορετικά σήματα (σχεδόν) το ίδιο φάσμα. Επομένως, δεν μπορούμε να χρησιμοποιήσουμε τον μετασχηματισμό Fourier για μη-στάσιμα σήματα, εκτός κι αν θέλουμε να ξέρουμε μονάχα τις συχνότητες που επικρατούν σε ένα σήμα χωρίς να ξέρουμε την χρονική στιγμή που αυτές υπάρχουν.

2.3 Μετασχηματισμός Fourier σύντομης χρονικής διάρκειας

Ο STFT δημιουργήθηκε με σκοπό να ξεπεράσει το πρόβλημα που δημιουργούσε ο FT στα μη-στάσιμα σήματα. Πιο συγκεκριμένα χωρίζει το μη-στάσιμο σήμα σε μικρά χρονικά διαστήματα ώστε το καθένα από αυτά να θεωρείται στάσιμο. Ο STFT ορίζεται ως εξής για συνεχείς συναρτήσεις:

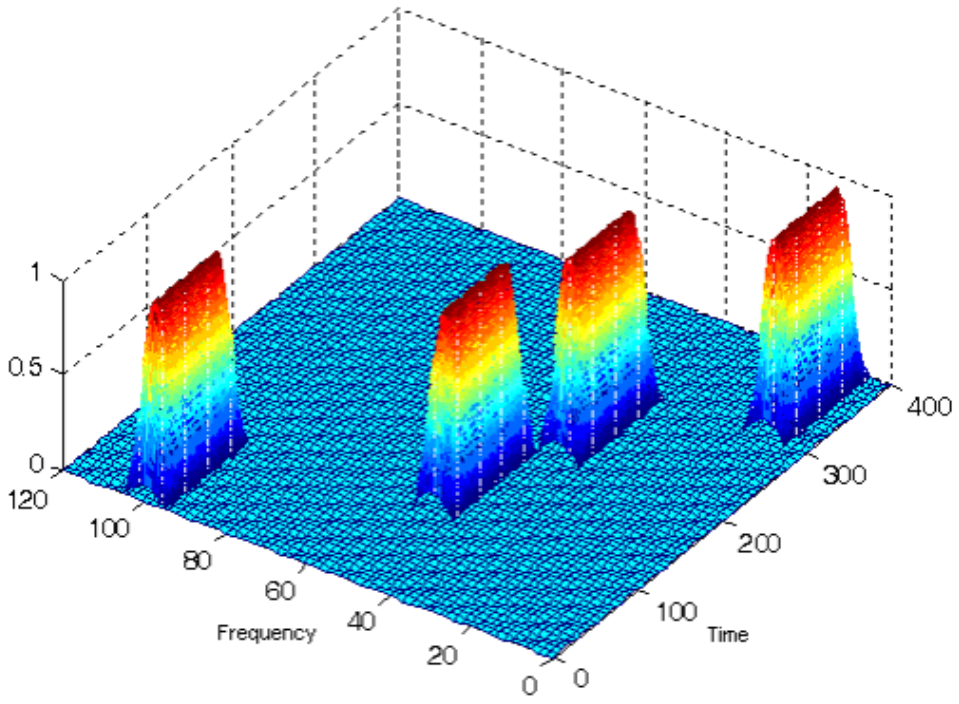
$$X(\tau, \omega) = \int_{-\infty}^{+\infty} x(t) w(t-\tau) e^{-j\omega t} dt \qquad (2.4)$$

όπου $x(t)$ είναι το σήμα μας και $w(t-\tau)$ είναι μια συνάρτηση παραθύρου (window function) συνήθως γκαουσιανό και κεντραρισμένο στο 0. Το t είναι η διάρκεια της συνάρτησης παραθύρου όπου έχει την τιμή 1 ενώ πριν και μετά μηδενίζεται. Αυτή εξαρτάται κάθε φορά από το διάστημα στο οποίο το σήμα μπορεί να θεωρηθεί στατικό. Το τ είναι το κέντρο όπου μετατοπίζεται κάθε φορά η w. Συνεπώς ο STFT μοιάζει πολύ με τον FT πολλαπλασιασμένος με την συνάρτηση παραθύρου. Ας δούμε ένα παράδειγμα [18]:

$$x(t) = \begin{cases} \cos(2\pi 100t), & t = 0 - 100\, ms \\ \cos(2\pi 50t), & t = 100 - 200\, ms \\ \cos(2\pi 40t), & t = 200 - 300\, ms \\ \cos(2\pi 10t), & t = 300 - 400\, ms \\ 0, & αλλού \end{cases}$$

Είναι το σήμα μας και η συνάρτηση παραθύρου w θα έχει t=100 ms.

Εφαρμόζοντας τον STFT βλέπουμε ότι θα έχουμε τέσσερις κορυφές για τέσσερις διαφορετικές συχνότητες που εντοπίζονται σε τέσσερα διαφορετικά χρονικά διαστήματα. Συνεπώς γνωρίζουμε τελικά και τις συχνότητες από τις οποίες αποτελείται το σήμα και σε ποια χρονικά διαστήματα αυτές υπάρχουν. Όλα αυτά φαίνονται στο *σχήμα 2.5:*



Σχήμα 2.5: Η 3-D γραφική παράσταση του πλάτους του σήματος στο χρόνο και στη συχνότητα

Φτάσαμε λοιπόν πιο μακριά σε σχέση με τον FT όμως θα δούμε ότι ακόμη και ο STFT παρουσιάζει πρόβλημα για τα μη-στάσιμα σήματα. Ας δούμε ένα δεύτερο παράδειγμα, έστω το σήμα:

$$x(t) = \begin{cases} \cos(2\pi 10t), & t = 0 - 100\ ms \\ \cos(2\pi 30t), & t = 100 - 130\ ms \\ \cos(2\pi 40t), & t = 130 - 140\ ms \\ \cos(2\pi 80t), & t = 140 - 180\ ms \\ 0, & αλλού \end{cases}$$

Τώρα πρέπει να γίνει η επιλογή της συνάρτησης παραθύρου. Θα πάρουμε t = 100 ms ώστε να καλύπτεται μια περίοδος του πρώτου διαστήματος με f_1 = 10 Hz, όμως μετά όπως βλέπουμε, το χρονικό διάστημα όπου επικρατούν οι άλλες

συχνότητες μειώνεται και δεν είναι σταθερό. Συνεπώς θα δημιουργηθεί πρόβλημα παίρνοντας τον STFT των υπόλοιπων συχνοτήτων με την συγκεκριμένη συνάρτηση παραθύρου. Επίσης δεν μπορούμε να πάρουμε μικρότερη χρονική διάρκεια για την *w* διότι θα δημιουργηθεί πρόβλημα με την πρώτη συχνότητα. Επομένως καταλήγουμε στο ίδιο πρόβλημα που είχαμε και πριν με τον FT για τα μη-στάσιμα σήματα.

Επιπλέον με τον STFT παρότι έχουμε αναπαράσταση συχνότητας-χρόνου δεν γνωρίζουμε την ακριβή αντιστοιχία τους, μονάχα σε ποια χρονικά διαστήματα υπάρχουν συγκεκριμένες περιοχές συχνοτήτων, όπως φαίνεται και στο *σχήμα 2.5*. Συνεπώς ο STFT εμφανίζει πρόβλημα ανάλυσης, το οποίο δεν υπήρχε στην ανάλυση στάσιμων σημάτων με τον FT. Αυτό το πρόβλημα προέρχεται από το θεώρημα εύρους ζώνης κατά το οποίο η μικρή διάρκεια (Δt) του σήματος στο πεδίο του χρόνου οδηγεί σε μεγάλο εύρος στο πεδίο συχνοτήτων (Δω). Αναλόγως με την επιλογή της συνάρτησης παραθύρου μεταβάλλεται και η δυνατότητα ανάλυσης, έτσι έχουμε για *w* μικρής χρονικής διάρκειας, καλή ανάλυση στο πεδίο του χρόνου και κακή ανάλυση στο πεδίο της συχνότητας και αντίστοιχα για *w* μεγάλης χρονικής διάρκειας, κακή ανάλυση στο πεδίο του χρόνου και καλή ανάλυση στο πεδίο της συχνότητας. Για αυτούς τους λόγους επομένως στραφήκαμε στα κυματίδια (wavelets).

2.4 Κυματίδια

2.4.1 Βασικά χαρακτηριστικά

Όπως έχουμε αναφέρει ο μετασχηματισμός των wavelets αποτελεί επέκταση του FT και χρησιμοποιείται για να ξεπεραστεί το πρόβλημα που δημιουργείται και από τον STFT στα μη-στάσιμα σήματα. Είναι ένα μαθηματικό εργαλείο που αναπτύχθηκε την δεκαετία του 1980 και έχει ευρύ πεδίο εφαρμογών, όπως στην ανάλυση και συμπίεση δεδομένων, στον θεωρητικό ηλεκτρομαγνητισμό κ.α. Το κυματίδιο (μικρό κύμα) ως ονομασία προκύπτει από την μορφή του που παρουσιάζει κυματισμό με μέση τιμή μηδέν και το «μικρό» προκύπτει από τη συνάρτηση παραθύρου που χρησιμοποιείται, η οποία είναι πεπερασμένης χρονικής διάρκειας.

Η βασική ιδέα του μετασχηματισμού αυτού αφορά στην ανάλυση ενός σήματος στο χρόνο και στη συχνότητα (χρονοσυχνοτική) [20].

2.4.2 Ο συνεχής μετασχηματισμός των κυματιδίων

Ο συνεχής μετασχηματισμός των κυματιδίων ή Continuous Wavelet Transform (CWT) λοιπόν, αναπτύχθηκε ως εναλλακτική προσέγγιση του STFT και η διαφορά με αυτόν είναι ότι η συνάρτηση παραθύρου *w* μπορεί να αλλάζει σε χρονική διάρκεια κατά την ανάλυση του σήματος. Ο CWT ορίζεται ως εξής:

$$CWT_x^{\psi}(\tau, s) = \Psi_x^{\psi}(\tau, s) = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{+\infty} x(t)\psi^* \left(\frac{t-\tau}{s}\right) dt \qquad (2.5)$$

όπου η $\psi^*(t)$ είναι η συνάρτηση παραθύρου και ονομάζεται μητρική (mother wavelet). Από αυτήν μπορούμε να κατασκευάσουμε μια σειρά από συναρτήσεις που χρησιμοποιούνται ώστε να αναλύουμε κάθε φορά ένα συγκεκριμένο κομμάτι του σήματος μας. Ο συντελεστής s συμβολίζει την κλίμακα (scale function) και ισχύει $s = \frac{1}{f}$, είναι το αντίστροφο της συχνότητας. Καθορίζει την διάρκεια του wavelet ψ^* και παρέχει την πληροφορία της συχνότητας για τον CWT. Πιο συγκεκριμένα όταν έχουμε μικρό scale έχουμε μεγάλες τιμές συχνοτήτων και αντίστοιχα για μεγάλα scale έχουμε μικρές συχνότητες. Ο συντελεστής τ είναι η χρονική μετατόπιση (translation) της συνάρτησης $\psi^*(t)$ κατά μήκος του σήματος για την ανάλυσή του. Παρέχει την πληροφορία του χρόνου στον CWT. Ο πολλαπλασιασμός με τον παράγοντα $\frac{1}{\sqrt{|s|}}$ είναι για λόγους κανονικοποίησης της ενέργειας, ώστε για κάθε scale να έχουμε την ίδια ενέργεια στο μετασχηματισμένο σήμα μας.

Η συνάρτηση κλίμακας λειτουργεί όπως ακριβώς και η κλίμακα στους χάρτες. Για μεγάλη τιμή του scale (χαμηλές συχνότητες) έχουμε μια γενική εικόνα του σήματος χωρίς λεπτομέρειες ενώ για μικρή τιμή scale (υψηλές συχνότητες) έχουμε λεπτομερή εικόνα και πληροφορία για το σήμα. Επίσης με την 'κλιμάκωση' (scaling) του σήματος μπορεί το σήμα μας να συμπιεστεί (μικρή κλιμάκωση) ή να διασταλεί (μεγάλη κλιμάκωση).

Οι διάφορες συναρτήσεις παραθύρου που θα έχουμε θα είναι λοιπόν διεσταλμένες ή συμπιεσμένες και μετακινημένες στο χρόνο εκδοχές του μητρικού wavelet και ονομάζονται θυγατρικές (daughter). Υπάρχουν πολλές συναρτήσεις που λειτουργούν ως βάση για το μητρικό wavelet, όπως το Morlet wavelet (που χρησιμοποιείται στα παρακάτω σχήματα), το Mexican hat και άλλα.

Όπως και στον FT υπάρχει και ο αντίστροφος μετασχηματισμός κυματιδιών, όπου για τον ICWT (Inverse CWT) έχουμε την ανακατασκευή του σήματος:

$$x(t) = \frac{1}{C_\psi^2} \iint_{-\infty}^{+\infty} \Psi_x^\psi(\tau, s) \frac{1}{s^2} \psi\left(\frac{t-\tau}{s}\right) d\tau ds \quad (2.6)$$

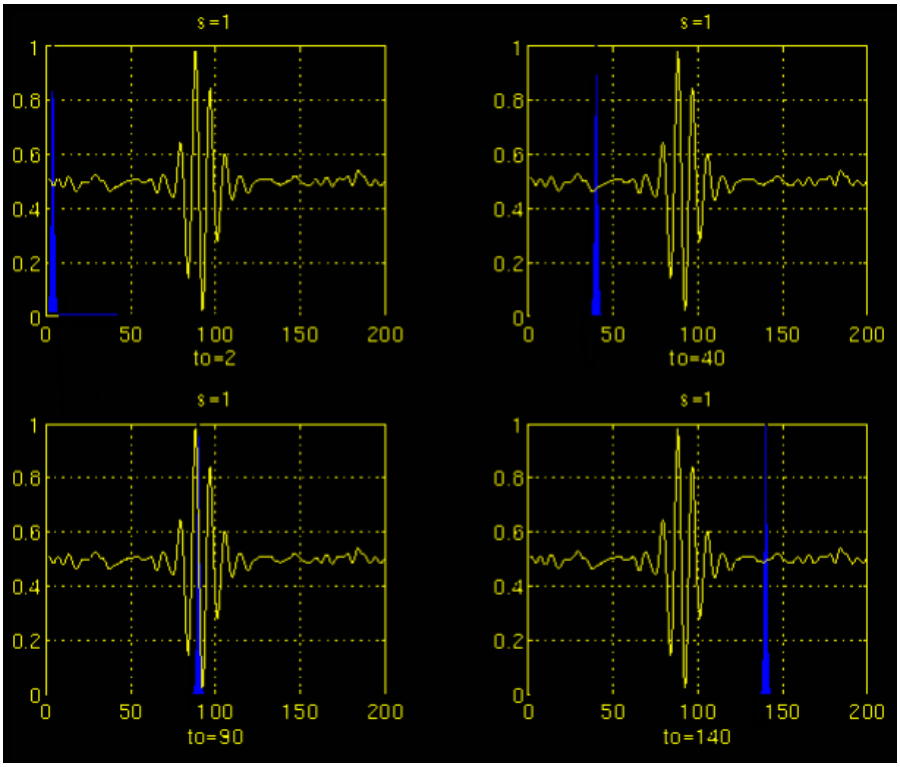
όπου το C_ψ είναι σταθερά κανονικοποίησης που εξαρτάται από το wavelet που χρησιμοποιούμε κάθε φορά. Έχουμε:

$$C_\psi = \left\{2\pi \int_{-\infty}^{+\infty} \frac{|\hat{\psi}(\xi)|^2}{\xi} d\xi\right\}^{\frac{1}{2}} < \infty \quad (2.7)$$

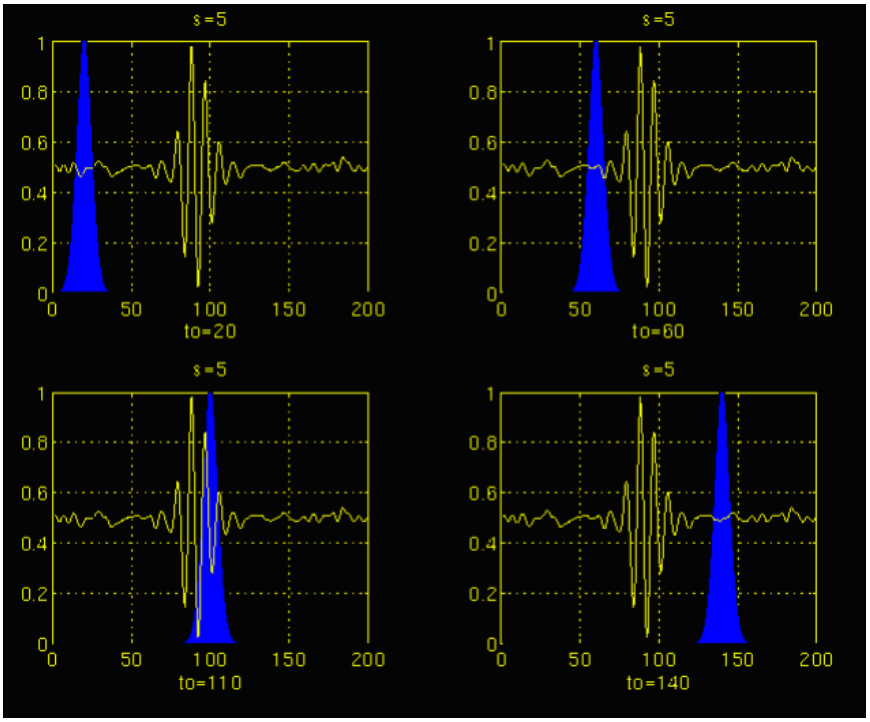
όπου το $\hat{\psi}(\xi)$ είναι ο FT του ψ(t) με $|\hat{\psi}(0)| = 0$.

Σύμφωνα με τον ορισμό του CWT, επιλέγουμε την κατάλληλη μητρικη συνάρτηση για την ανάλυσή μας και λαμβάνουμε αρχικά την τιμή $s = 1$ και θα υπολογίζουμε τον μετασχηματισμό σε όλο το σήμα για όλες τις μεγαλύτερες και μικρότερες τιμές του scale. Στο παράδειγμα που θα δούμε θα κάνουμε μόνο για τις μεγαλύτερες τιμές του scale άρα για $s > 1$, άρα θα μελετήσουμε το σήμα από τις υψηλές προς τις χαμηλές συχνότητες. Ξεκινάμε από την αρχή του σήματος άρα για $t = 0$ και αφού υπολογίσουμε την πρώτη τιμή δηλαδή για $CWT(\tau = 0, s = 1)$ μετατοπίζουμε το wavelet με $s = 1$ πάνω στην συνάρτηση του σήματος κατά $t = \tau$. Συνεχίζουμε την ίδια διαδικασία μέχρι να φτάσουμε στο τέλος του σήματος. Στη συνέχεια αυξάνουμε το s κατά λίγο και ξεκινάμε για το καινούριο αυτό scale την ίδια διαδικασία από την αρχή του σήματος. Αφού ολοκληρωθεί η διαδικασία για όλες τις τιμές του scale που θέλουμε τότε παίρνουμε τον CWT του σήματος. Αυτή η διαδικασία λοιπόν φαίνεται αναλυτικά στα παρακάτω σχήματα.

Σε κάθε χρονική τοποθεσία του παραθύρου πολλαπλασιάζουμε με το σήμα. Μετακινώντας το wavelet στο χρόνο εντοπίζουμε την χρονική πληροφορία για το σήμα μας και αλλάζοντας την κλίμακα εντοπίζουμε την πληροφορία συχνότητας για αυτό. Στο *σχήμα 2.6* φαίνεται ότι το σήμα έχει συχνότητες που εμφανίζονται σε παρόμοιο χρονικό εύρος (width) με αυτό της συνάρτησης παραθύρου με s=1 για χρόνο t=100 ms. Συνεπώς η τιμή του $CWT(t, s = 1)$ για αυτόν τον χρόνο θα είναι υψηλή ενώ στους υπόλοιπους χρόνους θα είναι χαμηλή ή μηδενική. Αν τώρα αυξήσουμε το scale το παράθυρό μας γίνεται πιο ευρύ και θα παρατηρήσουμε, σε αντίθεση με πριν, ότι γενικότερα για πιο υψηλές τιμές του scale θα έχουμε και υψηλές τιμές στον CWT σχεδόν σε όλο το υπόλοιπο σήμα και όχι για t=100 ms, διότι στο σήμα μας επικρατούν χαμηλές συχνότητες σε όλους τους χρόνους.

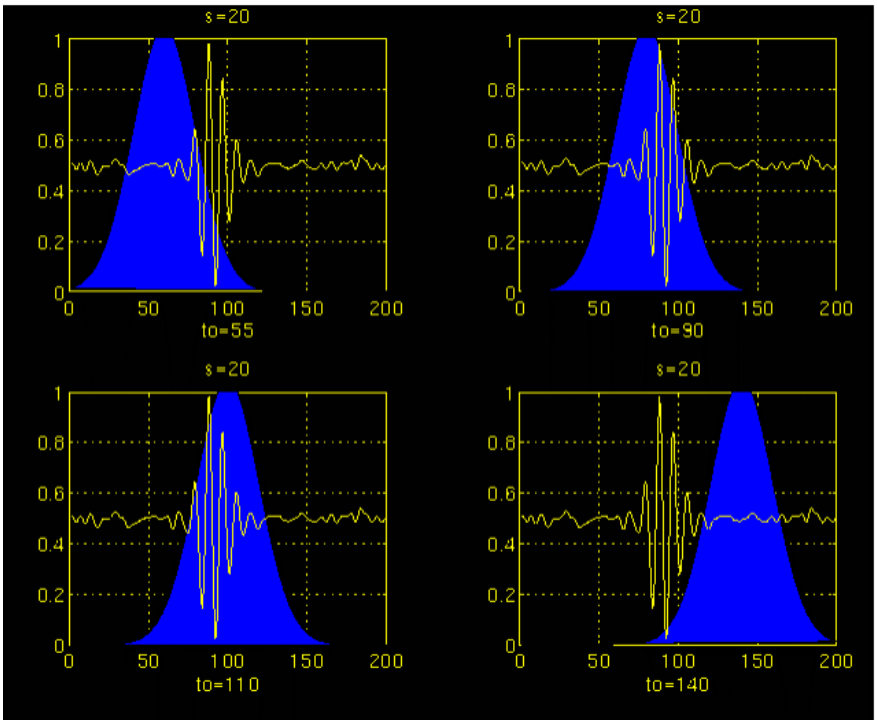


Σχήμα 2.6: Η ανάλυση του σήματός μας για s=1, την πιο μικρή τιμή της κλίμακας και άρα την πιο μεγάλη τιμή της συχνότητας. Με μπλε βλέπουμε το παράθυρο του wavelet που λόγω του μικρού scale είναι πολύ συμπιεσμένο (όσο στενή είναι δηλαδή η αναπαράσταση της μεγαλύτερης συχνότητας στο σήμα) . Το παράθυρο μετακινείται για 4 διαφορετικές χρονικές τιμές του τ.



Σχήμα 2.7: Ανάλυση του σήματος για s=5, το παράθυρο είναι πιο φαρδύ από πριν.

Μετακινείται πάλι για τέσσερεις χρονικές στιγμές του τ στο σήμα.



Σχήμα 2.8: Ανάλυση του σήματος με s=20 για 4 διαφορετικές χρονικές στιγμές τ του σήματος.

