ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

Αυτόματη Ανάκτηση Μουσικής Πληροφορίας με Έμφαση στο Ρυθμό

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Άγγελος Α. Γκιόκας

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Η/Υ Ε.Μ.Π.

Επιβλέπων Καθηγητής: Πέτρος Μαραγκός, Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2016

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

Αυτόματη Ανάκτηση Μουσικής Πληροφορίας με Έμφαση στον Ρυθμό

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Άγγελος Α. Γκιόκας

Συμβουλευτική Επιτροπή : Καθ. Π. Μαραγκός (Επιβλέπων)
Καθ. Α.Γ. Σταφυλοπάτης
Καθ. Π. Τσανάκας

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 21η Ιουλίου 2016.

.....	.....	.....
Π. Μαραγκός	Α.Γ. Σταφυλοπάτης	Π. Τσανάκας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.
.....	.....	.....
Ι. Ιωαννίδης	Β.Κατσούρος	Α. Ποταμιάνος
Καθ. Ε.Κ.Π.Α	Ερευν. Α', Ε.Κ. Αθηνά	Αναπλ. Καθηγητής Ε.Μ.Π.

.....
Ά. Πικράκης
Επικ. Καθηγητής ΠΑ.ΠΕΙ

Αθήνα, Ιούλιος 2016

Άγγελος Α. Γκιόκας

Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω όλους όσους μου συμπαραστάθηκαν και με βοήθησαν στη μακρά πορεία εκπόνησης της παρούσας διατριβής. Πρωτίστως τον επιβλέποντα Καθ. Γ. Καραγιάννη που είχα την μεγάλη τιμή να συνεργαστώ μαζί του και που προς μεγάλη λύπη μας άφησε νωρίς. Εξίσου μεγάλη ήταν και η τιμή μου να συνεργαστώ με τον Καθ. Π. Μαραγκό, που επιτέλεσε επιβλέπων της διατριβής μου στο τελευταίο μέρος της, και τον ευχαριστώ βαθέως. Θερμές ευχαριστίες επίσης σε όλα τα μέλη της επταμελούς επιτροπής. Στους συνάδελφους μου στο Ινστιτούτο Επεξεργασίας του Λόγου, και ιδιαίτερα στον Β. Κατσούρο που αποτέλεσε και σημαντικό επιστημονικό καθοδηγητή μου, καθώς και στους Β. Παπαβασιλείου και Φ. Σιμιστήρα. Ευχαριστώ επίσης τους Gerhard Widmer, Arthur Flexer, Jan Shlueter και Stefan Lattner από το Austrian Research Institute for Artificial Intelligence, τον Γιώργο Τζανετάκη από το University of Victoria και τον Άγγελο Πικράκη από το ΠΑ.ΠΕΙ. για την συνεργασία που είχαμε στη διάρκεια εκπόνησης της παρούσας διατριβής. Τέλος, θερμές ευχαριστίες σε συγγενείς και φίλους που ήταν τόσο κοντά αλλά και τόσο μακριά, και ιδιαίτερες ευχαριστίες στην Χάρις.

Περιεχόμενα

Κεφάλαιο 1 :Εισαγωγή **7**
1.1 Πρόλογος 7
1.2 Μουσικές Έννοιες και το Μουσικό Σήμα 9
1.2.1 Οι διαστάσεις του ήχου 9
1.2.2 Η μουσική νότα 11
1.2.3 Αρμονική οργάνωση της μουσικής 14
1.2.4 Χρονική οργάνωση της μουσικής και το πεντάγραμμο 17
1.2.5 Ρυθμικές αντικανονικότητες 23
Αντικανονικότητες μουσικού τέμπο 23
Αντικανονικότητες μουσικού παλμού 23
Αντικανονικότητες μουσικού μέτρου 24
1.3 Βιβλιογραφική Επισκόπηση 25
1.3.1 Εισαγωγή 25
1.3.2 Αυτόματη Εξαγωγή Τέμπο και Μουσικού Παλμού 26
1.3.3 Αυτόματη Αναγνώριση Μέτρου και Εξαγωγή Μουσικού Κλειδιού 28
1.3.4 Ρυθμική Κατηγοριοποίηση 28
Κεφάλαιο 2: Ανάλυση Περιοδικότητας Μουσικών Σημάτων **30**
2.1 Εισαγωγή 30
2.2 Ο μετασχηματισμός σταθερού Q 34
2.3 Διαχωρισμός αρμονικών / κρουστών πηγών 36
2.4 Εξαγωγή Συνάρτησης Έμφασης 42
2.5 Ανάλυση Περιοδικότητας 44
2.6 Η front-end επεξεργασία 50
Κεφάλαιο 3: Εξαγωγή Χαρακτηριστικών από τη Συνάρτηση Περιοδικότητας **56**
3.1 Εισαγωγή 56
3.2 Χειρονακτικά Χαρακτηριστικά 57
3.2.1 Κλιμάκωση της Συνάρτησης Περιοδικότητας 57
3.2.2 Κωδικοποίηση της Συνάρτησης Περιοδικότητας 58
3.3 Εξαγωγή Χαρακτηριστικών με Τεχνικές μη Επιβλεπόμενης Μάθησης 61
3.3.1 Principal Component Analysis 61
3.3.2 Restricted Boltzmann Machines 64
Κεφάλαιο 4: Αυτόματη Ρυθμική Κατηγοριοποίηση **69**
4.1 Εισαγωγή 69
4.2 Πειραματικά Αποτελέσματα 72
4.2.1 Αυτόματη Κατηγοριοποίηση Χορευτικής Μουσικής 72

4.2.2 Αυτόματη Εξαγωγή Μουσικού Κλειδιού 75
4.3 Ανάλυση της Επίδοσης των Χαρακτηριστικών 80
Κεφάλαιο 5: Εξαγωγή Μουσικού Τέμπο 83
5.1 Εισαγωγή 83
5.2 Εξαγωγή Τέμπο με Χρήση Μετρικών Σχέσεων 84
5.3 Εξαγωγή Τέμπο από την Μουσική Ταχύτητα με Χειρωνακτικά Χαρακτηριστικά .. 86
5.4 Εξαγωγή Τέμπο ως Πολλαπλά Προβλήματα Κατηγοριοποίησης με Αυτόματα Χαρακτηριστικά 90
5.5 Πειραματικά Αποτελέσματα 93
5.5.1 Εισαγωγή 93
5.5.3 Αξιολόγηση της μεθόδου εξαγωγής της μουσικής ταχύτητας και μουσικού τέμπο με χειρονακτικά χαρακτηριστικά και SVM ταξινομητή 98
5.5.4 Αξιολόγηση της μεθόδου εξαγωγής τέμπο με πολλαπλούς ταξινομητές σε αυτόματα χαρακτηριστικά. 101
5.5.5 Σύγκριση των μεθόδων με την διεθνή βιβλιογραφία 104
Κεφάλαιο 6 : Εξαγωγή Παλμού 110
6.1 Εισαγωγή 110
6.2 Αρχιτεκτονική της Προτεινόμενης Μεθόδου 112
6.3 Εξαγωγή Τοπικού Τέμπο και Μεταβολών 114
6.4 Συνάρτηση Έμφασης 118
6.5 Εξαγωγή Ακολουθίας Παλμών 118
6.6 Αξιολόγηση της Μεθόδου Εξαγωγής Ακολουθίας Παλμών 121
Κεφάλαιο 7: Προσεγγίζοντας έναν Ρυθμικό Μετασχηματισμό 126
7.1 Εισαγωγή 126
7.2 Προτεινόμενη Μέθοδος 128
7.2.1 Προεπισκόπηση Μεθόδου 128
7.2.2 Εξαγωγή και Ανακατασκευή Συνάρτησης Έμφασης 129
7.2.3 Ανάλυση Περιοδικότητας 129
7.2.4 Εκμάθηση Χαρακτηριστικών με RBM 133
7.2.5 Ρυθμική Κατηγοριοποίηση 133
7.3 Πειραματικά Αποτελέσματα 134
7.3.1 Αξιολόγηση της Αντιστρεψιμότητας 134
7.3.2 Πειραματικά Αποτελέσματα Κατηγοριοποίησης 136
7.3.3 Δειγματοληψία από Ρυθμικές Κλάσεις 138
Κεφάλαιο 8: Μουσική Ομοιότητα Βάσει Περιεχομένου 141
8.1 Εισαγωγή 141
8.2 Εύρεση Μουσικής Ομοιότητας 143
8.2.1 Προεπισκόπηση 143
8.2.2 Εξαγωγή Χαρακτηριστικών 144

8.2.3 Υπολογισμός Αποστάσεων Μουσικών Κομματιών 147
8.3 Πειραματικά Αποτελέσματα 148
8.4 Η Διαδικτυακή Πλατφόρμα Ανάκτησης Μουσικής 151
Κεφάλαιο 9: Συμπεράσματα και Σχολιασμός 155
9.1 Συνεισφορά της διατριβής 155
9.2 Κατευθύνσεις Μελλοντικής Έρευνας 157
Παραπομπές 159

ΠΕΡΙΛΗΨΗ

Σκοπός της παρούσας διατριβής είναι η ανάπτυξη τεχνικών για την αυτόματη ανάλυση μουσικών σημάτων. Μεγάλη έμφαση δίνεται στην αυτόματη ανάλυση του ρυθμού, ο οποίος αποτελεί ένα θεμελιώδες μέγεθος της μουσικής, αφού ορίζει την χρονική δομή και οργάνωση του μουσικού σήματος. Αποτελεί δομικό στοιχείο σε ένα σύστημα αυτόματης μεταγραφής (transcription) της μουσικής, ενώ η ρυθμική πληροφορία μπορεί να χρησιμοποιηθεί και σε άλλες σημαντικές εφαρμογές όπως η δεικτοδότηση και ανάκτηση βάσει περιεχομένου ο υπολογισμός ρυθμικής ομοιότητας και η μίξη σημάτων.

Εξέχουσα σημασία στην ανάπτυξη ενός συστήματος αυτόματης ανάλυσης ρυθμικού περιεχομένου είναι η συνάρτηση περιοδικότητας. Η συνάρτηση περιοδικότητας αποτελεί το «ρυθμικό φάσμα» ενός μουσικού σήματος, αφού μας δίνει την ισχύ των διάφορων περιοδικοτήτων. Ένα μεγάλο μέρος της παρούσας διατριβής είναι αφιερωμένο στην εξαγωγή και επεξεργασία μιας συνάρτησης περιοδικότητας.

Χρησιμοποιώντας μοντέρνες τεχνικές επεξεργασίας σήματος όπως τον διαχωρισμό πηγών, προτείνεται μια συνάρτηση περιοδικότητας, η οποία αποτελεί μια εύρωστη αναπαράσταση του ρυθμικού περιεχομένου. Στη συνέχεια γίνεται επεξεργασία της συνάρτησης περιοδικότητας με τεχνικές μη επιβλεπόμενης μάθησης για την εξαγωγή συμπαγών χαρακτηριστικών. Τα χαρακτηριστικά αυτά χρησιμοποιούνται σε δύο προβλήματα ρυθμικής κατηγοριοποίησης: την αυτόματη κατηγοριοποίηση βάσει ρυθμικής κλάσης και την εξαγωγή του χρονικού κλειδιού. Στη συνέχεια προτείνονται τρεις διαφορετικές μέθοδοι εξαγωγής του τέμπο από τη συνάρτηση περιοδικότητας και τα χαρακτηριστικά της καθώς και μια τεχνική εξαγωγής των θέσεων του μουσικού παλμού έχοντας γνώση του τέμπο.

Επίσης, η συνάρτηση περιοδικότητας τροποποιείται έτσι ώστε να είναι προσεγγιστικά αντιστρέψιμη, δηλαδή να είναι εφικτό να ανακατασκευαστεί ένα μουσικό σήμα από την συνάρτηση περιοδικότητας τέτοιο ώστε να διατηρεί τη ρυθμική δομή του αρχικού σήματος.

Τέλος, τα ρυθμικά χαρακτηριστικά πλαισιώνονται από χαρακτηριστικά «χροιάς» και «αρμονίας» προκειμένου να δημιουργηθεί ένα ολοκληρωμένο σύστημα μουσικής ομοιότητας βάσει περιεχομένου. Το σύστημα αυτό ενσωματώθηκε σε μια διαδικτυακή πλατφόρμα αναζήτησης μουσικής βάσει περιεχομένου.

Συνοψίζοντας, στα πλαίσια της παρούσας διατριβής αντιμετωπίστηκαν έξι διαφορετικά προβλήματα μουσικής ανάλυσης. Η εύρεση του μουσικού κλειδιού, η εύρεση χορευτικού στυλ, η εξαγωγή του τέμπο, η εξαγωγή του παλμού, και ο υπολογισμός της ομοιότητας βάσει περιεχομένου μεταξύ δύο μουσικών κομματιών. Επιπλέον, προτάθηκε τρόπος υπολογισμού μιας «αντιστρέψιμης» συνάρτησης περιοδικότητας. Οι προτεινόμενοι μέθοδοι αξιολογήθηκαν για όλα τα προβλήματα σε μεγάλο εύρος δεδομένων και συγκρινόμενες με άλλες μεθόδους αιχμής, πέτυχαν ανταγωνιστικά και σε αρκετές περιπτώσεις καλύτερα αποτελέσματα από οποιαδήποτε άλλη μέθοδο.

ABSTRACT

The purpose of this thesis is to develop techniques for automatic analysis of musical signals. Great emphasis is placed on automatic analysis of rhythm, which is a fundamental characteristic of music, as it describes the temporal structure and organization of the music signal. Rhythm is a structural element in an automated music transcription system, and the rhythmic information can be used in other important applications such as indexing and retrieval of music content calculation of rhythmic similarity and mixing of music signals.

Paramount importance to the development of an automated analysis system of rhythmic content is the periodicity function. The periodicity function is the "rhythmic spectrum" of a music signal, as it demonstrates the salience of different targeted periodicities. A large part of this thesis is dedicated to the extraction and processing of a periodicity function.

Using modern signal processing techniques such as source separation, we propose a periodicity function, which is a robust representation of rhythmic content. Then we apply unsupervised learning techniques on the periodicity function for extracting solid rhythm features. These features are used in two problems of rhythmic categorization, the automatic categorization in rhythmic classes and the extraction of time key. We propose three different tempo extraction methods based on the periodicity function and the extracted features, as well as a method for beat tracking.

Furthermore, the periodicity function is redefined so that it can be approximately reversible, i.e. it is possible to reconstruct a music signal from the periodicity function that maintains the rhythmic structure of the original signal.

Finally, the rhythmic features are extended with the incorporation of timbral and harmonic features in order to build a content-based music similarity system. This system was integrated in a content based music search web platform.

In summary, this thesis deals with six distinct music analysis problems, namely, music key extraction, dance style classification, tempo estimation, beat tracking, and content based similarity between music tracks. Furthermore, it proposes a method for the calculation of a "reversible" periodicity function. The proposed methods were evaluated for all the problems in a wide range of data sets and compared with other state of the art methods achieving competitive and in some cases even better results.

Κεφάλαιο 1 :Εισαγωγή

1.1 Πρόλογος

Η ευρεία κατανάλωση μουσικής μέσω του διαδικτύου έχει αυξήσει το ενδιαφέρον της ερευνητικής κοινότητας και της μουσικής βιομηχανίας για την Ανάκτηση Μουσικής Πληροφορίας (MIR: Music Information Retrieval). Μερικά από τα επιμέρους προβλήματα που αντιμετωπίζονται στο πλαίσιο της Ανάκτησης Μουσικής Πληροφορίας είναι τα εξής:

- Εξαγωγή χαρακτηριστικών χαμηλού επιπέδου από το μουσικό σήμα
- Εξαγωγή υψηλού επιπέδου περιγραφέων (descriptors) ή χαρακτηριστικών
- Μοντελοποίηση μουσικής ομοιότητας
- Αυτόματη μεταγραφή (transcription)
- Αυτόματη ομαδοποίηση και κατηγοριοποίηση μουσικών αποσπασμάτων
- Συστήματα αυτόματης σύστασης (recommendation) βάσει μουσικών προτιμήσεων (music taste)

Στην αρχή της ανάπτυξης των τεχνολογιών επεξεργασίας μουσικής πληροφορίας η ερευνητική κοινότητα δοκίμασε υπάρχουσες τεχνικές από το πεδίο της επεξεργασίας φωνής. Παρότι οι τεχνικές αυτές σημείωσαν κάποια σχετική επιτυχία δεν αποδείχτηκαν επαρκείς. Αυτό οφείλεται στο γεγονός της τελείως διαφορετικής φύσης του μουσικού σήματος από το σήμα φωνής. Το μουσικό σήμα ακολουθεί μια αυστηρή δομή στον χώρο χρόνου-συχνότητας που υποβάλλεται από τη μουσική σύνθεση. Σχετικά με την χρονική οργάνωση τα «ηχητικά γεγονότα» (sound events) όπως οι νότες συμβαίνουν σε (σχεδόν) προκαθορισμένες στιγμές όπως υπαγορεύει η ρυθμική/χρονική δομή ενός μουσικού κομματιού. Η μουσική αρμονία από την άλλη περιορίζει το συχνοτικό περιεχόμενο σε συγκεκριμένες ζώνες συχνοτήτων (π.χ. 12 κεντρικές συχνότητες ανά οκτάβα στη δυτική μουσική). Ενώ κάποιος θα μπορούσε να ισχυριστεί ότι οι περιορισμοί αυτοί θα διευκόλυναν την επεξεργασία και ανάλυση των μουσικών σημάτων, στην πραγματικότητα συμβαίνει το αντίθετο. Ο περιορισμός του χρονοσυχνοτικού χώρου σε στενές ζώνες δημιουργεί μεγάλες επικαλύψεις των ηχητικών πηγών. Δύο νότες συχνά θα ακουστούν ακριβώς την ίδια χρονική στιγμή, με πολύ μεγάλη επικάλυψη στο συχνοτικό περιεχόμενο. Η δυσκολία της ανάλυσης των μουσικών σημάτων επαληθεύεται και από ένα απλό παράδειγμα: ενώ ο άνθρωπος μπορεί να διακρίνει με πολύ μεγάλη ακρίβεια τη φωνή εκατοντάδων ανθρώπων, και μάλιστα σε πολύ δύσκολες συνθήκες (θόρυβος, χαμηλή ποιότητα τηλεφωνικής γραμμής κ.ο.κ.), ακόμα και πεπειραμένοι μουσικοί δυσκολεύονται να διακρίνουν μεταξύ τους τα περίπου 20 συμφωνικά όργανα [Eronen2001].

Μια άλλη ιδιαιτερότητα αποτελεί ο υποκειμενικός ανθρώπινος παράγοντας στην ακρόαση της μουσικής. Η μουσική αντίληψη διαφέρει από άνθρωπο σε άνθρωπο και οφείλεται σε διάφορους παράγοντες (κοινωνικούς, πολιτιστικούς, ψυχολογικούς). Ακόμα και ο ίδιος άνθρωπος μπορεί να αντιληφθεί με διαφορετικό τρόπο κάποια μουσικά χαρακτηριστικά ανάλογα με την ψυχολογική κατάσταση ή το περιβάλλον. Για παράδειγμα κάποιος μπορεί να αντιληφθεί διαφορετικά τον ρυθμό ενός μουσικού κομματιού ή ένα μουσικό κομμάτι να του προκαλέσει διαφορετικά συναισθήματα σε διαφορετικές ώρες της ημέρας ή σε διαφορετικές συνθήκες φυσικής κατάστασης. Ένα μουσικό κομμάτι μπορεί να χαρακτηριστεί ότι ανήκει στην κατηγορία «τζαζ» από έναν άνθρωπο ή rhythm ‘n

blues από κάποιον άλλο. Το ίδιο συμβαίνει και με άλλες πτυχές της μουσικής όπως η μουσική ομοιότητα δύο κομματιών, το μουσικό στυλ και το συναίσθημα/διάθεση (mood).

Σκοπός της παρούσας διδακτορικής διατριβής είναι η αυτόματη ανάλυση μουσικών σημάτων με έμφαση στο ρυθμικό περιεχόμενο και η επέκτασή της στην έννοια της μουσικής ομοιότητας. Ως πρώτο στάδιο ορίζεται η *συνάρτηση περιοδικότητας*, μία αναπαράσταση η οποία περιγράφει το ρυθμικό περιεχόμενο ενός μουσικού κομματιού. Η ανάλυση περιοδικότητας συνίσταται από διαδοχικές διεργασίες όπου από την κυματομορφή εξάγεται μία αναπαράσταση χρόνου – συχνότητας. Στην χρονοσυχνοτική αυτή αναπαράσταση εφαρμόζεται μία μέθοδος διαχωρισμού του σήματος σε «αρμονικές» και «κρουστές» πηγές. Από τις δύο συνιστώσες του σήματος εξάγονται δύο πολυδιάστατες ακολουθίες χαρακτηριστικών. Η ανάλυση περιοδικότητας επιτυγχάνεται με την επεξεργασία των ακολουθιών αυτών από μια συστοιχία ταλαντωτών.

Στη συνέχεια διερευνώνται μέθοδοι εξαγωγής πιο περιγραφικών και συμπαγών χαρακτηριστικών από το διάνυσμα περιοδικότητας. Πιο συγκεκριμένα, εξετάζονται και συγκρίνονται 3 είδη χαρακτηριστικών: α) χειρονακτικά χαρακτηριστικά, β) χαρακτηριστικά βάσει της τεχνικής «Ανάλυση σε Κύριες Συνιστώσες» (Principal Component Analysis) και γ) με την χρήση των Περιορισμένων Μηχανών Boltzmann (Restricted Boltzmann Machines, RBM).

Με την παραδοχή ότι τα μουσικά κομμάτια παρουσιάζουν σχεδόν σταθερό τέμπο (δηλ. διακυμάνσεις σε ένα περιορισμένο εύρος) αναπτύσσονται δύο μέθοδοι εξαγωγής του μουσικό τέμπο. Η πρώτη μέθοδος ενσωματώνει στοιχειώδεις κανόνες στις σχέσεις των μετρικών επιπέδων, ενώ η δεύτερη χρησιμοποιεί τεχνικές μηχανικής μάθησης, όπου το πρόβλημα εύρεσης του τέμπο μετασχηματίζεται σε πολλαπλά προβλήματα κατηγοριοποίησης. Στη συνέχεια, βάσει του εξαγόμενου τέμπο αναπτύσσεται μία μέθοδος εύρεσης του μουσικού παλμού η οποία βασίζεται σε δυναμικό προγραμματισμό. Οι προτεινόμενες μέθοδοι αξιολογήθηκαν σε μια πληθώρα συλλογών και επέδειξαν ανταγωνιστικές επιδόσεις συγκριτικά με τις μεθόδους που έχουν προταθεί στη διεθνή βιβλιογραφία.

Τα προτεινόμενα χαρακτηριστικά αποδεικνύονται πολύ καλοί περιγραφείς του ρυθμικού περιεχομένου ενός μουσικού κομματιού. Με χρήση συνήθων τεχνικών μηχανικής μάθησης και κατηγοριοποίησης, υλοποιήθηκε μια μέθοδος που αντιμετωπίζει δύο διαφορετικά (distinct) προβλήματα ρυθμικής ανάλυσης: (α) την εξαγωγή μέτρου (ή χρονικού κλειδιού) και (β) κατηγοριοποίηση με βάση το είδος χορευτικής μουσικής (Dance Style Classification).

Τα προτεινόμενα ρυθμικά χαρακτηριστικά συνδυάζονται με άλλα κοινά χαρακτηριστικά που συναντάμε στη βιβλιογραφία τα οποία περιγράφουν την μουσική χροιά, προκειμένου να υλοποιηθεί ένα ολοκληρωμένο σύστημα Ανάκτησης Μουσικής Πληροφορίας. Σε μια μεγάλη μουσική συλλογή αποτελούμενη από περίπου 130.000 μουσικά αποσπάσματα, υπολογίστηκαν οι «μουσικές αποστάσεις» μεταξύ των μουσικών κομματιών, αποκλειστικά βάσει του ακουστικού περιεχομένου (ρυθμικού και ηχοχρώματος). Έχοντας ως είσοδο κάποιο μουσικό κομμάτι (query), ο χρήστης μπορεί να πλοηγηθεί σε όμοια μουσικά κομμάτια σε ένα online ολοκληρωμένο σύστημα μουσικής ανάκτησης.

Τέλος, προτείνεται ένας ρυθμικός μετασχηματισμός που έχει το χαρακτηριστικό ότι είναι σχεδόν αναστρέψιμος. Η διαδικασία ανάλυσης περιοδικότητας τροποποιείται έτσι ώστε να είναι εφικτή η ανακατασκευή ενός σήματος από το διάνυσμα περιοδικότητας το οποίο να διατηρεί τις βασικές ρυθμικές ιδιότητες του αρχικού σήματος. Στη συνέχεια μια συστοιχία Μηχανών

Boltzmann «μαθαίνει» χαρακτηριστικά από το διάνυσμα περιοδικότητας, τα οποία χρησιμοποιούνται επιτυχώς στα προβλήματα της εξαγωγής μέτρου και της κατηγοριοποίησης με βάση το είδος της χορευτικής μουσικής. Η προτεινόμενη αρχιτεκτονική, δίνει επιπλέον τη δυνατότητα *δημιουργίας τυχαίων ακουστικών παραδειγμάτων* των διαφόρων κατηγοριών.

Συνοψίζοντας, οι βασικές συνεισφορές της παρούσας Διδακτορικής Διατριβής μπορούν να συνοψιστούν στα εξής:

- Προτείνεται μια εκλεπτυσμένη μέθοδος εξαγωγής του διανύσματος περιοδικότητας.
- Προτείνεται μια νέα μέθοδος διαχωρισμού του αρμονικού και του κρουστού περιεχομένου ενός μουσικού κομματιού.
- Ενσωματώνεται σε ένα σύστημα ρυθμικής ανάλυσης ο διαχωρισμός αρμονικού / κρουστού περιεχομένου του σήματος.
- Γίνεται εκτεταμένη μελέτη εξαγωγής συμπαγών και περιγραφικών χαρακτηριστικών του διανύσματος περιοδικότητας.
- Χρησιμοποιούνται τα εξαγόμενα χαρακτηριστικά για την αντιμετώπιση διαφορετικών προβλημάτων ρυθμικής ανάλυσης χωρίς καμία εκ των προτέρων γνώση για το πρόβλημα.
- Επιτυγχάνεται η δημιουργία ενός «κατά προσέγγιση αντιστρέψιμου» ρυθμικού μετασχηματισμού.

Όλες οι παραπάνω συνεισφορές δεν είχαν ξανασυναντηθεί πρότερα στη διεθνή βιβλιογραφία.

Η παρούσα διατριβή οργανώνεται ως εξής: Στη συνέχεια αυτού του Κεφαλαίου θα παρουσιαστούν κάποιες θεμελιώδεις έννοιες γύρω από την θεωρία της μουσικής καθώς και μια γενική βιβλιογραφική επισκόπηση του πεδίου της Ανάκτησης Μουσικής Πληροφορίας με έμφαση στην ανάλυση του ρυθμικού περιεχομένου. Στο Κεφάλαιο 2 παρουσιάζεται η προτεινόμενη μέθοδος εξαγωγής ανάλυσης περιοδικότητας, και στο Κεφάλαιο 3 η εξαγωγή χαρακτηριστικών από το διάνυσμα περιοδικότητας. Στο Κεφάλαιο 4 αξιολογούνται τα εξαγόμενα χαρακτηριστικά στην κατηγοριοποίηση ρυθμού και συγκεκριμένα στην κατηγοριοποίηση μέτρου και του είδους χορευτικής μουσικής. Τα Κεφάλαια 5 και 6 πραγματεύονται δύο βασικά προβλήματα ανάλυσης ρυθμού, της εξαγωγής τέμπο (tempo estimation) και της εξαγωγής παλμού (beat tracking) αντίστοιχα. Το Κεφάλαιο 7 πραγματεύεται τον αντίστροφο ρυθμικό μετασχηματισμό ενώ στο Κεφάλαιο 8 παρουσιάζεται το σύστημα ανάκτησης μουσικής πληροφορίας. Τέλος, το Κεφάλαιο 9 συνοψίζει με σχολιασμό και συμπεράσματα την διατριβή.

