Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Αυτόματη Καταγραφή Μουσικής

Διπλωματική Εργασία

της

Ελένης Τζιρίτα Ζαχαράτου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2013

Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Αυτόματη Καταγραφή Μουσικής

Διπλωματική Εργασία

της

Ελένης Τζιρίτα Ζαχαράτου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 15η Ιουλίου 2013.

.....	.....	.....
Πέτρος Μαραγκός	Στέφανος Κόλλιας	Κωνσταντίνος Τζαφέστας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2013

.....
Ελένη Τζιρίτα Ζαχαράτου
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Στον πατέρα μου,

Περίληψη

Η Ανάκτηση Μουσικής Πληροφορίας είναι ένας διεπιστημονικός τομέας έρευνας με αντικείμενο την ανάκτηση πληροφορίας από μουσικές ηχογραφήσεις. Η Αυτόματη Καταγραφή Μουσικής εντάσσεται στα πλαίσια της Ανάκτησης Μουσικής Πληροφορίας. Στόχος της είναι να εξάγει από ένα ακουστικό σήμα τα pitches των νοτών και τις χρονικές στιγμές που εμφανίζονται καθώς και να αναγνωρίσει από ποιό όργανο παράχθηκαν. Αφού εξαχθεί αυτή η πληροφορία πρέπει να αναπαρασταθεί σε μία μορφή που να είναι εύκολα αναγνωρίσιμη από τους μουσικούς και η οποία θα μπορούσε να χρησιμοποιηθεί για την αναπαραγωγή της αρχικής ηχογράφησης. Η Αυτόματη Καταγραφή Μουσικής, και κυρίως πολυφωνικής μουσικής, είναι ένα από τα προβλήματα της Ψηφιακής Επεξεργασίας Ηχητικών Σημάτων που παραμένει ανοιχτό. Μέχρι στιγμής καμία εφαρμογή δεν έχει καταφέρει να φτάσει τις ικανότητες ενός εκπαιδευμένου μουσικού.

Σκοπός αυτής της διπλωματικής εργασίας είναι η μελέτη και η ανάπτυξη τεχνικών και αλγορίθμων για την Αυτόματη Καταγραφή Μουσικής παιγμένης στο πιάνο. Η εκτενής ανασκόπηση της σχετικής με την Αυτόματη Καταγραφή Μουσικής βιβλιογραφίας συνοδεύεται από την παρουσίαση μίας πρωτότυπης μεθόδου ανίχνευσης της αρχής (onset) της νότας για μονοφωνική μουσική παιγμένη στο πιάνο και μίας πρωτότυπης μεθόδου εκτίμησης πολλαπλών τόνων για συγχορδίες παιγμένες στο πιάνο. Η ανίχνευση των onsets βασίζεται στον ενεργειακό τελεστή Teager-Kaiser και σε μία συστοιχία φίλτρων Gabor, που έχουν τις συχνότητες των κλειδιών του πιάνου σαν κεντρικές συχνότητες. Αντίστοιχα, η μέθοδος εκτίμησης πολλαπλών τόνων υπολογίζει τον DTFT του μουσικού σήματος στις συγκεκριμένες συχνότητες που αντιστοιχούν στις νότες του πιάνου. Ο βαθμός πολυφωνίας K είναι άγνωστος και επομένως ο προτεινόμενος αλγόριθμος στοχεύει αφενός στο να συνάγει το K και αφετέρου στο να βρει τις θεμελιώδεις συχνότητες. Αυτό γίνεται σειριακά, ξεκινώντας από μία μονοφωνική υπόθεση και συνεχίζοντας με εύλογους συνδυασμούς μεγαλύτερου βαθμού πολυφωνίας. Ως υποψήφιοι συνδυασμοί νοτών επιλέγονται αυτοί που ερμηνεύουν καλύτερα το παρατηρούμενο φάσμα καθώς και κάποιες ιδιότητες που προκύπτουν από τα δεδομένα με μία διαδικασία εκπαίδευσης. Ο τελικός συνδυασμός που επιλέγεται είναι αυτός με το ελάχιστο τετραγωνικό σφάλμα. Τα πειραματικά μας αποτελέσματα δείχνουν καλή απόδοση και για τις δύο μεθόδους.

Λέξεις-Κλειδιά Ανάκτηση Μουσικής Πληροφορίας, Αυτόματη Καταγραφή Μουσικής, Ανίχνευση του onset μίας νότας, Εκτίμηση Πολλαπλών Τόνων

Abstract

Music Information Retrieval is the interdisciplinary research area of retrieving information from music. Automatic Music Transcription is a field of Music Information Retrieval. It aims at extracting the pitches of notes and their timings from an audio signal as well as identifying which instrument generated them. After extracting that information it has to be represented in a form which is understandable to musicians and could be used to recreate the original audio. Automatic Music Transcription, especially for polyphonic music, is a problem of Audio Digital Signal Processing which remains open. Until today, an application that can reach the capabilities of a trained musician doesn't exist.

The goal of this diploma thesis is the study and the development of techniques and algorithms for the Automatic Piano Music Transcription. Apart from an extensive review of the relevant with Automatic Music Transcription literature, we present an onset detection method for monophonic piano music and a multiple fundamental frequency (F0) estimation method for polyphonic piano music. Onset detection is based on the Teager-Kaiser energy operator and a Gabor filter bank, having the frequencies of the piano keys as the central frequencies. Similarly, the multiple fundamental frequency estimation method evaluates the DTFT of the signal at the specific frequencies of the piano notes. The polyphony order K is unknown and thus the proposed algorithm aims at both inferring K and finding the fundamental frequencies. This is done sequentially, starting from a monophonic assumption and continuing with plausible combinations of higher polyphony order. Candidate combinations of notes are selected as best matching the spectrum and some properties learned from the data. The final combination which is selected is that of minimum squared error. Our experimental results show good performance for both methods.

Keywords Music Information Retrieval, Automatic Music Transcription, Onset Detection, Multiple Fundamental Frequency Estimation

Ευχαριστίες

Με την ευκαιρία της περάτωσης αυτής της διπλωματικής εργασίας και κατά συνέπεια των προπτυχιακών μου σπουδών στο Εθνικό Μετσόβιο Πολυτεχνείο, θα ήθελα να ευχαριστήσω όλους τους καθηγητές μου και τους ανθρώπους του ιδρύματος για τη συνεισφορά τους στη μακρά αυτή πορεία. Ειδικότερα θα ήθελα να ευχαριστήσω τον καθηγητή κ. Πέτρο Μαραγκό για τη δυνατότητα που μου έδωσε να δουλέψω κάτω από την επίβλεψη και την πολύτιμη καθοδήγησή του.

Εκτός από τους καθηγητές μου, θα ήθελα να ευχαριστήσω τον πατέρα μου που υπήρξε ανέκαθεν ο μέντοράς μου στην προσπάθειά μου να επιτύχω τους στόχους μου.

Κατά την εκπόνηση της διπλωματικής μου εργασίας, είχα την ευκαιρία να συζητήσω με τους κυρίους Γκιόκα και Κατσούρο, ερευνητές του ΙΕΛ, που μου παραχώρησαν και υλοποίησή τους στο MatLab. Ως εκ τούτου, θα ήθελα να τους ευχαριστήσω θερμά.

Τέλος θα ήθελα να εκφράσω την ευγνωμοσύνη μου σε ολόκληρη την οικογένειά μου, για την αγάπη και την υποστήριξή τους καθώς και στους κοντινούς φίλους μου και ιδιαίτερα στο φίλο μου Νίκο που ήταν δίπλα μου όλα τα χρόνια των σπουδών μου και μοιράστηκαν μαζί μου τόσο τις δύσκολες όσο και τις ευχάριστες στιγμές.

Περιεχόμενα

Κατάλογος σχημάτων

Κεφάλαιο 1

Εισαγωγή

Για τους ανθρώπους, ο ήχος δεν είναι απλά μια φυσική δόνηση. Μεταφέρει νοήματα και συναισθήματα. Μέσω της φωνής οι άνθρωποι μπορούν να επικοινωνούν μιλώντας μια γλώσσα, μέσω της μουσικής εκφράζονται καλλιτεχνικά ενώ οι περιβαλλοντικοί ήχοι συμβάλουν στο σχηματισμό της εικόνας που έχουμε για το περιβάλλον. Οι ανθρώπινες φυσιολογικές λειτουργίες είναι ιδιαίτερα αναπτυγμένες και η κατανόηση των ήχων και της σημασίας τους γίνεται αυτόματα και σε μεγάλο βαθμό ασυναίσθητα. Στην εποχή μας η πληροφορική έχει την δυνατότητα να μιμηθεί αυτή την ανθρώπινη λειτουργία κατανόησης των ήχων, με επιτυχία που φυσικά δε μπορεί να συγκριθεί με αυτή ενός ανθρώπου-ακροατή. Στην παρούσα διπλωματική εργασία, μας ενδιαφέρει η περίπτωση της μουσικής και κυρίως της μουσικής για πιάνο, για την οποία προσπαθούμε να εξάγουμε από μια ηχογράφηση τις νότες που έχουν παιχτεί και είναι παρούσες στον ήχο χρησιμοποιώντας εργαλεία από την επιστήμη της πληροφορικής και της επεξεργασίας σημάτων.

1.1 Διάρθρωση διπλωματικής εργασίας

Το κείμενο οργανώνεται σε έξι κεφάλαια. Στο *1ο Κεφάλαιο* γίνεται η περιγραφή του προβλήματος και παρουσιάζεται η περιοχή της ψηφιακής μουσικής επεξεργασίας και της αυτόματης μουσικής καταγραφής. Αναφέρονται τα συγκεκριμένα προβλήματα στα οποία εστιάζει η διπλωματική εντός αυτής της περιοχής καθώς και οι συνεισφορές της.

Το *2ο Κεφάλαιο* εισάγει τον αναγνώστη σε κάποιες βασικές έννοιες, τόσο περί μουσικής όσο και περί σημάτων και παραθέτει κάποιους βασικούς ορισμούς. Το υπόβαθρο που καλύπτεται σε αυτό το κεφάλαιο δεν είναι εκτενές, οπότε πιθανώς ο αναγνώστης να χρειαστεί να ανατρέξει και σε άλλες πηγές. Τέλος σε αυτό το κεφάλαιο, αναφέρονται κάποιοι σημαντικοί σταθμοί και τάσεις στην ιστορία της αυτόματης καταγραφής μουσικής.

Το *3ο Κεφάλαιο* καταπιάνεται με τις διάφορες πτυχές που αφορούν ένα σύστημα αυτόματης καταγραφής μονοφωνικής μουσικής. Παρουσιάζεται το πρόβλημα της εύρεσης της αρχής της νότας (onset detection) και το πρόβλημα εκτίμησης της θεμελιώδους συχνότητας και γίνεται μία ανασκόπηση κάποιων βασικών προσεγγίσεων που έχουν ακολουθηθεί για την επίλυση αυτών των προβλημάτων. Υλοποιείται ένα πλήρες σύστημα αυτόματης καταγραφής μονοφωνικής μουσικής με χρήση αλγορίθμων από τη βιβλιογραφία για την εύρεση των onsets και την εκτίμηση της θεμελιώδους συχνότητας, ενώ επιπλέον προτείνεται ένας καινούριος αλγόριθμος για την εύρεση των onsets ο οποίος κάνει χρήση του Teager τελεστή ενέργειας.

Το *4ο Κεφάλαιο* ασχολείται με μεθόδους εύρεσης ρυθμικής πληροφορίας και συγκεκριμένα με τα προβλήματα της εκτίμησης του tempo, του beat tracking και της αυτόματης εύρεσης του μέτρου. Εξετάστηκαν και υλοποιήθηκαν κάποιοι αλγόριθμοι από τη βιβλιογραφία για κάθε ένα από αυτά τα προβλήματα. Στο σημείο αυτό πρέπει να εκφραστούν ευχαριστίες στον κύριο Γκιόκα για την ευγενική παραχώρηση του πηγαίου κώδικα της μεθόδου που οι κύριοι Γκιόκας, Κατσούρος, Καραγιάννης και Σταφυλάκης προτείνουν στο [6] για την εκτίμηση του tempo.

Το *5ο Κεφάλαιο* το οποίο και αποτελεί την κύρια συνεισφορά της παρούσας εργασίας, αφορά μεθόδους εκτίμησης πολλαπλών τόνων. Γίνεται παρουσίαση του προβλήματος και βιβλιογραφική ανασκόπηση ενώ προτείνεται και μία πρωτότυπη μέθοδος για την εκτίμηση πολλαπλών τόνων.

Τέλος στο *6ο Κεφάλαιο* συνοψίζονται οι συνεισφορές της παρούσας εργασίας, ενώ προτείνονται και κάποιες πιθανές κατευθύνσεις για μελλοντική έρευνα.

1.2 Ορισμός του προβλήματος και εφαρμογές

Ονομάζουμε καταγραφή μουσικής μια συμβολική καταγραφή της εκτέλεσης ενός μουσικού κομματιού, το πέρασμα από μια μουσική ερμηνεία σε μια συμβολική περιγραφή. Με αυτή την έννοια, καταγραφή είναι η ανάλυση του ήχου που ακούμε με σκοπό την εξαγωγή πληροφοριών, του περιεχομένου δηλαδή εκείνου που έχει νόημα. Η αυτόματη καταγραφή της μουσικής είναι συγγενικός τομέας με αυτόν της αυτόματης αναγνώρισης φωνής, αν και η αναγνώριση φωνής έχει σίγουρα προσελκύσει τόσο ακαδημαϊκά όσο και εμπορικά, πολύ περισσότερο ενδιαφέρον κι έτσι έχει μελετηθεί περισσότερο. Τόσο η αυτόματη αναγνώριση φωνής όσο και η αυτόματη καταγραφή μουσικής, εφαρμόζονται σε δεδομένα του πραγματικού κόσμου, δεδομένα που παράγονται φυσικά. Έτσι και στις δύο περιπτώσεις, τα ξεχωριστά γεγονότα που μπορούν να ορίσουν δομές είναι περιορισμένα σε αριθμό. Στο λόγο έχουμε τα φωνήματα που χρησιμοποιούνται για τη δημιουργία λέξεων και προτάσεων και αντίστοιχα στη μουσική μεμονωμένοι ήχοι συνδυάζονται για τη δημιουργία μελωδιών, ρυθμών και τραγουδιών. Στον τομέα της αυτόματης αναγνώρισης φωνής, η καταγραφή μιας συζήτησης ή μιας ομιλίας έχει ως στόχο την εξαγωγή των λέξεων και των φράσεων. Η καταγραφή της μουσικής έχει πάνω από όλα την επιδίωξη να εκτιμήσει τις νότες που παίζονται και τις παραμέτρους τους: το ύψος τους, τη χρονική στιγμή της έναρξής τους, τη διάρκειά τους και τελικά πληροφορίες υψηλότερου επιπέδου τέτοιες όπως τα ρυθμικά σχήματα, το μέτρο ή ο οπλισμός. Μια βασική διαφορά μεταξύ λόγου και μουσικής είναι ότι ο λόγος είναι ουσιαστικά μονοφωνικός (ασχολούμαστε με έναν ομιλητή), ενώ η μουσική είναι συχνά πολυφωνική. Από την άλλη, ο λόγος διαφοροποιείται ταχύτερα και ταυτόχρονα οι ιδιότητες του ήχου που μεταφέρει πληροφορία φωνής είναι εγγενώς πολυδιάστατες, ενώ το pitch και η διάρκεια ενός (μονοφωνικού) μουσικού σήματος είναι μονοδιάστατες ποσότητες ([1]).

Η καταγραφή μιας ηχογράφησης αποτελεί και για τον άνθρωπο μία δύσκολη εργασία που γενικά απαιτεί κάποια σχετική εκπαίδευση. Όσο πιο πλούσια είναι η πολυφωνική πολυπλοκότητα της μουσικής σύνθεσης, τόσο μεγαλύτερη εμπειρία απαιτείται να έχει κάποιος στο μουσικό ύφος, στα υπό εξέταση όργανα και στη μουσική θεωρία. Η καταγραφή γίνεται αυτόματη όταν δεν πραγματοποιείται πλέον από έναν άνθρωπο αλλά από ειδικά σχεδιασμένο λογισμικό. Σε αυτό το πλαίσιο, το κομμάτι προς καταγραφή αντιπροσωπεύεται από ένα αρχείο ήχου -τύπου .wav ή .mp3 για παράδειγμα- και η καταγραφή που παράγεται παίρνει τη μορφή ενός αρχείου MIDI ή ισοδύναμου, που είναι κατάλληλο για την αναπαράσταση και την αποθήκευση της πληροφορίας που

εξάχθηκε. Παρόλα αυτά, οι έμπειροι μουσικοί είναι ικανοί για την επίλυση πλούσιων πολυφωνιών με μεγάλη ευελιξία όσον αφορά την ποικιλία των οργάνων και των μουσικών υφών καθιστώντας φανερό και αδιαμφισβήτητο το γεγονός ότι τα αυτόματα συστήματα καταγραφής υστερούν στην επίδοση συγκριτικά με τους ανθρώπους που έχουν μουσικές γνώσεις και κατάρτιση. Το κύριο πλεονέκτημα που έχουμε ως άνθρωποι είναι η μοναδική μας ικανότητα στην ταυτοποίηση προτύπων και η μνήμη μας, που μας επιτρέπουν να προβλέψουμε μελλοντικά γεγονότα. Η χρήση μνήμης στην αυτόματη καταγραφή συνήθως συνεπάγεται ένα τεράστιο υπολογιστικό κόστος. Δεν είναι τόσο δύσκολο να συμπεριληφθεί βραχεία μνήμη, αλλά η διατήρηση μιας μνήμης μακράς διαρκείας που σημαίνει να παραμένουν ενεργές πολλές υποθέσεις για διάφορα πλαίσια, έχει μεγάλο κόστος. Η επίλυση ορισμένων ασαφειών που οι άνθρωποι επιλύουν χρησιμοποιώντας τη μακράς διάρκειας μνήμη τους παραμένει μία πρόκληση.

Η αυτόματη καταγραφή μουσικής εντάσσεται στο πιο γενικό πλαίσιο της Ανάκτησης Μουσικής Πληροφορίας, ενός ολόκληρου τομέα έρευνας που σε αγγλική ορολογία λέγεται Music Information Retrieval (MIR). Αντικείμενο του MIR αποτελούν μια σειρά από θέματα που βασίζονται στην ανάκτηση πληροφοριών μετά από επεξεργασία του ακουστικού σήματος, σε αντίθεση με την αναζήτηση που βασίζεται σε βιβλιογραφικές πληροφορίες (όπως τίτλοι και ονόματα καλλιτεχνών από μία CDDB, μια online βάση δεδομένων με πληροφορίες CD). Αναφορικά, θέματα του MIR αποτελούν η αναζήτηση βάσει μιας μελωδίας (Research on Melody: Query by Humming -QBH), η αναζήτηση βάσει αποσπάσματος (Research on Music Fragments) και η αναζήτηση βάσει ομοιότητας μεταξύ μουσικών κομματιών (Research on Entire Musical Pieces). Άλλα θέματα ερευνών είναι η εκτίμηση του ρυθμού (Rhythm Tracking/Tempo και Beat Tracking), η αναγνώριση της μελωδίας (Melody Recognition) και η ανάλυση της μουσικής δομής. Το ερευνητικό πεδίο της Ανάκτησης Μουσικής Πληροφορίας είναι πλέον πολύ σημαντικό, απασχολεί πολλούς ερευνητές παγκοσμίως, περιλαμβάνει μία ευρεία γκάμα προβλημάτων και, από το 2000, έχει αποκτήσει το δικό του ετήσιο συνέδριο, το International Conference on Music Information Retrieval (ISMIR). Επίκεντρο της αναζήτησης μουσικής πληροφορίας είναι ο ήχος που παράγεται από την εκτέλεση ενός κομματιού. Αυτός ο ήχος, που εκπέμπεται σε μια δεδομένη στιγμή, είτε ηχογραφείται είτε αναλύεται άμεσα. Προέρχεται από μία οι περισσότερες ηχητικές πηγές και διαδίδεται σε ένα φυσικό μέσο, τον αέρα πιο συχνά, και σε ένα περιβάλλον, αίθουσα συναυλιών ή άλλο. Η ηχητική πηγή, ένα όργανο μουσικής για παράδειγμα, ελέγχεται -παίζεται- από έναν άνθρωπο, ο οποίος ερμηνεύει ένα έργο, το οποίο έχει συντεθεί εκ των προτέρων ή είναι αυτοσχεδιασμός. Η παραπάνω περιγραφή υποδεικνύει μια προσέγγιση υπό τη μορφή μιας αλυσίδας παραγωγής στην οποία κάθε στοιχείο έχει τη δική του συνεισφορά στο ηχητικό αποτέλεσμα και περιέχει συγκεκριμένες πληροφορίες που θα μπορούσαν να είναι αντικείμενο έρευνας στο MIR. Από την αναγνώριση των οργάνων μέχρι τον εντοπισμό των πηγών και την ταξινόμηση ανά είδος μουσικής για την παραγωγή λιστών αναπαραγωγής (γνωστών ως playlists), κάθε εφαρμογή ανάγεται στην απομόνωση των πληροφοριών που την αφορούν, και βρίσκονται θαμμένες μέσα στον ηχογραφημένο ήχο που είναι το κοινό σημείο όλων. Η αυτόματη καταγραφή μουσικής είναι ένα πρόβλημα του MIR, που εμπλέκει πολλά επιστημονικά πεδία, όπως επεξεργασία ηχητικών σημάτων, εκμάθηση μηχανής, επιστήμη υπολογιστών, ψυχοακουστική και μουσική αντίληψη (music perception), μουσική θεωρία και μουσική νόηση (music cognition). Οι πληροφορίες που αναζητούνται στο πλαίσιο της αυτόματης καταγραφής διαμορφώνουν ένα πεδίο που τοποθετείται στο επίπεδο των ηχητικών πηγών και που αποσκοπεί στην κατανόηση των νοτών που παίζονται με τους ρυθμούς τους, τα ύψη τους, τις αποχρώσεις τους, την άρθρωσή τους, την εύρεση του οργάνου από το οποίο προέρχονται, την αναγνώριση των συγχορδιών ή ακόμα την παρακολούθηση του tempo.

Η αυτόματη αναγνώριση και καταγραφή μουσικής έχει πολλές εφαρμογές. Η προφανής εφαρμογή της αυτόματης καταγραφής μουσικής δεν είναι άλλη από την βοήθεια που προσφέρει σε ένα μουσικό στην καταγραφή της μουσικής σημειογραφίας ενός κομματιού από μια ηχογράφηση (που είναι μια ιδιαίτερα χρονοβόρα εργασία όταν γίνεται με το χέρι) επιτρέποντας τη δημιουργία προγραμμάτων μετατροπής audio/παρτιτούρα ή audio/MIDI. Όμως η χρησιμότητά της αυτόματης καταγραφής ξεπερνάει κατά πολύ τον κύκλο των μουσικών. Η εξέλιξή της είναι στην πραγματικότητα στην καρδιά της επεξεργασίας ηχητικών σημάτων και των εργασιών στο MIR. Ο ολοένα και αυξανόμενος όγκος αρχείων ήχου έχει δημιουργήσει την ανάγκη μιας αρχειοθέτησης περισσότερο αυτόματης παρά χειρωνακτικής κι έτσι η αυτόματη καταγραφή αποτελεί μία βασική λειτουργία για τις διάφορες εφαρμογές του MIR παρέχοντας ουσιαστική βοήθεια στη διατήρηση ηχητικών αρχείων και στη στατιστική ανάλυσή τους. Αυτό συμβαίνει κυρίως σε εφαρμογές όπως η εξαγωγή της μελωδίας ή του τονικού περιεχομένου, η ταυτοποίηση των οργάνων, του μουσικού κομματιού, του ύφους, του καλλιτέχνη ή του συνθέτη, η ανίχνευση της δομής των κομματιών (εισαγωγή, ρεφραίν, κουπλέ) και η παραγωγή μουσικών περιλήψεων, η παρακολούθηση μιας παρτιτούρας σε πραγματικό χρόνο και η αντιστοίχισή της με μια ηχητική ροή, η ταξινόμηση ανά είδος. Γενικά, οι μέθοδοι καταγραφής μουσικής μπορούν επίσης να παρέχουν πληροφορίες σχετικές με τις νότες σε αλγορίθμους που δουλεύουν με συμβολικά μουσικά δεδομένα. Ενώ πολλές από αυτές τις εφαρμογές είναι ήδη μέρος της καθημερινότητάς μας, οι περισσότερες από αυτές αποτελούν τομείς έρευνας πολύ ενεργούς με επιδόσεις που επιδέχονται βελτίωση σε μεγάλο βαθμό και μεγάλες προοπτικές καινοτομίας, με πολυπόθητο στόχο να είναι η αναζήτηση και η πλοήγηση με βάση το περιεχόμενο σε μεγάλες βάσεις δεδομένων ήχων. Έτσι, σύντομα στο μέλλον θα μπορούσαμε να έχουμε έναν ραδιοφωνικό δέκτη ικανό να εντοπίσει ταυτόχρονα όλους τους σταθμούς που παίζουν τζαζ μουσική, κάτι που σίγουρα κάποιοι άνθρωποι θα εκτιμούσαν ιδιαιτέρως.

Επιγραμματικά λοιπόν, μεταξύ των πολυάριθμων πιθανών εφαρμογών ενός αυτομάτου συστήματος καταγραφής, κάποιες βασικές εφαρμογές είναι οι ακόλουθες:

- *Εργαλείο για μουσικούς και συνθέτες* το οποίο μπορούν να χρησιμοποιήσουν για να αναλύσουν αποδοτικά συνθέσεις τις οποίες διαθέτουν μόνο στη μορφή μιας ηχογράφησης.

- *Δομημένη κωδικοποίηση audio.* Μία συμβολική αναπαράσταση ενός ακουστικού σήματος επιτρέπει την επιλεκτική κωδικοποίηση του σήματος. Μία αναπαράσταση τύπου MIDI είναι εξαιρετικά συμπαγής και όμως διατηρεί την αναγνωρισιμότητα και τα χαρακτηριστικά ενός μουσικού κομματιού σε σημαντικό βαθμό. Σε μια δομημένη κωδικοποίηση audio, οι παράμετροι του ήχου πρέπει να κωδικοποιηθούν επίσης, αλλά το εύρος ζώνης παραμένει γύρω στα 2-3 kbit/s.

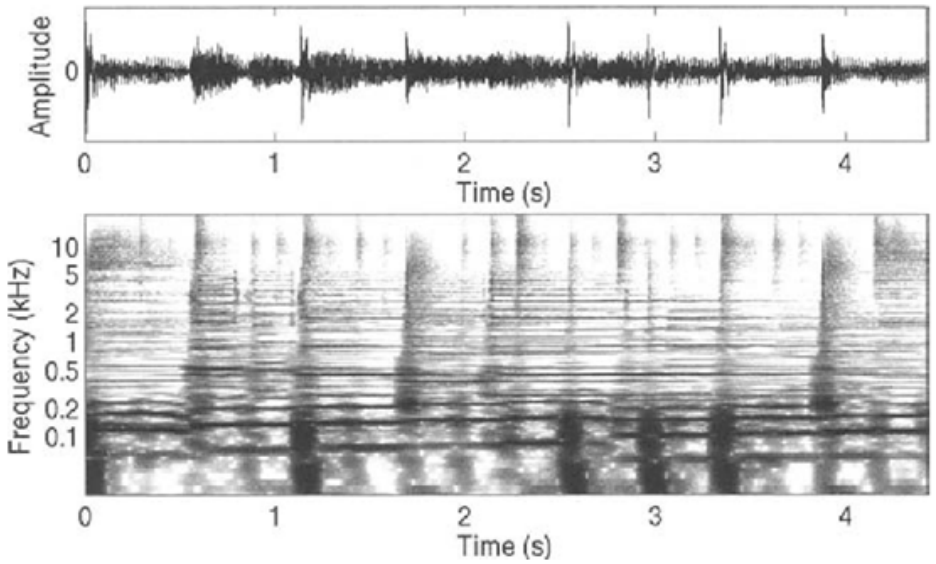
- *Αναζήτηση μουσικής πληροφορίας (Music Information Retrieval)* με βάση για παράδειγμα τη μελωδία ενός κομματιού.

- *Μουσικολογική ανάλυση.* Τα εργαλεία καταγραφής διευκολύνουν την ανάλυση αυτοσχεδιαστικής (π.χ. σόλο αυτοσχεδιασμοί στην τζαζ) και λαϊκής μουσικής που δεν υπάρχει πουθενά καταγεγραμμένη και τη διαχείριση εθνομουσικολογικών αρχείων. Η αυτόματη καταγραφή θα διευκόλυνε και μία μουσική ψυχολογική ανάλυση της εκτελούμενης μουσικής.

- *Ευέλικτη μουσική μίξη και επεξεργασία* αλλάζοντας την ενορχήστρωση, εφαρμόζοντας εφέ σε συγκεκριμένα μέρη, αλλάζοντας την ένταση διαφορετικών μερών ή εξάγοντας επιλεκτικά συγκεκριμένα όργανα και γενικότερα διασκευάζοντας με κάποιο τρόπο το κομμάτι πριν την ανασύνθεση ενός νέου κομματιού από την παρτιτούρα.

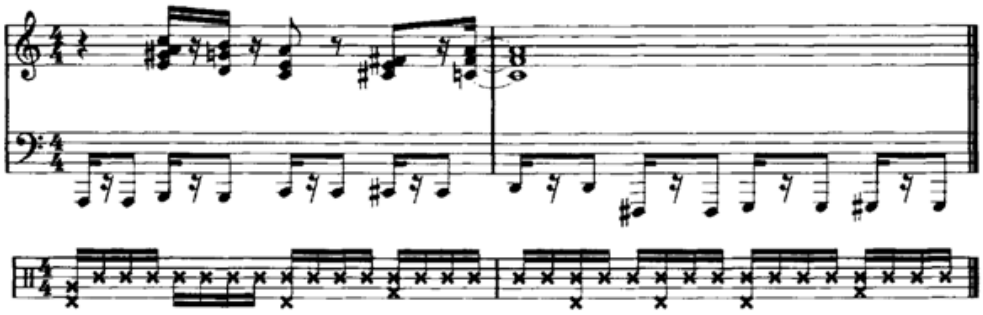
- *Διαδραστικά μουσικά συστήματα / αλληλεπίδραση ανθρώπου-μηχανής* που παρέχουν συνοδεία στο τραγούδι ή στην εκτέλεση ενός σολίστα, σε πραγματικό ή μη χρόνο, προγράμματα για τη δημιουργία παρτιτούρας ή ηλεκτρονικά παιχνίδια με μουσικό προσανατολισμό. Η καταγραφή του τραγουδιού έχει ιδιαίτερη σημασία εδώ.
- *Εξοπλισμός που σχετίζεται με τη μουσική*, όπως ο συγχρονισμός εφέ από φώτα με ένα μουσικό σήμα.
- *Εντοπισμός λογοκλοπής.*

Όπως είναι αρκετά προφανές, οι εφαρμογές ενός συστήματος αυτόματης καταγραφής μουσικής θα μπορούσαν να συγκριθούν με αυτές ενός συστήματος αυτόματης αναγνώρισης φωνής. Και τα δύο προβλήματα είναι ιδιαίτερα περίπλοκα και όπως έχει ήδη αναφερθεί έχουν πολλά κοινά σημεία, αλλά υπάρχει μεγάλο εμπορικό ενδιαφέρον για την κωδικοποίηση και την αναγνώριση της φωνής, σε αντίθεση με τη σχετικά μικρή έρευνα στην αναγνώριση μουσικής.



Σχήμα 1.1: Ένα ακουστικό μουσικό σήμα (πάνω) και η αναπαράστασή του στο χώρο χρόνου-συχνότητας (κάτω). Ανατύπωση από [1].

Αναφορικά με τους στόχους της καταγραφής, η φύση των πληροφοριών που αναζητούνται αλλάζει αντίστοιχα με την αναπαράστασή τους. Ο σκοπός της αυτόματης καταγραφής είναι να εξάγει από το ηχητικό σήμα μια αναπαράσταση που μπορεί να διαβαστεί και να ερμηνευτεί από έναν άνθρωπο. Τα συστήματα μουσικής σημειογραφίας επιτρέπουν τη δημιουργία μιας παρτιτούρας. Μια παρτιτούρα είναι ένας οδηγός για την εκτέλεση ενός κομματιού και μπορεί να αναπαρασταθεί με διαφορετικούς τρόπους. Τα συστήματα μουσικής σημειογραφίας έχουν εξελιχθεί μέσα στην ιστορία, από τα πρώτα ίχνη μελωδικής και ρυθμικής σημειογραφίας στην αρχαιότητα και τα πνεύματα στο Μεσαίωνα, μέχρι την κλασική σημειογραφία, που είναι σχετικά πρόσφατη και αποτελεί πάντα αντικείμενο καινοτομιών. Η σύγχρονη σημειογραφία που χρησιμοποιείται στη δυτική τονική μουσική είναι η πιο αναλυτική αναπαράσταση. Προκειμένου να εξαχθεί μια αναγνώσιμη παρτιτούρα από ένα σήμα, είναι απαραίτητο να εκτιμηθούν τα pitches, οι χρονικές στιγμές της αρχής κάθε νότας, οι διάρκειες των νοτών, το tempo, το μέτρο και η τονικότητα του μουσικού κομματιού. Τα σχήματα 1.1 και 1.2 προέρχονται από το [1] και δείχνουν την αναπαράσταση ενός παραδείγματος



Σχήμα 1.2: Μουσική σημειογραφία που αντιστοιχεί στο σήμα του σχήματος 1.1. Οι πάνω γραμμές πενταγράμμου δείχνουν τη σημειογραφία για τα μουσικά όργανα με pitch και οι κάτω γραμμές πενταγράμμου δείχνουν τη σημειογραφία για κρουστά όργανα. Ανατύπωση από [1].

μουσικού σήματος. Ξεχνώντας τις λεπτομέρειες, οι κύριες παραδοχές είναι ότι ο χρόνος κυλάει από αριστερά προς τα δεξιά και το pitch των νοτών υποδεικνύεται από την κατακόρυφη θέση τους στις γραμμές του πενταγράμμου. Στην περίπτωση των drums και των κρουστών, η κατακόρυφη θέση υποδεικνύει το όργανο και τον τρόπο κρούσης. Η ένταση (και το εφαρμοζόμενο όργανο στην περίπτωση οργάνων που έχουν pitch) δεν διευκρινίζεται συνήθως για μεμονωμένες νότες αλλά καθορίζεται για μεγαλύτερα μέρη.

Σχήμα 1.3: 'Piano-roll' αναπαράσταση ενός MIDI αρχείου που αντιστοιχεί στα όργανα με Pitch του σήματος του σχήματος 1.1. Οι διαφορετικές νότες είναι οργανωμένες στον κατακόρυφο άξονα και ο χρόνος κυλάει από αριστερά προς τα δεξιά. Ανατύπωση από [1].

Εκτός από τη συνηθισμένη μουσική σημειογραφία, η καταγραφή μπορεί να πάρει επίσης και άλλες μορφές. Για παράδειγμα, ένας κιθαρίστας μπορεί να θεωρήσει πιο βολικό να διαβάζει σύμβολα συγχορδιών που χαρακτηρίζουν το συνδυασμό των νοτών που πρέπει να παιχτούν με ένα πιο γενικό τρόπο. Επιπλέον η πρόοδος της μουσικής πληροφορικής εισήγαγε νέες μορφές για την έκδοση μιας παρτιτούρας και έτσι δημιουργήθηκαν μορφές όπως MIDI και SMF (Standard MIDI File) που έχουν το πλεονέκτημα της επικοινωνίας με άλλα ηλεκτρονικά όργανα. Κοινό σε όλες αυτές τις αναπαραστάσεις είναι ότι αποτυπώνουν μουσικά σημαντικές παραμέτρους που μπορούν να χρησιμοποιηθούν για την εκτέλεση ή την σύνθεση του υπό εξέταση μουσικού κομματιού. Από αυτήν την σκοπιά, η καταγραφή μουσικής μπορεί να θεωρηθεί σαν την ανακάλυψη της 'συνταγής', ή την αντίστροφη παραγωγή του 'πηγαίου κώδικα' ενός μουσικού σήματος. Το piano-roll είναι

μια φυσική αναπαράσταση ενός SMF (standart midi file) με το χρόνο στον οριζόντιο άξονα, και το pitch στον κατακόρυφο (Σχήμα 1.3). Οι νότες αναπαρίστανται χρησιμοποιώντας οριζόντιες ράβδους στο πλέγμα χρόνου-συχνότητας. Το piano-roll μπορεί να θεωρηθεί ότι είναι μία αναπαράσταση προσανατολισμένη στον ήχο που δείχνει όλες τις νότες που παίζονται σε κάθε στιγμή και η μετατροπή ενός μουσικού ηχητικού σήματος σε piano-roll αναπαράσταση -χωρίς ρυθμική πληροφορία- εξαρτάται άμεσα από την κυματομορφή. Η παρτιτούρα δεν αντιστοιχεί στην ηχογράφηση της ελεύθερης ερμηνείας ενός εκτελεστή με την ίδια ακρίβεια, αφού σε αυτήν οι χρονικές διάρκειες θα πρέπει να κβαντιστούν. Παρόλο που το MIDI format έχει κάποιους περιορισμούς, τα SMF's συνήθως θεωρούνται μια καλή αναπαράσταση στα πλαίσια της υπολογιστικής καταγραφής μουσικής. Στη συνέχεια η μουσική παρτιτούρα μπορεί να εξαχθεί από το αρχείο MIDI.

Μια πλήρης καταγραφή θα απαιτούσε την εύρεση του pitch, της χρονικής στιγμής και του οργάνου όλων των συμβάντων ήχου. Καθώς αυτό μπορεί να είναι πολύ δύσκολο, ή ακόμα και θεωρητικά αδύνατο σε κάποιες περιπτώσεις, ο στόχος συνήθως επαναπροσδιορίζεται ως εξής: καταγραφή όσων περισσότερων ήχων είναι δυνατό (πλήρης καταγραφή) ή καταγραφή μόνο ενός καλώς ορισμένου μέρους του μουσικού σήματος, για παράδειγμα της κυρίαρχης μελωδίας ή των πιο σημαντικών ήχων drums (μερική καταγραφή). Φυσικά εν γένει, ένα σύστημα μουσικής καταγραφής δεν μπορεί να δώσει την ακριβή παρτιτούρα που διάβασε αρχικά ο μουσικός. Τα μουσικά ηχητικά σήματα είναι συχνά εκφραστικές ερμηνείες και όχι απλές μηχανικές μεταφράσεις νοτών που διαβάζονται από ένα χαρτί. Μία συγκεκριμένη παρτιτούρα αποτελεί απλά έναν οδηγό για τον ερμηνευτή και μπορεί να εκτελεστεί από αυτόν με πολλούς διαφορετικούς τρόπους. Έτσι, το πρόβλημα της αυτόματης καταγραφής μουσικής δεν είναι καλώς ορισμένο και δεν έχει μία μοναδική λύση.

1.3 Συνοπτική περιγραφή των προβλημάτων που εξετάστηκαν στην παρούσα εργασία

Αυτή η διπλωματική εστιάζει στην αυτόματη καταγραφή μουσικής για πιάνο -ένα από τα πιο εξεζητημένα προβλήματα της αναγνώρισης μουσικής- και τα προβλήματα που σχετίζονται με αυτή. Συγκεκριμένα για την δημιουργία μιας παρτιτούρας, απαιτείται η εύρεση του pitch, της αρχής κάθε νότας (onset) και της διάρκειας των νοτών, το tempo, το μέτρο και η τονικότητα ενός μουσικού κομματιού. Η απόφαση να περιοριστεί η μελέτη μόνο στο όργανο του πιάνου παρακινήθηκε τόσο από το μεγάλο πλήθος σόλο ηχογραφήσεων για πιάνο όσο και τις επιστημονικές προκλήσεις που σχετίζονται με αυτό το όργανο. Κάποιες εργασίες υποδεικνύουν ότι η αυτόματη καταγραφή για πιάνο παραμένει μία από τις πιο δύσκολες περιπτώσεις συγκριτικά με την περίπτωση άλλων μουσικών οργάνων. Λόγοι που οδηγούν σε αυτή την αυξημένη δυσκολία είναι για παράδειγμα το μεγάλο εύρος θεμελιωδών συχνοτήτων του οργάνου, τα γρήγορα περάσματα νοτών που συναντώνται πολύ συχνά σε δεξιοτεχνικά κομμάτια για πιάνο και κάποια τυπικά χαρακτηριστικά του οργάνου, όπως η απόκλιση από την ακριβή αρμονικότητα, ο θόρυβος που εισάγεται στην αρχή κάθε νότας από το σφυράκι του πιάνου, το γεγονός ότι οι τα υψηλότερα πλήκτρα του πιάνου έχουν περισσότερες από μία χορδές για κάθε νότα οι οποίες δεν είναι κουρδισμένες επακριβώς στην ίδια συχνότητα και το γεγονός ότι η κρούση μίας χορδής προκαλεί κι άλλες ελεύθερες χορδές να ηχήσουν. Επιπλέον, αποτελεί ένα ανοιχτό ερώτημα εάν ένα τόσο γενικό θέμα όπως η αυτόματη καταγραφή, θα πρέπει να εξεταστεί μέσω μιας γενικής προσέγγισης, όπως έχει γίνει για αρκετές δεκαετίες, ή αποδομώντας το γενικό πρόβλημα σε πιο συγκεκριμένες υποεργασίες, όπως την εξαγωγή της μελωδικής γραμμής ή της γραμμής του μπάσου, τον διαχωρισμό των πηγών και την εξειδικευμένη ανά όργανο καταγραφή το οποίο αποτελεί αντικείμενο των πιο πρόσφατων ερευνών.

Στην παρούσα διπλωματική μελετήθηκαν διάφορες πτυχές του προβλήματος της αυτόματης καταγραφής μουσικής. Έτσι έχει αναπτυχθεί και περιγράφεται ένα πλήρες σύστημα αυτόματης αναγνώρισης και καταγραφής μονοφωνικής μουσικής καθώς και μία μέθοδος ανίχνευσης πολλαπλών τόνων. Ένα σύστημα καταγραφής αποτελείται από διάφορες βαθμίδες. Βασική και πρωταρχική βαθμίδα είναι αυτή της εύρεσης των onsets, βαθμίδα η οποία από την ακουστική κυματομορφή τεμαχίζει το σήμα σε νότες. Στη συνέχεια η τεμαχισμένη ηχητική κυματομορφή δίνεται ως είσοδος στη βαθμίδα εύρεσης του pitch η οποία έχει ως στόχο να επιλύσει κάθε πολυφωνικό στοιχείο σε ένα συγκεκριμένο ήχο/νότα. Στα επόμενα κεφάλαια, οι επιμέρους βαθμίδες ενός συστήματος αυτόματης καταγραφής μουσικής συζητούνται με μεγαλύτερη λεπτομέρεια.

Γενικά τα υποποβλήματα που μελετήθηκαν στα πλαίσια της διπλωματικής είναι τα:

1. Ανίχνευση της αρχής κάθε νότας
2. Εξαγωγή ύψους-αναγνώριση νότας σε μονοφωνική μουσική
3. Εύρεση του μουσικού ρυθμού και του μουσικού μέτρου
4. Αναγνώριση συγχορδιών-εξαγωγή των θεμελιωδών συχνοτήτων τους

Στην περίπτωση του συστήματος αυτόματης καταγραφής μονοφωνικής μουσικής, η είσοδος του συστήματος είναι ένα ηχητικό μουσικό σήμα και η έξοδός του είναι μια συμβολική αναπαράσταση αυτού του σήματος, που περιλαμβάνει τις νότες που το απαρτίζουν. Η διαδικασία της καταγραφής μπορεί να χωριστεί σε δύο βασικά στάδια: στη μετατροπή του ηχητικού σήματος σε piano-roll αναπαράσταση και στη μετατροπή της αναπαράστασης piano-roll σε συνήθη μουσική σημειογραφία. Στην παρούσα εργασία καλύπτεται το πρώτο στάδιο, ενώ όμως επιπλέον της πληροφορίας που περιέχεται στο piano-roll εξάγεται και ρυθμική πληροφορία για το μουσικό κομμάτι (tempo, beats, μέτρο). Άλλωστε, οι περισσότεροι συγγραφείς θεωρούν την αυτόματη καταγραφή μουσικής σαν το πρόβλημα της μετατροπής από audio σε piano-roll αφού από μόνο του είναι ένα ιδιαίτερα απαιτητικό πρόβλημα, ενώ το πρόβλημα της μετατροπής από το piano-roll στη συνήθη μουσική σημειογραφία θεωρείται ανεξάρτητο.

1.4 Συνοπτική περιγραφή των συνεισφορών της παρούσας εργασίας

Οι κύριες πρωτότυπες συνεισφορές αυτής της εργασίας είναι η πρόταση μίας καινούριας μεθόδου για την εκτίμηση πολλαπλών τόνων (multipitch estimation method) καθώς και μίας μεθόδου για την εύρεση των onsets σε μονοφωνική μουσική που βασίζεται στον Teager τελεστή ενέργειας. Εκτός από αυτά, υλοποιήθηκε ένα πλήρες σύστημα αναγνώρισης μονοφωνικής μουσικής για πιάνο συνδυάζοντας κατάλληλα μία βαθμίδα εύρεσης των onsets με μία βαθμίδα εκτίμησης του pitch, που και οι δύο στηρίζονται σε μεθόδους της βιβλιογραφίας που έχουν να επιδείξουν πολύ καλά αποτελέσματα ([7] και [8]) και έγιναν κάποια πειράματα-προσομοιώσεις σε αυτό το σύστημα. Τέλος, έγινε κάποια δουλειά και στην κατεύθυνση της εύρεσης ρυθμικής πληροφορίας, συνδυάζοντας τρεις σχετικές εργασίες που επιλέχτηκαν κατόπιν ανασκόπησης της αντίστοιχης βιβλιογραφίας και αφορούν την εύρεση της περιόδου του beat, την εύρεση των χρονικών στιγμών των beats και τελικά την εύρεση του μέτρου του κομματιού ([6], [9] και [4]). Όλα τα προβλήματα που θίγονται στα πλαίσια της διπλωματικής ορίζονται και περιγράφονται αναλυτικά ενώ γίνεται και μια ανασκόπηση στη σχετική βιβλιογραφία.