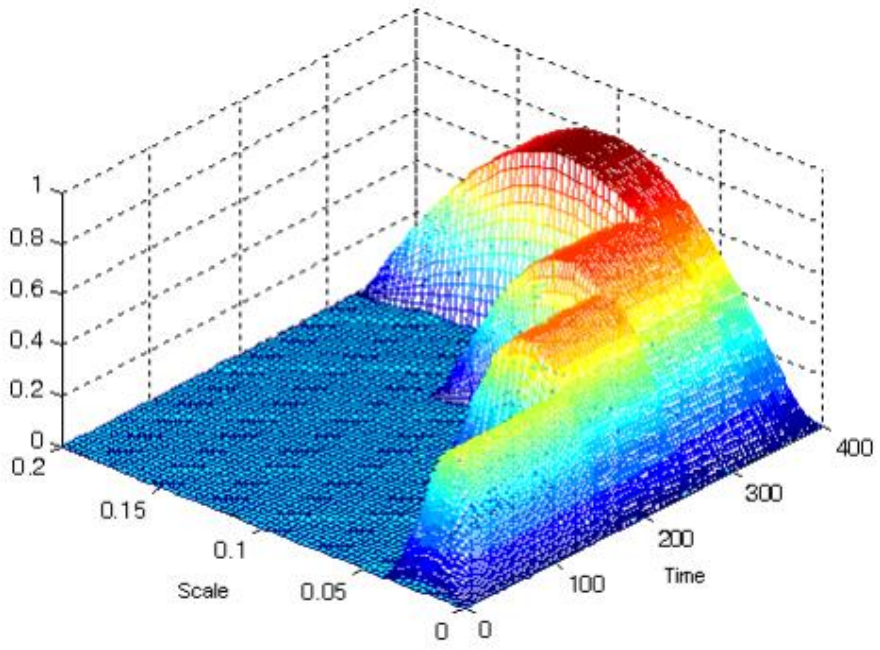
Όπως είπαμε λοιπόν και πριν όσο αυξάνεται το εύρος του παραθύρου ο μετασχηματισμός «αναγνωρίζει» τις πιο χαμηλές συχνότητες του σήματος. Ας δούμε άλλο ένα παράδειγμα [18]. Έχουμε το μη-στάσιμο σήμα:

$$x(t) = \begin{cases} \sin(2\pi 40t), & t = 0 - 250\ ms \\ \sin(2\pi 30t), & t = 250 - 500\ ms \\ \sin(2\pi 20t), & t = 500 - 750\ ms \\ \sin(2\pi 10t), & t = 750 - 1000\ ms \\ 0, & αλλού \end{cases}$$

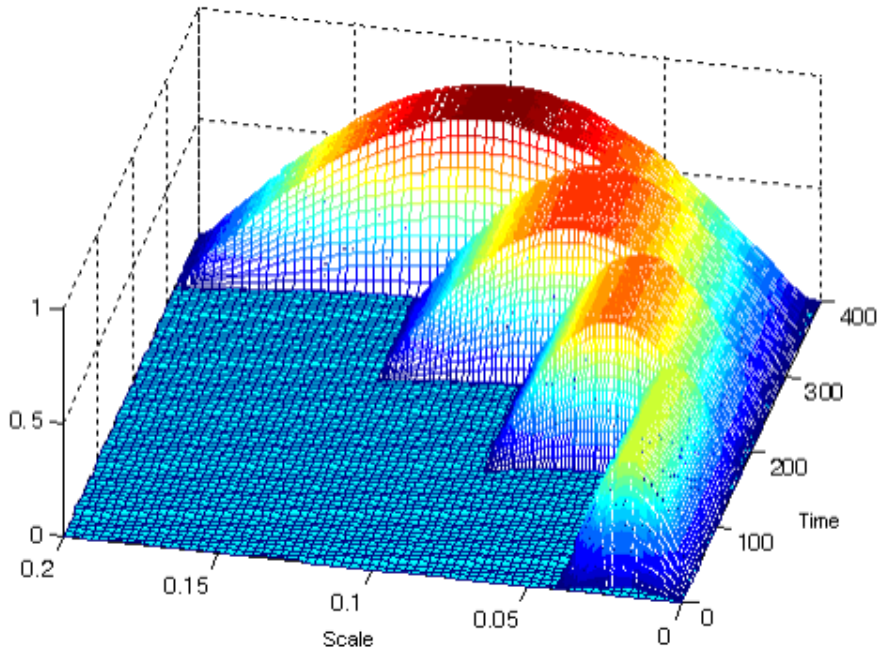
Ως μητρικό wavelet εδώ μπορούμε να χρησιμοποιήσουμε τη συνάρτηση

$$\psi(t) = \begin{cases} \cos(\pi t), -1 \le t \le 1 \\ 0, \quad αλλού \end{cases}$$

Εφαρμόζουμε τον CWT στο σήμα μας και κάνουμε την γραφική παράσταση του αποτελέσματος όπου στην τρισδιάστατη απεικόνιση οι άξονες αντιστοιχούν στο πλάτος ως προς την μετατόπιση στο χρόνο και την scale. Τα σχήματα που προκύπτουν φαίνονται παρακάτω.



Σχήμα 2.9: Αναπαράσταση του σήματος μετά από εφαρμογή CWT. Θυμόμαστε ότι στα χαμηλά scales έχουμε υψηλές συχνότητες άρα στην περιοχή κοντά στο 0 αντιστοιχεί η f_1=40Hz και όσο κινούμαστε στον άξονα του scale θα έχουμε μικρότερες συχνότητες

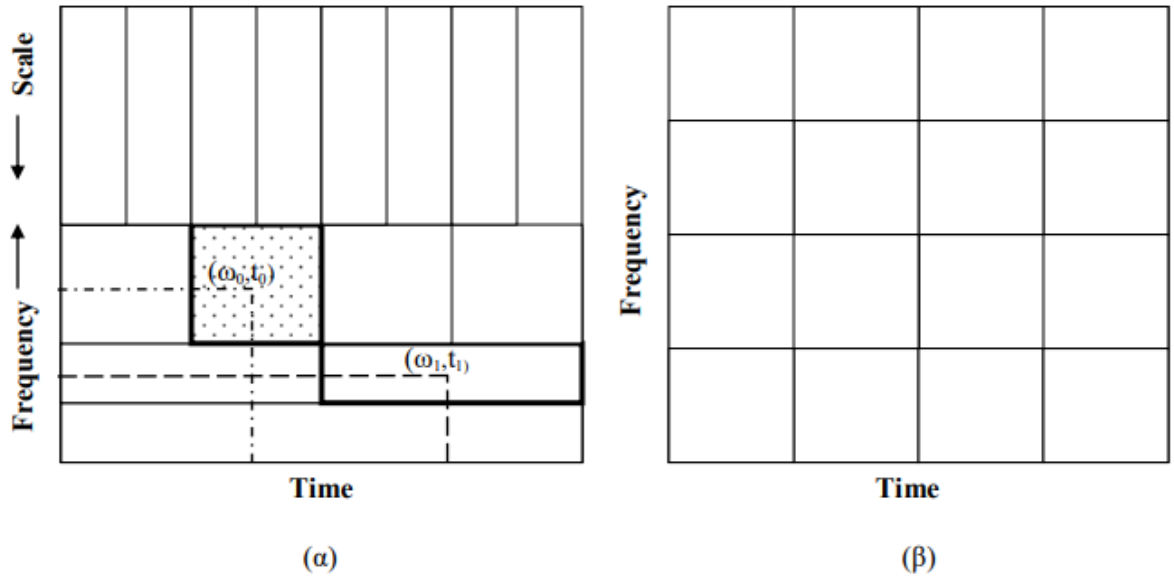


Σχήμα 2.10: Διαφορετική οπτική γωνία του *σχήματος 2.9* ώστε να φαίνεται καλύτερα ο άξονας του scale

Από τα δύο αυτά σχήματα προκύπτει ότι για χαμηλά scales και άρα υψηλές συχνότητες έχουμε καλή ανάλυση χρόνου και κακή ανάλυση συχνότητας, ενώ για υψηλά scales και άρα χαμηλές συχνότητες έχουμε χαμηλή ανάλυση χρόνου και υψηλή ανάλυση συχνότητας σε αντίθεση με τον STFT όπου η ανάλυση ήταν σταθερή ανεξαρτήτως του χρόνου και της συχνότητας. Πιο συγκεκριμένα όπως φαίνεται στο *σχήμα 2.10* για χαμηλά scales και άρα υψηλές συχνότητες φαίνεται να έχουμε καλή ανάλυση κλίμακας (εφόσον το διάστημα είναι στενό) και άρα η συχνότητα που είναι το αντίστροφο θα είναι πιο ασαφής και αντίστοιχα για τα υψηλά scales και συνεπώς χαμηλές συχνότητες.

2.4.3 Ανάλυση χρόνου-συχνότητας

Η ανάλυση χρόνου-συχνότητας θα φανεί καλύτερα στο παρακάτω σχήμα. Να θυμίσουμε από την *ενότητα 2.3* για τον STFT ότι δεν γνωρίζουμε για μια συγκεκριμένη χρονική στιγμή την ακριβή τιμή της συχνότητας αλλά ένα εύρος συχνοτήτων που αντιστοιχεί σε μια συγκεκριμένη χρονική περίοδο. Για αυτόν τον λόγο επίσης στραφήκαμε προς τον μετασχηματισμό των κυματιδίων.



Σχήμα 2.11: Στο (β) η ανάλυση χρόνου-συχνότητας για τον STFT. Στο (α) η ανάλυση χρόνου-συχνότητας ή κλίμακας για τον WT. Όσο πιο στενό το πλάτος του κάθε κουτιού τόσο καλύτερη είναι η ανάλυση, είτε χρόνου είτε συχνότητας

Στο *σχήμα 2.11* θεωρούμε ότι όσο πιο στενό είναι το κουτί τότε τόσο καλύτερη

είναι η ανάλυση. Βλέπουμε για τον STFT στο *(β)* ότι η ανάλυση χρόνου-συχνότητας είναι ίδια παντού ενώ για τον WT στο *(α)* έχουμε καλή ανάλυση συχνότητας για χαμηλές συχνότητες ή υψηλές κλίμακες (το κουτί είναι στενό στον άξονα της συχνότητας) αλλά κακή ανάλυση χρόνου (το κουτί είναι φαρδύ στον άξονα χρόνου) και αντίστροφα για υψηλές συχνότητες βλέπουμε κακή ανάλυση συχνότητας και καλή ανάλυση χρόνου.

Όσο αυξάνεται δηλαδή η συχνότητα (ή μειώνεται η κλίμακα) η ανάλυση του χρόνου βελτιώνεται και η ανάλυση συχνότητας χειροτερεύει. Το κέντρο κάθε κουτιού δηλώνει την θέση μιας συνάρτησης wavelet $\psi(t)$ ενώ το ίδιο το κουτί είναι το εύρος ανάλυσης του wavelet σε χρόνο και συχνότητα.

2.4.4 Ο μαθηματικός φορμαλισμός των κυματιδίων

Θα πρέπει να ισχύουν οι ακόλουθες ιδιότητες ώστε μια συνάρτηση $\psi(x)$ να είναι wavelet συνάρτηση [20].

- $\int \psi(x)dx = 0$, η μέση τιμή του σήματος είναι 0. Κατ επέκταση και $\int x^n\psi(x)dx = 0$ ώστε να μηδενίζονται όσο περισσότερες ροπές είναι δυνατόν, εδώ τάξης-n.
- $\|\psi(x)\| = 1$, η ενέργεια του σήματος είναι κανονικοποιημένη στη μονάδα
- Είναι κεντραρισμένη στη γειτονία του *x=0* και ισχύει και για το μητρικό wavelet και για τα θυγατρικά

Επίσης υπάρχει πεπερασμένο διάστημα [a,b] έτσι ώστε η wavelet συνάρτηση $\psi(x)$ να μηδενίζεται για όλα τα *x* που δεν ανήκουν στο διάστημα αυτό.
Εισάγουμε τώρα την έννοια της συνάρτησης κλίμακας (scaling) $\varphi(x)$ ή όπως αλλιώς ονομάζεται πατρικό κυματίδιο (father wavelet) το οποίο έχει άμεση σχέση με το mother wavelet. Από την $\varphi(x)$ μπορούμε να δημιουργήσουμε μια συνάρτηση wavelet ως εξής:
Πρέπει να ικανοποιείται η σχέση $\varphi(x) = \sum_{i=0}^{n} h_i\varphi(2x - i)$, όπου τα h_i είναι οι συντελεστές της scaling συνάρτησης, με *i=0,1,…,n.*
Άρα προκύπτει η wavelet συνάρτηση $\psi(x) = \sum_{i=0}^{n} w_i\varphi(2x - i)$, όπου w_i είναι οι συντελεστές της wavelet συνάρτησης.
Οι συναρτήσεις φ και ψ είναι ορθογώνιες μεταξύ τους και άρα ισχύει $< \psi, \varphi >= 0$. Με συνδυασμό των θυγατρικών wavelet των φ και ψ, οι οποίες έχουν μετατοπιστεί χρονικά και η διάρκεια τους έχει μεταβληθεί, μπορούμε να αναλύσουμε ένα σήμα ως εξής:

$$x(t) = \sum_{k=-\infty}^{+\infty} c_{m_0,k}\varphi_{m_0,k}(t) + \sum_{k=-\infty}^{+\infty}\sum_{m=m_0}^{+\infty} d_{m,k}\psi_{m,k}(t)$$

με $m \geq m_0$. Αυτή είναι η ανάλυση πολλαπλών επιπέδων (Multi-Resolution Analysis: MRA) με $c_{m_0,k} = \int_{-\infty}^{+\infty} x(t)\varphi_{m_0,k}(t)dt$ που είναι οι συντελεστές 'κατά προσέγγιση' και $d_{m,k} = \int_{-\infty}^{+\infty} x(t)\psi_{m,k}(t)dt$ που είναι οι 'λεπτομερείς' συντελεστές.

Ο WT που είδαμε και πριν από την (2.5) μπορεί να γραφτεί σαν μια πράξη συνέλιξης (convolution) του σήματος x με το wavelet ψ:

$$\Psi(\tau, s) = x * \bar{\psi}_s(\tau)$$

όπου $\bar{\psi}_s(\tau) = \frac{1}{\sqrt{s}}\psi^*(\frac{-t}{s})$

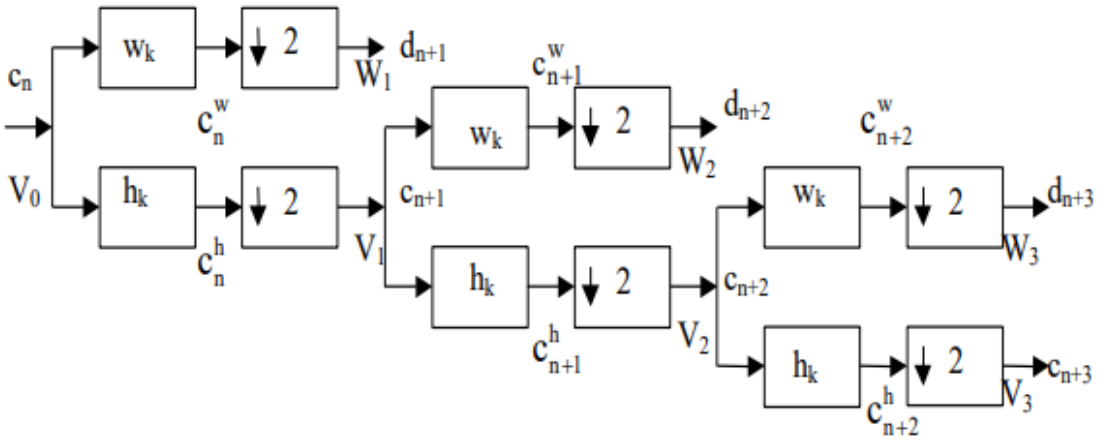
2.4.5 Ο διακριτός μετασχηματισμός wavelet

Η βασική ιδέα του DWT είναι ίδια με του CWT. Ο DWT λοιπόν είναι ένας γρήγορος αλγόριθμος που μετατρέπει ένα σήμα σε σύνολο wavelet συντελεστών, συνεπώς είναι η ανάλυση ενός σήματος μέσω WT που υλοποιείται με κώδικα στον Η/Υ.

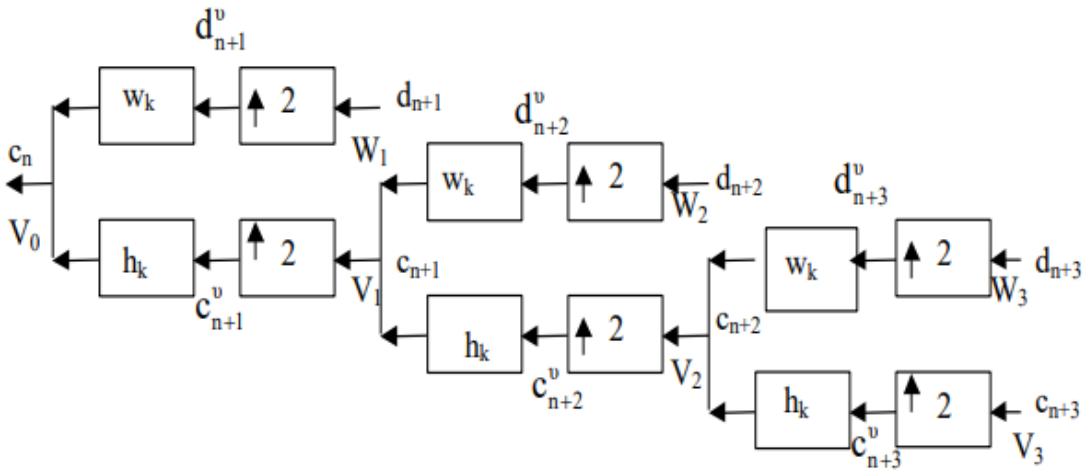
Στην *ενότητα 2.4.2* για τον υπολογισμό του CWT είχαμε δει ότι αλλάζουμε την κλίμακα (scale) του wavelet ψ και το μετακινούμε στον άξονα του χρόνου και στη συνέχεια υπολογίζουμε το εσωτερικό γινόμενο του ψ με το σήμα *x* και παίρνουμε το ολοκλήρωμα αυτού του γινομένου. Μέσω του DWT τώρα θα δούμε ότι μπορούμε να υπολογίσουμε το ίδιο αποτέλεσμα χωρίς να χρειαστούμε ολοκληρώματα. Οι συντελεστές wavelet συνεπώς θα προκύψουν με μια πράξη συνέλιξης μεταξύ του σήματος μας και των φίλτρων (υψηλής ή χαμηλής διέλευσης), τα οποία είναι συναρτήσεις scaling και wavelet. Έχουμε λοιπόν το αρχικό διακριτό σήμα c_n που φιλτράρεται από 2 φίλτρα ώστε στην έξοδο να έχουμε 2 σήματα τελικά. Με αυτόν τον τρόπο όμως καταλήγουμε με διπλάσιο αριθμό δειγμάτων από αυτόν που ξεκινήσαμε εφόσον τα 2 προκύπτοντα σήματα θα έχουν τον ίδιο αριθμό (περίπου) δειγμάτων με το αρχικό μας σήμα c_n, ανάλογα με το φίλτρο που έχουμε επιλέξει. Για την ανα-σύνθεση του αρχικού μας σήματος, εφόσον έχουμε περισσότερες πληροφορίες από όσες χρειαζόμαστε, θα απορρίψουμε επιλεκτικά ορισμένο αριθμό

δειγμάτων χωρίς να επηρεαστεί η ποιότητα ανάλυσης. Συνεπώς μειώνεται ο αριθμός των δειγμάτων του αρχικού σήματος.

Στη συνέχεια στο *σχήμα 2.12* θα δούμε την από-σύνθεση («αποσυνέλιξη») ενός σήματος ώστε να προκύψουν οι συντελεστές wavelet [18]. Χρειαζόμαστε τρεις παράγοντες: ένα φίλτρο εύρους w για ανάλυση των υψηλών συχνοτήτων, ένα φίλτρο εύρους *h* για ανάλυση των χαμηλών συχνοτήτων και την δυαδική υποδειγματοληψία (dyadic downsampling). Με τον όρο αυτόν εννοούμε την απόρριψη των ζυγών δειγμάτων από την ακολουθία δειγμάτων ενός σήματος. Υποδειγματοληψία κατά *n*, γενικότερα, σημαίνει ότι κάθε *n*-ιοστό δείγμα του σήματός μας απορρίπτεται. Αντίστοιχα υπάρχει και η υπερδειγματοληψία κατά *n* όπου εκεί προσθέτουμε ένα νέο δείγμα, συνήθως το μηδενικό (n=0), σε κάθε *n*-ιοστό δείγμα του σήματός μας. Αυτή χρησιμοποιείται κατά την ανα-σύνθεση ενός σήματος.



Σχήμα 2.12: Από-σύνθεση σήματος μέσω του DWT. Το σήμα χωρίζεται σε δύο σήματα αντίστοιχα και αυτό που φιλτραρίστηκε με το χαμηλής διέλευσης (low pass) φίλτρο υποβάλλεται ξανά στην ίδια διαδικασία. Προκύπτουν έτσι οι συντελεστές wavelet



Σχήμα 2.13: Ανα-σύνθεση του σήματος μέσω του DWT.

Τα σύνολα των wavelet συντελεστών είναι τα c_n, c_{n+1}, c_{n+2} κ.ο.κ. και προκύπτουν από:

$$c_{m,k} = \sum_n h_{n-2k} c_{m-1,k}\,,$$ οι χαμηλής διέλευσης και

$$d_{m,k} = \sum_n w_{n-2k} c_{m-1,k}\,,$$ οι υψηλής διέλευσης

Από την άλλη, για την ανα-σύνθεση του σήματος θα έχουμε:

$$c_{m-1,k} = \sum_n h_{k-2n} c_{m,n} + \sum_n w_{k-2n} d_{m,n}$$

Ο δείκτης *m* δηλώνει το επίπεδο ανάλυσης και ο δείκτης *k* τον αύξοντα αριθμό του στοιχείου. Το φίλτρο υψηλής διέλευσης *w* αποκόπτει όλες τις συχνότητες που είναι μικρότερες από το μισό της μέγιστης συχνότητας του σήματος, ενώ το φίλτρο χαμηλής διέλευσης *h* αποκόπτει όλες τις συχνότητες που είναι μεγαλύτερες από το μισό της μέγιστης συχνότητας του σήματος. Τελικά ο DWT διαχωρίζει το σήμα σε ένα σήμα χαμηλών συχνοτήτων, που ονομάζεται σύνολο προσεγγιστικών συντελεστών, και σε ένα σήμα υψηλών συχνοτήτων, που ονομάζεται σύνολο λεπτομερών συντελεστών, όπως ακριβώς είπαμε προηγουμένως στην ανάλυση πολλαπλών επιπέδων (MRA). Χρησιμοποιώντας τα φίλτρα μεταβάλλουμε την ανάλυση του σήματος ενώ με την δειγματοληψία αλλάζουμε την κλίμακα (scale) του σήματος. Με το φίλτρο *h* απορρίπτονται οι μισές συχνότητες του σήματος και άρα θεωρούμε ότι μειώνεται η ανάλυση του, εφόσον χάνονται πληροφορίες για αυτό. Όμως, κάνοντας υποδειγματοληψία στο σήμα, μετά την χρήση του φίλτρου, μεγαλώνει η κλίμακα και δεν αλλάζει η

ανάλυση του σήματος, εφόσον τα δείγματα που αποκόπτονται είναι πλεονάζοντα.

Συνεπώς βλέπουμε όλη την διαδικασία που περιγράψαμε με ένα παράδειγμα. Έχουμε ένα σήμα $c_0(n)$ με 512 διακριτά δείγματα και μέγιστη συχνότητα 1000Hz. Κάνουμε μια πρώτη απο-σύνθεση στο σήμα και προκύπτουν τα σήματα $c_1(n)$ και $d_1(n)$, τα οποία αποτελούνται το καθένα από 256 στοιχεία. Το $c_1(n)$ θα περιέχει τις συχνότητες 0-500Hz και το $d_1(n)$ τις συχνότητες 500-1000 Hz. Αυτή είναι η απο-σύνθεση πρώτου επιπέδου (first level decomposition). Στη συνέχεια, περνάμε το σήμα $c_1(n)$ από 2 φίλτρα και προκύπτουν 2 νέα σήματα, τα $c_2(n)$ και $d_2(n)$, με 125 στοιχεία. Το πρώτο θα περιέχει τις συχνότητες 0-250 Hz και το δεύτερο τις συχνότητες 250-500 Hz. Αυτή είναι η απο-σύνθεση δευτέρου επιπέδου και η διαδικασία συνεχίζεται μέχρι να φτάσουμε σε 2 σήματα που περιέχουν μονάχα ένα στοιχείο. Για την ανα-σύνθεση του σήματος συνδυάζουμε τα 2 σήματα c_n και d_n ώστε να αποκτήσουμε το σήμα του αμέσως ανωτέρου επιπέδου c_{n-1}.

Η από-σύνθεση μειώνει στο μισό την ανάλυση χρόνου αφού κάθε φορά έχουμε τα μισά δείγματα στο επόμενο επίπεδο, όμως η ανάλυση συχνότητας καλυτερεύει εφόσον το εύρος συχνοτήτων είναι το μισό από το αρχικό (πιο εστιασμένη ανάλυση). Συνεπώς έχουμε καλή ανάλυση χρόνου στις υψηλές συχνότητες με κακή ανάλυση συχνότητας και κακή ανάλυση χρόνου στις χαμηλές συχνότητες με καλή ανάλυση συχνότητας.