1.2 Μουσικές Έννοιες και το Μουσικό Σήμα

1.2.1 Οι διαστάσεις του ήχου

Ο όρος «διαστάσεις» του ήχου ανταποκρίνεται στις ποσότητες τις οποίες τις αντιλαμβάνεται το ανθρώπινο αυτί και που με αυτές μπορούμε να χαρακτηρίσουμε ή να ξεχωρίσουμε έναν ήχο. Οι διαστάσεις του ήχου θεωρούνται ότι είναι τέσσερις. Οι τρεις βασικές είναι η ένταση (loudness), η χρονική διάρκεια (duration) και το τονικό ύψος (pitch). Η τέταρτη και πιο σημαντική είναι η «χροιά» (timbre). Οι τρεις βασικές διαστάσεις μπορούν να οριστούν ως μονοδιάστατοι αριθμοί ενώ η χροιά θεωρείται ότι έχει πολλές διαστάσεις. Ωστόσο πολλοί ερευνητές έχουν δώσει και στο τονικό ύψος πολυδιάστατη υφή.

Ένταση

Η ένταση ήχου μπορεί να οριστεί ως μια ποσότητα σύμφωνα με την οποία ένας ήχος μεγαλύτερης έντασης από έναν άλλο ακούγεται από τον άνθρωπο πιο «ισχυρός». Ο επίσημος ορισμός (ANSI, 1973) δίνεται ως:

> *Ένταση είναι το χαρακτηριστικό γνώρισμα της αίσθησης της ακοής που έχει να κάνει με το ποιοι ήχοι μπορούν να ταξινομηθούν σε μια κλίμακα που ξεκινάει από το «απαλό» (soft) και φτάνει στο «έντονο» (loud).*

Η ένταση ενός ήχου εξαρτάται από την ακουστική ενέργεια που φτάνει στη θέση του ακροατή, τη διάρκειά του και το φασματικό περιεχόμενο. Ένας απλός αλλά αποτελεσματικός τρόπος να μετρηθεί η ένταση ενός ήχου είναι να αθροιστεί η ενέργεια σε συγκεκριμένες περιοχές του φάσματος που ονομάζονται «κρίσιμες ζώνες» (critical bands)

Διάρκεια

Η χρονική διάρκεια δεν έχει μελετηθεί τόσο όσο οι άλλες διαστάσεις του ήχου. Οι άνθρωποι αναγνωρίζουν ευκολότερα ήχους με μεγαλύτερη διάρκεια και αυτό μπορεί να παίζει κάποιο ρόλο στην αναγνώριση ήχων από τις μηχανές.

Τονικό ύψος

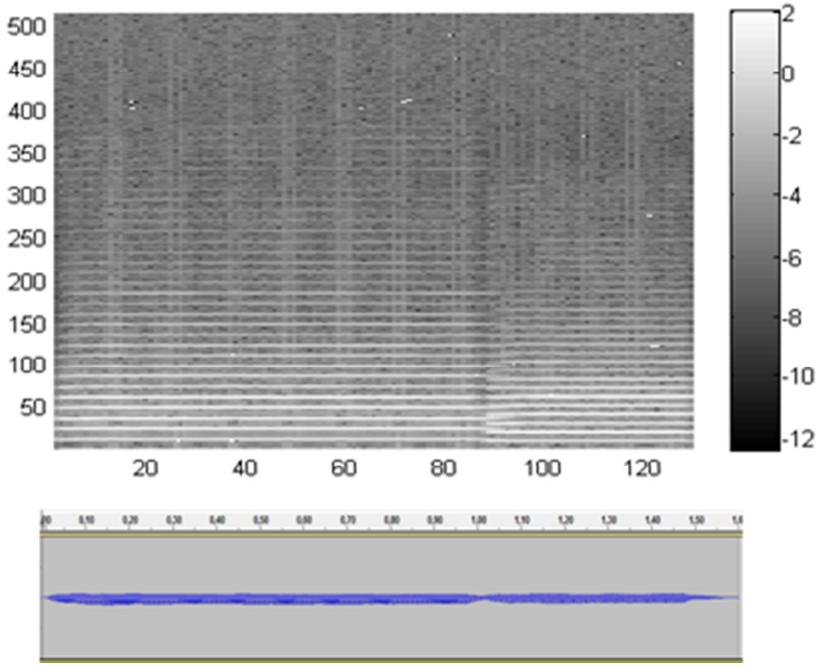
Το τονικό ύψος είναι η πιο σημαντική από τις τρεις βασικές διαστάσεις του ήχου. Ο επίσημος ορισμός κατά ANSI του τονικού ύψους είναι:

> *Το χαρακτηριστικό γνώρισμα της αίσθησης της ακοής με την οποία μπορούμε να κατατάξουμε ήχους σε μια κλίμακα από τον υψηλότερο στον χαμηλότερο*

Το τονικό ύψος πιο πρακτικά μπορεί να οριστεί και ως η συχνότητα του ημιτονοειδή ήχου που ταιριάζει καλύτερα σε έναν ήχο. Συνδέεται με την περιοδικότητα ενός ήχου και η συχνότητα του τονικού ύψους μπορεί να οριστεί ως ο αντίστροφος αριθμός της περιοδικότητας. Το τονικό ύψος ενός ήχου προσδιορίζει τον τόνο / νότα του («Λα», «Ντο»). Η αντίληψη του pitch είναι υποκειμενική και εξαρτάται τόσο από τη συχνότητα όσο και από το επίπεδο της πίεσης του ήχου. Η θεμελιώδης συχνότητα (F_0) είναι η αντίστοιχη φυσική σημασία του όρου και ορίζεται μόνο για περιοδικά ή σχεδόν περιοδικά σήματα. Σε αυτή τη περίπτωση, η θεμελιώδης συχνότητα ορίζεται ως το αντίστροφο της περιόδου και είναι στενά συνδεδεμένη με το pitch.

Χροιά

Τι είναι αυτό που μας κάνει να ξεχωρίζουμε τη φωνή ενός ανθρώπου από έναν άλλο; Τον ήχο μιας φλογέρας από αυτόν ενός φλάουτου; Αν μία νότα που ακούμε από κάποιο μουσικό όργανο είναι καθαρή ή παιγμένη από κάποιον αρχάριο; Κάθε προσπάθεια ορισμού της «ποιότητας» ενός ήχου, ή χαρακτηρισμού της πηγής του, βρίσκει μπροστά του την έννοια του «ηχοχρώματος» ή «χροιάς», που αποδίδεται από τον αγγλικό όρο «timbre». Σε αντίθεση με τα άλλα 3 βασικά χαρακτηριστικά ενός ήχου, η χροιά έχει ασαφή ορισμό και αποτελεί ακόμα και σήμερα πεδίο έρευνας και διαμάχης. Η περιγραφή που πλησιάζει έναν κοινά αποδεκτό ορισμό και που δεν έχει αμφισβητηθεί τις τελευταίες δεκαετίες είναι η εξής:



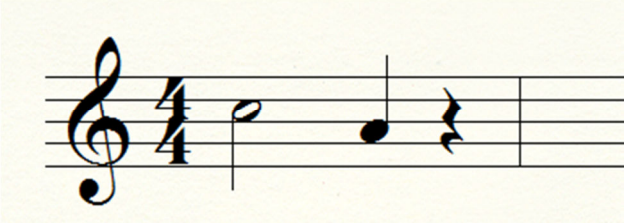
Σχήμα 1.1 Η κυματομορφή δύο διαδοχικών νοτών μιας τρομπέτας μαζί με το φασματογράφημα.

> *«Χροιά (Timbre) είναι εκείνη η ικανότητα της αίσθησης της ακοής με την οποία ένας ακροατής μπορεί να κρίνει ότι δύο ήχοι που παρουσιάζονται με τον ίδιο τρόπο και έχουν την ίδια ένταση και τόνο είναι διαφορετικοί. Η χροιά εξαρτάται κυρίως από το φάσμα της πηγής διέγερσης, αλλά επίσης από την κυματομορφή, την πίεση του αέρα, την θέση του φάσματος στις συχνότητες, και τα χρονικά χαρακτηριστικά της πηγής διέγερσης»* (*American Standards Association, 1960, p. 45*)

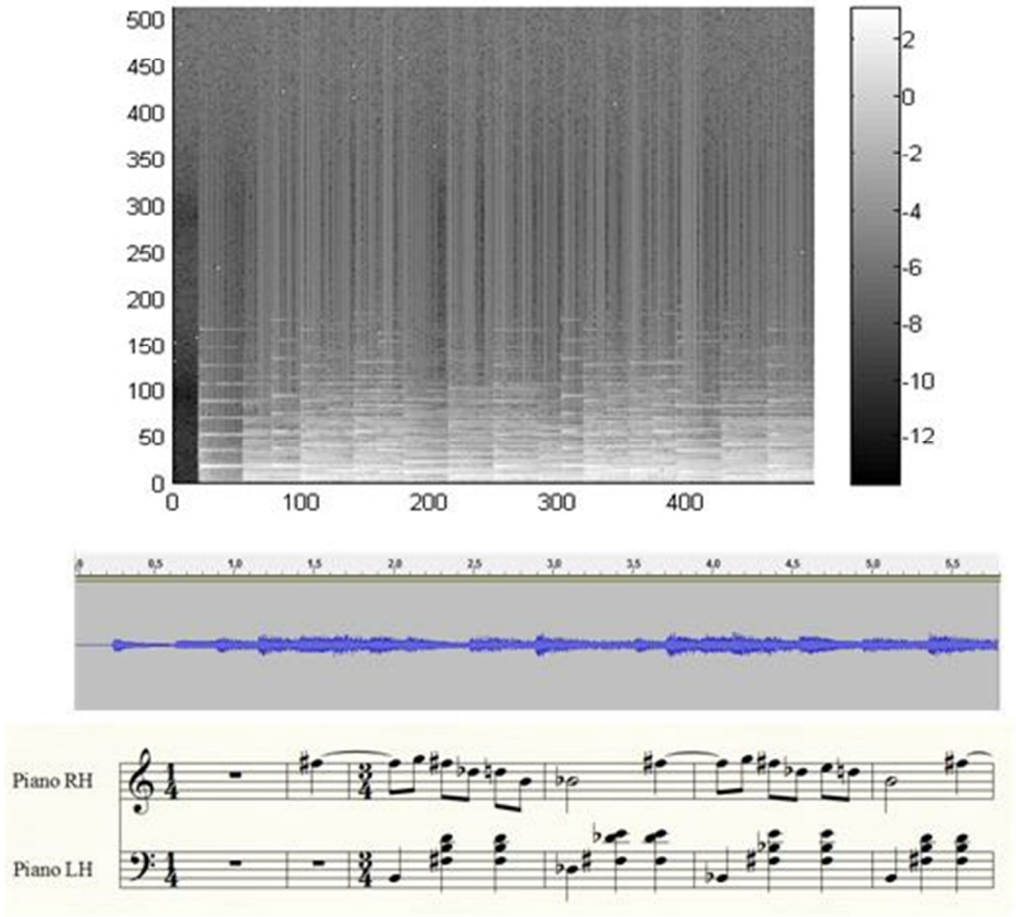
Όπως φαίνεται από τον παραπάνω ορισμό, η έννοια της χροιάς είναι ασαφής. Πιστεύεται ότι η χροιά, εφόσον αποδεκτούμε ότι μπορεί να χαρακτηριστεί από κάποια φυσική ποσότητα ή ιδιότητα, δεν είναι μονοδιάστατο μέγεθος όπως οι άλλες τρεις προαναφερθείσες ιδιότητες ενός ήχου, αλλά πολυδιάστατο. Μέχρι και τα μέσα του $20^{ου}$ αιώνα επικρατούσε η άποψη ότι η χροιά οφειλόταν αποκλειστικά στα στατικά φασματικά χαρακτηριστικά του ήχου, τόσο στο πλάτος όσο και στην φάση του. Πιο μοντέρνες θεωρίες όμως που είναι πλέον καθολικά αποδεκτές διατυπώνουν ότι και τα χρονικά χαρακτηριστικά (Temporal Features) είναι εξίσου, αν όχι πιο σημαντικά, από τα φασματικά. Η χροιά αποτελεί έννοια που σχετίζεται περισσότερο με την επιστήμη της ψυχοακουστικής, εκφράζοντας την αντιληπτική απόκριση του ανθρώπου σ'έναν ήχο. Όπως ο όρος «εμφάνιση» ή «όψη» χαρακτηρίζει ασαφώς μία εικόνα, έτσι και ο όρος «χροιά» χαρακτηρίζει έναν ήχο.

1.2.2 Η μουσική νότα

Η μουσική απαρτίζεται από ένα σύνολο μουσικών νοτών οι οποίες είναι οργανωμένες χρονικά. Στο Σχ. 1.1 φαίνεται η κυματομορφή ενός ήχου (μουσικό σήμα) και το αντίστοιχο φασματογράφημα, από μία τρομπέτα που παίζει δύο συνεχόμενες μεμονωμένες νότες.



Σχήμα 1.2 Η συμβολική αναπαράσταση στο πεντάγραμμο του σήματος του Σχ. 1.1



Σχήμα 1.3 Η κυματομορφή ενός αποσπάσματος μιας εκτέλεσης πιάνου μαζί με το φασματογράφημα και την αντίστοιχη συμβολική αναπαράσταση στο πεντάγραμμο.

Ο ήχος αυτός χαρακτηρίζεται από τα προαναφερθέντα μεγέθη και συμβολίζεται στο πεντάγραμμο όπως φαίνεται στο Σχ. 1.2. Η ενέργεια του σήματος σε συγκεκριμένες κρίσιμες ζώνες είναι ένας τρόπος να μετρηθεί η ένταση του ήχου.

Στο Σχ. 1.3 παρουσιάζεται ένα πιο πολύπλοκο μουσικό σήμα που αντιστοιχεί σε μία εκτέλεση πιάνου. Παρατηρούμε ότι το φασματικό περιεχόμενο σε αυτή τη περίπτωση είναι πολύ πιο πλούσιο, καθώς έχουμε επικαλύψεις τόσο στο χρόνο όσο και στη συχνότητα των μεμονωμένων νοτών που απαρτίζουν το μουσικό απόσπασμα.

Απλουστευτικά μπορούμε να θεωρήσουμε τον μηχανισμό παραγωγής του μουσικού ήχου ως ένα Γραμμικό Χρονικά Αναλλοίωτο (ΓΧΑ) σύστημα $h(n)$ που αντιστοιχεί στο σώμα του μουσικού οργάνου, το οποίο διεγείρεται από μία πηγή $x(n)$. Ο παραγόμενος ήχος $y(n)$ είναι η έξοδος του $h(n)$ όταν διεγείρεται από την πηγή $x(n)$:

$$y(n) = h(n) * x(n). \tag{1.1}$$

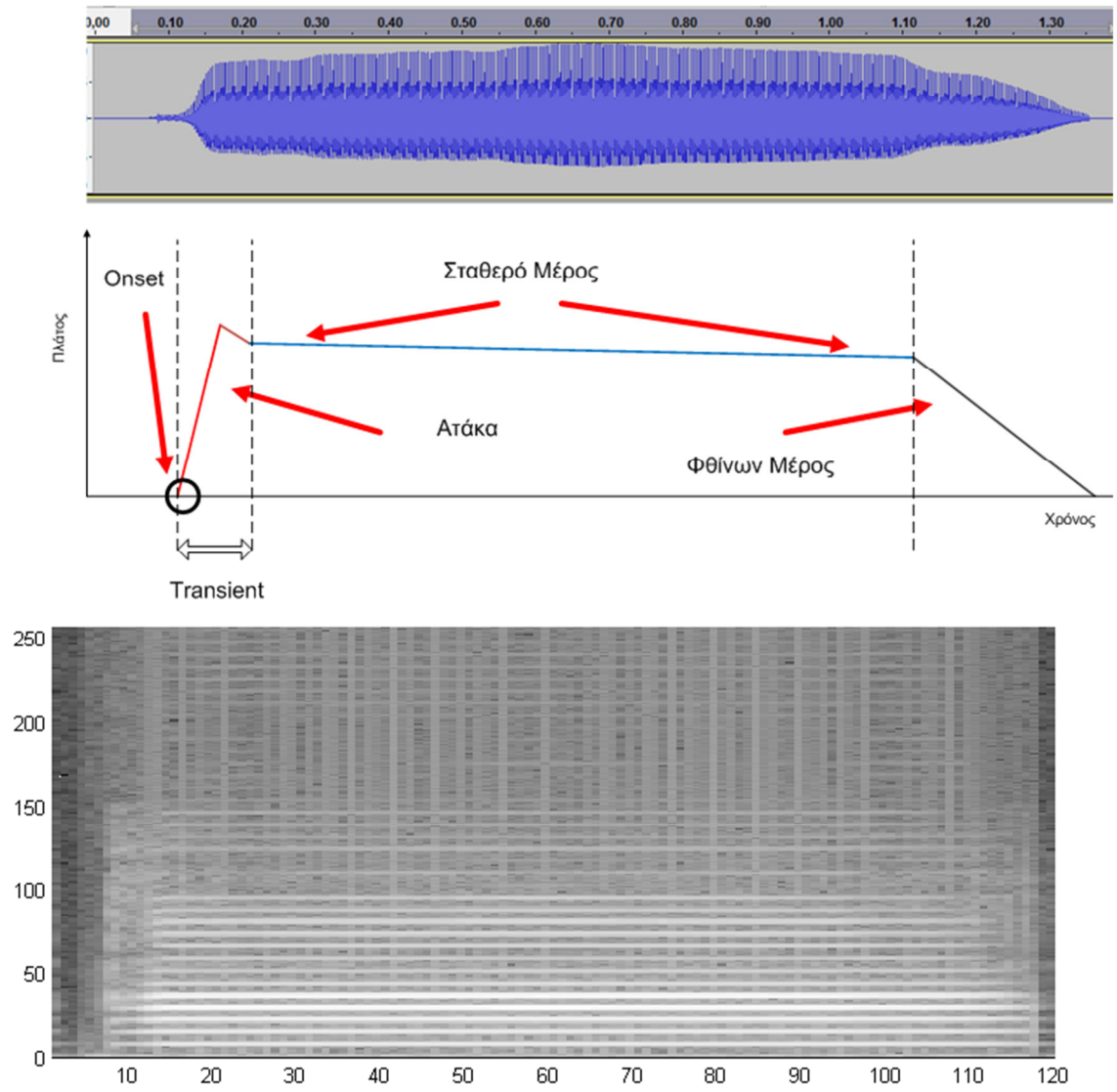
Προφανώς αυτή η μοντελοποίηση είναι απλουστευτική, καθώς αφενός οι αποκρίσεις των μουσικών οργάνων είναι προσεγγιστικά μόνο γραμμικές, αλλά υπάρχει και αλληλεπίδραση του ίδιου του μουσικού οργάνου με τον άνθρωπο, δηλαδή την πηγή. Επομένως υπάρχει εξάρτηση του $h(n)$ από το $x(n)$,

Υπάρχουν δύο κατηγορίες διεγέρσεων $x(n)$, που αντιστοιχούν σε δύο διαφορετικούς τύπους μηχανικής διέγερσης των μουσικών οργάνων. Η πρώτη κατηγορία περιλαμβάνει διεγέρσεις που έχουν μεγάλη διάρκεια και συνέχεια στο χρόνο, όπως είναι το φύσημα σε ένα πνευστό όργανο ή το γλίστρημα ενός δοξαριού στις χορδές του βιολιού. Τέτοιες διεγέρσεις μπορούν να προσεγγιστούν ως βηματικές συναρτήσεις. Η δεύτερη κατηγορία περιλαμβάνει διεγέρσεις που αποτελούνται από ένα μεμονωμένο κτύπημα, όπως για παράδειγμα το χτύπημα του πλήκτρου του πιάνου ή των χορδών μιας κιθάρας. Σε αυτή τη περίπτωση η διέγερση μπορεί να θεωρηθεί προσεγγιστικά η κρουστική συνάρτηση.

Ο παραπάνω διαχωρισμός γίνεται γιατί το είδος της διέγερσης αντανακλάται σε κάποιες από τις ιδιότητες των νοτών. Στο Σχ. 1.4 παρουσιάζεται η κυματομορφή μιας νότας παιγμένη από ένα όμποε (ξύλινο πνευστό κλασσικό όργανο), η γενική μορφή της περιβάλλουσας και το αντίστοιχο φασματογράφημα. Από το σχήμα της περιβάλλουσας μπορούμε να παρατηρήσουμε ότι η νότα αποτελείται από τρία διαδοχικά μέρη:

- «**Ατάκα**» **ή Attack**: Είναι το χρονικό διάστημα στην αρχή της νότας, στο οποίο η περιβάλλουσα του πλάτους του σήματος αυξάνει. Ξεκινάει την ίδια χρονική στιγμή με την διέγερση. Η η χρονική στιγμή έναρξης της νότας, ονομάζεται *onset*.
- «**Σταθερή Κατάσταση**» **(Steady State)**: Είναι το χρονικό διάστημα στο οποίο το σώμα του οργάνου έχει έρθει σε σταθερή κατάσταση και τα μεταβατικά φαινόμενα από την διέγερση έχουν εξαλειφθεί. Για παράδειγμα, στην σταθερή κατάσταση σε ένα πνευστό όργανο παρατηρούμε στάσιμα κύματα μέσα στον σωλήνα του πνευστού. Στην περίπτωση των οργάνων που διεγείρονται «κρουστικά» (πιάνο, κιθάρα) απουσιάζει η σταθερή κατάσταση και η φθίνουσα κατάσταση ακολουθεί αμέσως μετά το μεταβατικό στάδιο.
- «**Φθίνουσα Κατάσταση**» **(Decay State)**: Είναι το χρονικό διάστημα στο οποίο η διέγερση από τον εκτελεστή έχει σταματήσει και το μουσικό όργανο ταλαντώνεται ελεύθερα μέχρι να απελευθερωθεί όλη η ενέργεια στο περιβάλλον.

Σημαντική επίσης είναι και η έννοια του «μεταβατικού» ή «παροδικού» (transient) διαστήματος. Είναι δύσκολο να οριστεί επακριβώς αλλά ένας διαισθητικός ορισμός του μεταβατικού διαστήματος είναι το διάστημα στο οποίο παρατηρούνται γρήγορες και απρόβλεπτες μεταβολές του σήματος. Εναλλακτικά μπορούμε να πούμε ότι είναι το διάστημα των μεταβατικών φαινομένων μέχρι το μουσικό όργανο να έρθει στην σταθερή κατάσταση. Πρέπει να τονιστεί ότι όπως φαίνεται και στο Σχ. 1.4, το transient δεν ταυτίζεται με την «ατάκα».

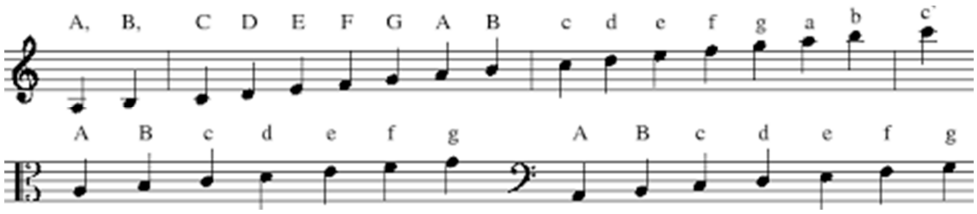


Σχήμα 1.4 Η νότα D#4 ενός όμποε (πνευστό ξύλινο κλασσικό όργανο).

Στο φασματογράφημα του Σχ. 1.4 παρατηρούμε τις αρμονικές του σήματος σε σχέση με τη θεμελιώδη συχνότητα που είναι ίση με ~311 Hz και αντιστοιχεί στον τόνο D#4. Παρατηρούμε ότι οι υψηλότερες αρμονικές θέλουν περισσότερο χρόνο να εμφανιστούν, ενώ κατά τη φάση φθίνουσας κατάστασης οι χαμηλόσυχνες συχνότητες φθίνουν πιο αργά. Επιπλέον, είναι φανερό ότι κατά την διάρκεια του μεταβατικού σταδίου δεν έχει διαμορφωθεί η αρμονική δομή του σήματος σε όλη την έκταση του φάσματος.

1.2.3 Αρμονική οργάνωση της μουσικής

Οι θεμελιώδεις συχνότητες των νοτών που συναντάμε στη μουσική δεν είναι τυχαίες, αλλά ακολουθούν μια αυστηρή δομή. Το συχνοτικό εύρος χωρίζεται σε οκτάβες και στη δυτική μουσική κάθε οκτάβα περιέχει 12 *τόνους*. Οι 12 τόνοι επαναλαμβάνονται σε κάθε οκτάβα και έχουν συγκεκριμένη ονομασία που είναι ίδια για όλες τις οκτάβες.



C	C#	D	D#	E	F	F#	G	G#	A	A#	B
Ντο	Ντο δίεση	Ρε	Ρε δίεση	Μι	Φα	Φα δίεση	Σολ	Σολ δίεση	Λα	Λαδίεση	Σι

Σχ. 1.5. Οι 12 φθόγγοι σε μία οκτάβα.

Κάθε τόνος περιγράφεται από ένα γράμμα που είναι η ταυτότητα του τόνου μέσα στην οκτάβα που ονομάζεται *φθόγγος*, και έναν αριθμό, που δηλώνει τον α/α της οκτάβας. Οι 12 φθόγγοι μέσα σε μία οκτάβα καθώς και η αναπαράστασή τους στο πεντάγραμμο παρουσιάζονται στο Σχ. 1.5. Επομένως με D3 συμβολίζουμε την νότα Ρε στην οκτάβα 3. Ο λόγος των κεντρικών συχνοτήτων δύο συνεχόμενων τόνων είναι ίσος πάντα με $\sqrt[12]{2}$. Επομένως η κατανομή των μουσικών συχνοτήτων είναι *λογαριθμική*.

Στον Πίνακα 1.1 παρουσιάζονται οι κεντρικές συχνότητες όλων των τόνων σε όλες τις οκτάβες. Παρατηρούμε ότι το ακουστικό εύρος των μουσικών τόνων είναι εννέα οκτάβες. Οι τόνοι που έχουν την ίδια ταυτότητα (ίδιο γράμμα: *φθόγγος*) δίνουν την ίδια αίσθηση του τόνου αλλά σε άλλη οκτάβα. Αυτό συμβαίνει γιατί οι θεμελιώδεις συχνότητες του ίδιου τόνου σε διαφορετικές οκτάβες θα έχουν λόγο που είναι ακέραιο πολλαπλάσιο του 2. Επειδή συχνά συγχέονται κάποιες έννοιες μεταξύ τους, όπως για παράδειγμα αυτές της νότας και του τόνου, σε αυτό το σημείο θα αποσαφηνιστούν κάποιες έννοιες καθώς και θα οριστούν νέες.