Κεφάλαιο 2

Υπόβαθρο και Ιστορική Ανασκόπηση

Αυτό το κεφάλαιο περιγράφει τις έννοιες της επεξεργασίας σημάτων και της μουσικής θεωρίας που απαιτούνται για την κατανόηση της βάσης αυτής της διπλωματικής. Επίσης παρουσιάζει μία επισκόπηση της προγενέστερης έρευνας που έχει γίνει στο χώρο της αυτόματης καταγραφής μουσικής.

2.1 Μουσικό υπόβαθρο

2.1.1 Χαρακτηριστικά των μουσικών σημάτων

Τα μουσικά σήματα είναι ένα υποσύνολο των ηχητικών σημάτων και έχουν συγκεκριμένα χαρακτηριστικά που μπορούν να ληφθούν υπόψιν για την ανάλυσή τους. Παρατίθενται οι ορισμοί για τις ποσότητες εκείνες τις οποίες αντιλαμβάνεται το ανθρώπινο αυτί οι οποίες παίζουν πρωτεύοντα ρόλο στο χαρακτηρισμό ή τον διαχωρισμό των ήχων. Αυτές οι ποσότητες είναι οι λεγόμενες τέσσερις διαστάσεις του ήχου και είναι οι ακόλουθες: η ένταση (loudness), η χρονική διάρκεια (duration), το τονικό ύψος (pitch) και η χροιά (timbre). Με τη βοήθεια του [1] οι προαναφερθείσες ποσότητες ορίζονται ως εξής:

Η αντίληψη της *δυναμικής ή ηχηρότητας* ενός ακουστικού σήματος δεν συνδέεται με έναν αντίστοιχα απλό τρόπο με τις φυσικές ιδιότητές του, και τα σχετικά υπολογιστικά μοντέλα για την αντίληψή της αποτελούν βασικό πεδίο έρευνας της ψυχοακουστικής. Παρόλα αυτά, κατά την επεξεργασία της μουσικής είναι βολικό να εκφράζουμε την ηχηρότητα ενός ακουστικού σήματος ως τη μέση τετραγωνική τιμή του (ισχύς) εκφρασμένη σε λογαριθμική κλίμακα (decibel).

Η αντιλαμβανόμενη *χρονική διάρκεια* ενός ήχου αντιστοιχεί λίγο ως πολύ στη φυσική διάρκειά του σε περιπτώσεις που αυτή μπορεί να προσδιοριστεί σαφώς.

Το *pitch* είναι μια ιδιότητα εύκολα αντιληπτή από τον άνθρωπο που επιτρέπει την κατηγοριοποίηση των ήχων σε μια κλίμακα συχνοτήτων μεταξύ χαμηλών και υψηλών τιμών. Ακριβέστερα, το pitch ορίζεται ως η συχνότητα μιας ημιτονοειδούς κυματομορφής που ταιριάζει με τον ήχο που αντιλαμβανόμαστε [10]. Το pitch είναι η συχνότητα του ήχου, όπως γίνεται αντιληπτή από τον άνθρωπο. Η συχνότητα και το pitch δεν ταυτίζονται αλλά συσχετίζονται με μη γραμμικό

τρόπο. Η αντίληψη του pitch είναι υποκειμενική και εξαρτάται τόσο από τη συχνότητα όσο και από το επίπεδο της πίεσης του ήχου. Η θεμελιώδης συχνότητα (F0) είναι η αντιστοιχούσα φυσική σημασία του όρου και ορίζεται μόνο για περιοδικά ή σχεδόν περιοδικά σήματα. Για αυτή την κατηγορία σημάτων, η θεμελιώδης συχνότητα ορίζεται ως το αντίστροφο της περιόδου και είναι στενά συνδεδεμένη με το pitch. Σε διφορούμενες καταστάσεις, η περίοδος που αντιστοιχεί στην αντιλαμβανόμενη συχνότητα είναι αυτή που επιλέγεται. Το pitch ενός σύνθετου τόνου μπορεί να γίνει αντιληπτό ακόμα κι αν λείπει η συχνοτική συνιστώσα που αντιστοιχεί στο F0 (απούσα θεμελιώδης). Σημειώνεται ότι αν η ύπαρξη της θεμελιώδους συχνότητας ήταν απαραίτητη, τότε η αντρική φωνή δε θα ήταν αντιληπτή μέσω τηλεφώνου, καθώς συχνότητες κάτω από τα 300 Hz συνήθως φιλτράρονται. Τα περισσότερα μουσικά όργανα που χρησιμοποιούνται στη δυτική μουσική παράγουν αρμονικούς ήχους καθώς βασίζονται σε έναν αρμονικό ταλαντωτή όπως είναι μια χορδή ή μια στήλη από αέρα. Το φάσμα αυτών των ήχων εμφανίζει μια σειρά από μερικές αρμονικές, σε τακτά διαστήματα. Σε έναν ιδανικά αρμονικό ήχο, οι αρμονικές είναι ακέραια πολλαπλάσια της θεμελιώδους συχνότητας. Επομένως η F0 ενός αρμονικού ήχου μπορεί να οριστεί ως ο μέγιστος κοινός διαιρέτης των αρμονικών συχνοτήτων. Όμως στην πραγματικότητα οι αρμονικές δεν είναι ακριβή πολλαπλάσια της θεμελιώδους συχνότητας. Αυτό το φαινόμενο είναι γνωστό ως αναρμονικότητα (inharmonicity) και εμφανίζεται όταν η μια μερική συχνότητα h δεν ισούται ακριβώς με hF0. Σύμφωνα με τους Fletcher και Rossing [11], οι αρμονικές συχνότητες σε μια χορδή πιάνου υπακούν (με κάποιες αποκλίσεις) τη φόρμουλα:

$$f_h = h f_0 \sqrt{1 + Bh^2} \tag{2.1}$$

Μια τυπική τιμή για τον παράγοντα αναρμονικότητας για τη μεσαία έκταση του πιάνου είναι B=0.0004, που είναι αρκετό για να ολισθήσει η 17η αρμονική στην ιδανική συχνότητα για την 18η αρμονική. Επίσης σημειώνεται ότι στην περίπτωση του πιάνου, ένα μέρος του ήχου μπορεί να είναι μη αρμονικό κι αυτό συμβαίνει κυρίως στην αρχή μιας νότας όπου μη αρμονικοί ήχοι ενδεχομένως να παραχθούν από το σφυράκι του πιάνου όπως αυτό χτυπάει μια χορδή. Ένας αρμονικός ήχος μπορεί να εκφραστεί ως ένα άθροισμα από H ημιτονοειδή μαζί με ένα μοντέλο σφάλματος ε :

$$x[n] = \sum_{h=1}^{H} A_h[n] cos(2\pi f_h n + \phi_h(0)) + \varepsilon[n] \tag{2.2}$$

όπου A_h είναι το πλάτος του h-οστού ημιτονοειδούς που μεταβάλλεται συναρτήσει του χρόνου, f_h είναι η συχνότητα του ημιτονοειδούς, και $\phi_h(0)$ είναι η αρχική φάση. Αγνοώντας το θόρυβο, ένας αρμονικός ήχος μπορεί εν γένει να περιγραφεί από το σχετικό ύψος των αρμονικών του και την εξέλιξή τους στο χρόνο. Αυτό είναι γνωστό ως το αρμονικό πρότυπο (ή φασματικό πρότυπο). Θεωρώντας μόνο το φασματικό πλάτος των αρμονικών σε ένα δεδομένο χρονικό πλαίσιο, ένα φασματικό πρότυπο μπορεί να οριστεί σαν ένα διάνυσμα που συμπεριλαμβάνει το πλάτος κάθε αρμονικής. Στους περισσότερους μουσικούς ήχους, οι πρώτες αρμονικές είναι αυτές που περιέχουν το μεγαλύτερο μέρος της ενέργειας του σήματος.

Το *ηχόχρωμα ή χροιά* μιάς νότας είναι στενά συνδεδεμένο με την πηγή από την οποία έχει παραχθεί ο εκάστοτε ήχος, χαρακτηριστικό γνώρισμα το οποίο μελετάται για το λόγο αυτό σε εργασίες κατηγοριοποίησης των μουσικών οργάνων [12][13]. Δύο ήχοι που αναπαράγονται από διαφορετικά όργανα, αν και μπορεί να είναι πανομοιότυποι όσον αφορά τις τιμές pitch και δυναμικής, παραμένουν εύκολα διαχωρίσιμοι λόγω του διαφορετικού τους ηχοχρώματος. Ο λόγος δεν μπορεί να στηριχθεί σε κάποια ξεχωριστή ιδιότητα του ακουστικού σήματος αλλά εξαρτάται κυρίως από το πόσο 'τραχύ' είναι το ενεργειακό φάσμα του σήματος στο χρόνο και τη

χρονική του συνάρτηση. Καθορίζεται ουσιαστικά από τη συμμετοχή των αρμονικών όρων πέραν της θεμελιώδους ή βασικής συχνότητας του ήχου και από τα χαρακτηριστικά κορυφώματα της περιβάλλουσας του φάσματος. Ενώ λοιπόν το pitch και η δυναμική ενός σήματος μπορούν με απλό τρόπο να αναπαρασταθούν σε κλίμακα δύο διαστάσεων, το ηχόχρωμα αποτελεί ουσιαστικά πολυδιάστατη παράμετρο, και αναπαρίσταται τυπικά από ένα χαρακτηριστικό διάνυσμα.

2.1.2 Τονική δομή

Στη σύγχρονη δυτική μουσική το φάσμα των συχνοτήτων χωρίζεται σε οκτάβες. Μία μουσική νότα μπορεί να αναπαρασταθεί χρησιμοποιώντας ένα γράμμα και έναν αριθμό οκτάβας. Για παράδειγμα, το C_3 αναφέρεται στη νότα C από την τρίτη οκτάβα. Οι νότες που διαφέρουν κατά μία οκτάβα έχουν το ίδιο όνομα, δίνουν τη αίσθηση του ίδιου τόνου με μια μεγαλύτερη ή μικρότερη οξύτητα, και οι συχνότητές τους είναι η μία διπλάσια της άλλης. Υπάρχουν διαφορετικοί τρόποι για τη διευθέτηση ενός αριθμού από μουσικές νότες μέσα σε μια οκτάβα και την ανάθεση μίας συχνότητας σε κάθε μία από αυτές. Στη δυτική μουσική, το πιο συνηθισμένο είναι το ισοσυγκερασμένο σύστημα των 12 τόνων, που διαχωρίζει κάθε οκτάβα σε 12 λογαριθμικά ίσα μέρη ή ημιτόνια, τα: A, A#, B, C, C#, D, D#, E, F, F#, G και G# . Ο λόγος των συχνοτήτων δύο διαδοχικών ημιτονίων είναι σταθερός, και ίσος με $2^{\frac{1}{12}}$, έτσι οι συχνότητες των μουσικών νοτών είναι λογαριθμικά κατανεμημένες στον άξονα των συχνοτήτων. Οι θεμελιώδεις συχνότητες των μουσικών νοτών φαίνονται στο σχήμα 2.1.

Note Frequency Table

Frequency in Hz

Based on formula: $Note_N = Note_{N-1} \times 2^{\uparrow}(1/12)$

Note in: / For:	Octave: 0	1	2	3	4	5	6	7	8
C	16.3516	32.7032	65.4064	130.813	261.626	523.251	1046.50	2093.00	4186.01
C#	17.3239	34.6478	69.2957	138.591	277.183	554.365	1108.73	2217.46	4434.92
D	18.3540	36.7081	73.4162	146.832	293.665	587.330	1174.66	2349.32	4698.64
D#	19.4454	38.8909	77.7817	155.563	311.127	622.254	1244.51	2489.02	4978.03
E	20.6017	41.2034	82.4069	164.814	329.628	659.255	1318.51	2637.02	5274.04
F	21.8268	43.6536	83.3071	174.614	349.228	698.456	1696.91	2793.83	5587.65
F#	23.1247	46.2493	92.4986	184.997	369.994	739.989	1479.98	2959.96	5919.91
G	24.4997	48.9994	97.9989	195.998	391.995	783.991	1567.98	3135.96	6271.93
G#	25.9565	51.9131	103.826	207.652	415.305	830.609	1661.22	3322.44	6644.88
A	27.5	55.0	110.0	220.0	440.0	880.0	1760.0	3520.0	7040.0
A#	29.1352	58.2705	116.541	233.082	466.164	932.328	1864.66	3729.31	7458.62
B	30.8671	63.7354	123.471	246.942	493.883	987.767	1975.53	3951.07	7902.13

Σχήμα 2.1: Θεμελιώδεις συχνότητες των νοτών. Ανατύπωση από το [2].

2.1.3 Ρυθμός

Ο ρυθμός χαρακτηρίζεται από πρότυπα μουσικών μονάδων που εμφανίζονται σε διαφορετικά ιεραρχικά ρυθμικά επίπεδα. Το μήκος των νοτών εξαρτάται από την διάρκεια για την οποία παίζονται, καθώς και από το μουσικό tempo και το μετρικό οπλισμό. Οι βασικές ρυθμικές μονάδες αποκαλούνται beats και ο ρυθμός επανάληψης αυτών των beats δίνει το tempo. Το tempo είναι αντιστρόφως ανάλογο της περιόδου του beat. Θεωρώντας μια περίοδο beat T_b εκφρασμένη σε seconds, το tempo μπορεί να υπολογιστεί ως $T = 60/T_b$. Στην περίπτωση δηλαδή που το tempo ισούται με 60, υπάρχει ένα beat ανά δευτερόλεπτο. Ο μετρικός οπλισμός, που δηλώνεται με ένα κλάσμα στην αρχή ενός κομματιού καθορίζει δύο χαρακτηριστικά της μουσικής. Ο αριθμητής δηλώνει τον αριθμό των μετρικών μονάδων μέσα σε ένα μέτρο (απόσταση μεταξύ δύο διαστολών[1]), δηλαδή τα πρότυπα με βάση τα οποία ομαδοποιούνται τα beats. Για παράδειγμα, ένα κομμάτι στο οποίο τα beats ομαδοποιούνται σε ζεύγη, "ένα,δύο,ένα-δύο,ένα-δύο..." συμβολίζεται ότι έχει δίσημο μέτρο [2]. Ο παρονομαστής δηλώνει ποιά είναι η μετρική μονάδα στην οποία αντιστοιχεί το κάθε beat. Ένας τυπικός μετρικός οπλισμός είναι τα 4-4, που αντιστοιχεί σε 4 beats ανά μέτρο, όπου το κάθε beat είναι μια νότα του ενός τετάρτου. Το μήκος των νοτών σε δευτερόλεπτα είναι σχετικό με το μήκος ενός μέτρου. Ο μετρικός οπλισμός και το tempo καθορίζουν τη διάρκεια τόσο του μέτρου όσο και των μεμονωμένων νοτών. Στο σχήμα 2.2 φαίνονται τα πολλαπλάσια και τα υποπολλαπλάσια ενός τετάρτου και οι μεταξύ τους σχέσεις.

Όνομα	Βασικές Αξίες	Παρεστιγμένα	Τρίηχα	Πεντάηχα
Ολόκληρο	𝅝 = 15 (60 ÷ 4)	𝅝. = 10 (30 ÷ 3)	^{3}𝅝= 22,5 (7,5 × 3)	
Μισό	𝅗𝅥 = 30 (60 ÷ 2)	𝅗𝅥. = 20 (60 ÷ 3)	^{3}𝅗𝅥 = 45 (15 × 3)	^{5}𝅗𝅥 = 37,5 (7,5 × 5)
Τέταρτο	♩ = 60	♩. = 40 (120 ÷ 3)	^{3}♩ = 90 (30 × 3)	^{5}♩ = 75 (15 × 5)
Όγδοο	♪= 120 (60×2)	♪. = 80 (240 ÷ 3)	^{3}♪= 180 (60 × 3)	^{5}♪= 150 (30 × 5)
Δέκατο έκτο	𝅘𝅥𝅯= 240 (60 × 4)	𝅘𝅥𝅯. = 160 (480 ÷ 3)	^{3}𝅘𝅥𝅯= 360 (120 × 3)	^{5}𝅘𝅥𝅯= 300 (60 × 5)
Τριακοστό δεύτερο	𝅘𝅥𝅰= 480 (60 × 8)	𝅘𝅥𝅰. = 320 (960 ÷ 3)	^{3}𝅘𝅥𝅰= 720 (240 × 3)	^{5}𝅘𝅥𝅰= 600 (120 × 5)
Εξηκοστό τέταρτο	𝅘𝅥𝅱= 960 (60 × 16)	𝅘𝅥𝅱. = 640 (1920 ÷ 3)	^{3}𝅘𝅥𝅱= 1440 (480 × 3)	^{5}𝅘𝅥𝅱= 1200 (240 × 5)

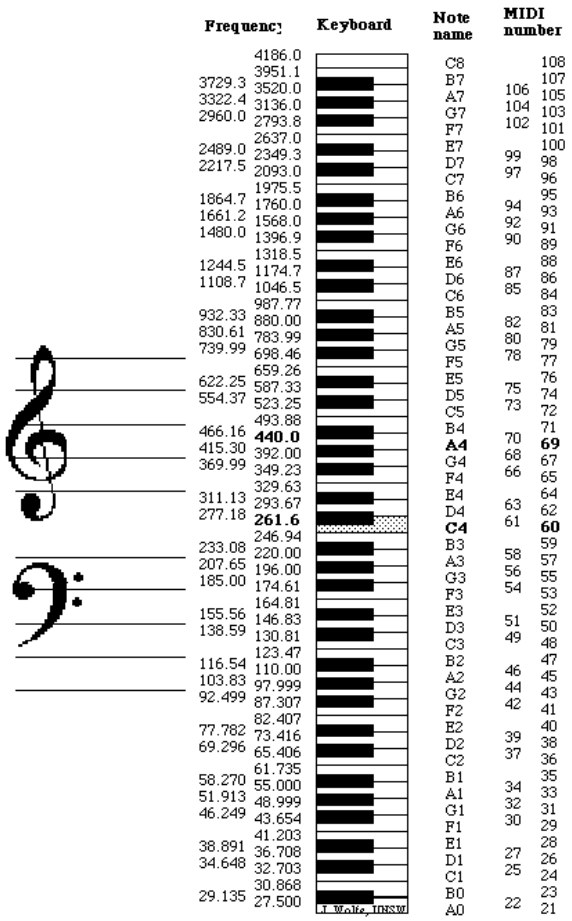
Σχήμα 2.2: Διάρκειες νοτών. Από το http://3euk1l4.blogspot.gr/2007/10/blog-post.html

2.1.4 Πρωτόκολλο MIDI

Το πρωτόκολλο MIDI (Musical Instrument Digital Interface) αναφέρεται στον εξ' αποστάσεως έλεγχο και επικοινωνία ανάμεσα σε ηλεκτρονικά μουσικά όργανα και άλλες συσκευές - όπως ηλεκτρονικούς υπολογιστές με λογισμικό μουσικής εγγραφής (sequencers), ρυθμομηχανές (drum-machines), δειγματολήπτες (samplers), συνθετητές με δυνατότητα μουσικής εγγραφής (workstation synthesizers), συσκευές συγχρονισμού - ανεξαρτήτως κατασκευαστή. Μηνύματα

[1] Διαστολές είναι οι κάθετες γραμμές που εκτείνονται από την 1η μέχρι την 5η γραμμή του πενταγράμμου.

[2] Η ομαδοποίηση των βασικών μετρικών μονάδων σε ζεύγη. Αντίστοιχα σε ένα τρίσημο μέτρο οι βασικές μετρικές μονάδες είναι ομαδοποιημένες σε τριάδες κτλ.



Σχήμα 2.3: Συχνότητες, ονόματα νοτών και MIDI αριθμοί. Από το http://www.phys.unsw.edu.au/jw/notes.html

για γεγονότα όπως το pitch και η ταχύτητα (velocity) μιας νότας μπορούν να μεταδοθούν χρησιμοποιώντας το πρωτόκολλο. Μπορεί επίσης να ελέγξει παραμέτρους όπως την ένταση (volume), το βιμπράτο, το ηχητικό panning, το μουσικό κλειδί, και σήματα ρολογιού που θέτουν το tempo. Το MIDI δεν παράγει κάποιο ήχο, αλλά μπορεί να χρησιμοποιηθεί για να ελέγξει ένα όργανο MIDI που θα παράγει το συγκεκριμένο ήχο.

Στο MIDI, το pitch μιας νότας κωδικοποιείται χρησιμοποιώντας έναν αριθμό (βλέπε σχήμα 2.3). Μία συχνότητα f μπορεί να μετατραπεί σε ένα MIDI picth αριθμό χρησιμοποιώντας την εξίσωση:

$$p = 69 + 12 \log_2 \frac{f}{440Hz} \tag{2.3}$$

Αντίστροφα, η συχνότητα f ενός δοσμένου MIDI pitch αριθμού n μπορεί να ανακτηθεί χρησιμοποιώντας την εξίσωση:

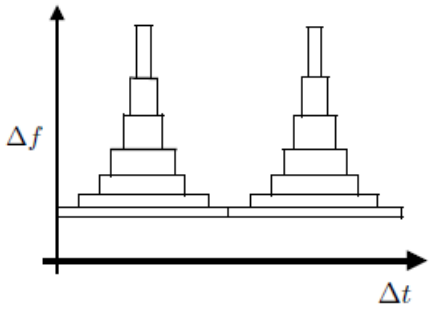
$$f = 440 * 2^{\frac{n-69}{12}} \tag{2.4}$$

Τα μηνύματα MIDI, μαζί με πληροφορίες συγχρονισμού, μπορούν να συλλεχθούν και να

αποθηκευτούν σε ένα SMF (standart midi file). Αυτή είναι η πιο εκτενής μορφή συμβολικού αρχείου στη μουσική για υπολογιστές. Το piano-roll είναι μια φυσική αναπαράσταση ενός SMF με το χρόνο στον οριζόντιο άξονα, και το pitch στον κατακόρυφο. Οι νότες αναπαρίστανται χρησιμοποιώντας οριζόντιες ράβδους στο πλέγμα χρόνου-συχνότητας. Παρόλο που το MIDI format έχει κάποιους περιορισμούς, τα SMF's συνήθως θεωρούνται μια καλή αναπαράσταση στα πλαίσια της υπολογιστικής καταγραφής μουσικής. Στη συνέχεια η μουσική παρτιτούρα μπορεί να εξαχθεί από το αρχείο MIDI. Βέβαια η συμβολική μουσική πληροφορία που αποθηκεύεται σε ένα SMF μπορεί να αναπαρασταθεί με διαφορετικούς τρόπους από ένα λογισμικό, κι έτσι η εξαγωγή μιας σύγχρονης παρτιτούρας από αυτή την πληροφορία δεν είναι ένα καλά ορισμένο πρόβλημα.

2.2 Θεωρητικό υπόβαθρο

2.2.1 Μετασχηματισμός σταθερού Q - Constant Q Transform (CQT)



Σχήμα 2.4: Μετασχηματισμός constant Q.

Όπως αναφέρθηκε και σε προηγούμενη ενότητα, οι θεμελιώδεις συχνότητες των μουσικών νοτών είναι λογαριθμικά κατανεμημένες στον άξονα των συχνοτήτων. Αναφέρεται ενδεικτικά ότι το εύρος της οκτάβας 3 είναι 220 Hz ενώ το εύρος της οκτάβας 8 είναι 7 kHz. Το γεγονός αυτό υποδεικνύει ότι για την ανάλυση ενός σήματος μουσικής απαιτείται μεγαλύτερη ακρίβεια στις χαμηλές συχνότητες και μικρότερη στις υψηλές.

Μια παραλλαγή του διακριτού μετασχηματισμού STFT (Short Time Fourier Transform) που πετυχαίνει σταθερό λόγο συχνότητας προς ακρίβεια ανάλυσης (όπως το ανθρώπινο αυτί) χρησιμοποιώντας μεταβλητό μήκος παραθύρου, είναι ο μετασχηματισμός constant Q, ένα υβρίδιο μεταξύ STFT και wavelet. Στο μετασχηματισμό constant Q ο λόγος της συχνότητας προς την ανάλυση είναι σταθερός και ίσος με Q. Αυτό σημαίνει ότι κάθε φασματική συνιστώσα k διαχωρίζεται με μεταβλητή ανάλυση συχνότητας $\Delta f_k = f_k/Q$. Ο μετασχηματισμός ορίζεται από τον τύπο:

$$X^Q[k] = \frac{1}{N[k]} \sum_{n=0}^{N[k]-1} x[n]w[n,k]e^{-j\frac{2\pi}{N[k]}Qn} \tag{2.5}$$

όπου $x[n]$ το σήμα, $N[k]$ είναι το μέγεθος του παραθύρου (σε δείγματα) που χρησιμοποιείται για τον υπολογισμό του μετασχηματισμού της συχνότητας k:

$$N[k] = f_s/\Delta f_k = (f_s/f_k)Q \tag{2.6}$$

Η συνάρτηση παραθύρωσης $w[n,k]$ που χρησιμοποιείται για να περιορίσει τη φασματική διαρροή έχει το ίδιο σχήμα αλλά διαφορετικό μήκος για κάθε συνιστώσα.

Ο μετασχηματισμός αυτός υπολογίζει το πλάτος σε ορισμένες συχνότητες του φάσματος λογαριθμικά κατανεμημένες, για κάθε μια από τις οποίες χρησιμοποιεί διαφορετικό μήκος παραθύρου για την επίτευξη της επιθυμητής ακρίβειας στη συχνότητα. Αυτό βέβαια έχει ως συνέπεια στις χαμηλές συχνότητες όπου το μήκος του παραθύρου είναι μεγαλύτερο, να είναι μικρότερη η ανάλυση στο πεδίο του χρόνου. Το μειονέκτημα αυτής της μεθόδου είναι ότι δεν εκμεταλλεύεται τη μεγαλύτερη ανάλυση στο χρόνο που μπορεί να επιτευχθεί χρησιμοποιώντας συντομότερα παράθυρα στις υψηλές συχνότητες, χάνοντας την κάλυψη στο πεδίο χρόνου-συχνότητας όπως φαίνεται στην εικόνα 2.4.

2.2.2 Φίλτρα στον τομέα της συχνότητας

Σε εφαρμογές όπου κάποιες συχνότητες είναι πιο σημαντικές από κάποιες άλλες, είναι χρήσιμο να μπορούμε να απομονώσουμε συγκεκριμένες συχνότητες. Τα φίλτρα στον τομέα της συχνότητας εξυπηρετούν την παροχή πληροφοριών για ορισμένες περιοχές συχνοτήτων και μπορούν να χρησιμοποιηθούν για να ενισχύσουν τις επιθυμητές συχνότητες και να αφαιρέσουν τις ανεπιθύμητες.

Μία συστοιχία φίλτρων διαχωρίζει το σήμα εισόδου σε διάφορες ζώνες συχνοτήτων χρησιμοποιώντας μία ακολουθία ζωνοπερατών φίλτρων. Η έξοδος μίας συστοιχίας φίλτρων είναι μία ακολουθία από φιλτραρισμένες τιμές, όπου κάθε μία αντιστοιχεί στο αποτέλεσμα του φιλτραρίσματος του φάσματος του σήματος εισόδου με ένα μεμονωμένο φίλτρο. Οι περισσότερες τράπεζες φίλτρων έχουν φίλτρα με τα άκρα τους να τοποθετούνται με τέτοιο τρόπο ώστε να συμπίπτουν με τις κεντρικές συχνότητες των γειτονικών φίλτρων.

2.2.3 Teager Τελεστής Ενέργειας

Θεωρούμε έναν ταλαντωτή με μάζα m και σταθερά ελατηρίου k. Τότε η μετατόπιση του ταλαντωτή υπακούει στην εξίσωση:

$$m\frac{d^2x}{dt^2} + kx = 0 \tag{2.7}$$

για την οποία η γενική λύση είναι ένα συνημίτονο:

$$x(t) = Acos(\omega_0 t + \theta) \tag{2.8}$$

με $\omega_0 = \sqrt{k/m}$.

Η συνολική ενέργεια του συστήματος είναι σταθερή και ίση με το άθροισμα της κινητικής και της δυναμικής ενέργειας:

$$E = \frac{1}{2}kx^2 + \frac{1}{2}m(\dot{x}^2) = \frac{1}{2}m\omega_0^2 A^2 \tag{2.9}$$

Έτσι, η ενέργεια του γραμμικού ταλαντωτή είναι ανάλογη τόσο του τετραγωνικού πλάτους όσο και της τετραγωνικής συχνότητας της ταλάντωσης. Με βάση τα παραπάνω οι Teager και Kaiser πρότειναν τον τελεστή Teager-Kaiser Ψ:

$$\Psi[x(t)] = \dot{x}^2(t) - x(t)\ddot{x}(t) \tag{2.10}$$

Όταν ο τελεστής εφαρμοστεί στο $x(t) = Acos(\omega_0 t + \theta)$, δίνει μία μέτρηση για την ενέργεια που εξαρτάται τόσο από το πλάτος όσο και από τη συχνότητα:

$$\Psi[x(t)] = A^2(t)\omega_0^2(t) = \frac{E_0}{(m/2)} \tag{2.11}$$

Περισσότερες λεπτομέρειες για τον Teager τελεστή ενέργειας μπορούν να βρεθούν στο [14].

2.3 Ιστορία της αυτόματης καταγραφής μουσικής

Γίνεται μία σύντομη αναφορά στην ιστορία της αυτόματης καταγραφής μουσικής από την δεκαετία του '70 και μετά όπως αυτή παρουσιάζεται στο πρώτο κεφάλαιο [1] του βιβλίου "Signal Processing Methods for Music Transcription" των Anssi Klapuri και Manuel Davy (2006) [15].

"Οι πρώτες απόπειρες για την αυτόματη καταγραφή πολυφωνικής μουσικής έγιναν τη δεκαετία του '70, όταν ο Moorer πρότεινε ένα σύστημα για την καταγραφή συνθέσεων 2 φωνών ([16],[17]). Οι Chafe et al. [18], Piszcalski [19] και Maher ([20],[21]) συνέχισαν τη δουλειά του Moorer κατά τη δεκαετία του '80. Τα πρώιμα αυτά συστήματα, περιορίζονταν στην αναγνώριση δύο το πολύ ταυτόχρονων φωνών ενώ οι σχέσεις των pitch που μπορούσαν αυτές να έχουν ήταν περιορισμένες με διαφόρους τρόπους. Όσον αφορά την ανάλυση του ρυθμού, ο πρώτος αλγόριθμος εκτίμησης του ρυθμού σε γενικού περιεχομένου ακουστικά σήματα προτάθηκε από τους Goto και Muraoka [22] την επόμενη δεκαετία, του '90, παρότι είχε προηγηθεί αξιοσημείωτος όγκος δουλειάς στην εκτίμηση του ρυθμού παραμετρικών δεδομένων από νότες καθώς και ο αλγόριθμος του Schloss [23] για την εκτίμηση του ρυθμού κρουστών οργάνων. Οι πρώτες απόπειρες καταγραφής των κρουστών οργάνων έγιναν στα μέσα του '80 από τον Schloss [23] και αργότερα από τον Bilmes [24]. Και οι δύο κατηγοριοποιούσαν διαφορετικού τύπου "χτύπους" σε συνεχείς ηχογραφήσεις. Η καταγραφή πολυφωνικών κρουστών ήχων πραγματοποιήθηκε αργότερα από τους Goto και Muraoka [25].

Από τις αρχές της δεκαετίας του '90, το ενδιαφέρον για την μουσική καταγραφή έχει αυξηθεί ταχύτατα και δεν είναι εφικτό να γίνει ο συνολικός απολογισμός της δουλειάς στα πλαίσια αυτής της εργασίας. Για το λόγο αυτό, θα αναφερθούν μόνο συγκεκριμένες τάσεις και επιτυχημένες προσεγγίσεις που έχουν σημειωθεί. Μία από αυτές είναι η αξιοποίηση στατιστικών μεθόδων (statistical methods). Σημαντικά παραδείγματα χρήσης στατιστικών μεθόδων στην ανάλυση των πολλαπλών pitch της πολυφωνικής μουσικής αποτελούν οι μέθοδοι που προτάθηκαν από τους Kashino et al [26], Goto [27], Davy και Godsill [28], και Ryynanen και Klapuri [29]. Στο πεδίο της εκτίμησης του ρυθμού, στατιστικές μέθοδοι εφαρμόστηκαν από τους Cemgil και Kappen [30], Hainsworth και MacLeod [31], και Klapuri et al. [32], ενώ στη μελέτη της καταγραφής κρουστών οργάνων από τους Gillet και Richard [33] και Paulus και Klapuri [34]. Η χρήση όμως στατιστικών μεθόδων αναγνώρισης προτύπων επικράτησε στο πεδίο της κατηγοριοποίησης των μουσικών οργάνων [35]. Μία ακόμη τάση αποτέλεσε η χρησιμοποίηση υπολογιστικών μοντέλων του ανθρώπινου ακουστικού συστήματος (computational models of the human auditory system). Τέτοιες τεχνικές για την μουσική καταγραφή παρουσιάστηκαν από τον Martin [36], και οδήγησαν στην πρόταση σχετικών μεθόδων για την ανάλυση των pitch πολυφωνικών ήχων από τους Karjalainen και Tolonen [37] και Klapuri [38], και μεθόδων για την εκτίμηση του ρυθμού από τον Scheirer [39], για παράδειγμα. Μια σπουδαία προσέγγιση ήταν αυτή της μοντελοποίησης του μουσικού τοπίου (Auditory Scene Analysis - ASA). Ο όρος ASA αναφέρεται στον τρόπο με τον οποίο οι άνθρωποι οργανώνουν τα φασματικά χαρακτηριστικά με τις αντίστοιχες πηγές

ήχων τους και αναγνωρίζουν ταυτόχρονους ήχους [40]. Οι αρχές του ASA εφαρμόστηκαν στην ανάλυση των pitch πολυφωνικών μουσικών σημάτων από τους Mellinger [41] και Kashimo και Tanaka [42], και αργότερα από τους Godsmark και Brown [43] και τους Sterian et al. [44]. Πιο πρόσφατα, κάποιες μέθοδοι μάθησης χωρίς επίβλεψη (unsupervised learning) έχουν προταθεί κατά τις οποίες ένας ελάχιστος αριθμός από πρωταρχικές υποθέσεις λαμβάνονται για το υπό ανάλυση σήμα. Μέθοδοι βασισμένες στην ανάλυση ανεξάρτητων συνιστωσών (independent component analysis) [45] προτάθηκαν από τον Casey ([46], [47]) και άλλες διαφορετικές μέθοδοι παρουσιάστηκαν αργότερα από τους Lepain [48], Smaragdis ([49], [50]), Abdallh ([51], [52]), Virtanen [53], FitzGerald ([54], [55]) και Paulus και Virtanen [56]. “

Στη συνέχεια της ίδιας ενότητας του ίδιου βιβλίου [1] γίνεται ένα σχόλιο πάνω στο πόσο εξελιγμένες είναι οι δυνατότητες που μας παρέχουν σήμερα οι διαθέσιμες τεχνικές για την αυτόματη καταγραφή μουσικής:

“ Ένα γενικού σκοπού πρακτικά εφαρμόσιμο σύστημα καταγραφής δεν έχει υπάρξει ακόμα. Παρόλα αυτά έχει επιτευχθεί μια κάποια επιτυχία στην καταγραφή πολυφωνικής μουσικής θέτοντας περιορισμούς στην πολυπλοκότητα. Στην καταγραφή οργάνων τα οποία παράγουν τόνους (pitch), τυπικοί περιορισμοί είναι ο μέγιστος αριθμός των ταυτόχρονων ήχων ([37], [28]), η απαγόρευση της παρεμβολής ντράμς και άλλων κρουστών οργάνων [57], ή το γεγονός ότι μόνο ένα συγκεκριμένο όργανο λαμβάνεται υπόψη [58]. Κάποια ελπιδοφόρα αποτελέσματα στην καταγραφή μουσικής του πραγματικού κόσμου από ηχογραφήσεις CD έχουν να επιδείξουν οι Goto [27] και οι Ryynanen και Klapuri [29]. Στην καταγραφή των κρουστών έχει επιτευχθεί αρκετά καλή ακρίβεια για κρουστικά κομμάτια που περιλαμβάνουν περιορισμένο αριθμό μουσικών οργάνων και άτονα μουσικά όργανα (μουσικά όργανα τα οποία δεν παράγουν τόνο/pitch αλλά κρότο) ([33], [56]). Επίσης υποσχόμενα αποτελέσματα έχουν παρουσιαστεί στην καταγραφή των μπάσων και του τυμπάνου σε πραγματικού κόσμου ηχογραφήσεις, αλλά αυτό είναι ένα πιο ανοικτό πρόβλημα (βλέπε Zils et al. [59], FitzGerald et al. [60], Yoshii et al. [61]). Η εκτίμηση του ρυθμού πολύπλοκων ακουστικών σημάτων μπορεί να γίνει με αρκετή αξιοπιστία με τις δεδομένες ανωτάτου επιπέδου τεχνικές, αλλά οι δυσκολίες παραμένουν ιδιαίτερα κατά την ανάλυση της κλασικής μουσικής και ρυθμικά περίπλοκου υλικού. Μια συγκριτική αξιολόγηση των εργασιών εκτίμησης του ρυθμού παρουσιάζεται στα [31], [32] και [62]. Οι έρευνες αναφορικά με την κατηγοριοποίηση των μουσικών οργάνων έχουν επικεντρωθεί στο επίπεδο των μεμονωμένων ήχων, παρότι πιο πρόσφατα επιχειρούνται και σε πολυφωνικά ακουστικά σήματα ([63], [64], [65], [66]). “

Κεφάλαιο 3

Σύστημα αυτόματης καταγραφής μονοφωνικής μουσικής

3.1 Ανίχνευση της αρχής (onset) της νότας

3.1.1 Περιγραφή προβλήματος και σχετικές εργασίες

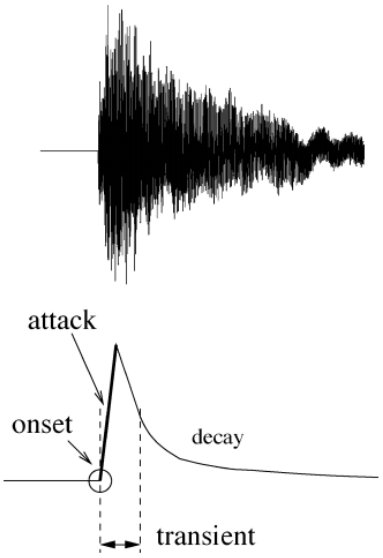
Ο όρος onset σε ένα μουσικό σήμα περιγράφει την αντιληπτή αρχή ενός διακριτού γεγονότος. Συνεπώς ο όρος ανίχνευση onset αναφέρεται στην ανίχνευση της αρχής ενός διακριτού γεγονότος σε ένα ηχητικό σήμα. Αυτά τα γεγονότα μπορεί να είναι ήχοι με pitch ή χωρίς. Ο στόχος ενός συστήματος ανίχνευσης των onsets είναι να διαχωρίσει το σήμα σε μικρότερες μονάδες (μεμονωμένες νότες, συγχορδίες, κτλ.). Η εύρεση των onsets είναι χρήσιμη σε διάφορες εφαρμογές, από την εκτίμηση του tempo και τον εντοπισμό των beats, μέχρι την εκτίμηση των θεμελιωδών συχνοτήτων, εφόσον η αρχή των νοτών μπορεί να χρησιμοποιηθεί για τον τεμαχισμό του σήματος.

Η πιο συνηθισμένη προσέγγιση του προβλήματος της εύρεσης των onsets είναι η αναζήτηση κάποιων "μεταβατικών" περιοχών στο σήμα. Μια μεταβατική περιοχή μπορεί να προσδιορίζεται από μία απότομη αύξηση της ενέργειας, μια αλλαγή στο φάσμα βραχέως χρόνου του σήματος, στις στατιστικές ιδιότητές του, κτλ.

Μια καλή επισκόπηση των διαφόρων τεχνικών για την εύρεση των onsets γίνεται στο [3]. Αρχικά, ξεκαθαρίζονται κάποιες βασικές έννοιες.

Στο σχήμα 3.1 φαίνεται πώς μπορούν να διαχωριστούν οι έννοιες transient, onset και attack.

- Το attack της νότας είναι το χρονικό διάστημα κατά το οποίο η περιβάλλουσα του πλάτους αυξάνει.
- Η έννοια του transient είναι πιο δύσκολο να περιγραφτεί επακριβώς. Άτυπα το transient μπορεί να οριστεί ως ένα σύντομο χρονικό διάστημα κατά το οποίο το σήμα μεταβάλλεται ταχέως κατά απρόβλεπτο τρόπο. Στην περίπτωση των ακουστικών οργάνων, το transient αντιστοιχεί στο χρονικό διάστημα κατά το οποίο εφαρμόζεται η διέγερση και στη συνέχεια η απόκριση μειώνεται αφήνοντας μια αργή εξασθένιση στις συχνότητες συντονισμού του οργάνου. Κατά τη διάρκεια του offset της νότας επίσης παρατηρείται μία περίοδος transient.



Σχήμα 3.1: Onset, Attack, Transient και Decay μιας νότας. Η εικόνα προέρχεται από το [3].

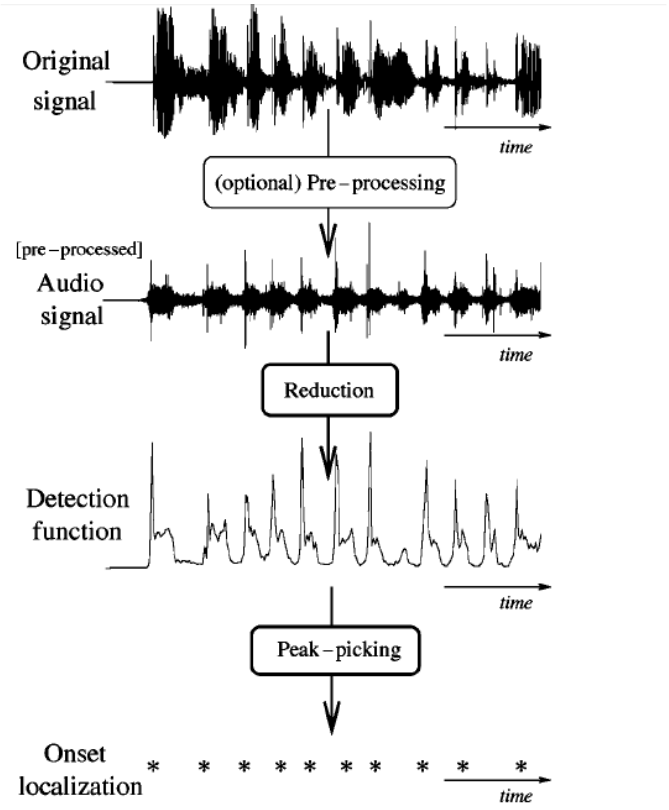
- Το onset μιας νότας είναι η χρονική στιγμή που ξεκινάει το transient (ή η στιγμή που μπορούμε να διακρίνουμε την αρχή του). Κάποιες εργασίες ορίζουν τα onsets σαν τη χρονική στιγμή που εμφανίζεται το attack.

Σε ένα πραγματικό σήμα (το οποίο εν γένει είναι πολυφωνικό και πολύ πιθανό να έχει υπερτεθεί και θόρυβος) οι παραπάνω έννοιες δεν μπορούν να διακριθούν με ακρίβεια. Έτσι, τα onsets δεν μπορούν να ανιχνευθούν απευθείας από το σήμα ή την παραγοντοποίησή του στο πεδίο του χρόνου. Πρέπει να βρεθεί ένα ενδιάμεσο σήμα που να ανταναικλά, σε μια απλοποιημένη μορφή, την τοπική δομή του αρχικού σήματος. Αυτό το ενδιάμεσο σήμα αποκαλείται συνάρτηση ανίχνευσης (detection function).

Στο σχήμα 3.2 παρουσιάζεται η διαδικασία που ακολουθείται από την πλειοψηφία των αλγορίθμων ανίχνευσης των onsets: από το αρχικό ηχητικό σήμα, το οποίο μπορεί να προ-επεξεργαστεί για τη βελτίωση της απόδοσης των επόμενων σταδίων, εξάγεται μια συνάρτηση ανίχνευσης των αλλαγών στο σήμα, στην οποία ο αλγόριθμος εύρεσης κορυφών (peak-picking algorithm) εφαρμόζεται για να εντοπίσει τα onsets. Στη συνέχεια αναλύονται με μεγαλύτερη ακρίβεια τα βήματα αυτά.

Προ-επεξεργασία

Ο όρος προ-επεξεργασία υποδεικνύει το μετασχηματισμό του αρχικού σήματος με σκοπό την εξασθένιση ή την ενίσχυση κάποιων πτυχών του σήματος, ώστε να απλοποιηθεί ο επιθυμητός στόχος. Η προεπεξεργασία είναι ένα προαιρετικό βήμα. Στην βιβλιογραφία αναφέρονται διάφορες μέθοδοι προ-επεξεργασίας που διευκολύνουν το πρόβλημα της ανίχνευσης των onsets. Τα πιο πολλά συστήματα σε αυτό το στάδιο μετασχηματίζουν το σήμα στο πεδίο της συχνότητας. Μία από τις πιο κλασικές μεθόδους είναι ο διαχωρισμός του σήματος σε πολλαπλές ζώνες συχνοτήτων και η ανεξάρτητη ανάλυση της πληροφορίας στις διαφορετικές ζώνες ([67], [68], [69], [70]



Σχήμα 3.2: Διάγραμμα ροής ενός τυπικού αλγορίθμου εύρεσης των onsets. Η εικόνα προέρχεται από το [3].

και [39]). Θεωρείται ότι η ανάλυση σε ζώνες συχνοτήτων προσδίδει ευρωστία στους αλγορίθμους ανίχνευσης των onsets. Ακόμα έχει χρησιμοποιηθεί μετασχηματισμός constant Q και wavelet αποσύνθεση.