Τα wavelets είναι χρήσιμα σε ευρύ πεδίο εφαρμογών, όπως για την αφαίρεση του θορύβου σε ένα σήμα ή μια εικόνα (denoising), τη συμπίεση και την επανασύνθεση εικόνων , αλλά και επίσης στον χώρο της βασικής πειραματικής φυσικής με ενδιαφέρουσες εφαρμογές [32], [33]. Για παράδειγμα οι αστρονόμοι χρησιμοποιούν τα wavelets για την ανάλυση εικόνων ώστε να εξάγουν «λεπτές» λεπτομέρειες π.χ. δακτύλιος του Αϊνστάιν (Einstein rings). Μπορούν να χρησιμοποιηθούν επίσης ως η βάση ενός αλγορίθμου συμπίεσης, όπως το JPEG 2000. Επιπλέον ο κοχλίας του αυτιού είναι σχεδιασμένος έτσι ώστε να εκτελεί έναν μετασχηματισμό wavelet του ήχου. Άλλη μια εφαρμογή των wavelets είναι στην απόσύνθεση (decomposition) της Κοσμικής Ακτινοβολίας Υποβάθρου (CMBR). Επίσης στο χρηματιστήριο φημολογείται ότι χρησιμοποιούν fractals, συνεπώς η ανάλυση μέσω wavelets είναι χρήσιμη, εφόσον μέσω των wavelets γίνεται ανίχνευση ομοιοτήτων και fractals.

Κεφάλαιο 3

Ο μετασχηματισμός διασκορπισμού κυματιδίων

3.1 Γενικά στοιχεία

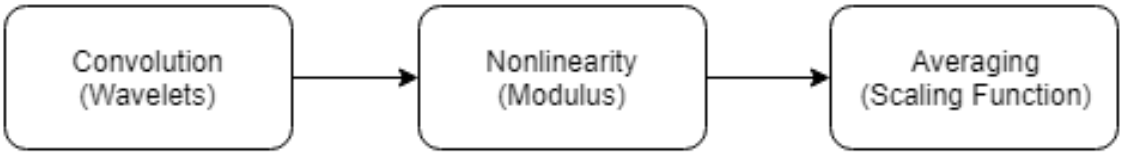
Σε αυτό το κεφάλαιο παρουσιάζεται ο μετασχηματισμός διασκορπισμού κυματιδίων (wavelet scattering transform: WST). Η πληροφορία του ήχου συνήθως δεν επηρεάζεται από την μεταφορά (translation και είναι σταθερή υπό την επίδραση μικρών διαφορομορφισμών που μπορούν να παραμορφώσουν τα σήματα. Ένας διαφορομορφισμός είναι μια αντιστρέψιμη συνάρτηση που αντιστοιχίζει μια διαφορίσιμη πολλαπλότητα σε μια άλλη έτσι ώστε τόσο η συνάρτηση όσο και το αντίστροφό της να είναι διαφορίσιμα. Δηλαδή, αν έχουμε δύο πολλαπλότητες M και N, μια διαφορίσιμη αντιστοίχηση $f: M \rightarrow N$ ονομάζεται διαφορομορφισμός αν είναι ένα προς ένα και το αντίστροφό της $f^{-1}: N \rightarrow M$ είναι επίσης διαφορίσιμο. Συνεπώς αναζητούμε αμετάβλητες προς την μεταφορά αναπαραστάσεις συναρτήσεων $L^2(R^d)$ που να είναι και συνεχείς στους διαφορομορφισμούς.

Συχνά στις υψηλές συχνότητες μπορούν να δημιουργηθούν αστάθειες στις παραμορφώσεις. Αυτές μπορούν να αποφθεχθούν αν ομαδοποιήσουμε τις συχνότητες σε δυαδικά πακέτα στον R^d με τον μετασχηματισμό wavelet. Όμως ο WT (Wavelet Transform) δεν είναι αμετάβλητος ως προς τη μεταφορά. Για να δημιουργήσουμε έναν τέτοιο τελεστή (operator) χρησιμοποιούμε την διαδικασία του scattering σε πολλαπλά μονοπάτια (paths), που να ικανοποιούν την σταθερότητα των κυματιδίων προς τους διαφορομορφισμούς [2].
Για το wavelet scattering μετασχηματισμό έχουμε:

$$\left| x * \psi_j \right| * \varphi_J(t) \tag{3.1}$$

Όπου:
x : δεδομένα
ψ_j : κυματίδια
φ_J : scaling συνάρτηση

Συνεπώς ο WST (Wavelet Scattering Transform) επεξεργάζεται τα δεδομένα σε διαφορετικά στάδια, τα οποία είναι άμεσα συνδεδεμένα μεταξύ τους. Το output του ενός γίνεται input του επόμενου και το κάθε στάδιο αποτελείται από τρεις πράξεις:



Σχήμα 3.1 : Οι 3 πράξεις που αποτελούν τον WST. Είναι το μέτρο (modulus) της συνέλιξης (convolution) του σήματος με τα wavelets και στη συνέχεια ο μέσος όρος ή ξανά συνέλιξη με την scaling συνάρτηση.

Για την πράξη της συνέλιξης (convolution) σε συνεχείς συναρτήσεις έχουμε:

$$(f * g)(t) = \int_{-\infty}^{+\infty} f(\tau) g(t-\tau) d\tau \tag{3.2}$$

Σε διακριτές συναρτήσεις είναι το άθροισμα:

$$(f * g)(n) = \sum_{-\infty}^{\infty} f(m) g(n-m) \tag{3.3}$$

Ένα κυματίδιο λειτουργεί στην πράξη της συνέλιξης όπως ακριβώς τα φίλτρα. Όμως εφόσον ο τελεστής wavelet μετακινείται με την μεταφορά (translation) η τελική απεικόνιση εξαρτάται από αυτήν, άρα η μετατόπιση του σήματος μετατοπίζει και τους wavelet συντελεστές. Συνεπώς η σύγκριση μεταξύ μετατοπισμένων σημάτων είναι δύσκολη και όσον αφορά την κατηγοριοποίηση, που μας αφορά στην μελέτη αυτή, είναι σημαντικό να έχουμε αμετάβλητες προς τη μεταφορά αναπαραστάσεις συναρτήσεων.

Η θεμελιώδης μητρική συνάρτηση που χρησιμοποιεί ο wavelet scattering μετασχηματισμός είναι το Morlet wavelet το οποίο είναι μιγαδική συνάρτηση [34]. Το complex modulus είναι μια μη-γραμμικότητα που εφαρμόζεται στους scattering συντελεστές που προκύπτουν, όπως θα δούμε παρακάτω, και τους

κάνει σταθερούς στους διαφορομορφισμούς και επίσης στην Ευκλίδεια μετρική L^2.

Τώρα θα δούμε τι είναι ο κινούμενος μέσος όρος (moving average) [14]. Είναι μια μορφή συνέλιξης που χρησιμοποιείται συχνά στην ανάλυση χρονικών σειρών για την εξομάλυνση του θορύβου στα δεδομένα (data). Αυτό γίνεται αντικαθιστώντας σε ένα κινούμενο παράθυρο (moving window) ένα σημείο ή μικρή περιοχή των δεδομένων με τον μέσο όρο των γειτονικών τιμών. Συνήθως ο κινούμενος μέσος όρος είναι ένα χαμηλής διέλευσης (low-pass) φίλτρο που αφαιρεί τις βραχυχρόνιες διακυμάνσεις σε ένα σήμα, ώστε να φαίνεται αυτό πιο καθαρό. Στην δική μας περίπτωση είναι το scaling φίλτρο. Ο τύπος για το moving average είναι:

$$\overline{x_i} = \frac{1}{2M+1}\sum_{j=-M}^{j=M} x[i+j] \qquad (3.4)$$

όπου υπολογίζεται ο μέσος όρος των σημείων μέσα στο κινούμενο παράθυρο που έχει διάστημα *[-M,+M]* και κέντρο x_i. Συνεπώς από το *σχήμα 3.1* όταν αναφέρεται στο averaging με την scaling συνάρτηση εννοεί ότι γίνεται συνέλιξη με αυτήν. Όποτε εφαρμόζεται η συνέλιξη στους συντελεστές wavelet με το φίλτρο φ αποκόπτονται οι υψηλές συχνότητες με σκοπό να εντοπιστεί το μεσοδιάστημα (bin) που καλύπτει το χαμηλότερο φάσμα. Οι υψηλές συχνότητες ανακτώνται όπως θα δούμε στην συνέχεια.

3.2 Μέθοδος

Πάμε να δημιουργήσουμε τον μετασχηματισμό scattering από την αρχή [1]. Το κυματίδιο $\psi(t)$ είναι ένα band-pass φίλτρο (αφήνει να περάσει συγκεκριμένο εύρος συχνοτήτων και τις υπόλοιπες τις απορρίπτει) με $\psi(0) = 0$. Για κάθε $\lambda > 0$, μπορούμε να γράψουμε ένα διεσταλμένο κυματίδιο με κεντρική συχνότητα λ στην μορφή: $\psi_\lambda(t) = \lambda\psi(\lambda t)$. (3.5)

Συμβολίζουμε με Q τον αριθμό των κυματιδίων σε κάθε οκτάβα και έχουμε $\lambda = 2^{\frac{k}{Q}}$, με $k \in Z$. Για τον ήχο θέλουμε καλή συχνότητα ανάλυσης (frequency resolution) άρα $\lambda \leq 2^{\frac{1}{8}}$.

Η ενέργεια του $\psi_\lambda(t)$ είναι συγκεντρωμένη γύρω από το 0, σε ένα διάστημα μήκους $\frac{2\pi Q}{\lambda}$. Η (3.5) ισχύει για $\lambda \geq \frac{2\pi Q}{T}$, διότι το μήκος του διαστήματος

πρέπει να είναι μικρότερο από την περίοδο *T*. Για $\lambda < \frac{2\pi Q}{T}$, δηλαδή το διάστημα χαμηλότερης συχνότητας $[0, \frac{2\pi Q}{T}]$ έχει *Q – 1* φίλτρα ψ_λ που ισαπέχουν μεταξύ τους με σταθερό εύρος συχνότητας $\frac{2\pi}{T}$. Τα χαμηλής συχνότητας φίλτρα ονομάζονται επίσης κυματίδια, για διευκόλυνση.

Άρα ο μετασχηματισμός wavelet είναι μια συνέλιξη του σήματος *x* με το φίλτρο χαμηλής διέλευσης *φ*, εύρους ζώνης συχνότητας $\frac{2\pi}{T}$ και στη συνέχεια ξανά συνελίξεις με όλα τα υψηλής συχνότητας κυματίδια ψ_λ για κάθε $\lambda \in \Lambda$. Το Λ είναι το πλέγμα (grid) όλων των κεντρικών συχνοτήτων λ των κυματιδίων.

$$Wx = (x * \varphi(t) , x * \psi_\lambda(t)) \qquad (3.6)$$

με $t \in R$ και $\lambda \in \Lambda$.

Το κυματίδιο ψ και το φίλτρο *φ* χαμηλής διέλευσης είναι κατασκευασμένα έτσι ώστε να καλύπτουν όλο τον άξονα της συχνότητας.

Έχουμε λοιπόν:

$$A(\omega) = |\varphi(\omega)|^2 + \frac{1}{2}\sum_{\lambda\in\Lambda}(|\psi_\lambda(\omega)|^2 + |\psi_\lambda(-\omega)|^2) \qquad (3.7)$$

για κάθε $\omega \in R$ και ισχύει : $1 - \alpha \leq A(\omega) \leq 1$ (3.8) , με $\alpha < 1$.

Ο wavelet μετασχηματισμός είναι σταθερός και αντίστροφος τελεστής (operator). Από την Ευκλείδεια νόρμα του *x* έχουμε $\|x\|^2 = \int |x|^2 dt$ και πολλαπλασιάζοντας στην (3.8) έχουμε:

$$(1 - \alpha)\|x\|^2 \leq \|Wx\|^2 \leq \|x\|^2 \qquad (3.9)$$

Και συνεπώς η νόρμα του Wx:

$$\|Wx\|^2 = \int |x * \varphi(t)|^2 dt + \sum_{\lambda\in\Lambda} \int |x * \psi_\lambda(t)|^2 dt \qquad (3.10)$$

Αθροίζει το τετράγωνο των συντελεστών.

Ένα σήμα δεν μπορεί να αναδημιουργηθεί από το μέτρο (modulus) του FT του, αλλά μπορεί να αναδημιουργηθεί από αυτήν του WT του. Αυτό ισχύει διότι το μέτρο (modulus) του τελεστή (operator) wavelet παρότι αφαιρεί την

σύνθετη φάση, δεν χάνει πληροφορία από το σήμα και συγκεκριμένα ο WT έχει περισσότερους συντελεστές από το αρχικό σήμα.

Από την (3.6) έχουμε ότι από την συνέλιξη του σήματος με το φίλτρο χαμηλής διέλευσης φ χάνονται οι υψηλές συχνότητές του. Αυτές μπορούν όμως να ανακτηθούν με μια νέα σειρά απόλυτων τιμών των συντελεστών των κυματιδίων που αλληλεπικαλύπτονται στα προηγούμενα. Αυτή η διαδικασία της τοποθέτησης με αλληλοεπικάλυψη, σε μορφή «καταρράκτη» (cascading: διαδικασία που δημιουργεί τον καταιγισμό), είναι ο μετασχηματισμός wavelet scattering.

Παίρνουμε μια τοπική παράμετρο του σήματος x που είναι αμετάβλητη στη μεταφορά με time-average (μέσος όρος των τιμών σε διαφορετικούς χρόνους) $S_0x(t) = x * \varphi(t)$ (3.11) και αφαιρεί όλες τις υψηλές συχνότητες.

Αυτές τις ανακτούμε με τον WT :

$$|W_1|x = (x * \varphi(t), |x * \psi_{\lambda 1}(t)|) \quad (3.12)$$

με $t \in R$ και $\lambda_1 \in \Lambda_1$

Τις υπολογίζουμε, λοιπόν, με τα κυματίδια $\psi_{\lambda 1}$ που έχουν συχνότητα ανάλυσης οκτάβας (octave frequency resolution) Q_1 , που για τα ηχητικά σήματα είναι $Q_1 = 8$. Συνεπώς έχουμε κυματίδια με την συχνότητα ανάλυσης ίδια με mel-frequency φίλτρα (περισσότερα για αυτά παρακάτω).

Τα ηχητικά σήματα έχουν χαμηλή ενέργεια στις χαμηλές συχνότητες και άρα παίρνουμε $S_0x(t) \approx 0$.

Οι πρώτης τάξης συντελεστές scattering είναι:

$$S_1x(t, \lambda_1) = |x * \psi_{\lambda 1}| * \varphi(t) \quad (3.13)$$

Τους υπολογίζουμε εφαρμόζοντας έναν δεύτερο μετασχηματισμό wavelet σε κάθε $|x * \psi_{\lambda 1}|$, που μας παρέχει συμπληρωματικούς υψηλής συχνότητας συντελεστές κυματιδίων:

$$|W_2||x * \psi_{\lambda 1}| = (|x * \psi_{\lambda 1}| * \varphi, ||x * \psi_{\lambda 1}| * \psi_{\lambda 2}|) \quad (3.14)$$

με $\lambda_2 \in \Lambda_2$.

Τα κυματίδια $\psi_{\lambda 2}$ έχουν συχνότητα ανάλυσης Q_2 διαφορετική από την Q_1. Την επιλέγουμε έτσι ώστε η πληροφορία του σήματος να συγκεντρώνεται σε όσο το δυνατόν λιγότερους συντελεστές κυματιδίων. Παίρνουμε τον μέσο όρο (average), που όπως εξηγήσαμε στην αρχή είναι ξανά μια συνέλιξη, αυτών των

συντελεστών με το φίλτρο φ, που μας εξασφαλίζει τοπική σταθερότητα σε χρονικές αλλαγές και προκύπτουν οι δεύτερης τάξης συντελεστές scattering:

$$S_2 x(t, \lambda_1, \lambda_2) = ||x * \psi_{\lambda 1}| * \psi_{\lambda 2}| * \varphi(t) \qquad (3.15)$$

Ξανά αυτοί οι συντελεστές υπολογίζονται εφαρμόζοντας έναν τρίτο μετασχηματισμό wavelet $|W_3|$ σε κάθε $||x * \psi_{\lambda 1}| * \psi_{\lambda 2}|$. Έχουμε συνελίξεις με νέα σειρά κυματιδίων $\psi_{\lambda 3}$ με συχνότητα ανάλυσης Q_3.

Επαναλαμβάνοντας την ίδια διαδικασία μπορούμε να βρούμε τους συντελεστές scattering *m*-οστής τάξης. Για κάθε $m \geq 1$ μπορούμε να γράψουμε:

$$U_m x(t, \lambda_1, \dots, \lambda_m) = | \dots ||x * \psi_{\lambda 1}| * \dots | * \psi_{\lambda m}(t)| \qquad (3.16)$$

Ο U_m είναι scattering χρονο-εξαρτώμενος τελεστής (propagator). Μια ταξινομημένη ακολουθία (sequence) $p = (\lambda_1, \lambda_2, \dots, \lambda_m)$ καλείται μονοπάτι (path). Το κενό σύνολο είναι το άδειο μονοπάτι. Έχουμε $U[\lambda] f = |f * \psi_\lambda|$ (3.17) για κάθε $f \in L^2(R^d)$. Συνεπώς ένας scattering propagator είναι προϊόν καθορισμένο από τα μονοπάτια μη-μεταθετικών τελεστών ως εξής:

$$U[p] = U[\lambda_m] \dots U[\lambda_2] U[\lambda_1] \qquad (3.18)$$

Η (3.17) είναι καλά καθορισμένη στον $L^2(R^d)$ διότι $||U[\lambda] f|| \leq \left|\left|\psi_\lambda\right|\right|_1 ||f||$ για κάθε $\lambda \in \Lambda_\infty$. Όπως φαίνεται και στην (3.16) ο scattering χρονο-εξαρτώμενος τελεστής είναι «καταρράκτης» από συνελίξεις και απόλυτες τιμές (modulus).

Κάθε *U[λ]* φιλτράρει τη συχνότητα στο εύρος συχνοτήτων που επικαλύπτεται από το ψ_λ και την αντιστοιχίζει σε χαμηλότερες συχνότητες μέσω του modulus. Άρα η ακολουθία $p = (\lambda_1, \lambda_2, \dots, \lambda_m)$ είναι λοιπόν μια μεταβλητή συχνότητας μονοπατιού.
Γυρνώντας πίσω στην (3.15) οι συντελεστές *m*-οστής τάξης προκύπουν ως εξής:

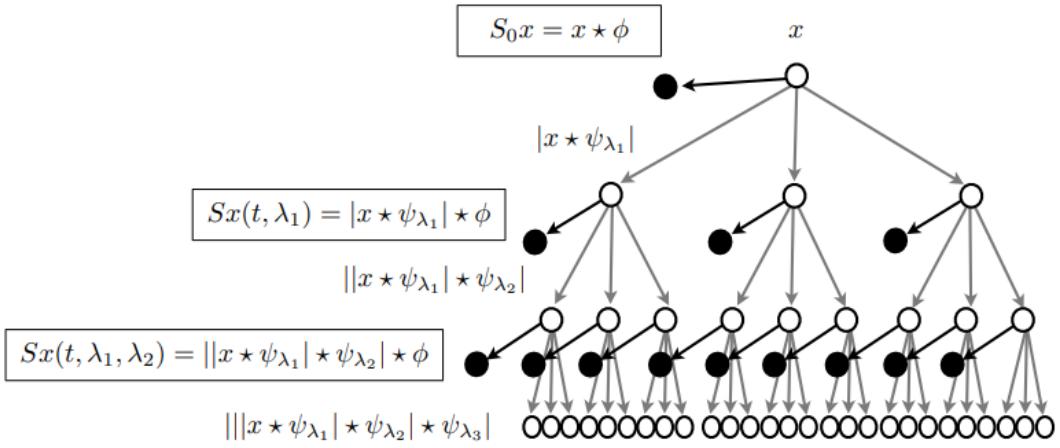
$$\begin{aligned} S_m x(t, \lambda_1, \dots, \lambda_m) &= | \dots ||x * \psi_{\lambda 1}| * \dots | * \psi_{\lambda m}| * \varphi(t) \\ &= U_m x(t, \lambda_1, \dots, \lambda_m) * \varphi(t) \end{aligned} \qquad (3.19)$$

Εφαρμόζοντας τον μετασχηματισμό $|W_{m+1}|$ στο $U_m x$ μπορούμε να υπολογίσουμε το $S_m x$ και το $U_{m+1} x$.

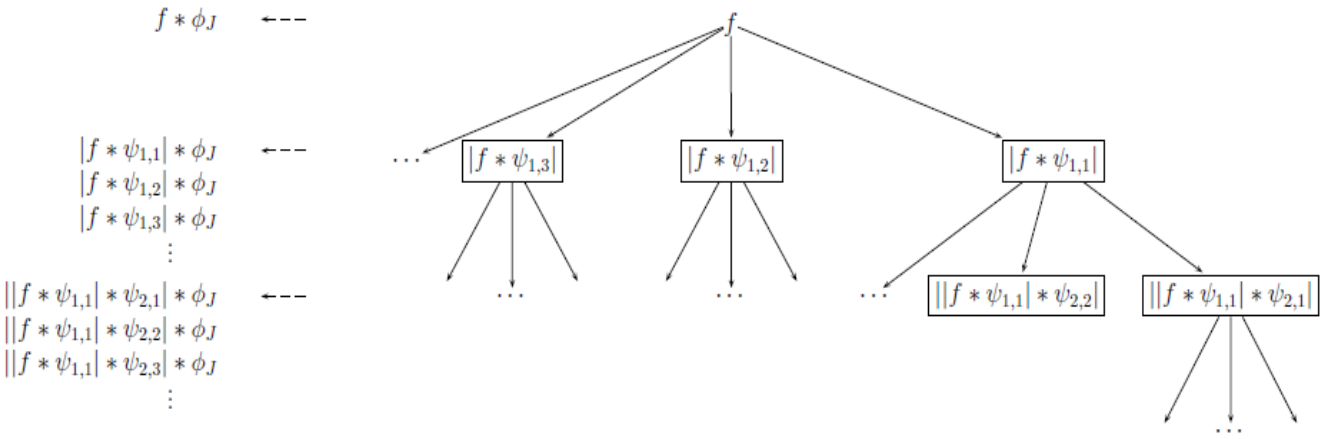
Αν θέσουμε λοιπόν για l μέγιστη τάξη με $U_0 x = x$ μπορούμε να υπολογίσουμε το WST για $0 \leq m \leq l$ ως εξής:

$$|W_{m+1}|U_m x = (S_m x\,, U_{m+1} x) \tag{3.20}$$

Ο μετασχηματισμός scattering φαίνεται στα παρακάτω σχήματα:



Σχήμα 3.2 : Ο μετασχηματισμός scattering. Από επαναλαμβανόμενη δράση των wavelet τελεστών $|W_m|$ υπολογίζονται «καταρράκτες» *m* συνελίξεων και μέτρων και προκύπτουν οι scattering συντελεστές $S_m x$.



Σχήμα 3.3: WST όπου φαίνονται περισσότερα μονοπάτια

Επίσης αν πάρουμε τον μετασχηματισμό για *m* φορές και βγάλουμε τους συντελεστές που δεν φιλτράρονται με το χαμηλής διέλευσης φίλτρο *φ* προκύπτει ένα διάνυσμα scattering τάξης *m+1* και χρόνου *t*:

$$S_J x(t) = \begin{pmatrix} x \star \phi_J(t) \\ |x \star \psi_{j_1}| \star \phi_J(t) \\ ||x \star \psi_{j_1}| \star \psi_{j_2}| \star \phi_J(t) \\ \vdots \\ ||\cdots|x \star \psi_{j_1}|\cdots| \star \psi_{j_m}| \star \phi_J(t) \end{pmatrix}_{j_1, j_2, \ldots < J+P}$$

Σχήμα 3.4 : Διάνυσμα scattering τάξης *m+1* και χρόνου *t*

Μπορούμε να δούμε ότι ο μετασχηματισμός scattering (ST) μοιάζει με τον μετασχηματισμό Fourier (FT), όπου όμως το μονοπάτι παίζει τον ρόλο μιας μεταβλητής της συχνότητας. Σε αντίθεση όμως με τον FT, ο ST είναι σταθερός υπό την επίδραση των διαφορομορφισμών, επειδή υπολογίζεται με επαναλαμβανόμενους wavelet μετασχηματισμούς και τελεστές modulus που είναι σταθεροί. Επίσης, όπως ο FT, και ο WST έχει αντίστροφο μετασχηματισμό.