- Με τον όρο ***νότα*** εννοούμε ένα συγκεκριμένο μουσικό γεγονός, που περιγράφεται από 4 κύρια χαρακτηριστικά (Σχ. 1.4):
  - Την χρονική στιγμή έναρξης (onset)
  - Την διάρκεια
  - Τον ***τόνο***
  - Το μουσικό όργανο
- Ο ***τόνος*** (π.χ. C5, Πίνακας 1.1) είναι το χαρακτηριστικό της νότας που συνδέεται μονοσήμαντα με την θεμελιώδη συχνότητα (pitch). Ωστόσο συμβαίνει πολύ συχνά στην καθομιλουμένη οι όροι τόνος και νότα να συγχέονται. Η φράση «η νότα ντο» είναι λάθος, ενώ «ο τόνος ντο» είναι το ακριβές.
- **Συγχορδία** είναι το φαινόμενο της μουσικής στο οποίο παίζονται ταυτόχρονα κάποιες νότες, είτε από το ίδιο είτε από διαφορετικά μουσικά όργανα.
- **Διάστημα** μεταξύ δύο τόνων είναι ο λόγος των θεμελιωδών συχνοτήτων των δύο τόνων. Τα διαστήματα μετρούνται σε **ημιτόνια**.
- **Ημιτόνιο** είναι ο λόγος των θεμελιωδών συχνοτήτων δύο διαδοχικών τόνων (π.χ. E5 – F5) και είναι πάντα ίσος με $\sqrt[12]{2}$.
- **Αρμονικές** ενός τόνου είναι οι αρμονικές συχνότητες (ακέραια πολλαπλάσια) της θεμελιώδους συχνότητας του τόνου. Κάθε όργανο που παράγει ένα τόνο παράγει και αρμονικές αυτού του τόνου

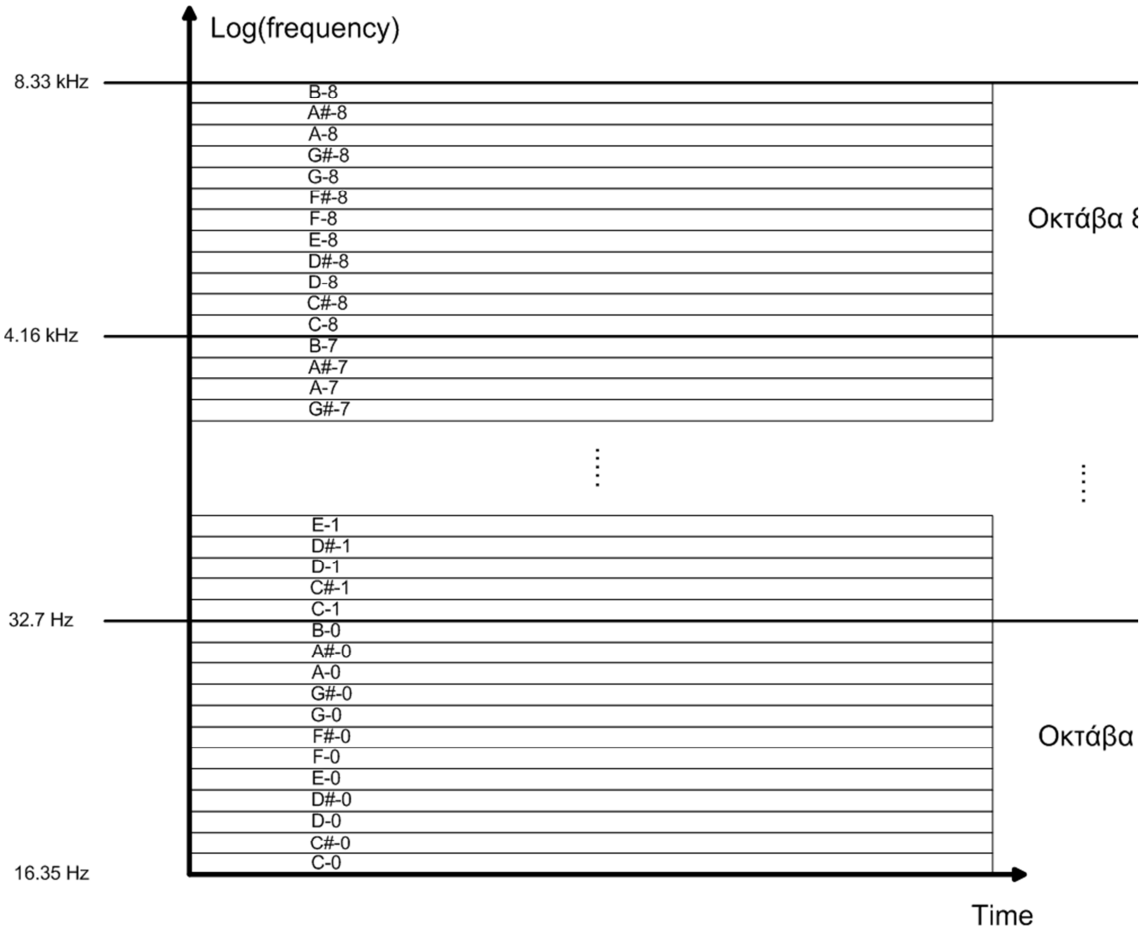
	Τόνος											
Οκτάβα	**C**	**C#**	**D**	**D#**	**E**	**F**	**F#**	**G**	**G#**	**A**	**A#**	**B**
0	16.35	17.32	18.35	19.45	20.60	21.83	23.12	24.50	25.96	27.50	29.14	30.87
1	32.70	34.65	36.71	38.89	41.20	43.65	46.25	49.00	51.91	55.00	58.27	61.74
2	65.41	69.30	73.42	77.78	82.41	87.31	92.50	98.00	103.8	110.0	116.5	123.5
3	130.8	138.6	146.8	155.6	164.8	174.6	185.0	196.0	207.7	220.0	233.1	246.9
4	261.6	277.2	293.7	311.1	329.6	349.2	370.0	392.0	415.3	440.0	466.2	493.9
5	523.3	554.4	587.3	622.3	659.3	698.5	740.0	784.0	830.6	880.0	932.3	987.8
6	1047	1109	1175	1245	1319	1397	1480	1568	1661	1760	1865	1976
7	2093	2217	2349	2489	2637	2794	2960	3136	3322	3520	3729	3951
8	4186	4435	4699	4978	5274	5588	5920	6272	6645	7040	7459	7902

Πίνακας 1.1: Οι θεμελιώδεις συχνότητες (Hz) των μουσικών τόνων σε όλες τις οκτάβες

(Σχ. 1.4). Η συχνότητα κάποιων αρμονικών συμπίπτουν με την συχνότητα άλλων τόνων: Για παράδειγμα η $3^{η}$ αρμονική του τόνου A2 (110 Hz) που είναι 330Hz αντιστοιχεί (σχεδόν) στον τόνο E4 (329.6 Hz). Στο φαινόμενο αυτό οφείλεται το γεγονός ότι κάποιες συγχορδίες δίνουν πολύ ευχάριστο ακουστικό αποτέλεσμα (όταν συμπίπτουν οι διαφορετικοί τόνοι με τις αρμονικές) ενώ σε άλλες περιπτώσεις το ακουστικό αποτέλεσμα μιας συγχορδίας μπορεί να είναι εξαιρετικά ενοχλητικό.

- **Κλίμακα** είναι ένα υποσύνολο των τόνων μιας οκτάβας, συνήθως αποτελούμενη από εφτά νότες (δυτική μουσική). Για παράδειγμα η κλίμακα «Ντο Μείζονα» αποτελείται από τις φθόγγους C-D-E-F-G-A-B. Κάθε μουσικό κομμάτι είναι «γραμμένο σε μία κλίμακα», το οποίο σημαίνει ότι μόνο οι εφτά από τους δώδεκα τόνους μέσα σε μία οκτάβα εμφανίζονται σε αυτό το κομμάτι. Αυτό ισχύει για όλες τις οκτάβες. Επομένως σε δύο κομμάτια «γραμμένα στην ίδια κλίμακα» θα συναντήσουμε τους ίδιους εφτά τόνους. Από τους $\binom{12}{7}$ πιθανούς συνδυασμούς δεν τους συναντάμε όλους ως κλίμακες στη μουσική, αλλά ένα αρκετά μικρότερο υποσύνολο που υπαγορεύουν μουσικοί κανόνες, οι οποίοι είναι γνωστοί και ως *αρμονία*. Οι μουσικοί αυτοί κανόνες έχουν στόχο να προσδώσουν ένα ευχάριστο (αρμονικό) ακουστικό αποτέλεσμα των συγχορδιών που προκύπτουν από την κλίμακα.

Συνοψίζοντας τα παραπάνω, καταλήγουμε ότι το φασματικό περιεχόμενο της μουσικής είναι πάρα πολύ συγκεντρωμένο στις συχνότητες που ορίζουν οι μουσικοί τόνοι: στις $12 \times 9 = 108$ θεμελιώδεις συχνότητες που ορίζουν οι 9 οκτάβες και 12 τόνοι ανά οκτάβα. Επομένως δημιουργείται ένα πλέγμα στο χώρο των συχνοτήτων στις 108 αυτές συχνότητες (Σχ. 1.6). Ιδανικά, οποιοδήποτε συχνοτικό περιεχόμενο πέραν αυτών των τιμών θα ακούγεται φάλτσο, με εξαίρεση τα κρουστά όργανα που δεν έχουν θεμελιώδη συχνότητα και η ενέργειά τους εκτείνεται σε όλο το φάσμα με συνεχή τρόπο.

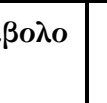


Σχήμα 1.6: Το πλέγμα συχνότητας της μουσικής

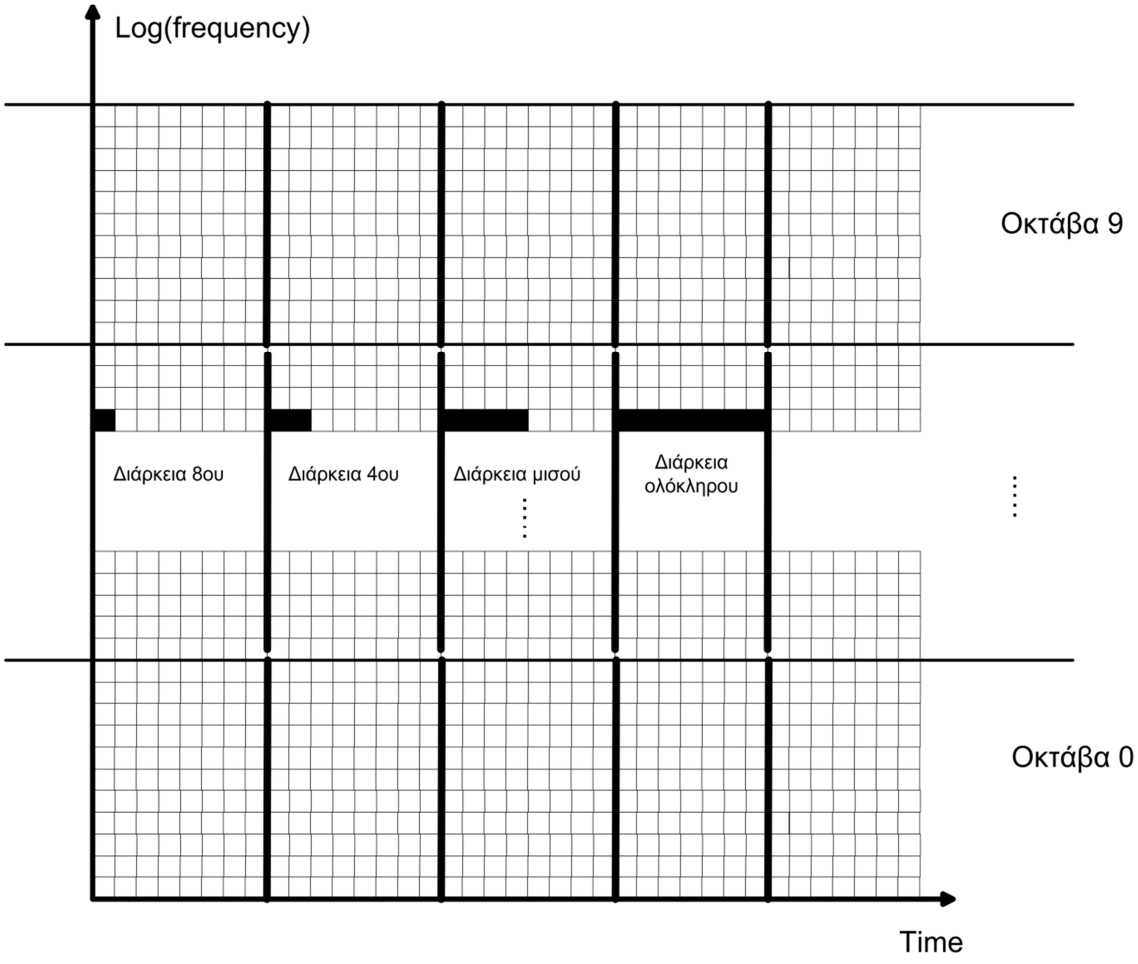
1.2.4 Χρονική οργάνωση της μουσικής και το πεντάγραμμο

Η μουσική συνίσταται από την χρονική οργάνωση νοτών. Κατά τη διάρκεια ενός μουσικού κομματιού υπάρχουν χρονικές στιγμές στις οποίες είναι παρούσες περισσότερες από μία νότες (συγχορδία), ή στιγμές που χαρακτηρίζονται από την απουσία οποιαδήποτε νότας (παύση). Όπως είδαμε στην προηγούμενη παράγραφο το τονικό (φασματικό) περιεχόμενο στη μουσική έχει πάρα πολύ συγκεκριμένη δομή με αποτέλεσμα την δημιουργία ενός «πλέγματος» στον χώρο της συχνότητας, πάνω στο οποίο βρίσκονται οι τόνοι. Κάτι αντίστοιχο συμβαίνει και στην χρονική οργάνωση της μουσικής. Οι νότες ενός κομματιού δεν εμφανίζονται σε τυχαίες χρονικές στιγμές ούτε έχουν τυχαίες διάρκειες. Οι διάρκειες των νοτών είναι υποπολλαπλάσια μιας θεμελιώδους ρυθμικής περιόδου T_r ενώ κάτι αντίστοιχο συμβαίνει και με τις στιγμές έναρξης. Για να παραστήσουμε με συμβολικό τρόπο τη διάρκεια μιας νότας στο πεντάγραμμο, κάθε νότα σε ένα μουσικό κομμάτι συμβολίζεται με μία *χρονική αξία*. Κάθε χρονική αξία έχει συγκεκριμένη *σχετική* χρονική διάρκεια συγκριτικά με το T_r. Στον Πίνακα 1.2 παρουσιάζονται οι κύριες χρονικές αξίες με τις διάρκειές τους. Μπορούμε να πούμε ότι οι χρονικές αξίες είναι το ανάλογο των τόνων στο πεδίο του χρόνου.

Το γεγονός ότι οι διάρκειες και οι χρόνοι είναι κλάσματα μιας θεμελιώδους ρυθμικής περιόδου δημιουργεί και στον χώρο του χρόνου ένα πλέγμα αντίστοιχο με τον χώρο της συχνότητας (Σχ. 1.7).

Αξία	**Ολόκληρο**	**Μισό**	**Τέταρτο**	**Όγδοο**	**Δέκατο Έκτο**	**Τριακοστό Δεύτερο**
Σύμβολο						
Διάρκεια	T_r	$T_r / 2$	$T_r / 4$	$T_r / 8$	$T_r / 16$	$T_r / 32$

Πίνακας 1.2: Οι μουσικές χρονικές αξίες



Σχήμα 1.7: Το πλέγμα χρόνου - συχνότητας της μουσικής: Οι έντονες κάθετες γραμμές δηλώνουν τα όρια του μέτρου (8/8) ενώ οι υπόλοιπες τον βασικό παλμό.

Πλέον, έχοντας ορίσει το ρυθμικό πλέγμα, κάθε μουσικό γεγονός δεν μπορεί παρά να συμβεί πάνω σε αυτό το πλέγμα. Στη πραγματικότητα ωστόσο υπάρχουν χρονικές αποκλίσεις που οφείλονται στην μουσική έκφραση. Στη συνέχεια ακολουθούν ορισμοί βασικών εννοιών που σχετίζονται με το πεντάγραμμο και τη χρονική οργάνωση της μουσικής (Σχήμα 1.8).

Οι τόνοι στο πεντάγραμμο

Το πεντάγραμμο είναι ένας συμβολικός τρόπος να αναπαρασταθούν τα μουσικά γεγονότα πάνω στο μουσικό πλέγμα χρόνου-συχνότητας (Σχ. 1.8α,). Το μουσικό πεντάγραμμο μπορεί να θεωρηθεί ανάλογο της αναπαράστασης χρόνου συχνότητας. Οι οριζόντιες γραμμές ορίζουν το τονικό πλέγμα, όπως οι νότες γράφονται πάνω στις γραμμές και στα διαστήματα μεταξύ των γραμμών του πενταγράμμου με μοναδικό τρόπο, κάθε διάστημα ή γραμμή στο πεντάγραμμο αντιστοιχεί σε έναν μοναδικό τόνο. Αν ένας τόνος γράφεται σε μία γραμμή, ο επόμενος τόνος (1 ημιτόνιο πάνω) γράφεται στο διάστημα πάνω από τη γραμμή αυτή. Επειδή οι τόνοι είναι περισσότεροι από αυτούς που μπορούν να γραφτούν στο πεντάγραμμο συχνά χρησιμοποιούνται βοηθητικές γραμμές πάνω και κάτω από αυτό.

Οι χρονικές αξίες στο πεντάγραμμο

Η χρονική *αξία* που χρησιμοποιείται για να αναπαρασταθεί (ολόκληρο, μισό, τέταρτο κ.λπ.) μία νότα δηλώνει την διάρκειά της. Στο Σχ. 1.8α η πρώτη νότα είναι η E5 με διάρκεια «μισού», επομένως ίση με $T_r/2$. Ακολουθεί η νότα C5 με διάρκεια $T_r/8$ (αξία ογδόου) κ.ο.κ. Οι αντίστοιχες διάρκειες στον άξονα του χρόνου φαίνονται στην κυματομορφή (Σχ. 1.8γ).

Στιγμές έναρξης νότας στο πεντάγραμμο

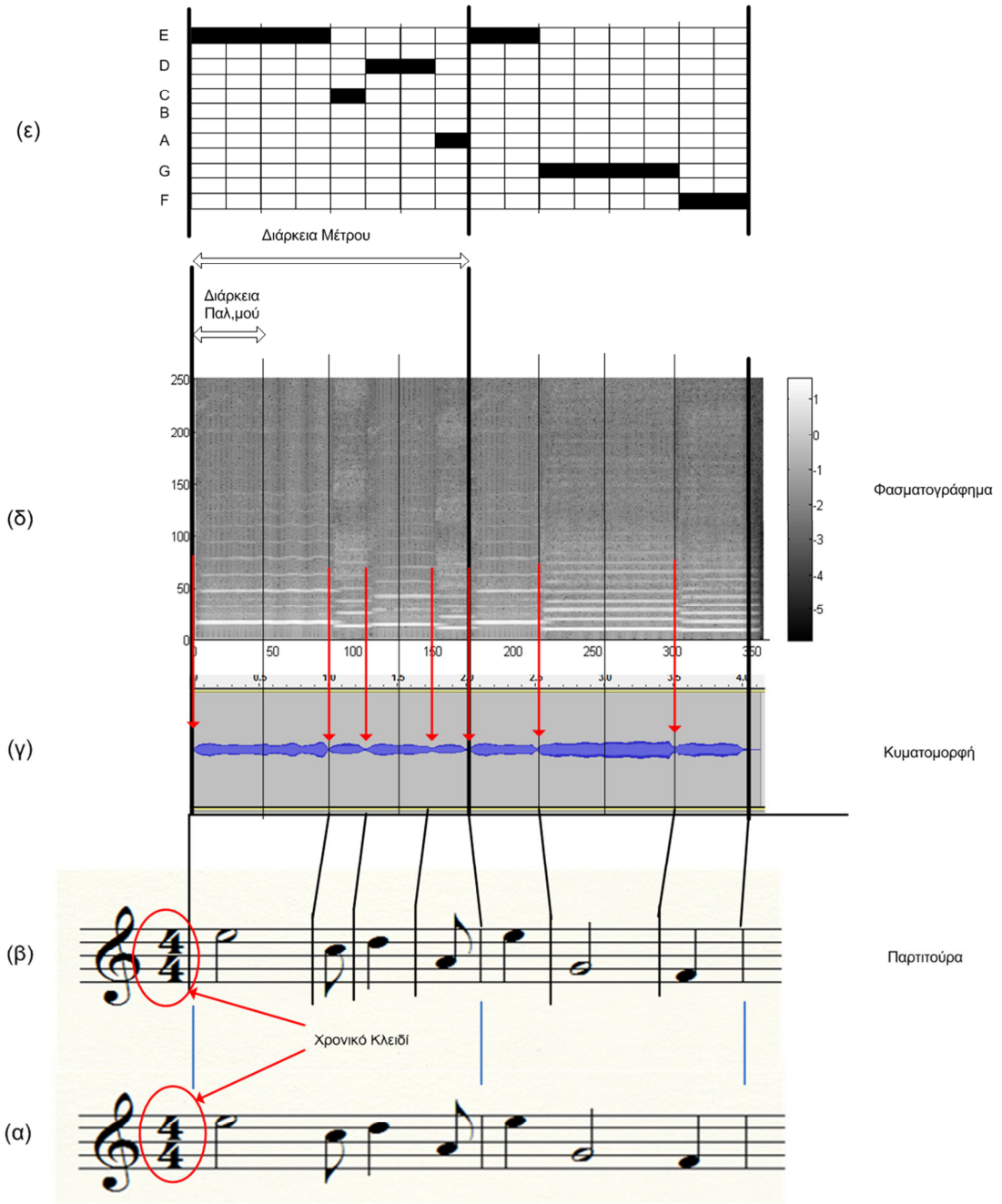
Η στιγμή έναρξης κάθε νότας υποδηλώνεται *έμμεσα* από την οριζόντια θέση της στο πεντάγραμμο σε σχέση και με τις *διάρκειες των προηγούμενων νοτών*. Με αυτόν τον τρόπο ορίζεται η χρονική στιγμή των μουσικών γεγονότων σε ένα μουσικό απόσπασμα. Πρέπει να τονιστεί ότι η οριζόντια θέση μιας νότας στο πεντάγραμμο *δεν υποδηλώνει αυστηρά* την χρονική στιγμή έναρξής της, όπως φαίνεται και στο Σχ. 1.8α-β. Ωστόσο συνηθίζεται στη γραφή στο πεντάγραμμο οι μεγαλύτερης αξίας μουσικοί φθόγγοι να έχουν *μεγαλύτερη οριζόντια έκταση* από αυτές με μικρότερο.

Το μουσικό μέτρο

Ένα μουσικό κομμάτι στο πεντάγραμμο χωρίζεται με περιοδικές κάθετες γραμμές (Σχ. 1.8α) Το χρονικό διάστημα μεταξύ των δύο κάθετων γραμμών ονομάζεται **μέτρο** (bar). Το μέτρο είναι το χρονικό δομικό στοιχείο ενός κομματιού. Η συνολική διάρκεια των νοτών σε ένα μέτρο πρέπει να είναι σταθερή σε κάθε κομμάτι Το μέτρο δεν είναι απλά μια τυχαία σύμβαση προκειμένου να διαιρεθεί η μουσική σε μέρη. Συνήθως αποτελεί την ελάχιστη χρονική μονάδα στην οποία συναντάμε κάποια επαναληπτικότητα στη μουσική. Γι' αυτόν τον λόγο στην αρχή του μέτρου συναντάμε «έντονα» μουσικά γεγονότα όπως αλλαγή συγχορδίας ή ισχυρές νότες (σε ένταση ή σε μελωδική σημασία).

Το χρονικό κλειδί

Το χρονικό κλειδί (Σχ. 1.8α-β) ορίζει τη χρονική διάρκεια του μέτρου. Αποτελεί ένα κλάσμα του οποίου η τιμή ορίζει τη διάρκεια του μέτρου σε σχέση με το T_r. Για παράδειγμα η διάρκεια ενός μέτρου με χρονικό κλειδί 4/4 έχει διάρκεια ίση με $4/4 \cdot T_r = T_r$. Εναλλακτικά μπορεί να ερμηνευτεί ότι έχει διάρκεια ίση με 4 (αριθμητής) αξίες ενός τετάρτου (παρονομαστής), δηλ. $4 \cdot \frac{T_r}{4} = T_r$.



Σχήμα 1.8 : Από κάτω προς τα πάνω: α) Πεντάγραμμο με μέτρο 4/4. Με μπλε γραμμές συμβολίζονται τα όρια των μέτρων. β) Το πεντάγραμμο έχει χωριστεί στα διαστήματα στα οποία ακούγονται οι μεμονωμένες νότες. γ) Η κυματομορφή της εκτέλεσης της μελωδίας που συμβολίζεται στο πεντάγραμμο από φλάουτο. Με έντονες κάθετες γραμμές σημειώνονται τα όρια του μέτρου, ενώ οι υπόλοιπες κάθετες γραμμές υποδεικνύουν τις θέσεις των μουσικών παλμών. Τα κόκκινα βέλη δείχνουν τις χρονικές στιγμές της έναρξης κάθε νότας στη κυματομορφή. Οι γραμμές μεταξύ πενταγράμμου και κυματομορφής δείχνουν την αντιστοίχιση των χρονικών στιγμών έναρξης νότας από το πεντάγραμμο στον πραγματικό χρόνο της εκτέλεσης. δ) Το φασματογράφημα με σήμανση του μέτρου και των παλμών όπως στην κυματομορφή. ε) Η συμβολική αναπαράσταση της μελωδίας στο πλέγμα χρόνου-συχνότητας.

Ο παρονομαστής ορίζει την βασική χρονική αξία του μέτρου (π.χ. τέταρτο για 4/4, όγδοο για 7/8 κ.ο.κ.) ενώ ο αριθμητής ορίζει την διάρκεια του μέτρου βάσει αυτής της βασικής χρονικής αξίας. Ένα κομμάτι με χρονικό κλειδί 3/8 έχει μέτρο με διάρκεια τρεις φορές την διάρκεια ενός ογδόου, ή αλλιώς $3 \cdot T_r/8$ αφού το όγδοο έχει διάρκεια $T_r/8$. Το χρονικό κλειδί σχετίζεται άμεσα με το ρυθμικό αποτέλεσμα ενός μουσικού κομματιού. Για παράδειγμα τα «βαλς» κομμάτια έχουν συνήθως μέτρο 3/4, στη μουσική ροκ συναντάμε πολύ συχνά τα 4/4 και στα «μπλουζ» τα 12/8. Οι ελληνικοί παραδοσιακοί χοροί χαρακτηρίζονται έντονα από το μέτρο (π.χ. καλαματιανός 7/8, πεντοζάλης 2/4).

Ο μουσικός παλμός

Κάθε μουσικό μέτρο (bar) χωρίζεται σε ισόχρονα διαστήματα ίσα με την διάρκεια της βασικής χρονικής αξίας (παρονομαστής χρονικού κλειδιού). Το πλήθος των διαστημάτων αυτών είναι ίσος με τον αριθμητή του χρονικού κλειδιού. Για παράδειγμα, σε ένα κομμάτι με χρονικό κλειδί 3/8, το μέτρο χωρίζεται σε τρία ισόχρονα διαστήματα διάρκειας ενός ογδόου. Η αρχή των διαστημάτων αυτών ορίζουν τον *μουσικό παλμό*. Ο μουσικός παλμός (beat) ταυτίζεται με τις χρονικές στιγμές που θα αντιστοιχούσαν στην έναρξη των νοτών αν το μέτρο απαρτιζόταν από διαδοχικές νότες με διάρκεια ίση με τη βασική χρονική αξία. Εναλλακτικά μπορούμε να πούμε ότι ο μουσικός παλμός είναι οι χρονικές στιγμές τις οποίες θα χτύπαγε ένας ιδανικός μετρονόμος κατά την εκτέλεση ενός κομματιού.