Εξαγωγή συνάρτησης ανίχνευσης

Η συνάρτηση ανίχνευσης είναι ένας μετασχηματισμός του ηχητικού σήματος που δείχνει την εμφάνιση των transients στο αρχικό σήμα. Είναι η βασική διαδικασία στους περισσότερους αλγορίθμους ανίχνευσης των onsets. Υπάρχουν δύο κατηγορίες τεχνικών για την εξαγωγή της συνάρτησης ανίχνευσης: οι ντετερμινιστικές τεχνικές που βασίζονται στη χρήση χαρακτηριστικών του σήματος, χρονικών ή φασματικών (με ή χωρίς χρήση πληροφορίας για τη φάση) και οι στατιστικές τεχνικές που βασίζονται στην υπόθεση ότι το σήμα μπορεί να περιγραφεί από ένα στατιστικό μοντέλο [71]. Κάποια παραδείγματα συναρτήσεων ανίχνευσης είναι η περιβάλλουσα του σήματος στο πεδίο του χρόνου [23], η ενέργεια βραχέος χρόνου [22], η φασματική ροή (εξηγείται στη συνέχεια) ([72], [69], [73] και [74]) και η μεταβολή της φάσης ([75] και [7]).

Επιλογή Κορυφών

Εάν η συνάρτηση ανίχνευσης έχει σχεδιαστεί κατάλληλα, τα onsets και άλλα απότομα γεγονότα θα οδηγήσουν σε καλά εντοπισμένα αναγνωρίσιμα χαρακτηριστικά στη συνάρτηση ανίχνευσης. Συνήθως αυτά τα χαρακτηριστικά είναι τοπικά μέγιστα, συχνά αντικείμενα κάποιου βαθμού μεταβλητότητας στο μέγεθος και το σχήμα, και επισκιάζονται από 'θόρυβο' που είτε είναι πραγματικός θόρυβος που υπάρχει στο σήμα, ή άλλες πτυχές του σήματος που δεν σχετίζονται με τα onsets. Επομένως, ένας εύρωστος αλγόριθμος επιλογής κορυφών είναι απαραίτητος για να εκτιμήσει τους χρόνους των onsets του σήματος. Η διαδικασία της επιλογής κορυφών από μια συνάρτηση ανίχνευσης μπορεί να χωριστεί σε 3 στάδια: μετα-επεξεργασία, κατωφλίωση, και μια τελική διαδικασία απόφασης.

Μετα-επεξεργασία: Όπως και η προ-επεξεργασία, η μετα-επεξεργασία είναι ένα προαιρετικό βήμα που εξαρτάται και από τη μέθοδο που ακολουθήθηκε για την εύρεση της συνάρτησης ανίχνευσης. Ο σκοπός της μετα-επεξεργασίας είναι να διευκολύνει τα προβλήματα της κατωφλίωσης και της επιλογής κορυφών αυξάνοντας την ομοιομορφία και τη συνεκτικότητα των χαρακτηριστικών της συνάρτησης ανίχνευσης που σχετίζονται με onsets, ιδανικά μετασχηματίζοντάς τα σε απομονωμένα, εύκολα ανιχνεύσιμα τοπικά μέγιστα. Μέθοδοι μετα-επεξεργασίας είναι το smoothing για την αφαίρεση του θορύβου, και διαδικασίες που χρειάζονται για την επιτυχή επιλογή των παραμέτρων κατωφλίωσης (κανονικοποίηση, αφαίρεση της DC συνιστώσας).

Κατωφλίωση: Για κάθε τύπο συνάρτησης ανίχνευσης και ακόμα και μετά τη μετα-επεξεργασία, θα υπάρχει ένας αριθμός κορυφών που δε θα σχετίζονται με τα onsets. Συνεπώς, είναι απαραίτητο να οριστεί ένα κατώφλι που θα χωρίζει τα γεγονότα σε αυτά που σχετίζονται με ένα onset και σε αυτά που δεν σχετίζονται. Το κατώφλι μπορεί να είναι σταθερό ή προσαρμοστικό.

Τελική επιλογή κορυφών: Μετά τη μετα-επεξεργασία και την κατωφλίωση της συνάρτησης ανίχνευσης, το μόνο που μένει για την επιλογή κορυφών είναι ο προσδιορισμός των τοπικών μεγίστων που υπερβαίνουν το καθορισμένο κατώφλι.

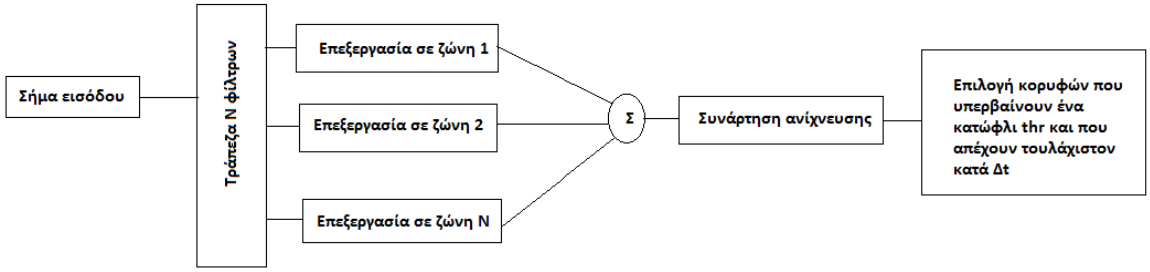
3.1.2 Σχόλια πάνω στις μεθόδους ανίχνευσης των onsets

Σε πολλές περιπτώσεις, οι αλγόριθμοι ανίχνευσης των onsets επιχειρούν να εντοπίσουν απότομες μεταβολές στην ενέργεια σε μια αναπαράσταση χρόνου-συχνότητας. Επίσης, κάποια συστήματα συνδυάζουν την πληροφορία της ενέργειας λαμβάνοντας υπόψιν και τη φάση, επιχειρώντας να εντοπίσουν απότομες μεταβολές τόσο στο πλάτος όσο και στη φάση ([74], [72], [7] και [69]). Με το συνδυασμό της πληροφορίας ενέργειας και φάσης λαμβάνεται υπόψιν τόσο ο κρουστικός όσο και ο αρμονικός χαρακτήρας του οργάνου. Η ανίχνευση των onsets μπορεί να βασιστεί και στο pitch, ειδικά για ήχους που δεν έχουν τόσο έντονο onset όσο αυτό των κρουστικών οργάνων (όργανα στα οποία συμπεριλαμβάνεται και το πιάνο). Τέλος, κάποιες άλλες προσεγγίσεις εστιάζουν στον απευθείας εντοπισμό των transients στο σήμα, υποθέτοντας ότι σχεδόν όλα τα transients που εμφανίζονται αντιστοιχούν σε onsets. Τέλος, υπάρχουν κάποιες άλλες μέθοδοι που δεν ακολουθούν το γενικό σχήμα που παρουσιάζεται στο 3.2. Αντί αυτού, ακολουθούν τεχνικές εκμάθησης μηχανής (επιβλεπόμενη και μη επιβλεπόμενη μάθηση) για να ταξινομήσουν κάθε πλαίσιο σε onset ή όχι onset.

Το να βρεθεί ένας αλγόριθμος ανίχνευσης των onsets γενικού σκοπού είναι πολύ δύσκολο. Τα διαφορετικά όργανα έχουν onsets με διαφορετικά χαρακτηριστικά, κι έτσι κανένας απλός αλγόριθμος δεν μπορεί να είναι βέλτιστος για μία ανίχνευση των onsets γενικού σκοπού.

3.1.3 Ο αλγόριθμος που εφαρμόζεται

Ο αλγόριθμος που υλοποιήθηκε ακολουθεί το διάγραμμα ροής που φαίνεται στο σχήμα 3.3. Αρχικά, το σήμα χωρίζεται σε ζώνες συχνοτήτων. Επεξεργαζόμαστε κάθε ζώνη συχνοτήτων για να βρεθεί η χρονική στιγμή και η ένταση του onset συστατικού της. Μετά, οι συνεισφορές από όλες τις ζώνες συχνοτήτων αθροίζονται. Το σύστημα βασίζεται στη γενική υπόθεση ότι η εμφάνιση ενός onset στο ηχητικό σήμα οδηγεί σε μια αλλαγή στο συχνοτικό περιεχόμενο του σήματος σε συγκεκριμένες συχνοτικές συνιστώσες.



Σχήμα 3.3: Διάγραμμα ροής του αλγορίθμου εύρεσης των onsets που εφαρμόζεται.

Συστοιχία φίλτρων

Χρησιμοποιείται μια συστοιχία φίλτρων που χωρίζει το σήμα σε μη επικαλυπτόμενες ζώνες όπως στο [70].

Ο Shreier ήταν ο πρώτος που τόνισε το γεγονός ότι ένας αλγόριθμος εύρεσης των onsets θα πρέπει να ακολουθεί το ανθρώπινο ηχητικό σύστημα, αντιμετωπίζοντας κάθε ζώνη συχνοτήτων ξεχωριστά και συνδυάζοντας στο τέλος τα αποτελέσματα [39].

Η συστοιχία φίλτρων αποτελείται από μια ομάδα φίλτρων που είναι περίπου κρίσιμης ζώνης (critical band) και καλύπτει τις συχνότητες από 44Hz έως 18kHz. Το πλάτος των φίλτρων κρίσιμης ζώνης μεταβάλλεται με τη συχνότητα. Τα χαμηλότερα 3 από τα απαιτούμενα φίλτρα είναι ζωνοπερατά φίλτρα μιας οκτάβας. Τα εναπομείναντα φίλτρα είναι ζωνοπερατά φίλτρα 1/3 οκτάβας. Συγκεκριμένα οι 21 συχνότητες αποκοπής των φίλτρων που χρησιμοποιήθηκαν είναι οι:

44 Hz, 88 Hz, 176 Hz, 352 Hz, 443.49 Hz, 558.77 Hz, 704 Hz, 886.98 Hz, 1117.53 Hz, 1408 Hz, 1773.97 Hz, 2235.06 Hz, 2816 Hz, 3547.94 Hz, 4470.12 Hz, 5632 Hz, 7095.88 Hz, 8940.24 Hz, 11264 Hz, 14191.75 Hz και 17880.49 Hz

Συνάρτηση ανίχνευσης

Στη συνέχεια, σε κάθε ζώνη συχνοτήτων χωριστά υπολογίζεται μια συνάρτηση ανίχνευσης. Η συνάρτηση ανίχνευσης είναι στο μιγαδικό πεδίο και συνδυάζει πληροφορία και για την ενέργεια

και για τη φάση του σήματος (βλέπε [7]).

Έστω ο μετασχηματισμός βραχέως χρόνου του σήματος, με τη χρήση ενός παραθύρου Hamming $w(m)$. Με τη βοήθεια του μετασχηματισμού υπολογίζεται μία αναπαράσταση χρόνου-συχνότητας. Εάν $X(n,k)$ αντιστοιχεί στην k-οστή συχνοτική συνιστώσα (frequency bin) του n-οστού πλαισίου, τότε:

$$X(n,k) = \sum_{m=-\frac{N}{2}}^{\frac{N}{2}-1} x(hn+m)w(m)e^{-\frac{2j\pi mk}{N}} \tag{3.1}$$

όπου N το μέγεθος του παραθύρου και h το hop size (αριθμός δειγμάτων μεταξύ δύο διαδοχικών παραθύρων fft).

Η φασματική ροή (spectral flux) είναι ένας τρόπος υπολογισμού της τοπικής φασματικής μεταβολής μεταξύ διαδοχικών πλαισίων. Δεδομένου ενός φασματογραφήματος η φασματική ροή είναι η απόσταση μεταξύ του φάσματος των διαδοχικών πλαισίων. Οι κορυφές στην καμπύλη της φασματικής ροής υποδεικνύουν τις χρονικές θέσεις σημαντικών αντιθέσεων στο φασματογράφημα. Με τη βοήθεια της φασματικής ροής ανιχνεύονται απότομες θετικές αλλαγές στην ενέργεια του σήματος που υποδεικνύουν τα attack μέρη νέων νοτών. Μετράει τη διαφορά στο πλάτος σε κάθε συχνοτική συνιστώσα, και αν περιοριστεί μόνο σε θετικές αλλαγές και αθροιστεί κατά μήκος όλων των συχνοτήτων, δίνει :

$$SF(n) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} H(|X(n,k)| - |X(n-1,k)|) \tag{3.2}$$

όπου $H(x) = \frac{x+|x|}{2}$ είναι η συνάρτηση ανόρθωσης μισού κύματος.

Η στιγμιαία συχνότητα μπορεί να υπολογιστεί από την πρώτη διαφορά στη φάση του $X(n,k)$. Έστω $\psi(n,k)$ η φάση του $X(n,k)$, δηλαδή $X(n,k) = |X(n,k)|e^{j\psi(n,k)}$ όπου $-\pi < \psi(n,k) \leq \pi$. Τότε η στιγμιαία συχνότητα δίνεται από την :

$$\psi'(n,k) = \psi(n,k) - \psi(n-1,k) \tag{3.3}$$

αντιστοιχισμένη στο διάστημα $(-\pi, \pi]$. Η αλλαγή στη στιγμιαία συχνότητα, που είναι μία ένδειξη ενός πιθανού onset, δίνεται από τη δεύτερη διαφορά της φάσης:

$$\psi''(n,k) = \psi'(n,k) - \psi'(n-1,k) \tag{3.4}$$

αντιστοιχισμένη επίσης στο διάστημα $(-\pi, \pi]$.

Το πλάτος και η φάση μπορούν να θεωρηθούν από κοινού στην αναζήτηση αναχωρήσεων από τη συμπεριφορά σταθερής κατάστασης υπολογίζοντας το αναμενόμενο πλάτος και φάση της τρέχουσας συχνοτικής συνιστώσας $X(n,k)$ βασιζόμενοι στις δύο προηγούμενες συχνοτικές συνιστώσες $X(n-1,k)$ και $X(n-2,k)$. Η αναμενόμενη τιμή $X_T(n,k)$ της τρέχουσας συνιστώσας εκτιμάται θεωρώντας σταθερό πλάτος και ρυθμό αλλαγής φάσης:

$$X_T(n,k) = |X(n-1,k)|e^{\psi(n-1,k)+\psi'(n-1,k)} \tag{3.5}$$

και στη συνέχεια παίρνουμε στο μιγαδικό πεδίο τη συνάρτηση του αθροίσματος των απολύτων αποκλίσεων από τις αναμενόμενες τιμές :

$$d(n) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |X(n,k) - X_T(n,k)| \tag{3.6}$$

Αυτή η ποσότητα υπολογίζεται σε κάθε ζώνη συχνοτήτων. Χρησιμοποιείται η νόρμα L1.

Κατωφλίωση και επιλογή κορυφών

Οι μικρές spurious κορυφές απομακρύνονται, χρησιμοποιώντας ένα φίλτρο ενδιάμεσης τιμής (median filter). Για παράθυρο μήκους l $\{w_i, i = -M...M\}$, που περιλαμβάνει τον ίδιο αριθμό προηγούμενων και επόμενων δειγμάτων, και θετική σταθερά C υπολογίζεται το:

$$\delta(n) = C * median|d(n-M)|, ..., |d(n+M)|, M = \frac{l-1}{2} \tag{3.7}$$

Το φιλτράρισμα ενδιάμεσης τιμής ακολουθείται από ανόρθωση μισού κύματος προκειμένου να επιλεχθούν οι κορυφές που υπερβαίνουν το δυναμικό κατώφλι που υπολογίστηκε με τη βοήθεια του φίλτρου ενδιάμεσης τιμής. Για να εξασφαλίσουμε ακριβή ανίχνευση, το μήκος του φίλτρου ενδιάμεσης τιμής πρέπει να είναι μεγαλύτερο από το μέσο πλάτος των κορυφών της συνάρτησης $d(n)$.

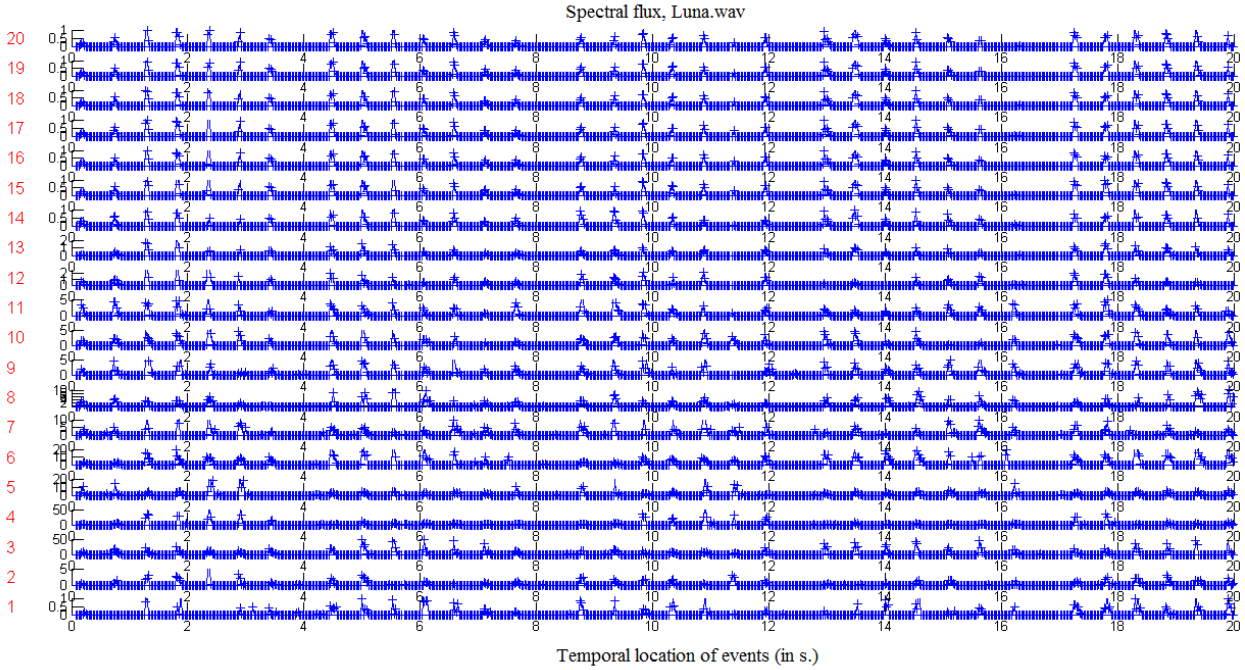
Τέλος το αποτέλεσμα από τις διαφορετικές ζώνες συχνοτήτων συνδυάζεται και δίνει την τελική συνάρτηση ανίχνευσης ενώ ακολουθεί το στάδιο επιλογής των κορυφών από τη συνάρτηση ανίχνευσης. Ένα τοπικό μέγιστο θα θεωρείται κορυφή, εάν η κανονικοποιημένη τιμή του ξεπερνάει ένα κατώφλι *thr*. Η κανονικοποιημένη τιμή κυμαίνεται μεταξύ 0 (το ελάχιστο) και 1 (το μέγιστο). Επίσης, έχει οριστεί μία ελάχιστη απόσταση Δt μεταξύ δύο διαδοχικών onset κι εάν έχουν βρεθεί δύο onsets που απέχουν λιγότερο από Δt τότε το μικρότερο από αυτά απορρίπτεται.

Στο Σχήμα 3.4 φαίνεται η φασματική ροή σε 20 ζώνες πριν συνδυαστούν οι ζώνες μεταξύ τους ενώ στα Σχήματα 3.5 και 3.6 φαίνονται οι τελικές συναρτήσεις ανίχνευσης για δύο διαφορετικά κομμάτια, με σημειωμένες επάνω τις κορυφές που έχουν επιλεγεί ως onsets.

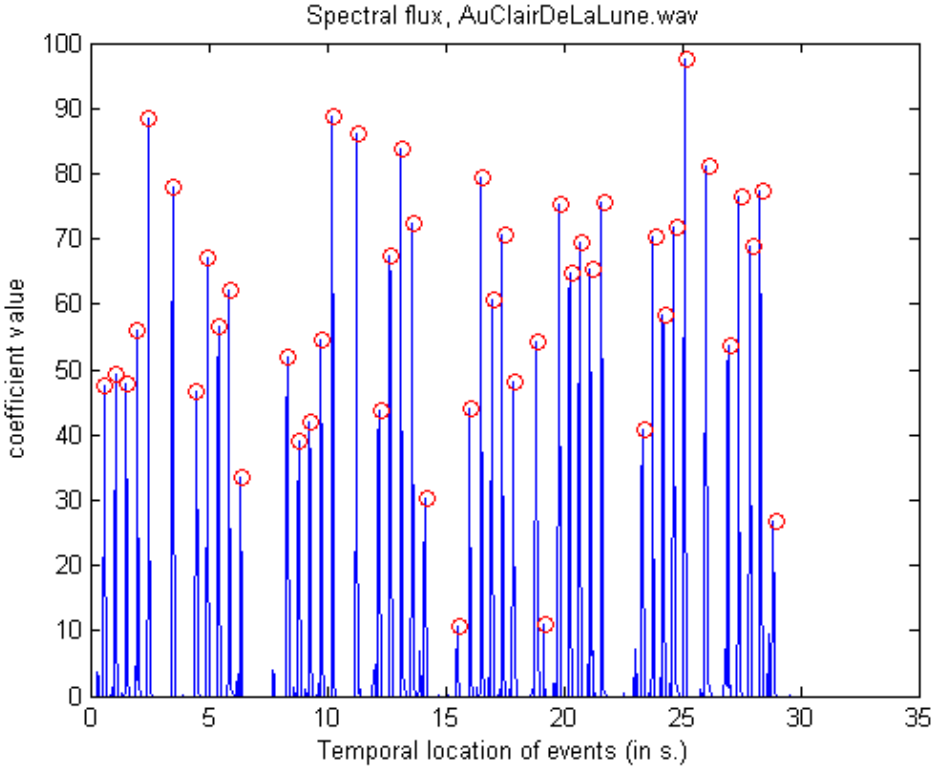
Αποτελέσματα εφαρμογής του αλγορίθμου

Για την αξιολόγηση της μεθόδου, δημιουργήθηκε και χρησιμοποιήθηκε μία μικρή βάση δεδομένων με πραγματικές μονοφωνικές ηχογραφήσεις πιάνου. Τα αποτελέσματα είναι πολύ ικανοποιητικά και φαίνονται στον πίνακα 3.1. Έχουν υπολογιστεί για μέγεθος παραθύρου $l = 0.2s$, θετική σταθερά $C = 1.9$, $\Delta t = 92.8ms$ και κατώφλι $thr = 0.1$. Διατηρώντας σταθερά στις παραπάνω τιμές τα l, C και Δt, μελετήθηκε πώς επιδρά στην απόδοση η τιμή του κατωφλίου (Σχήματα 3.7, 3.8). Για τιμή κατωφλίου ίση με 0.1 επιτυγχάνεται ένας καλός συμβιβασμός μεταξύ precision και recall.

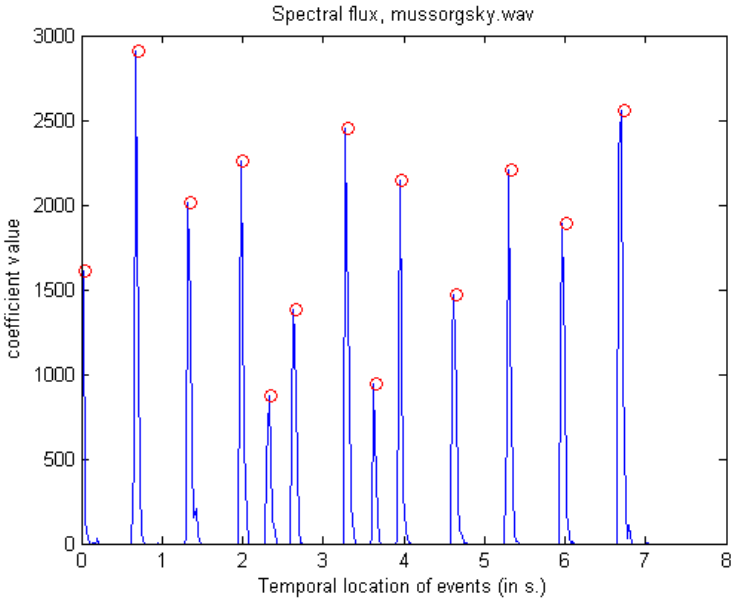
Τα αποτελέσματα αυτά σαφώς δείχνουν την αποτελεσματικότητα της μεθόδου που εφαρμόστηκε. Όμως, τα δεδομένα που χρησιμοποιήθηκαν αποτελούν σχετικά απλές περιπτώσεις χωρίς μεγάλη ποικιλία, οπότε δεν μπορούν να χρησιμοποιηθούν για μια πιο γενική αξιολόγηση. Παρόλα αυτά, τέτοια είναι τα δεδομένα που δίνονται ως είσοδο σε ένα σύστημα αυτόματης καταγραφής



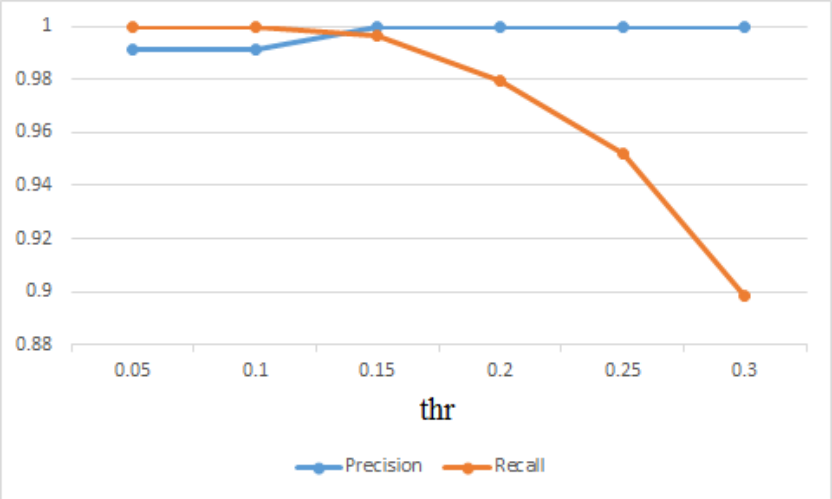
Σχήμα 3.4: Αποσύνθεση σε ζώνες συχνοτήτων, Φεγγαράκι μου λαμπρό



Σχήμα 3.5: Spectral flux, Au Clair De La Lune



Σχήμα 3.6: Spectral flux, Mussorgksy Promenade



Σχήμα 3.7: Οι καμπύλες του precision και του recall για τη μέθοδο ανίχνευσης των onsets με τη βοήθεια της φασματικής ροής ως συνάρτηση του κατωφλίου, διατηρώντας σταθερές τις υπόλοιπες παραμέτρους.

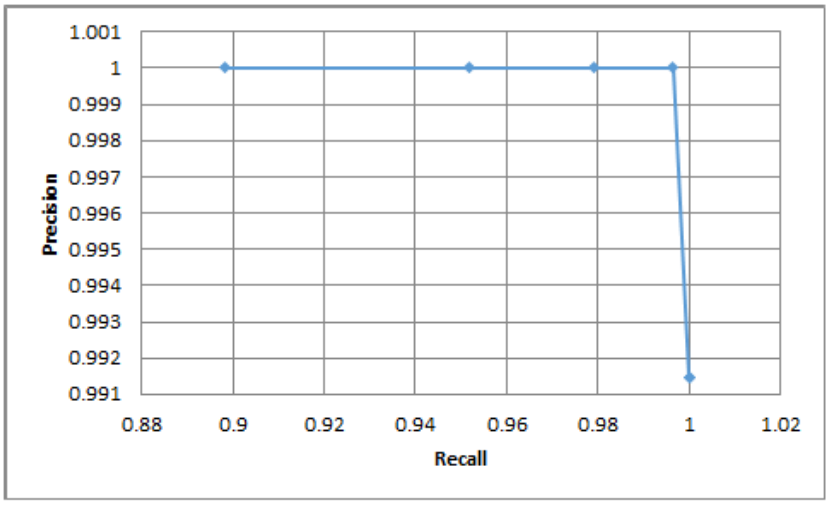
μονοφωνικής μουσικής παιγμένης στο πιάνο, οπότε θεωρήθηκε ότι στα πλαίσια της παρούσας διπλωματικής δεν υπήρχε λόγος να δοκιμαστεί η μέθοδος με διαφορετικά δεδομένα.

3.2 Εκτίμηση του τονικού ύψους (pitch) της νότας

Η εκτίμηση του τονικού ύψους έχει αποτελέσει αντικείμενο πολλών ερευνών κι έχουν προταθεί διάφορα συστήματα, τόσο στο πεδίο του χρόνου όσο και στο πεδίο της συχνότητας. Οι περισσότερες μέθοδοι αναπτύχθηκαν αρχικά για σήματα φωνής, αλλά αποδείχθηκε ότι μπορούν να

Κομμάτι	TP	FP	FN	Pr%	Re%	F-m%
Au Clair De La Lune	44	2	0	95.65	100	97.78
Fur Elise	51	0	0	100	100	100
Harry Potter theme	60	0	0	100	100	100
Long Long Ago	72	0	0	100	100	100
Mussorgksy Promenade	13	0	0	100	100	100
Chopin Valse	236	0	0	100	100	100
Τι Χαρά	49	2	0	96.08	100	98
Φεγγαράκι μου λαμπρό	56	1	0	98.25	100	99.12
Σύνολο	581	5	0	99.15	100	99.57

Πίνακας 3.1: Αποτελέσματα του onset detection με τη βοήθεια της φασματικής ροής. Ο πίνακας δείχνει τον αριθμό των onsets που ανιχνεύθηκαν σωστά (True Positives -TP), τα False Positives (FP), τα False Negatives (FN), το Precision (Pr), το Recall (Re) και το F-measure (F-m).



Σχήμα 3.8: Η καμπύλη Precision-Recall για τη μέθοδο ανίχνευσης των onsets με τη βοήθεια της φασματικής ροής ως συνάρτηση του κατωφλίου. Χωρίς να θυσιάσουμε precision, μπορούμε να πετύχουμε 99.66% recall. Για 100% recall, το precision πέφτει ελάχιστα στο 99.15%.

εφαρμοστούν με επιτυχία και σε αρμονικούς μουσικούς ήχους. Το τονικό ύψος είναι όπως είδαμε μία από τις τέσσερις διαστάσεις του ήχου. Σχετίζεται με ήχους περιοδικούς ή σχεδόν περιοδικούς και χαρακτηρίζεται από την περίοδό του, ή τη θεμελιώδη συχνότητα (F_0), αντίστροφη της περιόδου. Συνήθως χρησιμοποιούμε τον όρο pitch σαν συνώνυμο της θεμελιώδους συχνότητας, αλλά στην πραγματικότητα υπάρχει μία διάκριση μεταξύ τους: το pitch αναφέρεται στην *αντιλαμβανόμενη* θεμελιώδη συχνότητα. Το αυτί έχει την ικανότητα να κβαντίσει με μεγάλη ακρίβεια αυτό το μέγεθος, του οποίου η αριθμητική τιμή εκφράζεται σε hertz που είναι μια φυσική κλίμακα, ή σε mel που είναι μια κλίμακα υποκειμενική η οποία μετράει το πώς αντιλαμβανόμαστε τον ήχο. Προφανώς η θεμελιώδης συχνότητα είναι αυτή που μας επιτρέπει να οργανώσουμε τους ήχους σε κλίμακα οπότε η εκτίμηση της θεμελιώδους συχνότητας και η αντιστοίχιση της σε μία μουσική νότα είναι η πιο σημαντική απαίτηση που έχουμε από ένα σύστημα αυτόματης καταγραφής μουσικής.

3.2.1 Μέθοδοι εκτίμησης

Η εκτίμηση του μονοφωνικού τονικού ύψους συνίσταται στην εκτίμηση της θεμελιώδους συχνότητας ενός ήχου περιοδικού ή σχεδόν περιοδικού. Η μεγάλη πλειοψηφία των τεχνικών εκτίμησης μπορούν να χωριστούν σε δύο κατηγορίες: τις προσεγγίσεις στο πεδίο της συχνότητας και τις προσεγγίσεις στο πεδίο του χρόνου. Στο πεδίο της συχνότητας, ο ήχος παρουσιάζει διακριτές κορυφές που ακολουθούν μία κατανομή σχεδόν αρμονική: αρκεί λοιπόν να υπολογιστούν αυτές οι κορυφές για να πάρουμε τη θεμελιώδη συχνότητα. Στο πεδίο του χρόνου, υπολογίζεται συνήθως η θεμελιώδης περίοδος, ισοδύναμο μέγεθος της θεμελιώδους συχνότητας, ψάχνοντας την πιο μικρή μη μηδενική χρονική μετατόπιση για την οποία το αρχικό σήμα και η μετατοπισμένη εκδοχή του ταυτίζονται. Πάντως η αυτόματη καταγραφή μονοφωνικών σημάτων, είναι πρακτικά ένα λυμένο πρόβλημα εφόσον αρκετοί αλγόριθμοι έχουν προταθεί που είναι αξιόπιστοι, εμπορικά εφαρμόσιμοι και λειτουργούν σε πραγματικό χρόνο.

Στο πεδίο της **συχνότητας** έχουν χρησιμοποιηθεί ιδιαιτέρως δύο συναρτήσεις ανίχνευσης του pitch:

- το φασματικό άθροισμα $S(f)$, που ορίζεται ως

$$S(f) = 20 \log \sum_{h=1}^{H} |X(hf)| \tag{3.8}$$

- το φασματικό γινόμενο $P(f)$, που ορίζεται ως

$$P(f) = 20 \log \prod_{h=1}^{H} |X(hf)| \tag{3.9}$$

όπου f η θεμελιώδης συχνότητα, X ο διακριτός μετασχηματισμός Fourier του σήματος και θεωρούμε τις H πρώτες αρμονικές.

Πολλές μέθοδοι στο πεδίο της συχνότητας στηρίζονται σε αυτές τις συναρτήσεις καθώς και σε παραλλαγές τους. Σε κάποιες εργασίες όπως στην [76] και στην [77] το πρόβλημα προσεγγίζεται συγκρίνοντας το μετρούμενο φάσμα με ένα δεδομένο φασματικό πρότυπο. Το φασματικό πρότυπο μπορεί να είναι είτε ένα συγκεκριμένο φασματικό μοντέλο ή μία ακολουθία από ισαπέχουσες φασματικές συνιστώσες, κάτι που αποκαλείται συχνά φασματική χτένα. Σε αυτήν την περίπτωση επιχειρείται να βρεθεί μία θεμελιώδης συχνότητα της οποίας οι αρμονικές ερμηνεύουν βέλτιστα το παρατηρούμενο φάσμα. Αρκετές από αυτές τις μεθόδους στηρίζονται εν μέρει και στην πρότερη γνώση της πηγής του ήχου. Παίρνοντας τον αντίστροφο μετασχηματισμό Fourier του λογαρίθμου του μέτρου του μετασχηματισμού Fourier ενός σήματος λαμβάνουμε το cepstrum του. Μία πλήρης μεθοδολογία για τη χρήση του cepstrum στην εκτίμηση της θεμελιώδους συχνότητας σημάτων φωνής δίνεται στο [78]. Η μέγιστη τιμή του cepstrum είναι αυτή που καθορίζει το pitch του σήματος.

Στο πεδίο του **χρόνου**, η αυτοσυσχέτιση του σήματος είναι η βασική μέθοδος για την εκτίμηση του pitch, αφού μπορεί να ανιχνεύσει κρυμμένες περιοδικότητες σε ένα σήμα. Πολλές μέθοδοι στο πεδίο του χρόνου στηρίζονται στην αυτοσυσχέτιση και σε παραλλαγές της. Οι de Cheveigné και Kawahara δείχνουν τη σχέση που συνδέει την αυτοσυσχέτιση με την AMDF (Average Magnitude Difference Function) και αναπτύσσουν τη μέθοδο εκτίμησης του τονικού

ύψους YIN [79] και ο Klapuri ([38], [80]) στηρίζεται στην αρχή της αυτοσυσχέτισης αλλάζοντας τελείως το βήμα της αναστροφής του μετασχηματισμού Fourier. Κάποιοι αλγόριθμοι εύρεσης του pitch βασίζονται και σε μετρήσεις κορυφών και κοιλάδων στον τομέα του χρόνου, ή σε μετρήσεις των zero-crossings, προσπαθώντας έπειτα να βρουν ποιες μετρήσεις είναι αυτές που αντιστοιχούν στην πραγματική περίοδο του pitch.

Τέλος, η μέτρηση της περιοδικότητας μπορεί να γίνει χρησιμοποιώντας από κοινού τεχνικές στο πεδίο του χρόνου και στο πεδίο της συχνότητας, όπως στο [81].

3.2.2 Ο αλγόριθμος που εφαρμόζεται

Εξ' ορισμού, η θέση του μεγίστου της συνάρτησης αυτοσυσχέτισης ενός ήχου δίνει μια καλή εκτίμηση της περιόδου του pitch του ήχου. Όμως, η δειγματοληψία και η παραθύρωση προκαλούν προβλήματα στον ακριβή προσδιορισμό της θέσης και του ύψους του μεγίστου. Για την εύρεση του pitch κάθε νότας, χρησιμοποιείται ο αλγόριθμος που περιγράφεται στο [8], ο οποίος αντιμετωπίζει αυτά τα προβλήματα.

Για ένα σήμα $x(t)$ που είναι στατικό, η αυτοσυσχέτιση $r_x(\tau)$ σαν συνάρτηση της καθυστέρησης (lag) τ ορίζεται ως :

$$r_x \equiv \int x(t)x(t+\tau)\mathrm{d}t \tag{3.10}$$

Αυτή η συνάρτηση έχει ένα ολικό μέγιστο για $\tau = 0$. Αν υπάρχουν και άλλα ολικά μέγιστα εκτός του 0, το σήμα καλείται περιοδικό και υπάρχει μία καθυστέρηση T_0, που αποκαλείται περίοδος, έτσι ώστε όλα αυτά τα μέγιστα να είναι τοποθετημένα σε καθυστερήσεις nT_0 για κάθε ακέραιο n, με $r_x(nT_0) = r_x(0)$. Η θεμελιώδης περίοδος F_0 αυτού του περιοδικού σήματος ορίζεται ως $F_0 = \frac{1}{T_0}$.

Για ένα μη στατικό σήμα, η αυτοσυσχέτιση βραχέως χρόνου για μια χρονική στιγμή t εκτιμάται από ένα σύντομο παραθυρομένο τμήμα του σήματος κεντραρισμένο γύρω από το t. Έτσι εκτιμάται η τοπική θεμελιώδης συχνότητα $F_0(t)$.
Η διαδικασία εκτίμησης της τοπικής θεμελιωδούς συχνότητας περιγράφεται στη συνέχεια.

Παίρνουμε από το σήμα $x(t)$ ένα κομμάτι διάρκειας , κεντραρισμένο γύρω από το t_{mid}. Αφαιρούμε από αυτό το κομμάτι τη μέση τιμή του μ_x και πολλαπλασιάζουμε το αποτέλεσμα με ένα παράθυρο $w(t)$, ώστε να πάρουμε το παραθυρομένο σήμα:

$$\alpha_t = (x(t_{mid} - \frac{1}{2}T + t) - \mu_x)w(t) \tag{3.11}$$

Η συνάρτηση παραθύρου $w(t)$ είναι συμμετρική γύρω από το $t = \frac{1}{2}T$ και μηδέν οπουδήποτε αλλού εκτός του διαστήματος $[0, T]$. Έχει επιλεχθεί ένα παράθυρο Hanning, που δίνεται από την:

$$w(t) = \frac{1}{2} - \frac{1}{2}\cos\frac{2\pi t}{T} \tag{3.12}$$

Η κανονικοποιημένη αυτοσυσχέτιση $r_\alpha(\tau)$ του παραθυρομένου σήματος είναι μια συμμετρική συνάρτηση της καθυστέρησης τ:

$$r_\alpha(\tau) = r_a(-\tau) = \frac{\int_0^{T-\tau} \alpha(t)\alpha(t+\tau)\,\mathrm{d}x}{\int_0^T \alpha^2(t)\,\mathrm{d}x} \tag{3.13}$$

Η κανονικοποιημένη αυτοσυσχέτιση $r_w(\tau)$ του παραθύρου υπολογίζεται με αντίστοιχο τρόπο με αυτό της εξίσωσης 3.13. Η κανονικοποιημένη αυτοσυσχέτιση ενός παραθύρου Hanning είναι:

$$r_w(\tau) = (1 - \frac{|\tau|}{T})(\frac{2}{3} + \frac{1}{3}\cos\frac{2\pi\tau}{T}) + \frac{1}{2\pi}\sin\frac{2\pi|\tau|}{T} \tag{3.14}$$

Για να εκτιμηθεί η αυτοσυσχέτιση $r_x(\tau)$ του αρχικού τμήματος του σήματος, η αυτοσυσχέτιση $r_\alpha(\tau)$ του παραθυρομένου σήματος διαιρείται με την αυτοσυσχέτιση $r_w(\tau)$ του παραθύρου:

$$r_x(\tau) \approx \frac{r_\alpha(\tau)}{r_w(\tau)} \tag{3.15}$$

Η εκτίμηση αυτή είναι ακριβής για το σταθερό σήμα $x(t) = 1$ (χωρίς φυσικά την αφαίρεση της μέσης τιμής). Για περιοδικά σήματα φέρνει τις κορυφές της συνάρτησης αυτοσυσχέτισης πολύ κοντά στο 1. Η ακρίβεια του αλγορίθμου εξαρτάται από την αξιοπιστία της εκτίμησης (3.15), η οποία εξαρτάται απευθείας από το σχήμα του παραθύρου. Το παράθυρο Hanning δίνει καλύτερα αποτελέσματα συγκριτικά με άλλα παράθυρα.

Στην υλοποίηση του αλγορίθμου, οι αυτοσυσχετίσεις του παραθυρομένου σήματος και του παραθύρου υπολογίζονται αριθμητικά μέσω FFT. Αυτό είναι δυνατό επειδή η αυτοσυσχέτιση μπορεί να βρεθεί υπολογίζοντας το μετασχηματισμό Fourier του παραθυρομένου σήματος που δίνει στο πεδίο της συχνότητας:

$$\tilde{\alpha}(\omega) = \int \alpha(t)e^{-i\omega t}\mathrm{d}t \tag{3.16}$$

και στη συνέχεια υπολογίζοντας τον αντίστροφο μετασχηματισμό Fourier της πυκνότητας ισχύος $|\tilde{\alpha}(\omega)|^2$ έχουμε:

$$r_\alpha(\tau) = \int |\tilde{\alpha}(\omega)|^2 e^{i\omega t}\frac{\mathrm{d}\omega}{2\pi} \tag{3.17}$$

Θεωρώντας ένα σήμα συνεχούς χρόνου $x(t)$ που δεν περιέχει συχνότητες μεγαλύτερες μιας συγκεκριμένης συχνότητας f_{max}, μπορούμε να πάρουμε δείγματα αυτού του σήματος σε τακτά χρονικά διαστήματα $\Delta t \leq \frac{1}{2f_{max}}$ ώστε να γνωρίζουμε τις τιμές x_n σε ισαπέχουσες χρονικές στιγμές t_n :

$$x_n = x(t_n); t_n = t_0 + n\Delta t \tag{3.18}$$

Δεν χάνουμε δεδομένα με αυτή τη δειγματοληψία επειδή μπορούμε να ανακατασκευάσουμε το αρχικό σήμα ως :

$$x(t) = \sum_{n=-\infty}^{+\infty} x_n \frac{\frac{\sin\pi(t-t_n)}{\Delta t}}{\frac{\pi(t-t_{n)}}{\Delta t}} \tag{3.19}$$

Η αυτοσυσχέτιση που υπολογίζεται από το δειγματοληπτημένο σήμα είναι και αυτή μία δειγματοληπτημένη συνάρτηση:

$$r_n = r(n\Delta\tau)$$

Υπάρχει ένα τοπικό μέγιστο στην αυτοσυσχέτιση μεταξύ (m-1)Δτ και (m+1)Δτ εάν :

$$r_m > r_{m-1} \text{ και } r_m > r_{m+1} \tag{3.20}$$

Αντίστοιχα με την (3.19) μπορούμε με παρεμβολή σε ένα πεπερασμένο αριθμό δειγμάτων N στα αριστερά και στα δεξιά, χρησιμοποιώντας και πάλι ένα παράθυρο Hanning για να έχουμε μηδενική παρεμβολή στα άκρα να πάρουμε:

$$\begin{aligned} r(\tau) \approx & \sum_{n=1}^{N} r_{n_r-n} \frac{\sin \pi(\varphi_l + n - 1)}{\pi(\varphi_l + n - 1} (\frac{1}{2} + \frac{1}{2} \cos \frac{\pi(\varphi_l + n - 1)}{\varphi_l + N}) + \\ & \sum_{n=1}^{N} r_{n_l+n} \frac{\sin \pi(\varphi_r + n - 1)}{\pi(\varphi_r + n - 1)} (\frac{1}{2} + \frac{1}{2} \cos \frac{\pi(\varphi_r + n - 1)}{\varphi_r + N}) \end{aligned} \tag{3.21}$$

όπου $n_l \equiv$ ο μεγαλύτερος ακέραιος $\leq \frac{\tau}{\Delta\tau}; n_r \equiv n_l + 1; \phi_l \equiv \frac{\tau}{\Delta\tau} - n_l; \phi_r \equiv 1 - \phi_l$

Στην υλοποίηση, το N είναι το μικρότερο μεταξύ των 500 και του μεγαλύτερου αριθμού για τον οποίο το $(n_l + N)\Delta\tau$ είναι μικρότερο από το μισό του μήκους του παραθύρου. Αυτό επειδή η εκτίμηση της αυτοσυσχέτισης δεν είναι αξιόπιστη για καθυστερήσεις μεγαλύτερες από το μισό μήκος παραθύρου, αν υπάρχουν μόνο λίγες περίοδοι σε κάθε παράθυρο. Οι θέσεις και τα ύψη των μεγίστων της εξίσωσης 3.21 μπορούν να καθοριστούν με μεγάλη ακρίβεια (αναζητούνται μεταξύ $(m-1)\Delta\tau$ και $(m+1)\Delta\tau$).