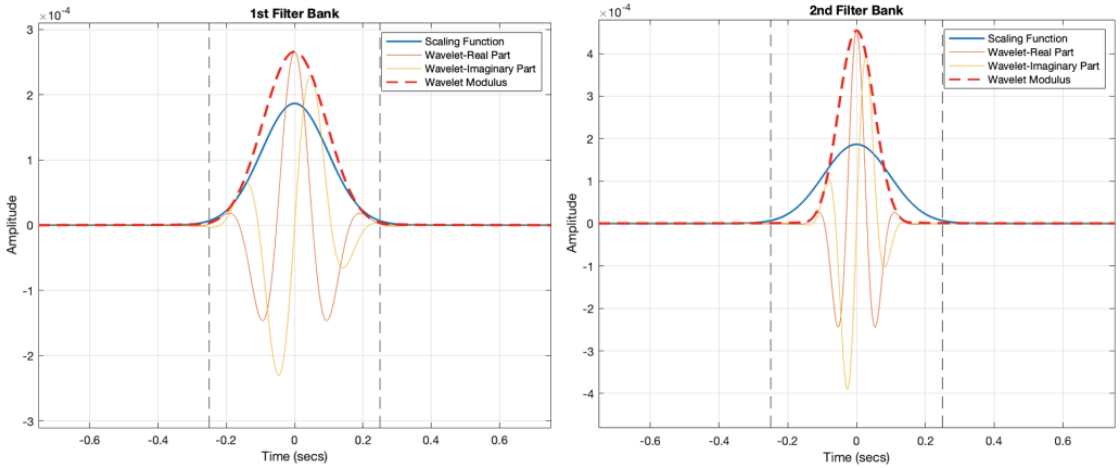
3.3 Αμετάβλητη κλίμακα και συχνότητα ανάλυσης

Όπως θα δούμε και παρακάτω στον κώδικα για να εξάγουμε πληροφορίες από τα δεδομένα στο πρόγραμμα που θα χρησιμοποιήσουμε, στην προκειμένη περίπτωση στη MATLAB, καθορίζουμε κάποιες παραμέτρους. Αυτές είναι οι εξής τρεις:

- Το μέγεθος της αμετάβλητης κλίμακας (invariance scale)
- Την τάξη του ST ή όπως αναφέρεται αλλιώς wavelet filter bank (στρώση/σειρά φίλτρων των κυματιδίων). Στις περισσότερες εφαρμογές χρησιμοποιούμε μόνο *m=2* filter banks διότι εκεί συγκεντρώνεται η μεγαλύτερη ποσότητα της ενέργειας του σήματος (για μικρό averaging scale)
- Τον αριθμό των κυματιδίων ανά οκτάβα σε κάθε filter bank

Αρχικά να εξηγήσουμε το invariance scale [6]. Όπως αναφέραμε και προηγουμένως (3.1) το φίλτρο φ_J είναι scaling συνάρτηση. Το σύστημα είναι αμετάβλητο σε μεταφορές πάνω από την χρονική τιμή (support) της φ_J . Η τιμή αυτή λοιπόν, καθορίζει την σταθερότητα του συστήματος στο χωροχρόνο.

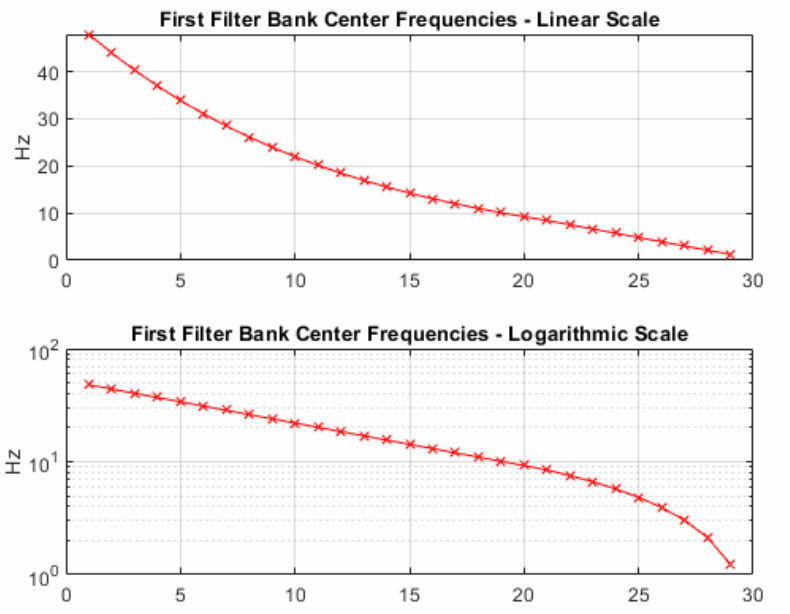
Εφόσον μελετάμε ηχητικά σήματα που είναι χρονοεξαρτώμενα, το invariance scale θα είναι χρονική διάρκεια. Η scaling συνάρτηση είναι κατά βάση ένα γκαουσιανό φίλτρο χαμηλής διέλευσης.



(a) 1st-order filter bank (b) 2nd-order filter bank

Σχήμα 3.5: Scattering κυματίδια στην πρώτη και δεύτερη τάξη του ST. Το invariance scale είναι *0.5 s* και στις 2 περιπτώσεις. Στο (a) τα κυματίδια είναι περιορισμένα στο invariance scale και το time support τους είναι ίσο με αυτό. Στο (b) τα κυματίδια είναι ακόμα πιο περιορισμένα στο χρόνο και το time support τους είναι πιο μικρό από το invariance scale

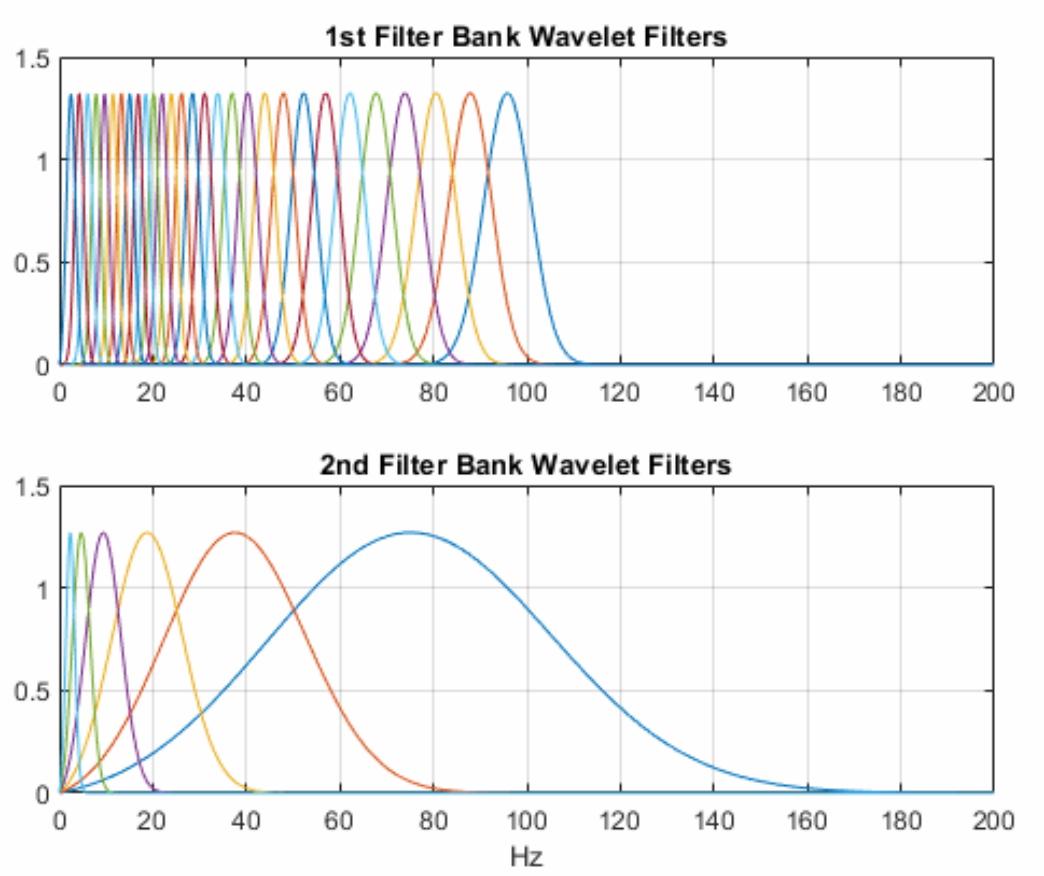
Το invariance scale επηρεάζει επίσης την απόσταση των κεντρικών συχνοτήτων λ των κυματιδίων στα filter banks.



Σχήμα 3.6: Οι κεντρικές συχνότητες των κυματιδίων στο πρώτο filter bank του ST. Στην

πρώτη εικόνα έχουμε γραμμική απεικόνιση και στην δεύτερη λογαριθμική. Στην γραμμική απεικόνιση είναι οι χαμηλές συχνότητες και στην λογαριθμική οι υψηλές.

Στη συνέχεια εξηγούμε την συχνότητα ανάλυσης ή αλλιώς τον αριθμό των κυματιδίων ανά οκτάβα. Η συχνότητα ανάλυσης Q εκτιμάται σε κάθε στρώση (layer) *m* ώστε να παραχθούν αραιοί συντελεστές κυματιδίων στην επόμενη στρώση. Αυτό διατηρεί καλύτερα την πληροφορία του σήματος. Για τα ηχητικά σήματα επομένως, επιλέγεται το $Q_1 = 8$ κυματίδια ανά οκτάβα, διότι μας παρέχει αραιές αναπαραστάσεις του σήματος μουσικής. Στην δεύτερη τάξη επιλέγεται το $Q_2 = 1$ που καθορίζονται κυματίδια που χαρακτηρίζουν καλύτερα τα transients (δυνατός ήχος μικρής διάρκειας) και attacks (ο χρόνος που χρειάζεται ένας ήχος να πάει σε ένταση από 0-100%) σε ένα μουσικό κομμάτι.

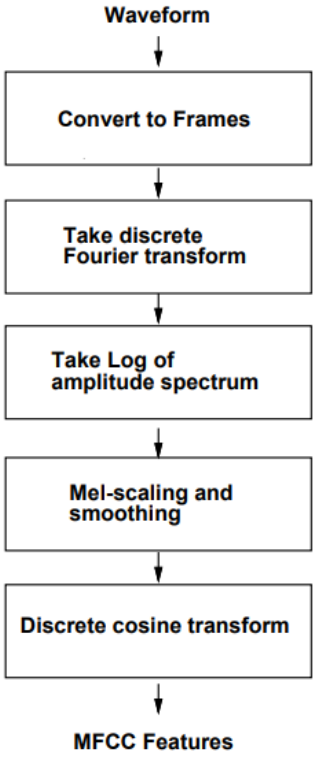


Σχήμα 3.7 : Τα φίλτρα κυματιδίων με invariance scale 1 δευτερόλεπτο και συχνότητα 200Hz. Στο πρώτο είναι $Q_1=8$ και στο δεύτερο $Q_2=1$.

Όπως αναφέραμε και προηγουμένως στις περισσότερες εφαρμογές χρησιμοποιούμε μέχρι $m=2$ τάξης scattering μετασχηματισμούς. Αυτό διότι όσο η τάξη m αυξάνεται η ενέργεια του σήματος συγκλίνει στο 0. Η ενέργεια είναι συγκεντρωμένη σε μονοπάτια ελάττωσης της συχνότητας και το $U[\lambda]$ την «σπρώχνει» σε χαμηλές συχνότητες.

3.4 Mel-Frequency Cepstral Coefficients (MFCCs)

Τα MFCCs, που είναι μετασχηματισμοί συνημιτόνων των mel-frequency spectral coefficients (MFSCs), χρησιμοποιούνται από πολλούς ταξινομητές (classifiers) μουσικής και λόγου εδώ και κάποια χρόνια [3]. Έχουν ως χαρακτηριστικό μικρής χρονικής διάρκειας φάσματα. Η τεχνική που χρησιμοποιείται αποτελείται από δύο μορφής φίλτρα, γραμμικού διαχωρισμού και λογαριθμικού διαχωρισμού. Η Mel κλίμακα αντιστοιχίζει την κανονική συχνότητα με την οξύτητα (pitch) που αντιλαμβανόμαστε, εφόσον το ανθρώπινο αυτί δεν αντιλαμβάνεται το pitch ως γραμμικό. Η κλίμακα για τα γραμμικά φίλτρα πάει κάτω από 1000Hz, ενώ αυτή των λογαριθμικών φίλτρων πάνω από 1000Hz [11].



Σχήμα 3.8: Διαδικασία δημιουργίας MFCCs

Αρχικά χωρίζουμε το σήμα σε μικρότερα χρονικά κομμάτια ώστε να θεωρείται το καθένα από αυτά στάσιμο. Συνήθως γίνεται με την βοήθεια συνάρτησης παραθύρου (windowing function). Στη συνέχεια παίρνουμε τον διακριτό Fourier μετασχηματισμό για κάθε ένα από τα μικρότερα παράθυρα. Ύστερα παίρνουμε τον λογάριθμο του πλάτους του φάσματος (spectrum) διότι η ένταση ενός σήματος που αντιλαμβανόμαστε είναι λογαριθμικής κλίμακας. Στη συνέχεια φιλτράρουμε αυτόν τον λογάριθμο με φίλτρα της mel κλίμακας και το φάσμα γίνεται πιο ομαλό (smooth). Τελευταίο βήμα είναι στα διανύσματα του φάσματος mel να εφαρμόσουμε τον διακριτό μετασχηματισμό συνημιτόνου (discrete cosine transform).

Σε σταθερό χρονικό διάστημα τα MFCCs υπολογίζουν την συχνότητα ενέργειας του σήματος. Όμως, χάνεται πληροφορία σε σήματα που δεν είναι στατικά (non-stationary) σε αυτό το διάστημα. Για να μειωθεί αυτή η απώλεια, παίρνουμε μικρά χρονικά παράθυρα διάρκειας *T* = 23 ms τα οποία θεωρούνται τοπικά στατικά. Στα μη-στατικά σήματα παρόλα αυτά παραμένει το πρόβλημα και για αυτό χρησιμοποιούμε τον wavelet scattering μετασχηματισμό.
Ένα mel-frequency spectrogram (απεικόνιση του φάσματος των συχνοτήτων ενός σήματος) βρίσκει τον μέσο όρο (average) της ενέργειας του spectrogram με τα mel φίλτρα ψ_λ όπου λ είναι η κεντρική συχνότητα σε κάθε $\psi_\lambda(\omega)$:

$$Mx(t,\lambda) = \frac{1}{2\pi}\int |x(t,\omega)|^2 |\psi_\lambda(\omega)|^2 d\omega \qquad (3.21)$$

Όπου το spectrogram γράφεται ως εξής:

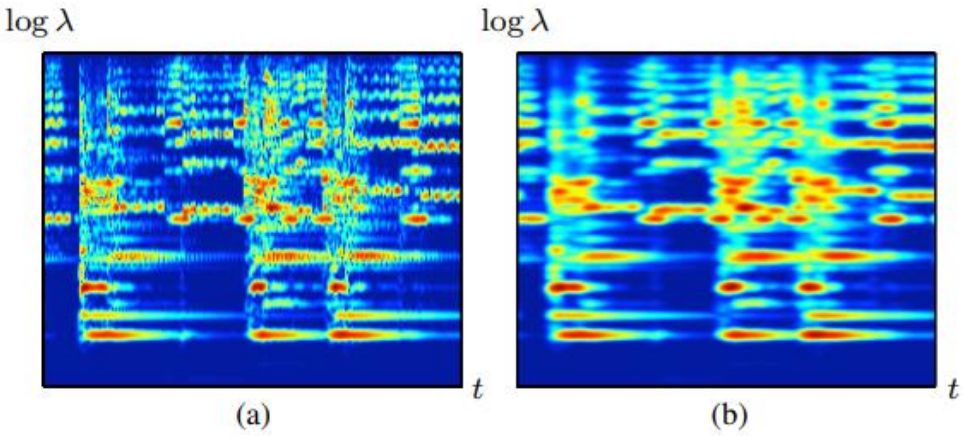
$$|x(t,\omega)| = |\int x(u)\varphi(u-t)e^{-i\omega t}\,du| \qquad (3.22)$$

Τα band-pass φίλτρα ψ_λ έχουν σταθερό εύρος συχνοτήτων Q σε υψηλές συχνότητες. Η τιμή της συχνότητας έχει εύρος της τάξης $\frac{\lambda}{Q}$ με το κέντρο στο λ. Στις χαμηλές συχνότητες το εύρος συχνοτήτων δεν είναι σταθερό, αλλά ίσο με $\frac{2\pi}{T}$. Η κλίμακα mel μας παρέχει ισορροπία στην χρονική παραμόρφωση (time warp) αλλά χάνει πληροφορία. Η (22) είναι ο FT της $x_t(u) = x(u)\varphi(u-t)$ και έτσι μπορεί να προκύψει από την (21):

$$\begin{aligned} Mx(t,\lambda) &= \int |x_t * \psi_\lambda(\upsilon)|^2 d\upsilon \\ &= \int |\int x(u)\varphi(u-t)\psi_\lambda(\upsilon-u)du\,|^2 d\upsilon \end{aligned} \qquad (3.23)$$

Και για $\lambda \gg \frac{Q}{T}$ το $\varphi(t)$ θεωρείται σταθερό και έχουμε:

$$Mx(t,\lambda) = \int \left| \int x(u)\psi_\lambda(v-u)du \right|^2 |\varphi(u-t)|^2 dv$$
$$= |x * \psi_\lambda|^2 * |\varphi|^2(t) \quad (3.24)$$



Σχήμα 3.9: (a) Το scalogram $log|x * \psi_\lambda(t)|^2$ για μουσικό σήμα. (b) Το averaged scalogram $|x * \psi_\lambda|^2 * |\varphi|^2(t)$ με χαμηλής διέλευσης φίλτρο φ διάρκειας *T=190 ms*

Παρατηρούμε ότι στην εικόνα (b) χάνεται πληροφορία.

Επίσης από την (3.24) προκύπτει ότι τα MFCCs είναι ίσα με το τετράγωνο των scattering συντελεστών κυματιδίων όπως φαίνεται στην (3.1). Όμως μεγάλοι συντελεστές κυματιδίων αυξάνονται κατά πολύ όταν υπολογίζουμε το τετράγωνό τους. Προς αποφυγή της ενίσχυσης των απομακρυσμένων αυτών κυματιδίων, αφαιρούμε το τετράγωνο δίχως να αποτελεί πρόβλημα. Έτσι παίρνουμε τους συντελεστές scattering που όπως δείξαμε και προηγουμένως λύνουν το πρόβλημα της πληροφορίας που χάνεται από τα MFCCs σε μεγάλης χρονικής διάρκειας σήματα.

3.5 Ιδιότητες του μετασχηματισμού scattering

3.5.1 Ισορροπία σε χρονική παραμόρφωση

Ο μετασχηματισμός Fourier δεν είναι σταθερός σε παραμορφώσεις διότι αν διαστείλουμε ένα ημιτονοειδές σήμα, το αποτέλεσμα θα είναι ένα καινούριο ημιτονοειδές σήμα με διαφορετική συχνότητα και το οποίο θα είναι ορθογώνιο προς το αρχικό σήμα. Από την άλλη τα mel-frequency spectrograms παραμένουν σταθερά σε χρονικές παραμορφώσεις όταν παίρνουμε τον μέσο όρο (average) τους με μια συχνότητα.

Το ίδιο ισχύει και για τον μετασχηματισμό scattering, εφόσον τα κυματίδια είναι σταθερά σε χρονικές παραμορφώσεις. Υπάρχει $C > 0$ ώστε

$$||\psi_\lambda - \psi_{\lambda,\tau}|| \leq C||\psi_\lambda||sup_t|\tau'(t)| \quad (3.25)$$

για κάθε λ και $\tau(t)$ και $\psi_{\lambda,\tau} = \psi_\lambda(t - \tau(t))$ (3.26) χρονική παραμόρφωση του κυματιδίου ψ_λ . Άρα ομοίως προκύπτει για τον μετασχηματισμό scattering. Αν έχουμε χρονική παραμόρφωση του σήματος $x_\tau(t) = x(t - \tau(t))$ με $|\tau'(t)| < 1$ και $sup_t|\tau(t)| \ll T$, η μέγιστη χρονική μεταφορά είναι πολύ μικρότερη από την περίοδο. Εν τέλει προκύπτει:

$$\big||Sx_\tau - Sx|\big| \leq Csup|\tau'(t)|||x|| \quad (3.27)$$

Η σταθερά C είναι μέτρο ισορροπίας (σταθερότητας) του μετασχηματισμού.

3.5.2 Διατήρηση της ενέργειας

Ένας scattering μετασχηματισμός παραμένει σταθερός με την προσθήκη θορύβου.

$$\|Sx - Sx'\| \leq \|x - x'\| \quad (3.28)$$

Με αυτήν την ιδιότητα από την (3.20) μπορούμε να πάρουμε:

$$\|U_m x\|^2 = \|S_m x\|^2 + \|U_{m+1} x\|^2 \quad (3.29)$$

και για $0 \leq m \leq l$ παίρνουμε:

$$\|x\|^2 = \|Sx\|^2 + \|U_{l+1} x\|^2 \quad (3.30)$$

Όσο το l αυξάνεται το $\|U_{l+1} x\|$ τείνει στο 0, συνεπώς για $l = \infty$ τότε $\|x\| = \|Sx\|$.

Το modulus των συντελεστών των κυματιδίων «σπρώχνουν» την ενέργεια σε χαμηλές συχνότητες. Συνεπώς και ο μετασχηματισμός scattering «σπρώχνει» την ενέργεια του U_m σε χαμηλές συχνότητες, η οποία αποθηκεύεται από το χαμηλής διέλευσης φίλτρο φ.

Στον πίνακα βλέπουμε τα ποσοστά της ενέργειας $\frac{\|S_m x\|^2}{\|x\|^2}$ που αποθηκεύονται σε κάθε τάξη του μετασχηματισμού scattering.

T	$m = 0$	$m = 1$	$m = 2$	$m = 3$
23 ms	0.0%	94.5%	4.8%	0.2%
93 ms	0.0%	68.0%	29.0%	1.9%
370 ms	0.0%	34.9%	53.3%	11.6%
1.5 s	0.0%	27.7%	56.1%	24.7%

Πίνακας 3.1: Οι τιμές της ενέργειας υπολογισμένες για x σήματα από ένα dataset λόγου [1] ως συνάρτηση τάξης *m* και averaging scale *T*. Για *m=1* είναι $Q_1=8$ και για *m=2,3* είναι $Q_2=Q_3=1$.

Τα ηχητικά σήματα έχουν χαμηλή ενέργεια στις χαμηλές συχνότητες και $S_0 x \approx 0$ συνεπώς το μεγαλύτερο ποσοστό ενέργειας για *T=23 ms* αποθηκεύεται στους πρώτης τάξης scattering συντελεστές. Για αυτό και τα mel-frequency spectrograms χρησιμοποιούνται όταν έχουμε παράθυρα τόσο μικρής χρονικής διάρκειας. Όσο το *T* αυξάνεται όμως μεγαλύτερα ποσοστά ενέργειας αποθηκεύονται σε πιο υψηλής τάξης συντελεστές. Για τα σήματα όμως που θα μελετήσουμε και εμείς ισχύει ότι $T < 1{,}5\ s$ άρα οι τρίτης τάξης συντελεστές δεν μας ενδιαφέρουν.

Κεφάλαιο 4

Κατηγοριοποίηση μουσικών ειδών μέσω του διανύσματος των 30 διαστάσεων

4.1 Εισαγωγή

Σε αυτό το κεφάλαιο περιγράφουμε και αναλύουμε την μελέτη που έχει γίνει από τους G. Tzanetakis και P. Cook στο [8]. Η μέθοδος που χρησιμοποιείται είναι συνδυασμός του STFT, των MFCCs και του WT για την εξαγωγή διαφόρων χαρακτηριστικών. Στο έκτο κεφάλαιο θα συγκρίνουμε τα αποτελέσματα αυτής με την μελέτη που έγινε στο [7] για wavelet scattering, η οποία και είναι η εργασία με την οποία θα ασχοληθούμε εις βάθος.

Θα μελετήσουμε λοιπόν την κατηγοριοποίηση μουσικών κομματιών και τραγουδιών σε συγκεκριμένα μουσικά είδη. Ως μουσικό είδος, όπως αναφέραμε και στο πρώτο κεφάλαιο, εννοούμε την περιγραφή των χαρακτηριστικών ενός μουσικού κομματιού [12]. Με αυτόν τον όρο μπορούμε να οργανώσουμε βάσεις δεδομένων και μουσικές βιβλιοθήκες αλλά και γενικότερα να περιγράψουμε την μουσική ώστε να διευκολυνθούμε στην εύρεση τραγουδιών ή και καλλιτεχνών. Συνεπώς μπορούμε να ξεχωρίσουμε διάφορα χαρακτηριστικά στα τραγούδια, να τα ομαδοποιήσουμε σε ένα συγκεκριμένο είδος και να συγκρίνουμε τις ομοιότητες και τις διαφορές τους με τα υπόλοιπα μουσικά είδη. Με αυτόν τον τρόπο μπορούμε να επιλέγουμε ευκολότερα τα τραγούδια και κομμάτια που μας φαίνονται πιο ευχάριστα στο άκουσμα. Γενικά τα

μουσικά είδη δεν έχουν αυστηρούς περιορισμούς εφόσον προκύπτουν από παράγοντες πολιτισμικούς, ιστορικούς, μάρκετινγκ αλλά και αλληλεπίδρασης γενικότερα με το ευρύτερο κοινό [8]. Τυπικά δημιουργήθηκαν από ειδικούς στον τομέα ανθρώπους που κατηγοριοποιούσαν με την ακοή. Σε αυτή την μελέτη όμως θα γίνει αυτόματη κατηγοριοποίηση, συνεπώς θα δούμε ένα πλαίσιο (framework) που θα αναπτύσσει και θα αξιολογεί χαρακτηριστικά που περιγράφουν την μουσική. Τέτοια χαρακτηριστικά είναι: το ηχόχρωμα, ο ρυθμός και η οξύτητα (pitch).