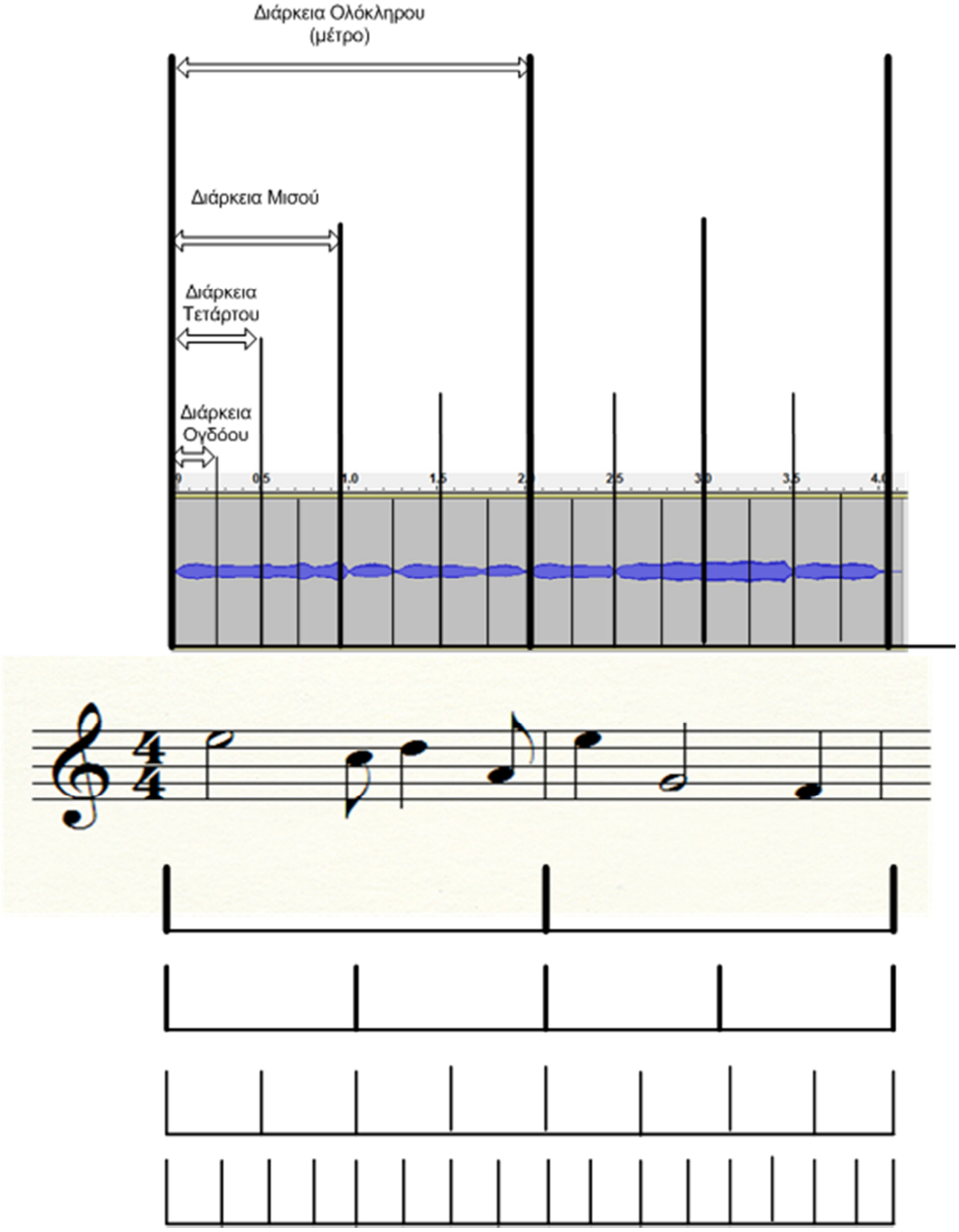
Το μουσικό τέμπο

Το *μουσικό τέμπο* είναι ένας πραγματικός θετικός αριθμός (στην πράξη όμως συναντιέται κυρίως ως φυσικός) ο οποίος αντιστοιχεί στη συχνότητα του παλμού. Το τέμπο μετριέται σε κτύπους ανά λεπτό (beats per minute - BPM) και μας δείχνει πόσο *γρήγορα* παίζεται ένα κομμάτι.

Το Σχήμα 1.8 συνοψίζει όλες οι παραπάνω έννοιες, Στο Σχήμα 1.8(α) φαίνεται η μελωδία στο πεντάγραμμο και τα όρια των μέτρων έχουν επισημανθεί με μπλε γραμμές. Στο Σχήμα 1.8(β) το πεντάγραμμο έχει χωριστεί στις μεμονωμένες νότες με κάθετες γραμμές. Το διάστημα μεταξύ δύο γραμμών παριστάνει το χρονικό διάστημα που ακούγεται η νότα όπως αυτό *αναπαρίσταται στο πεντάγραμμο*. Στο Σχήμα 1.8(γ) παρουσιάζεται η κυματομορφή της εκτέλεση της μελωδίας από φλάουτο. Με παχιές κάθετες γραμμές φαίνονται τα όρια του μέτρου, ενώ με ελαφριές κάθετες γραμμές οι θέσεις του μουσικού παλμού. Τα κόκκινα βέλη δείχνουν τις χρονικές στιγμές της έναρξης κάθε νότας στην κυματομορφή. Οι γραμμές μεταξύ πενταγράμμου και κυματομορφής δείχνουν την στοίχιση των χρονικών στιγμών έναρξης νότας από το πεντάγραμμο στον πραγματικό χρόνο. Παρατηρούμε ότι δεν υπάρχει απευθείας αντιστοίχιση του χρονικού άξονα στο πεντάγραμμο με τον πραγματικό χρονικό άξονα του σήματος. Στο Σχήμα 1.8(δ) παρουσιάζεται το αντίστοιχο φασματογράφημα. Το χρονικό διάστημα μεταξύ των ορίων του μέτρου είναι σταθερή, και ίση με τέσσερις φορές τη διάρκεια του παλμού. Τέλος, στο Σχήμα 1.8(ε) βλέπουμε τις θέσεις των νοτών της συγκεκριμένης μελωδίας στο πλέγμα χρόνου συχνότητας.

Μετρικά επίπεδα

Οι χρονικές στιγμές του παλμού όπως αυτοί παρουσιάζονται στα Σχ. 1.8(γ), (δ) ορίζουν ένα χρονικό πλέγμα σαν αυτό που παρουσιάστηκε στο Σχ. 1.7. Ωστόσο και το ίδιο το μέτρο ορίζει ένα άλλο πλέγμα στον χρόνο, όπως φαίνεται στα Σχ. 1.7 και Σχ.1.8(ε).



Σχήμα 1.9 : Χρονικό κλειδί 4/4. Κάθε μέτρο (μπάρα) έχει διάρκεια 4 τέταρτα. Παρουσιάζονται 3 μετρικά επίπεδα (από κάτω προς τα πάνω): Επίπεδο ογδόου, επίπεδο τετάρτου (βασικό μετρικό επίπεδο), επίπεδο μισού, και ολοκλήρου (επίπεδο μέτρου)

Ο βασικός παλμός και το μέτρο σε αυτή τη περίπτωση αποτελούν διαφορετικά *μετρικά επίπεδα*. Τα μετρικά επίπεδα είναι διαφορετικές χρονικές κλίμακες οργάνωσης της μουσικής, που έχουν ιεραρχική σχέση μεταξύ τους. Το μετρικό επίπεδο που έχει διάρκεια ίση με το μέτρο ονομάζεται *μετρικό επίπεδο μέτρου* ενώ το μετρικό επίπεδο του παλμού ονομάζεται *βασικό μετρικό επίπεδο*. Κάθε

μετρική δομή ορίζει διάφορα μετρικά επίπεδα που είναι οργανωμένα ιεραρχικά. Κάθε μετρικό επίπεδο ορίζει έναν παλμό με περίοδο ίση με την διάρκεια της αντίστοιχης χρονικής αξίας (π.χ. μετρικό επίπεδο τέταρτου, μισού κ.ο.κ.). Στο Σχήμα 1.9 φαίνονται τέσσερα μετρικά επίπεδα για το μουσικό σήμα. του Σχ. 1.8. Το επίπεδο ολοκλήρου (που ταυτίζεται με το επίπεδο μέτρου επειδή το χρονικό κλειδί είναι 4/4, δηλ. το μέτρο έχει διάρκεια ίση με ένα «ολόκληρο»), το επίπεδο του μισού, το επίπεδο του τετάρτου (που ταυτίζεται με το βασικό μετρικό επίπεδο αφού ο παρονομαστής του κλειδιού είναι 4) και το επίπεδο του ογδόου. Τα μετρικά επίπεδα είναι έννοια ύψιστης σημασίας για την ρυθμική ανάλυση μουσικών σημάτων, αφού ο ρυθμός ενός κομματιού αποτελεί σύνθεση επιμέρους ρυθμών στα διάφορα μετρικά επίπεδα. Η απόσταση/διάρκεια μεταξύ δύο διαδοχικών παλμών στο *βασικό* μετρικό επίπεδο ορίζει την περίοδο που αντιστοιχεί στο μουσικό τέμπο.

Ενώ υπάρχει ισχυρή συσχέτιση μεταξύ του παλμού και του τέμπο δεν υπάρχει απόλυτη συμφωνία για το αν το μουσικό τέμπο επάγεται από το μουσικό παλμό ή το αντίστροφο. Δηλαδή το αν ο ανθρώπινος εγκέφαλος πρωτίστως αναγνωρίζει τον μουσικό παλμό και από αυτόν δημιουργείται η πιο «γενική» αίσθηση του τέμπο ή το αντίστροφο, δηλ. αντιλαμβάνεται πρώτα μια γενική αίσθηση του ρυθμού, από τον οποίο σε δεύτερη φάση εξάγει και τις θέσεις των παλμών. Αν και αυτή η συζήτηση ξεπερνά το επιστημονικό πεδίο της παρούσας διατριβής, αναφέρεται καθώς αντανακλάται στον ίδιο τον σχεδιασμό αλγορίθμων αυτόματης εξαγωγής τέμπο και παλμού (Κεφ. 5, 6).

1.2.5 Ρυθμικές αντικανονικότητες

Με τον όρο ρυθμική αντικανονικότητα εννοούμε οποιαδήποτε αλλοίωση, προσωρινή ή μόνιμη των ρυθμικών ιδιοτήτων που παρουσιάστηκαν προηγουμένως, όπως για παράδειγμα μια μεταβολή του μουσικού τέμπο σε κάποια χρονική στιγμή. Για τις τρεις χρονικές ιδιότητες της μουσικής (τέμπο, παλμός, μέτρο) έχουμε και τρεις κατηγορίες ανωμαλιών. Στο Σχ. 1.10 παρουσιάζονται οι συχνότερες αντικανονικότητες που συναντώνται στη μουσική.

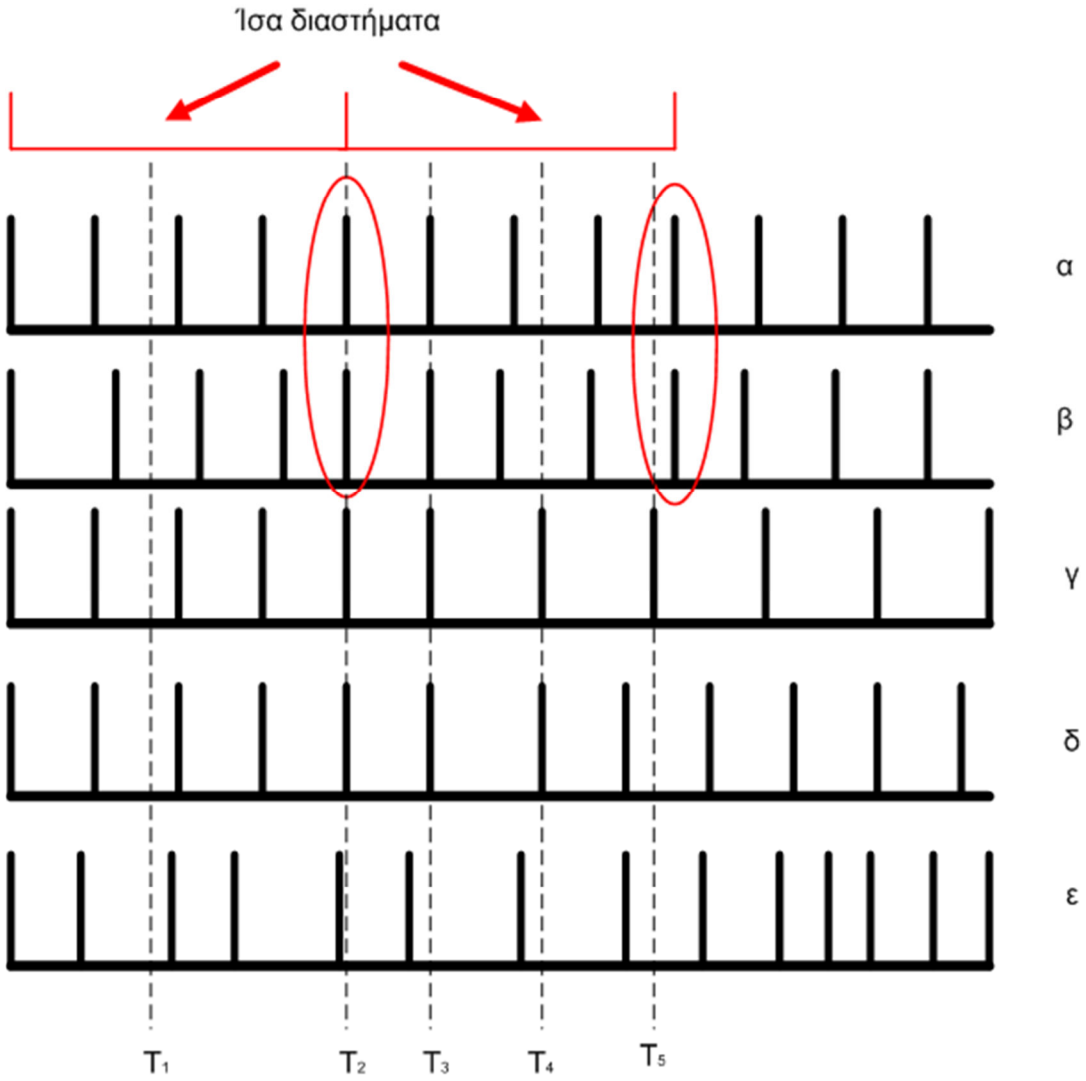
Αντικανονικότητες μουσικού τέμπο

Το μουσικό τέμπο μπορεί να υποστεί τις παρακάτω αλλοιώσεις κατά τη διάρκεια ενός κομματιού:

- Απότομη αλλαγή από ένα σταθερό τέμπο σε ένα άλλο σταθερό τέμπο (Σχήμα 1.10γ)
- Σταδιακή αλλαγή του τέμπο από μία τιμή σε μία άλλη
- Συνεχώς μεταβαλλόμενο τέμπο (Σχήμα 1.10ε)

Αντικανονικότητες μουσικού παλμού

Οι αντικανονικότητες του μουσικού παλμού αν και έχουν στενή σχέση με τις αλλοιώσεις του τέμπο, δηλαδή αλλοιώσεις του τέμπο προκαλούν αλλοιώσεις και του μουσικού παλμού, οι αλλοιώσεις του μουσικού παλμού μπορούν να οριστούν και ανεξάρτητα από το τέμπο. Στην περίπτωση σταθερού τέμπο οι αντικανονικότητες του μουσικού παλμού μπορούν να συνοψιστούν ως εξής:



Σχήμα 1.10: Χρονική ακολουθία παλμών για το χρονικό κλειδί 4/4. α) Κανονική ακολουθία: όλες οι θέσεις είναι ισαπέχουσες. β) Σταθερό τέμπο, αλλά με αποκλίσεις μέσα στο μέτρο. γ) Απότομη αλλαγή τέμπο την χρονική στιγμή Τ3. δ) Μεταπήδηση φάσης την χρονική στιγμή Τ4 ε) Ακολουθία παλμών με συνεχώς μεταβαλλόμενο τέμπο.

- Μετακίνηση (πήδημα) φάσης: Χωρίς αλλαγή του τέμπο, εμφανίζεται ένα άλμα στη φάση, συνήθως κατά ½ της περιόδου (βλ. Σχήμα 1.10δ). Η μετακίνηση φάσης προκαλείται συνήθως από προσωρινή αλλαγή μέτρου.
- Μεταβολή φάσης που οφείλεται κυρίως σε μουσική έκφραση όπου η διάρκεια κάποιων μουσικών παλμών αποκλίνει από την θεμελιώδη περίοδο. Η συνολική τους όμως διάρκεια παραμένει σταθερή και σύμφωνα με το τέμπο. (βλ. Σχήμα 1.10β).

Αντικανονικότητες μουσικού μέτρου

Οι αντικανονικότητες του μουσικού μέτρου περιλαμβάνουν την προσωρινή ή μόνιμη αλλαγή του μουσικού μέτρου. Αν και τέτοιου είδους αντικανονικότητες

μπορεί να μην επηρεάζουν άμεσα το τέμπο ή τον μουσικό παλμό στο βασικό μετρικό επίπεδο, μπορεί να επηρεάζουν τον παλμό στα άλλα (υψηλότερα) μετρικά επίπεδα. Για παράδειγμα στην μετάβαση από το μέτρο 4/4 στο 3/4, τα μετρικά επίπεδα μισού και ολόκληρου στα 4/4 παύουν να είναι ρυθμικά σχετικά με το μέτρο 3/4. Επομένως ένα σύστημα εξαγωγής παλμού που έχει «κλειδώσει» στο μετρικό επίπεδο του μισού, θα έχει πρόβλημα να ακολουθήσει την μεταβολή του μέτρου. Κάτι τέτοιο όμως δεν θα συνέβαινε στην περίπτωση που θα ακολουθούσε το βασικό μετρικό επίπεδο.

Παρατηρώντας το Σχ. 1.10 μπορούμε να εξάγουμε και συμπεράσματα για τους περιορισμούς που έχει ένα αιτιατό (casual) σύστημα εξαγωγής τέμπο και παλμού. Στην περίπτωση (β) του Σχ. 1.10 την χρονική στιγμή T1 παρακολουθώντας τις προηγούμενες θέσεις των παλμών δεν μπορούμε να γνωρίζουμε αν οι ακολουθίες (α) και (β) επιτυγχάνουν το ίδιο τέμπο. Αυτό μπορεί να γίνεται μόνο την χρονική στιγμή T2. Στις περιπτώσεις (γ) και (δ) αντίστοιχα, την χρονική στιγμή T4 δεν μπορούμε να γνωρίζουμε αν η αλλαγή της περιόδου οφείλεται σε αλλαγή του τέμπο ή σε μετακίνηση της φάσης. Αυτό μπορεί να γίνει (με αβεβαιότητα) μετά τη χρονική στιγμή T5. Ένας εμπειρικός κανόνας για την χρονική διάρκεια της παρατήρησης που απαιτείται για να μπορεί κάποιος με βεβαιότητα να αντιληφθεί μια αλλαγή στις ρυθμικές ιδιότητες ενός μουσικού κομματιού είναι περίπου ίση με την διάρκεια ενός μέτρου, αφού μόνο τότε μπορεί να παρατηρηθεί ολόκληρη η ρυθμική δομή.

1.3 Βιβλιογραφική Επισκόπηση

1.3.1 Εισαγωγή

Ο ρυθμός είναι μία θεμελιώδης ιδιότητα της μουσικής μαζί με την μελωδία, την αρμονία και την ενορχήστρωση, και αντιστοιχεί στην χρονική οργάνωση της μουσικής. Μια δεσπόζουσα έννοια της χρονικής οργάνωσης της μουσικής είναι η μετρική δομή, η οποία προκύπτει από τα διάφορα μετρικά επίπεδα [Cooper1963]. Κάθε μετρικό επίπεδο αντιστοιχεί σε έναν παλμό σε διαφορετικές χρονικές κλίμακες. Τα μετρικά επίπεδα όπως είδαμε στην προηγούμενη ενότητα είναι οργανωμένα ιεραρχικά, και μια σειρά από κανόνες που καθορίζουν τις σχέσεις των μετρικών επιπέδων έχουν οριστεί στην «Παραγωγική Θεωρία της Τονικής Μουσικής» (Generative Theory of Tonal Music, [Lerdalh1985]. Το προτιμώμενο μετρικό επίπεδο ενός κομματιού, δηλαδή το μετρικό επίπεδο για το οποίο ένας ακροατής θα συγχρόνιζε ένα βηματισμό ή τον κτύπο των ποδιών του, το οποίο συναντάται με την ονομασία *tactus*, είναι υποκειμενικό και επηρεάζεται από διάφορους παράγοντες, όπως η μουσική εκπαίδευση, η ηλικία, και το περιβάλλον ακρόασης ([Drake,2000a],[Drake,2000b], [Lapidaki2000]). Μαζί με το *tactus*, τα πιο σημαντικά μετρικά επίπεδα είναι το tatum, το οποίο είναι το γρηγορότερο μετρικό επίπεδο σε ένα κομμάτι καθώς και το μετρικό επίπεδο του μέτρου (bar) [Parncutt1994]. Δεδομένου ενός μετρικού επιπέδου, ορίζεται ένα τέμπο για αυτό το επίπεδο. Όπως αναφέρθηκε στην προηγούμενη ενότητα, το τέμπο ενός κομματιού είναι η συχνότητα του παλμού για το μετρικό επίπεδο που αντιστοιχεί στον παρονομαστή του χρονικού κλειδιού. Αυτός ο ορισμός του τέμπο αντιστοιχεί στο *επισημειωμένο* (notated) τέμπο και πρέπει να διαχωριστεί από την έννοια του *αντιληπτικού* (perceptual) τέμπο, το οποίο αντιστοιχεί στο τέμπο του tactus. Προφανώς όταν το μετρικό επίπεδο του tactus συμπίπτει με το μετρικό επίπεδο του παρονομαστή του μουσικού κλειδιού, τα δύο αυτά τέμπο είναι ίδια.

Οι παραπάνω ιδιότητες του της μετρικής δομής, υπαγορεύουν και τα πιο σημαντικά προβλήματα αυτόματης ανάλυσης μουσικής, τα οποία είναι η αυτόματη εύρεση του τέμπο, των θέσεων των παλμών και του μουσικού κλειδιού. Η αυτόματη εύρεση της αρχής του κάθε μέτρου (downbeat tracking), ή εναλλακτικά η εύρεση του πιο ισχυρού παλμού σε κάθε μέτρο, παρότι μπορεί να θεωρηθεί ένας συνδυασμός εύρεσης παλμού και χρονικού κλειδιού, συνήθως συναντάται ως ένα αυτόνομο πρόβλημα. Ένα πιο πρόσφατο πρόβλημα αυτόματης ανάκτησης ρυθμικής πληροφορίας, είναι η εύρεση της «αντιληπτικής ταχύτητας» [Elowsson2013], (*perceptual speed*) ενός κομματιού, ή η κατηγοριοποίηση σε προκαθορισμένες κλάσεις ταχύτητας, όπως για παράδειγμα «αργό», «μέτριο» και «γρήγορο» [Peeters2012]. Παρόλο που η αντιληπτική ταχύτητα δεν συνδέεται άρρηκτα με το τέμπο [Madison2010], μπορεί να χρησιμοποιηθεί ως επιπρόσθετη πληροφορία για να επιλεχθεί το σωστό μετρικό επίπεδο ενός συστήματος εξαγωγής τέμπο ή παλμού, μειώνοντας έτσι τα περίφημα «λάθη οκτάβας». Τα λάθη οκτάβας προκύπτουν όταν ένα σύστημα θεωρήσει ως σωστό τέμπο ένα πολλαπλάσιο ή υποπολλαπλάσιο του σωστού τέμπο. Πέρα από την ρυθμική ανάλυση όπως περιγράφηκε ως τώρα, υπάρχουν πληθώρα προβλήματα που σχετίζονται με αυτήν, όπως για παράδειγμα η κατηγοριοποίηση χορού [Gouyon2004] και η εύρεση ρυθμικής ομοιότητας [Antonopoulos2007].

1.3.2 Αυτόματη Εξαγωγή Τέμπο και Μουσικού Παλμού

Οι πρώτες μέθοδοι αυτόματης ανάλυσης ρυθμού ήταν σχεδιασμένες για να επεξεργάζονται συμβολικές αναπαραστάσεις μουσικής, όπως για παράδειγμα λίστες onsets ή αρχεία MIDI ([Parncutt1994],[Rosenthal1992],[Desain1992]). Μερικά χρόνια αργότερα προτάθηκαν μέθοδοι αυτόματης ανάλυσης ρυθμού που χειρίζονταν σήματα ήχου, και ήταν κυρίως στοχευμένες στην εύρεση του τέμπο και των θέσεων των παλμών, όπως για παράδειγμα στις ([Scheirer1998],[Goto1998]). Οι περισσότερες από αυτές τις μεθόδους δεν χρησιμοποίησαν τεχνικές εκμάθησης των παραμέτρων τους από δεδομένα, αλλά αποτελούνταν κυρίως από μια ακολουθία προκαθορισμένων υπολογιστικών βημάτων.

Πιο σύγχρονες μέθοδοι εξαγωγής τέμπο και μουσικού παλμού χρησιμοποίησαν έναν πιθανοτικό φορμαλισμό. Στο [Klapuri2006] προτάθηκε ένα χρονικά αναλλοίωτο Μπεϋζιανό Δίκτυο (Time-Invariant Bayesian Network) για την μοντελοποίηση των τριών βασικών μετρικών επιπέδων: των tactus, tatum και του μέτρου. Οι Davies και Plumbley [Davies2007] πρότειναν μια μέθοδο εξαγωγής παλμού που υιοθετεί ένα Μπεϋζιανό Δίκτυο δύο καταστάσεων για να χειριστεί ασυνέχειες του παλμού που οφείλονται σε αλλαγές των μετρικών επιπέδων. Παρόμοια, οι Peeters και Papadopoulos [Peeters2011b, Papadopoulos2011] υιοθέτησαν ένα Μπεϋζιανό Δίκτυο όπου ο παλμός αναπαρίσταται ως μια κρυφή μεταβλητή (hidden state) προκειμένου να μοντελοποιήσουν τον παλμό στο βασικό μετρικό επίπεδο και τον παλμό στο επίπεδο του μέτρου (beats, downbeats αντίστοιχα). Παρότι όλες οι προαναφερθέντες μέθοδοι έκαναν χρήση πιθανοτικού φορμαλισμού, οι τιμές των παραμέτρων αυτών των μοντέλων ήταν καθορισμένες με χειρονακτικό τρόπο.

Η πρώτη προσπάθεια να χρησιμοποιηθεί ένα σχήμα μηχανικής μάθησης για την εξαγωγή του τέμπο προτάθηκε στο [Seyerlehner2007]. Όρισαν το πρόβλημα της αυτόματης εξαγωγής τέμπο ως πρόβλημα κατηγοριοποίησης: πρότειναν έναν ταξινομητή κοντινότερου γείτονα (k-Nearest Neighbor Classifier: kNN) πάνω σε δύο αναπαραστάσεις περιοδικότητας: (α) την συνάρτηση αυτοσυσχέτισης (Autocorrelation Function, ACF) και (β) στα Πρότυπα Διακύμανσης (Fluctuation

Patterns). Για κάθε κομμάτι, το τέμπο αποφασίστηκε ως το τέμπο του κοντινότερου γείτονα βάσει αυτών των δύο αναπαραστάσεων. Σε μια παρόμοια προσέγγιση, οι Eronen και Klapuri [Eronen2010] χρησιμοποίησαν χαρακτηριστικά αναπαράστασης χρώματος μαζί με την ανακλιμάκωση του διανύσματος περιοδικότητας. Στο ίδιο άρθρο, πρότειναν επίσης ένα υποσύστημα κατηγοριοποίησης μουσικής ταχύτητας (αργό, μέτριο και γρήγορο) βάσει του εξαγόμενου τέμπο. Μια αξιοσημείωτη μέθοδος για την εξαγωγή παλμού είναι η εργασία των Bock και Schedl [Bock2011]. Χρησιμοποίησαν ένα Διπλής Κατεύθυνσης Μακράς Βραχείας Μνήμης Νευρωνικό Δίκτυο (Bidirectional Long-Short Term Memory Neural Network - BLSTM) για την εξαγωγή παλμού. Η προτεινόμενη μέθοδος παρουσίασε πολύ υψηλά ποσοστά ακρίβειας, τα υψηλότερα στην βιβλιογραφία διεθνώς μέχρι και σήμερα.