Συνοψίζοντας τα παραπάνω, ο αλγόριθμος έχει ως εξής:

Βήμα 1. Προεπεξεργασία: για να αφαιρεθεί ο πλευρικός λοβός του μετασχηματισμού Fourier του παραθύρου Hanning για τα συστατικά του σήματος που βρίσκονται κοντά στη συχνότητα του Nyquist, εκτελείται μία ελαφριά υπερδειγματοληψία ως εξής: κάνε ένα FFT σε όλο το σήμα, φίλτραρε με πολλαπλασιασμό στο πεδίο της συχνότητας γραμμικά προς το μηδέν από το 95% της συχνότητας του Nyquist στο 100% της συχνότητας του Nyquist, κάνε έναν αντίστροφο FFT μιας τάξης μεγαλύτερο από τον πρώτο FFT.

Βήμα 2. Υπολόγισε την τιμή της ολικής απόλυτης κορυφής του σήματος (βλέπε βήμα 3.3).

Βήμα 3. Επειδή η μέθοδος είναι μέθοδος ανάλυσης βραχέως χρόνου, η ανάλυση εκτελείται για ένα μικρό αριθμό τμημάτων (*frames*) που εξάγονται από το σήμα σε βήματα που δίνονται από την παράμετρο *TimeStep*. Σε κάθε frame, αναζητούμε το πολύ *MaximumNumberOfCanditatesPerFrame* ζευγάρια lag-ύψους που είναι καλοί υποψήφιοι για την περιοδικότητα σε αυτό το frame. Αυτός ο αριθμός συμπεριλαμβάνει τον *unvoiced* υποψήφιο, που είναι πάντα παρών. Τα επόμενα βήματα γίνονται για κάθε frame:

Βήμα 3.1 Πάρε ένα τμήμα του σήματος. Το μήκος αυτού του τμήματος (το μήκος του παραθύρου) καθορίζεται από την παράμετρο *MinimumPitch*, που αντιστοιχεί στην χαμηλότερη θεμελιώδη συχνότητα που θέλουμε να ανιχνευτεί. Το παράθυρο θα πρέπει να είναι τόσο μεγάλο ώστε να περιλαμβάνει τρεις περιόδους. Π.χ. εάν το *MinimumPitch* είναι 75 Hz, το μήκος του παραθύρου είναι 40ms.

Βήμα 3.2 Αφαίρεσε την τοπική μέση τιμή.

Βήμα 3.3 Ο πρώτος υποψήφιος είναι ο άφωνος (unvoiced) υποψήφιος, που είναι πάντα παρών. Η ισχύς αυτού του υποψηφίου υπολογίζεται με βάση δύο χαλαρές παραμέτρους κατωφλίωσης, τις *SilenceThreshold* και *VoicingThreshold*. Π.χ. αν το *VoicingThreshold* είναι 0.4 και το *SilenceThreshold* είναι 0.05, το πλαίσιο έχει πολλές πιθανότητες να χαρακτηριστει ως άφωνο (βήμα 4) εάν δεν υπάρχουν κορυφές της συνάρτησης αυτοσυσχέτισης πάνω από 0.4 περίπου ή εάν η τοπική απόλυτη κορυφή είναι μικρότερη από περίπου 0.05 φορές την ολική απόλυτη κορυφή, που υπολογίστηκε στο βήμα 2.

Βήμα 3.4 Πολλαπλασίασε με τη συνάρτηση παραθύρου (εξίσωση 3.11).

Βήμα 3.5 Πρόσθεσε μηδενικά για μήκος όσο το μισό του παραθύρου (επειδή χρειαζόμαστε τιμές της αυτοσυσχέτισης μέχρι και το μισό μήκος παραθύρου για την παρεμβολή).

Βήμα 3.6 Πρόσθεσε μηδενικά μέχρι ο αριθμός δειγμάτων να είναι δύναμη του 2.

Βήμα 3.7 Κάνε ένα FFT (διακριτή εκδοχή της εξίσωσης 3.16).

Βήμα 3.8 Τετραγώνισε τα δείγματα στο πεδίο της συχνότητας.

Βήμα 3.9 Κάνε ένα FFT (διακριτή εκδοχή της εξίσωσης 3.17). Αυτό δίνει μια δειγματοληπτημένη εκδοχή της $r_\alpha(\tau)$.

Βήμα 3.10 Διαίρεσε με την αυτοσυσχέτιση του παραθύρου, που υπολογίστηκε μια φορά με τα βήματα 3.5 έως 3.9 (εξίσωση 3.15). Αυτό δίνει μια δειγματοληπτημένη εκδοχή του $r_x(\tau)$.

Βήμα 3.11 Βρες τις θέσεις και τα ύψη των μεγίστων της συνεχούς εκδοχής του $r_x(\tau)$, που δίνεται από την εξίσωση (3.21). Οι μόνες θέσεις που εξετάζονται για τα μέγιστα είναι αυτές που παράγουν ένα pitch μεταξύ *MinimumPitch* και *MaximumPitch*. Η παράμετρος *MaximumPitch* πρέπει να είναι μεταξύ του *MinimumPitch* και της συχνότητας του Nyquist. Οι μόνοι υποψήφιοι που δεν απορρίπτονται είναι ο άφωνος υποψήφιος, που έχει τοπική ισχύ ίση με:

$$R \equiv VoicingThreshold + max(0, 2 - \frac{(localabsolutepeak)/(globalabsolutepeak)}{SilenceThreshold/(1 + VoicingThreshold)}) \quad (3.22)$$

και οι έμφωνοι υποψήφιοι με τις υψηλότερες (*MaximumNumberOfCandidatesPerFrame* μείον 1) τιμές τοπικής ισχύος:

$$R \equiv r(\tau_{max}) - OctaveCost \log_2(MinimumPitch\tau_{max}) \quad (3.23)$$

Η παράμετρος *OctaveCost* ευνοεί υψηλότερες θεμελιώδεις συχνότητες. Ένας από τους λόγους ύπαρξης αυτής της παραμέτρου είναι ότι για τέλεια περιοδικά σήματα όλες οι κορυφές είναι εξίσου υψηλές και θα πρέπει να διαλέξουμε αυτή με τη μικρότερη καθυστέρηση. Άλλοι λόγοι για αυτή την παράμετρο είναι ανεπιθύμητες πτώσεις οκτάβας που οφείλονται σε πρόσθετο θόρυβο. Τέλος, μια σημαντική χρήση αυτής της παραμέτρου, έγκειται στην διαφορά μεταξύ της ακουστικής θεμελιώδους συχνότητας και του αντιλαμβανομένου pitch.

Αφού εκτελέσουμε το βήμα 3 για κάθε πλαίσιο, μένουμε με ένα αριθμό ζευγαριών συχνότητας-ισχύος F_{ni}, R_{ni}, όπου ο δείκτης n παίρνει τιμές από 1 έως τον αριθμό των πλαισίων, και το i είναι μεταξύ 1 και τον αριθμό των υποψηφίων σε κάθε πλαίσιο. Ο τοπικά καλύτερος υποψήφιος σε κάθε πλαίσιο είναι αυτός με το μεγαλύτερο R. Όμως καθώς μπορούμε να έχουμε αρκετούς σχεδόν ισοδύναμους υποψήφιους σε οποιοδήποτε πλαίσιο, μπορούμε να τρέξουμε σε αυτά τα ζεύγη έναν αλγόριθμο εύρεσης του ολικού μονοπατιού, ο σκοπός του οποίου είναι να ελαχιστοποιήσει τον αριθμό των τυχαίων αποφάσεων μεταξύ έμφωνου-άφωνου και των μεγάλων αλμάτων στη συχνότητα:

Βήμα 4. Για κάθε πλαίσιο n, p_n είναι ένας αριθμός μεταξύ 1 και του αριθμού των υποψηφίων για αυτό το πλαίσιο. Οι τιμές $\{p_n | 1 \leq n \leq \text{αριθμό πλαισίων }\}$ ορίζουν ένα μονοπάτι δια μέσου των υποψηφίων : $\{(F_{np_n}, R_{np_n}) | 1 \leq n \leq \text{αριθμό πλαισίων }\}$. Σε κάθε δυνατό μονοπάτι αναθέτουμε ένα κόστος:

$$cost(\{p_n\}) = \sum_{n=2}^{NoFrames} transitionCost(F_{n-1,p_{n-1}}, F_{np_n}) - \sum_{n=1}^{NoFrames} R_{np_n} \tag{3.24}$$

όπου το transitionCost ορίζεται ως ($F = 0$ σημαίνει άφωνο)

$$transitionCost(F_1, F_2) = \begin{cases} 0 & \text{Εάν } F_1 = 0 \text{ και } F_2 = 0 \\ VoicedUnvoicedCost & \text{Εάν } F_1 = 0 \text{ xor } F_2 = 0 \\ OctaveJumpCost * |\log_2 \frac{F_1}{F_2}| & \text{Εάν } F_1 \neq 0 \text{ και } F_2 \neq 0 \end{cases} \tag{3.25}$$

Το ολικά βέλτιστο μονοπάτι είναι αυτό με το μικρότερο κόστος. Αυτό το μονοπάτι μπορεί να συμπεριλαμβάνει κάποιους υποψηφίους που είναι τοπικά δεύτερης επιλογής. Το φθηνότερο μονοπάτι μπορεί να βρεθεί με τη βοήθεια του δυναμικού προγραμματισμού.

Ο αλγόριθμος εύρεσης του ολικού μονοπατιού μπορεί να αντιμετωπίσει κάποια τοπικά λάθη οκτάβας. Αν θέλουμε να επιλέγεται ο τοπικά καλύτερος υποψήφιος σε κάθε πλαίσιο, τότε οι παράμετροι VoicedUnvoicedCost και OctaveJumpCost πρέπει να τεθούν ίσες με μηδέν.

Η εκτίμηση του pitch πραγματοποιείται με βραχυπρόθεσμη (short-term) ανάλυση η οποία πραγματοποιείται για καθένα από τα αποσπάσματα (frames) τα οποία ανακτώνται από το σήμα με βήμα ίσο με την παράμετρο *TimeStep*. Για τις απαιτήσεις της παρούσας εργασίας, το βήμα δειγματοληψίας, ή διαφορετικά το εύρος του παραθύρου που μελετάται είναι ίσο με την τιμή 0.01sec. Το αποτέλεσμα του αλγορίθμου που περιγράφτηκε είναι η τιμή του pitch σε Hz σε κάθε παράθυρο. Αυτή μετατρέπεται εύκολα σε αριθμό midi, στρογγυλοποιώντας στην πλησιέστερη μουσική συχνότητα σύμφωνα με τη σχέση (2.3). Εάν ως είσοδος στον αλγόριθμο έχει δοθεί ένα σήμα που αντιστοιχεί σε μία μόνο νότα τότε ενώ το pitch σε Hz δεν έχει την ίδια τιμή σε όλα τα παράθυρα, ο αριθμός midi διατηρεί την ίδια τιμή στο steady μέρος της νότας. Έτσι έχοντας ως στόχο την επιλογή ενός και μόνο midi αριθμού, υπολογίζεται ο midi αριθμός που αντιστοιχεί σε κάθε παράθυρο και έπειτα επιλέγεται η τιμή του midi που εμφανίζεται πιο συχνά.

Αποτέλεσμα εφαρμογής του αλγορίθμου

Γενικά ο αλγόριθμος αποδίδει καλά, εμφανίζοντας πολύ μεγάλα ποσοστά επιτυχίας αναγνώρισης. Χρησιμοποιώντας σαν δεδομένα 205 δείγματα μεμονωμένων νοτών από το σύνολο δειγμάτων

ISOL της βάσης MAPS [5] για τα όργανα και τις συνθήκες ηχογράφησης που αντιστοιχούν στους κωδικούς ENSTDkCl και SptkBGCl με τις θεμελιώδεις συχνότητες που περιέχονται στα δείγματα να εκτείνονται από το Do 0 στα 33 Hz μέχρι και το Si 5 στα 1865 Hz είχαμε ρυθμό λαθών μόλις 3.41%.

Το πιο συνηθισμένο λάθος είναι η εκτίμηση του pitch να είναι κατά μία οκτάβα χαμηλότερη από την πραγματική, δηλαδή η εκτιμώμενη αριθμητική τιμή της θεμελιώδους συχνότητας να είναι υποδιπλάσια της πραγματικής (octave errors). Αυτό είναι ένα πρόβλημα από το οποίο υποφέρουν όλες οι μέθοδοι ανίχνευσης του pitch που στηρίζονται στη συνάρτηση αυτοσυσχέτισης, εφόσον στη συνάρτηση εμφανίζονται κι άλλες κορυφές εκτός από αυτή που αντιστοιχεί στη θεμελιώδη συχνότητα. Νότες που διαφέρουν κατά μία οκτάβα δίνουν την αίσθηση του ίδιου τόνου με μια μεγαλύτερη ή μικρότερη οξύτητα και οι συχνότητές τους είναι η μια διπλάσια της άλλης. Η παράμετρος OctaveCost έχει οριστεί ίση με 0.06 αντί της default τιμής της που ήταν 0.01, προκειμένου να μειωθεί αυτό το φαινόμενο και τα αποτελέσματα να είναι καλύτερα. Ένα άλλο λάθος που εμφανίζεται επίσης, είναι η εσφαλμένη εκτίμηση κατά ένα ημιτόνιο, κυρίως στις υψηλότερες συχνότητες. Αυτό το σφάλμα εν μέρει μπορεί να οφείλεται και στον τρόπο με τον οποίο παίρνουμε τον midi αριθμό (σχέση μετατροπής 2.3 από Hz σε αριθμό midi και επιλογή του αριθμού που εμφανίζεται πιο συχνά μεταξύ των διαφορετικών παραθύρων).

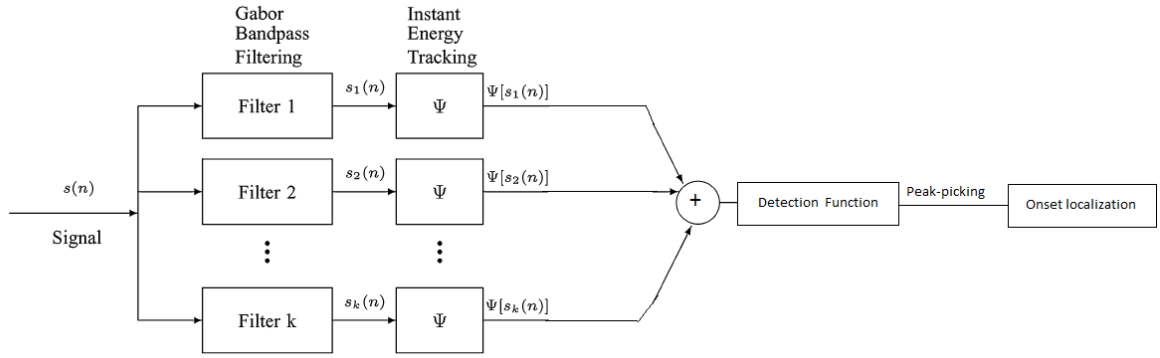
3.3 Ανίχνευση της αρχής (onset) της νότας σε μονοφωνική μουσική για πιάνο με τη χρήση του Teager τελεστή ενέργειας

3.3.1 Περιγραφή αλγορίθμου

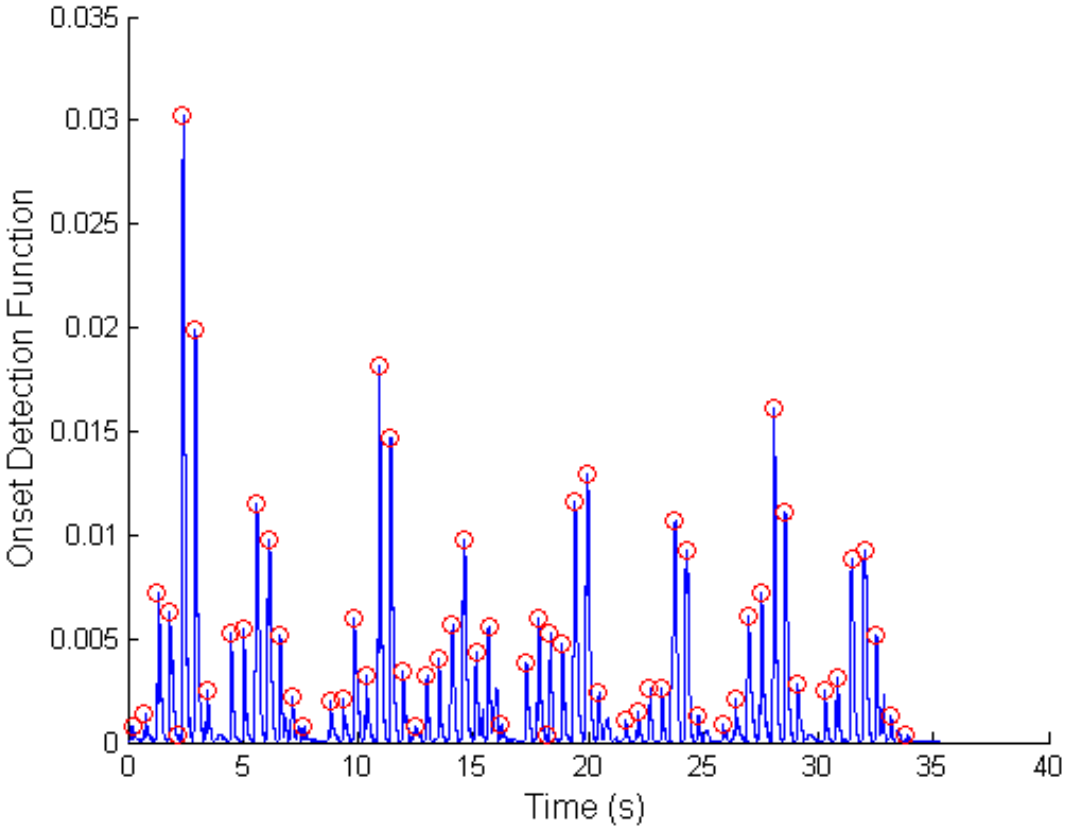
Στα πλαίσια αυτής της διπλωματικής εργασίας αναπτύχθηκε μία πρωτότυπη μέθοδος ανίχνευσης της αρχής (onset) της νότας σε μονοφωνική μουσική για πιάνο η οποία βασίζεται στη χρήση του Teager τελεστή ενέργειας (2.2.3). Το μπλοκ διάγραμμα της εικόνας 3.9 δείχνει πώς ανιχνεύονται τα onsets με τη βοήθεια του Teager τελεστή. Το σήμα φιλτράρεται από μία συστοιχία 88 ζωνοπερατών Gabor φίλτρων οι κεντρικές συχνότητες των οποίων αντιστοιχούν στις 88 διαφορετικές συχνότητες των πλήκτρων ενός πιάνου. Έπειτα ο διακριτός ενεργειακός τελεστής Teager-Kaiser εφαρμόζεται στην έξοδο κάθε φίλτρου και τα αποτελέσματα όλων των τελεστών από όλες τις ζώνες αθροίζονται. Τέλος, βρίσκονται οι κορυφές της συνάρτησης που προκύπτει και επιλέγονται κατάλληλα οι κορυφές που αντιστοιχούν σε onsets. Κάποια αποτελέσματα της εφαρμογής της μεθόδου φαίνονται στα σχήματα 3.10 και 3.11.

3.3.2 Αποτελέσματα εφαρμογής του αλγορίθμου

Για την αξιολόγηση της μεθόδου, χρησιμοποιήθηκαν πραγματικές ηχογραφήσεις πιάνου, με μία μόνο φωνή. Τα δεδομένα είναι τα ίδια που χρησιμοποιήθηκαν και για την αξιολόγηση της μεθόδους εύρεσης των onsets με χρήση της φασματικής ροής. Τα αποτελέσματα φαίνονται στον πίνακα 3.2. Κρατάμε μόνο εκείνες τις κορυφές της συνάρτησης ανίχνευσης που αντιπροσωπεύουν απότομες αλλαγές. Το κατώφλι για την ανίχνευση μίας κορυφής είναι η διάμεσος (median) τιμή όλων των κορυφών και το ελάχιστο χρονικό διάστημα μεταξύ δύο διαδοχικών κορυφών ορίστηκε ως 90.7 ms. Διατηρώντας σταθερό το χρονικό διάστημα μεταξύ δύο διαδοχικών κορυφών, μελετήθηκε πώς επιδρά στην απόδοση η τιμή του κατωφλίου (Σχήματα 3.12, 3.13), θέτοντας διάφορα κατώφλια τα οποία εκφράζονται ως ένα ποσοστό της διαμέσου τιμής όλων των κορυφών.



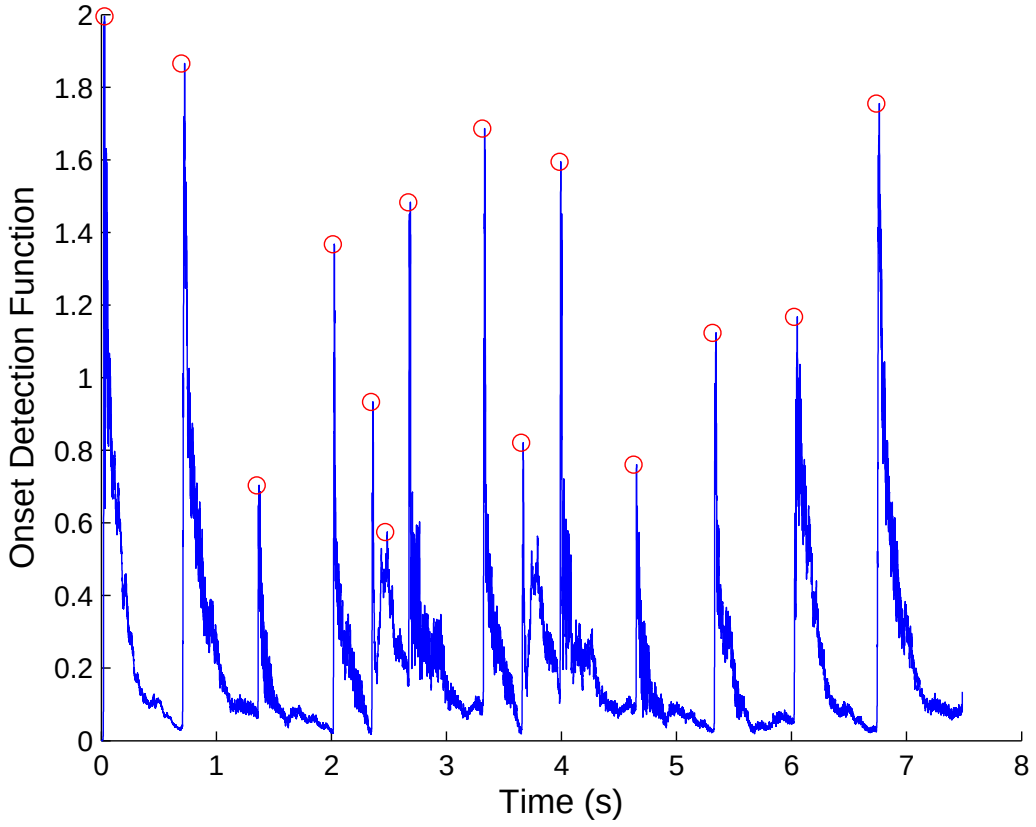
Σχήμα 3.9: Φιλτράρισμα με μια συστοιχία φίλτρων Gabor, εφαρμογή του διακριτού τελεστή Teager στην έξοδο κάθε ζώνης, εύρεση της συνάρτησης ανίχνευσης με την άθροιση της πληροφορίας από κάθε ζώνη και προσδιορισμός των onsets από τις κορυφές της συνάρτησης.



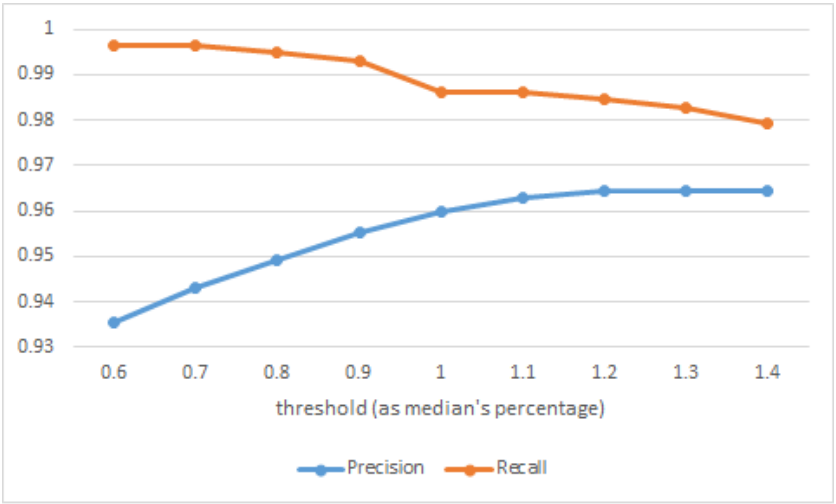
Σχήμα 3.10: Teager Onset Detection Function, Φεγγαράκι μου λαμπρό

Όσο αυξάνεται το κατώφλι, μειώνονται τα false positives αλλά παράλληλα αυξάνονται τα false negatives, συνεπώς το precision αυξάνεται ενώ το recall μειώνεται.

Το κύριο μειονέκτημα της μεθόδου και η κύρια πηγή σφαλμάτων είναι η μέθοδος κατωφλίω-



Σχήμα 3.11: Teager Onset Detection Function, Mussorgksy Promenade

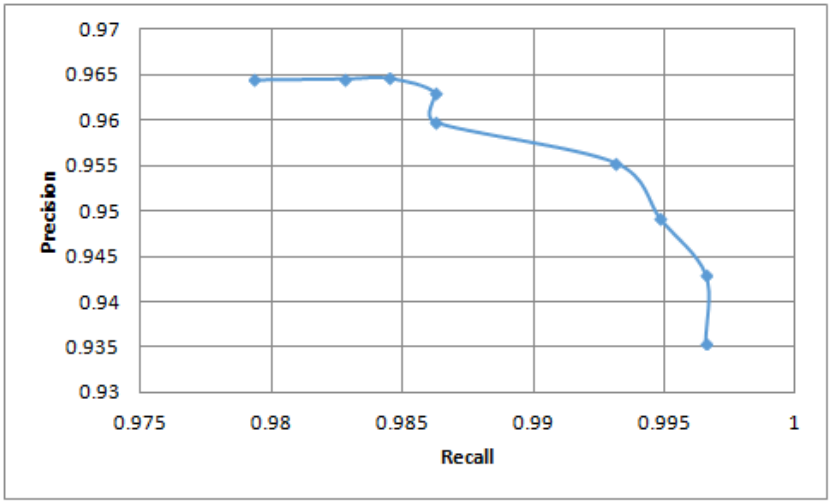


Σχήμα 3.12: Οι καμπύλες του precision και του recall για τη μέθοδο ανίχνευσης των onsets που βασίζεται στον τελεστή Teager ως συνάρτηση του κατωφλίου. Το εκάστοτε κατώφλι, αποτελεί ένα ποσοστό της διαμέσου τιμής όλων των κορυφών της συνάρτησης ανίχνευσης.

Ενότητα 3.3: Ανίχνευση της αρχής (onset) της νότας σε μονοφωνική μουσική για πιάνο με τη χρήση του Teager τελε

Κομμάτι	TP	FP	FN	Pr%	Re%	F-m%
Au Clair De La Lune	44	7	0	86.27	100	92.63
For Elise	50	0	1	100	98.04	99.01
Harry Potter theme	60	3	0	95.24	100	97.56
Long Long Ago	66	0	6	100	91.67	95.65
Mussorgksy Promenade	13	1	0	92.86	100	96.30
Chopin Valse	236	9	0	96.33	100	98.13
Τι Χαρά	48	1	1	97.96	97.96	97.96
Φεγγαράκι μου λαμπρό	56	3	0	94.92	100	97.39
Σύνολο	573	24	8	95.98	98.62	97.28

Πίνακας 3.2: Αποτελέσματα του onset detection με τη βοήθεια του τελεστή Teager. Ο πίνακας δείχνει τον αριθμό των onsets που ανιχνεύθηκαν σωστά (True Positives -TP), τα False Positives (FP), τα False Negatives (FN), το Precision (Pr), το Recall (Re) και το F-measure (F-m).



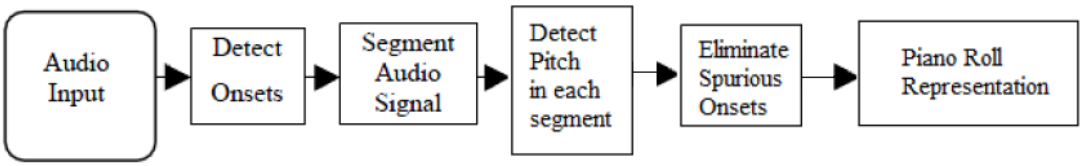
Σχήμα 3.13: Η καμπύλη Precision-Recall για τη μέθοδο ανίχνευσης των onsets που βασίζεται στον τελεστή Teager ως συνάρτηση του κατωφλίου, για τις ίδιες τιμές κατωφλίου με το Σχήμα 3.12

.

σης που ακολουθείται. Είναι φανερό από τις γραφικές παραστάσεις ότι εν γένει στη συνάρτηση η οποία αθροίζει την ενέργεια Teager από κάθε ζώνη υπάρχει η πληροφορία για τα onsets οπότε αυτό που έχει κρίσιμη σημασία είναι το πώς θα ανακτηθεί αυτή η πληροφορία με αυτόματο τρόπο. Κάθε μουσικό σήμα έχει τα δικά του χαρακτηριστικά και ιδιαιτερότητες, οπότε μία γενική μέθοδος επιλογής των κορυφών όπως αυτή που εφαρμόζουμε, είναι αναμενόμενο να προκαλεί λάθη σε ορισμένες περιπτώσεις. Κατάλληλη μετα-επεξεργασία της συνάρτησης και προσαρμοστική κατωφλίωση θα μπορούσε ενδεχομένως να βελτιώσει τα αποτελέσματα. Μία μελέτη της επίδρασης της μεθόδου επιλογής των κορυφών στην ανίχνευση των onsets παρουσιάζεται στο [82].

3.4 Το πλήρες σύστημα αυτόματης καταγραφής μονοφωνικής μουσικής

Τα διάφορα μέρη του πλήρους συστήματος φαίνονται στο σχήμα 3.14.



Σχήμα 3.14: Διάγραμμα του συστήματος καταγραφής μονοφωνικής μουσικής.

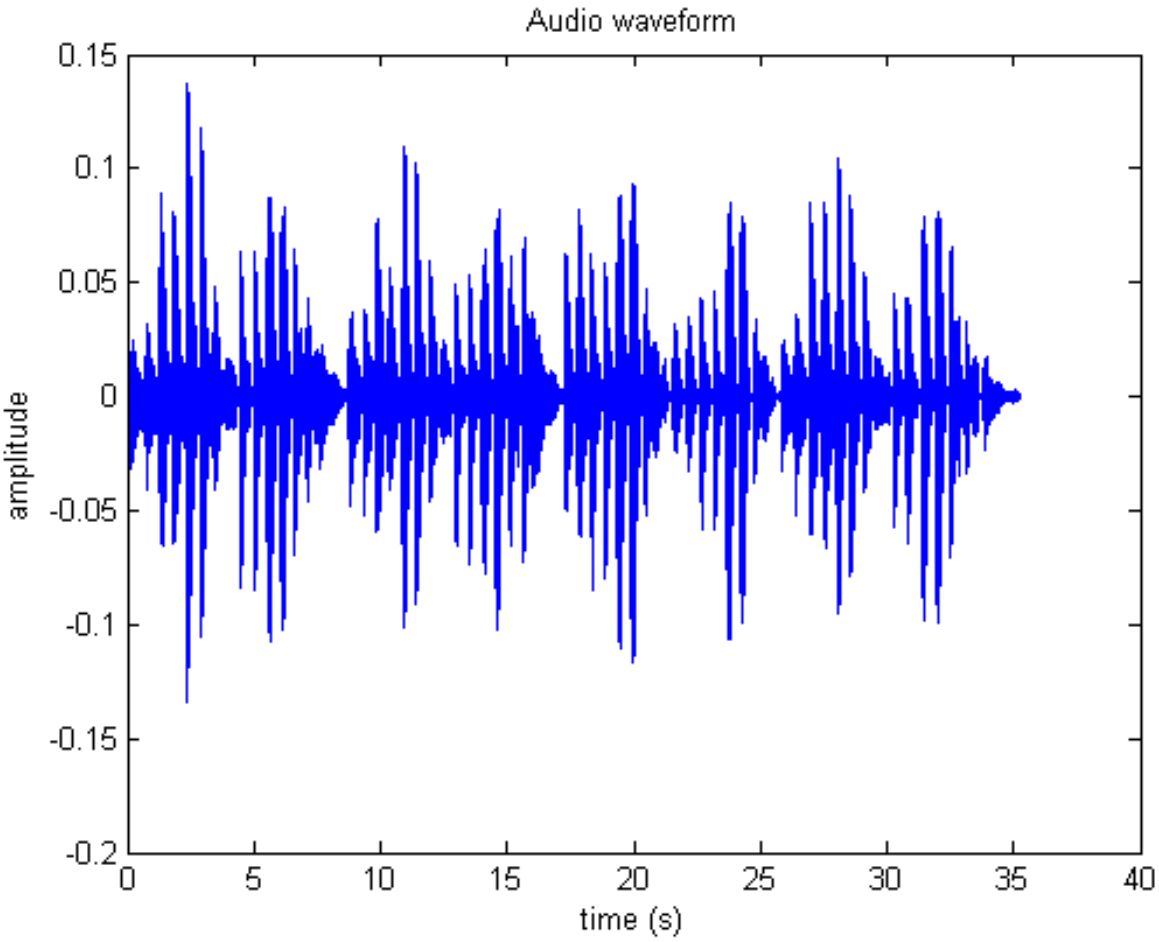
Η ανίχνευση των onsets γίνεται με τη μέθοδο που περιγράφεται στο 3.1.3 ενώ η εκτίμηση του pitch γίνεται με τη μέθοδο που περιγράφεται στο 3.2.2. Η κατάτμηση του μουσικού σήματος της εισόδου γίνεται με βάση τους χρόνους των onsets που έχουν υπολογιστεί και η εκτίμηση του pitch γίνεται χωριστά σε κάθε τμήμα. Τέλος όταν πλέον είναι διαθέσιμη και η πληροφορία για το pitch, επιχειρείται να διορθωθούν κάποια επιπλέον onsets που έχουν βρεθεί, εξετάζοντας αν δύο διαδοχικές νότες που έχουν το ίδιο pitch έχουν διάρκεια μεγαλύτερη από το μισό της διάρκειας ενός beat. Η διάρκεια ενός beat βρίσκεται με τη βοήθεια ενός αλγορίθμου που περιγράφεται στη συνέχεια στο 4.2. Αν κάποια από τις δύο διαδοχικές νότες του ίδιου pitch δεν έχει διάρκεια μεγαλύτερη του 1/2 της διάρκειας του beat, τότε οι δύο αυτές νότες συγχωνεύονται σε μία.

Στην υλοποίηση χρησιμοποιούνται διάφορες συναρτήσεις του MIR Toolbox (βλέπε σχετικό παράρτημα) όπως οι: miraudio, mirfilterbank, mirflux, mirsum, mirpeaks και mirsegment. Επίσης χρησιμοποιείται το MIDI Toolbox [83] για την αναπαράσταση σε piano roll και οι συναρτήσεις matrix2midi.m και writemidi.m από το [1] για την εγγραφή του αποτελέσματος σε αρχείο MIDI.

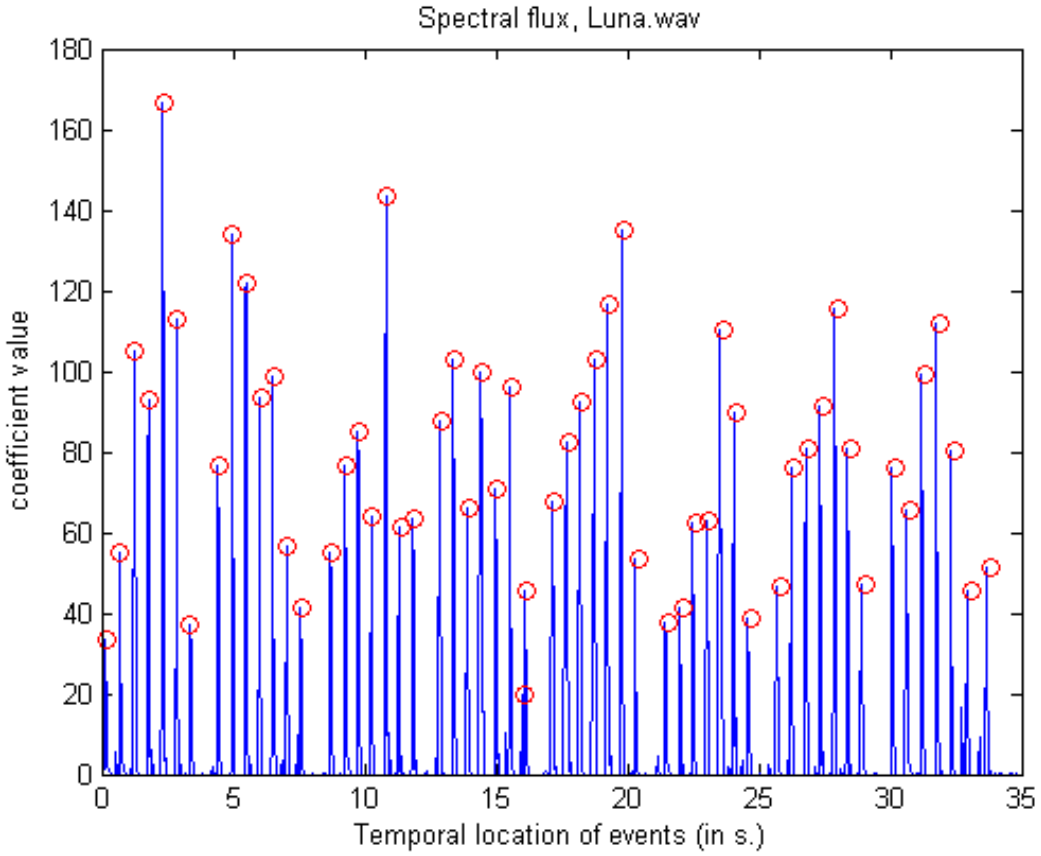
Στα σχήματα 3.15, 3.16, 3.17 και 3.18 φαίνονται οι έξοδοι των διαφόρων σταδίων του συστήματος για ένα μουσικό κομμάτι, το παιδικό τραγούδι "Φεγγαράκι μου λαμπρό".

Στα σχήματα 3.19, 3.20, 3.21 και 3.22 φαίνονται κάποιες ακόμα έξοδοι του συστήματος (αναπαραστάσεις σε piano roll), με σημειωμένα επάνω τα λάθη που έχουν γίνει.

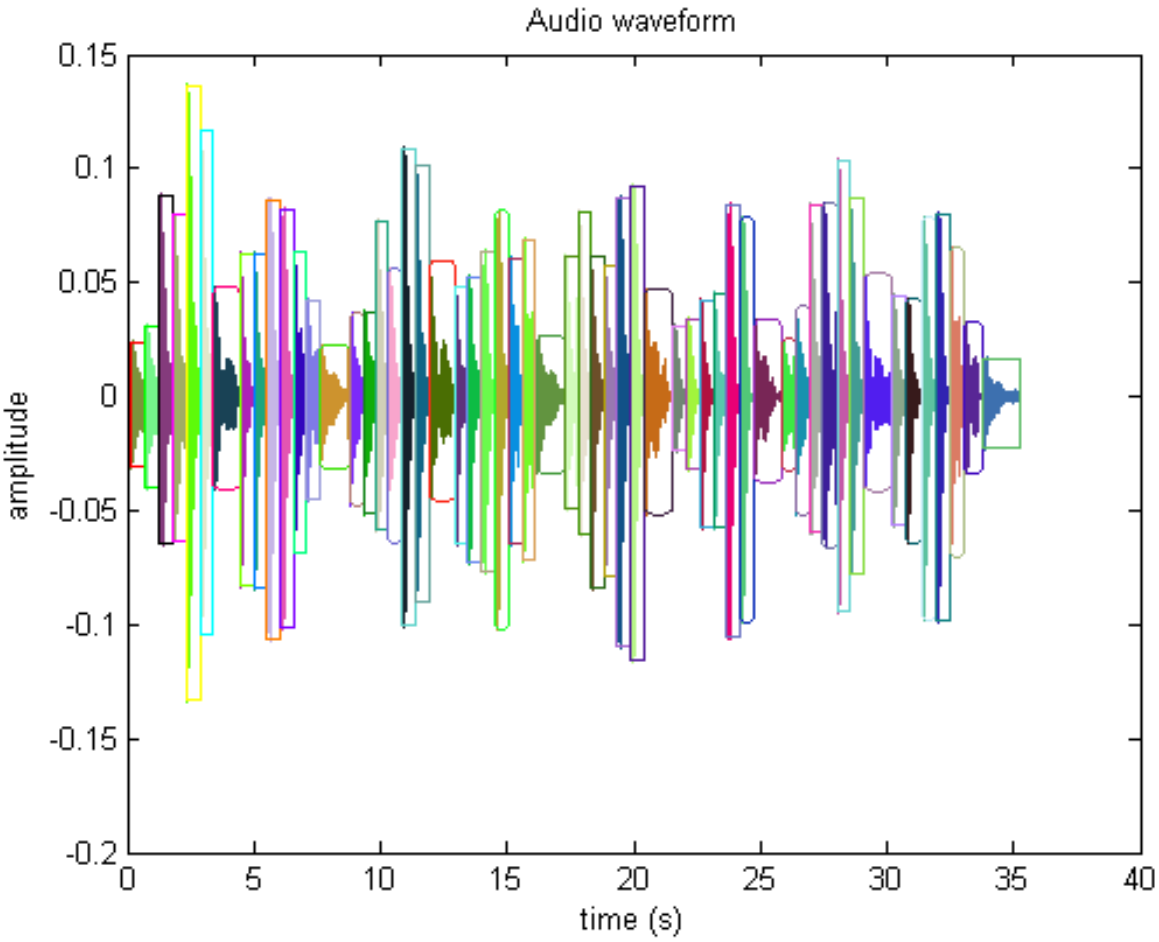
[1] http://www.kenschutte.com/midi



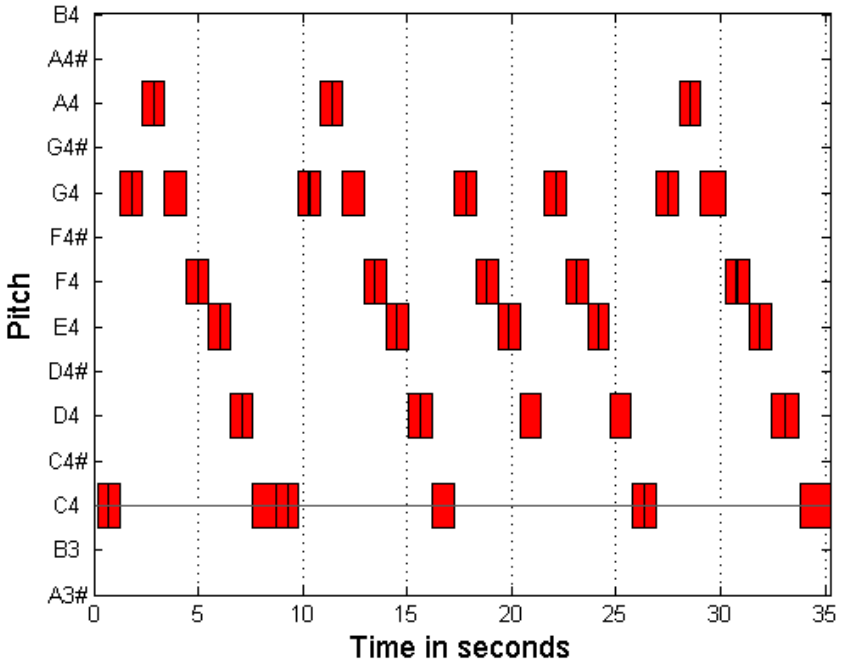
Σχήμα 3.15: Σήμα εισόδου, Φεγγαράκι μου λαμπρό



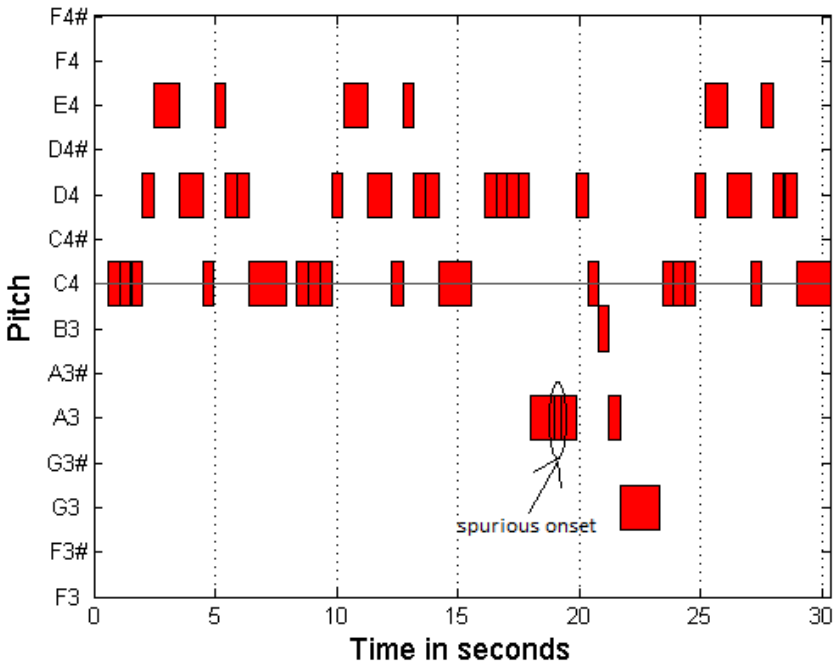
Σχήμα 3.16: Spectral flux, Φεγγαράκι μου λαμπρό