4.2 Εξαγωγή χαρακτηριστικών

4.2.1 Χαρακτηριστικά ηχοχρώματος

Το ηχόχρωμα (timbre) είναι η ιδιαίτερη χροιά του ήχου που δεν έχει σχέση με την έντασή του. Τα χαρακτηριστικά του βασίζονται στον short time Fourier Transform (STFT) και στα Mel-frequency cepstral coefficients (MFCCs) και υπολογίζονται για κάθε διάστημα ήχου μικρής χρονικής διάρκειας. Αυτά αναφέρονται παρακάτω.

Φασματικό κέντρο βάρους (Spectral Centroid)

Είναι το κέντρο βάρους του μεγέθους (magnitude) του φάσματος του STFT.

$$C_t = \frac{\sum_{n=1}^{N} M_t[n] \cdot n}{\sum_{n=1}^{N} M_t[n]} \tag{4.1}$$

όπου $M_t[n]$ είναι το πλάτος (μέγεθος) του μετασχηματισμού Fourier σε διάστημα χρόνου *t* και frequency bins (το μεσοδιάστημα ανάμεσα στα δείγματα της συχνότητας που έχουμε) *n*. Το centroid είναι μονάδα μέτρησης του σχήματος του φάσματος (spectral shape) και οι υψηλές τιμές του αντιστοιχούν σε πιο «αραιό» φάσμα με υψηλές συχνότητες.

Φασματική συχνότητα αποκοπής (Spectral Rolloff)

Είναι η συχνότητα R_t κάτω από την οποία είναι συγκεντρωμένο το 85% του πλάτους κατανομής (magnitude distribution)

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 \cdot \sum_{n=1}^{N} M_t[n] \tag{4.2}$$

Όπως και το centroid, το rolloff είναι μια άλλη μονάδα μέτρησης του σχήματος του φάσματος.

Ροή Φάσματος (Spectral Flux)

Είναι η διαφορά των κανονικοποιημένων διαδοχικών πλατών των φασματικών κατανομών (spectral distributions) στο τετράγωνο

$$F_t = \sum_{n=1}^{N}(N_t[n] - N_{t-1}[n])^2 \qquad (4.3)$$

όπου $N_t[n]$ και $N_{t-1}[n]$ είναι τα κανονικοποιημένα πλάτη του μετασχηματισμού Fourier στο τωρινό διάστημα χρόνου *t* και στο προηγούμενο διάστημα *t-1,* αντίστοιχα. Κανονικοποίηση (normalization) είναι η διαδικασία μετατροπής των δεδομένων σε μία ακολουθία κανονικών μορφών, οι οποίες αποτελούνται από απλές και σαφείς σχέσεις που δεν περιέχουν επαναλήψεις. Το spectral flux είναι μονάδα μέτρησης της ποσότητας της τοπικής αλλαγής του φάσματος.

Μηδενισμοί στο πεδίο του χρόνου (Time Domain Zero Crossings)

$$Z_t = \frac{1}{2}\sum_{n=1}^{N}|sign(x[n]) - sign(x[n-1])| \qquad (4.4)$$

όπου η συνάρτηση $sign(x)$ παίρνει την τιμή 1 για θετικά ορίσματα (arguments) και 0 για αρνητικά και το $x[n]$ είναι το σήμα για διάστημα χρόνου (time domain) *t.* Τα time domain zero crossings είναι μονάδα μέτρησης του θορύβου ενός σήματος.

Mel-Frequency Cepstral Coefficients (MFCCs)

Είναι βασισμένα στον STFT όπως είδαμε και στο προηγούμενο κεφάλαιο. Παίρνουμε τον λογάριθμο του πλάτους του φάσματος και φιλτράροντας τον λογάριθμο με την βοήθεια της κλίμακας mel, τα FFT bins ομαδοποιούνται και γίνονται πιο ομαλά (smooth), το φάσμα γίνεται πιο ομαλό. Τέλος, εφαρμόζεται στα διανύσματα του φάσματος mel διακριτός συνημιτονοειδής μετασχηματισμός. Για την κατηγοριοποίηση των ειδών μουσικής που μελετάμε οι πρώτοι 5 συντελεστές MFCCs βγάζουν καλά αποτελέσματα, ενώ για ηχητικό σήμα που περιέχει ομιλία χρειάζονται 13 συντελεστές.

Παράθυρα ανάλυσης και υφής (Analysis and Texture Windows)

Τα analysis windows είναι μικρά χρονικά τμήματα που χωρίζουμε το σήμα και τα μελετάμε ξεχωριστά. Μερικά από αυτά πιθανώς αλληλεπικαλύπτονται. Πρέπει να είναι αρκετά μικρά ώστε το σήμα να θεωρείται στάσιμο και έτσι η συχνότητα του πλάτους του φάσματος να παραμένει σχετικά σταθερή.

Από την άλλη, η 'υφή' (texture) των ήχων προκύπτει σαν αποτέλεσμα από πολλαπλά φάσματα μικρής χρονικής διάρκειας με διαφορετικά χαρακτηριστικά. Για παράδειγμα στην ομιλία τα φωνήεντα και τα σύμφωνα έχουν πολύ διαφορετικά φασματικά χαρακτηριστικά. Ο όρος texture window συνεπώς χρησιμοποιείται για να περιγράψει ένα πιο μεγάλο παράθυρο (τμήμα) και αντιστοιχεί στον ελάχιστο χρόνο που χρειάζεται ώστε να αναγνωρίσουμε έναν ιδιαίτερο ήχο ή μια μουσική 'υφή'. Το texture window περιέχει τα τωρινά διανύσματα χαρακτηριστικών (feature vectors) αλλά και συγκεκριμένο αριθμό από παρελθοντικά διανύσματα. Στη συγκεκριμένη μελέτη, χρησιμοποιείται ένα analysis window των 23 ms (με 512 δείγματα και ρυθμό δειγματοληψίας 22050Hz) και ένα texture window του 1 s (με 43 analysis windows).

Χαρακτηριστικό χαμηλής ενέργειας (Low-Energy Feature)

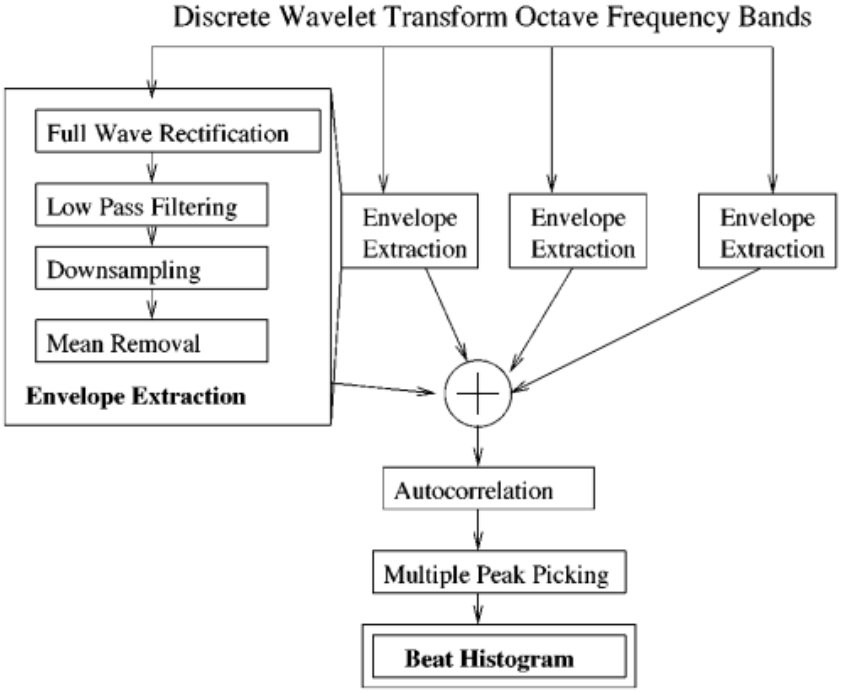
Είναι το μοναδικό χαρακτηριστικό ηχοχρώματος που βασίζεται στο texture window και όχι στο analysis window. Είναι το ποσοστό του analysis window που έχει χαμηλότερη Μέση Τετραγωνική Ρίζα (Root Mean Square: RMS) ενέργειας από την μέση RMS ενέργεια σε όλο το texture window. Για παράδειγμα σε ένα τραγούδι που περιέχει φωνητικά στοιχεία, αν έχουμε παύσεις ή σιωπές τότε η τιμή της low-energy (χαμηλής ενέργειας) θα είναι υψηλή, ενώ αν έχουμε έγχορδα που παίζουν αδιάκοπα τότε η τιμή της low-energy θα είναι χαμηλή.

4.2.2 Χαρακτηριστικά ρυθμού

Αυτά βασίζονται στην εκτίμηση και ανίχνευση του παλμού (beat), τον σταθερό χτύπο, δηλαδή, ενός μουσικού κομματιού που εμφανίζεται ανά τακτά χρονικά διαστήματα. Ως χαρακτηριστικά του ρυθμού μπορούν να θεωρηθούν τα εξής: η τακτικότητα (regularity) του ρυθμού, η σχέση του κύριου παλμού και των υποπαλμών, η σχετική ισχύς της των υποπαλμών στον κύριο παλμό. Βασικό χαρακτηριστικό συνεπώς του ρυθμού είναι να εντοπίσουμε τις περιοχές όπου η περιοδικότητα είναι εμφανής.

Η τεχνική που θα χρησιμοποιηθεί εδώ για την εξαγωγή των ρυθμικών χαρακτηριστικών είναι ο μετασχηματισμός κυματιδίων. Όπως έχουμε ξανά αναφέρει ο WT (wavelet transform) είναι καλύτερη τεχνική για την ανάλυση σήματος διότι σε σχέση με τον STFT (short-time Fourier transform) ξεπερνάει κάποια προβλήματα ανάλυσης. Στον STFT έχουμε μια ομοιόμορφη χρονική ανάλυση σε όλες τις συχνότητες, ενώ στον WT έχουμε για υψηλές συχνότητες καλή ανάλυση χρόνου και κακή ανάλυση συχνότητας και το ανάποδο για χαμηλές συχνότητες, δηλαδή καλή ανάλυση συχνότητας και κακή ανάλυση χρόνου.

Χρησιμοποιούμε πιο συγκεκριμένα τον DWT (discrete wavelet transform) που μας παρέχει μια καλή αναπαράσταση του σήματος σε χρόνο και συχνότητα.



Σχήμα 4.1: Ο υπολογισμός του Beat Histogram

Στο *Σχήμα 4.1* λοιπόν βλέπουμε το διάγραμμα του αλγόριθμου για την ανάλυση του παλμού (beat analysis). Ξεκινώντας με τον DWT το σήμα αποσυντίθεται σε έναν συγκεκριμένο αριθμό από εύρη συχνοτήτων των οκτάβων (octave frequency bands). Ύστερα εξάγεται ξεχωριστά για κάθε εύρος συχνότητας (band) η αλλαγή του πλάτους στο χρόνο (amplitude envelope). Αυτή η αλλαγή επηρεάζει την αντίληψή μας για το ηχόχρωμα. Temporal envelope γενικότερα είναι οι διάφορες αλλαγές (συχνότητας, πλάτους ή οξύτητας) ενός ηχητικού σήματος στο χρόνο. Συνεπώς αυτό το amplitude envelope το πετυχαίνουμε με 3 βήματα από το *Σχήμα 4.1:*

- **Full Wave Rectification**

$$y[n] = |x[n]| \quad (4.5)$$

Εξάγουμε τις αλλαγές (envelope) του σήματος και όχι το ίδιο το σήμα.

- **Φίλτρο χαμηλής διέλευσης (Low Pass Filtering)**

$$y[n] = (1 - a)x[n] + ay[n - 1] \quad (4.6)$$

όπου το α=0.99 και χρησιμοποιείται για να εξομαλύνει τα envelope.

- **Μείωση δειγματοληψίας/συμπίεση (Downsampling)**

$$y[n] = xk[n] \quad (4.7)$$

όπου *k=16*. Συμπιέζοντας λοιπόν το σήμα μπορούμε να μειώσουμε τον χρόνο υπολογισμού της αυτοσυσχέτισης (autocorrelation), λόγω της μεγάλης περιοδικότητας της ανάλυσης παλμών. Ως αυτοσυσχέτιση εννοούμε την συσχέτιση ενός σήματος με ένα καθυστερημένο αντίγραφο του εαυτού του.
Συνεπώς συμπιέζουμε το σήμα μας στα εύρη συχνοτήτων των οκτάβων (octave frequency bands).

- **Αφαίρεση του Μέσου (Mean Removal)**
Το σήμα μας είναι κεντραρισμένο στο 0 για την αυτοσυσχέτιση.
- **Βελτιωμένη/αυξημένη αυτοσυσχέτιση (Enhanced Autocorrelation)**
Αθροίζουμε τις αλλαγές πλάτους κάθε ζώνης συχνοτήτων (band) και υπολογίζουμε την αυτοσυσχέτιση του αθροίσματος.

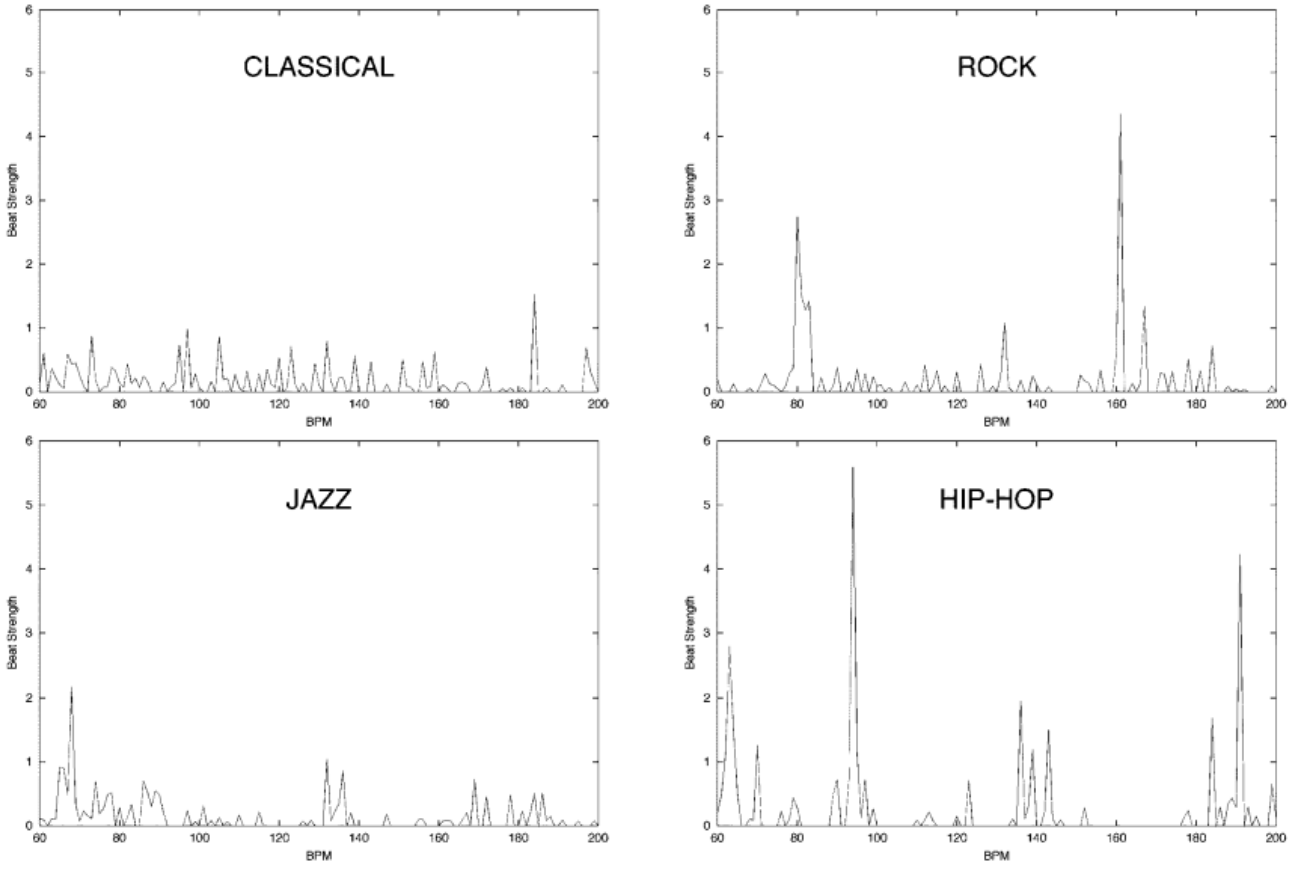
$$y[k] = \frac{1}{N}\sum_n x[n]x[n - k] \quad (4.8)$$

Οι αιχμές/κορυφές (peaks) που συναντάμε στην συνάρτηση της autocorrelation είναι οι χρόνοι όπου το σήμα μας μοιάζει περισσότερο με τον καθυστερημένο εαυτό του. Αυτές οι χρονικές καθυστερήσεις των κορυφών αντιστοιχούν στην περιοδικότητα των παλμών.

- **Εντοπισμός των κορυφών και υπολογισμός του ιστογράμματος**
Οι πρώτες τρεις κορυφές από την enhanced autocorrelation συνάρτηση προστίθενται στο παλμικό ιστόγραμμα (beat histogram: BH). Το διάστημα ανάμεσα στις κορυφές είναι η περίοδος των παλμών που μετριέται σε παλμό-ανά-λεπτό (beats-per-minute: bpm) και έχουμε από 40 έως 200 bpm. Το πλάτος του κάθε παλμού προστίθεται στο BH και με αυτόν τον τρόπο στις μεγάλες κορυφές του BH είναι τα σημεία όπου το σήμα είναι παρόμοιο με τον καθυστερημένο εαυτό του.

Στο *σχήμα 4.2* το κλασσικό BH είναι απόσπασμα από το κομμάτι '*La mer*' του *Claude Debussy*. Λόγω της πολυπλοκότητας των πολλών μουσικών οργάνων της ορχήστρας το σήμα δεν εμφανίζει ομοιότητα με τον καθυστερημένο εαυτό του

(self-similarity) και συνεπώς δεν έχουμε κάποια έντονη κορυφή. Στο τζαζ ΒΗ, που είναι απόσπασμα από live εκτέλεση της *Dee Dee Bridgewater*, βλέπουμε δύο, όχι πολύ έντονες κορυφές, στα 70 και 140 bpm. Στο ροκ ΒΗ, απόσπασμα από το τραγούδι *'Come together'* των *Beatles*, βλέπουμε δύο έντονες κορυφές στα 80 και 160 bpm. Γενικά στη ροκ μουσική έχουμε πιο έντονους παλμούς και αυτό φαίνεται και στο ΒΗ της. Το τελευταίο ΒΗ αντιστοιχεί σε χιπ-χοπ τραγούδι της *Neneh Cherry*. Εδώ βλέπουμε αρκετές κορυφές, κάποιες πολύ πιο ισχυρές από τις άλλες. Συνεπώς καταλαβαίνουμε το πόσο ρυθμικά είναι τα χιπ-χοπ τραγούδια.



Σχήμα 4.2: Τα Beat Histogram 4 διαφορετικών ειδών μουσικής

Συγκρίνοντας λοιπόν τα beat histogram διαφορετικών ειδών μουσικής μεταξύ τους παρατηρούμε κάποιες διαφορές που μας βοηθάνε να τα ξεχωρίσουμε κιόλας μεταξύ τους. Παρακάτω θα δούμε κάποια χαρακτηριστικά των ΒΗ που βοηθάνε στην αυτόματη κατηγοριοποίηση των ειδών μουσικής.

- Το σχετικό πλάτος (διαιρεμένο με το άθροισμα όλων των πλατών στο ΒΗ) του πρώτου και του δεύτερου παλμού A_0, A_1 στο histogram.

- Ο λόγος *RA* του πλάτους του δεύτερου παλμού προς το πλάτος του πρώτου.
- Οι περίοδοι του πρώτου και του δεύτερου παλμού P_1, P_2 σε bpm.
- Το άθροισμα *SUM* σε όλο το histogram. Με το *SUM* καταλαβαίνουμε την «δύναμη» των παλμών.

Εφαρμόζουμε τον DWT σε παράθυρο 65536 δειγμάτων με εύρος συχνοτήτων 22050Hz που αντιστοιχεί σε 3 s. Στη συνέχεια «προχωράμε» το παράθυρο κατά 32768 δείγματα και με αυτό το μεγαλύτερο παράθυρο βλέπουμε που επαναλαμβάνεται το σήμα.

4.2.3 Χαρακτηριστικά οξύτητας/τονικότητας

Είναι η συχνότητα που αντιστοιχεί σε κάθε νότα και καθορίζει την οξύτητα (pitch) ή το τονικό της ύψος. Η νότα λα έχει f= 440 Hz και πάνω από αυτήν την συχνότητα θεωρείται ότι έχουμε υψηλή οξύτητα (high pitch), ενώ αντίστοιχα κάτω από την f έχουμε χαμηλή οξύτητα.

Η μέθοδος που χρησιμοποιείται μοιάζει με τον προηγούμενο αλγόριθμο για τον υπολογισμό του beat histogram σε πιο μικρά χρονικά διαστήματα. (Για τους παλμούς έχουμε από 0.5 s έως 1.5 s ενώ για την οξύτητα από 2 ms έως 50 ms). Έχουμε εντοπισμό πολλαπλών οξυτήτων (multiple pitch). Το σήμα μας αποσυντίθεται σε δύο εύρη συχνοτήτων (frequency bands), κάτω και πάνω από 1000 Hz, και εξάγουμε τις αλλαγές του πλάτους (amplitude envelope) και στα δύο. Αυτό γίνεται με την βοήθεια του half-wave rectification και του φιλτραρίσματος με φίλτρο χαμηλής διέλευσης. Ύστερα, όπως και προηγουμένως, αθροίζουμε τις αλλαγές (envelopes) και υπολογίζεται η συνάρτηση αυτοσυσχέτισης (autocorrelation). Οι κορυφές που ξεχωρίζουν στην συνάρτηση αυτοσυσχέτισης αντιστοιχούν στις οξύτητες που κυριαρχούν στο σήμα μας και οι πρώτες τρεις από αυτές δημιουργούν το pitch histogram (PH). Αυτή τη φορά χρησιμοποιούμε ένα pitch analysis window με 512 δείγματα σε εύρος συχνοτήτων 22050 Hz που αντιστοιχεί σε περίπου 23 ms.

Οι συχνότητες που αντιστοιχούν σε κάθε κορυφή του PH μετατρέπονται σε οξύτητες έτσι ώστε κάθε μεσοδιάστημα (bin) του PH να αντιστοιχεί σε μια νότα με συγκεκριμένη τονικότητα/οξύτητα. Στο PH τις νότες τις γράφουμε μετατρέποντάς τες μέσω του συστήματος MIDI. Έτσι έχουμε:

$$n = 12log_2 \frac{f}{440} + 69 \qquad (4.9)$$

όπου λοιπόν *f* είναι η συχνότητα και *n* είναι ο αριθμός MIDI ή αριθμός bin.

Δημιουργούνται τελικά δύο εκδοχές του PH. Με την σχέση (4.9) δημιουργείται το unfolded PH και με την (4.10) το folded PH.

$$c = n \; mod 12 \tag{4.10}$$

όπου c είναι το folded histogram bin. Από αυτήν την σχέση όλες οι νότες αντιστοιχίζονται σε μια μοναδική οκτάβα και από εκεί προκύπτει και ο αριθμός 12, έχουμε 12 ημιτόνια σε μια οκτάβα. Ημιτόνιο ονομάζουμε την πιο μικρή απόσταση μεταξύ δύο νοτών, για παράδειγμα το ρε με το ρε# απέχουν ένα ημιτόνιο. Δύο ημιτόνια αποτελούν έναν τόνο.
Οι πληροφορίες που παίρνουμε από τα δύο PH είναι λίγο διαφορετικές. Από το unfolded βρίσκουμε το φάσμα τονικότητας (pitch range) ενός τραγουδιού ενώ από το folded βρίσκουμε τις αρμονικές.