Όπως προαναφέρθηκε, οι περισσότερες μέθοδοι εξαγωγής τέμπο και παλμού πάσχουν από τα «λάθη οκτάβας» (octave errors), δηλαδή τείνουν να υπολογίζουν τέμπο που είναι κάποιο πολλαπλάσιο ή κλάσμα του αληθούς τέμπο. Για παράδειγμα, για κάποιο μουσικό κομμάτι με τέμπο 120 BPM, κάποιος αλγόριθμος μπορεί να προτείνει το τέμπο 60 BPM ως αληθές. Αυτό δεν μπορεί να θεωρηθεί πάντα λάθος καθώς κάποιοι ακροατές ακούγοντας το κομμάτι μπορεί να χτυπούσαν αυθόρμητα το πόδι τους με συχνότητα 60 BPM. Σε αυτή την περίπτωση, το αποτέλεσμα ενός αλγορίθμου που υπολογίζει 60 BPM ως αληθές τέμπο, δεν μπορεί να αξιολογηθεί ως λάθος. Επομένως, κάποια υποπολλαπλάσια ή πολλαπλάσια του αληθούς τέμπο μπορούν να θεωρηθούν και αυτά ως αληθή τέμπο. Αυτό το παράδειγμα αναδεικνύει και την διαφορά μεταξύ του *αντιληπτικού* με του *επισημειωμένου* τέμπο. Για την πιο δίκαιη αξιολόγηση των μεθόδων αυτόματης εξαγωγής τέμπο, έχουν προταθεί μετρικές αξιολόγησης που λαμβάνουν τα λάθη οκτάβας ως σωστά, όπως για παράδειγμα η μετρική *accuracy2*, η μετρική P-Score που χρησιμοποιείται στον διαγωνισμό MIREX[1], ή οι μετρικές AMLt, AMLc [Davies2009] που χρησιμοποιούνται για την εξαγωγή παλμού. Παρόλα αυτά όμως δεν μπορούν να θεωρηθούν όλα τα πολλαπλάσια και τα υποπολλαπλάσια του αληθούς τέμπο σωστά. Επομένως τα λάθη οκτάβας αποτελούν μία από τις κύριες αδυναμίες των μεθόδων αυτόματης ανάλυσης ρυθμού. Στη βιβλιογραφία υπάρχουν διάφορες ευριστικές μέθοδοι για την αντιμετώπιση αυτού του προβλήματος, όπως για παράδειγμα στα [Smith2010], [Tzanetakis2013], [Dixon2001]. Άλλες μέθοδοι χρησιμοποιούν εκ των προτέρων κατανομές του τέμπο [Klapuri2006], ή την υιοθέτηση μοντέλων μετρικών σχέσεων [Peeters2005]. Όπως προαναφέρθηκε, ένας τρόπος μείωσης των λαθών οκτάβας είναι η ενσωμάτωση της γνώσης της *μουσικής ταχύτητας* ενός κομματιού. Οι Hockman και Fujinaja [Hockman2010] πρότειναν ένα σύστημα που κατηγοριοποιεί την μουσική ταχύτητα ανεξάρτητα από το πρόβλημα αυτόματης ανάλυσης ρυθμού. Χρησιμοποιώντας μονάχα φασματικά χαρακτηριστικά και τον AdaBoost αλγόριθμο, επετεύχθη ποσοστό 96% σωστής αναγνώρισης ανάμεσα στις δύο κλάσεις «αργό» και «γρήγορο». Τα δεδομένα που χρησιμοποίησαν ήταν τραγούδια από το YouTube και οι δύο κατηγορίες είχαν εξαχθεί αυτόματα βάσει των ετικετών (tags) από τους χρήστες. Στο [Peeters2012] προτάθηκε μία μέθοδος εξαγωγής μουσικής ταχύτητας και αντιληπτικού τέμπο χρησιμοποιώντας παλινδρόμηση (regression) πάνω σε Μίξη Γκαουσσιανών Μοντέλων (Gaussian Mixture Models). Οι Elowsson και Friberg [Elowsson2013] χρησιμοποίησαν ένα μοντέλο της αντίληψης μουσικής ταχύτητας βασιζόμενο σε χαρακτηριστικά των onsets για τον υπολογισμό του μουσικού τέμπο.

[1] http://www.music-ir.org/mirex/wiki/MIREX_HOME

1.3.3 Αυτόματη Αναγνώριση Μέτρου και Εξαγωγή Μουσικού Κλειδιού

Το χρονικό κλειδί (Time Signature) αλλά και το μέτρο (δηλαδή η μετρική δομή γενικότερα) αποτελούν ένα από τα κύρια ρυθμικά χαρακτηριστικά της μουσικής που προσφέρουν πολύτιμη πληροφορία για άλλα ρυθμικά προβλήματα όπως για παράδειγμα η ρυθμική ομοιότητα μεταξύ μουσικών κομματιών, η αποσαφήνιση των μετρικών επιπέδων και γενικότερα η ρυθμική αντίληψη. Στο [Klapuri2006] οι συγγραφείς πρότειναν ένα πλήρες πιθανοτικό πλαίσιο για την μοντελοποίηση της ρυθμικής δομής. Παρόμοιες μεθόδους συναντούμε για μη Δυτική μουσική, όπως για παράδειγμα στο [Holzapfel2014], όπου προτάθηκε ένα Μπεϋζιανό μοντέλο που δοκιμάστηκε σε μουσικές συλλογές προερχόμενες από διάφορες χώρες.

Στην διεθνή βιβλιογραφία ο όρος *υπολογισμού του μέτρου* (meter estimation) πολλές φορές περιορίζεται στην *κατηγοριοποίηση μουσικού κλειδιού* (time-signature classification) ή στην *εξαγωγή της αρχής του μέτρου* (downbeat tracking). Οι Toiviainen και Erola [Toiviainen2006] παρουσίασαν μια μέθοδο κατηγοριοποίησης μουσικού κλειδιού στηριζόμενοι σε μελωδικά χαρακτηριστικά που είχαν εξαχθεί από αρχεία MIDI χρησιμοποιώντας έναν Multiple Discriminant Analysis ταξινομητή. Στο [Robine2005] προτάθηκε το Προφίλ Μετρικών Σχέσεων (Meter Class Profile) για την εξαγωγή του μουσικού κλειδιού από κυματομορφές που είχαν συνθεθεί από MIDI αρχεία. Στο [Frierer2004] προτάθηκε η χρήση Γκαουσσιανών στη θέση των onsets για την εξαγωγή του παλμού και του μουσικού κλειδιού, ενώ στο [Krebs2013] συναντάμε πάλι Μπεϋζιανά Δίκτυα για την μοντελοποίηση του παλμού και της αρχής του κάθε μέτρου (beats,downbeats).

Κάποιες μέθοδοι εύρεσης τέμπο η παλμού, ενσωματώνουν ένα υποσύστημα ταξινόμησης μουσικού κλειδιού. Στο [Eck2005] για παράδειγμα, προτείνεται ένα απλό μοντέλο για τον υπολογισμό της πιθανοφάνειας (likelihood) διπλού (duple) ή τριπλού (triple) μουσικού κλειδιού, προκειμένου να επιλεχθεί το σωστό τέμπο από την συνάρτηση περιοδικότητας. Παρόμοια, ο Peeters [Peeters2007] ορίζει μία *κατάσταση τέμπο* για να ορίσει διπλά, τριπλά ή σύνθετα (compound) μουσικά μέτρα στα πλαίσια μιας μεθόδου εξαγωγής τέμπο. Εφεξής, στην παρούσα διατριβή με τον όρο *εύρεση μέτρου* θα εννοούμε την κατηγοριοποίηση ή εύρεση του μουσικού κλειδιού.

1.3.4 Ρυθμική Κατηγοριοποίηση

Παρότι το μέτρο μαζί με το τέμπο αποτελούν πολύτιμη ρυθμική πληροφορία και μπορεί να είναι αρκετά περιγραφικές μιας ρυθμικής κλάσης, ο όρος ρυθμική κατηγοριοποίηση δεν είναι πάντοτε σαφής στη διεθνή βιβλιογραφία. Κάποιες μέθοδοι ορίζουν την ρυθμική κατηγοριοποίηση πολύ κοντά στην έννοια της εξαγωγής μέτρου, όπως για παράδειγμα στα [Antonopoulos2007] και [Pikrakis2004], όπου κάθε ρυθμική κατηγορία αντιστοιχούσε σε συγκεκριμένο χρονικό κλειδί και για ένα στενό εύρος τέμπο. Αρκετές μέθοδοι χρησιμοποιούν τον όρο ρυθμική κατηγοριοποίηση για να περιγράψουν την ταξινόμηση σε είδη χορευτικών ρυθμών (dance music classification). Οι πιο πολλές από αυτές έχουν αξιολογηθεί στην συλλογή Ballroom που εμφανίστηκε σε διαγωνισμό στο πλαίσιο του συνεδρίου ISMIR 2004, και που αποτελείται από 8 χορευτικές κλασεις δυτικής μουσικής. Οι Gouyon και Dixon [Gouyon2004b] χρησιμοποίησαν μίξεις Γκαουσιανών σε διάφορα είδη χαρακτηριστικών, όπως στο ιστόγραμμα της περιοδικότητας, το τέμπο, τα διαστήματα μεταξύ των στιγμών έναρξης νοτών (Inter Onset Intervals, IOI) και σε φασματικά χαρακτηριστικά. Οι ίδιοι

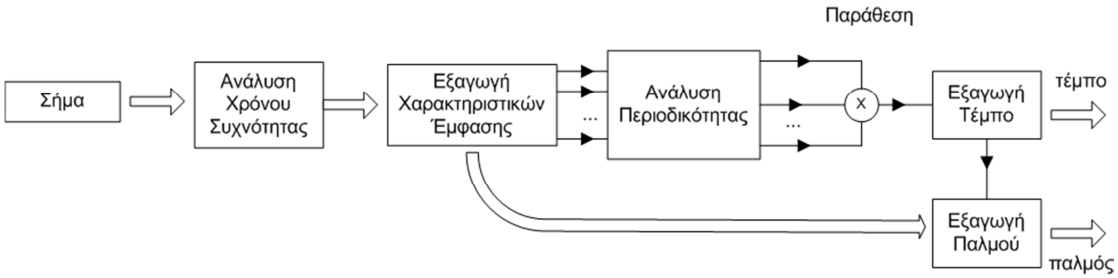
συγγραφείς [Dixon2004] ανέφεραν 96% αναγνώριση για τις 8 κατηγορίες ρυθμών στη συλλογή Ballroom υιοθετώντας τον αλγόριθμο AdaBoost πάνω σε ταξινομητές εφαρμοσμένους σε πληθώρα φασματικών και χρονικών χαρακτηριστικών. Στο [Gouyon2004c] τρία είδη χαρακτηριστικών (χαρακτηριστικά IOI, τέμπο και η συνάρτηση περιοδικότητας) χρησιμοποιήθηκαν για τη δημιουργία ενός ταξινομητή ρυθμού ενώ ο Peeters [Peeters2005] υιοθέτησε εξ ολοκλήρου φασματικά χαρακτηριστικά περιοδικότητας. Πέρα από προαναφερθέντες μεθόδους που όλες εφαρμόστηκαν στην συλλογή Ballroom, υπάρχουν και άλλες που σχεδιάστηκαν και αξιολογήθηκαν για μη δυτική μουσική, όπως για παράδειγμα στο [Pikrakis2004] παρουσιάστηκε μία μέθοδος ρυθμικής κατηγοριοποίησης ελληνικής μουσικής ή στο [Ranjani2013], όπου ένα σύστημα κατηγοριοποίησης καρνατικής μουσικής (ν. Ινδία) περιελάμβανε ένα υποσύστημα ρυθμικής κατηγοριοποίησης.

Κεφάλαιο 2: Ανάλυση Περιοδικότητας Μουσικών Σημάτων

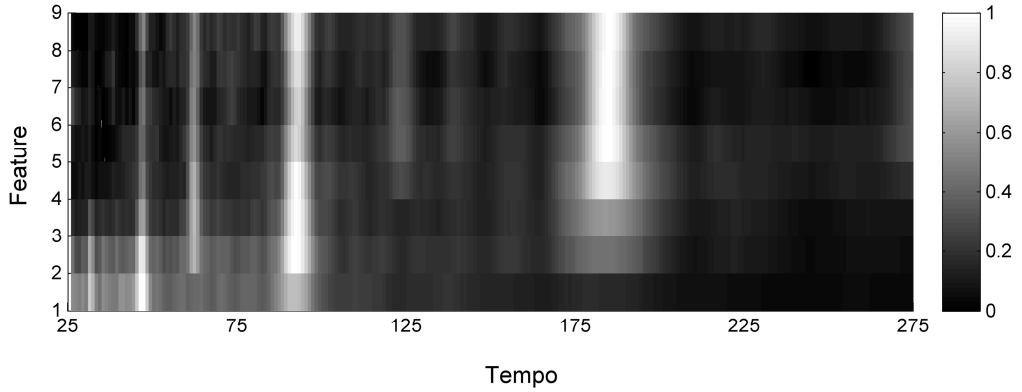
2.1 Εισαγωγή

Ο υπολογισμός μιας *συνάρτησης περιοδικότητας* (periodicity function) είναι από τα σημαντικότερα στοιχεία ενός συστήματος ανάλυσης ρυθμού [Gouyon2005]. Το πεδίο ορισμού της συνάρτησης περιοδικότητας είναι η συχνότητα των παλμών και συνήθως μετριέται σε πλήθος παλμών ανά λεπτό (beats per minute, BPM) ή Hz. Εναλλακτικά μπορούμε τα θεωρήσουμε ότι το πεδίο ορισμού της συνάρτησης περιοδικότητας είναι τα υποψήφια τέμπο. Η συνάρτηση περιοδικότητας αναπαριστά την ισχύ των ρυθμικών περιοδικοτήτων, οι οποίες είναι συνήθως μεταξύ 30 και 300 BPM ή ισοδύναμα μεταξύ 0.5 και 5 Hz [Noorden1999]. Όπως και στην περίπτωση του φασματογραφήματος, αν η συνάρτηση περιοδικότητας υπολογιστεί σε συνεχόμενα χρονικά διαστήματα προκύπτει μια αναπαράσταση χρόνου-περιοδικότητας, όπως για παράδειγμα το τεμπογράφημα [Cemgil2000]. Το εύρος των συχνοτήτων της συνάρτησης περιοδικότητας (0.5-5 Hz) είναι εκτός του φάσματος ακοής του ανθρώπου (20-20kHz). Πώς όμως τότε ο άνθρωπος αντιλαμβάνεται αυτές τις συχνότητες, δηλαδή τον ρυθμό; Όταν ακούμε τον κτύπο ενός μετρονόμου, έστω στα 120 BPM (2 Hz), αυτό που συμβαίνει δεν είναι κάποια φασματική ανάλυση στον κοχλία του αυτιού, ο οποίος μετά διεγείρει ένα νεύρο του εγκεφάλου που μας δίνει απευθείας την αντίληψη της συχνότητας των 2 Hz, όπως θα συνέβαινε στην περίπτωση ενός ήχου με συχνότητα εντός του ακουστικού εύρους του ανθρώπου. Στην ρυθμική αντίληψη, η φασματική ανάλυση στο ανθρώπινο αυτί περιορίζεται στην αναγνώριση μεμονωμένων γεγονότων (δηλαδή των κτύπων του μετρονόμου, ή των μεμονωμένων νοτών στην περίπτωση μιας μουσικής εκτέλεσης). Στην συνέχεια αναλαμβάνει ο ίδιος ο εγκέφαλος με μια δομή χρονικής μνήμης η οποία επεξεργάζεται την ακολουθία αυτών των γεγονότων και μας παρέχει την αντίληψη του ρυθμού [Levitin1996].

Αυτά τα δύο στάδια ανάλυσης από τον άνθρωπο αντανακλούνται και στις περισσότερες μεθόδους αυτόματης ρυθμικής ανάλυσης, οι οποίες έχουν κάποια κοινά χαρακτηριστικά. Στο Σχ. 2.1 παρουσιάζεται μια γενική μορφή ενός συστήματος ανάλυσης περιοδικότητας. Αρχικά, εφαρμόζεται στο ακουστικό σήμα κάποια ανάλυση χρόνου-συχνότητας, όπως ο Μετασχηματισμός Fourier Βραχέως Χρόνου (Short-Time Fourier Transform), ο οποίος προκύπτει με την εφαρμογή του Μετασχηματισμού Fourier σε διαδοχικά παράθυρα του μουσικού σήματος. Στην συνέχεια εξάγονται κάποια χαρακτηριστικά ως προς το χρόνο, τα οποία συνήθως αναφέρονται ως *χαρακτηριστικά έμφασης* (accent features) *ή συνάρτηση έμφασης* (accent function) [Klapuri2006]. Τα χαρακτηριστικά έμφασης υπολογίζονται για διάφορες φασματικές ζώνες και στο πλαίσιο της ρυθμικής ανάλυσης μπορούν να θεωρηθούν *πολυδιάστατα χρονικά* χαρακτηριστικά.. Στη συνέχεια για κάθε διάσταση της συνάρτησης έμφασης ακολουθεί η ανάλυση περιοδικότητας. Το τελικό αποτέλεσμα αυτής της διαδικασίας είναι ένα σύνολο διανυσμάτων των οποίων οι τιμές αντιστοιχούν στο πλάτος ή την ισχύ των εξεταζόμενων ρυθμικών περιοδικοτήτων, ή απλούστερα μία συνάρτηση περιοδικότητας για κάθε διάσταση των χαρακτηριστικών έμφασης. (βλ. Σχήμα 2.2).



Σχήμα 2.1 Γενική μορφή ενός τυπικού συστήματος αυτόματης ρυθμικής ανάλυσης.



Σχήμα 2.2 Συνάρτηση περιοδικότητας κομματιού για διάφορα χαρακτηριστικά έμφασης. Το αληθές τέμπο είναι 93 BPM. Πέραν των 93 BPM παρατηρούνται κορυφές σε πολλαπλάσια και σε υποπολλαπλάσια αυτού (~47 και 185 BPM)

Σε πολλές περιπτώσεις επιπρόσθετα βήματα επεξεργασίας της συνάρτησης περιοδικότητας (εφεξής ΣΠ) μπορούν να εφαρμοστούν, όπως για παράδειγμα την παράθεση ή συνδυασμό των διάφορων ΣΠ, ή την κανονικοποίησή τους ([Klapuri2006], [Ellis2007]) βάσει κάποιας πρότερης κατανομής τέμπο ([Parncutt1994], [Noordenan1999]). Μετά την εξαγωγή και επεξεργασία της ΣΠ μπορούμε να θεωρήσουμε εκ νέου τα χρονικά χαρακτηριστικά, π.χ. στις μεθόδους εύρεσης παλμού είναι σύνηθες μετά την εξαγωγή της ΣΠ, να υπολογιστεί το τέμπο ενός κομματιού και στη συνέχεια δεδομένου αυτού του τέμπο να υπολογίσουμε τον παλμό από τα χρονικά χαρακτηριστικά.

Γενικά μπορούμε να ισχυριστούμε ότι σε ένα σύστημα αυτόματης ρυθμικής ανάλυσης συναντάμε δύο είδη χαρακτηριστικών, τα χρονικά (συνάρτηση έμφασης) και τα φασματικά (συνάρτηση περιοδικότητας), όπου τα δεύτερα υπολογίζονται από τα πρώτα. Ο Πίνακας 2.1 παρουσιάζει ποια προβλήματα αυτόματης ανάλυσης ρυθμού μπορούν να αντιμετωπιστούν χρησιμοποιώντας χρονικά, φασματικά ή και τα δύο είδη χαρακτηριστικών. Από την $1^{η}$ γραμμή του Πίνακα 2.1 μπορούμε να συμπεράνουμε την σημασία της ΣΠ, αφού δεν υπάρχει πρόβλημα που να μην περιλαμβάνει τον υπολογισμό της ΣΠ, με εξαίρεση μία μέθοδο κατηγοριοποίησης ρυθμού που χρησιμοποίησε αποκλειστικά χρονικά χαρακτηριστικά [Dixon2004]. Η $2^{η}$ γραμμή του πίνακα μας δείχνει ποια προβλήματα μπορούν να αντιμετωπιστούν μονάχα με τα φασματικά χαρακτηριστικά, δηλαδή τη ΣΠ, χωρίς την γνώση των χρονικών χαρακτηριστικών.

Χαρατηριστικά	Τέμπο	Παλμός	Αρχή Μέτρου	Μέτρο	Κατηγορία Ρυθμού
Χρονικά	Όχι	Όχι	Όχι	Όχι	Ναι
Φασματικά	Ναι	Όχι	Όχι	Ναι	Ναι
Και τα δύο	Ναι	Ναι	Ναι	Ναι	Ναι

Πίνακας 2.1: Ικανότητα των χρονικών/φασματικών χαρακτηριστικών για την αντιμετώπιση προβλημάτων αυτόματης ρυθμικής ανάλυσης.

Βλέπουμε ότι με την χρήση μονάχα της ΣΠ μπορούμε να εξάγουμε το τέμπο και την ρυθμική κατηγορία [Peeters2005]. Το ίδιο συμβαίνει και με το μέτρο, αφού τέτοια πληροφορία μπορεί να βρεθεί στον λόγο των κορυφών της ΣΠ (π.χ. δύο εξέχουσες κορυφές στα 40 και 120 BPM υποδηλώνει μέτρο 3/4). Για να εξάγουμε όμως τον παλμό και την αρχή του κάθε μέτρου χρειαζόμαστε και τα χρονικά χαρακτηριστικά ($3^{η}$ γραμμή Πίνακας 2.1).

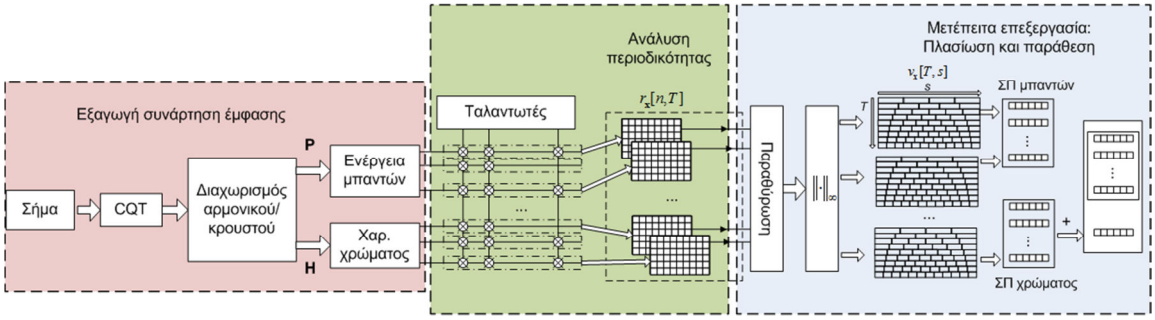
Διάφορες μέθοδοι έχουν προταθεί για τα χαρακτηριστικά έμφασης. Συνήθως ο υπολογισμός τους περιλαμβάνει την εξαγωγή χρονικών ακολουθιών διαφόρων φασματικών χαρακτηριστικών όπως η ενέργεια φασματικών περιοχών, η μεταβολή των φάσεων κ.λπ. Η Φασματική Μιγαδική Διαφορά (Spectral Complex Difference) [Bello2005a], [Davies2007], [Duxbury2002] υπολογίζεται ως το άθροισμα της L_2 νόρμας της διαφοράς των συντελεστών Fourier.

$$SD[n] = \sum_k (H(X[n,k] - X[n-1,k]))^2 \qquad (2.1)$$

όπου $H(\cdot)$ είναι η συνάρτηση ανόρθωσης και $X[n,k]$ είναι ο STFT στη χρονική στιγμή n και συχνότητα k. Άλλοι συγγραφείς χρησιμοποιούν την L_1 νόρμα αντί της L_2 [Marsi1996].

Άλλες μέθοδοι περιλαμβάνουν την ανάλυση του σήματος σε φασματικές περιοχές (μπάντες) πριν από οποιαδήποτε επεξεργασία. Ο Scheirer [Scheirer1998] έκανε το εξής πείραμα. Από μία σειρά μουσικών κομματιών εξήγαγε την ενέργεια σε διάφορες μπάντες και οι αντίστοιχες περιβάλλουσες χρησιμοποιήθηκαν για να διαμορφώσουν λευκό θόρυβο. Βάσει ακουστικών πειραμάτων με ακροατές, προέκυψε ότι το παραγόμενο σήμα διατηρούσε το ρυθμικό περιεχόμενο του αρχικού σήματος, ή πιο απλά, οι ακροατές μπορούσαν να αντιληφθούν τον ρυθμό. Ο αριθμός των μπαντών που προέκυψε ότι είναι κρίσιμος για να διατηρηθεί το ακουστικό περιεχόμενο ήταν μόλις πέντε. Ωστόσο η ανάλυση σε μπάντες αποτυχαίνει να αναπαραστήσει ασθενείς μουσικές μεταβολές, κυρίως αρμονικών αλλαγών (που οφείλονται κυρίως σε αλλαγές στη φάση και όχι στην ενέργεια), όπως για παράδειγμα συμβαίνει στην κλασσική μουσική.

Πολλές μέθοδοι εκτελούν παραγώγιση και ημιανόρθωση του σήματος των χαρακτηριστικών [Klapuri2006], [Ellis2007] πριν από οποιαδήποτε ρυθμική ανάλυση, καταδεικνύοντας ότι μόνο θετικές μεταβολές των φασματικών μεγεθών αντιστοιχούν σε σχετικά με τον ρυθμό γεγονότα, ενώ είναι συνήθης η υιοθέτηση της λογαριθμικής παραγώγου του σήματος που μας δίνει τις μεταβολές του σήματος σε σχέση με το επίπεδο έντασής του [Klapuri2006], [Alonso2007]. Άλλες μέθοδοι χρησιμοποιούν μεθόδους ανίχνευσης «έναρξης νότας» (onset). Η ανίχνευση νότας περιλαμβάνει μια συνεχή «συνάρτηση ανίχνευσης» (onset detection function) η οποία δείχνει την έμφαση κάθε χρονικής στιγμής σχετικά με τον αν αποτελεί onset ή όχι.



Σχήμα 2.3 Διάγραμμα των σταδίων επεξεργασίας της προτεινόμενης μεθόδου ανάλυσης περιοδικότητας.

Η συνάρτηση αυτή αναλύεται συνήθως με κάποιον αλγόριθμο ανίχνευσης κορυφών, προκειμένου να εξαχθούν οι διακριτοί χρόνοι έναρξης νότας. Για την ανάλυση περιοδικότητας χρησιμοποιούνται είτε απευθείας οι συναρτήσεις ανίχνευσης [Peeters2005] είτε οι διακριτοί χρόνοι onset [Dixon2001]. Ωστόσο, τα πειραματικά αποτελέσματα δείχνουν ότι οι συνεχείς αναπαραστάσεις επιτυγχάνουν καλύτερα αποτελέσματα [Gouyon2006].

Η ανάλυση περιοδικότητας περιλαμβάνει συνήθως κάποια συνάρτηση αυτό-ομοιότητας [Davies2007], [Seyerlehner2007] όπως για παράδειγμα την συνάρτηση αυτοσυσχέτισης (autocorrelation)

$$A[m] = \sum_{n=0}^{N-1} x[n] \cdot x[n+m]. \qquad (2.2)$$

Η τιμή της συνάρτησης αυτοσυσχέτισης σε κάθε χρονική στιγμή μας δείχνει μια εκτίμηση της ισχύος του τέμπο με περίοδο ίση με m. Πολλές μέθοδοι χρησιμοποιούν ομάδες ζωνοπερατών φίλτρων (filter-banks) με κεντρικές συχνότητες που αντιστοιχούν στις συχνότητες των τέμπο στόχων [Scheirer1998] [Klapuri2006]. Η απόκριση των φίλτρων ερμηνεύεται ως η ισχύς ή το πλάτος των υποψήφιων τέμπο. Στην περίπτωση που έχουμε διακριτή ως προς τον χρόνο αναπαράσταση, όπως για παράδειγμα στην περίπτωση λίστας από onsets, η συνάρτηση περιοδικότητας μπορεί να υπολογιστεί ως το ιστόγραμμα των μεσοδιαστημάτων των χρόνων έναρξης των νοτών (Inter-Onset-Intervals, IOI) [Dixon2001] [Oliveira2012]. Τα διαστήματα με την μεγαλύτερη τιμή αντιστοιχούν και σε ποιο εξέχουσες ρυθμικές περιοδικότητες.