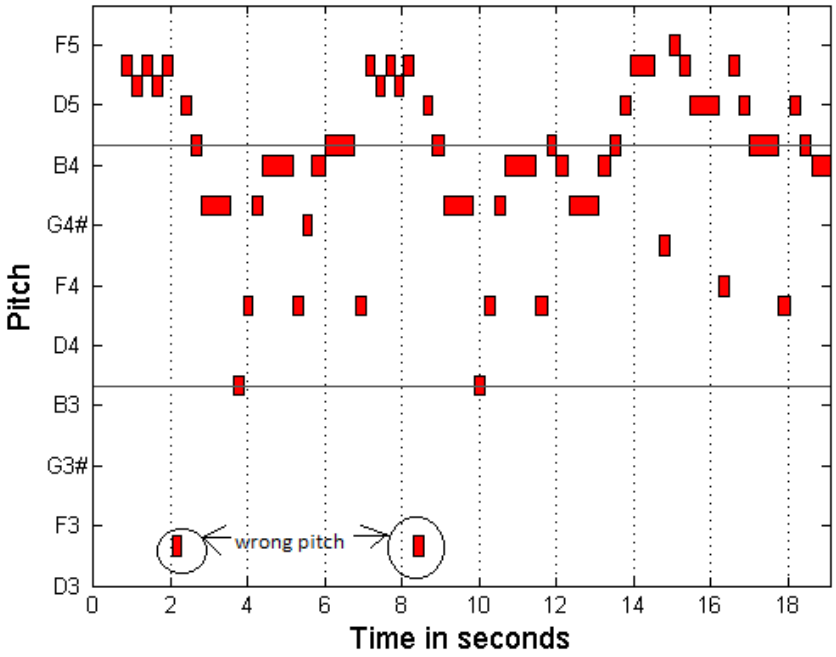
Σχήμα 3.17: Τεμαχισμένο σήμα εισόδου, Φεγγαράκι μου λαμπρό



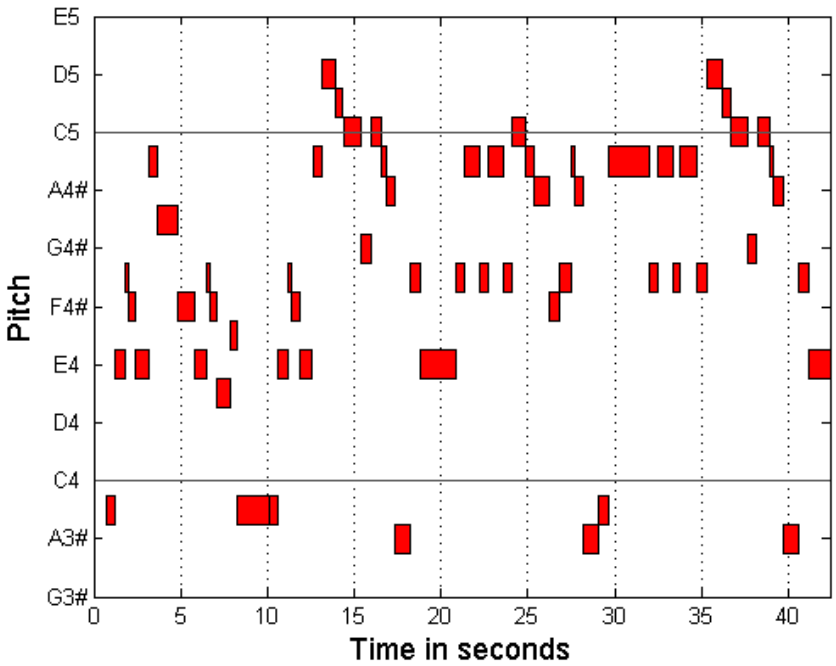
Σχήμα 3.18: Piano Roll - Φεγγαράκι μου λαμπρό



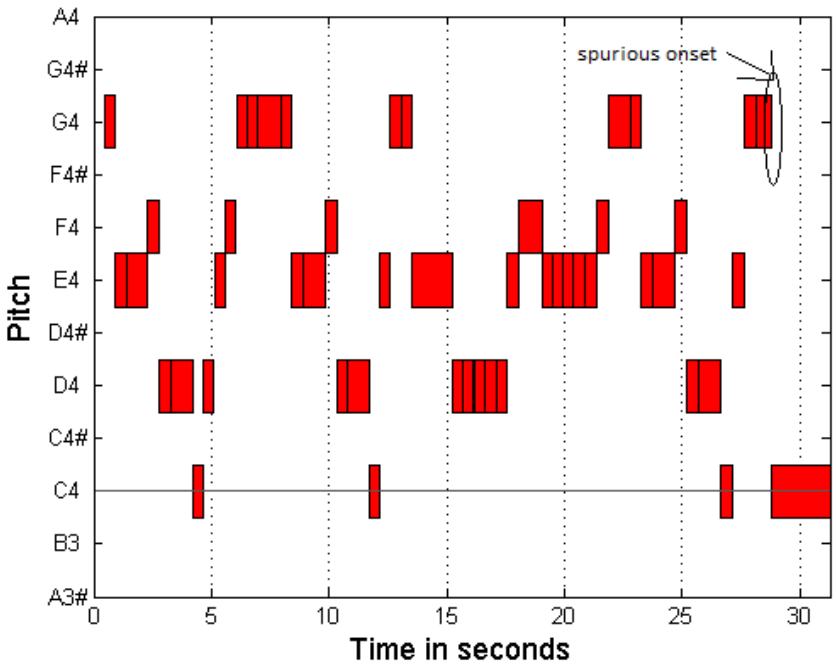
Σχήμα 3.19: Piano Roll - Au Clair De La Lune



Σχήμα 3.20: Piano Roll - Fur Elise



Σχήμα 3.21: Piano Roll - Harry Potter



Σχήμα 3.22: Piano Roll - Τι χαρά

Κεφάλαιο 4

Αυτόματη εύρεση ρυθμικής πληροφορίας

Ο ρυθμός είναι ένα από τα βασικότερα στοιχεία της μουσικής ενώ είναι αντιληπτός ακόμα και από άτομα που δεν έχουν μουσική παιδεία (σχεδόν οποιοσδήποτε είναι σε θέση να χτυπήσει παλαμάκια με τη μουσική). Η ανάλυση του ρυθμού είναι ένα κρίσιμο βήμα για την περιγραφή και την κατανόηση ενός μουσικού κομματιού. Ένα μουσικό κομμάτι διέπεται από μία ρυθμική δομή: τα μουσικά γεγονότα εκδηλώνονται σε συγκεκριμένες χρονικές στιγμές, έχουν συγκεκριμένη χρονική διάρκεια, συγκροτούν ομάδες στο χρόνο, επαναλαμβάνονται περιοδικά. Όπως εξηγείται και στην ενότητα 2.1.3 τα beats είναι μία ακολουθία από ισαπέχοντες παλμούς που εμφανίζονται περιοδικά στη μουσική. Είναι οι χρονικές στιγμές στις οποίες θα χτυπούσαμε το πόδι μας παράλληλα με τη μουσική. Το tempo ενός κομματιού είναι το αντίστροφο της περιόδου του beat. Αντί για συχνότητα σε Hz, το εκφράζουμε συχνά σε beats ανά λεπτό (Beats Per Minute). Από τη σκοπιά της αυτόματης καταγραφής μουσικής, η εκτίμηση της ρυθμικής πληροφορίας έγκειται στη χρονική κατάτμηση της μουσικής με βάση κάποια κριτήρια και επιτυγχάνεται ενισχύοντας και ανακαλύπτοντας τις εγγενείς περιοδικότητες ενός μουσικού κομματιού.

4.1 Σχετικές εργασίες

Στη βιβλιογραφία υπάρχουν διάφοροι αλγόριθμοι που εξάγουν αυτόματα ρυθμική πληροφορία από ένα μουσικό σήμα και καλύπτουν διαφορετικές εφαρμογές, όπως την εύρεση του ρυθμού, τον προσδιορισμό των beats και την εύρεση του μουσικού μέτρου. Οι πρώτες προσεγγίσεις επικεντρώνονταν σε σήματα MIDI, όμως πλέον οι σύγχρονες προσεγγίσεις εφαρμόζονται απευθείας σε πολυφωνική μουσική. Ο Scheirer [39] προτείνει μία μέθοδο που συνδέει μία συστοιχία φίλτρων με ένα σύνολο από φίλτρα τύπου χτένας. Απλούστερες μέθοδοι παρουσιάστηκαν από τον Seppannen [84] χρησιμοποιώντας την πληροφορία των onsets ή από τον Τζανετάκη [85] στα πλαίσια της ταξινόμησης σε μουσικό είδος. Μία άλλη προσέγγιση παρουσιάστηκε επίσης από τον Goto [86] για την εξαγωγή της ιεραρχικής δομής των beat. Για την ανάλυση της περιοδικότητας και την εκτίμηση του tempo, πολύ συχνά χρησιμοποιείται μία συνάρτηση ανίχνευσης από αυτές που είδαμε ότι χρησιμοποιούνται και για την εύρεση των onsets. Η μέθοδος της αυτοσυσχέτισης στηρίζεται στη σύγκριση της συνάρτησης ανίχνευσης με μία χρονικά ολισθημένη εκδοχή της και ανιχνεύει τις περιοδικές ομοιότητες ([87], [88]). Μία άλλη μέθοδος που χρησιμοποιείται ευρέως, βασίζεται σε μία συστοιχία από συντονιστές, όπου μία συνάρτηση ανίχνευσης συγκρίνεται

με πρότυπα από ισαπέχουσες κορυφές που καλύπτουν ένα μεγάλο εύρος περιόδων και φάσεων ([32],[39]). Σε κάποιες άλλες προσεγγίσεις, ο μετασχηματισμός Fourier βραχέως χρόνου χρησιμοποιείται για την εξαγωγή μίας αναπαράστασης χρόνου-συχνότητας της συνάρτησης ανίχνευσης ([89],[90],[91]). Εδώ, η συνάρτηση ανίχνευσης συγκρίνεται με πρότυπα που αποτελούνται από ημιτονοειδείς πυρήνες, με κάθε έναν να αντιπροσωπεύει μία συγκεκριμένη συχνότητα. Όλες αυτές οι μέθοδοι στοχεύουν στην αποκάλυψη των περιοδικών ιδιοτήτων της συνάρτησης ανίχνευσης από τις οποίες μπορεί να εκτιμηθεί το tempo ή η δομή των beats. Όσον αφορά το tempo, σημειώνεται ότι στις περισσότερες περιπτώσεις δεν είναι σταθερό κατά τη διάρκεια ολόκληρου του κομματιού. Οι ηχογραφήσεις αποτελούν εκφραστικές ερμηνείες κι έτσι το tempo μεταβάλλεται, πολλές φορές ακόμα και με απότομο ή ακραίο τρόπο. Οι περισσότερες μέθοδοι απλά επιστρέφουν ένα καθολικό tempo για ολόκληρη την ηχογράφηση. Με τις διαφοροποιήσεις στο tempo, το πρόβλημα της τοπικής εκτίμησης του tempo (εκτίμηση για κάθε σημείο στο χρόνο) γίνεται ένα πολύ δύσκολο ή ακόμα και μη καλώς ορισμένο πρόβλημα. Και το πρόβλημα του προσδιορισμού των θέσεων των beats (beat tracking) είναι ένα δύσκολο πρόβλημα, παρόλο που για τον άνθρωπο αποτελεί μία πολύ φυσική και εύκολη διαδικασία. Ειδικά στην περίπτωση μεγάλων διαφοροποιήσεων στο tempo το πρόβλημα δυσκολεύει πολύ. Το πρόβλημα περιπλέκεται κι άλλο επειδή θεωρείται ότι υπάρχουν διάφορα μετρικά επίπεδα τα οποία συμβάλουν στην αντίληψη που έχει ο άνθρωπος για το beat. Οι περισσότερες προσεγγίσεις εστιάζουν στον προσδιορισμό των παλμών στο επίπεδο tactus (ο ρυθμός στον οποίο κτυπάμε το πόδι μας) ([9],[87],[91]). Επίσης, δεν είναι απαραίτητο η θέση ενός beat να αντιστοιχεί σε onset κάποιας νότας. Οι μέθοδοι για beat tracking δουλεύουν καλύτερα σε μουσική ποπ και ροκ όπου ένα δυνατό και σταθερό beat είναι παρόν. Επαναλαμβανόμενα πρότυπα τονισμών διαμορφώνουν χαρακτηριστικές ομάδες παλμών, που καθορίζουν το μέτρο ενός μουσικού κομματιού. Εδώ, κάθε ομάδα ξεκινάει με ένα έντονα τονισμένο beat και αποτελείται από όλους τους παλμούς μέχρι τον επόμενο τονισμό. Η αυτόματη εξαγωγή του μέτρου είναι ένα δύσκολο πρόβλημα. Μία από τις πρώτες προσεγγίσεις για την επίλυσή του περιγράφεται στο ([92]), όπου ο αριθμός των beats μεταξύ των τονισμών εκτιμάται για την διαφοροποίηση μεταξύ κομματιών με τρίσημο και κομματιών με δίσημο μέτρο. Ένας άλλος τρόπος για την εξαγωγή χαρακτηριστικών που σχετίζονται με το ρυθμό, είναι η εξέταση διαστημάτων που ορίζονται μεταξύ διαδοχικών onsets ή beats, τα οποία λέγονται inter-onset-intervals (IOIs). Με βάση το ιστόγραμμα των διαρκειών των IOIs που εμφανίζονται, κάποιος μπορεί να εξάγει υποθέσεις για την περίοδο του beat, το tempo, και το μέτρο ([93], [94]). Το μειονέκτημα αυτών των προσεγγίσεων είναι ότι βασίζονται σε έναν σαφή εντοπισμό ενός διακριτού συνόλου από θέσεις onset και beat - ένα βήμα επιρρεπές σε λάθη. Για την αντιστάθμιση αυτών των λαθών, διάφορες προσεγγίσεις έχουν προταθεί που εκτιμούν από κοινού τις παραμέτρους που σχετίζονται με το ρυθμό ([95],[96]). Η πληροφορία για το ρυθμό και το tempo χρησιμοποιείται ευρέως στα προβλήματα ταξινόμησης ενός κομματιού με βάση το είδος του. Επίσης χρησιμεύει στον αυτόματο ρυθμικό συγχρονισμό πολλαπλών οργάνων, καναλιών ή μουσικών κομματιών (για μίξεις ή καραόκε) και σε γραφικά υπολογιστών που καθοδηγούνται από το beat (π.χ. ψηφιακοί χορευτές).

4.2 Εκτίμηση του tempo

Μία συγκριτική παρουσίαση και αξιολόγηση κάποιων αλγορίθμων εκτίμησης του μουσικού tempo γίνεται στο [62]. Οι αλγόριθμοι που σχεδιάζονται για να εκτιμήσουν το tempo ενός μουσικού κομματιού, στηρίζονται στην ίδια βασική αρχή: αρχικά από τα ηχητικά δεδομένα εξάγεται μία ακολουθία χαρακτηριστικών σε συνάρτηση με το χρόνο η οποία υπακούει την επικρατούσα ρυθμική πληροφορία και αποτυπώνει όλη τη σχετική με το ρυθμό πληροφορία του μουσικού κομματιού. Έπειτα ακολουθεί η εξαγωγή του tempo από αυτήν την ακολουθία χαρακτηριστικών. Η

εξαγωγή του tempo μπορεί να γίνει σε δύο στάδια: πρώτα εξάγεται ένα περιοδικό διάνυσμα κι έπειτα από αυτό το διάνυσμα επιλέγεται το ζητούμενο tempo (συνήθως επιλέγοντας τις κορυφές του διανύσματος). Η ακολουθία χαρακτηριστικών μπορεί να είναι μία συνάρτηση ανίχνευσης των onsets ή άλλα χαρακτηριστικά του σήματος, υπολογισμένα σε ένα μειωμένο ρυθμό δειγματοληψίας. Για παράδειγμα, αν τα χαρακτηριστικά είναι μια συνάρτηση ανίχνευσης, τότε πρόκειται για μια λίστα από τις χρονικές στιγμές και τα πλάτη των onsets των νοτών, ενώ μια άλλη πιθανή ακολουθία χαρακτηριστικών θα μπορούσε να είναι οι τιμές της μέσης ενέργειας, υπολογισμένες σε διαδοχικά πλαίσια διάρκειας 10ms ή 20ms, ή η παράγωγος της ενέργειας σε διαφορετικές ζώνες συχνοτήτων. Υπάρχουν ενδείξεις ότι μια ακολουθία χαρακτηριστικών η οποία έχει ληφθεί με ένα σχετικά συνεχή (διαδοχικά επικαλυπτόμενα πλαίσια) τρόπο από το σήμα εισόδου αποτελεί καλύτερη αναπαράσταση σαν πρώτο βήμα για την εύρεση του tempo συγκριτικά με την εύρεση των onsets των διακριτών ηχητικών γεγονότων σαν πρώτο βήμα. Πολλοί αλγόριθμοι επεξεργάζονται ξεχωριστά έναν αριθμό από διαφορετικές ζώνες συχνοτήτων, ανακαλύπτοντας ξεχωριστά τις περιοδικότητες στην κάθε μία και συνδυάζουν στο τέλος τα επιμέρους αποτελέσματα. Για αυτό το βήμα π.χ. ο αλγόριθμος του Scheirer [39] χρησιμοποιεί μία συστοιχία έξι IIR φίλτρων ενώ στο [97] συναντάται μία συστοιχία οκτώ φίλτρων.

Στα πλαίσια της διπλωματικής, χρησιμοποιήθηκε ο αλγόριθμος των Γκιόκα, Κατσούρου, Καραγιάννη και Σταφυλάκη ([6]) για την εκτίμηση του tempo, για τον οποίο οι συγγραφείς διέθεσαν και την υλοποίηση. Ο αλγόριθμος είναι πρόσφατος και κατετάγη πρώτος στον διαγωνισμό του MIREX 2011 για την εκτίμηση του tempo, ξεπερνώντας την επίδοση όλων των άλλων αλγορίθμων που είχαν υποβληθεί. Σε αυτόν τον αλγόριθμο χρησιμοποιείται διαχωρισμός του ηχητικού σήματος σε κρουστικό και αρμονικό μέρος, με σκοπό την εξαγωγή των filterbank ενεργειών από το κρουστικό μέρος και των chroma χαρακτηριστικών από το αρμονικό μέρος. Η ανάλυση της περιοδικότητας γίνεται με τη συνέλιξη της ακολουθίας των χαρακτηριστικών με μια συστοιχία συντονιστών. Το ζητούμενο tempo υπολογίζεται από το παραγόμενο διάνυσμα περιοδικότητας ενσωματώνοντας πληροφορία για τις μετρικές σχέσεις.

4.2.1 Εξαγωγή χαρακτηριστικών

Υπολογίζεται ο constant-Q μεταχηματισμός (CQT) του ηχητικού σήματος, χρησιμοποιώντας 12 κάδους συχνοτήτων ανά οκτάβα, με 25Hz/5kHz ελάχιστες/μέγιστες συχνότητες και ένα παράθυρο Hanning με 50% επικάλυψη. Οι κεντρικές συχνότητες f_k δίνονται από τη σχέση $f_k = 2^{k/b} f_{min}$ όπου b είναι ο αριθμός των φίλτρων ανά οκτάβα και f_{min} είναι η ελάχιστη κεντρική συχνότητα που λαμβάνεται υπόψιν. Οι κάδοι συχνοτήτων είναι ευθυγραμμισμένοι με τα pitches της δυτικής κλίμακας και στη συνέχεια εφαρμόζεται δικυβική παρεμβολή/αποδεκάτιση σε ένα σταθερό και ίσο με 200Hz ρυθμό πλαισίων. Έτσι προκύπτει το φασματογράφημα λογαριθμικής συχνότητας $S = \{W_{i,f}\}$ όπου $W_{i,f}$ είναι ο CQT και τα i, f είναι ο χρονικός και ο συχνοτικός δείκτης αντίστοιχα.

Για την ενίσχυση της εξαγωγής των επιθυμητών χαρακτηριστικών, εφαρμόζεται στο S ο αλγόριθμος αρμονικής/κρουστικής διάσπασης που περιγράφεται στο [98]. Ο αλγόριθμος αυτός είναι ιδιαίτερα γρήγορος και απλός. Γίνεται φιλτράρισμα με ένα φίλτρο μέσης τιμής κατά μήκος των διαδοχικών πλαισίων για την καταστολή των κρουστικών γεγονότων και την ενίσχυση των αρμονικών συστατικών, ενώ γίνεται φιλτράρισμα με ένα φίλτρο μέσης τιμής και κατά μήκος των κάδων συχνοτήτων για την ενίσχυση των κρουστικών γεγονότων και την καταστολή των αρμονικών συστατικών. Τα δύο φασματογραφήματα που προκύπτουν μετά το φιλτράρισμα, χρησιμοποιούνται για την παραγωγή μασκών που στη συνέχεια εφαρμόζονται στο αρχικό φασματογράφημα

για να διαχωρίσουν το αρμονικό από το κρουστικό μέρος του σήματος. Αυτή η τεχνική βασίζεται στη διαισθητική ιδέα ότι σταθερά αρμονικά ή στάσιμα συστατικά διαμορφώνουν οριζόντιες κορυφογραμμές στο φασματογράφημα, ενώ τα κρουστικά συστατικά διαμορφώνουν κατακόρυφες κορυφογραμμές με ευρυζωνική απόκριση συχνότητας.

Τα φίλτρα μέσης τιμής λειτουργούν αντικαθιστώντας ένα δεδομένο δείγμα σε ένα σήμα με τη διάμεσο των τιμών του σήματος σε ένα παράθυρο γύρω από το δείγμα. Δεδομένου ενός διανύσματος εισόδου $x(n)$ τότε $y(n)$ είναι η έξοδος ενός φίλτρου μήκους l όπου l ο αριθμός των δειγμάτων πάνω στα οποία το φιλτράρισμα μέσης τιμής λαμβάνει χώρα. Όταν το l είναι περιττό, το φίλτρο μέσης τιμής μπορεί να οριστεί ως:

$$\mathrm{y(n)} = \mathrm{median}\{x(n-k:n+k), k=(l-1)/2\} \tag{4.1}$$

Στην πραγματικότητα, το αρχικό δείγμα αντικαθίσταται από τη μεσαία τιμή που προκύπτει από μια ταξινομημένη λίστα των δειγμάτων στη γειτονιά του αρχικού δείγματος. Στις περιπτώσεις όπου το l είναι άρτιο, η διάμεσος προκύπτει ως το μέσο των δύο τιμών που βρίσκονται στη μέση της ταξινομημένης λίστας. Τα φίλτρα μέσης τιμής είναι αποτελεσματικά στο να αφαιρούν κρουστικό θόρυβο επειδή δεν εξαρτώνται από τιμές που είναι ακραίες σε σχέση με τις τυπικές τιμές της περιοχής γύρω από το αρχικό δείγμα.

Δεδομένου του πλάτους του φασματογραφήματος S σαν είσοδο, δηλώνοντας το i-οστό χρονικό πλαίσιο ως S_i και την h-οστή φέτα συχνότητας ως S_h, ένα κρουστικά-ενισχυμένο πλαίσιο φασματογραφήματος, το οποίο συμβολίζουμε με P_i, μπορεί να δημιουργηθεί εφαρμόζοντας φιλτράρισμα μέσης τιμής στο S_i:

$$P_i = M\{S_i, l_{perc}\} \tag{4.2}$$

όπου το σύμβολο M δηλώνει φιλτράρισμα μέσης τιμής και l_{perc} είναι το μήκος του φίλτρου. Τα επιμέρους κρουστικά-ενισχυμένα πλαίσια συνδυάζονται στη συνέχεια αποφέροντας ένα κρουστικά ενισχυμένο φασματογράφημα P. Ομοίως, μία αρμονικά-ενισχυμένη φέτα συχνοτήτων ενός φασματογραφήματος, την οποία συμβολίζουμε με H_i, μπορεί να ληφθεί φιλτράροντας με φίλτρο μέσης τιμής τη φέτα συχνοτήτων S_h:

$$H_i = M\{S_h, l_{harm}\} \tag{4.3}$$

όπου το σύμβολο M δηλώνει φιλτράρισμα μέσης τιμής και l_{harm} είναι το μήκος του φίλτρου. Οι φέτες συνδυάζονται στη συνέχεια αποφέροντας ένα αρμονικά ενισχυμένο φασματογράφημα H. Τα μήκη των φίλτρων στην υλοποίηση ισούνται με 10.

Τα δύο φασματογραφήματα που προκύπτουν χρησιμοποιούνται για τη δημιουργία μασκών που μπορούν έπειτα να εφαρμοστούν στο αρχικό φασματογράφημα. Χρησιμοποιούνται soft μάσκες που ορίζονται ως:

$$M_{H_{h,i}} = \frac{H_{h,i}^p}{H_{h,i}^p + P_{h,i}^p} \tag{4.4}$$

$$M_{P_{h,i}} = \frac{P_{h,i}^p}{H_{h,i}^p + P_{h,i}^p} \tag{4.5}$$

όπου το p ορίζει τη δύναμη στην οποία υψώνεται κάθε ξεχωριστό στοιχείο του φασματογραφήματος. Η τιμή του p τίθεται ίση με 2.

Τέλος τα αρμονικά/κρουστικά συστατικά H και P αντίστοιχα προκύπτουν από το S ως εξής:

$$H = S \otimes M_H \tag{4.6}$$

$$P = S \otimes M_P \tag{4.7}$$

όπου το $\otimes$ ορίζει τον πολλαπλασιασμό στοιχείο προς στοιχείο.

Για κάθε χρονικό δείκτη i αθροίζουμε τα πλάτη των κάδων συχνοτήτων στο H που αντιστοιχούν στα 12 ημιτόνια της δυτικής μουσικής κλίμακας με σκοπό τον υπολογισμό του 12-διάστατου διανύσματος x_{ch}:

$$x_{ch}^{k}[i] = \sum_{f_k \in F_k} H_{i,f_k,k=1..12} \tag{4.8}$$

όπου F_k είναι οι κάδοι που αντιστοιχούν στον τόνο k.

Με έναν αντίστοιχο τρόπο, οι filterbank ενέργειες εξάγονται από το P με 8 λογαριθμικής κλίμακας, ίσου εύρους, επικαλυπτόμενα τριγωνικά φίλτρα. Οι filterbank ενέργειες δηλώνονται ως x_{fl}.

4.2.2 Ανάλυση περιοδικότητας

Τα διανύσματα χαρακτηριστικών διαφορίζονται και συνελίσσονται με μια συστοιχία συντονιστών στο εύρος [30,500] bpm. Η βασική μονάδα ταλάντωσης έχει υιοθετηθεί ως κρουστική απόκριση του συντονιστή. Η εξίσωση της μονάδας ταλάντωσης δίνεται από την:

$$r(i) = 1 + \tanh(\gamma * (\sin(2\pi\psi_t i) - 1)) \tag{4.9}$$

όπου ψ_t είναι η συχνότητα του tempo t. Η παράμετρος γ ονομάζεται κέρδος εξόδου (γ=8). Οι έξοδοι των συντονιστών τεμαχίζονται χρησιμοποιώντας ένα τετραγωνικό παράθυρο 8πλάσιο του μήκους της ζητούμενης περιόδου tempo και με μια περίοδο επικάλυψη. Υπολογίζουμε το salience του tempo t στο τμήμα s ως τη μέγιστη τιμή πλάτους στην έξοδο του αντίστοιχου συντονιστή στο τμήμα s.

Αυτή η διαδικασία χρησιμοποιείται για τον υπολογισμό των 'tempogram' πινάκων $TG^v(t,s)$ για κάθε διάνυσμα χαρακτηριστικών v. Η επίδραση της ιδιότητας του constant Q έχει ως αποτέλεσμα λιγότερα τμήματα για μικρότερα tempi, έτσι γίνεται time wrapping στις γραμμές του TG^v για να έχουν το ίδιο μέγεθος που είναι ίσο με το μέγεθος του ταχύτερου tempo. Έπειτα, για κάθε δείκτη τμήματος s, τα tempograms αθροίζονται για κάθε τάξη χαρακτηριστικών (filterbank/chroma) ξεχωριστά, έχοντας ως αποτέλεσμα τους πίνακες TG^{ch} και TG^{fl} για τα chroma και τα filterbank χαρακτηριστικά αντίστοιχα. Για την εκτίμηση ενός καθολικού διανύσματος περιοδικότητας για ολόκληρο το μουσικό κομμάτι, έστω T_{gl}, τα TG^{ch} και TG^{fl} αθροίζονται κατά μήκος όλων των τμημάτων και μετά πολλαπλασιάζονται:

$$T_{gl}(t) = (\sum_s TG^{fl}(t,s))(\sum_s TG^{ch}(t,s)) \tag{4.10}$$

4.2.3 Επιλέγοντας το σωστό μετρικό επίπεδο

Η μέθοδος κάνει την υπόθεση ότι οι κορυφές του καθολικού διανύσματος περιοδικότητας που σχετίζονται μουσικά με τα πραγματικά δεδομένα, είναι ακέραια πολλαπλάσια μίας συγκεκριμένης

τιμής tempo. Υπολογίζουμε το θεμελιώδες tempo T_0 ως

$$T_0 = arg\max_t \{\sum_{k=1}^{4} T_{gl}(kt)\} \tag{4.11}$$

Έπειτα περιμένουμε ότι το T_{gl} έχει κορυφές στο ζητούμενο tempo καθώς επίσης και σε ακέραια πολλαπλάσια του T_0. Κατά αντιστοιχία με το {μέτρο, tactus, tatum } ιεραρχικό μοντέλο για τις σχέσεις μεταξύ των beat, στην προσέγγισή μας θεωρούμε ένα μοντέλο με δύο tempi με τιμές T_s, T_f (αργό, γρήγορο) με την υπόθεση ότι το T_s είναι αυτό που γίνεται πιο εύκολα αντιληπτό, ενώ το T_f είναι κατά πάσα πιθανότητα το διπλάσιο, το τριπλάσιο ή το τετραπλάσιο του T_s. Ορίζουμε την από κοινού salience $J_s(T_s, T_f)$ των T_s, T_f ως:

$$J_s(T_s, T_f) = [T_{gl}(T_s) + T_{gl}(T_f)] \sum_{i=2..4} e^{-(T_f/T_s - i)^2/(\rho i)^2} \tag{4.12}$$

Είναι προφανές ότι το J_s γίνεται μεγαλύτερο καθώς οι saliences των T_s, T_f αυξάνονται και όταν το δεύτερο είναι το διπλάσιο, το τριπλάσιο ή το τετραπλάσιο του πρώτου. Το τελικό tempo T είναι το T_s που μεγιστοποιεί το J_s και είναι πολλαπλάσιο του T_0, δηλαδή:

$$T = arg\max_{iT_0} \{J_s(iT_0, kT_0), iT_0, kT_0 \in \{30, ...500\}\} \tag{4.13}$$

4.3 Beat tracking

Beat tracking είναι η εξαγωγή από το μουσικό ηχητικό σήμα μιας ακολουθίας από χρονικές στιγμές που θα μπορούσαν να αντιστοιχούν στο πότε ένας άνθρωπος-ακροατής θα κτυπούσε το πόδι του. Πρέπει να ικανοποιούνται δύο συνθήκες. Από τη μια πλευρά θα πρέπει οι χρονικές στιγμές γενικά να αντιστοιχούν σε στιγμές στον χρόνο όπου υπάρχει ένδειξη για beat, για παράδειγμα από το onset μιας νότας που παίζεται από ένα από τα όργανα. Από την άλλη πλευρά, το σύνολο των beats θα πρέπει να ανταναχλά ένα χρονικά σταθερό διάστημα μεταξύ των beats, εφόσον είναι τα τακτά διαστήματα μεταξύ των χρονικών στιγμών των beats που καθορίζουν το μουσικό ρυθμό. Αυτοί οι δυαδικοί περιορισμοί μπορούν να αξιοποιηθούν για την εφαρμογή δυναμικού προγραμματισμού για την εύρεση ενός συνόλου από χρονικές στιγμές beat που αντανακλούν το tempo ενώ επίσης αντιστοιχούν σε στιγμές υψηλού 'onset strength' στην συνάρτηση ανίχνευσης των onsets που έχει εξαχθεί από το ηχητικό σήμα. Αρκεί να οριστεί μία αντικειμενική συνάρτηση που επιδιώκει να μεγιστοποιήσει τόσο τα 'onset strengths' σε κάθε υποτιθέμενη χρονική στιγμή ενός beat και τη συνέπεια των διαστημάτων μεταξύ των beats με το σταθερό tempo που έχει υπολογιστεί εκ των προτέρων. Η μέθοδος αυτή που προτείνεται στο [9], είναι ιδιαίτερα απλή και υπολογιστικά αποδοτική. Ο κώδικας της μεθόδου διατίθεται στο διαδίκτυο[1]. Αναφέρεται ότι ο στόχος που μας ενδιαφέρει περισσότερο στα πλαίσια της διπλωματικής είναι η εύρεση του μέτρου, οπότε δεν είναι ανάγκη να προσδιοριστούν οι χρονικές στιγμές που εμφανίζονται τα beats με πάρα πολύ μεγάλη ακρίβεια, αρκεί μια καλή εκτίμηση η οποία και θα χρησιμοποιηθεί στη συνέχεια για την εύρεση του μέτρου.

4.3.1 Αλγόριθμος δυναμικού προγραμματισμού

Θεωρούμε ότι έχουμε ένα σταθερό tempo που δίνεται εκ των προτέρων. Ο στόχος του beat tracker είναι να δημιουργήσει μία ακολουθία από χρονικές στιγμές beats που αντιστοιχούν τόσο

[1] http://labrosa.ee.columbia.edu/projects/beattrack/

στα αντιλαμβανόμενα onsets του ηχητικού σήματος ενώ την ίδια στιγμή συνιστούν ένα τακτικό ρυθμικό μοτίβο από μόνα τους. Μπορούμε να ορίσουμε μία απλή αντικειμενική συνάρτηση που συνδυάζει και τους δύο αυτούς στόχους:

$$C(\{t_i\}) = \sum_{i=1}^{N} O(t_i) + \alpha \sum_{i=2}^{N} F(t_i - t_{i-1}, \tau_p) \tag{4.14}$$

όπου $\{t_i\}$ είναι η ακολουθία από τα N beats που βρέθηκαν από τον ανιχνευτή και $O(t)$ η συνάρτηση ανίχνευσης των onsets που έχει υψηλές στιγμές στις χρονικές στιγμές που θα αποτελούσαν καλές επιλογές για τα beats με βάση τις τοπικές ακουστικές ιδιότητες, το α ρυθμίζει την ισορροπία της σημαντικότητας των δύο όρων και το $F(\Delta t, \tau_p)$ είναι μια συνάρτηση που μετράει τη συνέπεια μεταξύ ενός διαστήματος Δt ανάμεσα στα beats και της ιδανικής απόστασης τ_p που ορίζεται από το tempo-στόχο. Για παράδειγμα, χρησιμοποιούμε μια απλή συνάρτηση τετραγωνικού λάθους που εφαρμόζεται στο λογαριθμικό λόγο του πραγματικού προς το ιδανικό διάστημα, δηλαδή:

$$F(\Delta t, \tau_p) = -(\log \frac{\Delta t}{\tau})^2 \tag{4.15}$$

που έχει μέγιστη τιμή 0 όταν $\Delta t = \tau$, γίνεται αυξανόμενα αρνητική για μεγαλύτερες αποκλίσεις, και είναι συμμετρική σε έναν λογαριθμικό ως προς το χρόνο άξονα, ώστε $F(k\tau, \tau) = F(\tau/k, \tau)$.

Στη συνέχεια θεωρούμε ότι ο χρόνος είναι κβαντισμένος σε ένα κατάλληλο πλέγμα.

Η ιδιότητα κλειδί της αντικειμενικής συνάρτησης είναι ότι η χρονική ακολουθία με το καλύτερο score μπορεί να βρεθεί αναδρομικά. Για να βρούμε το μεγαλύτερο δυνατό score $C^*(t)$ όλων των ακολουθιών που τελειώνουν τη χρονική στιγμή t, ορίζουμε την αναδρομική σχέση:

$$C^*(t) = O(t) + \max_{\tau=0..t} \{\alpha F(t - \tau, \tau_p) + C^*(\tau)\} \tag{4.16}$$

Αυτή η εξίσωση στηρίζεται στην παρατήρηση ότι το καλύτερο score τη χρονική στιγμή t είναι το τοπικό onset strength, συν το καλύτερο score του προηγούμενου beat τ που μεγιστοποιεί το άθροισμα αυτού του καλύτερου score και του κόστους μετάβασης από εκείνη τη χρονική στιγμή. Ενώ υπολογίζουμε το $C^*(t)$, καταγράφουμε επίσης την χρονική στιγμή του προηγούμενου beat που έδωσε το βέλτιστο score:

$$P^*(t) = arg \max_{\tau=0..t} \{\alpha F(t - \tau, \tau_p) + C^*(\tau)\} \tag{4.17}$$

Στην πραγματικότητα είναι απαραίτητο να ψάξουμε μόνο ένα περιορισμένο εύρος από τ εφόσον ο γρήγορα αυξανόμενος όρος ποινής F καθιστά απίθανο η βέλτιστη στιγμή του προηγούμενου beat να είναι μακριά από το $t - \tau_p$. Έτσι η αναζήτηση γίνεται για $\tau = t - 2\tau_p ... t - \tau_p/2$.

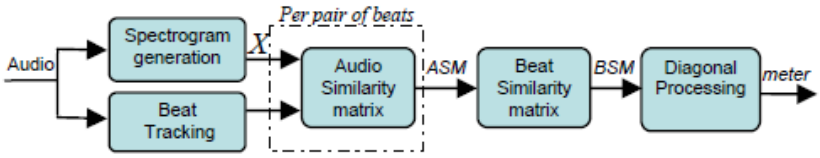
Για να βρούμε το σύνολο των beats που βελτιστοποιούν την αντικειμενική συνάρτηση για μια δεδομένη συνάρτηση ανίχνευσης, ξεκινάμε υπολογίζοντας τα C^* και P^* για κάθε στιγμή ξεκινώντας από το 0. Μόλις ολοκληρωθεί αυτή η διαδικασία, κοιτάμε για τη μεγαλύτερη τιμή του C^*; έτσι έχουμε την τελική εμφάνιση beat t_N- όπου το N, ο συνολικός αριθμός των beats, δεν είναι ακόμα γνωστός σε αυτό το σημείο. Έπειτα με 'οπισθοδρόμηση' μέσω του P^*, βρίσκουμε το προηγούμενο beat $t_{N-1} = P^*(t_N)$, και σταδιακά δουλεύοντας προς τα πίσω, φτάνουμε στην αρχή του σήματος. Αυτό δίνει ολόκληρη τη βέλτιστη ακολουθία $\{t_i\}^*$. Χάρη στο δυναμικό προγραμματισμό, ερευνήθηκε αποδοτικά ολόκληρο το εκθετικού μεγέθους σύνολο από όλες τις δυνατές

ακολουθίες beat σε γραμμικό χρόνο. Αυτό ήταν εφικτό διότι στην αντικειμενική συνάρτηση γεγονότα μεταγενέστερα του t_i δε μπορούν να επηρεάσουν τη συνεισφορά των προγενέστερων γεγονότων στο κόστος.

Η μέθοδος προϋποθέτει τη γνώση της συνάρτησης ανίχνευσης των onsets την οποία και υπολογίζουμε με τον αλγόριθμο που περιγράφεται στην ενότητα 3.1.3 καθώς και του tempo που βρίσκεται με τον τρόπο που περιγράφουμε στην προηγούμενη ενότητα 4.2.

4.4 Αυτόματη εύρεση του μέτρου

Για την εύρεση του μουσικού μέτρου, χρησιμοποιήθηκε η μέθοδος που παρουσιάζεται στο [4]. Η προσέγγιση βασίζεται στη δημιουργία ενός πίνακα ομοιότητας του beat (Beat Similarity Matrix), ο οποίος παρέχει πληροφορίες σχετικά με την ομοιότητα μεταξύ οποιωνδήποτε δύο beats ενός κομματιού μουσικής. Η επαναληπτική δομή που συναντάει κανείς στα περισσότερα μουσικά κομμάτια εξερευνάται με την επεξεργασία του πίνακα ομοιότητας του beat με σκοπό τον προσδιορισμό συναφών μοτίβων στα beats σε διαφορετικά μέρη του κομματιού. Με βάση αυτή την αρχή, μπορούν να βρεθούν με αποτελεσματικότητα τόσο δίσημα και τρίσημα μέτρα όσο και σύνθετα. Η χρήση των θέσεων των beats και τεχνικών δυναμικού προγραμματισμού επιτρέπει την ανίχνευση συναφών μουσικών μοτίβων που δημιουργούνται από beats με μέτριες αποκλίσεις στο tempo.



Σχήμα 4.1: Block διάγραμμα του συστήματος εύρεσης του μέτρου. Ανατύπωση από το [4].

Στο σχήμα 4.1 φαίνεται το block διάγραμμα του συστήματος. Αρχικά, παράγεται ένα φασματογράφημα του ηχητικού σήματος. Έπειτα, πίνακες ηχητικής ομοιότητας (Audio Similarity Matrices - ASM) υπολογίζονται συγκρίνοντας τα πλαίσια στο φασματογράφημα ανά κάθε δύο beats του κομματιού μουσικής. Κατόπιν, ένας πίνακας ομοιότητας του beat κατασκευάζεται χρησιμοποιώντας μέτρα ομοιότητας που παράγονται από τους επιμέρους πίνακες ηχητικής ομοιότητας. Τέλος, η ύπαρξη συναφών μοτίβων στα beats εξερευνάται με την επεξεργασία των διαγωνίων του πίνακα ομοιότητας beat.

4.4.1 Επιμέρους δομικά στοιχεία

Φασματογράφημα

Παράγεται ένα φασματογράφημα από παραθυρομένα πλαίσια μήκους $L = 1024$ δειγμάτων και hop size $H = 512$ δειγμάτων (το μισό μήκος του πλαισίου). Για καλύτερη υπολογιστική αποδοτικότητα, κρατάμε μόνο τους κάδους συχνοτήτων στο εύρος $1..S$, όπου το S αντιστοιχεί στον κάδο που είναι τοποθετημένος στα 5000 Hz.

Beat tracking

Το beat tracking γίνεται με τη μέθοδο που περιγράφτηκε στην ενότητα 4.3. Οι θέσεις των beats χρησιμοποιούνται στη συνέχεια με σκοπό τη σύγκριση κάθε δύο ζευγαριών από beats χρησιμοποιώντας τα αντίστοιχα πλαίσια του φασματογραφήματος.

ASM για κάθε δύο beats

Ένας ASM κατασκευάζεται συγκρίνοντας τα πλαίσια του φασματογραφήματος κάθε ζεύγους από beats. Για τη μέτρηση της ομοιότητας μεταξύ δύο πλαισίων $m = a$ και $m = b$, χρησιμοποιείται η μέτρηση της ευκλείδειας απόστασης:

$$ASM(a,b) = \sum_{k=1}^{S}[X(a,k) - X(b,k)]^2$$
$$\text{όπου} \tag{4.18}$$
$$X(m,k) = abs(\sum_{n=0}^{L-1} x(n+mH)w(n) * e^{-j(2\pi/N)kn})$$

$w(n)$ είναι ένα παράθυρο hanning που επιλέγει ένα μήκους L μπλοκ από το σήμα εισόδου $x(n)$, και m, N, k είναι ο δείκτης του πλαισίου, το μήκος του FFT και ο αριθμός του κάδου αντίστοιχα.

Έπειτα χρησιμοποιείται δυναμικός προγραμματισμός με σκοπό να βρεθεί το καλύτερο μονοπάτι μεταξύ της πάνω αριστερά και της κάτω δεξιά γωνίας του ASM που ελαχιστοποιεί το συνολικό κόστος ομοιότητας. Με σκοπό να βρεθεί το βέλτιστο μονοπάτι μέσα στον ASM για δύο beats x και y με μήκη lx και ly πλαίσια αντίστοιχα, δημιουργείται ένας πίνακας μετάβασης M. Έτσι, το $M_{i,j}$ αντιπροσωπεύει το ελάχιστο κόστος που χρειάζεται για να βρεθούμε στη θέση $[i, j]$ του ASM από την πάνω αριστερή γωνία του πίνακα:

$$M_{i,j} = ASM_{i,j} + min(M_{i-1,j-1}, M_{i-1,j}, M_{i,j-1}) \tag{4.19}$$

Η ομοιότητα μεταξύ δύο beats x και y δίνεται από το $S = M_{lx,ly}$. Με τη βοήθεια του δυναμικού προγραμματισμού ξεπερνιούνται οι δυσκολίες που οφείλονται στη διαφορετική διάρκεια που μπορεί να έχουν δύο beats και στην διαφορά της εκτιμώμενης θέσης του beat από την πραγματική.

Beat Similarity Matrix (BSM)

Το μέτρο ομοιότητας S μεταξύ κάθε ζευγαριού από beats ενός κομματιού μουσικής χρησιμοποιείται για την αναδρομική κατασκευή ενός πίνακα ομοιότητας beat. Έτσι, το $BSM(x, y)$ αντιστοιχεί στην ομοιότητα μεταξύ δύο beat x και y. Ο πίνακας BSM είναι συμμετρικός γύρω από την κύρια διαγώνιο, οπότε αρκεί ο υπολογισμός μόνο του μισού πίνακα.

Επεξεργασία των διαγωνίων και υπολογισμός του τύπου του μέτρου

Η ύπαρξη συναφών μετρικών δομών σε ένα κομμάτι μουσικής ερευνάται με την επεξεργασία των διαγωνίων του BSM. Κάθε διαγώνιος αντιπροσωπεύει την ομοιότητα μεταξύ beats που χωρίζονται από ένα διαφορετικό αριθμό από beats. Αυτή η ομοιότητα μετράται υπολογίζοντας το μέσο

όρο των στοιχείων κάθε διαγωνίου του BSM. Έπειτα, η συνάρτηση που προκύπτει αντιστρέφεται προκειμένου να δώσει μια συνάρτηση που εμφανίζει κορυφές στις διαγώνιες στις οποίες τα συστατικά τους εμφανίζουν μέγιστη ομοιότητα:

$$\begin{aligned} d_i &= mean(diag(BSM_i)) \\ d &= -d + max(|d|) \end{aligned} \tag{4.20}$$

όπου $diag(BSM_i)$ αντιστοιχεί στη διαγώνιο i του BSM.