Φτιάχνουμε καινούριο folded histogram bin ώστε τα διπλανά histogram bins να έχουν απόσταση μιας 5^{ης} (= απόσταση τριών τόνων + ενός ημιτόνιο ή εφτά ημιτονίων) μεταξύ τους. Με αυτόν τον τρόπο μπορούμε να εκφράσουμε καλύτερα την σχέση της τονικής συγχορδίας (1^{η} βαθμίδα) με την δεσπόζουσα (5^{η} βαθμίδα) που συχνά η πρώτη ακολουθάει την πέμπτη στα τραγούδια και απέχουν μια 5^{η} μεταξύ τους.

$$c' = (7 \times c) \; mod 12 \tag{4.11}$$

όπου c' είναι το καινούριο folded histogram bin.
Το pitch histogram μπορεί να παρουσιάσει και αυτό μικρές διαφορές για διαφορετικά μουσικά είδη. Για παράδειγμα στα κλασσικά και τζαζ κομμάτια έχουμε συχνή εναλλαγή νοτών και συνεπώς αλλαγή στην τονικότητα, κάτι που δεν ισχύει τόσο σε ροκ και ποπ τραγούδια. Η μορφή του PH λοιπόν των ροκ και ποπ τραγουδιών θα έχει λιγότερες κορυφές αλλά πιο έντονες σε σχέση με τα κλασσικά και τζαζ κομμάτια.

Μερικά χαρακτηριστικά που μπορούμε να εξάγουμε από τα PH για κατηγοριοποίηση των σημάτων σε συγκεκριμένα είδη μουσικής φαίνονται παρακάτω:

- Το πλάτος της μεγαλύτερης κορυφής στο folded histogram FA_o. Συνήθως αντιστοιχεί στην τονική συγχορδία. Όπως αναφέραμε και πριν, για τραγούδια που δεν έχουν μεγάλες εναλλαγές στην αρμονικότητά τους η κορυφή θα είναι πιο έντονη.

- Η περίοδος της μεγαλύτερης κορυφής στο unfolded histogram UP_0. Αντιστοιχεί στο οκταβικό φάσμα (octave range) της βασικής τονικότητας του τραγουδιού.
- Η περίοδος της μεγαλύτερης κορυφής στο folded histogram FP_0. Αντιστοιχεί στις αρμονικές του τραγουδιού.
- Η ενδιάμεση τονικότητα ανάμεσα στις 2 μεγαλύτερες κορυφές στο folded histogram IPO_1. Αντιστοιχεί στην σχέση των βασικών συγχορδιών (συνήθως τονική και δεσπόζουσα).
- Το άθροισμα *SUM* σε όλο το histogram. Με το άθροισμα καταλαβαίνουμε την «δύναμη» των pitch.

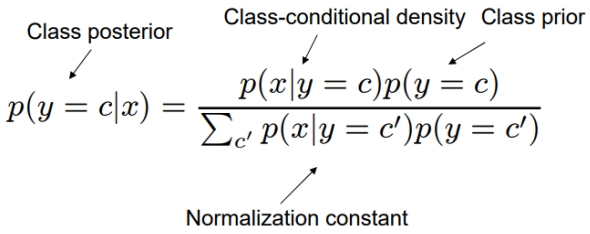
4.3 Αποτελέσματα κατηγοριοποίησης

Η βάση δεδομένων που χρησιμοποιείται είναι το GTZAN [9] που περιέχει 1000 τραγούδια από 30 s το καθένα και χωρισμένα σε 10 κατηγορίες: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae και rock. Κάθε κατηγορία περιέχει 100 τραγούδια.
Για την κατηγοριοποίηση των τραγουδιών χρησιμοποιήθηκαν διάφοροι στατιστικοί ταξινομητές αναγνώρισης μορφών (statistical pattern recognition classifiers). Μέσω αυτών εκτιμούμε την συνάρτηση πυκνότητας πιθανότητας (probability density function) για τα διανύσματα χαρακτηριστικών σε κάθε κατηγορία/είδος. Οι ταξινομητές συνεπώς είναι οι:

- Simple Gaussian classifier (GS)
- Gaussian mixture model classifier (GMM)
- K-nearest neighbor classifier (K-NN)

Λίγα λόγια για τους ταξινομητές [30], [31].
Για τους Gaussian classifiers, θεωρούμε ότι έχουμε μια βάση δεδομένων για εξάσκηση (training dataset) που κάνει δυαδική ταξινόμηση (χωρίζει στις τιμές/κλάσεις 1 και 2). Ο σκοπός είναι να προσδιοριστεί η κλάση όπου ανήκουν τα νέα δεδομένα *x* (new data). Θέλουμε να εκτιμήσουμε δηλαδή τα $p(y=1|x)$ και $p(y=2|x)$. Το *x* παίρνει την τιμή της κλάσης με την μεγαλύτερη πιθανότητα. Στο *σχήμα 4.3* βλέπουμε αυτήν την εκτίμηση.
Η τελική κλάση (class posterior) $p(y=c|x)$ είναι τα $p(y=1|x)$ και $p(y=2|x)$.
Η πυκνότητα πιθανότητας κλάσης (class-conditional density) $p(x|y=c)$ είναι γκαουσιανή/κανονική κατανομή. Λόγω αυτού καλείται και ο ταξινομητής

γκαουσιανός.
Η προηγούμενη κλάση (class prior) *p(y=c)* είναι η αναλογία του αριθμού της κλάσης/ αριθμό του συνολικού δείγματος. Το άθροισμα που βλέπουμε στο *σχήμα 4.3* είναι σταθερά κανονικοποίησης (normalization constant).



Σχήμα 4.3: Θεώρημα του Bayes για εκτίμηση της πιθανότητας της κατάταξης των νέων δεδομένων

Για τον k-NN τώρα έχουμε ότι είναι ένας επιβλεπόμενος αλγόριθμος μηχανικής μάθησης, δηλαδή βασίζεται στα ταξινομημένα input δεδομένα (data) ώστε να μάθει σε μια συνάρτηση να παράγει τα κατάλληλα output όταν της δίνονται μη-ταξινομημένα input. Ο συγκεκριμένος αλγόριθμος υποθέτει ότι παρόμοια πράγματα υπάρχουν σε κοντινή απόσταση. Η διαδικασία είναι η παρακάτω: 1) Τα δεδομένα φορτώνονται, 2) θέτουμε το *K* ως έναν συγκεκριμένο αριθμό γειτόνων, 3) για κάθε στοιχείο των δεδομένων υπολογίζουμε την απόσταση αυτού με κάποιου ζητούμενου στοιχείου και βάζουμε την απόσταση και τον δείκτη του στοιχείου σε μια ταξινομημένη συλλογή, 4) ταξινομούμε την συλλογή από χαμηλότερες σε υψηλότερες τιμές των αποστάσεων, 5) επιλέγουμε τα πρώτα *K* από την συλλογή, 6) παίρνουμε τις ετικέτες (labels) των επιλεγμένων *K*, 7) για την κατηγοριοποίηση παίρνουμε πίσω την κατάσταση (mode) των ετικετών *K*.

Επιστρέφοντας στην κατηγοριοποίηση μας, τα αποτελέσματα υπολογίζονται χρησιμοποιώντας ένα μοναδικό διάνυσμα 30 διαστάσεων για όλο το αρχείο των τραγουδιών. Αυτό προκύπτει με βάση τα χαρακτηριστικά που αναλύσαμε προηγουμένως. Έτσι έχουμε για τα χαρακτηριστικά του ηχοχρώματος 9 διαστάσεις των STFT και 10 των MFCCs άρα 19 διαστάσεις, ο ρυθμός έχει 6 διαστάσεις και η τονικότητα/οξύτητα έχει 5 διαστάσεις. Άρα συνολικά έχουμε 30 διαστάσεις. Από την βάση δεδομένων χρησιμοποιούνται το 10% των τραγουδιών για testing και το υπόλοιπο 90% για training. Η επιλογή γίνεται τυχαία. Υπολογίζουμε το confusion matrix των ειδών.

GENRE CONFUSION MATRIX

	cl	co	di	hi	ja	ro	bl	re	po	me
cl	**69**	0	0	0	1	0	0	0	0	0
co	0	**53**	2	0	5	8	6	4	2	0
di	0	8	**52**	11	0	13	14	5	9	6
hi	0	3	18	**64**	1	6	3	26	7	6
ja	26	4	0	0	**75**	8	7	1	2	1
ro	5	13	4	1	9	**40**	14	1	7	33
bl	0	7	0	1	3	4	**43**	1	0	0
re	0	9	10	18	2	12	11	**59**	7	1
po	0	2	14	5	3	5	0	3	**66**	0
me	0	1	0	1	0	4	2	0	0	**53**

Πίνακας 4.1: Το confusion matrix των μουσικών ειδών. Στη διαγώνιο φαίνεται το ποσοστό ακρίβειας (ή 'επιτυχίας') της κατάταξης, εάν δηλαδή τα τραγούδια αντιστοιχήθηκαν στην σωστή κατηγορία/είδος.

Το ποσοστό στο confusion matrix υπολογίζεται στα 100 τραγούδια. Στο έκτο κεφάλαιο θα εξηγήσουμε και θα ερμηνεύσουμε αναλυτικότερα τα αποτελέσματα αυτού.

	Genres(10)
Random	10
RT GS	44 ± 2
GS	59 ± 4
GMM(2)	60 ± 4
GMM(3)	61 ± 4
GMM(4)	61 ± 4
GMM(5)	61 ± 4
KNN(1)	59 ± 4
KNN(3)	60 ± 4
KNN(5)	56 ± 3

Πίνακας 4.2: Τα ποσοστά ακρίβειας όλης της μεθόδου χρησιμοποιώντας τους διάφορους ταξινομητές. Μέγιστο είναι το **61%**.

Στη συνέχεια υπολογίζουμε με τον GS ταξινομητή τα ποσοστά ακρίβειας για όλη την μέθοδο αλλά και τα διάφορα χαρακτηριστικά ξεχωριστά.

	Genres
RND	10
PHF(5)	23
BHF(6)	28
STFT(9)	45
MFCC(10)	47
FULL(30)	59

Πίνακας 4.3: Το ποσοστό ακρίβειας όλης τη μεθόδου στα 100 τραγούδια με GS classifier. Το RND είναι η τυχαία (random) ταξινόμηση. Το PHF είναι για το pitch histogram και αντίστοιχα το BHF για το beat histogram. Τα STFT και MFCC είναι για τα χαρακτηριστικά του ηχοχρώματος. Στην τελευταία γραμμή βλέπουμε το ολικό ποσοστό ακρίβειας **59%**.

Από τον *πίνακα 4.3* βλέπουμε ότι η τυχαία ταξινόμηση είναι και η πιο χαμηλή. Τα χαρακτηριστικά του ηχοχρώματος έχουν καλύτερα αποτελέσματα από αυτά των τονικοτήτων και του ρυθμού, τα οποία είναι αρκετά χαμηλά. Παρόλα αυτά ξεπερνάνε το ποσοστό της τυχαιότητας οπότε μας προσφέρουν πληροφορία για τα μουσικά είδη. Το ολικό ποσοστό ακρίβειας σε όλο το σύστημα είναι 59%. Το μέγιστο ποσοστό όμως που έφτασε το σύστημα φαίνεται στον *πίνακα 4.2* με τον Gaussian Mixture Model ταξινομητή να φτάνει στο 61%. Συνεπώς υπάρχει χώρος για βελτίωση, όπως θα δούμε και παρακάτω.

Κεφάλαιο 5

Κατηγοριοποίηση μουσικών ειδών μέσω της μεθόδου διασκορπισμού των κυματιδίων

5.1 Εισαγωγή

Σε αυτό το κεφάλαιο περιγράφουμε αναλυτικά τον κώδικα της αναφοράς [7] και παρουσιάζουμε την εκτέλεση αυτού. Η βασική ιδέα λοιπόν, όπως και στο προηγούμενο κεφάλαιο, είναι να ταξινομήσουμε μουσικά κομμάτια και τραγούδια ανάλογα με το μουσικό είδος στο οποίο ανήκουν. Η μέθοδος που χρησιμοποιούμε εδώ είναι η wavelet time scattering transform. Αφού τρέξουμε τον κώδικα προκύπτει ένας confusion matrix ώστε να διαπιστώσουμε το ποσοστό των τραγουδιών που βρέθηκαν στην «σωστή» τους κατηγορία (είδος) και συνεπώς έτσι να δούμε πόσο καλή είναι η μέθοδος που χρησιμοποιήσαμε για το συγκεκριμένο πρόβλημα. Τα αποτελέσματα θα εκτιμηθούν αναλυτικά και θα συγκριθούν με άλλες μεθόδους όπως και με την προηγούμενη, του διανύσματος 30-D, στο επόμενο κεφάλαιο.

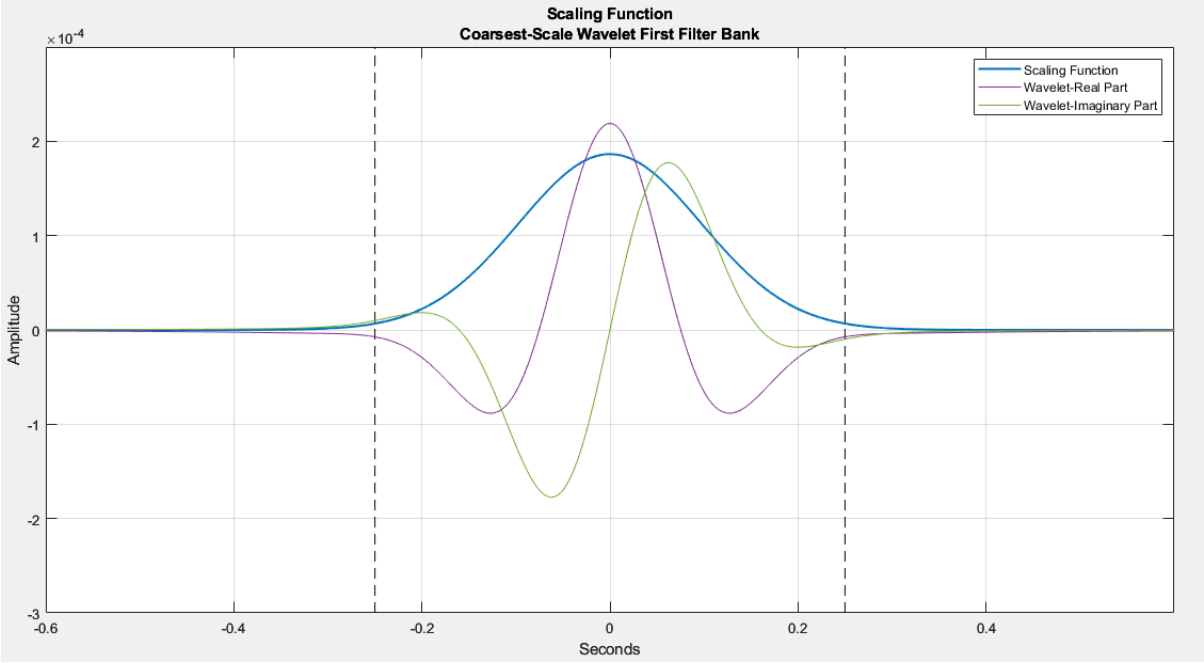
5.2 Βάση δεδομένων (GTZAN Dataset)

Ξεκινάμε από την βάση δεδομένων που θα χρησιμοποιήσουμε από το [9]. Αυτή χρησιμοποιείται σε πολλές εφαρμογές αλλά και στο [8] όπως είδαμε στο προηγούμενο κεφάλαιο, που είναι και οι δημιουργοί του. Αποτελείται από έναν φάκελο που περιέχει άλλους 10 υποφακέλους (subfolders). Κάθε υποφάκελος περιέχει 100 τραγούδια από ένα συγκεκριμένο είδος, διαφορετικό για κάθε φάκελο, τα οποία έχουν διάρκεια 30 δευτερολέπτων και συχνότητα δειγματοληψίας 22050Hz. Ο καθένας από τους 10 φακέλους ονομάζεται με βάση το μουσικό είδος που περιέχει, αυτά είναι τα εξής: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae και rock.

5.3 Πλαίσιο υλοποίησης

5.3.1 Αμετάβλητη κλίμακα

Χρησιμοποιούμε την MATLAB για την διεξαγωγή της μελέτης μας.
Όπως αναφέραμε εκτενώς και στο *κεφάλαιο 3*, οι παράμετροι που ορίζουμε είναι η διάρκεια της αμετάβλητης κλίμακας (invariance scale), ο αριθμός m των filter banks και ο αριθμός Q των κυματιδίων ανά οκτάβα.
Για τις περισσότερες εφαρμογές χρησιμοποιούμε $m=2$ filter banks (ή τάξεις του ST) και έτσι θα συμβεί και σε αυτήν. Στο πρώτο filter bank $m=1$ θα έχουμε $Q_1=8$ κυματίδια ανά οκτάβα και στο δεύτερο $m=2$ θα έχουμε $Q_2=1$.
Την αμετάβλητη κλίμακα (invariance scale) την θέτουμε 0.5 δευτερόλεπτα, που αντιστοιχεί σε λίγο περισσότερα από 11000 δείγματα για την συχνότητα δειγματοληψίας 22050Hz. Θέτουμε μήκος σήματος 2^{19} δείγματα.



Σχήμα 5.1: Η γραφική παράσταση της scaling συνάρτησης φ σε συνδυασμό με το πραγματικό και το φανταστικό μέρος των κυματιδίων $\psi_{\lambda 1}$ για το πρώτο filter bank. Παρατηρούμε ότι τα time support (χρονική τιμή) των συναρτήσεων δεν ξεπερνούν τα 0.5 s της invariance scale.

5.3.2 Βάση δεδομένων του ήχου

Όπως αναφέραμε και στην *ενότητα 5.2* η βάση δεδομένων που χρησιμοποιούμε είναι η GTZAN [9]. Κατεβάζουμε λοιπόν από εκεί το αρχείο με τους 10 φακέλους που ο καθένας έχει το όνομα του συγκεκριμένου είδους μουσικής που περιέχει. Στον κώδικα, που φαίνεται στο τέλος της εργασίας, στην εντολή `'IncludeSubFolders'` θέτουμε αληθοτιμή `true`, ώστε η βάση δεδομένων να χρησιμοποιήσει τους υποφακέλους (subfolders) και θέτουμε στην εντολή `'LabelSource'` το `'foldernames'` έτσι ώστε να ονομάσουμε τα δεδομένα με τους τίτλους των υποφακέλων. Πρέπει να προσέξουμε να έχουμε επιλέξει το σωστό μονοπάτι (path) που οδηγεί στον βασικό φάκελο που περιέχει τους υπόλοιπους υποφακέλους και ονομάζεται ‘genres’. Όλες οι εντολές φαίνονται στον κώδικα στο τέλος.

```
>> countEachLabel(ads)

ans =

  10×2 table

      Label       Count
    _________     _____

    blues          100
    classical      100
    country        100
    disco          100
    hiphop         100
    jazz           100
    metal          100
    pop            100
    reggae         100
    rock           100
```

Πίνακας 5.1: Τρέχω την εντολή ‘countEachLabel(ads)’ και προκύπτουν τα είδη που περιέχει ο φάκελος ‘genre’ και ο αριθμός των τραγουδιών που περιέχει κάθε είδος.

Συνεπώς προκύπτουν τα 10 μουσικά είδη με τα 100 τραγούδια το καθένα.

5.3.3 Σειρές Training και Test

Χωρίζουμε τα τραγούδια στην τύχη, με την βοήθεια της συνάρτησης ‘shuffle’, σε δύο σειρές (sets). Μια θα είναι για εξάσκηση και ανάπτυξη του ταξινομητή (classifier) μας και η άλλη θα είναι για να τον τεστάρουμε. Το 80%, λοιπόν, των δεδομένων μας θα είναι για training και το 20% για testing. Και ύστερα χρησιμοποιούμε την συνάρτηση ‘splitEachLabel’ για να κάνουμε τον διαχωρισμό των τραγουδιών σε 80-20. Με αυτήν την εντολή είμαστε σίγουροι ότι όλα τα είδη αντιπροσωπεύονται ισάξια.

```
>> countEachLabel(adsTrain)

ans =

  10×2 table

      Label       Count
    _________     _____

    blues          80
    classical      80
    country        80
    disco          80
    hiphop         80
    jazz           80
    metal          80
    pop            80
    reggae         80
    rock           80
```

```
>> countEachLabel(adsTest)

ans =

  10×2 table

      Label       Count
    _________     _____

    blues          20
    classical      20
    country        20
    disco          20
    hiphop         20
    jazz           20
    metal          20
    pop            20
    reggae         20
    rock           20
```

Πίνακας 5.2: Αριστερά έχουμε το training set και δεξιά το testing set

Συνεπώς έχουμε 800 τραγούδια για training και 200 τραγούδια για test.

Η συνάρτηση 'audioDatastore' λειτουργεί με tall arrays (ψηλές σειρές/ψηλούς πίνακες) στην MATLAB. Αυτοί είναι πίνακες με περισσότερες γραμμές από όσες χωράνε στην μνήμη, δηλαδή τους χρησιμοποιούμε όταν θέλουμε να μελετήσουμε δεδομένα τα οποία ανήκουν σε βάσεις δεδομένων που μπορεί να έχουν εκατομμύριες γραμμές. Άρα δημιουργώντας tall arrays και για τις δύο σειρές (sets) που έχουμε ξεκινάει να τρέχει παράλληλα η 'parallel pool (parpool) MATLAB' για τον υπολογισμό.

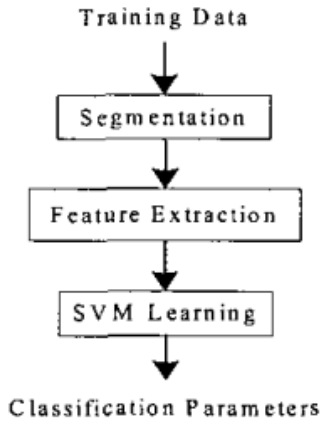
Τώρα θέλουμε να πάρουμε τα χαρακτηριστικά του scattering. Για να γίνει αυτό χρησιμοποιούμε την συνάρτηση 'helperscatfeatures' που υπάρχει στο Appendix του κώδικα. Αυτή λοιπόν η συνάρτηση μας επιστρέφει τον πίνακα των χαρακτηριστικών του wavelet time scattering για το σήμα που μελετάμε. Παίρνουμε τον φυσικό λογάριθμο των συντελεστών scattering και ο πίνακας υπολογίζεται για 2^{19} δείγματα (το μήκος του σήματός μας) σε κάθε μουσικό δείγμα (audio file). Υπολογίζουμε αυτά τα scattering χαρακτηριστικά και για το training και το testing. Αφού γίνει ο υπολογισμός, πακετάρουμε τα scattering χαρακτηριστικά για κάθε περίπτωση σε έναν πίνακα.

Ο μετασχηματισμός scattering του ηχητικού σήματος βγάζει 341 μονοπάτια (paths) και κάθε σειρά των πινάκων 'TrainFeatures' και 'TestFeatures' είναι ένα χρονικό παράθυρο scattering (scattering time window) σε αυτά τα paths. Για κάθε μουσικό δείγμα έχουμε 32 τέτοια χρονικά παράθυρα. Συνεπώς ο πίνακας για το training data θα έχει διαστάσεις 25600 x 341 και ο αντίστοιχος του test

data θα έχει διαστάσεις 6400 x 341. Ο αριθμός των γραμμών στους πίνακες προκύπτει από τον πολλαπλασιασμό του αριθμού των τραγουδιών που χρησιμοποιούνται με τα 32 παράθυρα ($32 \cdot 800 = 25600$ και $32 \cdot 200 = 6400$). Υπάρχουν λοιπόν 200 παραδείγματα του test set και 32 παράθυρα για κάθε παράδειγμα. Δημιουργούμε μια 'ετικέτα' είδους (genre label) για κάθε ένα από αυτά τα 32 παράθυρα, για κάθε σειρά δηλαδή στον πίνακα χαρακτηριστικών wavelet scattering για τα δεδομένα του training.