Στην παρούσα εργασία προτείνεται μια διαδικασία ανάλυσης περιοδικότητας με τρία βασικά συστατικά (Σχ. 2.3). Αντί του συνήθη μετασχηματισμού Fourier βραχέως χρόνου (Short Time Fourier, STFT) υιοθετείται ο μετασχηματισμός σταθερού Q, ο οποίος έχει ιδιότητες που τον καθιστούν πιο κατάλληλο για την αναπαράσταση μουσικών σημάτων. Στη συνέχεια, χρησιμοποιείται μία τεχνική διαχωρισμού πηγών που χωρίζει το σήμα σε δύο συνιστώσες, μία συνιστώσα που περιέχει το «κρουστό» περιεχόμενο (κρουστά όργανα) και μία συνιστώσα με το «αρμονικό» περιεχόμενο (μελωδικά όργανα). Από κάθε συνιστώσα εξάγεται ένα διάνυσμα χαρακτηριστικών το οποίο περιλαμβάνει τις ενέργειες μπαντών από την συνιστώσα των κρουστών και τα χαρακτηριστικά χρώματος από το αρμονικό μέρος. Αυτά τα στάδια συνιστούν την διαδικασία εξαγωγής της συνάρτησης έμφασης. Στη συνέχεια, κάθε συνιστώσα της συνάρτησης έμφασης επεξεργάζεται από ένα σύστημα ανάλυσης περιοδικότητας που είναι βασισμένο στη συνέλιξη με ταλαντωτές, μια μέθοδος η οποία όπως θα δείξουμε στη συνέχεια υπερτερεί έναντι άλλων μεθόδων εξαγωγής ΣΠ καθώς έχει την ικανότητα να μοντελοποιεί και υποτυπώδεις μετρικές σχέσεις. Οι έξοδοι των ταλαντωτών στη συνέχεια

επεξεργάζονται με παραθύρωση και παράθεση των επιμέρους ΣΠ για την εξαγωγή του τελικού διανύσματος περιοδικότητας. Στη συνέχεια αυτού του κεφαλαίου θα περιγραφούν τα επιμέρους στοιχεία της προτεινόμενης ΣΠ. Στην Ενότητα 2.2 θα περιγραφεί ο μετασχηματισμός σταθερού Q, στην Ενότητα 2.3 η τεχνική διαχωρισμού αρμονικών/κρουστών πηγών, ενώ στην Ενότητα 2.4 θα περιγραφούν οι προτεινόμενες συναρτήσεις έμφασης. Η ανάλυση περιοδικότητας και η μετέπειτα επεξεργασία των ΣΠ περιγράφονται στις Ενότητες 2.5 και 2.6 αντίστοιχα.

2.2 Ο μετασχηματισμός σταθερού Q

Ο Μετασχηματισμός Σταθερού Q (Constant Q Transform, CQT) προτάθηκε από τον Brown [Brown1991] και αποτελεί συνδυασμό του Μετασχηματισμού Fourier Βραχέως Χρόνου (Shot-Time Fourier Transform, STFT) και του Wavelet μετασχηματισμού.

Το κίνητρο για την εισαγωγή του CQT ήταν η αδυναμία του STFT να αναπαριστά τα μουσικά σήματα επαρκώς. Αυτό οφείλεται κυρίως στο τετραγωνικό πλέγμα του STFT στον χώρο χρόνου-συχνότητας. Συγκεκριμένα, ο STFT έχει τις παρακάτω δύο ιδιότητες:

- Χρησιμοποιεί σταθερό παράθυρο χρόνου για όλες τις συχνότητες.
- Χρησιμοποιεί γραμμική τοποθέτηση των συχνοτήτων (ή φίλτρων) στο φάσμα.

Αντιθέτως, η φύση των μουσικών σημάτων έχει τις εξής ιδιαιτερότητες:

- Η κατανομή των συχνοτήτων της δυτικής μουσικής είναι γεωμετρική (12 τόνοι/οκτάβα).
- Λόγω τόσο της κατασκευής των ηχητικών πηγών όσο και της ανθρώπινης αντίληψης οι υψίσυχνες νότες εμφανίζουν γρήγορες μεταβολές ως προς τον χρόνο, ενώ οι χαμηλές συχνότητες πιο αργές μεταβολές. Σχετικά με τις ηχητικές πηγές (μουσικά όργανα) η φυσική τους κατασκευή υπαγορεύει τις παραπάνω ιδιότητες. Τα πιο «μπάσα» όργανα είναι μεγαλύτερα σε μέγεθος, επομένως εμφανίζουν μεγαλύτερους χρόνους μεταβατικών φαινομένων μέχρι να ακουστεί μια νότα. Αντίθετα τα μικρά όργανα έχουν πολύ γρηγορότερη απόκριση. Σχετικά με την ανθρώπινη αντίληψη, το ανθρώπινο αυτί χρειάζεται περισσότερο χρόνο για να αντιληφθεί μία χαμηλόσυχνη νότα από μια υψίσυχνη.

Τα δύο αυτά βασικά στοιχεία των μουσικών σημάτων καθιστούν τον STFT ανεπαρκή για την ανάλυση μουσικών σημάτων. Για παράδειγμα, ένας STFT σε συχνότητα δειγματοληψίας 44.1 kHz και με τυπική τιμή μήκους παραθύρου 23ms δίνει 512 ομοιόμορφα κατανεμημένες συχνότητες στο διάστημα [0, 22.050] Hz. Για τους 12 τόνους της μουσικής οκτάβας [55, 110] Hz αντιστοιχεί μονάχα μία συχνότητα Fourier ενώ για την οκτάβα [1.760, 3.520] Hz αντιστοιχούν 40 συχνότητες Fourier.

Ο STFT περιλαμβάνει την παραθύρωση του σήματος και στη συνέχεια τον υπολογισμό του FFT ή ισοδύναμα, για κάθε συχνότητα Fourier το σήμα παραθυρώνεται και σε κάθε παράθυρο υπολογίζεται το εσωτερικό γινόμενο του σήματος με τη μιγαδική ποσότητα $\exp\{-j(2\pi f_k)\}$ με τις συχνότητες Fourier f_k να είναι ομοιόμορφα κατανεμημένες στο φάσμα. Για το διακριτό σήμα $x[n]$, ο STFT $X[m,k]$ μπορεί να γραφτεί ως

$$X[m,k] = \sum_{n=-N_k/2}^{N_k/2} x[n]a_k(n-m+N_k/2), k = 0..N_k-1 \quad (2.3)$$

όπου

$$a_k[l] = {N_k}^{-1} w\left(\frac{l}{N_k}\right) \exp\left\{-j\left(\frac{2\pi l k}{N_k}\right)\right\} = {N_k}^{-1} w\left(\frac{l}{N_k}\right) \exp\left\{-j\left(\frac{2\pi l f_k}{f_s}\right)\right\}, f_k = \frac{k f_s}{N_k}. \quad (2.4)$$

Με f_s συμβολίζεται η συχνότητα δειγματοληψίας, f_k είναι οι k κεντρικές συχνότητες Fourier ομοιόμορφα κατανεμημένες στο διάστημα $[0, f_s/2)$ και $w(\cdot)$ κάποιο παράθυρο στο διάστημα [0,1]. Το N_k συμβολίζει το μήκος του παραθύρου και είναι σταθερό για όλες τις $a_k(\cdot)$. Οι συναρτήσεις $\{a_k(\cdot)\}_k$ αποτελούν ορθογώνιο σύνολο όταν $m = i \cdot N_k/2, i = 1, 2, ..$, δηλαδή η ολίσθηση του παραθύρου ανάλυσης είναι το μισό του μήκους ανάλυσης και είναι οι συναρτήσεις βάσεις του μετασχηματισμού.

Αντίθετα στον CQT οι συχνότητες είναι γεωμετρικά κατανεμημένες στο φάσμα ώστε να επιτυγχάνεται σταθερό πλήθος συχνοτήτων ανά οκτάβα. Επιπλέον η παραθύρωση γίνεται με μεταβλητό μήκος παραθύρου, αντιστρόφως ανάλογο με την συχνότητα, με αποτέλεσμα να επιτυγχάνεται σταθερό πλήθος «κύκλων» σε κάθε παράθυρο. Οι συναρτήσεις $\bar{a}_k(\cdot)$ για τον CQT γράφονται ως:

$$\bar{a}_k(l) = {N_k}^{-1} w\left(\frac{l}{N_k}\right) \exp\left\{-j\left(\frac{2\pi l f_k}{f_s}\right)\right\} \quad (2.5)$$

με $f_k = f_1 \cdot 2^{\frac{k-1}{B}}$ και $f_k < f_{max}$ όπου B είναι το πλήθος των συχνοτήτων/οκτάβα και f_1, f_{max} η μικρότερη / μεγαλύτερη συχνότητα του μετασχηματισμού. Ο συντελεστής ποιότητας Q_k του k στοιχείου ορίζεται ως

$$Q_k = \frac{f_k}{\Delta f_k} = \frac{N_k f_k}{\Delta\omega f_s} \quad (2.6)$$

όπου Δf_k είναι η -3dB συχνότητα της απόκρισης του φίλτρου $\bar{a}_k$ και $\Delta\omega$ η αντίστοιχη συχνότητα για την απόκριση συχνότητας του παραθύρου $w(\cdot)$. Εξ ορισμού επιλέγεται σταθερό $Q_k = Q$ για όλα τα k στοιχεία και προκύπτει

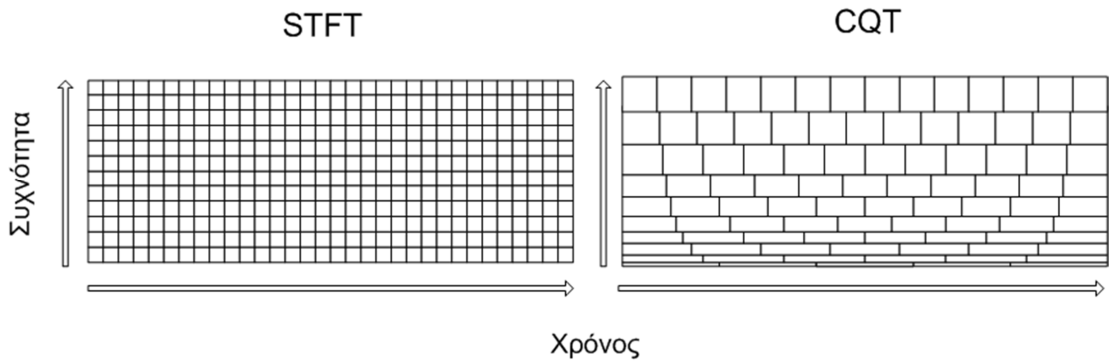
$$N_k = Q\Delta\omega \frac{f_s}{f_k} = \acute{Q} \frac{f_s}{f_k}. \quad (2.7)$$

Επομένως το μήκος των παραθύρων είναι αντιστρόφως ανάλογο με την κεντρική συχνότητα του κάθε στοιχείου. Αυτό έχει ως αποτέλεσμα την επίτευξη σταθερού πλήθος κύκλων σε κάθε στοιχείο, το οποίο ισούται με $N_k f_k/f_s = \acute{Q}$. Ο CQT γράφεται αναλυτικά

$$CQT[m,k] = \sum_{n=-N_k/2}^{N_k/2} x[n]\bar{a}_k(n-m+N_k/2), k = 0..N_k-1 \quad (2.8)$$

Στο Σχήμα 2.4 παρουσιάζεται το πλέγμα χρόνου συχνότητας για τον STFT και τον CQT.

Φαίνεται καθαρά ότι ενώ στον STFT το πλέγμα είναι ομοιόμορφα κατανεμημένο στους δύο άξονες, στην περίπτωση του CQT έχουμε υψηλότερη ανάλυση συχνότητας και χαμηλότερη ανάλυση χρόνου στις χαμηλές συχνότητες, ενώ στις υψηλές συχνότητες έχουμε υψηλότερη ανάλυση χρόνου και χαμηλότερη ανάλυση στον άξονα της συχνότητας. Πρέπει να σημειωθεί ότι όπως και στον STFT, έτσι και στον CQT το εμβαδόν κάθε στοιχείου του πλέγματος είναι σταθερό.



Σχήμα 2.4: Το πλέγμα χρόνου-συχνότητας για τον STFT και τον CQT.

Λόγω του μεταβλητού μήκους παραθύρου για κάθε κεντρική συχνότητα, το πλήθος των τιμών του CQT είναι διαφορετικό για κάθε k, γεγονός που δυσκολεύει την ταυτόχρονη χρονική ανάλυση όλων των κεντρικών συχνοτήτων. Προκειμένου να αποκτήσουν ίδιο μήκος N, εφαρμόζουμε κυβική παρεμβολή για κάθε γραμμή (δηλ. κεντρική συχνότητα) του CQT, με αποτέλεσμα μια κανονικοποιημένη μορφή του CQT:

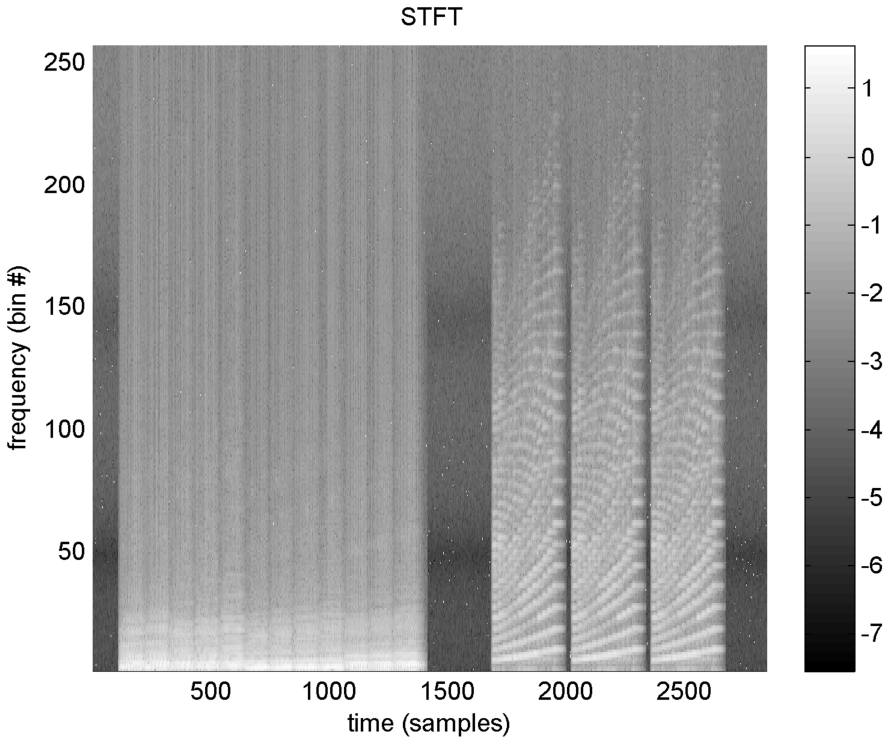
$$\overline{\mathrm{CQT}}[m,k] = cubic(\overline{\mathrm{CQT}}(\cdot,\mathrm{k}),N) \tag{2.9}$$

όπου η συνάρτηση $cubic(\mathbf{x},N)$ επιστρέφει ένα νέο διάνυσμα μήκους N το οποίο έχει εξαχθεί με κυβική παρεμβολή από το διάνυσμα $\mathbf{x}$.

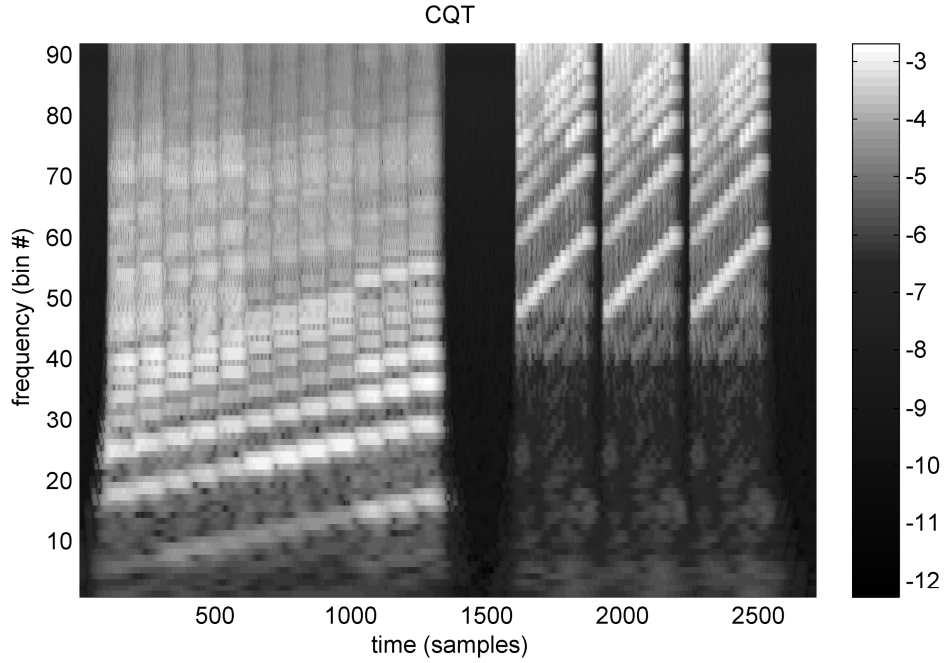
Στα σχήματα 2.5 και 2.6 παρουσιάζονται αντίστοιχα ο STFT και ο κανονικοποιημένος CQT ενός πρότυπου σήματος, το οποίο αποτελείται από δύο μέρη. Την χρωματική κλίμακα ενός κοντραμπάσου σε μία οκτάβα, ακολουθούμενη από την εκτέλεση τρεις φορές της ίδιας χρωματικής κλίμακας από ένα βιολί (σε πολύ υψηλότερη οκτάβα) και με πολύ μεγαλύτερη ταχύτητα. Ο CQT κανονικοποιήθηκε (Εξ. 2.9) έτσι ώστε να έχει το ίδιο πλήθος πλαισίων ανά μονάδα χρόνου με τον STFT (~200 frames/s). Όπως φαίνεται και στις δύο εκτελέσεις, ο CQT υπερτερεί φανερά του STFT. Στην περίπτωση του κοντραμπάσου, παρατηρούμε ότι ο STFT αποτυγχάνει εντελώς να αναπαραστήσει τις χαμηλόσυχνες νότες, αφού όπως προαναφέρθηκε έχει πολύ χαμηλή συχνοτική ανάλυση σε αυτές. Αντίθετα ο CQT είναι αποτελεσματικός και στις δύο εκτελέσεις.

2.3 Διαχωρισμός αρμονικών / κρουστών πηγών

Προκειμένου να αναπαρασταθεί το μουσικό σήμα με όσο το δυνατόν ανεξάρτητες και συμπληρωματικές συνιστώσες, πριν από οποιαδήποτε ρυθμική επεξεργασία προτείνεται ο διαχωρισμός του μουσικού σήματος σε «αρμονικές» και «κρουστές» πηγές. Το αρμονικό μέρος του σήματος, που περιλαμβάνει όλα τα αρμονικά (μη κρουστά) μουσικά όργανα θα περιέχει ρυθμική πληροφορία σχετικά με την μελωδία και κυρίως τις απαλές αλλαγές. Αντιθέτως, το κρουστό μέρος θα περιέχει ρυθμική πληροφορία που οφείλεται κυρίως σε απότομες μεταβολές της συνολικής ενέργειας του σήματος, όπως δηλαδή συμβαίνει στα κρουστά όργανα.



Σχήμα 2.5: Ο STFT μετασχηματισμός του πρότυπου σήματος. Συχνότητα δειγματοληψίας σήματος 44.1kHz, μήκος παραθύρου Fourier 512 samples, ολίσθηση 210 samples.



Σχήμα 2.6: Ο CQT μετασχηματισμός του πρότυπου σήματος. $f_1 = 25$ Hz, $f_{max} = 5\ kHz$, $B = 12$. Επειδή $B = 12$ (12 bins/οκτάβα) κάθε συχνότητα αντιστοιχεί σε έναν τόνο της δυτικής μουσικής κλίμακας.

Ο διαχωρισμός αρμονικού/κρουστού σήματος συναντάται στην βιβλιογραφία κυρίως ως μέθοδος διαχωρισμού και μεταγραφής των ντραμς. Στην εργασία

[FitzGerald2009] οι συγγραφείς πρότειναν μοντέλα παραγοντοποίησης τανυστών (Tensor Factorization Model) για να διαχωρίσουν τα ντραμς από πολυφωνική μουσική, ενώ στην [Gilliet2008] οι συγγραφείς ανέλυσαν το σήμα σε δύο συνιστώσες, μία αρμονική και μία συνιστώσα θορύβου. Η δεύτερη συνιστώσα ερμηνεύτηκε ως ήχος των ντραμς. Οι Yoshii, Goto, και Okuno [Yoshii2007] πρότειναν ένα σύστημα αναγνώρισης ντραμς όπου μετά την καταστολή του αρμονικού περιεχομένου στο φασματογράφημα, ακολουθεί το ταίριασμα με πρότυπους ήχους (templates). Οι Helen και Virtanen [Helen2005] εφάρμοσαν παραγοντοποίηση σε μη αρνητικούς πίνακες (Non Negative Matrix Factorization, NMF) και μηχανές διανυσμάτων υποστήριξης (Support Vector Machines) για τον διαχωρισμό των ντραμς από πολυφωνική μουσική.

Εργασία ορόσημο στον διαχωρισμό αρμονικών/κρουστών πηγών αποτελούν οι εργασίες [Ono2008a,b]. Με την παραδοχή ότι οι κρουστοί/αρμονικοί ήχοι προκαλούν κάθετες/οριζόντιες γραμμές στο φασματογράφημα αντίστοιχα, η εξαγωγή των δύο συνιστωσών πραγματοποιείται με την αμοιβαία ελαχιστοποίηση των κάθετων και οριζόντιων παραγώγων μιας συμπιεσμένης μορφής του φασματογραφήματος. Σε επόμενη εργασία τους οι ίδιοι συγγραφείς [Duong2011] έλαβαν υπόψη και την χωρική και συχνοτική συνέχεια των δύο συνιστωσών σε στερεοφωνικές ηχογραφήσεις για να εξάγουν το τελικό αποτέλεσμα. Βάσει της ίδιας παραδοχής, ο FitzGerald [FitzGerald2011] υιοθέτησε τη χρήση φίλτρων μέσων (median filters) στις γραμμές /στήλες του φασματογραφήματος για να εξαχθούν μάσκες για την αρμονική/κρουστή συνιστώσα αντίστοιχα. Στο [Thoshkahna2011] οι συγγραφείς επέκτειναν την εργασία του FitzGerald προτείνοντας μια επιπλέον μέθοδο επεξεργασίας στην κρουστή συνιστώσα. Οι «αρμονικές διαρροές» (harmonic leaks) εξαλείφθηκαν με την ανακατασκευή του σήματος χρησιμοποιώντας την περιβάλλουσα του σήματος στις επιμέρους μπάντες.

Στη παρούσα εργασία τροποποιήθηκε η μέθοδος [FitzGerald2011] προς δύο κατευθύνσεις. Την χρησιμοποίηση μορφολογικών τελεστών στον CQT του σήματος αντί φίλτρων μέσου και την χρήση διάφορων δομικών στοιχείων. Τα φίλτρα αυτά δεν έχουν ξαναχρησιμοποιηθεί για την ανάλυση μουσικών σημάτων και τα πειραματικά αποτελέσματα δείχνουν ότι αποτελούν μία αποδοτική και εύκολα υλοποιήσιμη τεχνική για τον διαχωρισμό αρμονικών/κρουστών πηγών.

Έστω $\mathbf{X} = X[m,k]$ μια αναπαράσταση χρόνου-συχνότητας ενός σήματος. Αυτή η αναπαράσταση μπορεί να είναι είτε ο STFT είτε ο CQT. Αντί της χρήσης φίλτρων μέσων θα υιοθετήσουμε τα μη γραμμικά φίλτρα διάβρωσης (Erosion), διαστολής (Dilation), ανοίγματος (Opening) και κλεισίματος (Closing). Για την περιγραφή των φίλτρων αυτών ο αναγνώστης παραπέμπεται στο [Serra1982].

Με βάση την παραδοχή ότι τα κρουστά/αρμονικά μέρη του μουσικού σήματος «εμφανίζονται» ως κατακόρυφες/οριζόντιες γραμμές στο φασματογράφημα αντίστοιχα, η διαίσθηση μας οδηγεί στο συμπέρασμα ότι η εφαρμογή των παραπάνω μη γραμμικών φίλτρων με δομικά στοιχεία κατακόρυφες και οριζόντιες γραμμές θα προκαλούσε κάποιου είδους διαχωρισμό των δύο συνιστωσών του σήματος.

Στο Σχήμα 2.7(α) παρουσιάζεται ο CQT του πρότυπου σήματος της προηγούμενης παραγράφου στο οποίο έχει προστεθεί το χτύπημα ενός τυμπάνου (ταμπούρου) (Σχ. 2.7(β)), συγχρονισμένο ρυθμικά με τα μουσικά όργανα έτσι ώστε να τα επικαλύπτει. Ο CQT της παράθεσης των δύο σημάτων παρουσιάζεται στο Σχ. 2.7(γ). Είναι φανερό ότι τα χτυπήματα του τυμπάνου παρουσιάζονται ως κάθετες γραμμές στον CQT. Στα Σχήματα 2.7 (δ)-(ε) παρουσιάζεται ο CQT όταν εφαρμόζεται σε αυτόν ένας από τους τελεστές που προαναφέρθηκαν, σε αυτή την

περίπτωση ο τελεστής ανοίγματος (opening), χρησιμοποιώντας κάθετη/οριζόντια γραμμή ως δομικό στοιχείο. Παρατηρούμε ότι η χρήση του οριζόντιου στοιχείου απαλείφει το κρουστό περιεχόμενο, ενώ η χρήση του κάθετου στοιχείου το αρμονικό περιεχόμενο. Ορίζοντας ως $\widehat{\mathbf{H}}_r = \hat{H}_r[m,k]$ το αποτέλεσμα της χρήσης φίλτρων με οριζόντιο δομικό στοιχείο, $\widehat{\mathbf{P}}_r = \hat{P}_r[m,k]$ με κάθετο δομικό στοιχείο και με r τον τύπο του φίλτρου, και υπό τον περιορισμό ότι τα δύο επιμέρους σήματα πρέπει να αθροίζουν στο αρχικό σήμα, υιοθετούμε τις μάσκες $\mathbf{M}_r^p$, $\mathbf{M}_r^p$ ως εξής:

$$M_r^p[m,k] = \frac{(\hat{H}_r[m,k])^2}{(\hat{H}_r[m,k])^2+(\hat{P}_r[m,k])^2}$$
$$M_r^h[m,k] = \frac{(\hat{P}_r[m,k])^2}{(\hat{H}_r[m,k])^2+(\hat{P}_r[m,k])^2} \quad (2.10)$$

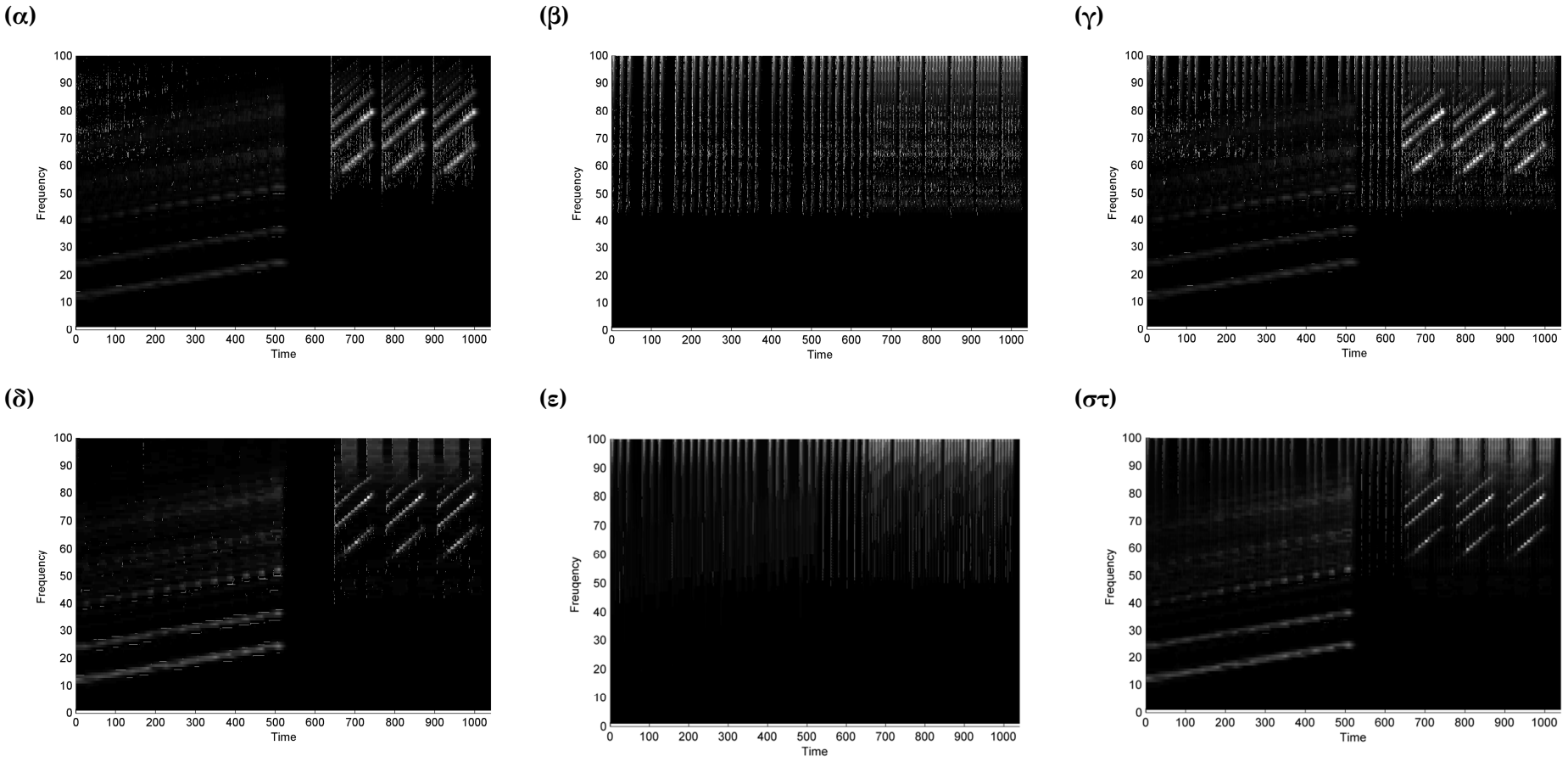
όπου $r \in \{E, D, O, C\}$, με τα E, D, O και C να αντιστοιχούν στα Erosion, Dilation, Opening και Closure. Τότε, οι CQT των δύο συνιστωσών $\mathbf{H}_r$ (harmonic) και $\mathbf{P}_r$ (percussive) θα είναι ίσες με

$$\mathbf{H}_r = \mathbf{M}_r^p \circ \mathbf{X}, \quad \mathbf{P}_r = \mathbf{M}_r^h \circ \mathbf{X}. \quad (2.11)$$

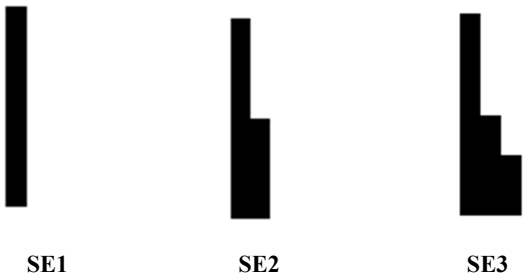
Όταν τα δομικά στοιχεία είναι κάθετες/οριζόντιες γραμμές τότε οι τελεστές διάβρωσης, διαστολής, ανοίγματος και κλεισίματος είναι ισοδύναμες με την εφαρμογή min, max, maxmin και minmax φίλτρων στις γραμμές και τις στήλες του φασματογραφήματος αντίστοιχα. Αφού υπάρχει μια στενή σχέση μεταξύ των μορφολογικών φίλτρων και των φίλτρων μέσων [Maragos1987], η προτεινόμενη μέθοδος είναι παρεμφερής στην μέθοδο του FitzGerald. Στην παρούσα διατριβή διερευνήσαμε την επέκταση των δομικών στοιχείων των φίλτρων σε παραπάνω από μία διαστάσεις. Παρόλο που οι κρουστοί ήχοι παρουσιάζουν κάθετες γραμμές στο φασματογράφημα, αυτές οι γραμμές τείνουν να είναι πιο παχιές στις χαμηλότερες συχνότητες. Για το δείξουμε αυτό, υιοθετήσαμε πιο σύνθετα δομικά στοιχεία για να ενισχύσουμε το κρουστό μέρος.