Με σκοπό την επίλυση αμφιλεγόμενων περιπτώσεων, όπου οι ομοιότητες στο d δεν διακρίνονται εύκολα, δίνεται στις κορυφές του d ένα επιπλέον βάρος.

Ένας μεγάλος αριθμός από υποψήφια μέτρα λαμβάνεται υπόψιν στη συγκεκριμένη προσέγγιση. Περιλαμβάνει δίσημα και τρίσημα μέτρα, που συμβολίζονται με $c = 2$ και $c = 3$, απλά πολλαπλάσια δίσημων και τρίσημων μέτρων που συμβολίζονται με $c = 4$, 6 και 8 καθώς και σύνθετα μέτρα που συμβολίζονται ως $c = 5$, 7, 9 και 11. Με σκοπό να συνυπολογιστούν πολλαπλάσια για κάθε υποψήφιο μέτρο c, που αντιστοιχεί σε συναφή μοτίβα από beats τοποθετημένα σε διαφορετικά μουσικά μέτρα, ένα σταθμισμένο φίλτρο χτένα εφαρμόζεται στην συνάρτηση d. Η προκύπτουσα συνάρτηση που συμβολίζεται ως T_c δίνει περισσότερο βάρος σε μουσικά μέτρα που είναι κοντινά διαχωρισμένα ως εξής:

$$T_c = \sum_{p=1}^{lt} \frac{d(pc)}{1 - \frac{p-1}{lt}} \; c = 2, ..., 11 \tag{4.21}$$

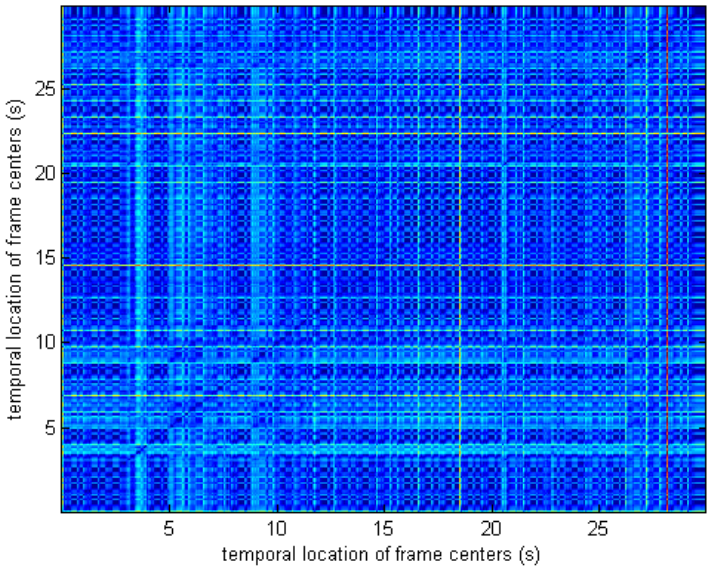
όπου το lt αντιστοιχεί στο $\lfloor nb/11 \rfloor$ και το nb είναι ο αριθμός των beats του μουσικού κομματιού.

Το ζητούμενο μέτρο c προσδιορίζεται από τη μεγαλύτερη κορυφή της συνάρτησης T_c.

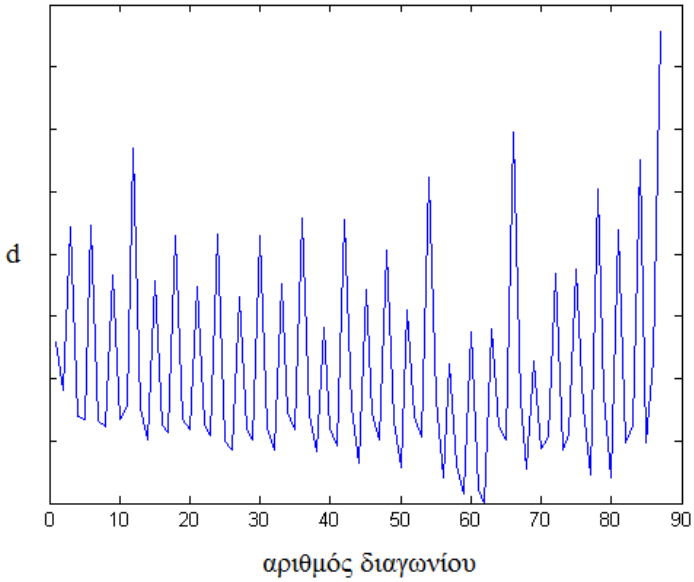
4.4.2 Αποτέλεσμα εκτέλεσης

Στον κώδικα της υλοποίησης χρησιμοποιούνται κάποιες συναρτήσεις του MIR Toolbox (βλέπε σχετικό παράρτημα) και συγκεκριμένα οι mirframe, mirspectrum και mirsimatrix. Οι γραφικές παραστάσεις που ακολουθούν (4.2, 4.3, 4.4), αφορούν το κομμάτι train15 από τα δεδομένα για την αξιολόγηση του beat tracking και του tempo estimation του MIREX 2006.

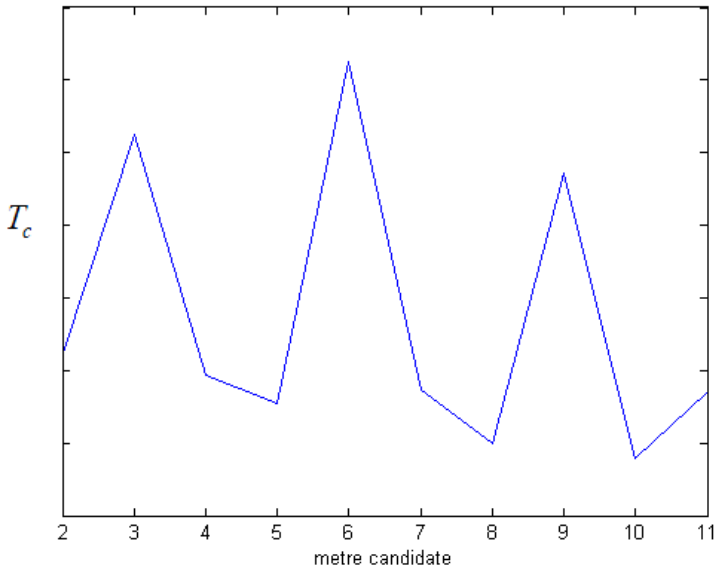
Για το συγκεκριμένο κομμάτι, θα ήταν μάλλον σωστότερη η επιλογή ενός μέτρου στα 3. Το λάθος όμως δεν είναι σημαντικό εφόσον το μέτρο που επιλέγεται είναι το διπλάσιο του σωστού και οφείλεται κυρίως στο γεγονός ότι ο αλγόριθμος εκτίμησης του tempo επέστρεψε ως πιο πιθανό ένα tempo που είναι διπλάσιο του πραγματικού. Άλλωστε, οι απλοί άνθρωποι αντιλαμβάνονται το ρυθμό σε διαφορετικές τιμές και η αυτόματη εκτίμηση του tempo είναι επιρρεπής σε λάθη όπου επιστρέφεται το διπλάσιο ή το μισό tempo αντί του πραγματικού.



Σχήμα 4.2: ASM του κομματιού train15.



Σχήμα 4.3: Η συνάρτηση d για το παράδειγμα train15. Είναι φανερό ότι η μέθοδος βρίσκει μεγάλη ομοιότητα σε μουσικά μέτρα που χωρίζονται από πολλαπλάσια των 3 beats.



Σχήμα 4.4: Παράδειγμα ανίχνευσης του μέτρου. Ξεχωρίζει η κορυφή στο υποψήφιο μέτρο c=6, που αντιστοιχεί σε ομαδοποίηση των 6 beats ανά μουσικό μέτρο.

Κεφάλαιο 5

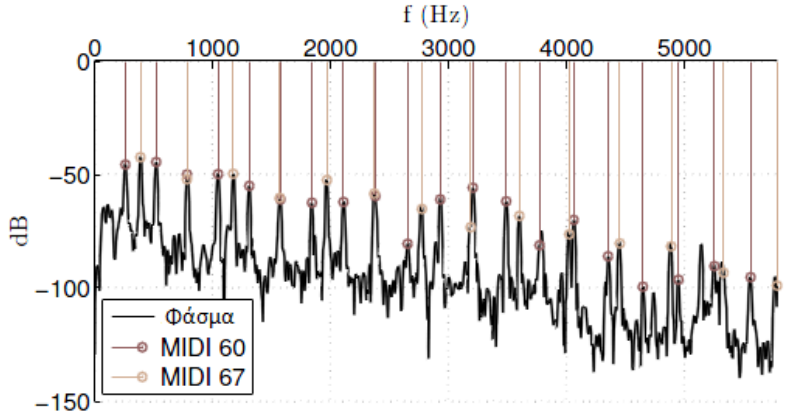
Μέθοδος εκτίμησης πολλαπλών τόνων

5.1 Περιγραφή προβλήματος

Οι μέθοδοι εκτίμησης ενός μόνο τόνου f_0 επιχειρούν να προσδιορίσουν τη θεμελιώδη συχνότητα σε σήματα με το πολύ μία αρμονική πηγή να ηχεί σε κάθε στιγμή. Μία μέθοδος ανίχνευσης πολλαπλών τόνων υποθέτει ότι μπορούν να υπάρχουν περισσότερες από μία αρμονικές πηγές στο σήμα εισόδου. Το πολυφωνικό πρόβλημα είναι πολύ πιο περίπλοκο από το μονοφωνικό κι έχει αποδειχθεί γενικά ότι οι μέθοδοι εκτίμησης ενός απλού τόνου δεν είναι κατάλληλες για την εκτίμηση πολλαπλών τόνων. Πρώτον, ο αριθμός των πηγών M πρέπει να προσδιοριστεί. Σε αντίθεση με την εκτίμηση ενός τόνου, όπου η μόνη απόφαση που πρέπει να γίνει είναι αν υπάρχει ήχος ή σιωπή, στην εκτίμηση πολλαπλών τόνων δεν γνωρίζουμε εκ των προτέρων πόσες νότες ηχούν ταυτόχρονα σε μία δεδομένη χρονική στιγμή. Το να βρεθεί ο βαθμός πολυφωνίας, είναι από μόνο του ένα δύσκολο πρόβλημα. Δεύτερον, υπάρχει αλληλεπίδραση μεταξύ των διαφορετικών πηγών κι έτσι προκαλείται επικάλυψη στα φάσματα των διαφορετικών θεμελιωδών συχνοτήτων, όπως φαίνεται στο Σχήμα 5.1. Για παράδειγμα, ας θεωρήσουμε δύο νότες σε σχέση οκτάβας που παίζονται ταυτόχρονα. Εφόσον η υψηλότερη νότα με τις αρμονικές της επικαλύπτεται πλήρως από τις αρμονικές της χαμηλότερης νότας, χρειάζεται κάποια άλλου είδους πληροφορία (όπως π.χ. η αναμενόμενη ενέργεια σε κάθε αρμονική για ένα συγκεκριμένο όργανο) για να συνάγουμε την παρουσία δύο νοτών. Οι ήχοι που παράγονται από διαφορετικά όργανα, διαφέρουν αρκετά όσον αφορά τα αρμονικά τους πρότυπα και επιπλέον ακόμα και για το ίδιο όργανο παρατηρούνται σημαντικές διαφορές από τις χαμηλές στις υψηλές νότες. Συγκριτικά πάντως με τη φωνή, οι τιμές f_0 στη μουσική είναι χρονικά πιο σταθερές. Είναι πιο δύσκολο να εντοπιστούν οι f_0 τεσσάρων ομιλητών που μιλούν ταυτόχρονα παρά να εκτελεστεί μουσική καταγραφή για ένα μουσικό κομμάτι τεσσάρων φωνών [99].

5.1.1 Αρμονική επικάλυψη

Κάθε ν-οστή αρμονική μίας αρμονικής πηγής α επικαλύπτει κάθε κ-οστή αρμονική μιας αρμονικής πηγής b όταν $f_\alpha = \frac{\mu}{\kappa} f_b$, όπου ν και κ είναι θετικοί ακέραιοι αριθμοί. Αυτό στη δυτική μουσική δεν αποτελεί ειδική περίπτωση, αλλά αντιθέτως είναι ένα φαινόμενο που εμφανίζεται



Σχήμα 5.1: Φάσμα δύο νοτών σε σχέση 5ης. Επικάλυψη των αρμονικών του Do 3 (MIDI 60) των οποίων η τάξη είναι πολλαπλάσιο του 3 με τις αρμονικές του Sol 3 (MIDI 67) των οποίων η τάξη είναι πολλαπλάσιο του 2. Σχήμα από το [5].

ιδιαίτερα συχνά.

Όταν δύο ήχοι υπερτίθενται, οι αντίστοιχες κυματομορφές αθροίζονται. Όταν υπάρχει αρμονική επικάλυψη, δύο απλά αρμονικά σήματα με την ίδια συχνότητα, αλλά διαφορετικά πλάτη και φάσεις προστίθενται. Αυτό παράγει ένα άλλο απλό αρμονικό σήμα με την ίδια συχνότητα αλλά διαφορετικό πλάτος το οποίο εξαρτάται από τη διαφορά στη φάση των επιμέρους σημάτων.

5.2 Σχετικές εργασίες

Πολλές διαφορετικές τεχνικές έχουν προταθεί για την εκτίμηση των πολλαπλών τόνων. Οι διαφορετικές τεχνικές μπορούν να κατηγοριοποιηθούν με βάση διάφορα κριτήρια, όπως η αναπαράσταση του σήματος που χρησιμοποιούν (αναπαράσταση στο πεδίο του χρόνου, STFT, wavelets, τράπεζες φίλτρων, κτλ), η γενικότητά τους (κάποιες τεχνικές έχουν σχεδιαστεί για ένα συγκεκριμένο όργανο ενώ κάποιες άλλες μπορούν να χρησιμοποιηθούν για να αναλύσουν γενικούς αρμονικούς ήχους), η ικανότητά τους να μοντελοποιήσουν μεταβαλλόμενα ηχοχρώματα ή ο τρόπος με τον οποίο εκτιμούν την αλληλεπίδραση μεταξύ των πηγών (επαναληπτική ή από κοινού εκτίμηση). Στις επόμενες υποενότητες παρουσιάζονται μερικές από τις διαφορετικές προσεγγίσεις που συναντώνται στη βιβλιογραφία. Η παρουσίαση αυτή δεν είναι σε καμία περίπτωση εκτενής, απλά παρουσιάζονται συνοπτικά κάποιες βασικές κατευθύνσεις. Επίσης σε αρκετές περιπτώσεις υπάρχει επικάλυψη μεταξύ των διαφορετικών προσεγγίσεων, με αποτέλεσμα κάποιες από τις εργασίες που έχει θεωρηθεί ότι υπάγονται σε μία προσέγγιση να έχουν στοιχεία και από άλλες προσεγγίσεις.

5.2.1 Μέθοδοι επεξεργασίας σημάτων

Αυτές οι μέθοδοι προσπαθούν να ενισχύσουν και να εξάγουν τις θεμελιώδεις συχνότητες εφαρμόζοντας μετασχηματισμούς επεξεργασίας σημάτων στο σήμα εισόδου. Συνεπώς, κύριο ρόλο για την εκτίμηση των συχνοτήτων σε αυτές τις προσεγγίσεις παίζει ο μετασχηματισμός του σήματος που επιλέγεται τελικά (όπως π.χ. η συνάρτηση αυτοσυσχέτισης στο πεδίο του χρόνου, cepstrum, φασματική αυτοσυσχέτιση, συστοιχία φίλτρων). Οι θεμελιώδεις συχνότητες προκύ-

πτουν έπειτα από κατάλληλη επεξεργασία της μετασχηματισμένης αναπαράστασης του σήματος. Ενδεικτικά αναφέρονται οι εργασίες των Tolonen και Karjalainen [37], Peeters [81] και Zhou et al. [100].

5.2.2 Μέθοδοι επαναληπτικής ακύρωσης

Στην περίπτωση ενός πολυφωνικού μείγματος είναι πολλές φορές πιο εύκολο να βρεθεί η πιο πιθανή υποψήφια συχνότητα μεταξύ όλων όσες είναι παρούσες, χρησιμοποιώντας για παράδειγμα ένα ενεργειακό κριτήριο. Εάν αυτή η συχνότητα αφαιρεθεί από το μείγμα, το υπόλοιπο θεωρητικά περιέχει πλέον μία νότα λιγότερη και μπορούμε εκ νέου να ψάξουμε την πιο πιθανή υποψήφια συχνότητα μεταξύ αυτών που απομένουν. Αυτή η προσέγγιση απαιτεί την υλοποίηση τεσσάρων εργασιών: την επιλογή της πιο πιθανής υποψήφιας συχνότητας από το σήμα, την εκτίμηση της συνεισφοράς της, την αφαίρεσή της από το σήμα και τον υπολογισμό της συνθήκης τερματισμού. Η επιλογή της πιο πιθανής υποψήφιας συχνότητας στηρίζεται συχνά σε ενεργειακά κριτήρια, που αντιστοιχούν μερικές φορές σε μια μέθοδο εκτίμησης που έχει αναπτυχθεί για το μονοφωνικό πρόβλημα. Για παράδειγμα το μέγιστο του φασματικού γινομένου (3.9) είναι μια καλή υποψήφια συχνότητα. Από τη στιγμή που έχει βρεθεί η υποψήφια συχνότητα, η εκτίμηση του σήματος που της αντιστοιχεί και η αφαίρεσή της από το σήμα του μείγματος γίνεται με προσεγγιστικό τρόπο, καθώς η φασματική επικάλυψη μεταξύ των νοτών δεν επιτρέπει τον πλήρη διαχωρισμό της συνεισφοράς της κάθε μίας. Μία λύση είναι η εκμετάλλευση της πληροφορίας που υπάρχει στη φασματική περιβάλλουσα, καμπύλη που συνδέει στον τομέα της συχνότητας τα πλάτη ή τις ενέργειες των αρμονικών μίας νότας. Μία νότα μπορεί να χαρακτηριστεί όχι μόνο συναρτήσει της ενέργειας των αρμονικών της, αλλά και σε σχέση με κάποια όρια, κάποιους περιορισμούς στις σχετικές ενέργειες μεταξύ των αρμονικών. Με αυτόν τον τρόπο, σε περίπτωση φασματικής επικάλυψης, το πλάτος ή η ενέργεια μιας αρμονικής μπορεί να καθοριστεί σαν μια μέση ή ενδιάμεση τιμή των πλατών των γειτονικών αρμονικών. Τέλος, η συνθήκη τερματισμού, είναι γενικά δύσκολο να υπολογιστεί και περιλαμβάνει ένα όριο στην ενέργεια ή στο λόγο σήματος προς θόρυβο για το σήμα που έχει απομείνει. Σχετικές εργασίες που ακολουθούν μία τέτοιου είδους προσέγγιση είναι οι [101], [80], [102] και [103].

5.2.3 Μέθοδοι από κοινού εκτίμησης

Αυτές οι μέθοδοι συγκρίνουν έναν αριθμό από πιθανές υποθέσεις που αποτελούνται από συνδυασμούς f_0, ώστε να επιλέξουν τον καλύτερο. Η μεθοδολογία συνίσταται στη χρήση φίλτρων τύπου χτένας για την καταστολή των αρμονικών που ακολουθούν την αρμονική κατανομή που αντιστοιχεί στις θεμελιώδεις συχνότητες που έχει υποτεθεί ότι υπάρχουν στο συνδυασμό και τον υπολογισμό της ενέργειας του σήματος που απομένει. Η ενέργεια αυτή θα είναι ελάχιστη στην περίπτωση που οι θεμελιώδεις συχνότητες έχουν επιλεγεί σωστά. Η εκτίμηση των θεμελιωδών συχνοτήτων με αυτόν τον τρόπο, είναι πιο αποτελεσματική από τις επαναληπτικές μεθόδους στις οποίες η επιλογή της υποψήφιας συχνότητας και η αφαίρεσή της από το φάσμα είναι προσεγγιστική. Όμως, οι μέθοδοι από κοινού εκτίμησης δεν είναι εφαρμόσιμες υπό την παρουσία υψηλού βαθμού πολυφωνίας: σε αυτήν την περίπτωση ο αριθμός των συνδυασμών των θεμελιωδών συχνοτήτων που πρέπει να ελεγχθούν είναι αυξημένος. Για την ακρίβεια, έστω ότι υπάρχουν N πιθανές νότες και μέγιστη τιμή πολυφωνίας P. Η επαναληπτική εκτίμηση θα αναζητά μία νότα μεταξύ των N σε κάθε επανάληψη, συνεπώς θα έχουμε N P υπολογισμούς το μέγιστο. Σε μια μέθοδο από κοινού εκτίμησης θα πρέπει να ελεγχθούν όλοι οι συνδυασμοί ανά 1,2,....,P νότες μεταξύ των N, δηλαδή $\sum_{p=0}^{P}\binom{N}{p}$. Μία τέτοια πολυπλοκότητα (2^N στην περίπτωση που P=N) παραμένει υποφερτή για μικρό βαθμό πολυφωνίας αλλά δεν είναι ρεαλιστική στην περίπτωση της

μουσικής εν γένει. Το πλαίσιο των μπαεσιανών (bayesian) προσεγγίσεων προσφέρει τεχνικές πιο αποδοτικές για τη σύγκλιση προς τη λύση. Σχετικές εργασίες είναι οι [104], [105] και [5].

5.2.4 Μπαεσιανά μοντέλα

Πρόσφατα, η εισαγωγή μπαεσιανών προσεγγίσεων για την αυτόματη καταγραφή μουσικής επέτρεψε να δοθεί στο πρόβλημα ένα θεωρητικό στατιστικό πλαίσιο καθώς και να διευρυνθούν οι δυνατότητες μοντελοποίησης. Το σήμα περιγράφεται με έναν ενοποιημένο και συστηματικό τρόπο από τυχαίες μεταβλητές που αντιπροσωπεύουν κάθε στοιχείο προς μοντελοποίηση: την κατανομή των συχνοτήτων, τα πλάτη, τον θόρυβο, αλλά ακόμη και την πολυφωνία, τις νότες, τα όργανα, την τονικότητα, κτλ. Αυτές οι τυχαίες μεταβλητές είναι αλληλένδετες κι έτσι συνθέτουν ένα δίκτυο που χαρακτηρίζεται από την επιλογή των τυχαίων κατανομών τους. Το πρόβλημα επομένως λύνεται μέσω της θεωρίας της μπαεσιανής εκτίμησης, με την εκτίμηση των βέλτιστων μοντέλων και παραμέτρων δεδομένου του σήματος που παρατηρείται. Σχηματικώς, εφόσον μοντελοποιούμε το παρατηρούμενο σήμα x με ένα στατιστικό μοντέλο ϑ, η επίλυση του προβλήματος συνίσταται στην εύρεση των πιο πιθανών παραμέτρων, δηλαδή στη μεγιστοποίηση της a posteriori πιθανότητας $p(\vartheta \mid x)$ ως προς το ϑ. Εφαρμόζοντας τον κανόνα του Bayes, αυτή η πιθανότητα είναι ανάλογη του $p(x \mid \vartheta)p(\vartheta)$, γινόμενο της πιθανοφάνειας $p(x \mid \vartheta)$ των δεδομένων και της a priori πιθανότητας $p(\vartheta)$. Αυτές οι δύο συναρτήσεις είναι εκείνες ακριβώς που έχουν επιλεγεί για τη μοντελοποίηση του σήματος: η a priori πιθανότητα ερμηνεύει την γνώση που έχουμε για τις παραμέτρους - όπως την κατανομή των πλατών, τις φάσεις ή τον αριθμό των συνιστωσών - ενώ η πιθανοφάνεια θέτει τη σχέση μεταξύ των παραμέτρων και του σήματος - το σήμα είναι για παράδειγμα ένα άθροισμα ημιτονοειδών και θορύβου.

Η αυτόματη καταγραφή εμπλέκει ένα μεγάλο αριθμό παραμέτρων και αλληλένδετων μεταβλητών. Η μπαεσιανή προσέγγιση έχει το πλεονέκτημα ότι προσφέρει ένα αυστηρό στατιστικό πλαίσιο για την μοντελοποίηση του συνόλου του συστήματος. Τεχνικές επίλυσης όπως οι μέθοδοι Monte-Carlo είναι επομένως ιδιαίτερα αποτελεσματικές για την εκτίμηση των αγνώστων παραμέτρων λαμβάνοντας από κοινού υπόψιν όλες τις αλληλοεξαρτήσεις. Αναλυτική περιγραφή και χρήση τέτοιων τεχνικών περιέχεται στις εργασίες [106] και [107]. Μπαεσιανές προσεγγίσεις για τη μουσική καταγραφή προτείνονται στα [108], [109], [110] και [111].

5.2.5 Προσεγγίσεις με εκπαίδευση

Έχουν προταθεί διάφοροι αλγόριθμοι για την αυτόματη καταγραφή μουσικής που χρησιμοποιούν εκπαίδευση, επιβλεπόμενη ή μη.

Στο [112] συναντάται η χρήση κρυφών μαρκοβιανών μοντέλων (Hidden Markov Models). Η προσέγγιση στηρίζεται στη χρήση μιας τεχνικής εκτίμησης των θεμελιωδών συχνοτήτων με βάση ένα πλαίσιο του σήματος. Σε κάθε πλαίσιο υπολογίζονται οι πέντε πιο πιθανές θεμελιώδεις συχνότητες με τις αντίστοιχες πιθανότητές τους, καθώς και οι πέντε κυρίαρχες συχνότητες για τη φάση του attack με τις αντίστοιχες εντάσεις των attacks. Για κάθε νότα n και κάθε πλαίσιο t, κατασκευάζεται ένα διάνυσμα με τις παρατηρήσεις $o_{n,t}$ με βάση τις παραμέτρους του πλαισίου για να χρησιμοποιηθεί σε ένα HMM. Το διάνυσμα αποτελείται από τρεις παραμέτρους: το χάσμα μεταξύ της θεμελιώδους συχνότητας που βρέθηκε και της θεωρητικής, την πιθανότητα της θεμελιώδους συχνότητας και την τελική ένταση ενός attack του οποίου η θεμελιώδης συχνότητα είναι κοντά στο n. Το HMM είναι τριών καταστάσεων: η πρώτη κατάσταση αντιστοιχεί στο attack της νότας, η δεύτερη στο sustain της νότας και η τρίτη στον θόρυβο ενώ υπάρχουν

περιορισμοί στις δυνατές μεταβάσεις μεταξύ των καταστάσεων. Η πιθανοφάνεια των παρατηρήσεων μοντελοποιείται με ένα GMM (Gaussian Mixture Model, μοντέλο μείγματος γκαουσιανών).

Έχουν υλοποιηθεί και άλλες προσεγγίσεις που βασίζονται στην παραμετροποίηση και την εκτίμηση των θεμελιωδών συχνοτήτων. Και το σύστημα [113] υιοθετεί την αποδόμηση του προβλήματος σε δύο φάσεις, τη συχνοτική ανάλυση η οποία ακολουθείται από χρονική ανάλυση, όπου αυτή τη φορά η παραμετροποίηση γίνεται με SVM's (Support Vector Machines, μηχανές διανυσμάτων υποστήριξης).

Το SONIC [1] είναι ένα σύστημα για την αυτόματη καταγραφή μουσικής για πιάνο που βασίζεται στη μέθοδο που περιγράφεται στο [58], η οποία είναι μια μέθοδος επιβλεπόμενης μάθησης που χρησιμοποιεί νευρωνικά δίκτυα.

Στο [114] χρησιμοποιούνται γενετικοί αλγόριθμοι για την αυτόματη καταγραφή πολυφωνικής μουσικής για πιάνο. Ένας γενετικός αλγόριθμος αποτελείται από ένα σύνολο από υποψήφιες λύσεις (ή χρωμοσώματα) τα οποία εξελίσσονται μέσω κληρονομικότητας, επιλογής, μετάλλαξης και διασταύρωσης μέχρι ένα κριτήριο τερματισμού. Σε κάθε γενιά, αποτιμάται μία αντικειμενική συνάρτηση (fitness function) για κάθε χρωμόσωμα, και τα καλύτερα χρωμοσώματα επιλέγονται για να συνεχίσουν να εξελίσσονται. Τελικά, το καλύτερο χρωμόσωμα επιλέγεται ως λύση. Στη μέθοδο που προτείνεται στο [114], κάθε χρωμόσωμα αντιστοιχεί σε μία ακολουθία από γεγονότα νοτών, όπου κάθε νότα έχει pitch, onset, διάρκεια και ένταση. Η αρχικοποίηση του πληθυσμού βασίζεται στις παρατηρούμενες STFT κορυφές. Η αντικειμενική συνάρτηση για ένα χρωμόσωμα προκύπτει από τη σύγκριση του αρχικού STFT με το STFT από συντεθειμένες εκδοχές των χρωμοσωμάτων, δεδομένου ενός οργάνου. Η μέθοδος απαιτεί την a priori γνώση του οργάνου που πρέπει να συντεθεί.

Άλλες προσεγγίσεις εφαρμόζουν εκπαίδευση απευθείας στο προς ανάλυση σήμα. Η μη αρνητική παραγοντοποίηση πίνακα (Non-negative Matrix Facorization, NMF) είναι μία σχετική τεχνική. Συνίσταται στην προσέγγιση ενός μη αρνητικού πίνακα Y σαν ένα γινόμενο δύο μη αρνητικών πινάκων W και H, με τέτοιο τρόπο ώστε να ελαχιστοποιηθεί το σφάλμα ανακατασκευής. Στα πλαίσια της αυτόματης καταγραφής μουσικής, ο Y είναι τα φασματικά δεδομένα, ο H αντιστοιχεί στα φασματικά μοντέλα (συναρτήσεις βάσης), και ο W περιέχει τα βάρη, δηλαδή την εξέλιξη της έντασης συναρτήσει του χρόνου. Σχετικές εργασίες είναι οι [50], [115], [116] και [117]. Η ανάλυση ανεξάρτητων συνιστωσών (Independent Component Analysis, ICA)[45], σχετίζεται στενά με την NMF. Στην ICA ένα μοντέλο σήματος εκφράζεται ως $x = Wh$, όπου τα x και h είναι ν-διάστατα πραγματικά διανύσματα και ο W είναι ένας αντιστρέψιμος πίνακας. Η ICA προσπαθεί να διαχωρίσει ένα πολυμεταβλητό σήμα σε προσθετικές υποσυνιστώσες υποθέτοντας ότι οι υποσυνιστώσες είναι μη γκαουσιανά σήματα και ότι είναι στατιστικά ανεξάρτητες η μία από την άλλη. Είναι μια ειδική περίπτωση τυφλού χωρισμού πηγής (blind source separation). Η διαφορά της σε σχέση με την NMF είναι οι διαφορετικοί περιορισμοί που τίθενται στους πίνακες που παραγοντοποιούνται. Μέθοδοι βασισμένες στην ανάλυση ανεξάρτητων συνιστωσών και στη μη αρνητική παραγοντοποίηση πίνακα συζητούνται στο [53].

[1] http://lgm.fri.uni-lj.si/SONIC/

5.2.6 Συστήματα μαυροπίνακα

Τα συστήματα μαυροπίνακα έχουν εφαρμοστεί και στην καταγραφή της μουσικής. Το όνομα στηρίζεται στην ιδέα ότι μία ομάδα εμπειρογνωμόνων συγκεντρώνεται σε ένα δωμάτιο με έναν μαυροπίνακα. Οι εμπειρογνώμονες συνεργάζονται για την επίλυση ενός προβλήματος χρησιμοποιώντας το μαυροπίνακα ως κοινό χώρο εργασίας. Κάθε εμπειρογνώμονας έχει την ευκαιρία να συνεισφέρει στην επίλυση του προβλήματος. Η διαδικασία συνεχίζεται μέχρι να επιλυθεί το πρόβλημα, δηλαδή ουσιαστικά μέχρις ότου οι εμπειρογνώμονες να είναι ικανοποιημένοι με τις υποθέσεις που βρίσκονται στο μαυροπίνακα δεδομένου ενός περιθωρίου λάθους. Μία βασική προϋπόθεση σε αυτή τη μέθοδο είναι ότι κάθε εμπειρογνώμονας έχει εξειδικευμένη γνώση που είναι συμπληρωματική της γνώσης των άλλων συμμετεχόντων. Επειδή κάθε εμπειρογνώμονας θεωρείται ότι συμπληρώνει τους άλλους, δεν δημιουργείται κλίμα ανταγωνισμού. Υπάρχει και ένας δρομολογητής που καθορίζει τη σειρά με την οποία οι εμπειρογνώμονες επιτρέπεται να δράσουν.

Στη γενική αρχιτεκτονική μαυροπίνακα που χρησιμοποιείται στην καταγραφή μουσικής, η ιεραρχία στο μαυροπίνακα διαμορφώνεται με αυξανόμενη αφαιρετικότητα, περνώντας από τα δεδομένα εισόδου στο χαμηλότερο επίπεδο, στις νότες σε ένα μεσαίο επίπεδο και καταλήγοντας στις συγχορδίες. Συστήματα μαυροπίνακα για την αυτόματη καταγραφή μουσικής έχουν προταθεί στα [118], [119] και [120].

5.3 Πρόταση μιας νέας μεθόδου εκτίμησης πολλαπλών τόνων

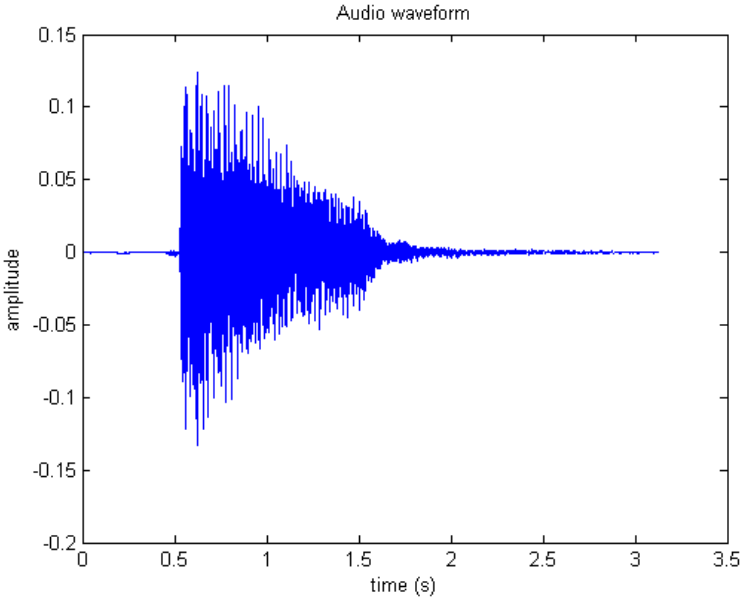
Στα πλαίσια αυτής της διπλωματικής εργασίας αναπτύχθηκε μία μέθοδος εκτίμησης πολλαπλών τόνων που παράγονται στο πιάνο. Ο αλγόριθμος που έχουμε αναπτύξει, σχετίζεται με τις μεθόδους επεξεργασίας σημάτων, μιας και κάνει ανάλυση φάσματος σε ορισμένες συχνότητες και εκτίμηση παραμέτρων ενός μοντέλου. Ένα κοινό της μεθόδου μας με τις μεθόδους επαναληπτικής ακύρωσης και από κοινού εκτίμησης, είναι ότι αφαιρώντας νότες από το φάσμα (όχι μία-μία όπως στις μεθόδους επαναληπτικής ακύρωσης αλλά ολόκληρους συνδυασμούς όπως στις μεθόδους από κοινού εκτίμησης, ξεκινώντας από συνδυασμούς μίας μόνο νότας και συνεχίζοντας με συνδυασμούς περισσότερων νοτών) ελέγχουμε το σήμα που απομένει. Όπως και στις μεθόδους επαναληπτικής ακύρωσης προσπαθούμε να προσδιορίσουμε και να εκμεταλλευτούμε τη σχέση που συνδέει στον τομέα της συχνότητας τα πλάτη των αρμονικών μιας νότας, χωρίς βέβαια αυτό να αποτελεί καθοριστικό βήμα της μεθόδου μας. Η σημαντική διαφορά μας σε σχέση με αυτές τις μεθόδους είναι ότι σε κάθε περίπτωση κάνουμε μία υπόθεση ως προς το βαθμό πολυφωνίας και τις νότες και παίρνουμε αυτή την υπόθεση ως δεδομένη, οπότε μέσα σε ένα συνδυασμό δεν χρειάζεται να ξεχωρίσουμε τη συνεισφορά κάθε μίας νότας. Αυτό είναι και το κοινό σημείο μας με τις μεθόδους από κοινού εκτίμησης. Όλα αυτά εξηγούνται με περισσότερες λεπτομέρειες στις επόμενες ενότητες όπου περιγράφεται ο αλγόριθμος που προτείνουμε για την εύρεση πολλαπλών τόνων παιγμένων στο πιάνο.

Ο διακριτός μετασχηματισμός Fourier του σήματος υπολογίζεται στις συγκεκριμένες συχνότητες που αντιστοιχούν σε νότες του πιάνου. Ο βαθμός πολυφωνίας δεν θεωρείται γνωστός και έτσι ο αλγόριθμος που προτείνεται έχει στόχο αφενός να προσδιορίσει τον βαθμό της πολυφωνίας (ο οποίος για λόγους απλότητας στα πλαίσια της διπλωματικής έχει περιοριστεί σε μέγιστη τιμή ίση με 3 αλλά μπορεί να επεκταθεί άμεσα και σε μεγαλύτερες τιμές) κι αφετέρου να βρει τις θεμελιώδεις συχνότητες. Αυτό γίνεται εξετάζοντας ποιες μπορεί να είναι οι υποψήφιες συχνό-

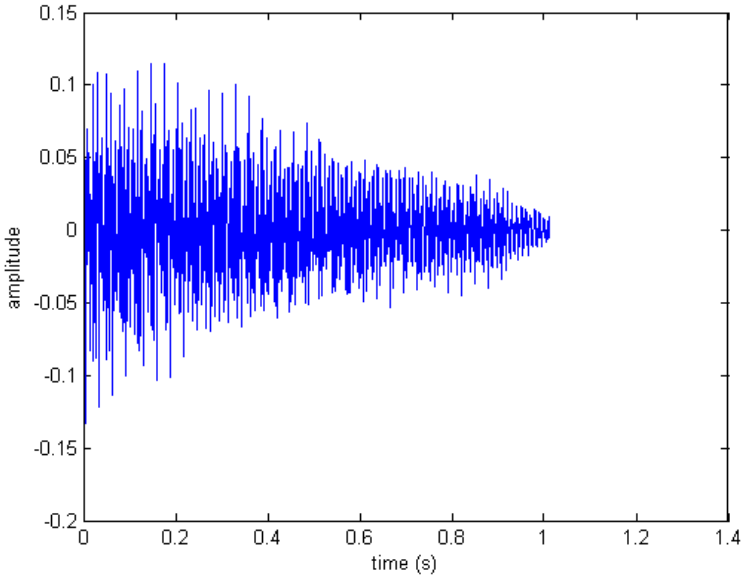
τητες με δεδομένο κάθε φορά βαθμό πολυφωνίας κι επιλέγοντας έπειτα την επικρατέστερη από όλες τις υποψήφιες λύσεις. Διαδοχικά ξεκινάμε από μία μονοφωνική υπόθεση και συνεχίζουμε με εύλογες υποθέσεις υψηλότερου βαθμού πολυφωνίας. Οι υποψήφιοι συνδυασμοί (που μπορούν να αποτελούνται από μία, δύο ή τρεις νότες) επιλέγονται με κριτήριο το κατά πόσο ερμηνεύουν το παρατηρούμενο φάσμα και κατά πόσο πληρούν κάποια χαρακτηριστικά που προκύπτουν με μία διαδικασία εκπαίδευσης. Τελικά επικρατέστερη θεωρείται η λύση εκείνη η οποία επαληθεύει στο μεγαλύτερο βαθμό ένα απλό ημιτονοειδές μοντέλο με την έννοια του ελάχιστου τετραγωνικού σφάλματος. Εφαρμόζονται διάφορα κριτήρια προκειμένου να επιλεγούν κατάλληλα οι υποψήφιοι συνδυασμοί από το φάσμα και να περιοριστεί ο αριθμός τους. Τα αποτελέσματα που προέκυψαν ήταν ιδιαίτερα ικανοποιητικά.

5.3.1 Ανάγνωση εισόδου

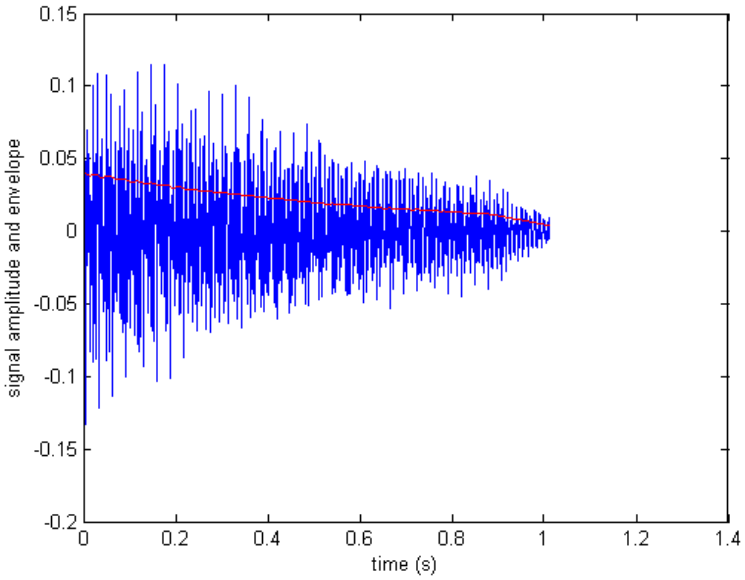
Ο αλγόριθμος δέχεται ως είσοδο είτε μία μόνο απομονωμένη νότα είτε μία συγχορδία με βαθμό πολυφωνίας 2 ή 3 (Σχήμα 5.2). Το σήμα εισόδου δεν χρησιμοποιείται αυτούσιο, αλλά κόβεται κατάλληλα με στόχο να μείνει ένα τμήμα του που να είναι απαλλαγμένο από θόρυβο και να έχει όσο το δυνατό πιο αρμονικά χαρακτηριστικά και ταυτόχρονα ικανή διάρκεια ώστε να μπορούν να μετρηθούν αυτά τα αρμονικά χαρακτηριστικά. Ορίζουμε σαν αρχή του τμήματος μία χρονική στιγμή που θεωρούμε ότι βρίσκεται πριν ή μέσα στο attack της νότας και είναι η χρονική στιγμή που η τιμή του σήματος ξεπερνάει το 5% της μέγιστης τιμής του. Ορίζουμε σαν τέλος του τμήματος τη χρονική στιγμή που η τιμή της περιβάλλουσας του σήματος γίνεται μικρότερη του 10% της τιμής που είχε στην αρχή του τμήματος κι έτσι όταν το πλάτος του σήματος γίνεται αμελητέο δεν το λαμβάνουμε πλέον υπόψιν. Σε κάθε περίπτωση φροντίζουμε να εξασφαλίσουμε ότι η διάρκεια του τμήματος που κρατάμε είναι τουλάχιστον 10 ms. Τέλος από αυτό το τμήμα αφαιρείται η μέση τιμή του (Σχήμα 5.3).



Σχήμα 5.2: Σήμα εισόδου. Το συγκεκριμένο σήμα είναι μία συγχορδία με δύο νότες και συγκεκριμένα τις MIDI νότες 44 και 53.



Σχήμα 5.3: Το σήμα του σχήματος 5.2 αφού κοπεί στην αρχή και το τέλος και αφαιρεθεί η μέση τιμή.



Σχήμα 5.4: Η περιβάλλουσα του σήματος του σχήματος 5.3 υπερτιθέμενη στο σήμα.

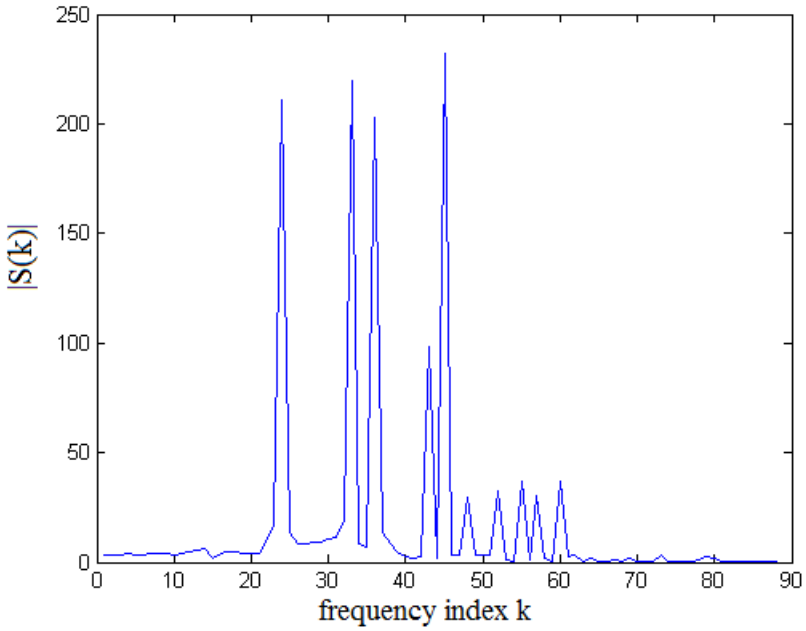
5.3.2 Υπολογισμός της απόκρισης συχνότητας

Το πιάνο είναι ένα όργανο με πλήκτρα οπότε οι νότες που μπορούν να παιχτούν σε αυτό είναι συγκεκριμένες και οι διαφορετικές συχνότητές τους μπορούν να θεωρηθούν εκ των προτέρων δεδομένες. Γι' αυτό το λόγο υπολογίζεται η απόκριση συχνότητας για τις δεδομένες και γνωστές εκ των προτέρων συχνότητες των πλήκτρων του πιάνου, αφού δεν έχει νόημα να εξετάζεται η

απόκριση σε συχνότητες οι οποίες δεν είναι δυνατό να παραχθούν από το πιάνο. Οι διαφορετικές δυνατές συχνότητες είναι 88 και απέχουν μεταξύ τους κατά $2^{\frac{1}{12}}$ (απόσταση ενός ημιτονίου). Το φάσμα υπολογίζεται σε εύρος από 27.5 Hz έως 4186.01 Hz, δηλαδή σε όλο το εύρος του πιάνου. Έτσι η σχέση που δίνει το φάσμα S είναι η :

$$S(k) = |\sum_{n} s(n) e^{-i2\pi \frac{f_k}{F_s} n}| \tag{5.1}$$

όπου $s(n)$ είναι το τμήμα του σήματος που επεξεργαζόμαστε, F_s η συχνότητα δειγματοληψίας και f_k η k-οστή συχνότητα, $k = 1, 2, ..., 88$, για $k = 1$, $f(1) = 27.5$ και δύο διαδοχικές συχνότητες διαφέρουν κατά $2^{\frac{1}{12}}$.