Ο ταξινομητής (classifier) που θα χρησιμοποιήσουμε για την κατανομή των τραγουδιών είναι ο πολλαπλών τάξεων (multi-class) SVM (Support Vector Machine). Είναι μια τεχνική μηχανικής μάθησης που χρησιμοποιείται συχνά για αναγνώριση μορφών (pattern recognition) [13]. Ο τρόπος που λειτουργεί είναι να μετασχηματίζει τα input διανύσματα, όπου σε αυτήν την μελέτη το input data είναι το training data $(\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{n})$ και τα μουσικά είδη όπου ανήκει το training data $(\mathrm{y}_1, \dots, \mathrm{y}_{10})$, σε έναν χώρο μεγάλης διάστασης (feature space) μέσω ενός μη γραμμικού μετασχηματισμού Φ και στη συνέχεια να κάνει έναν γραμμικό διαχωρισμό σε αυτόν τον χώρο (feature space). Συνεπώς έτσι καθορίζει σε ποιο μουσικό είδος θα κατανεμηθούν τα training data. Για την κατασκευή, λοιπόν, ενός μη γραμμικού SVM ταξινομητή αντικαθιστούμε τα input $< \mathrm{x}, \mathrm{y} >$ με μια συνάρτηση Kernel $\mathrm{K}(\mathrm{x}, \mathrm{y})$ (συνάρτηση πυρήνα). Υπάρχουν τρία είδη Kernel για την κατασκευή διαφορετικών μηχανών μάθησης. Εμείς θα χρησιμοποιήσουμε τον polynomial kernel (πυρήνα πολυωνύμου) βαθμού *d*.

$$\mathrm{K}(\mathrm{x}, \mathrm{y}) = (< x, y > +1)^{\mathrm{d}} \qquad (5.1)$$



Σχήμα 5.2: SVM learning process, όπου segmentation είναι ο διαχωρισμός των δεδομένων του training σε τμήματα

Στο *σχήμα 5.2* συνεπώς βλέπουμε την διαδικασία για τον SVM ώστε να γίνει η σωστή ταξινόμηση των μουσικών σημάτων στα αντίστοιχά τους μουσικά είδη.

5.4 Αποτελέσματα της σειράς 'test'

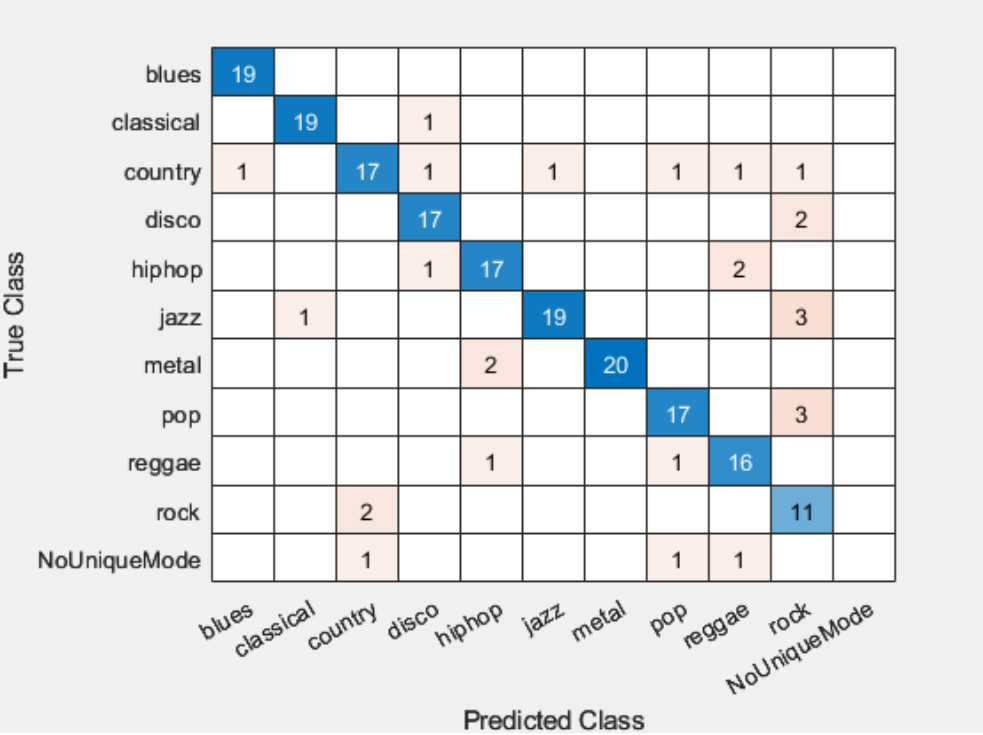
Χρησιμοποιούμε το SVM στον μετασχηματισμό scattering για το training set για να προβλέψουμε τα αποτελέσματα στο test set. Όπως αναφέραμε και προηγουμένως έχουμε 32 χρονικά παράθυρα για κάθε ηχητικό σήμα στον ST. Χρησιμοποιούμε την συνάρτηση 'helperMajorityVote' για να προβλέψουμε το είδος των ηχητικών σημάτων. Αυτή η συνάρτηση μας επιστρέφει την κατάσταση (mode) της προβλεπόμενης 'ετικέτας' μουσικού είδους (class label) για τον αριθμό διανυσμάτων των χαρακτηριστικών, όπου εδώ είναι τα 32 παράθυρα που έχει κάθε μουσικό δείγμα. Αν δεν βρέθηκε μοναδική αντιστοιχία, κάποια μοναδική κατάσταση δηλαδή που να ανήκουν τα 32 παράθυρα ενός συγκεκριμένου τραγουδιού, τότε το συγκεκριμένο μουσικό παράδειγμα κατατάσσεται στην ετικέτα 'NoUniqueMode' που αποτελεί και το σφάλμα του ταξινομητή μας. Τα αποτελέσματα του test data τα παίρνουμε σε ένα confusion matrix, που φαίνεται στον *πίνακα 5.3*. Θυμόμαστε ότι κάθε είδος περιέχει 20 τραγούδια. Επίσης τρέχουμε την εντολή 'testAccuracy' που θα μας βγάλει το συνολικό ποσοστό ακριβείας, για όλο data set που χρησιμοποιήθηκε και συνεπώς θα έχουμε μια γενική εικόνα για την λειτουργία και αποτελεσματικότητα του προγράμματός μας.

```
>> testAccuracy = sum(eq(TestVotes,adsTest.Labels))/numTestSignals*100

testAccuracy =

    86
```

Σχήμα 5.3: Το ποσοστό ακρίβειας συνεπώς για αυτό το παράδειγμα προέκυψε **86%.**

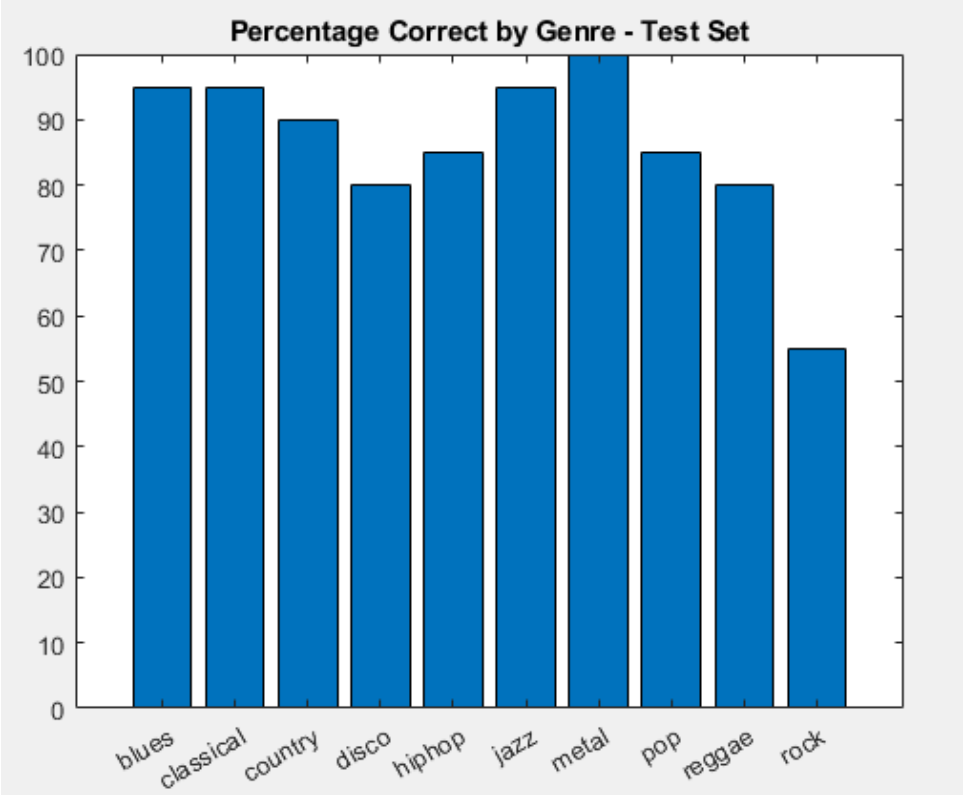
Στο confusion matrix, που φαίνεται στον *πίνακα 5.3*, οι γραμμές αντιστοιχούν στο προβλεπόμενο είδος και οι στήλες στο πραγματικό είδος. Στη διαγώνιο μπορούμε να δούμε το ποσό της σωστής κατανομής, όπου δηλαδή το πραγματικό και το προβλεπόμενο είδος ταυτίζονται. Παρατηρούμε καλά αποτελέσματα για τα περισσότερα μουσικά είδη. Η rock έχει την χειρότερη

ακρίβεια και σχεδόν τα μισά παραδείγματα έχουν πέσει εκτός, αυτό λόγω του ευρέος φάσματος του συγκεκριμένου είδους μουσικής. Θα αναφερθούμε πιο αναλυτικά σε αυτά στο επόμενο κεφάλαιο, όπου και θα τα συγκρίνουμε με τα confusion matrix που προέκυψαν από την μελέτη στο [8].



Πίνακας 5.3: Το confusion matrix. Από την διαγώνιο του πίνακα προκύπτει ότι η ακρίβεια για κάθε είδος, πέραν της rock, είναι αρκετά καλή. Θυμόμαστε ότι τα τραγούδια που χρησιμοποιήθηκαν για το test set είναι 20.

Τέλος, παίρνουμε το ποσοστό ακρίβειας για κάθε είδος ξεχωριστά σε ένα ραβδόγραμμα και βλέπουμε για άλλη μια φορά ότι η μέθοδος ανάλυσης που χρησιμοποιήσαμε ήταν πολύ ικανοποιητική για την κατηγοριοποίηση των τραγουδιών στα περισσότερα μουσικά είδη. Αν εξαιρέσουμε την ροκ που έχει ποσοστό ακρίβειας 55% όλα τα υπόλοιπα είδη είναι πάνω από 80%.



Σχήμα 5.4: Τα ποσοστά ακρίβειας για κάθε είδος ξεχωριστά.

Το ποσοστό ακρίβειας που προέκυψε για την εκτέλεση του κώδικα στο [7] ήταν 88%, ενώ σε εμάς 86%. Δεν έχουν μεγάλη διαφορά όμως αυτό το 2% προκύπτει από την τυχαία επιλογή τραγουδιών στα training και test sets. Συνειδητοποιούμε για ακόμα μια φορά την δυσκολία ταξινόμησης των τραγουδιών στις «σωστές» τους κατηγορίες.

Παρόλα αυτά, η μέθοδος wavelet scattering έβγαλε πολύ καλά αποτελέσματα και συγκριτικά με την μέθοδο του προηγούμενου κεφαλαίου υπερτερεί κατά πολύ. Στο επόμενο κεφάλαιο θα συγκρίνουμε αυτές τις δύο μεθόδους μεταξύ τους, εξηγώντας πιο αναλυτικά τα αποτελέσματα που προέκυψαν.

Κεφάλαιο 6

Σύγκριση των αποτελεσμάτων

6.1 Εισαγωγή

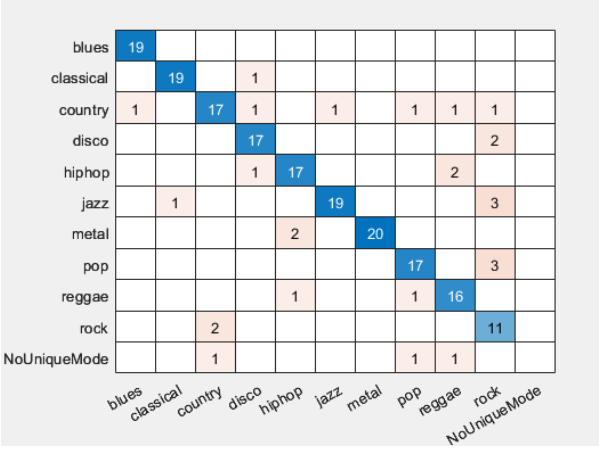
Σε αυτό το κεφάλαιο θα συγκρίνουμε τα αποτελέσματα της μεθόδου Wavelet Time Scattering, με αυτά της μεθόδου με το διάνυσμα των 30 διαστάσεων που είδαμε. Αυτό γίνεται εύκολα εφόσον και οι δύο μέθοδοι χρησιμοποιούν την ίδια βάση δεδομένων, GTZAN. Στη συνέχεια θα δούμε τα αποτελέσματα της μελέτης του [15], όπου χρησιμοποιούνται διάφορες μέθοδοι βασισμένες στα wavelets, και θα τα συγκρίνουμε και αυτά με τα προηγούμενα. Τέλος, θα δούμε την ακρίβεια της ανθρώπινης ακοής όσον αφορά στην κατάταξη τραγουδιών σε μουσικά είδη και θα εξάγουμε τα συμπεράσματά μας.

6.2 Σύγκριση αποτελεσμάτων της μεθόδου wavelet scattering με την μέθοδο του διανύσματος 30-D

Αρχικά βλέποντας το ποσοστό ακρίβειας για κάθε μέθοδο έχουμε για wavelet scattering 86% και για την μέθοδο του διανύσματος των 30 διαστάσεων 61%. Βλέπουμε ήδη μια ξεκάθαρη «νίκη» για την μέθοδο scattering. Παρακάτω θα εξηγήσουμε αναλυτικά τα αποτελέσματα των confusion matrices σε κάθε περίπτωση. Ας τους ξαναδούμε διπλά-δίπλα.

GENRE CONFUSION MATRIX

	cl	co	di	hi	ja	ro	bl	re	po	me
cl	**69**	0	0	0	1	0	0	0	0	0
co	0	**53**	2	0	5	8	6	4	2	0
di	0	8	**52**	11	0	13	14	5	9	6
hi	0	3	18	**64**	1	6	3	26	7	6
ja	26	4	0	0	**75**	8	7	1	2	1
ro	5	13	4	1	9	**40**	14	1	7	33
bl	0	7	0	1	3	4	**43**	1	0	0
re	0	9	10	18	2	12	11	**59**	7	1
po	0	2	14	5	3	5	0	3	**66**	0
me	0	1	0	1	0	4	2	0	0	**53**



Σχήμα 6.1: Αριστερά ο confusion matrix από την μέθοδο του διανύσματος των 30 διαστάσεων. Δεξιά ο confusion matrix από το wavelet scattering.

Να θυμίσουμε ότι στον confusion matrix στα αριστερά είναι τα τραγούδια που βρέθηκαν στη σωστή κατηγορία για 100 τραγούδια, ενώ στα δεξιά για 20 τραγούδια. Ένας confusion matrix λειτουργεί ως εξής. Στον πίνακα οι γραμμές αντιστοιχούν στο προβλεπόμενο είδος και οι στήλες στο πραγματικό είδος, έτσι λοιπόν στη διαγώνιο μπορούμε να δούμε κατά πόσο το πραγματικό και το προβλεπόμενο είδος ταυτίζονται. Ας πάρουμε για παράδειγμα την κλασσική μουσική. Αριστερά βλέπουμε ένα ποσοστό 69% το οποίο βρέθηκε στο σωστό είδος. Όμως μπορούμε να δούμε και ένα 26% που βρέθηκε στη τζαζ και το τελευταίο 5% που βρέθηκε στη ροκ μουσική. Αντίστοιχα δεξιά βλέπουμε 19/20 τραγούδια = 95% που βρέθηκε στην σωστή κατηγορία και μόνο το 5% που βρέθηκε σε διαφορετικό είδος και συγκεκριμένα στη τζαζ. Για τον αριστερό πίνακα παρατηρούμε ότι μόνο για την κλασσική μουσική τύχαινε ένα σχετικά μεγάλο ποσοστό να συγκεντρωθεί σε διαφορετικό είδος. Στα υπόλοιπα είδη παρότι το ποσοστό επιτυχίας τους είναι μικρότερο από την κλασσική μουσική, οι λάθος αντιστοιχήσεις συνέβησαν σε πολλαπλά είδη. Το υψηλότερο ποσοστό στον αριστερό πίνακα τον έχει η τζαζ με 75% ενώ στον δεξιό πίνακα η metal μουσική με 100% επιτυχία. Αυτό συμβαίνει διότι χρησιμοποιείται διαφορετική μέθοδος ανάλυσης στις δύο μελέτες επομένως αναλόγως με τα διαφορετικά μουσικά είδη μια συγκεκριμένη μέθοδος θα είναι πιο ισχυρή (καλή) για το κάθε είδος. Στο confusion matrix του WST πλην της ροκ μουσικής όλα τα υπόλοιπα είδη έχουν ποσοστό ακρίβειας από 80% και πάνω. Εκεί όπου συμφωνούν και οι δύο πίνακες είναι για την κατάταξη της ροκ μουσικής όπου και στους δύο έχει το πιο χαμηλό ποσοστό ακρίβειας με 40% στον αριστερό πίνακα και μόλις 55% στον δεξιό.

Η ροκ μουσική έχει τόσο ευρύ φάσμα τραγουδιών και συνθετών και τόσες υποκατηγορίες που ανήκουν σε αυτήν που για αυτόν τον λόγο είναι και το πιο δύσκολο είδος προς ταξινόμηση.

Όσον αφορά την κλασσική και την τζαζ μουσική είναι λογικό να συγχέονται μιας και χρησιμοποιούν πολύ παρόμοια μουσικά όργανα, αλλά και μερικά κλασσικά κομμάτια με ισχυρό ρυθμό θα μπορούσαν να ταξινομηθούν λανθασμένα ως τζαζ τραγούδια.

6.3 Σύγκριση με διάφορες μεθόδους που χρησιμοποιούν WT

Στο [15] έχουμε ακόμη μια μελέτη πιο πρόσφατη για κατηγοριοποίηση της μουσικής σε συγκεκριμένα είδη όπου και χρησιμοποιείται ξανά η βάση δεδομένων GTZAN από το [9]. Σε αυτήν την μελέτη λοιπόν, χρησιμοποιούνται διάφορες μέθοδοι ξεχωριστά, για σύγκριση. Αυτές είναι οι: Conventional 1-D feature extraction using Fourier Transform, Mel Spectograms, Discrete Wavelet Transform, Extracting and Down sampling scalogram features, Dual-tree complex wavelet transform και wavelet scattering transform.

Αναφέρουμε περιληπτικά κάποια στοιχεία για τις διάφορες μεθόδους ανάλυσης που χρησιμοποιούνται. Τον Discrete Wavelet Transform (DWT) τον είδαμε αναλυτικά στο *κεφάλαιο 2* στην *ενότητα 2.4.5*. Τα scalogram είναι η γραφική παράσταση (συνάρτηση χρόνου-κλίμακας) της απόλυτης τιμής του CWT ενός σήματος. Με τη διαδικασία της υποδειγματοληψίας (downsampling) του ηχητικού σήματος προκύπτει μια δισδιάστατη scalogram εικόνα που στη συνέχεια χρησιμοποιείται ως input σε ένα Convolutional Neural Network (CNN) μοντέλο. Το Dual-tree complex wavelet transform (DTCWT) είναι μια βελτιωμένη έκδοση του DWT. Χρησιμοποιεί δύο πραγματικούς DWTs, ο πρώτος DWT δίνει το πραγματικό τμήμα του μετασχηματισμού και ο δεύτερος DWT δίνει το φανταστικό τμήμα του μετασχηματισμού. Ο DTCWT είναι χρήσιμος όταν έχουμε περίπλοκους μετασχηματισμούς wavelet.

Οι ταξινομητές (classifiers) που χρησιμοποιήθηκαν είναι οι: Random Forest (RD), Gradient boosting (GB), Kernel SVM και άλλοι, αλλά οι 2 πρώτοι είναι οι κυρίαρχοι. Τα αποτελέσματα για όλες τις μεθόδους προκύπτουν στον *πίνακα 6.1*.

Θα δούμε γενικότερα ότι οι μέθοδοι που περιέχουν wavelets υπερτερούν στις υπόλοιπες όσον αφορά την ταξινόμηση της μουσικής. Πιο συγκεκριμένα όπως φαίνεται από τον *πίνακα 6.1* και εδώ η καλύτερη μέθοδος με το

μεγαλύτερο ποσοστό ακρίβειας στο 88% είναι ο Wavelet scattering μετασχηματισμός. Η 2% διαφορά που προέκυψε με την δική μας εκτέλεση μπορεί να οφείλεται στην τυχαιότητα της επιλογής τραγουδιών από την λίστα GTZAN και συνεπώς μικρή αλλαγή στα αποτελέσματα. Η Dual-tree complex wavelet transform επίσης φαίνεται να βγάζει πολύ καλά αποτελέσματα με 85% ακρίβεια, όμως δεν θα ασχοληθούμε αναλυτικότερα με αυτήν στην παρούσα εργασία.

SUMMARY OF RESULTS

FEATURE EXTRACTION	TEST ACCURACY (%)
1-D FT	67
Mel Spectrogram	73
DWT	**81.5**
Scalogram	**60**
DTCWT	**85**
Wavelet Scattering	**88**

Πίνακας 6.1: Αποτελέσματα ποσοστών ακρίβειας για τις διάφορες μεθόδους

Στον *πίνακα 6.2* φαίνονται τα χαρακτηριστικά για ανάλυση μουσικής που χρησιμοποιούν τον 1-D Fourier Transform. Πολλά από αυτά τα έχουμε δει στο προηγούμενο κεφάλαιο αναλυτικά.

1-D FEATURE SET USING FOURIER TRANSFORM
MFCCs
Spectral Centroid
Chroma frequencies
Spectral roll-off
Root Mean Square
Zero Crossing Rate
Spectral Bandwidth

Πίνακας 6.2: Χαρακτηριστικά που προκύπτουν από 1-D FT

Βλέπουμε λοιπόν από τον *πίνακα 6.1* ότι τα χαρακτηριστικά που προκύπτουν από τον 1-D FT έχουν το χαμηλότερο ποσοστό ακρίβειας στο 67%. Και το

ποσοστό ακρίβειας του Discrete wavelet μετασχηματισμού είναι στο 81.5%. Ο DWT εδώ χρησιμοποιείται σαν filter-bank που σημαίνει ότι χωρίζει το σήμα σε μικρότερες περιοχές συχνοτήτων (frequency sub-bands) και συνεπώς έχουμε «καταρράκτη» (cascade) από υψηλής διέλευσης και χαμηλής διέλευσης (high-pass and low-pass) φίλτρα. Στο *κεφάλαιο 4* είχαμε χρησιμοποιήσει αυτές τις δύο μεθόδους συνδυαστικά, με τον DWT να μας βγάζει το Beat Histogram των σημάτων, αλλά και με τα έξτρα χαρακτηριστικά από τα Pitch Histogram. Εκεί το ποσοστό ακρίβειας ήταν πολύ χαμηλότερο από αυτά που βλέπουμε εδώ, εφόσον έφτανε μόλις το 61%. Αυτό μπορεί να οφείλεται, πέρα από το γεγονός ότι ο DWT (που βγάζει και τα καλύτερα αποτελέσματα) χρησιμοποιούνταν διαφορετικά, στο ότι χρησιμοποιούνται διαφορετικοί ταξινομητές (classifiers). Πιθανώς αυτοί που χρησιμοποιούνται στην συγκεκριμένη μελέτη, αλλά και ο SVM που χρησιμοποιήσαμε στο προηγούμενο κεφάλαιο είναι αρκετά καλύτεροι. Τα Mel spectrogram και scalogram με 73% και 60% ποσοστό ακρίβειας αντίστοιχα δεν θα μελετηθούν στην παρούσα εργασία.