Ορίζουμε τα δομικά στοιχεία SE1, SE2 και SE3 (Structure Elements) ως τις δυαδικές εικόνες που αντιστοιχούν στα σχήματα που φαίνονται στο Σχήμα 2.8. Τα πειραματικά αποτελέσματα δείχνουν ότι αυτά τα δομικά στοιχεία συλλαμβάνουν καλύτερα τα κρουστά στοιχεία. Κάτι αντίστοιχο δεν παρατηρήθηκε για την αρμονική συνιστώσα, δηλαδή η χρήση αντίστοιχων οριζόντιων στοιχείων για την εξαγωγή του αρμονικού μέρους δεν βελτίωσε την ποιότητα διαχωρισμού.

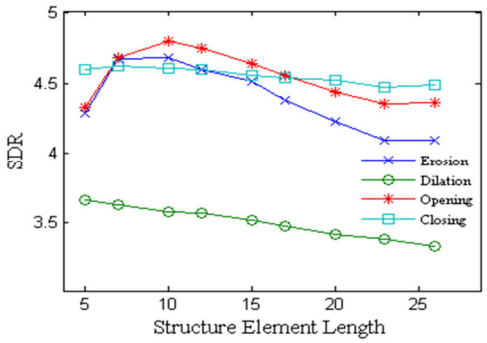
Λόγω του γεγονότος ότι ο CQT έχει μόνο προσεγγιστικά αντίστροφο, προκειμένου να αξιολογήσουμε την προτεινόμενη μέθοδο διαχωρισμού χρησιμοποιήσαμε τον STFT ως χρονοσυχνοτική αναπαράσταση. Η συχνότητα δειγματοληψίας ήταν 44.1 kHz, το μήκος παραθύρου 1024 δείγματα με 512 δείγματα ολίσθηση παραθύρου τύπου Hanning. Η αξιολόγηση έγινε στα δεδομένα αναφοράς SISEC 2008/2010 [Vincent2012]. Ως μετρική αξιολόγησης χρησιμοποιήθηκε ο λόγος Σήματος – Παραμόρφωσης (Signal to Distortion Ratio, SDR) όπως έχει οριστεί στο [Vincent2006].



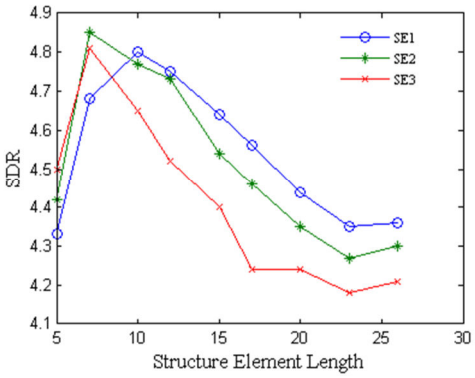
Σχήμα 2.7: Ο CQT του (α) αρμονικού σήματος, (β) του κρουστού σήματος και (γ) της μίξης τους. Στα (δ) και (ε) φαίνεται το αποτέλεσμα του τελεστή «ανοίγματος» (opening) με οριζόντιες (δ) και κάθετες (ε) γραμμές. Είναι ξεκάθαρη η ομοιότητα των (α) και (β) με τα (δ) και (ε) αντίστοιχα. Στο (στ) παρουσιάζεται η μίξη των (δ) και (ε).



Σχήμα 2.8. Δομικά στοιχεία για την εξαγωγή του κρουστού μέρους.



(α)

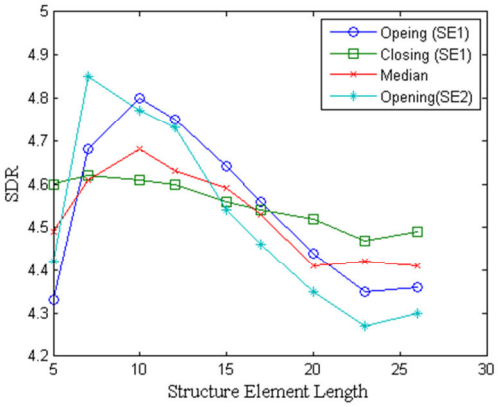


(β)

(α) Ποιότητα διαχωρισμού για τα 4 φίλτρα για διάφορες τιμές του μήκους του δομικού στοιχείου SE1.

(β) Ποιότητα διαχωρισμού για τα τρία δομικά στοιχεία για διάφορες τιμές του μήκους τους με χρήση του φίλτρου ανοίγματος.

(γ) Ποιότητα διαχωρισμού της προτεινόμενης μεθόδου συγκριτικά με τη χρήση φίλτρων μέσων.



(γ)

Σχήμα 2.9. Αποτελέσματα ποιότητας διαχωρισμού κρουστών/αρμονικών πηγών.

Στο Σχ. 2.9(α) παρουσιάζεται η ποιότητα του διαχωρισμού (SDR) για τα τέσσερα διαφορετικά φίλτρα, με χρήση του στοιχείου SE1 για διάφορες τιμές του μήκους του. Η καλύτερη επίδοση επιτεύχθηκε με την χρήση του φίλτρου ανοίγματος, και με βέλτιστο μήκος στο εύρος 7 έως 17. Το φίλτρο κλεισίματος είναι λιγότερο ευαίσθητο στο μήκος του SE, αλλά το SDR είναι περίπου 0.2 dB χαμηλότερο. Σε κάποιες περιπτώσεις η διάβρωση επιτυγχάνει καλύτερη επίδοση από το κλείσιμο, αλλά το SDR μειώνεται γρηγορότερα όσο το μήκος του SE μεγαλώνει. Η χρήση της διαστολή από την άλλη είναι υποδεέστερη προσέγγιση, αφού το SDR είναι τουλάχιστον 1 dB λιγότερο από τις άλλες περιπτώσεις.

Για να δείξουμε την επίδραση των διαφορετικών δομικών στοιχείων μετρήσαμε το SDR για τα SE1, SE2, SE3 για διάφορες τιμές μήκους χρησιμοποιώντας το φίλτρο ανοίγματος (Σχ. 2.9(β)). Αυτό που παρατηρούμε

είναι ότι με τη χρήση των SE2 και SE3 η μεγαλύτερη τιμή SDR επιτυγχάνεται για μικρότερες τιμές μήκους συγκριτικά με το SE1. Επομένως μπορούμε να ισχυριστούμε ότι τα πιο σύνθετα SE αναπαριστούν καλύτερα τα κρουστά μέρη του σήματος. Ωστόσο όσο το μέγεθος του SE μεγαλώνει, η επίδοση μειώνεται, επειδή τα δομικά αυτά στοιχεία προκαλούν παραμόρφωση του φασματογραφήματος. Στο Σχήμα 2.9(γ) παρουσιάζεται η επίδοση της προτεινόμενης μεθόδου συγκρινόμενη με τη χρήση φίλτρων μέσων. Η προτεινόμενη μέθοδος υπερτερεί κατά 0.2 dB για ένα μεγάλο εύρος μηκών του SE. Επιπλέον πρέπει να σημειωθεί ότι το υπολογιστικό κόστος αυτών των φίλτρων είναι μικρότερο από τα φίλτρα μέσων. Τα φίλτρα μέσων χρειάζονται την διάταξη τουλάχιστον των μισών τιμών, ενώ τα μορφολογικά φίλτρα απαιτούν την εύρεση της ελάχιστης/μέγιστης τιμής.

2.4 Εξαγωγή Συνάρτησης Έμφασης

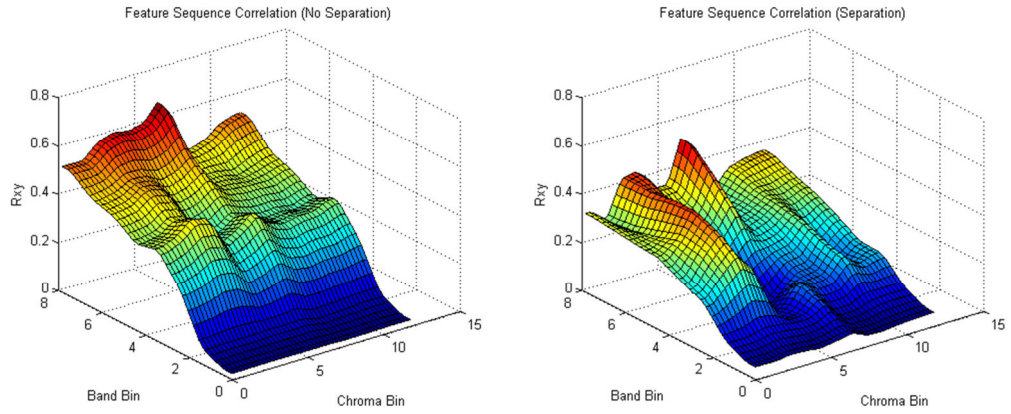
Η συνάρτηση έμφασης αποτελεί την κύρια είσοδο ενός συστήματος αυτόματης ρυθμικής ανάλυσης. Αποτελεί την ενδιάμεση αναπαράσταση ανάμεσα στο αρχικό ηχητικό σήμα και στην ανάλυση περιοδικότητας. Όπως αναφέρθηκε στην Ενότητα 2.1 υπάρχει μια πληθώρα χαρακτηριστικών έμφασης που συναντώνται στην βιβλιογραφία για την αναπαράσταση των ρυθμικών συνιστωσών του μουσικού σήματος. Η παρούσα εργασία πραγματεύεται δύο κύριες κατηγορίες χαρακτηριστικών έμφασης, τα χαρακτηριστικά ενέργειας φασματικών μπαντών και τα χαρακτηριστικά χρώματος. Και τα δύο είδη χαρακτηριστικών εξάγονται από την αναπαράσταση χρόνου συχνότητας, που στη συγκεκριμένη περίπτωση είναι ο CQT και για κάθε πλαίσιο ανάλυσης (frame) m. Επομένως οι συναρτήσεις έμφασης έχουν το ίδιο μήκος ως προς τον χρόνο με τον CQT. Τα χαρακτηριστικά ενέργειας μπαντών υπολογίζονται με τριγωνικά φίλτρα ισοκατανεμημένα στη λογαριθμική κλίμακα στο φάσμα. Στη συνέχεια ο λογάριθμος της ενέργειας υπολογίζεται από την έξοδο του κάθε φίλτρου. Λόγω του ότι οι συχνότητες του CQT είναι γεωμετρικά κατανεμημένες, τα φίλτρα αυτά είναι γραμμικά στην κλίμακα συχνοτήτων του CQT. Αν $\mathbf{X} = X[m,k]$ είναι ο CQT και $\mathbf{\Phi} = \Phi[k,i]$ είναι ο πίνακας που περιγράφει τις αποκρίσεις συχνότητας των i φίλτρων, τότε η ενέργεια της i μπάντας για κάθε πλαίσιο (frame) ανάλυσης m υπολογίζεται ως:

$$E[m,i] = \sum_k X[m,k] \cdot \Phi[k,i] \tag{2.12}$$

Τα χαρακτηριστικά ενέργειας $\mathbf{x}_e^i$ προκύπτουν με λογαρίθμιση της ενέργειας της (2.12). Όπως αναφέρθηκε και στην Ενότητα 2.1 συνήθως οι ακολουθίες των χαρακτηριστικών έμφασης παραγωγίζονται ως προς το χρόνο [Klapuri2006] καθώς οι μεταβολές στο σήμα είναι αυτές που περιέχουν την ρυθμική πληροφορία. Τα τελικά χαρακτηριστικά ενέργειας $\mathbf{x}_e^i$ προκύπτουν με παραγώγιση της ενέργειας με το παρακάτω μη-αιτιατό φίλτρο: $\mathbf{x}_{ch}^i \cup \mathbf{x}_e^k$

$$\dot{x}_e^i[m] = \frac{1}{L(L+1)} \sum_{k=-L}^{k=L} x_e^i[m+k]/k \tag{2.13}$$

Η χρήση λογαρίθμου της ενέργειας και η παραγώγιση μας δίνουν την μεταβολή του σήματος σε σχέση με το επίπεδο έντασής του [Klapuri2006], [Alonso2007] του για κάθε μπάντα i και σε κάθε χρονική στιγμή m. Τα χαρακτηριστικά ενέργειας αναμένουμε να αναπαραστήσουν πλατιές μεταβολές της ενέργειας οι οποίες οφείλονται κυρίως σε χτυπήματα κρουστών, ή ισχυρά onsets.



(α) (β)

Σχήμα 2.10. Συσχέτιση χαρακτηριστικών χρώματος με τα χαρακτηριστικά ενέργειας α) χωρίς διαχωρισμό πηγών β) μετά τον διαχωρισμό πηγών.

Για την αναπαράσταση πιο εκλεπτυσμένων χαρακτηριστικών, όπως απαλές μελωδικές αλλαγές οι οποίες δεν παρουσιάζουν αισθητές μεταβολές στην ενέργεια του φάσματος υιοθετούμε τα «χαρακτηριστικά χρώματος». Τα χαρακτηριστικά χρώματος χρησιμοποιούνται κυρίως για την αναγνώριση συγχορδιών [Ni2012] και την αναγνώριση μουσικού κλειδιού (Music Key Extraction) [Papadopoulos2011]. Ωστόσο η διαίσθησή μας υπαγορεύει ότι πέραν της μελωδικής περιέχουν και ρυθμική πληροφορία. Τα χαρακτηριστικά χρώματος χρησιμοποιήθηκαν για πρώτη φορά στο πλαίσιο της ρυθμικής ανάλυσης για το πρόβλημα εξαγωγής μουσικού τέμπο από τους Eronen και Klapuri [Eronen2010]. Συνιστούν ένα διάνυσμα 12 διαστάσεων για κάθε πλαίσιο ανάλυσης (frame) m. Κάθε μία διάσταση αντιστοιχεί σε έναν από τους 12 φθόγγους της δυτικής κλίμακας. Η τιμή του διανύσματος χρώματος για κάθε φθόγγο ισούται με την ενέργεια του φάσματος για τις συχνότητες που αντιστοιχούν σε αυτόν τον φθόγγο. Πιο συγκεκριμένα, γράφουμε

$$x_{ch}^{j}[m] = \sum_{k \in K(j)} X[m,k], j = 1..12 \qquad (2.14)$$

όπου με $K(j)$ συμβολίζεται το πλήθος των συχνοτήτων του CQT που αντιστοιχούν στο χρώμα j.

Όμοια με τα χαρακτηριστικά ενέργειας, λογαριθμίζουμε και στη συνέχεια παραγωγίζουμε για να εξάγουμε τα τελικά χαρακτηριστικά χρώματος. Βάσει των συμπερασμάτων της προηγούμενης παραγράφου σε σχέση με τον διαχωρισμό της αρμονικής και κρουστής συνιστώσας από το αρχικό σήμα, αναμένουμε ότι κάθε μία από τις δύο συνιστώσες του σήματος θα συνεισφέρει περισσότερο και σε ένα από τα δύο είδη χαρακτηριστικών. Είναι φυσική υπόθεση ότι οι φασματικές ενέργειες είναι πιο έκδηλες στο κρουστό μέρος, ενώ τα χαρακτηριστικά χρώματος στο αρμονικό μέρος.

Για να το δείξουμε αυτό, ας θεωρήσουμε το πρότυπο σήμα της Ενότητας 2.3 (Σχ. 2.6). Από αυτό εξήχθησαν τα χαρακτηριστικά ενέργειας/χρώματος (α) απευθείας από τον CQT και (β) μετά από τον κρουστό /αρμονικό διαχωρισμό. Στόχος μας είναι να δείξουμε ότι τα δύο είδη χαρακτηριστικών είναι πιο «ασυσχέτιστα» μετά τον διαχωρισμό των πηγών. Και για τις δύο περιπτώσεις (διαχωρισμός /μη διαχωρισμός πηγών) υπολογίστηκε η συσχέτιση (correlation)

κάθε ακολουθίας χαρακτηριστικών της μίας κατηγορίας με όλα τα χαρακτηριστικά της άλλης ως:

$$R = \left[r_{i,j}\right]_{i,j},\ r_{i,j} = \frac{<\mathbf{x}_{ch}^{j} - \bar{\mathbf{x}}_{ch}^{j}, \mathbf{x}_{e}^{i} - \bar{\mathbf{x}}_{e}^{i}>}{\left\|\mathbf{x}_{ch}^{j} - \bar{\mathbf{x}}_{ch}^{j}\right\| \cdot \left\|\mathbf{x}_{e}^{i} - \bar{\mathbf{x}}_{e}^{i}\right\|} \quad (2.15)$$

Στο Σχήμα 2.10 παρουσιάζεται η συσχέτιση των δύο ειδών χαρακτηριστικών με ή χωρίς τον διαχωρισμό πηγών.

Η συνολική μέση συσχέτιση των δύο οικογενειών χαρακτηριστικών είναι 0.28 όταν υπολογίζουμε τα χαρακτηριστικά κατευθείαν από τον CQT, ενώ μειώνεται σε 0.2 όταν εφαρμόζεται ο διαχωρισμός πηγών. Τόσο το γεγονός ότι (α) η συσχέτιση και στις δύο περιπτώσεις είναι σχετικά χαμηλή, αλλά και (β) ότι μικραίνει αισθητά όταν εφαρμόζεται ο διαχωρισμός πηγών, επαληθεύουν αφενός ότι οι δύο τύποι χαρακτηριστικών περιέχουν συμπληρωματική ρυθμική πληροφορία, αφετέρου ότι ο διαχωρισμός πηγών κάνει τις δύο κατηγορίες λιγότερο σχετιζόμενες μεταξύ τους.

2.5 Ανάλυση Περιοδικότητας

Όπως αναφέρθηκε και προηγουμένως, η ανάλυση περιοδικότητας είναι μια διαδικασία όπου μια (πολυδιάστατη) ακολουθία των χαρακτηριστικών έμφασης απεικονίζεται σε ένα διάνυσμα(τα) του οποίου το πεδίο ορισμού είναι η περιοδικότητα (ή η συχνότητα) του ρυθμού. Συνοψίζοντας την επισκόπηση της Ενότητας 2.1 οι κύριες τεχνικές ανάλυσης της περιοδικότητας μπορούν να απαριθμηθούν ως εξής:

Inter-Onsets-Intervals (IOIs) Histogram

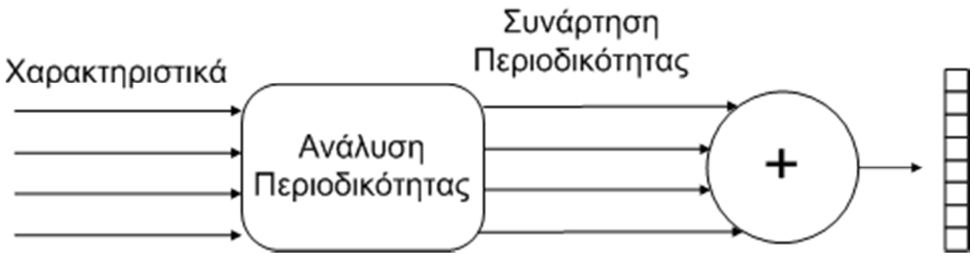
Αυτή η μέθοδος βρίσκει εφαρμογή στη λίστα των onsets που υπάρχουν σε ένα μουσικό απόσπασμα. Στην περίπτωση συμβολικής εισόδου (π.χ. αρχεία MIDI), η εξαγωγή των onsets γίνεται ευθέως και χωρίς λάθη από την είσοδο. Στην περίπτωση αρχείων ήχου, μεσολαβεί ένα στάδιο ανίχνευσης onsets. Προκειμένου να εξαχθεί το διάνυσμα περιοδικότητας, υπολογίζονται τα διαστήματα μεταξύ των onsets, και το ιστόγραμμά τους θεωρείται ως το διάνυσμα περιοδικότητας. Το πεδίο ορισμού του ιστογράμματος αντιστοιχεί σε ρυθμικές περιόδους.

Συνάρτηση αυτοσυσχέτισης (auto-correlation function, ACF)

Η συνάρτηση αυτοσυσχέτισης (Εξ. 2.2) είναι αρκετά δημοφιλής [Davies2007] [Seyerlehner2007] ως συνάρτηση περιοδικότητας. Όπως και στην περίπτωση των IOIs Histogram, το πεδίο ορισμού είναι οι χρονικές καθυστερήσεις, οι οποίες αντιστοιχούν σε ρυθμικές περιοδικότητες. Η τιμή της συνάρτησης στην καθυστέρηση (lag) τ δίνει μια εκτίμηση της "ισχύος" του τέμπο που αντιστοιχεί σε περίοδο τ.

Συνέλιξη με ταλαντωτές

Μια αρκετά δημοφιλής μέθοδος εξαγωγής της περιοδικότητας είναι η συνέλιξη με μια συστοιχία ταλαντωτών. Κάθε ταλαντωτής έχει συχνότητα ταλάντωσης που αντιστοιχεί σε κάποιο από τα υποψήφια τέμπο. Η τιμή της εξόδου των ταλαντωτών με κεντρικές συχνότητες τις συχνότητες των τέμπο στόχων μας δίνει μια εκτίμηση του πόσο κυρίαρχο είναι κάθε τέμπο στο μουσικό σήμα.



Σχήμα 2.11 Ανάλυση περιοδικότητας ανά συνιστώσα χαρακτηριστικών.

Ένα άλλο πολύ σημαντικό στοιχείο των συστημάτων ανάλυσης περιοδικότητας, είναι η χωριστή περιοδική ανάλυση κάθε διάστασης της συνάρτησης έμφασης και ο μετέπειτα συνδυασμός τους [Shceirer1998, Gouyon2005]. Αν $x^i[m]$ είναι η πολυδιάστατη συνάρτησης έμφασης, η γενική μορφή της συνάρτησης περιοδικότητας είναι

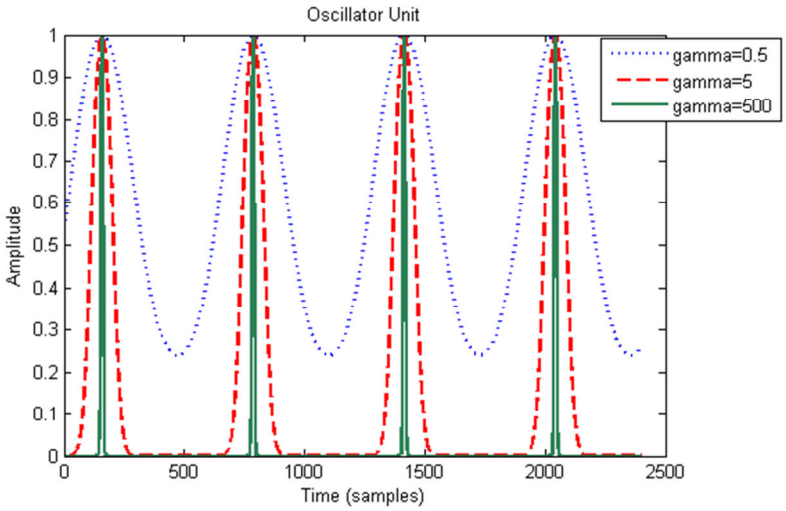
$$v_{\mathbf{x}}[T] = \sum_i pf(x^i[m], T) \tag{2.16}$$

όπου $pf(\cdot)$ είναι η συνάρτηση περιοδικότητας που μπορεί να είναι μία από οποιαδήποτε από τις προαναφερθείσες και T το τέμπο. Στο Σχήμα 2.10 παρουσιάζεται η γενική μορφή της περιοδικής ανάλυσης (μεσαίο μπλοκ του Σχ. 2.3).

Μια πρώτη προσέγγιση που θα μπορούσε να ακολουθήσει κανείς για τον υπολογισμό ενός διανύσματος περιοδικότητας είναι να εφαρμόσει απευθείας τον DFT στις ακολουθίες των χαρακτηριστικών. Προκύπτει όμως το εξής ερώτημα. Θα αντιστοιχούσε κάθε συχνότητα του DFT σε κάποια ρυθμική συχνότητα; Και ποια θα πρέπει να είναι η συχνότητα δειγματοληψίας στις συχνότητες ώστε να επιτύχουμε επαρκή διακριτική ανάλυση (resolution) ώστε να αναπαραστήσουμε όλες τις συχνότητες που μας ενδιαφέρουν; Ας θεωρήσουμε το ακόλουθο παράδειγμα. Έστω $x^i[m]$ μια συνάρτηση έμφασης ενός σήματος σε συχνότητα δειγματοληψίας 200 Hz. Επιθυμούμε την ρυθμική ανάλυση σε παράθυρα μήκους 6 s. Αυτή η επιλογή έχει να κάνει με δύο ανταγωνιστικά φαινόμενα που συναντούμε πάντα στις αναλύσεις χρόνου συχνότητας. Χρειαζόμαστε όσο το δυνατόν μεγαλύτερο παράθυρο για καλύτερη ανάλυση συχνότητας, και όσο το δυνατόν μικρότερο παράθυρο για καλύτερη δυνατή χρονική ανάλυση, ώστε να παρατηρούμε τις μεταβολές του τέμπο. Η επιλογή των 6 s αποτελεί μια τυπική επιλογή που επαρκεί για το συγκεκριμένο παράδειγμα. Οι συχνότητες του DFT σε παράθυρο με μήκος $N = 6 \cdot 200 = 1200$ samples θα είναι

$$f_k = f_s \cdot \frac{k}{N},\ k = 1..\frac{N}{2} \tag{2.17}$$

Η διακριτική ανάλυση συχνότητας τότε θα είναι $f_s/N = 0.167$ Hz και το εύρος των συχνοτήτων [0.167, 100] Hz. Το επιθυμητό εύρος ανάλυσης είναι 30 BPM έως 300 BPM που αντιστοιχεί στο εύρος [0.5, 5] Hz, επομένως με τον DFT υπολογίζουμε συχνότητες εκτός του ρυθμικού εύρους. Αν περιορίζαμε τον DFT στις συχνότητες που μας ενδιαφέρουν, αυτό θα ήταν ισοδύναμο με το να υπολογίζαμε το εσωτερικό γινόμενο του σήματος με τις συναρτήσεις βάσεως των μιγαδικών συνημίτονων του μετασχηματισμού για αυτές τις συχνότητες.