Σχήμα 5.5: Τα πλάτη του διακριτού μετασχηματισμού Fourier του σήματος του σχήματος 5.3 όπως προκύπτουν από τη σχέση 5.1. Ο δείκτης k της συχνότητας διατρέχει και τις 88 συχνότητες του πιάνου.

Η σχέση που συνδέει την πραγματική συχνότητα f_k με τον ακέραιο δείκτη k της συχνότητας είναι η:

$$f_k = 440(2^{\frac{k-49}{12}}). \tag{5.2}$$

Σημειώνεται ότι το k+20 αντιστοιχεί στον αριθμό MIDI της νότας με συχνότητα f_k.

5.3.3 Έλεγχος ύπαρξης μίας επικρατούσας συχνότητας

Προτού γίνει κάποια επιπλέον διερεύνηση, εξετάζεται κατά πόσο είναι εμφανής η ύπαρξης μίας και μόνο επικρατούσας κορυφής στο φάσμα. Γι' αυτό το σκοπό εξετάζεται η σχέση μεταξύ του μεγίστου του φάσματος και της διασποράς που υπάρχει στις υπόλοιπες θέσεις του φάσματος. Αν

η ισχύς στο μέγιστο είναι πολύ μεγάλη και η διασπορά στις υπόλοιπες θέσεις πολύ μικρή, τότε θεωρείται ότι υπάρχει μία μόνο επικρατούσα κορυφή και δεν είναι δυνατή η ανίχνευση κάποιας επιπλέον συχνότητας. Σε αυτή την περίπτωση δεν γίνεται τίποτα από όλα όσα περιγράφονται στη συνέχεια, θεωρείται ότι υπάρχει μόνο μία θεμελιώδης συχνότητα παρούσα (αυτή που αντιστοιχεί στο μέγιστο του φάσματος) κι επιστρέφεται άμεσα ως έξοδος. Η σχέση μεταξύ της ισχύος στο μέγιστο και της διασποράς στις υπόλοιπες θέσεις η οποία υποδεικνύει ότι μόνο μία συχνότητα είναι παρούσα, βρέθηκε με μία διαδικασία εκπαίδευσης.

5.3.4 Υποψήφιοι συνδυασμοί με μία νότα

Αναζητούνται υποψήφιες θεμελιώδεις συχνότητες με δεδομένο ότι ο βαθμός πολυφωνίας είναι 1. Γι' αυτό γίνεται μία κατάλληλη επιλογή από το φάσμα. Παρατηρείται ότι μία υποψήφια συχνότητα πρέπει να έχει αισθητή ισχύ και θεωρείται ότι αφήνει μόνο λευκό θόρυβο όταν αφαιρείται από το φάσμα. Επομένως η τυπική απόκλιση του απομένοντος φάσματος αναμένεται να είναι μικρή. Για κάθε μία από τις 88 συχνότητες του πιάνου υπολογίζονται τα ακόλουθα:

- η συνολική ισχύς της P ως το άθροισμα του πλάτους της μαζί με τα πλάτη όλων των σημείων του φάσματος που αντιστοιχούν στις θέσεις των αρμονικών της [2]. Μιας και το φάσμα έχει υπολογιστεί μέχρι και τη συχνότητα 4186.01 Hz και λαμβάνονται υπόψιν μόνο όσες αρμονικές είναι παρούσες στο φάσμα που έχουμε υπολογίσει, οι αρμονικές που έχουν συχνότητα υψηλότερη από 4186.01 Hz αγνοούνται. Επίσης σημειώνεται ότι στην πραγματικότητα αγνοούνται και άλλες αρμονικές, γιατί κάποιες αρμονικές δεν αντιστοιχούν σε υπαρκτή νότα του πιάνου συνεπώς δεν έχει υπολογιστεί η απόκρισή τους, δηλαδή δεν τους αντιστοιχεί καμία από τις 88 θέσεις στο φάσμα. Η απόλυτη ισχύς που υπολογίστηκε για κάθε κορυφή, διαιρείται με το άθροισμα της ισχύος σε κάθε δυνατή θέση του φάσματος κι έτσι παίρνουμε ένα ποσοστό που δείχνει πόσο μέρος της παρατηρούμενης ισχύος μπορεί να ερμηνεύσει η κάθε υποψήφια συχνότητα (σχετική ισχύς).

- η τυπική απόκλιση σ των εναπομείναντων πλατών στο φάσμα, αφού αφαιρεθεί η υποψήφια θεμελιώδης συχνότητα με τις αρμονικές της.

- ο αριθμός των αρμονικών n που συνεισφέρουν ουσιαστικά στη συνολική ισχύ της υποψήφιας συχνότητας.

- η σταθμισμένη από τα πλάτη μέση τιμή της κατανομής που ακολουθεί μία υποψήφια θεμελιώδης συχνότητα με τις αρμονικές της ως προς τους συχνοτικούς δείκτες k (κέντρο της κατανομής).

- η τοπική κυρτότητα στο φάσμα για την υποψήφια θεμελιώδη και κάθε αρμονική της:

$$c(k) = \frac{S(k)}{S(k-1) + S(k+1)}. \tag{5.3}$$

  Η μέση τιμή του c δίνει ένα μέτρο p της παρουσίας κορυφών μεταξύ της υποψήφιας θεμελιώδους και όλων των αρμονικών της.

Μετά από τον υπολογισμό των παραπάνω χαρακτηριστικών για όλες τις θέσεις στο φάσμα, γίνονται οι ακόλουθοι έλεγχοι:

[2] Στην παρούσα ενότητα, όπου γίνεται αναφορά σε ισχύ, εννοείται η απόλυτη τιμή του πλάτους.

- Οι τυπικές αποκλίσεις σ που αντιστοιχούν στις υποψήφιες συχνότητες ταξινομούνται σε αύξουσα σειρά, και ανιχνεύονται οι αισθητές μεταβολές ανάμεσα σε δύο διαδοχικές τιμές τυπικών αποκλίσεων. Υποψήφιες θεωρούνται όλες οι νότες πριν την πρώτη απότομη μεταβολή που ανιχνεύεται.

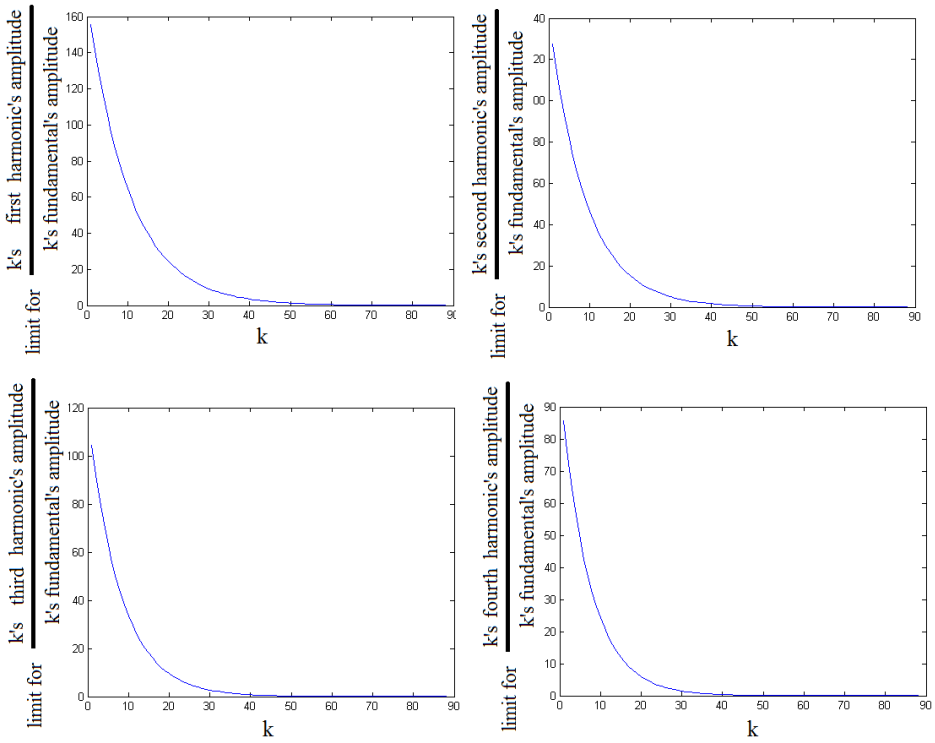
- Οι υποψήφιες νότες με χαμηλό p απορρίπτονται. Επιπλέον, οι υποψήφιες νότες που δεν πληρούν μία σχέση που συσχετίζει το κέντρο της κατανομής με τον αριθμό n απορρίπτονται επίσης. Η σχέση βρέθηκε έπειτα από εκπαίδευση σε πραγματικά δεδομένα.

Κατόπιν από όλες τις υποψήφιες συχνότητες που έχουν επιλεγεί με αυτόν τον τρόπο, απορρίπτονται κάποιες εφόσον δεν πληρούν ορισμένες συνθήκες που θέτουν όρια στο πλάτος των αρμονικών μιας θεμελιώδους που θεωρείται υποψήφια (κι εφόσον έχει υποτεθεί ότι δεν υπάρχουν άλλες θεμελιώδεις συχνότητες παρούσες που θα μπορούσαν να υπερτεθούν) και σε λόγους μεταξύ της υποψήφιας θεμελιώδους και των τριών πρώτων αρμονικών της. Με την εκπαίδευση σε πραγματικά δεδομένα βρήκαμε τις διάφορες τιμές αυτών των λόγων, και θέσαμε ένα όριο με βάση τις μέγιστες παρατηρούμενες τιμές. Το μέγιστο επιτρεπτό πλάτος για κάθε τάξης αρμονική υπολογίζεται σε σχέση με το ύψος και το πλάτος της θεμελιώδους. Κάποια παραδείγματα των σχέσεων που διαμορφώνονται φαίνονται στα Σχήματα 5.6 και 5.7.

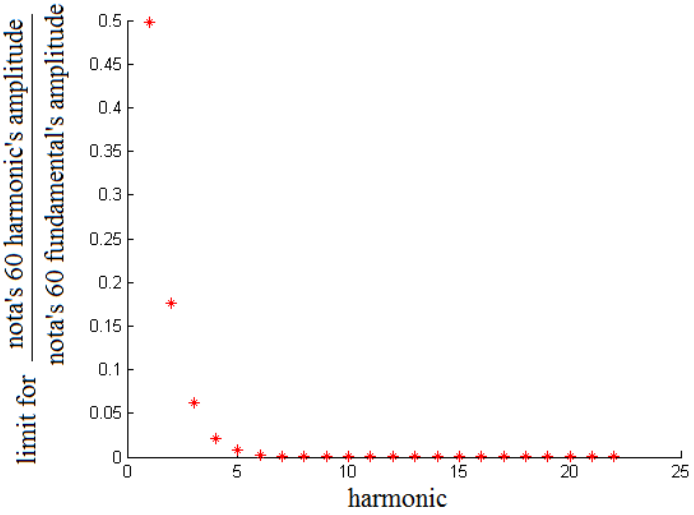
Τα όρια του πλάτους για τις αρμονικές προκύπτουν από μία διαδικασία εκπαίδευσης, συνδυάζοντας όλα τα διαθέσιμα δεδομένα εκπαίδευσης μαζί ώστε να προκύψει ένα γενικό μοντέλο τεσσάρων παραμέτρων. Αρχικά για κάθε νότα που υπάρχει στα δεδομένα μας και για κάθε αρμονική της που υπάρχει στο φάσμα, βρίσκουμε το μέγιστο λόγο ισχύος αρμονικής προς θεμελιώδη που εμφανίζεται μεταξύ όλων των δεδομένων για τη συγκεκριμένη νότα. Εξετάζοντας στη συνέχεια για κάθε αρμονική πώς μεταβάλλεται αυτός ο λόγος συναρτήσει του δείκτη κάθε θεμελιώδους συχνότητας του πιάνου (ο οποίος σχετίζεται με το λογάριθμο της θεμελιώδους συχνότητας όπως φαίνεται στη σχέση 5.2) βλέπουμε ότι η σχέση που συνδέει τον λογάριθμο του λόγου με το δείκτη της συχνότητας μπορεί να προσεγγιστεί με μία ευθεία γραμμή. Μέσω γραμμικής παλινδρόμησης υπολογίζεται αυτή η ευθεία γραμμή. Επειδή ψάχνουμε ένα άνω όριο για τον λόγο της ισχύος της αρμονικής προς τη θεμελιώδη, μετακινούμε τη γραμμή που βρίσκουμε με γραμμική παλινδρόμηση σε τέτοια θέση ώστε να βρίσκεται ανάμεσα στο μεσαίο και το μέγιστο σφάλμα απόκλισης από τα πραγματικά δεδομένα. Τα αποτελέσματα για την 1η αρμονική φαίνονται στο Σχήμα 5.8. Με αυτόν τον τρόπο, για κάθε τάξης αρμονική υπολογίζουμε δύο συντελεστές. Όμως μεταξύ τους οι αρμονικές θα πρέπει να έχουν μία συναρτησιακή σχέση, οπότε οι συντελεστές τους δεν μπορούν να είναι ασυσχέτιστοι. Γι' αυτό το λόγο, χρησιμοποιώντας πάλι γραμμική παλινδρόμηση, ταιριάζουμε δύο ευθείες σε αυτούς τους συντελεστές (μία ευθεία για κάθε τάξης συντελεστή) τις οποίες έπειτα μετακινούμε βάσει των μεγίστων σφαλμάτων απόκλισης όπως φαίνεται στο Σχήμα 5.9. Έτσι καταλήγουμε τελικά σε ένα μοντέλο που μπορεί να χαρακτηριστεί με τέσσερις μόνο παραμέτρους.

5.3.5 Επιλογή νοτών που θα χρησιμοποιηθούν σε συνδυασμούς ανά δύο και ανά τρία

Προτού προχωρήσουμε στην επιλογή συχνοτήτων υποψήφιων για να χρησιμοποιηθούν σε συνδυασμούς ανά δύο και ανά τρία, ελέγχουμε αν μπορούν να υπάρξουν τέτοιοι συνδυασμοί. Συγκεκριμένα, ελέγχουμε την πιο ψηλή κορυφή από τα διάφορα υπόλοιπα που αφήνουν στο φάσμα οι συνδυασμοί με μία συχνότητα που έχουμε θεωρήσει και αν η χαμηλότερη από αυτές τις κορυφές (προσέγγιση min-max) είναι αρκετά χαμηλή, θεωρούμε ότι δεν μπορούν να υπάρξουν συνδυασμοί με περισσότερες από μία νότες, οπότε οι μόνοι υποψήφιοι συνδυασμοί που λαμβάνονται υπόψιν



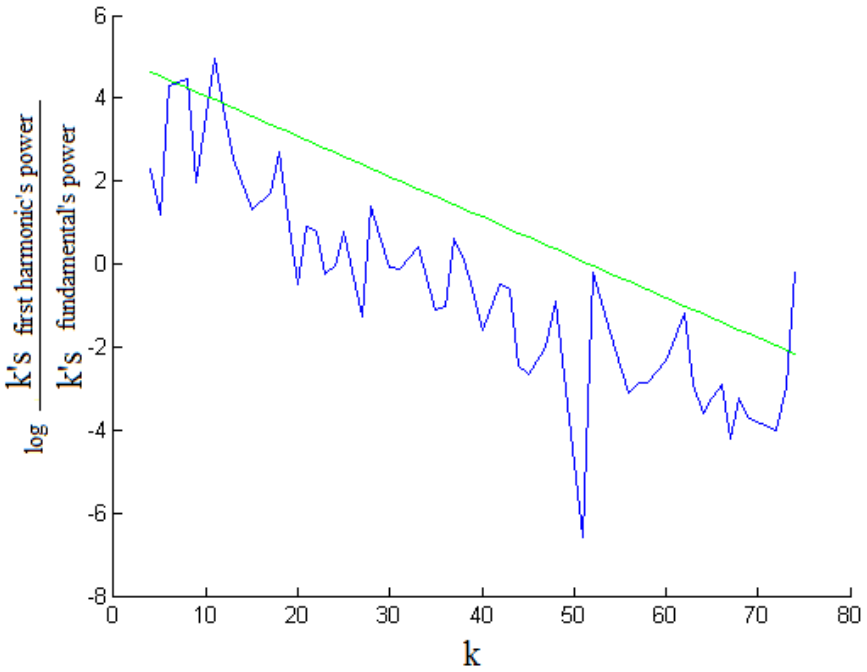
Σχήμα 5.6: Το όριο του πλάτους της 1ης, 2ης, 3ης και 4ης αρμονικής σε σχέση με το πλάτος της θεμελιώδους για κάθε μία από τις 88 θεμελιώδεις.



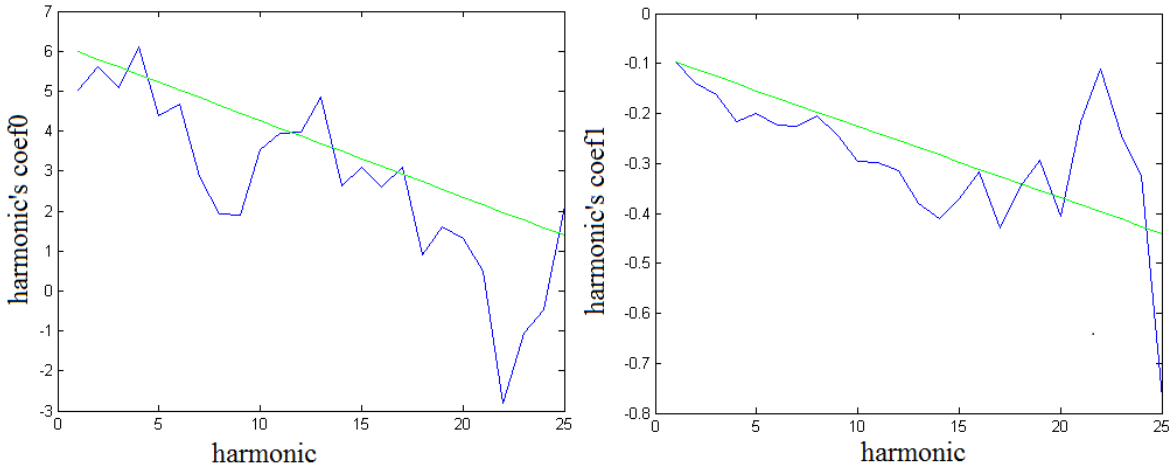
Σχήμα 5.7: Τα όρια των λόγων των πλατών των 22 πρώτων αρμονικών του φάσματος ως προς τη θεμελιώδη για τη νότα με $k = 60$.

είναι αυτοί με τη μία νότα που υπολογίσαμε προηγουμένως. Διαφορετικά η επιλογή των συχνοτήτων συνεχίζεται όπως περιγράφεται στη συνέχεια.

Για την εύρεση των συχνοτήτων που θα χρησιμοποιηθούν σε συνδυασμούς, υπολογίζεται όπως προηγουμένως το ποσοστό της συνολικής ισχύος που μπορεί να ερμηνεύσει η κάθε νότα.



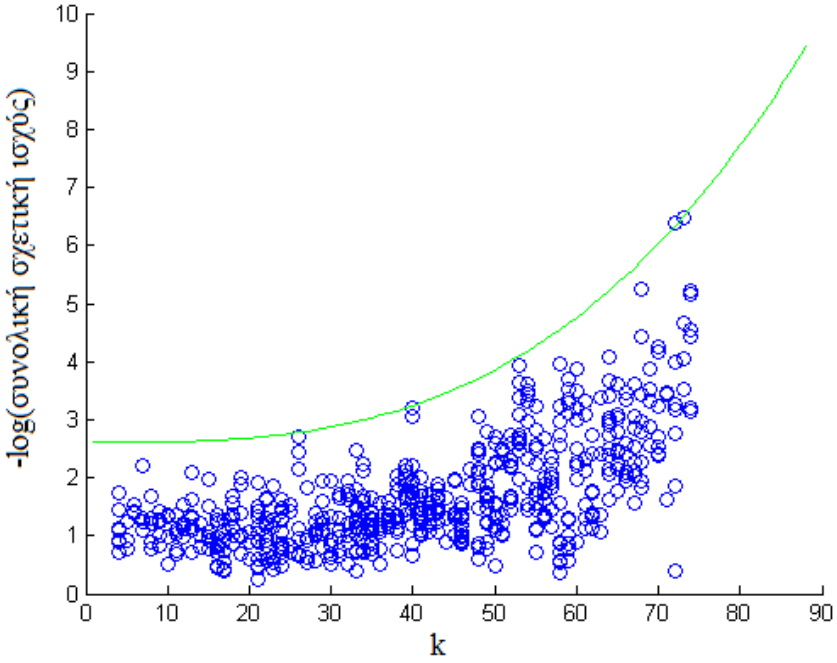
Σχήμα 5.8: Λογάριθμος του λόγου ισχύος της πρώτης αρμονικής ως προς τη θεμελιώδη συναρτήσει της νότας. Με μπλε χρώμα απεικονίζονται τα πραγματικά δεδομένα ενώ με πράσινο χρώμα φαίνεται η γραμμική προσέγγιση που επιλέγεται.



Σχήμα 5.9: Τιμή του συντελεστή μηδενικής τάξης (αριστερά) και πρώτης τάξης (δεξιά) συναρτήσει της τάξης της αρμονικής για τις 22 πρώτες αρμονικές του φάσματος. Με μπλε χρώμα απεικονίζονται οι συντελεστές που υπολογίστηκαν για κάθε αρμονική ενώ με πράσινο χρώμα φαίνεται η γραμμική προσέγγιση που επιλέγεται.

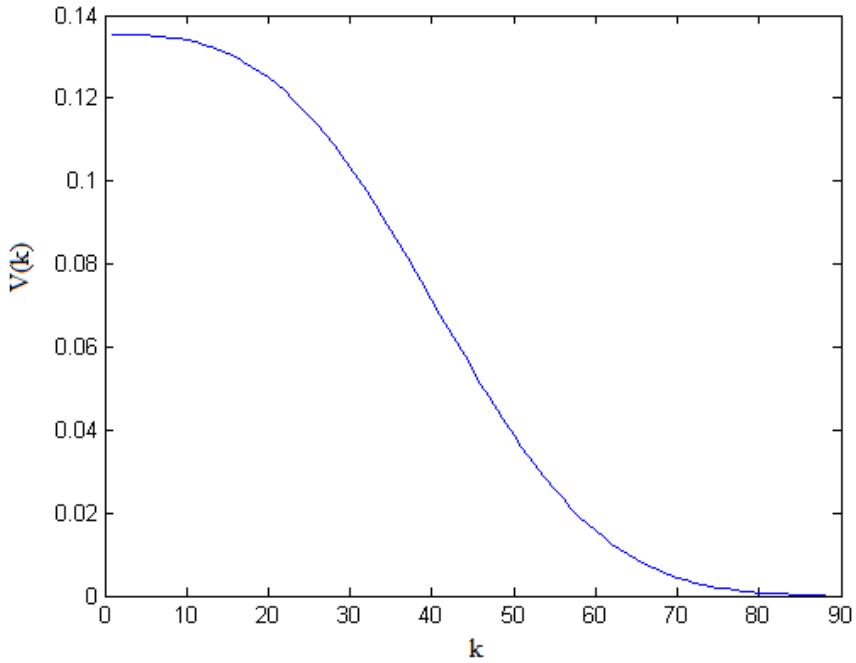
Στη συνέχεια, για να μπορούμε να εξάγουμε ασφαλή συμπεράσματα από αυτά τα ποσοστά, τα κανονικοποιούμε. Η κανονικοποίηση των τιμών γίνεται με τη βοήθεια μίας συνάρτησης που έχει προκύψει από πραγματικά δεδομένα σαν αποτέλεσμα εκπαίδευσης. Με την εκπαίδευση πάνω σε

πραγματικά δεδομένα συγχορδιών δύο και τριών νοτών βλέπουμε ποιες είναι οι πραγματικές τιμές για τα ποσοστά της ισχύος μιας νότας συναρτήσει της συχνότητας της νότας και έτσι εξάγουμε ένα όριο για το ποσοστό της ισχύος κάθε νότας συναρτήσει της συχνότητάς της (Σχήμα 5.10). Με αυτόν τον τρόπο λαμβάνεται υπόψιν ο διαφορετικός αριθμός των αρμονικών που συνυπολογίστηκαν για τη μέτρηση της ισχύος ανάλογα με τη συχνότητα καθώς και άλλες διαφορές που παρατηρήθηκαν στην ισχύ της απόκρισης ανάλογα με τη συχνότητα και εικάζεται ότι οφείλονται αφενός στα εγγενή χαρακτηριστικά του οργάνου κι αφετέρου στην ηχογράφηση (π.χ. επίδραση της συνάρτησης μεταφοράς του μικροφώνου). Η συνάρτηση που προκύπτει τελικά είναι εκθετικής μορφής και φαίνεται στο Σχήμα 5.11.



Σχήμα 5.10: Οι μπλε κύκλοι αντιπροσωπεύουν τον αρνητικό λογάριθμο της συνολικής σχετικής ισχύος για ένα σύνολο πραγματικών δεδομένων ενώ η πράσινη καμπύλη περιγράφεται από τη σχέση: $10^{-5}k^3 + 2.5 - \log(0.9), k \in [1, 88]$.

Σε πρώτο στάδιο, υποψήφιες να χρησιμοποιηθούν σε συνδυασμό είναι όλες οι συχνότητες των οποίων η σχετική κανονικοποιημένη ισχύς είναι μεγαλύτερη του 80% για τους συνδυασμούς ανά δύο και 90% για τους συνδυασμούς ανά τρία και που εμφανίζουν κορυφές στην κατανομή τους. Παρόλο που πλέον έχει υποτεθεί ότι δεν υπάρχει μόνο μία υποψήφια θεμελιώδης συχνότητα, οπότε εμφανίζεται κατά πάσα πιθανότητα μίξη των αρμονικών, υπάρχουν κάποιες σχέσεις που ισχύουν κατά αντιστοιχία με την περίπτωση που είχαμε μία μόνο υποψήφια συχνότητα. Για κάθε μία από τις 7 το πολύ πρώτες αρμονικές της υποψήφιας θεμελιώδους που υπάρχουν στο φάσμα έχει οριστεί ένα μέγιστο επιτρεπτό πλάτος, λαμβάνοντας υπόψιν την τάξη της αρμονικής, καθώς και το ύψος και το πλάτος της θεμελιώδους. Υποψήφιες συχνότητες των οποίων οι αρμονικές υπερβαίνουν αυτό το πλάτος, απορρίπτονται. Το όριο του πλάτους δεν είναι το ίδιο με αυτό που θεωρήσαμε προηγουμένως στην περίπτωση που υποθέσαμε ότι υπάρχει μόνο μία συχνότητα παρούσα, γιατί τώρα λαμβάνεται υπόψιν και η πιθανότητα της υπέρθεσης. Η εκπαίδευση πάντως



Σχήμα 5.11: Συνάρτηση κανονικοποίησης της σχετικής ισχύος για κάθε μία από τις 88 νότες του πιάνου. Η εξίσωση της συνάρτησης είναι η: $V(k) = e^{-10^{-5}k^3-2.5}, k \in [1, 88]$.

γίνεται με αντίστοιχο τρόπο αλλά πάνω σε διαφορετικά δεδομένα που αφορούν συγχορδίες με δύο και τρεις νότες και όχι μεμονωμένες νότες.

5.3.6 Υποψήφιοι συνδυασμοί με δύο νότες

Υποθέτοντας ότι ο βαθμός πολυφωνίας είναι δύο, δημιουργούνται όλοι οι δυνατοί συνδυασμοί ανά δύο από τις υποψήφιες για να χρησιμοποιηθούν σε συνδυασμό συχνότητες. Αυτή η προσέγγιση, όπου έχουμε υποθέσει ένα δεδομένο βαθμό πολυφωνίας, έχει το πλεονέκτημα ότι στην περίπτωση που οι δύο θεμελιώδεις έχουν κοινές αρμονικές, δεν καθίσταται αναγκαίο να βρεθεί το ποσοστό της συνεισφοράς κάθε θεμελιώδους στο παρατηρούμενο πλάτος μιας κοινής αρμονικής. Αρκεί σε κάθε συνδυασμό να συνυπολογιστούν μία φορά οι κοινές αρμονικές, όπως και γίνεται. Για κάθε συνδυασμό σε αντιστοιχία με την περίπτωση που είχαμε μόνο μία νότα, υπολογίζονται κάποια μεγέθη τα οποία στη συνέχεια θα χρησιμοποιηθούν για τον αποκλεισμό κάποιων συνδυασμών: η συνολική σχετική ισχύς P, ως το άθροισμα των πλατών των δύο υποψήφιων θεμελιωδών συχνοτήτων με τις αρμονικές τους (αθροίζοντας μόνο μία φορά το πλάτος των κοινών αρμονικών) διαιρεμένο με το άθροισμα της ισχύος σε όλες τις θέσεις του φάσματος, η τυπική απόκλιση σ στο φάσμα που απομένει αν αφαιρέσουμε τις νότες (θεμελιώδεις και αρμονικές) του εκάστοτε υποψήφιου συνδυασμού, οι αρμονικές n που έχουν ουσιαστική ισχύ και συνεισφέρουν στη συνολική ισχύ του συνδυασμού και το κέντρο της κατανομής της θεμελιώδους με τις αρμονικές της. Η διαλογή των συνδυασμών γίνεται αφενός με βάση την απόκλιση σ που εμφανίζεται στο φάσμα που απομένει αν αφαιρεθεί ο συνδυασμός από αυτό και αφετέρου σε δεύτερη φάση με βάση την ισχύ του συνδυασμού, εφόσον ένας "καλός" συνδυασμός πρέπει να έχει από μόνος του αρκετή ισχύ, ενώ αυτά που αφήνει όταν αφαιρεθεί να είναι θόρυβος. Για να είναι αμερόληπτη και πιο ουσιαστική η σύγκριση όσον αφορά την απόκλιση, γίνεται μεταξύ όσων συνδυασμών έχουν τον

ίδιο αριθμό αρμονικών. Για κάθε μήκος συνδυασμού:

- απορρίπτονται οι συνδυασμοί για τους οποίους η τυπική απόκλιση είναι μεγάλη και αυτοί που δεν ικανοποιούν μία σχέση που συσχετίζει το κέντρο της κατανομής με τον αριθμό n.
- γίνονται δεκτοί ως καλύτεροι υποψήφιοι οι συνδυασμοί που αντιστοιχούν σε χαμηλή τυπική απόκλιση σ.
- μεταξύ των καλύτερων υποψήφιων που επιλέχτηκαν στο προηγούμενο βήμα, απορρίπτονται αυτοί που έχουν χαμηλή ισχύ P.

Επίσης ελέγχεται αν οι συνδυασμοί που έχουν επιλεγεί πληρούν κάποια κριτήρια όσον αφορά τις αρμονικές. Πρώτα βρίσκεται πόσες αρμονικές είναι κοινές και για τις δύο συχνότητες και πόσες αρμονικές ανήκουν μόνο στη μία ή μόνο στην άλλη συχνότητα. Για τις αρμονικές που ανήκουν μόνο σε μία συχνότητα, ελέγχεται κατά πόσο το πλάτος τους είναι μικρότερο από το μέγιστο επιτρεπτό πλάτος που έχει οριστεί και για την περίπτωση που ήταν μόνο μία συχνότητα παρούσα. Για τις αρμονικές που ανήκουν και στις δύο υποψήφιες συχνότητες το μέγιστο επιτρεπτό πλάτος ορίζεται ως το άθροισμα των μεγίστων επιτρεπτών πλατών που προκύπτουν εξαιτίας κάθε μίας εκ των δύο συχνοτήτων και ελέγχεται κατά πόσο το πλάτος τους είναι μικρότερο από αυτό το νέο όριο του πλάτους. Σε κάθε περίπτωση εξασφαλίζεται ότι το όριο που θέτουμε για το πλάτος μιας αρμονικής δεν πέφτει κάτω από ένα κατώφλι. Αν το πλάτος κάποιας αρμονικής δεν πληροί αυτή τη συνθήκη, ο συνδυασμός απορρίπτεται. Τέλος γίνεται και ένας έλεγχος της κατανομής της ισχύος με κριτήριο τις κεντρικές ροπές της κατανομής των αρμονικών. Με μία διαδικασία εκπαίδευσης βρίσκονται για μία νότα οι περιορισμοί που ισχύουν για τις ροπές (π.χ. Σχήματα 5.12 και 5.13) και κατόπιν για κάθε μία από τις νότες του συνδυασμού μπορεί να ελεγχθεί αν οι ροπές που υπολογίζονται με βάση τις αρμονικές που ανήκουν μόνο σε αυτή ικανοποιούν τους περιορισμούς. Σε περίπτωση που κάποια από τις νότες του συνδυασμού δεν ικανοποιεί τους περιορισμούς, ο συνδυασμός απορρίπτεται.

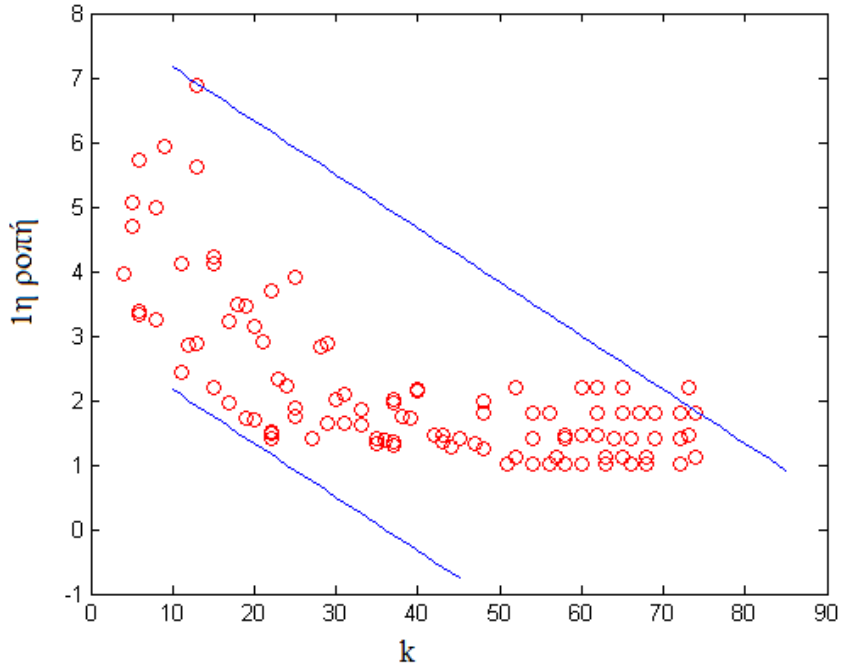
5.3.7 Υποψήφιοι συνδυασμοί με τρεις νότες

Όσα περιγράφτηκαν για τους υποψήφιους συνδυασμούς με δύο συχνότητες επεκτείνονται με ανάλογο τρόπο και στην περίπτωση υποψήφιων συνδυασμών με τρεις συχνότητες. Η διαφορά είναι ότι εκτός των ελέγχων που γίνονται στην περίπτωση των συνδυασμών ανά δύο, ένας συνδυασμός ανά τρία γίνεται δεκτός όταν η σχετική ισχύς του συνδυασμού ξεπερνάει ένα ελάχιστο όριο. Πλέον που έχουμε τρεις νότες η συνολική ισχύς του συνδυασμού είναι σίγουρα αισθητή (και εφόσον δεν εξετάζουμε συγχορδίες με μεγαλύτερο βαθμό πολυφωνίας) και είναι ασφαλές να τεθεί ένα τέτοιο όριο.

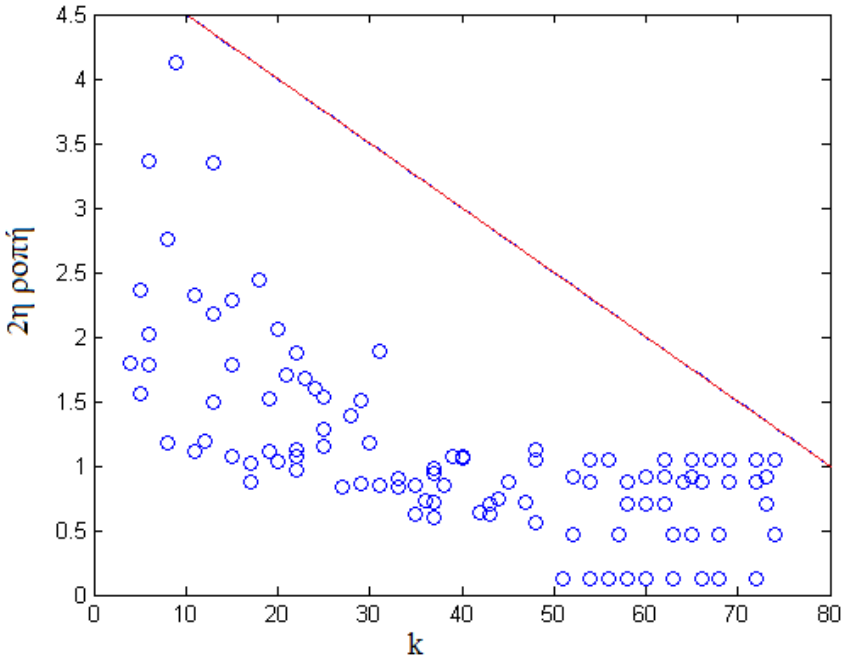
5.3.8 Ημιτονοειδές μοντέλο και υπολογισμός ενός σκορ για κάθε συνδυασμό

Η αξιολόγηση όλων των διαφορετικών υποψήφιων συνδυασμών γίνεται με κριτήριο το κατά πόσο επαληθεύουν ένα απλό ημιτονοειδές μοντέλο του σήματος, σύμφωνα με το οποίο το σήμα μπορεί να γραφτεί σαν ένας γραμμικός συνδυασμός ημιτόνων και συνημιτόνων:

$$\hat{s}(n) = \sum_{j=1}^{C} \{x_{j,1} A(n) \cos(2\pi \frac{f_j}{F_s} n) + x_{j,2} A(n) \sin(2\pi \frac{f_j}{F_s} n)\} \tag{5.4}$$



Σχήμα 5.12: Οι κύκλοι αντιπροσωπεύουν τις τιμές που παίρνει η πρώτη ροπή της κατανομής των αρμονικών μίας νότας σε ένα σύνολο δεδομένων που χρησιμοποιούμε για εκπαίδευση.



Σχήμα 5.13: Οι κύκλοι αντιπροσωπεύουν τις τιμές που παίρνει η δεύτερη ροπή της κατανομής των αρμονικών μίας νότας σε ένα σύνολο δεδομένων που χρησιμοποιούμε για εκπαίδευση.

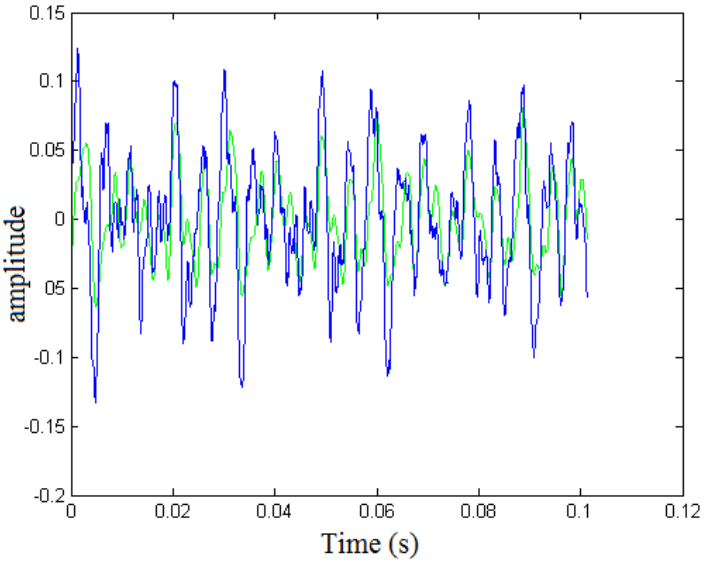
όπου C ο συνολικός αριθμός των διαφορετικών συχνοτήτων (θεμελιώδεις και αρμονικές τους), A η περιβάλλουσα του σήματος (στο πεδίο του χρόνου) που διαμορφώνει το πλάτος, f οι διαφορετικές συχνότητες που συνθέτουν το σήμα (θεμελιώδεις και αρμονικές) και F_s η συχνότητα δειγματοληψίας.

Το πλάτος του σήματος σε κάθε χρονική στιγμή είναι μεταβλητό, δεν μπορεί να θεωρηθεί σταθερό γιατί το χρονικό παράθυρο στο οποίο δουλεύουμε είναι αρκετά μεγάλο. Κανονικά λοιπόν θα έπρεπε να υπάρχει ένα μοντέλο για την εκτίμηση αυτού του μεταβλητού πλάτους. Για λόγους απλότητας όμως, έχει υποτεθεί ότι η περιβάλλουσα διαμορφώνει το πλάτος ανεξαρτήτως της συχνότητας κι έτσι με τη βοήθεια της περιβάλλουσας παίρνουμε μία εκτίμηση για το πώς μεταβάλλεται το πλάτος.

Το μοντέλο από το οποίο υπολογίζεται το συνθετικό σήμα είναι ένας απλός γραμμικός συνδυασμός (εξίσωση 5.4), του οποίου οι άγνωστοι παράμετροι $x_{j,1}, x_{j,2}$ μπορούν να προσδιοριστούν με τη μέθοδο των ελαχίστων τετραγώνων. Το σκορ κάθε υποψήφιου συνδυασμού δίνεται από την εξίσωση 5.5 η οποία συμπεριλαμβάνει το σφάλμα εκτίμησης του μοντέλου συν δύο όρους που "τιμωρούν" τους συνδυασμούς των οποίων τα μοντέλα έχουν πολλές παραμέτρους (έχουν θεωρηθεί πολλές συχνότητες) ώστε να δίνεται προτεραιότητα στα μοντέλα που κάνουν εκτίμηση με λιγότερες παραμέτρους.

$$\text{score} = \frac{\overline{(\hat{s}(n) - s(n))^2}}{\overline{s(n)^2}} + 0.01 \log_2(C) + 0.001 \log_2(K), \tag{5.5}$$

όπου K είναι ο βαθμός πολυφωνίας.
Μεταξύ όλων των υποψήφιων συνδυασμών που υπολογίστηκαν, επιλέγεται αυτός με το ελάχιστο σκορ.



Σχήμα 5.14: Το πραγματικό σήμα με μπλε χρώμα και το συνθετικό σήμα που προκύπτει με βάση το μοντέλο με πράσινο χρώμα.

Στο ημιτονοειδές μοντέλο, για κάθε υποψήφια θεμελιώδη συχνότητα προστίθεται και ένας

Πολυφωνία	Precision%	Recall%	F-measure%
1	84.48	95.15	89.50
2	79.82	78.38	79.09
3	83.37	72.49	77.55

Πίνακας 5.1: Επίδοση της προτεινόμενης μεθόδου στη βάση δεδομένων MAPS.

αριθμός από τις αρμονικές της. Εν γένει, για κάθε θεμελιώδη νότα κρατάμε τόσες αρμονικές όσες με την παρουσία τους επιδρούν αισθητά στη συνολική ισχύ για τη συγκεκριμένη νότα. Εξασφαλίζουμε ότι συγκρίνουμε τους συνδυασμούς επί ίσοις όροις παίρνοντας για όλους συγκρίσιμο αριθμό αρμονικών.

5.3.9 Αποτελέσματα

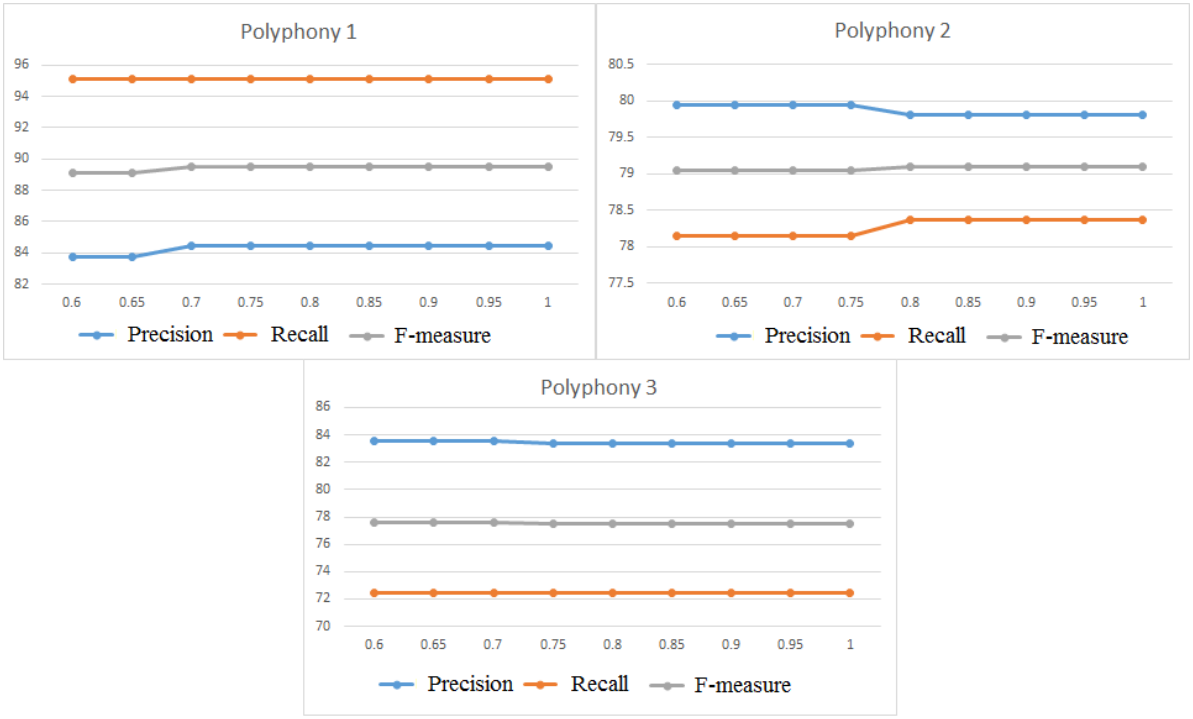
Ο αλγόριθμος που περιγράφεται υλοποιήθηκε στο MatLab. Σαν δεδομένα χρησιμοποιήσαμε 1028 δείγματα μεμονωμένων νοτών και συγχορδιών δύο ή τριών νοτών από τα σύνολα δειγμάτων ISOL, RAND, και UCHO της βάσης MAPS [5] για τα όργανα και τις συνθήκες ηχογράφησης που αντιστοιχούν στους κωδικούς ENSTDkCl και SptkBGCl. Οι θεμελιώδεις συχνότητες που περιέχονται σε αυτά τα δείγματα, εκτείνονται από το C1 στα 32.7032 Hz μέχρι και το A#6 στα 1864.66 Hz. Τα μισά από αυτά τα δείγματα χρησιμοποιήθηκαν για το training που έγινε και τα άλλα μισά δείγματα χρησιμοποιήθηκαν στο testing. Τα αποτελέσματα (precision, recall και F-measure) φαίνονται στον πίνακα 5.1.