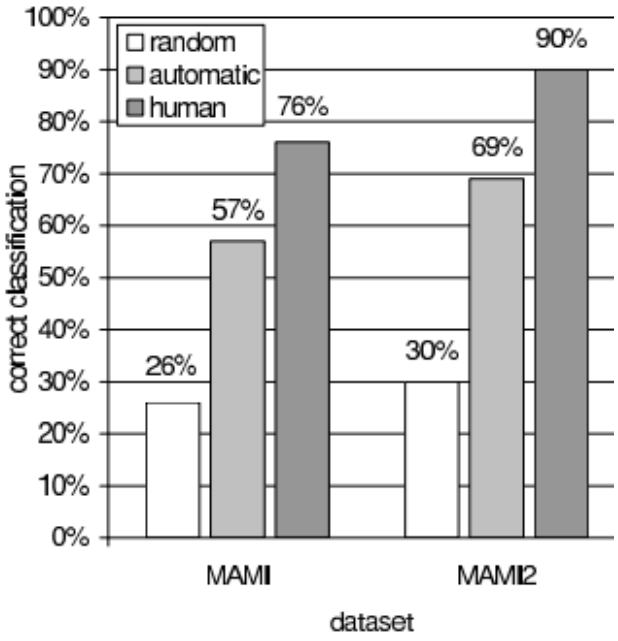
6.4 Η ανθρώπινη ακοή/ ο ανθρώπινος παράγοντας

Όπως αναφέρθηκε και προηγουμένως πριν δημιουργηθούν οι αυτόματες μέθοδοι για κατηγοριοποίηση της μουσικής σε διαφορετικά είδη, αυτή η διαδικασία γινόταν μέσω της ακοής από ειδικούς και το μουσικό είδος καθοριζόταν από τις διάφορες εταιρείες μουσικής εξαρχής. Στο [16] βλέπουμε ένα πείραμα που έγινε τη δεκαετία του 1990 σε 52 πρωτοετείς φοιτητές ψυχολογίας για την κατάταξη διαφόρων τραγουδιών που άκουγαν σε συγκεκριμένα προτεινόμενα είδη. Οι φοιτητές αυτοί δεν είχαν κάποια μουσική εκπαίδευση, απλώς άκουγαν συχνά μουσική που τους άρεσε στην καθημερινότητά τους. Τα 10 είδη μου χρησιμοποιήθηκαν ως βάση δεδομένων ήταν τα εξής: Blues, Classical, Country, Dance, Jazz, Latin, Pop, Rhythm and blues (R&B), Rap, Rock. Τα χρόνια που ακολούθησαν πολλά από αυτά τα είδη άλλαξαν ή συγχωνεύτηκαν σε άλλα, αλλά την χρονιά του πειράματος ήταν από τα επικρατέστερα είδη στην αγορά. Κάθε είδος περιείχε 8 τραγούδια, 4 από τα οποία ήταν μόνο με όργανα και τα υπόλοιπα 4 και με φωνή. Η συνολική διάρκειά τους ήταν 3 s αλλά οι φοιτητές μπορεί να άκουγαν διάρκειες των 250, 325, 400, 475 και 3000 ms.

Τα προκύπτοντα αποτελέσματα είναι τα εξής:
Για την διάρκεια των 3000 ms το ποσοστό ακρίβειας ήταν **70%**. Η τυχαία ταξινόμηση ανέρχεται στα 10%. Για μικρότερες διάρκειες των τραγουδιών το

ποσοστό ακρίβειας ήταν 54.62% για τραγούδια που περιείχαν μόνο όργανα και 50.04% για τραγούδια και με φωνή. Αυξάνοντας πάνω από 3 s τα τραγούδια δεν καλυτέρευε το αποτέλεσμα. Επίσης έγινε ένα μικρό πείραμα με φοιτητές που σπούδαζαν θεωρία της μουσικής και τα αποτελέσματά τους δεν είχαν διαφορά από αυτά των φοιτητών ψυχολογίας, συνεπώς δεν έχει σημασία η γνώση μουσικής για την εκτέλεση αυτού του πειράματος. Βλέπουμε λοιπόν ότι τα αποτελέσματα του ανθρώπινου παράγοντα σε σχέση με το WST είναι χειρότερα. Παρόλα αυτά, δεν μπορεί να γίνει άμεση σύγκριση, εφόσον δεν χρησιμοποιήθηκε η ίδια βάση δεδομένων.

Στο [17] έχουμε ένα πείραμα ξανά που αυτήν την φορά συγκρίνει την μέθοδο του *κεφαλαίου 4* του διανύσματος των 30 διαστάσεων (G. Tzanetakis) με την ανθρώπινη ακοή. Η βάση δεδομένων που χρησιμοποιείται είναι η MAMI που περιέχει 160 τραγούδια ολόκληρης της διάρκειάς τους. Η κατηγοριοποίηση αυτής της βάσης δεδομένων έγινε από 27 ανθρώπους που άκουσαν 30 δευτερόλεπτα και από τα 160 τραγούδια και τα κατηγοριοποιούσαν σε 6 διαφορετικά είδη: classical, dance, pop, rap, rock ή άλλο. Για κάθε τραγούδι καθοριζόταν ο αριθμός των ψήφων Q για το μουσικό είδος στο οποίο αυτό μπορεί να ανήκει. Ο μέγιστος αριθμός των ψήφων για το είδος κατάτασσε το τραγούδι στο συγκεκριμένο είδος. Έτσι προέκυψε η εξής κατάταξη για το MAMI dataset: 24 classical, 18 dance, 69 pop, 8 rap, 25 rock και 16 άλλο. Για κάθε άτομο γινόταν σύγκριση της απάντησής του με το επιλεγμένο είδος του τραγουδιού και τελικά το ποσοστό ακρίβειας για την ανθρώπινη ακοή ήταν 76% κατά μέσο όρο, με ποσοστά που ξεκινούσαν από 57% έως 86% για κάθε άνθρωπο ξεχωριστά. Φαίνεται από τα παραπάνω ότι η υποκειμενικότητα του κάθε ανθρώπου για κάθε μουσικό είδος μπορεί να επηρεάσει τα αποτελέσματα. Συνεπώς δημιουργείται μια καινούρια βάση δεδομένων MAMI2, όπου από το MAMI dataset επιλέγονται τα 98 τραγούδια που δεν ανήκουν στην κατηγορία ‘άλλο’ και ταυτόχρονα ο μέγιστος αριθμός των ψήφων Q_{max} να είναι πάνω από 18. Αυτή η βάση έβγαλε πολύ καλύτερα αποτελέσματα αφού ο μέσος όρος του ποσοστού ακρίβειας έφτασε το 90%.



Σχήμα 6.2: Σύγκριση των αποτελεσμάτων της αυτόματης με της ανθρώπινης μεθόδου για τα 2 διαφορετικά dataset.

Για την μέθοδο του G. Tzanetakis ισχύουν τα ίδια που αναφέραμε στο *κεφάλαιο 4*. Το ποσοστό στην μια περίπτωση είναι 57%, πολύ κοντά στο 61% που είδαμε προηγουμένως. Στην δεύτερη περίπτωση όπου και η βάση των δεδομένων είναι λιγότερο υποκειμενική έχουμε και καλύτερο ποσοστό ακρίβειας στο 69%. Η τυχαία κατανομή παραμένει χαμηλά και στις δύο περιπτώσεις. Επομένως βλέπουμε για δεύτερη φορά ότι ο ανθρώπινος παράγοντας εξάγει καλύτερα αποτελέσματα από την μέθοδο του διανύσματος των 30 διαστάσεων του *κεφαλαίου 4*, συνεπώς για αυτό η μέθοδος του wavelet scattering φάνηκε πολύ χρήσιμη για την κατηγοριοποίηση τραγουδιών σε μουσικά είδη.

Συμπεράσματα

Σε αυτήν την διπλωματική εργασία μελετήσαμε διάφορες μεθόδους ανάλυσης για την κατηγοριοποίηση μουσικών κομματιών και τραγουδιών στα μουσικά είδη που αυτά αντιστοιχούν. Είδαμε ότι κάποιες λειτούργησαν πολύ καλά με μεγάλο ποσοστό ακρίβειας, ενώ άλλες λιγότερο καλά, παρόλα αυτά το ποσοστό ακρίβειας τους ήταν πάντα πάνω από το 50% και πάντα πάνω από την τυχαία κατανομή (random). Οι μέθοδοι ανάλυσης που είδαμε στην αυτόματη κατηγοριοποίηση είναι η μέθοδος του διανύσματος των 30-D και ο wavelet scattering transform, με έμφαση στον τελευταίο.

Αρχικά η κατηγοριοποίηση μέσω της ανθρώπινης ακοής χρησιμοποιήθηκε σαν βάση για την αυτόματη κατηγοριοποίηση παρά τα μειονεκτήματα που μπορεί να έχει αυτή, όπως το πρόβλημα της υποκειμενικότητας του κάθε ανθρώπου και του πως το κάθε άτομο αντιλαμβάνεται τα μουσικά είδη αλλά έχει και τις δικές του προτιμήσεις για αυτά, όπως είδαμε και στο προηγούμενο κεφάλαιο. Δίνουμε λοιπόν βαρύτητα στις αυτόματες μεθόδους ανάλυσης.

Η επιλογή του κατάλληλου ταξινομητή (classifier) είναι πολύ σημαντική, όπως φάνηκε και στην *ενότητα 6.3*. Υπάρχουν κάποιοι ταξινομητές που για την συγκεκριμένη μελέτη της ταξινόμησης ηχητικών σημάτων σε μουσικά είδη, λειτουργούν πολύ καλύτερα από άλλους, όπως για παράδειγμα ο Support Vector Machine. Όμως ο SVM χάνει πολύ χρόνο στην διαδικασία του training, ειδικά αν υπάρχουν πολλά δείγματα σε αυτό [13]. Για αυτό και είναι σημαντική η επιλογή της κατάλληλης συνάρτησης πυρήνα (kernel), ώστε να καθοριστούν οι σχετικές παράμετροι. Επομένως υπάρχει χώρος για βελτίωση προς εκείνη την κατεύθυνση.

Επιπροσθέτως, για καλύτερα αποτελέσματα της κατηγοριοποίησης, πιο ακριβή δηλαδή, θα πρέπει να εξερευνηθούν περισσότερα μουσικά χαρακτηριστικά για την μελέτη του μουσικού περιεχομένου.

Επιπλέον πολλές φορές για να κατηγοριοποιήσουμε μουσικά είδη, επιλέγουμε συγκεκριμένους ερμηνευτές που θεωρούμε ότι ανήκουν σε μονάχα ένα μουσικό είδος και ταξινομούμε όλα τα άλμπουμ και τραγούδια τους σε μια κατηγορία. Αυτό μπορεί να ισχύει για κάποιους ερμηνευτές (αν και ποτέ πλήρως), αλλά κάποιοι άλλοι μπορεί να είχαν μεγάλη γκάμα τραγουδιών σε όλη την διάρκεια της καριέρας τους και συνεπώς να ανήκουν σε περισσότερα μουσικά είδη. Για αυτό και είναι σημαντικό η κατηγοριοποίηση να γίνεται σε κάθε τραγούδι ξεχωριστά και όχι σε ερμηνευτές.

Πρέπει να προσέξουμε επίσης το πρόβλημα της λανθασμένης ταξινόμησης. Το να κατηγοριοποιήσουμε ένα τραγούδι hard rock ως heavy metal δεν είναι τόσο σοβαρό όσο το να το κατηγοριοποιούσαμε ως jazz ή disco [29]. Πρέπει συνεπώς, να λάβουμε υπόψη αυτό το πρόβλημα κατά την διάρκεια της διαδικασίας του training και της αξιολόγησης, για να προκύψουν ακριβή αποτελέσματα.

Να θυμίσουμε όμως εδώ ότι γενικά η κατηγοριοποίηση τραγουδιών σε μουσικά είδη είναι δύσκολη διαδικασία λόγω του ευρέως φάσματος που έχουν πολλά είδη, όπως η ροκ, και ειδικότερα πλέον με την ύπαρξη πολλών υποκατηγοριών για το κάθε είδος. Οι μελέτες που παρουσιάστηκαν σε αυτήν την εργασία είχαν σχετικά μικρές βάσεις δεδομένων με μόλις 10 μουσικά είδη που αποτελούν και τα πιο «βασικά», ενώ θα ήταν ενδιαφέρον ένα πείραμα πάνω σε περισσότερες κατηγορίες και υποκατηγορίες τους. Επίσης πολλά μουσικά είδη μπορούν να αλληλεπικαλύπτονται και τελικά κάποια τραγούδια να ανήκουν σε περισσότερα είδη. Ακόμη και οι ίδιες οι εταιρείες μουσικής διαφωνούν σε πολλές περιπτώσεις για το που ακριβώς ταξινομούνται τα μουσικά σήματα. Οι μέθοδοι αυτόματης κατηγοριοποίησης, και κυρίως η μέθοδος του wavelet scattering όπου εστιάσαμε, μπορούν να επεκταθούν περεταίρω, εφόσον βασίζονται σε συγκεκριμένα χαρακτηριστικά για το κάθε ηχητικό σήμα. Ακόμα και με την συνεχή εξέλιξη στην βιομηχανία της μουσικής και την παραγωγή νέων ήχων και μουσικών και κατά συνέπεια μουσικών ειδών, οι αυτόματες μέθοδοι ανάλυσης, με την επιλογή κατάλληλων ταξινομητών της μηχανικής μάθησης μπορούν να παράξουν καλά αποτελέσματα ώστε να μας διευκολύνουν στην κατηγοριοποίηση της μουσικής αλλά και στην εύρεση καινούριων μουσικών κομματιών που να ταιριάζουν στην συγκεκριμένη προτίμησή του καθενός.

Παράρτημα Α

Κώδικας στη MATLAB

```
sf =
waveletScattering('SignalLength',2^19,'SamplingFrequency',22050,...
    'InvarianceScale',0.5);
[fb,f,filterparams] = filterbank(sf);
phi = ifftshift(ifft(fb{1}.phift));
psiL1 = ifftshift(ifft(fb{2}.psift(:,end)));
dt = 1/22050;
time = -2^18*dt:dt:2^18*dt-dt;
scalplt = plot(time,phi,'linewidth',1.5);
hold on
grid on
ylimits = [-3e-4 3e-4];
ylim(ylimits);
plot([-0.25 -0.25],ylimits,'k--');
plot([0.25 0.25],ylimits,'k--');
xlim([-0.6 0.6]);
xlabel('Seconds'); ylabel('Amplitude');
wavplt = plot(time,[real(psiL1) imag(psiL1)]);
legend([scalplt wavplt(1) wavplt(2)],{'Scaling Function','Wavelet-
Real Part','Wavelet-Imaginary Part'});
title({'Scaling Function';'Coarsest-Scale Wavelet First Filter
Bank'})
hold off


ads = audioDatastore('Data/genres','IncludeSubFolders',true,...
     'LabelSource','foldernames');

countEachLabel(ads)
```

```
rng(100);
ads = shuffle(ads);
[adsTrain,adsTest] = splitEachLabel(ads,0.8);

Ttrain = tall(adsTrain);
Ttest = tall(adsTest);

scatteringTrain =
cellfun(@(x)helperscatfeatures(x,sf),Ttrain,'UniformOutput',false);
scatteringTest =
cellfun(@(x)helperscatfeatures(x,sf),Ttest,'UniformOutput',false);

TrainFeatures = gather(scatteringTrain);
TrainFeatures = cell2mat(TrainFeatures);

TestFeatures = gather(scatteringTest);
TestFeatures = cell2mat(TestFeatures);



numTimeWindows = 32;
trainLabels = adsTrain.Labels;
numTrainSignals = numel(trainLabels);
trainLabels = repmat(trainLabels,1,numTimeWindows);
trainLabels =
reshape(trainLabels',numTrainSignals*numTimeWindows,1);

testLabels = adsTest.Labels;
numTestSignals = numel(testLabels);
testLabels = repmat(testLabels,1,numTimeWindows);
testLabels = reshape(testLabels',numTestSignals*numTimeWindows,1);

template = templateSVM(...
    'KernelFunction', 'polynomial', ...
    'PolynomialOrder', 3, ...
    'KernelScale', 'auto', ...
    'BoxConstraint', 1, ...
    'Standardize', true);
Classes = {'blues','classical','country','disco','hiphop','jazz',...
    'metal','pop','reggae','rock'};
classificationSVM = fitcecoc(...
    TrainFeatures, ...
    trainLabels, ...
    'Learners', template, ...
    'Coding', 'onevsone','ClassNames',categorical(Classes));

predLabels = predict(classificationSVM,TestFeatures);
[TestVotes,TestCounts] =
helperMajorityVote(predLabels,adsTest.Labels,categorical(Classes));
```

```
testAccuracy = sum(eq(TestVotes,adsTest.Labels))/numTestSignals*100

confusionchart(TestVotes,adsTest.Labels)

cm = confusionmat(TestVotes,adsTest.Labels);
cm(:,end) = [];
genreAccuracy = diag(cm)./20*100;
figure;
bar(genreAccuracy)
set(gca,'XTickLabels',Classes);
xtickangle(gca,30);
title('Percentage Correct by Genre - Test Set');
```

Supporting Functions

helperMajorityVote

```
function [ClassVotes,ClassCounts] =
helperMajorityVote(predLabels,origLabels,classes)

% This function is in support of wavelet scattering examples
only. It may

% change or be removed in a future release.


% Make categorical arrays if the labels are not already
categorical

predLabels = categorical(predLabels);

origLabels = categorical(origLabels);

% Expects both predLabels and origLabels to be categorical
vectors

Npred = numel(predLabels);

Norig = numel(origLabels);

Nwin = Npred/Norig;

predLabels = reshape(predLabels,Nwin,Norig);

ClassCounts = countcats(predLabels);

[mxcount,idx] = max(ClassCounts);

ClassVotes = classes(idx);
```

```
% Check for any ties in the maximum values and ensure they are
marked as
% error if the mode occurs more than once
modecnt = modecount(ClassCounts,mxcount);
ClassVotes(modecnt>1) = categorical({'NoUniqueMode'});
ClassVotes = ClassVotes(:);

%-----------------------------------------------------------------
----------
    function modecnt = modecount(ClassCounts,mxcount)
        modecnt = Inf(size(ClassCounts,2),1);
        for nc = 1:size(ClassCounts,2)
            modecnt(nc) = histc(ClassCounts(:,nc),mxcount(nc));
        end
    end
end
```

helperscatfeatures

```
function features = helperscatfeatures(x,sf)
% This function is in support of wavelet scattering examples
only. It may
% change or be removed in a future release.

features = featureMatrix(sf,x(1:2^19),'Transform','log');
features = features(:,1:8:end)';
end
```

Βιβλιογραφία

[1] Anden J. , Mallat S., " Deep scattering spectrum", 2014, IEEE Transactions on Signal Processing, Vol. 62, 16, pp. 4114-4128

[2] Mallat. S., "Group invariant scattering", 2012, Communications on Pure and Applied Mathematics, Vol. 65, 10, pp. 1331-1398

[3] Andén J., Mallat S., "Multiscale scattering for audio classification" , Proceedings of the International Society for Music Information Retrieval (Miami, 2011), 657–662.
Available at: http://ismir2011.ismir.net/papers/PS6-1.pdf

[4] Bruna J., Mallat S., "Classification with scattering operators" , 2011, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011), 1561–1566. doi:10.1109/CVPR.2011.5995635

[5] Kanalici E., Bilgin G., "Music Genre Classification via Sequential Wavelet Scattering Feature Learning" , Knowledge Science, Engineering and Management: 12th International Conference, KSEM 2019, Athens, Greece, August 28–30, 2019, Proceedings, Part II Pages 365–372.
Available at: https://doi.org/10.1007/978-3-030-29563-9_32

[6] MathWorks Help Center, "Wavelet Scattering" , R2020a wavelet .
Available at:
https://www.mathworks.com/help/releases/R2020a/wavelet/ug/wavelet-scattering.html

[7] MathWorks Help Center, "Music genre classification using wavelet time scattering" , R2020a wavelet examples. Available at:
https://www.mathworks.com/help/releases/R2020a/wavelet/examples/music-genre-classification-using-wavelet-scattering.html

[8] Tzanetakis G. and Cook P., 2002, "Musical genre classification of audio signals" , IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 5, pp. 293-302.

[9] *GTZAN Genre Collection*. Available at :

http://marsyas.info/downloads/datasets.html

[10] B. Logan, "Mel Frequency Cepstral Coefficients for Music Modeling," in ISMIR, 2000.

[11] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani, Md. Saifur Rahman, "Speaker identification using Mel-frequency cepstral coefficients", 3rd International Conference on Electrical & Computer Engineering ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh

[12] Α. Ζλατίντση, "Επεξεργασία σημάτων μουσικής και εφαρμογές αναγνώρισης" , Διδακτορική διατριβή, Εθνικό Μετσόβιο Πολυτεχνείο, Δεκέμβριος 2013, Αθήνα, Ελλάδα

[13] Changsheng Xu, N.C. Maddage, Xi Shao, Fang Cao, Qi Tian, "Musical genre classification using support vector machines", 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03), Hong Kong

[14] Rohini Gupta, "Implementation of the Moving Average Filter Using Convolution", blog by Pat Reed's group at Cornell University, September 2018. Available at: https://waterprogramming.wordpress.com/2018/09/04/implementation-of-the-moving-average-filter-using-convolution/

[15] Pranav Vijaya Kumar Rao, Vishwas Nagesh Moolimani, "ECG analysis based feature extraction using wavelet transform for music genre classification" , GitHub, February 2020. Available at: https://github.com/vnageshm/Genre_Classification-Wavelets/blob/master/Project_Report.pdf

[16] Robert O. Gjerdingen, David Perrott, "Scanning the Dial: The Rapid Recognition of Music Genres", Journal of New Music Research 2008, Vol. 37, No. 2, pp. 93–100

[17] S. Lippens, J.P Martens, M. Leman, B. Baets, H. Meyer, "A comparison of human and automatic musical genre classification", Conference Paper in Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on · June 2004

[18] Α. Γαραντζιώτης, Ι. Κ. Χατζηλάου, "Από τα FOURIER στα WAVELETS: Μια

Εισαγωγική Παρουσίαση" , ΣΧΟΛΗ ΝΑΥΤΙΚΩΝ ΔΟΚΙΜΩΝ Τομέας Ηλεκτροτεχνίας και Ηλεκτρονικών Υπολογιστών Εργαστήρια Ηλεκτροτεχνίας, Μάιος 2009

[19] Robi Polikar, "The Wavelet Tutorial: The Engineer's Ultimate Guide to Wavelet Analysis" , Rowan University, Ιούνιος 1996.
Available at: https://users.rowan.edu/~polikar/WTtutorial.html

[20] Σ. Μαλτέζος, "Εισαγωγή στη θεωρία των wavelets" , Συμπληρωματικές σημειώσεις του μαθήματος ανάλυσης σήματος του 6ου εξαμήνου της σχολής Ε.Μ.Φ.Ε. του ΕΜΠ, 2020, Αθήνα

[21] Θ. Αλεξόπουλος, "Εισαγωγή στην Ανάλυση Σήματος", Πανεπιστημιακές εκδόσεις ΕΜΠ, 2010, Αθήνα

[22] Amara Graps, "An Introduction to Wavelets", IEEE Computational Science and Engineering, 1995

[23] Gotz Paschmann, Patrick W. Daly, "Analysis Methods for Multi-Spacecraft Data", The International Space Science Institute, Bern, 1998

[24] Epperson Gordon, "music", *Encyclopedia Britannica*, 31 Aug. 2022.
Available at: https://www.britannica.com/art/music

[25] Rachel Becker, Stephanie Przybylek, Sasha Blakeley, "What is music?", Study.com, October 2021. Available at:
https://study.com/academy/lesson/what-is-music-definition-terminology-characteristics.html

[26] Gunther Schuller, "Jazz music", Encyclopedia Britannica, 6 October 2022.
Available at: https://www.britannica.com/art/jazz

[27] Natalie Sarrazin, "Music and the child", Chapter 2, Open SUNY Textbooks, June 15, 2016.
Available at: https://milnepublishing.geneseo.edu/music-and-the-child/chapter/chapter-2/

[28] Julia Harvey-Trappel, "What is Rhythm in music", Jooya Teaching resources, 2021.
Available at: https://juliajooya.com/2020/12/21/what-is-rhythm-in-music/

[29] Cory McKay and Ichiro Fujinaga, "Musical genre classification: Is it worth pursuing and how can it be improved?", Music Technology, Schulich School of Music, McGill University, Montreal, Quebec, Canada, January 2006

[30] Rina Buoy, "Understanding Gaussian Classifier", Published in 'Medium', June 2019. Available at: https://medium.com/swlh/understanding-gaussian-classifier-6c9f3452358f

[31] Onel Harrison, "Machine Learning Basics with the K-Nearest Neighbors Algorithm", Published in 'Towards Data science', September 2018. Available at: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[32] James Monk, "Signal Processing with Wavelets", Niels Bohr Institute, University of Copenhagen, 2016. Available at: https://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2016/WaveletStats.pdf

[33] Α. Προσπαθόπουλος, "Wavelets: Ένα πανίσχυρο μαθηματικό εργαλείο με πολλές εφαρμογές", Ινστιτούτο Ωκεανογραφίας ΕΛΚΕΘΕ, Απρίλιος 2009. Available at: https://onedrive.live.com/Edit.aspx?resid=BE78FE3563A87FA!1065

[34] Lihan Yao, "A ConvNet that works well with 20 samples: Wavelet Scattering", Published in 'Towards Data science', July 2019. Available at: https://towardsdatascience.com/a-convnet-that-works-on-like-20-samples-scatter-wavelets-b2e858f8a385