Η προτεινόμενη μέθοδος είναι αρκετά απλή και έχει καλή επίδοση. Η χειρότερη επίδοση παρατηρείται για βαθμό πολυφωνίας 3, εξαιτίας του χαμηλού recall συγκριτικά με τις άλλες περιπτώσεις. Αυτό συμβαίνει επειδή μιας και δεν υπάρχει το περιθώριο να επιλεγούν περισσότερες από τρεις συχνότητες μέσα σε ένα συνδυασμό, κάποιες φορές "παραμερίζονται" συχνότητες οι οποίες υπάρχουν από άλλες που δεν είναι σωστές (αλλά συνήθως σχετίζονται αρμονικά με τις σωστές συχνότητες). Γενικά στις περισσότερες περιπτώσεις όπου δε βρέθηκαν σωστά οι συχνότητες, αντί αυτών βρίσκονται συχνότητες που σχετίζονται αρμονικά με τις σωστές. Ειδικά σε συνδυασμούς όπου συνυπάρχουν δύο νότες σε σχέση οκτάβας, όλες οι αρμονικές της δεύτερης νότας είναι αρμονικές και της πρώτης με αποτέλεσμα η μέθοδός μας συχνά να προτιμάει έναν λάθος συνδυασμό που έχει μόνο την πρώτη νότα, αφού αυτός θα εμπεριέχει ήδη τη συνεισφορά της δεύτερης λόγω της επικάλυψης των αρμονικών. Πιστεύουμε ότι αυτά τα προβλήματα θα μπορούσαν να περιοριστούν με κατάλληλη εκπαίδευση. Πέρα από τα αναμενόμενα όρια που πρέπει να έχουν οι αρμονικές στην απόκριση τα οποία αξιοποιούμε θα μπορούσε να μελετηθεί η απόκριση περαιτέρω και ενδεχομένως να γίνει εξαρχής μία βαθμολογία των συνδυασμών με βάση τις αποκρίσεις. Έπειτα θα μπορούσε να εξεταστεί κατά πόσο είναι δυνατόν να υπάρχει όντως ο συνδυασμός για τον οποίο το ημιτονοειδές μοντέλο έδωσε το βέλτιστο σκορ και σε κάποιες περιπτώσεις να επιλέγονται ως επικρατέστεροι κάποιοι συνδυασμοί οι οποίοι ναι μεν δε θα έχουν το βέλτιστο σκορ, όμως θα ερμηνεύουν καλύτερα το παρατηρούμενο φάσμα.

Η πιο χρονοβόρα διαδικασία, είναι η δοκιμή των διαφορετικών μοντέλων γι' αυτό είναι σημαντικό να περιοριστεί με κατάλληλη εκπαίδευση ο αριθμός τους ακόμα περισσότερο ώστε να μη χρειαστεί να δοκιμαστούν πολλά. Αυτή είναι μία κατεύθυνση στην οποία θα θέλαμε να εστιάσουμε μελλοντικά. Αν αντιμετωπιστεί καλύτερα αυτό το πρόβλημα, τότε θα είναι πιο εύκολο να επεκταθεί η μέθοδος και για μεγαλύτερους βαθμούς πολυφωνίας.

Σημειώνεται ότι η εκπαίδευση έχει σημαντικό ρόλο στη μέθοδό μας, όμως γίνεται για χαρακτηριστικά που είναι αναλλοίωτα ως προς το όργανο κι έτσι είναι αρκετά γενική.

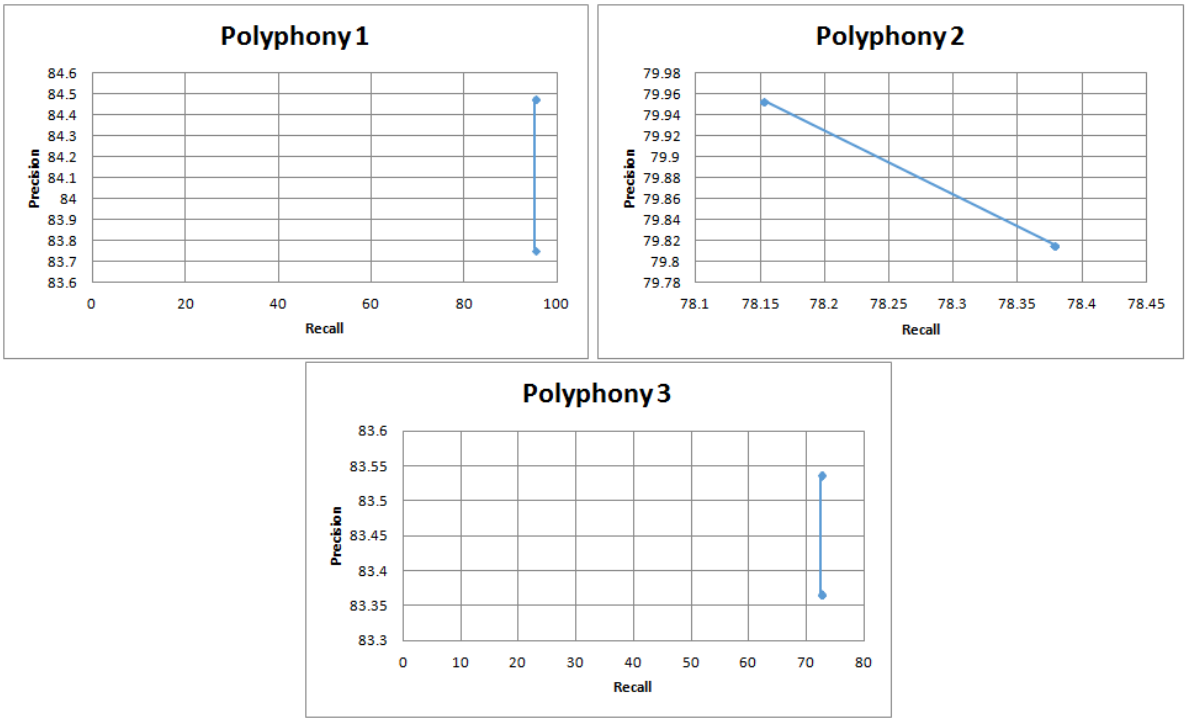
Επιπλέον, εξετάσαμε πώς επηρεάζονται το precision, το recall και το F-measure αν μεταβάλουμε την παράμετρο που εκφράζει το κατώφλι που πρέπει να ξεπερνάει η σχετική κανονικοποιημένη ισχύς μίας νότας ώστε να θεωρηθεί υποψήφια μέσα σε συνδυασμούς ανά δύο ή ανά τρία (ενότητα 5.3.5) ενώ σχεδιάστηκε και η precision-recall καμπύλη για τις ίδιες περιπτώσεις. Τα αποτελέσματα φαίνονται στα Σχήματα 5.15, 5.16, 5.17 και 5.18.



Σχήμα 5.15: Η μεταβολή των precision, recall και F-measure για βαθμούς πολυφωνίας 1,2 και 3 ως συνάρτηση του ποσοστού της συνολικής ισχύος που η κανονικοποιημένη ισχύς μιας νότας πρέπει να ξεπερνάει ώστε να θεωρηθεί υποψήφια μέσα σε συνδυασμούς ανά δύο.

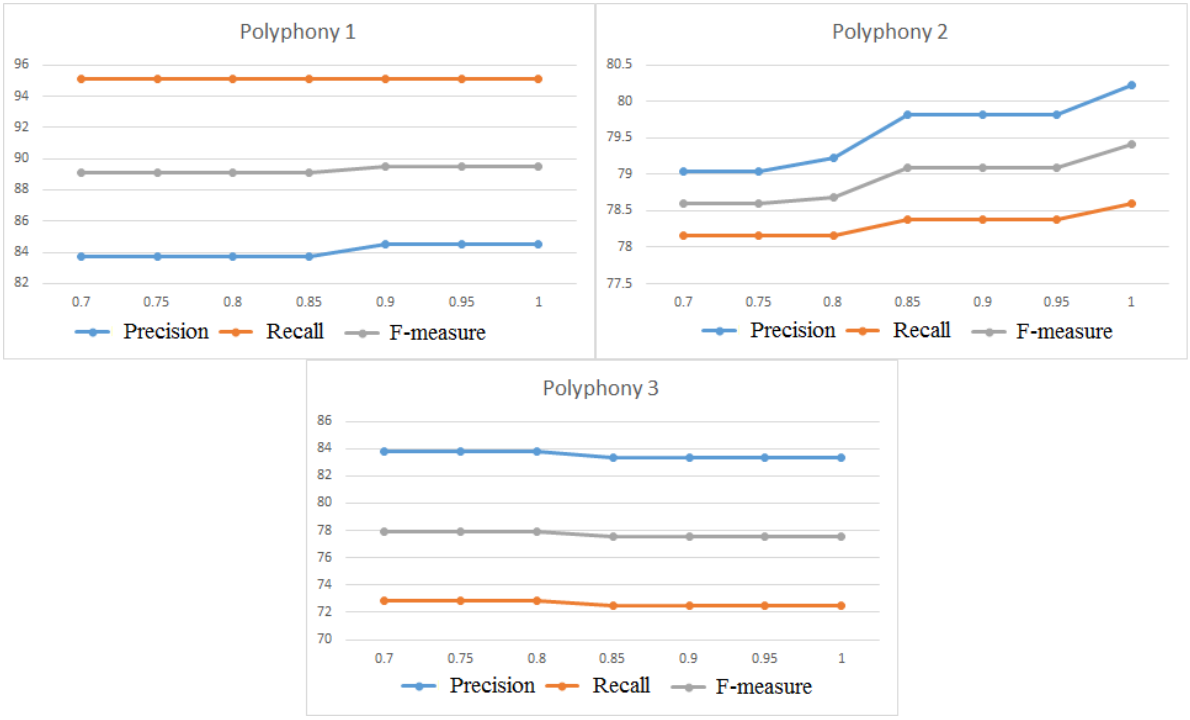
Όσον αφορά τα Σχήματα 5.15, 5.16 και την περίπτωση πολυφωνίας 1 παρατηρείται ότι ενώ το recall μένει σταθερό, το precision αυξάνεται με την αύξηση του κατωφλίου. Αυτό συμβαίνει διότι πλέον έχοντας ένα πιο αυστηρό κατώφλι, συνδυασμοί που περιέχουν τη σωστή συχνότητα συν μία συχνότητα επιπλέον (αρμονική της σωστής συχνότητας κατά πάσα πιθανότητα) απορρίπτονται και μένουν οι σωστοί συνδυασμοί που περιέχουν μόνο τη μία σωστή συχνότητα. Για την περίπτωση πολυφωνίας 2, όταν συνδυασμοί με δύο νότες αποκλείονται λόγω του αυστηρότερου κατωφλίου, τότε συχνά παίρνουν τη θέση τους συνδυασμοί με τρεις νότες, δύο από τις οποίες είναι οι ζητούμενες. Τέλος, στην περίπτωση πολυφωνίας 3, η επίδοση επηρεάζεται ελάχιστα. Μειώνεται λίγο το precision το οποίο υποδεικνύει ότι κάνοντας μία αυστηρότερη επιλογή των συνδυασμών με δύο νότες, το σύστημα εκεί που έβρισκε δύο νότες αντί για τρεις, τώρα βρίσκει τρεις νότες, από τις οποίες όμως συχνά η τρίτη νότα δεν είναι η σωστή.

Όσον αφορά τα Σχήματα 5.17, 5.18 και την περίπτωση πολυφωνίας 1 παρατηρείται ότι ενώ το

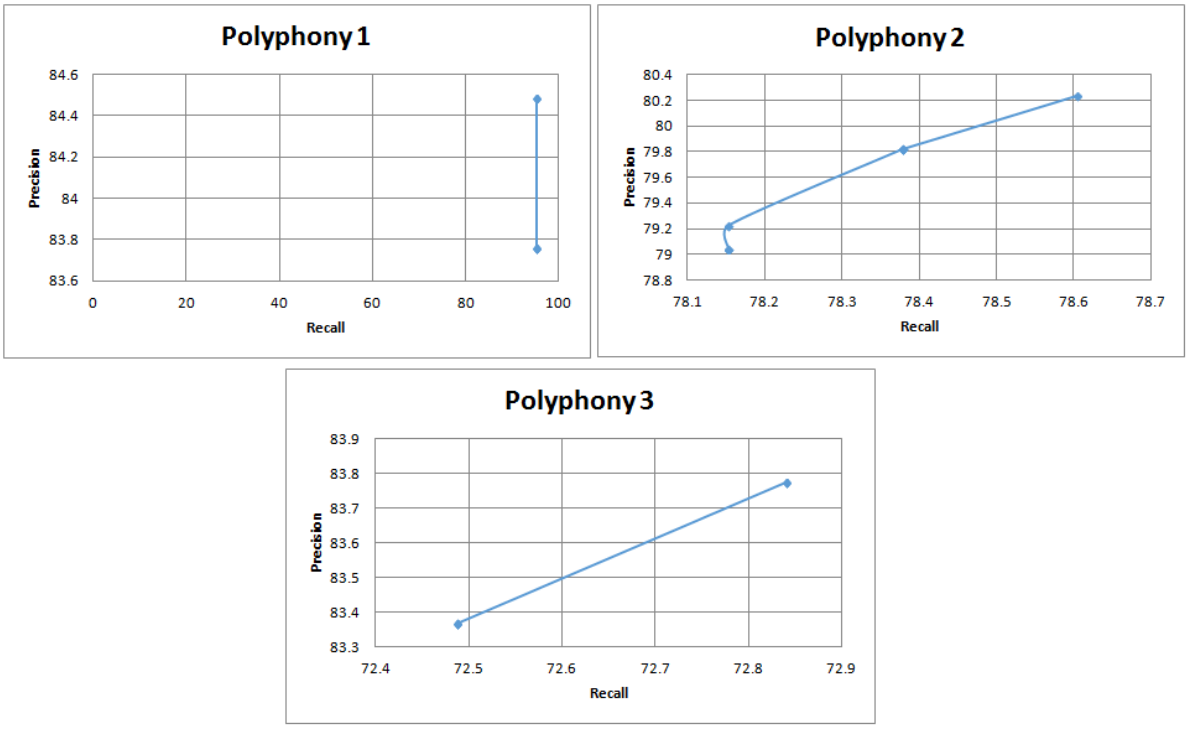


Σχήμα 5.16: Καμπύλη Precision-Recall για βαθμούς πολυφωνίας 1,2 και 3 ως συνάρτηση του ποσοστού της συνολικής ισχύος που η κανονικοποιημένη ισχύς μιας νότας πρέπει να ξεπερνάει ώστε να θεωρηθεί υποψήφια μέσα σε συνδυασμούς ανά δύο.

recall μένει σταθερό, το precision αυξάνεται με την αύξηση του κατωφλίου. Αυτό συμβαίνει διότι πλέον έχοντας ένα πιο αυστηρό κατώφλι, συνδυασμοί που περιέχουν τη σωστή συχνότητα συν δύο συχνότητες επιπλέον απορρίπτονται και μένουν οι σωστοί συνδυασμοί που περιέχουν μόνο τη μία σωστή συχνότητα. Για την περίπτωση πολυφωνίας 2, κάνοντας μία αυστηρότερη επιλογή των συνδυασμών με τρεις νότες, μειώνεται η πιθανότητα το σύστημα να κάνει λάθος παίρνοντας ένα συνδυασμό με τρεις νότες στη θέση ενός συνδυασμού με δύο νότες. Τέλος, στην περίπτωση πολυφωνίας 3, αυξάνοντας το κατώφλι μειώνονται λίγο τόσο το precision όσο και το recall, γιατί αποκλείονται κάποιες σωστές συχνότητες οι οποίες αντικαθίστανται από συχνότητες που δεν υπάρχουν πραγματικά και αφού δεν θεωρούμε συνδυασμούς μεγαλύτερης πολυφωνίας από 3, δεν υπάρχει η δυνατότητα να εμφανιστούν αυτές οι συχνότητες σε συνδυασμούς μεγαλύτερης πολυφωνίας.



Σχήμα 5.17: Η μεταβολή των precision, recall και F-measure για βαθμούς πολυφωνίας 1,2 και 3 ως συνάρτηση του ποσοστού της συνολικής ισχύος που η κανονικοποιημένη ισχύς μιας νότας πρέπει να ξεπερνάει ώστε να θεωρηθεί υποψήφια μέσα σε συνδυασμούς ανά τρία.



Σχήμα 5.18: Καμπύλη Precision-Recall για βαθμούς πολυφωνίας 1,2 και 3 ως συνάρτηση του ποσοστού της συνολικής ισχύος που η κανονικοποιημένη ισχύς μιας νότας πρέπει να ξεπερνάει ώστε να θεωρηθεί υποψήφια μέσα σε συνδυασμούς ανά τρία.

Κεφάλαιο 6

Συμπεράσματα

6.1 Συμβολή της διπλωματικής εργασίας

Στην παρούσα διπλωματική εργασία εξετάστηκε το πρόβλημα της Αυτόματης Καταγραφής Μουσικής. Το ενδιαφέρον για αυτό το πρόβλημα, που σχετίζεται με τον τομέα της Ψηφιακής Επεξεργασίας Ηχητικών Σημάτων και παραμένει πάντα ανοιχτό, είναι ολοένα αυξανόμενο τα τελευταία χρόνια. Μελετήθηκαν διάφορες πτυχές που απαρτίζουν το σύνθετο αυτό θέμα. Πιο συγκεκριμένα, η κύρια συμβολή της εργασίας αφορά στην παρουσίαση δύο πρωτότυπων μεθόδων, μία για την αυτόματη ανίχνευση της αρχής (onset) της νότας σε μονοφωνικά αποσπάσματα μουσικής για πιάνο και μία μέθοδο εκτίμησης πολλαπλών τόνων για συγχορδίες παιγμένες στο πιάνο. Οι μέθοδοί μας είναι σχετικά απλές, ενώ δίνουν πολύ ενθαρρυντικά αποτελέσματα. Και οι δύο μέθοδοι εκμεταλλεύονται τη συχνοτική αναπαράσταση των κλειδιών του πιάνου: για την εύρεση των onsets χρησιμοποιείται μία συστοιχία Gabor φίλτρων, που έχει ως κεντρικές συχνότητες τις συχνότητες του πιάνου και για την εκτίμηση πολλαπλών τόνων ο DTFT υπολογίζεται ακριβώς σε αυτές τις συχνότητες του πιάνου. Στη συνέχεια, γίνεται κατάλληλη επεξεργασία αυτών των συχνοτικών αναπαραστάσεων χρησιμοποιώντας κάποιες αποδοτικές τεχνικές επεξεργασίας σημάτων, με σκοπό την επίτευξη του στόχου της ανίχνευσης των onsets στην πρώτη περίπτωση και της εκτίμησης πολλαπλών τόνων στη δεύτερη. Η υλοποίηση αυτών των μεθόδων έγινε στο προγραμματιστικό περιβάλλον MatLab.

Γενικότερα η διπλωματική εργασία είχε ακόμα τις εξής κατευθύνσεις:

• Επισκόπηση της βιβλιογραφία σχετικά με τα προβλήματα της ανίχνευσης της αρχής (onset) της νότας, της εκτίμησης του τονικού ύψους (pitch) της νότας, της αυτόματης εύρεσης ρυθμικής πληροφορίας (εκτίμηση tempo, beat tracking, αυτόματη εύρεση του μέτρου) και της εκτίμησης πολλαπλών τόνων.

• Υλοποίηση σε MatLab ενός πλήρους συστήματος αναγνώρισης μονοφωνικής μουσικής που εξάγει την Piano Roll αναπαράσταση ενός μουσικού σήματος, συνδυάζοντας κατάλληλα μία βαθμίδα εύρεσης των onsets ([7]) με μία βαθμίδα εκτίμησης του pitch ([8]).

• Υλοποίηση σε MatLab της μεθόδου που περιγράφεται στο [4] για την αυτόματη εύρεση του μέτρου, η οποία απαιτεί τη χρήση μίας μεθόδου για την εύρεση της περιόδου των beats (επιλέχτηκε

η [6]) και μίας μεθόδου για την εύρεση των χρονικών στιγμών των beats (επιλέχτηκε η [9]).

6.2 Κατευθύνσεις για μελλοντική έρευνα

Παρόλο που η παρούσα εργασία έδωσε ικανοποιητικά αποτελέσματα, υπάρχουν περιθώρια για μελλοντική έρευνα.

• Η μέθοδος εκτίμησης πολλαπλών τόνων μπορεί να δώσει καλύτερα αποτελέσματα και να επεκταθεί και σε μεγαλύτερο βαθμό πολυφωνίας κάνοντας ευρύτερη χρήση εκπαίδευσης για διάφορα κατάλληλα επιλεγμένα χαρακτηριστικά.

• Τα πειράματα και η εκπαίδευση που έγιναν, είναι σχετικά μικρής κλίμακας, καθώς είχαμε λίγα δείγματα από κάθε νότα και λίγους συνδυασμούς. Για την επιβεβαίωση της μεθόδου, κρίνουμε ότι είναι απαραίτητο να γίνουν πειράματα σε μεγαλύτερη έκταση.

• Η μέθοδος ανίχνευσης των onsets, κάνοντας χρήση μίας μεγαλύτερης βάσης δεδομένων, θα μπορούσε να επωφεληθεί από μία πιο προσαρμοστική κατωφλίωση της συνάρτησης ανίχνευσης.

Τέλος, ιδιαίτερο ενδιαφέρον παρουσιάζει η ενσωμάτωση της μεθόδου εκτίμησης πολλαπλών τόνων που περιγράφεται στην ενότητα 5.3 σε ένα πλήρες σύστημα αναγνώρισης πολυφωνικής μουσικής. Θα μπορούσαν να χρησιμοποιηθούν και a priori μουσικές γνώσεις ώστε να οριστούν κατάλληλα πιθανότητες για τους διάφορους συνδυασμούς νοτών και τις ακολουθίες συνδυασμών, μιας και δεν είναι όλες οι περιπτώσεις ισοπίθανες. Αυτή η παρατήρηση δίνει μία νέα διάσταση στο πρόβλημα προς την οποία θα μπορούσε να στραφεί η μελλοντική έρευνα.

Παράρτημα Α΄

MIR Toolbox

Το MIR toolbox [1] που χρησιμοποιήθηκε στα πλαίσια της διπλωματικής είναι ένα ολοκληρωμένο σύστημα λειτουργιών γραμμένο σε Matlab από τους Olivier Lartillot και Petri Toiviainen και είναι αφιερωμένο στην εξαγωγή μουσικών χαρακτηριστικών από ηχητικά αρχεία. Μεταξύ άλλων, χρησιμοποιώντας το MIRtoolbox, μπορούν να εξαχθούν χαρακτηριστικά που σχετίζονται με το pitch, τη χροιά, την τονικότητα ή το ρυθμό. Το toolbox επίσης συμπεριλαμβάνει συναρτήσεις για στατιστική ανάλυση, κατάτμηση και ομαδοποίηση. Είναι σχεδιασμένο με έναν αρθρωτό τρόπο, όπου οι διαφορετικοί αλγόριθμοι μπορούν να αποδομηθούν σε μικρότερες και πιο στοιχειώδεις συναρτήσεις και μηχανισμούς. Αυτή η προσέγγιση επιτρέπει στους χρήστες να αλληλεπιδράσουν με αυτά τα ελάχιστα μπλοκ και να τα συνδυάσουν με ένα διαφορετικό και πρωτότυπο τρόπο για την παραγωγή νέων χαρακτηριστικών. Όλες οι μέθοδοι εξαγωγής χαρακτηριστικών δέχονται αρχεία ήχου ως είσοδο, ή κάποιο άλλο ενδιάμεσο αποτέλεσμα από προηγούμενες λειτουργίες. Η ίδια σύνταξη μπορεί επίσης να χρησιμοποιηθεί για την ανάλυση μεμονωμένων αρχείων ήχου, ένα σύνολο αρχείων, έναν φάκελο γεμάτο με αρχεία ήχου, μία σειρά από ηχητικά αποσπάσματα, πολυκαναλικά σήματα κτλ.

Ο μεγάλος αριθμός από χαρακτηριστικά χαμηλού αλλά και υψηλού επιπέδου σε συνδυασμό με την προσαρμοστική σύνταξη για τη δημιουργία νέων συναρτήσεων καθώς και τις υπάρχουσες μεθόδους εξόδου, που επιτρέπουν όχι μόνο την εξαγωγή πληροφοριών αλλά και την οπτική αναπαράστασή τους, καθιστούν αυτό το toolbox ιδιαίτερα χρήσιμο. Η τεκμηρίωση που παρέχεται είναι καλή. Το toolbox για να λειτουργήσει χρειάζεται το περιβάλλον του Matlab καθώς και το Signal Processing Toolbox [2] [121].

[1] https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox
[2] http://www.mathworks.com/products/signal/

Βιβλιογραφία

[1] A. Klapuri, "Introduction to music transcription," in Klapuri and Davy [15], pp. 3–20.

[2] K. J. Hildon, *The Complete Commodore Inner Space Anthology.* Transactor Publishing Incorporated, 1985.

[3] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 1035–1047, Sept. 2005.

[4] M. Gainza, "Automatic musical meter detection," in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '09, (Washington, DC, USA), pp. 329–332, IEEE Computer Society, 2009.

[5] V. Emiya, *Transcription automatique de la musique de piano.* These, Télécom ParisTech, Oct. 2008.

[6] A. Gkiokas, V. Katsouros, G. Carayannis, and T. Stafylakis, "Music Tempo Estimation and Beat Tracking by Applying Source Separation and Metrical Relations," in *Proceedings of the 37th IEEE International Conference on Acoustics, Speech and Signal Processing*, (Kyoto, Japan), March 2012.

[7] J. P. Bello, C. Duxbury, M. Davies, and M. Sandler, "On the use of phase and energy for musical onset detection in the complex domain," *Signal Processing Letters, IEEE*, vol. 11, no. 6, pp. 553–556, 2004.

[8] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *IFA Proceedings 17*, pp. 97–110, 1993.

[9] D. P. W. Ellis, "Beat tracking by dynamic programming," *J. New Music Research*, vol. 2007, pp. 51–60, 2007.

[10] W. M. Hartmann, "Pitch, periodicity, and auditory organization.," *The Journal of the Acoustical Society of America*, vol. 100, pp. 3491–3502, Dec. 1996.

[11] N. Fletcher and T. Rossing, *The Physics of Musical Instruments.* Springer, 1988.

[12] S. Handel, "Timbre perception and auditory object identification," [13], pp. 425–460.

[13] B. Moore, *Hearing.* Handbook of Perception and Cognition, Academic Press, 1995.

[14] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing*, vol. 41, no. 10, pp. 3024–3051, 1993.

[15] A. Klapuri and M. Davy, eds., *Signal Processing Methods for Music Transcription*. New York: Springer, 2006.

[16] J. A. Moorer, *On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer*. PhD thesis, Stanford University, Stanford, CA, 1975.

[17] J. A. Moorer, "On the transcription of musical sound by computer," *Computer Music Journal*, pp. 1(4):32–38, 1977.

[18] C. Chafe, D. A. Jaffe, K. Kashima, B. Mont-Reynaud, and J. O. Smith, "Techniques for note identification in polyphonic music," in *Proceddings of the 1985 International Computer Music Conference, Burnaby, B.C., Canada*, no. STAN-M-29, International Computer Music Association, International Computer Music Association, 1985.

[19] M. Piszczalski, *A computational model of music transcription*. PhD thesis, Ann Arbor, MI, USA, 1986.

[20] R. C. Maher, *An Approach for the Separation of Voices in Composite Musical Signals*. PhD thesis, University of Illinois, IL, USA, 1989.

[21] R. C. Maher, "Evaluation of a method for separating digitized duet signals," *J. Audio Eng. Soc.*, vol. 38(12), December 1990.

[22] M. Goto and Y. Muraoka, "A beat tracking system for acoustic signals of music," in *Proc. of the Second ACM Intl. Conf. on Multimedia*, pp. 365–372, 1994.

[23] W. A. Schloss, *On the Automatic Transcription of Percussive Music: From Acoustic Signal to High Level Analysis*. PhD thesis, Stanford University, CA, USA, May 1985.

[24] J. Bilmes, "Timing is of the essence: Perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm," Master's thesis, MIT, Cambridge, MA, 1993.

[25] M. Goto and Y. Muraoka, "A sound source separation system for percussion instruments," in *Transactions of the Institute of Electronics, Information and Communication Engineers*, vol. J77-D-II, pp. 901–911, 1994.

[26] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka, "Organization of Hierarchical Perceptual Sounds: Music Scene Analysis with Autonomous Processing Modules and a Quantitative Information Integration Mechanism," in *IJCAI*, pp. 158–164, 1995.

[27] M. Goto, "A predominant-FO estimation method for real-world musical audio signals: Map estimation for incorporating prior knowledge about FOs and tone models," in *Proc. Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis*, 2001.

[28] M. Davy and S. J. Godsill, "Bayesian harmonic models for musical signal analysis," in *Bayesian Statistics 7*, Oxford University Press, 2002.

[29] M. Ryynänen and A. Klapuri, "Polyphonic music transcription using note event modeling," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, NY), 2005.

[30] A. T. Cemgil and B. Kappen, "Monte carlo methods for tempo tracking and rhythm quantization," *Journal of Artificial Intelligence Research*, vol. 18, pp. 45–81, 2003.

[31] S. W. Hainsworth and M. D. Macleod, "Particle filtering applied to musical tempo tracking," *EURASIP J. Appl. Signal Process.*, vol. 2004, pp. 2385–2395, Jan. 2004.

[32] A. P. Klapuri, A. J. Eronen, and J. T. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 342–355, Jan. 2006.

[33] O. Gillet and G. Richard, "Automatic Transcription of Drum Loops," in *Proceedings of the 2004 IEEE Conference on Acoustics, Speech and Signal Processings (ICASSP'04)*, May 2004.

[34] J. Paulus and A. Klapuri, "Conventional and periodic n-grams in the transcription of drum sequences," *Multimedia and Expo, IEEE International Conference on*, vol. 2, pp. 737–740, 2003.

[35] P. Herrera, G. Peeters, and S. Dubnov, "Automatic classification of musical instrument sounds," *Journal of New Music Research*, vol. 32, 2003.

[36] K. D. Martin, "Automatic Transcription of Simple Polyphonic Music: Robust Front End Processing," Tech. Rep. Technical Report No. 399, Massachusetts Institute of Technology (MIT) Media Lab, 1996.

[37] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 708–716, Nov. 2000.

[38] A. P. Klapuri, "A perceptually motivated multiple-F0 estimation method," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 291–294, 2005.

[39] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.

[40] A. S. Bregman, *Auditory Scene Analysis*. MIT Press, 1990.

[41] D. K. Mellinger, *Event Formation and Separation in Musical Sound*. PhD thesis, Stanford University, CA, USA, 1991.

[42] K. Kashino and H. Tanaka, "A sound source separation system with the ability of automatic tone modeling," in *International Computer Music Conference*, 1993.

[43] D. Godsmark and G. J. Brown, "A blackboard architecture for computational auditory scene analysis," 1999.

[44] A. Sterian, M. H. Simoni, and G. H. Wakefield, "Model-Based Musical Transcription," in *International Computer Music Conference (ICMC)*, (Beijing, China), International Computer Music Association, 1999.

[45] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Netw.*, vol. 13, pp. 411–430, May 2000.

[46] M. Casey, *Auditory Group Theory with Applications to Statistical Basis Methods for Structured Audio*. PhD thesis, 1998.

[47] M. A. Casey, "Separation of Mixed Audio Sources by Independent Subspace Analysis," tech. rep., MERL, 2000.

[48] P. Lepain, "Polyphonic pitch extraction from musical signals," *Journal of New Music Research*, vol. 28, no. 4, pp. 296–309, 1999.

[49] P. Smaragdis, *Redundancy reduction for computational audition, a unifying approach*. PhD thesis, 2001. AAI0803468.

[50] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 177–180, 2003.

[51] S. A. Abdallah, *Towards Music Perception by Redundancy Reduction and Unsupervised Learning in Probabilistic Models*. PhD thesis, King's College London, London, UK, 2002.

[52] S. M. Abdallah and M. D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," in *ISMIR'04*.

[53] T. Virtanen, "Unsupervised learning methods for source separation in monaural music signals," in Klapuri and Davy [15], pp. 267–296.

[54] D. FitzGerald, *Automatic Drum Transcription and Source Separation*. PhD thesis, Dublin Institute of Technology, Dublin, Ireland, 2004.

[55] D. Fitzgerald, B. Lawlor, and E. Coyle, "Prior subspace analysis for drum transcription," in *114th AES Convention Amsterdam March 22 nd –25 th 2003*, 2003.

[56] J. Paulus and T. Virtanen, "Drum transcription with nonnegative spectrogram factorisation," in *EUSIPCO*, pp. 4–8, 2005.

[57] H. Kameoka, T. Nishimoto, and S. Sagayama, "Separation of harmonic structures based on tied gaussian mixture model and information criterion for concurrent sounds," in *International Conference on Acoustics, Speech, and Signal Processing*, pp. 297–300, 2004.

[58] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," *IEEE Transactions on Multimedia*, vol. 6, no. 3, pp. 439–449, 2004.

[59] A. Zils, F. Pachet, O. Delerue, and F. Gouyon, "Automatic extraction of drum tracks from polyphonic music signals," in *Proceedings of the First International Symposium on Cyber Worlds (CW'02)*, pp. 179–, 2002.

[60] D. FitzGerald, R. Lawlor, and E. Coyle, "Drum transcription in the presence of pitched instruments using prior subspace analysis," in *Irish Signals & Systems Conference 2003, Limerick, Ireland*, jul 2003.

[61] K. Yoshii, M. Goto, and H. G. Okuno, "Drum sound identification for polyphonic music using template adaptation and matching methods," in *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing, Jeju, Korea*, 2004.

[62] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, "An experimental comparison of audio tempo induction algorithms," *Trans. Audio, Speech and Lang. Proc.*, vol. 14, pp. 1832–1844, Sept. 2006.

[63] K. Kashino and H. Murase, "A sound source identification system for ensemble music based on template adaptation and music stream extraction," *Speech Commun.*, vol. 27, pp. 337–349, Apr. 1999.

[64] A. L. Berenzweig and D. P. W. Ellis, "Locating singing voice segments within music signals," in *Proc. IEEE Workshop on Apps. of Sig. Proc. to Audio and Acous.*, pp. 119–122, 2001.

[65] J. Eggink and G. J. Brown, "Instrument Recognition in Accompanied Sonatas and Concertos," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, (Montreal, Canada), pp. 217–220, May 2004.

[66] E. Vincent and X. Rodet, "Instrument identification in solo and ensemble music using independent subspace analysis," in *Proc. ISMIR*, pp. 576–581, 2004.

[67] M. Goto and Y. Muraoka, "Beat Tracking based on Multiple-agent Architecture – A Real-time Beat Tracking System for Audio Signals," in *Proc. of the Second International Conference on Multiagent Systems (ICMAS)*, December 1996.

[68] M. Gainza, B. Lawlor, and E. Coyle, "Onset Detection Using Comb Filters," in *Workshop on Applications of Signal Processing to Audio and Acoustics, 2005*, 2005.

[69] C. Duxbury, M. Sandler, and M. Davis, "A Hybrid Approach to Musical Note Onset Detection," in *Proc. Digital Audio Effects Workshop (DAFx)*, 2002.

[70] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proceedings of the Acoustics, Speech, and Signal Processing,on 1999 IEEE International Conference - Volume 06*, ICASSP '99, (Washington, DC, USA), pp. 3089–3092, IEEE Computer Society, 1999.

[71] C. C. Toh, B. Zhang, and Y. Wang, "Multiple-Feature Fusion Based Onset Detection for Solo Singing Voice," in *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08)*, (Philadelphia, PA, USA), Sept. 2008.

[72] S. Dixon, "Onset detection revisited," in *Proceedings of the 9th International Conference on Digital Audio Effects*, pp. 133–137, 2006.

[73] S. Hainsworth and M. Macleod, "Onset Detection in Musical Audio Signals," in *Proc. of the International Computer Music Conference (ICMC)*, (Singapore), Sept. 2003.

[74] N. Collins, "A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions," in *AES Convention 118*, pp. 28–31, 2005.

[75] J. Bello, G. Monti, and M. Sandler, "Phase-Based Note Onset Detection for Music Signals," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, (Hong Kong), Mar. 2003.

[76] J. Brown, M. I. of Technology. Media Laboratory. Vision, and M. Group, *Musical Fundamental Frequency Tracking Using a Pattern Recognition Method*. M.I.T. Media Lab Vision and Modeling Group technical report, Vision and Modeling Group, Media Laboratory, Massachusetts Institute of Technology, 1993.

[77] B. Doval and X. Robet, "Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and hmms," in *Proceedings of the 1993 IEEE international conference on Acoustics, speech, and signal processing: plenary, special, audio, underwater acoustics, VLSI, neural networks - Volume I*, ICASSP'93, (Washington, DC, USA), pp. 221–224, IEEE Computer Society, 1993.

[78] A. M. Noll, "Cepstrum pitch determination," *The Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 293–309, 1967.

[79] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[80] A. Klapuri, "Multipitch estimation and sound separation by the spectral smoothness principle," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 5, pp. 3381–3384, 2001.

[81] G. Peeters, "Music Pitch Representation by Periodicity Measures Based on Combined Temporal and Spectral Representations," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, (Toulouse, France), pp. 53–56, May 2006.

[82] C. Rosão, R. Ribeiro, and D. M. de Matos, "Influence of peak selection methods on onset detection," in *ISMIR* (F. Gouyon, P. Herrera, L. G. Martins, and M. Müller, eds.), pp. 517–522, FEUP Edições, 2012.

[83] T. Eerola and P. Toiviainen, *MIDI Toolbox: MATLAB Tools for Music Research*. Jyväskylä, Finland: University of Jyväskylä, 2004.

[84] J. Seppänen, "Tatum grid analysis of musical signals," 2001.

[85] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 293–302, July 2002.

[86] M. Goto and Y. Muraoka, "An audiobased realtime beat tracking system and its applications," Proceedings of the International Computer Music Conference, 1998.

[87] M. E. P. Davies and M. D. Plumbley, "Context-Dependent Beat Tracking of Musical Audio," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 1009–1020, Mar. 2007.

[88] J. Seppänen, A. Eronen, and J. Hiipakka, "Joint Beat and Tatum Tracking from Music Signals," in *Proceedings of the 7th International Conference on Music Information Retrieval* (K. Lemström, A. Tindale, and R. Dannenberg, eds.), (Victoria, Canada), University of Victoria, Oct. 2006. `http://ismir2006.ismir.net/PAPERS/ISMIR0683_Paper.pdf`.

[89] P. Grosche and M. Müller, "A mid-level representation for capturing dominant tempo and pulse information in music recordings," in *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009, Kobe International Conference Center, Kobe, Japan, October 26-30, 2009* (K. Hirata, G. Tzanetakis, and K. Yoshii, eds.), pp. 189–194, International Society for Music Information Retrieval, 2009.

[90] P. Grosche and M. Müller, "Extracting predominant local pulse information from music recordings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1688–1701, 2011.

[91] G. Peeters, "Template-based estimation of time-varying tempo," *EURASIP J. Appl. Signal Process.*, vol. 2007, pp. 158–158, Jan. 2007.

[92] F. Gouyon and P. Herrera, "Determination of the Meter of musical audio signals: Seeking recurrences in beat segment descriptors," in *Proc. 114th Convention of the Audio Engineering Society*, (Amsterdam, The Netherlands), 2003.

[93] S. E. Dixon, "Automatic extraction of tempo and beat from expressive performances," *Journal of New Music Research*, vol. 30, pp. 39–58, 2001.

[94] S. Dixon, "Evaluation of the Audio Beat Tracking System BeatRoot," *Journal of New Music Research*, vol. 36, pp. 39–50, Mar. 2007.

[95] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *Journal of New Music Research*, vol. 30, no. 2, pp. 159–171, 2001.

[96] N. Degara, A. Pena, M. E. P. Davies, and M. D. Plumbley, "Note onset detection using rhythmic structure," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, 14-19 March 2010, Sheraton Dallas Hotel, Dallas, Texas, USA*, pp. 5526–5529, IEEE, 2010.

[97] J. Paulus and A. Klapuri, "Measuring the Similarity of Rhythmic Patterns," in *Proc. of the Third International Conference on Music Information Retrieval* (M. Fingerhut, ed.), (Paris, France), pp. 150–156, Oct 2002.

[98] D. FitzGerald, "Harmonic/percussive separation using median filtering," in *13th International Conference on Digital Audio Effects (DAFX10)*, (Graz, Austria), 2010.

[99] A. Klapuri, *Signal Processing Methods for the Automatic Transcription of Music*. PhD thesis, Tampere University of Technology, Finnland, March 2004.

[100] R. Zhou, J. D. Reiss, M. Mattavelli, and G. Zoia, "A computationally efficient method for polyphonic pitch estimation," *EURASIP J. Adv. Signal Process*, vol. 2009, pp. 28:1–28:11, Jan. 2009.

[101] A. P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 804–816, 2003.

[102] J. Yin, T. Sim, Y. Wang, and A. Shenoy, "Music Transcription Using an Instrument Model," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, pp. 217–220, IEEE, 2005.

[103] C. Cao, M. Li, J. Liu, and Y. Yan, "Multiple f0 estimation in polyphonic music," in *MIREX (2007), multiple f0 estimation and tracking contest*, 2007.

[104] A. de Cheveigné, "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," *Acoustical Society of America Journal*, vol. 93, pp. 3271–3290, June 1993.

[105] C. Yeh, A. Robel, and X. Rodet, "Multiple fundamental frequency estimation of polyphonic music signals," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, vol. 3, pp. iii/225–iii/228 Vol. 3, 2005.

[106] A. T. Cemgil, *Bayesian Music Transcription.* PhD thesis, Radboud University of Nijmegen, 2004.

[107] A. Doucet and X. Wang, "Monte Carlo methods for signal processing: a review in the statistical signal processing context," *Signal Processing Magazine, IEEE*, vol. 22, no. 6, pp. 152–170, 2005.

[108] P. J. Walmsley, S. J. Godsill, and P. J. W. Rayner, "Polyphonic pitch tracking using joint bayesian estimation of multiple frame parameters," in *Proc. IEEE Workshop on Audio and Acoustics, Mohonk*, 1999.

[109] M. Davy and S. J. Godsill, "Bayesian harmonic models for musical signal analysis," in *Bayesian Statistics 7*, Oxford University Press, 2002.

[110] A. T. Cemgil, H. J. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 679–694, Mar. 2006.

[111] M. Davy, S. J. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *Journal of the Acoustical Society of America*, vol. 119, pp. 2498–2517, 2006.

[112] M. P. Ryynänen and A. Klapuri, "Polyphonic music transcription using note event modeling," in *In Proc. 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 319–322, 2005.

[113] G. E. Poliner and D. P. W. Ellis, "A discriminative model for polyphonic piano transcription," in *EURASIP Journal on Advances in Signal Processing*, p. 2007, 2007.

[114] G. Reis, N. Fonseca, F. Fernandez, and A. Ferreira, "A genetic algorithm approach with harmonic structure evolution for polyphonic music transcription," *The 8th IEEE International Symposium on Signal Processing and Information Technology*, pp. 491–496, 2008.

[115] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 1066–1074, Mar. 2007.

[116] E. Vincent, N. Bertin, and R. Badeau, "Two nonnegative matrix factorization methods for polyphonic pitch transcription," in *2007 Music Information Retrieval Evaluation eXchange (MIREX)*, (Vienna, Austria), 2007.

[117] N. Bertin, R. Badeau, and G. Richard, "Blind Signal Decompositions for Automatic Transcription of Polyphonic Music: NMF and K-SVD on the Benchmark," *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1, Apr. 2007.

[118] K. D. Martin, "A blackboard system for automatic transcription of simple polyphonic music," Tech. Rep. Technical Report #385, Massachusetts Institute of Technology (MIT) Media Lab, July 1996.

[119] J. Bello and M. Sandler, "Blackboard system and top-down processing for the transcription of simple polyphonic music," in *COST G-6 Conference on Digital Audio Effects (DAFx-00)*, (Verona, Italy), 2000.

[120] G. Monti and M. Sandler, "Automatic polyphonic piano note extraction using fuzzy logic in a blackboard system," in *Proceedings Digital Audio Effects Workshop (DAFx)*, pp. 39–44, 2002.

[121] O. Lartillot and P. Toiviainen, "MIR in Matlab (II): A toolbox for musical feature extraction from audio," in *International Conference on Music Information Retrieval*, pp. 127–130, 2007.