Σχήμα 2.12. Απόκριση του ταλαντωτή ως προς τον χρόνο για γ = 0.5, 5, και 500.

Το ερώτημα είναι κατά πόσο είναι επαρκείς αυτές οι συναρτήσεις. Μήπως υπάρχουν συναρτήσεις πέραν των μιγαδικών συνημίτονων που θα περιγράφουν καλύτερα το ρυθμικό περιεχόμενο; Επειδή δεν μας ενδιαφέρει η επιστροφή στον χώρο του χρόνου από τον χώρο της συχνότητας, δεν είναι ανάγκη αυτές οι νέες συναρτήσεις να είναι ορθογώνιες. Επόμενη σκέψη είναι ότι και το εσωτερικό γινόμενο θα μπορούσε να αντικατασταθεί από κάποια άλλη, πιο διαισθητική πράξη, όπως για παράδειγμα τη συνέλιξη.

Βάσει του παραπάνω συλλογισμού, προκύπτει η ανάλυση του σήματος από μια συστοιχία φίλτρων. Μια πρώτη τέτοια προσέγγιση, ήταν η υιοθέτηση συστοιχίας φίλτρων που έχουν κρουστική απόκριση περιοδικές συναρτήσεις Dirac, όπως αναφέρεται στη βιβλιογραφία [Scheirer1998, Klapuri2006]. Στη παρούσα εργασία, εμπνευσμένοι από τον ταλαντωτή που παρουσίασαν οι Large και Kolen [Large1994], υιοθετήσαμε μια συστοιχία φίλτρων με δομικά στοιχεία αυτόν τον ταλαντωτή στις συχνότητες - τέμπο στόχους. Η κρουστική απόκριση αυτού του ταλαντωτή ορίζεται ως

$$o_T[m] = \left[1 + \tanh\left(\gamma(\cos(2\pi f_T m) - 1)\right)\right][u(m - Q_0\tau_T) - u(m)]. \qquad (2.18)$$

Τα f_T , τ_T είναι η συχνότητα και η περίοδος που αντιστοιχούν στο τέμπο T, ενώ η παράμετρος γ ονομάζεται κέρδος εξόδου (output gain). Q_0 είναι ο αριθμός των κύκλων που επιτυγχάνει η κρουστική απόκριση του ταλαντωτή και $u(\cdot)$ είναι η συνάρτηση βήματος. Λόγω του ότι η παράγωγος της υπερβολικής εφαπτομένης πλησιάζει τη μονάδα σε περιοχή του μηδέν, όσο το γ μικραίνει η $o_T[m]$ προσεγγίζει ένα υπερυψωμένο ημιτονοειδές, ενώ όσο μεγαλώνει προσεγγίζει συνάρτηση ισαπέχοντων παλμών Dirac (Σχήμα 2.12). Επιπλέον όσο η συχνότητα ταλάντωσης αυξάνει, το εύρος των αποκρίσεων μειώνεται ώστε οι κρουστικές αποκρίσεις όλων των φίλτρων να έχουν την ίδια ενέργεια ανά χρονικό διάστημα. Έτσι αποφεύγεται η τάση πόλωσης (bias) στα διάφορα τέμπο. Η τιμή της συνάρτησης περιοδικότητας $v[T]$ για το τέμπο T και τη συνάρτηση έμφασης $x^i[m]$ ισούται με την μέγιστη τιμή της απόκρισης του ταλαντωτή $o_T[\cdot]$ με είσοδο το$x^i[m]$:

$$v[T] = max\{x^i[m] * o_T[m]\}. \qquad (2.19)$$

Επειδή τα $o_T[m]$ είναι συμμετρικά ως προς τον χρόνο ($o_T[m] = o_T[M-m]$) όπου $1..M$ το πεδίο ορισμού του $o_T m$, η συνέλιξη είναι ίση με την συσχέτιση. Μια εναλλακτική ερμηνεία της μέγιστης τιμής εξόδου του ταλαντωτή, είναι το καλύτερο "ταίριασμα" του ταλαντωτή με το σήμα εισόδου, δηλαδή η μέγιστη τιμή της συσχέτισης των δύο σημάτων.

- Για να δείξουμε το πλεονέκτημα των φίλτρων αυτών, υιοθετούμε ένα κύκλο πειραμάτων με κάποια πρότυπα σήματα εισόδου για μια σειρά τεχνικών ανάλυσης. Συγκεκριμένα χρησιμοποιούμε τα παρακάτω (μονοδιάστατα) σήματα εισόδου:Ημίτονο σταθερής συχνότητας
- Τριγωνικό παλμό σταθερής συχνότητας
- Τετραγωνικό παλμό σταθερής συχνότητας
- Ακολουθία ισαπέχουσων dirac
- Ένα μουσικό σήμα πραγματικών ντραμς σταθερού ρυθμού
- Χτύπημα ογδόων τυμπάνου με τονισμό στο τέταρτο

Για τα παραπάνω σήματα αξιολογούμε την συνάρτηση περιοδικότητας των ακόλουθων μεθόδων.

- Συνάρτηση Αυτοσυσχέτισης
- Εσωτερικό γινόμενο με μιγαδικά εκθετικών συχνοτήτων
- Συνέλιξη με τράπεζα dirac φίλτρων
- Συνέλιξη με την τράπεζα ταλαντωτών (Εξ. 2.18)

Όλα τα παραπάνω σήματα / φίλτρα έχουν συχνότητα δειγματοληψίας 200 Hz, τέμπο 120 BPM και μήκος 8 s. Στο Σχήμα 2.13 (α) έως (στ) φαίνεται η συνάρτηση περιοδικότητας για τα έξι πρότυπα εισόδου και για τις τέσσερις συναρτήσεις περιοδικότητας που αναφέραμε.

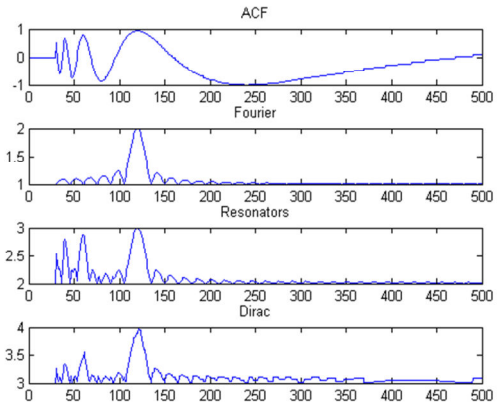
Η ανάλυση περιοδικότητας με την συνάρτηση αυτοσυσχέτισης δείχνει ότι η κύρια περίοδος, δηλαδή το τέμπο της διέγερσης (120 BPM), εμφανίζει την μεγαλύτερη κορυφή στο διάνυσμα της περιοδικότητας. Επιπλέον παρατηρούμε ότι εμφανίζονται και κορυφές σε υποπολλαπλάσια αυτού, δηλαδή στα 60, 40, και 30 BPM, με φθίνων πλάτος. Κάτι τέτοιο είναι θεμιτό, αφού όλα τα πρότυπα σήματα προσδίδουν την ρυθμική αίσθηση και αυτών των μουσικών τέμπο. Για παράδειγμα ένα κτύπος στα 120 BPM μπορεί να δώσει την ρυθμική αίσθηση των 30 BPM αν οι κτύποι ομαδοποιηθούν ανά 4, ή την αίσθηση των 40 BPM αν ομαδοποιηθούν ανά 3. Χαρακτηριστικό παράδειγμα είναι η αντίληψη των κτύπων ενός ρολογιού. Οι περισσότεροι άνθρωποι ομαδοποιούν τον ήχο του ρολογιού ανά δύο κτύπους, προκαλώντας έτσι την αυταπάτη του «τικ-τακ» παρόλο που όλοι οι κτύποι είναι ίδιοι. Κάτι αντίστοιχο ισχύει και για τα πολλαπλάσια του 120 BPM. Ένας κτύπος στα 120 BPM μπορεί να θεωρηθεί ένας κτύπος στα 240 BPM όπου οι μισοί κτύποι απουσιάζουν. Στην περίπτωση όμως της ACF ως ΣΠ δεν παρατηρούμε κάτι τέτοιο για τα πολλαπλάσια του 120 BPM, κάτι που θα ήταν επιθυμητό. Επίσης πρέπει να αναφερθεί ότι λόγω του ότι η αυτοσυσχέτιση υπολογίζεται σε ισοκατανεμημένες καθυστερήσεις στο πεδίο του χρόνου, οι αντίστοιχες περιοδικότητες θα είναι κατανεμημένες λογαριθμικά. Λόγω της αραιής κατανομής των μεγάλων τιμών τέμπο, παρατηρούμε πολύ πλατιές κορυφές με πλάτος ανάλογο της συχνότητας.

Στην περίπτωση της συνέλιξης με ταλαντωτές με κρουστική απόκριση ακολουθία Dirac, παρατηρούμε ότι εμφανίζονται κορυφές σε πολλαπλάσια του μουσικού τέμπο 120 BPM. Ωστόσο σε πολλές περιπτώσεις αυτά τα πολλαπλάσια είναι ισχυρότερα από την κεντρική συχνότητα (βλ. Σχ. 2.13(γ). Επιπλέον παρατηρούμε ότι τα σχετιζόμενα με τον ρυθμό υποπολλαπλάσια της κεντρικής συχνότητας (30, 40, 60 BPM) είναι ασθενέστερα σε σχέση με την συνάρτηση αυτοσυσχέτισης. Επιπλέον στις πολύ μεγάλες τιμές τέμπο έχουμε σφάλματα κβαντισμού της περιόδου του παλμού.

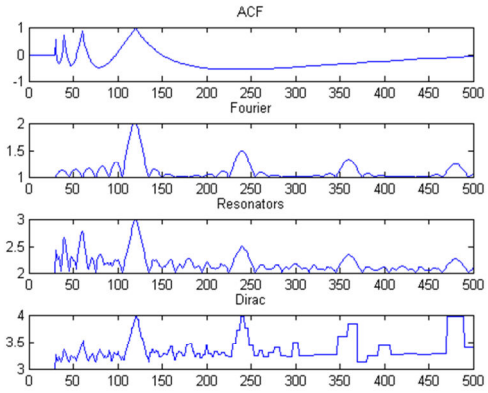
Η ανάλυση περιοδικότητας με την υιοθέτηση των προτεινόμενων ταλαντωτών παρουσιάζει αρκετές ομοιότητες με αυτή της χρήσης απευθείας του μετασχηματισμού Fourier (όπως π.χ. στις περιπτώσεις τριγωνικού, τετραγωνικού παλμού). Ωστόσο η διαδικασία συνέλιξης με τους ταλαντωτές αναδεικνύει καλύτερα τις ρυθμικές μετρικές σχέσεις αφού παρατηρούμε κορυφές τόσο στα πολλαπλάσια όσο και στα υποπολλαπλάσια του κεντρικού τέμπο. Αυτό αντίθετα δεν συμβαίνει στην περίπτωση του μετασχηματισμού Fourier. Σε κάποιες περιπτώσεις υπάρχει έμφαση στα πολλαπλάσια την κεντρικής συχνότητας, όπως για παράδειγμα στον τετραγωνικό παλμό και στην ακολουθία Dirac. Ωστόσο οι κορυφές αυτές όμως δεν οφείλονται απαραίτητα στην ανάλυση του ρυθμικού περιεχομένου του σήματος, αλλά στην ίδια την φύση του μετασχηματισμού. Για παράδειγμα, οι κορυφές που παρουσιάζονται στην περίπτωση του τετραγωνικού παλμού οφείλονται στην ανάλυση σε σειρά Fourier του τετραγωνικού παλμού, ο οποίος αναλύεται στις περιττές αρμονικές της κεντρικής συχνότητας. Αυτό είναι ακόμα περισσότερο έκδηλο στην περίπτωση της ημιτονοειδούς διέγερσης, στην οποία εμφανίζεται μονάχα μία προεξέχουσα κορυφή. Αυτή η ανεπιθύμητη ιδιότητα στην ρυθμική ανάλυση μουσικών σημάτων οφείλεται στην ορθογωνιότητα των αρμονικών συναρτήσεων του μετασχηματισμού Fourier. Επιπλέον στη περίπτωση ανάλυσης περιοδικότητας με τον μετασχηματισμό Fourier δεν παρατηρούμε καθόλου κορυφές σε υποπολλαπλάσια του κεντρικού τέμπο.

Αντίθετα, η συνέλιξη με τους ταλαντωτές φαίνεται περισσότερο αποτελεσματική σε όλες τις περιπτώσεις, τόσο των τεχνητών σημάτων (Σχ. 2.13 (α)-(δ)) τόσο και των πραγματικών. Στην περίπτωση (ε) βλέπουμε ότι η υψηλότερη κορυφή παρατηρείται στα 120 BPM, ενώ εξέχουσες κορυφές παρατηρούνται στα 60 και στα 240 BPM, ενώ αντίθετα στην περίπτωση του Fourier η πιο εξέχουσα κορυφή παρατηρείται στα 240 BPM. Ακόμα πιο έντονο είναι το φαινόμενο αυτό στην περίπτωση (στ), όπου ο Fourier μετασχηματισμός "κλειδώνει" στην γρηγορότερη συχνότητα χτυπήματος των όγδοων, σε αντίθεση με την τράπεζα ταλαντωτών όπου οι κορυφές στις συχνότητες τετάρτου/ογδόου είναι εξίσου εξέχουσες.

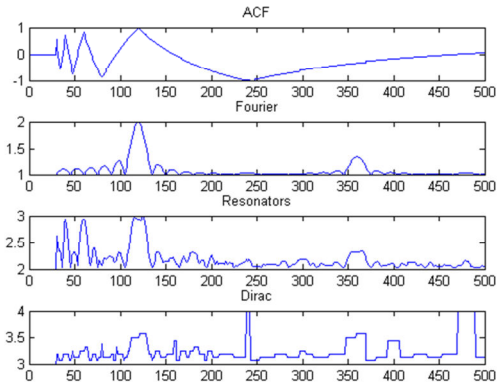
Φαίνεται λοιπόν ότι η διαδικασία συνέλιξης του σήματος έμφασης με τους προτεινόμενους ταλαντωτές έχει συγκεκριμένα πλεονεκτήματα σε σχέση με τις υπόλοιπες μεθόδους. Ένα από τα βασικά πλεονεκτήματα είναι η αναπαράσταση μετρικών σχέσεων στο μουσικό σήμα, οι οποίες παρουσιάζονται ως κορυφές των πολλαπλασίων και των υποπολλαπλάσιων του κεντρικού τέμπο στη ΣΠ. Αυτό το πλεονέκτημα επαληθεύεται και από τα πειραματικά αποτελέσματα που θα παρουσιαστούν στα Κεφάλαια 4 και 5.



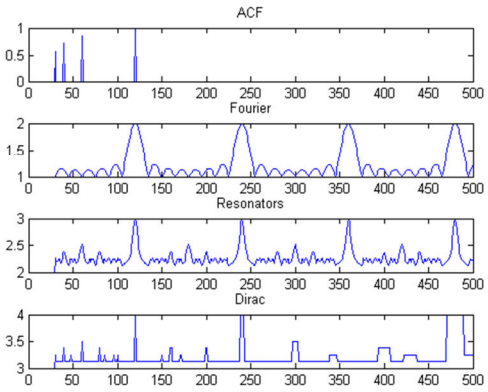
α) Ημίτονο



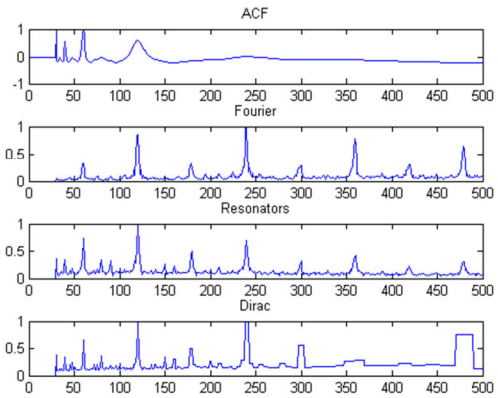
β) Τριγωνικός Παλμός



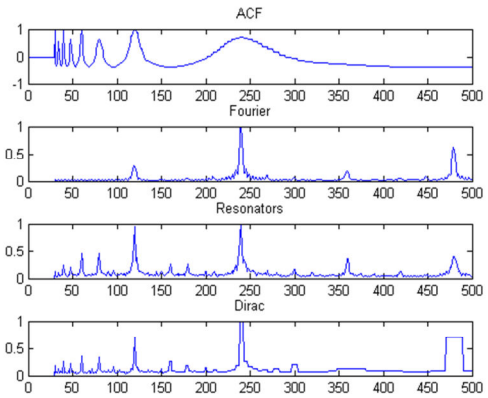
γ) Τετραγωνικός Παλμός



δ) Ακολουθία Dirac



ε) Ντραμς 4/4



στ) Όγδοα τυμπάνου με τονισμό στο 4°

Σχήμα 2.13. Ανάλυση περιοδικότητας για 6 σήματα εισόδου και τις 4 συγκρινόμενες μεθόδους (ACF: autocorrelation function, Fourier, Resonators, Dirac).

2.6 Η front-end επεξεργασία

Στην προηγούμενη ενότητα αναλύσαμε διάφορες μεθόδους ανάλυσης περιοδικότητας, και αναδείξαμε τα πλεονεκτήματα της προσέγγισης που χρησιμοποιεί την τράπεζα των ταλαντωτών. Η ανάλυση της προηγούμενης παραγράφου περιορίστηκε σε πρότυπα συνθετικά σήματα, τα οποία θεωρήθηκαν ως συνάρτηση έμφασης μίας διάστασης. Επιπλέον, το μήκος των σημάτων εισόδου ήταν μικρό (8s) και δεν υποβλήθηκε σε κατάτμηση. Σε πρακτικό επίπεδο όμως τα μουσικά σήματα είναι μεγαλύτερης διάρκειας και μπορεί να παρουσιάζονται μικρές ή μεγαλύτερες μεταβολές του ρυθμικού περιεχομένου. Ένα σύστημα αυτόματης ανάλυσης ρυθμού, όπως συμβαίνει με κάθε σύστημα ανάλυσης χρόνου-συχνότητας, υπόκειται στην απροσδιοριστία χρόνου – συχνότητας (time-frequency uncertainty principle). Για παράδειγμα, δεν μπορούμε να υπολογίσουμε όλα τα ρυθμικά χαρακτηριστικά σε αυθαίρετα μικρό παράθυρο. Για να βρούμε κάποιο μουσικό τέμπο, χρειαζόμαστε κάποια ελάχιστη χρονική διάρκεια ανάλυσης του μουσικού αποσπάσματος. Το ίδιο συμβαίνει και με τις μεταβολές του τέμπο. Η μεταπήδηση από ένα (γρήγορο) τέμπο σε ένα άλλο (γρήγορο) τέμπο χρειάζεται λιγότερο χρόνο παρατήρησης σε σχέση με μια μετάβαση σε ένα αργό τέμπο. (βλ. Ενότητα 1.3). Επομένως απαιτείται κάποιος ειδικός χειρισμός των σημάτων εισόδου και της συνάρτησης περιοδικότητας ώστε να περιγράφει το ρυθμικό περιεχόμενο ενός μουσικού κομματιού και κατόπιν να εξαχθεί το μουσικό τέμπο και ο μουσικός παλμός.

Η παράγραφος αυτή εξετάζει τη διαδικασία από την οποία θα προκύψει η τελική συνάρτηση περιοδικότητας χρησιμοποιώντας την ανάλυση που περιγράφτηκε στην Ενότητα 2.5. σε πραγματικά μουσικά αποσπάσματα από τα οποία έχει εξαχθεί η πολυδιάστατη συνάρτηση έμφασης που περιγράφηκε στις Ενότητες 2.2-2.4. Περιλαμβάνει τα εξής βήματα:

- Συνέλιξη των συναρτήσεων έμφασης με τους ταλαντωτές
- Κατάτμηση (segmentation) των εξόδων των ταλαντωτών
- Εξαγωγή συνάρτησης περιοδικότητας για κάθε χαρακτηριστικό ανά τμήμα ανάλυσης
- Συνδυασμός των συναρτήσεων περιοδικότητας
- Κανονικοποίηση

Επιπλέον αναλύονται και επιπρόσθετες ιδιότητες της προτεινόμενης μεθόδου που δεν έχουν να κάνουν με την φύση των ταλαντωτών, αλλά με την διαδικασία κατάτμησης και κανονικοποίησης.

Στο Σχήμα 2.14 παρουσιάζεται η διαδικασία εξαγωγής της τελικής συνάρτησης περιοδικότητας από τα χαρακτηριστικά έμφασης. Έστω $\mathbf{x}_e^i = \mathrm{x}_e^i[m]$, $\mathbf{x}_{ch}^j = \mathrm{x}_{ch}^j[m]$, τα δύο πολυδιάστατα σήματα έμφασης που έχουν εξαχθεί από τα χαρακτηριστικά ενέργειας και χρώματος αντίστοιχα (Παρ. 2.4). Τα σήματα αυτά όπως προαναφέρθηκε παραγωγίζονται με ένα μη-αιτιατό φίλτρο παραγώγισης (Εξ. 2.13) τάξης (order) L. Η παραγώγιση μετά από λογαρίθμηση των χαρακτηριστικών έμφασης είναι σε συμφωνία με τη βιβλιογραφία [Klapuri1999] και περιγράφει τις μεταβολές του σήματος ως προς το επίπεδο έντασής του. Στη συνέχεια υπολογίζεται η συνέλιξη καθενός από τα $\mathbf{x} \in \{\mathbf{x}_{ch}^i \cup \mathbf{x}_e^j\}$ με τους ταλαντωτές (Εξ. 2.18). Η παράμετρος Q_0 καθορίζει το μήκος των μη μηδενικών τιμών της κρουστικής απόκρισης του ταλαντωτή, το οποίο είναι ανάλογο της περιόδου συντονισμού. Στη συνέχεια, για κάθε υποψήφιο τέμπο $T \in [T_{min}, T_{max}]$,

όπου T_{min}, T_{max} το ελάχιστο και μέγιστο τέμπο ανάλυσης αντίστοιχα, και για κάθε χαρακτηριστικό έμφασης $\mathbf{x} \in \{\mathbf{x}_{ch}^{i} \cup \mathbf{x}_{e}^{j}\}$, υπολογίζουμε την έξοδο του αντίστοιχου ταλαντωτή, την οποία συμβολίζουμε με $r_{\mathbf{x}}[n,T]$. Στην τράπεζα ταλαντωτών, χρησιμοποιούμε ένα ταλαντωτή για κάθε ακέραιη τιμή των τέμπο στόχων, δηλαδή $T = T_{min} + k, k = 0 \dots T_{max} - T_{min}$. Επομένως το πρώτο βήμα για την εξαγωγή της τελικής συνάρτησης περιοδικότητας είναι η συνέλιξη κάθε συνιστώσας με την τράπεζα των ταλαντωτών, όπως φαίνεται από την ακόλουθη εξίσωση:

$$r_{\mathbf{x}}[m,T] = (\mathbf{x} * \mathbf{o}_T)[m] \tag{2.19}$$

όπου $\mathbf{x} \in \{\mathbf{x}_{ch}^{i} \cup \mathbf{x}_{e}^{k}\}$ το διάνυσμα έμφασης. Στη συνέχεια, τα $r_{\mathbf{x}}[m,T]$ χωρίζονται σε τμήματα (segments) χρησιμοποιώντας τετραγωνικό παράθυρο του οποίο το μήκος και η ολίσθηση είναι ανάλογα της περιόδου ταλάντωσης τ_T του κάθε ταλαντωτή. Επομένως τα $r_{\mathbf{x}}[n,T]$ κατατέμνονται με διαφορετικό μήκος παραθύρου για κάθε τέμπο T. Τα προκύπτοντα τμήματα τα συμβολίζουμε με $r_x^{s,T}[m]$, όπου $s = 1..S_T$ είναι ο δείκτης του τμήματος, m είναι ο δείκτης του πλαισίου (frame) μέσα στο τμήμα s και S_T είναι το πλήθος των τμημάτων για το τέμπο T. Το S_T εξαρτάται από το T λόγω του μεταβλητού μήκους παραθύρου. Η ισχύς του τέμπο T για το τμήμα s και την συνάρτηση έμφασης $\mathbf{x}$ ισούται το μέγιστο της απόλυτης τιμής της απόκρισης $r_x^{s,T}[m]$ σε όλο το τμήμα

$$v_{\mathbf{x}}[T,s] = \max_m\left(\left|r_x^{s,T}[m]\right|\right), T \in [T_{\min,} T_{\max,}]. \tag{2.20}$$

Με την παραδοχή «σταθερού» ρυθμού για ένα μουσικό απόσπασμα, η τελική συνάρτηση περιοδικότητας $v_{\mathbf{x}}[T]$ μπορεί να υπολογιστεί ως ο μέσος όρος των $v_{\mathbf{x}}[T,s]$ ως προς τα τμήματα s.

Το γεγονός ότι η κρουστική απόκριση των ταλαντωτών έχει σταθερό μήκος κύκλων (είναι ανάλογο της περιόδου του κατά Q_0), προσδίδει στη συγκεκριμένη ανάλυση περιοδικότητας μια πολύ επιθυμητή ιδιότητα να είναι αναλλοίωτη σε χρονικές κλιμακώσεις.

Για δύο σήματα τα οποία είναι κλιμακούμενες εκδόσεις της ίδιας κυματομορφής, η ανάλυση περιοδικότητας θα δώσει ακριβώς την ίδια συνάρτηση περιοδικότητας, κλιμακούμενη κατά τον ίδιο λόγο. Δηλαδή

$$v_{x(at)}[T,s] = v_{x(t)}[a^{-1}T,s] \leftrightarrow v_{x(at)}[aT,s] = v_{x(t)}[T,s]. \tag{2.21}$$

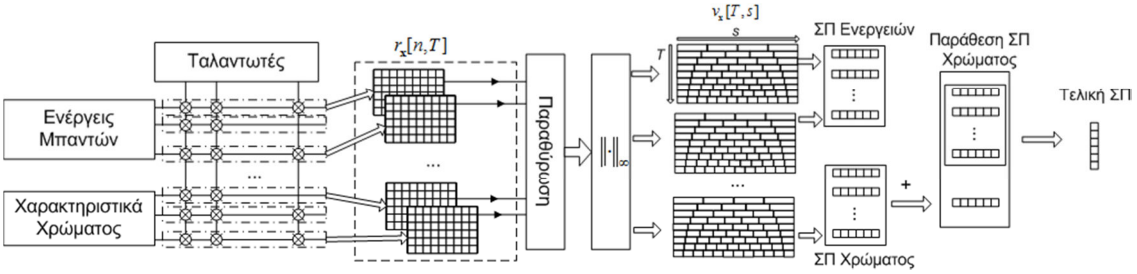
Η ιδιότητα (2.21) είναι εύκολο να αποδειχθεί για πραγματικά σήματα. Πράγματι, έστω τα σήματα εισόδου $x(t), x(at)$ με $a > 1$, δηλαδή το δεύτερο σήμα εισόδου μια πιο «γρήγορη» εκδοχή του πρώτου. Τότε για κάποιο παράθυρο ανάλυσης $[t_{o,}t_1]$, με δείκτη s, ο $1^{ος}$ όρος της (2.21) γράφεται για το τέμπο T,

$$v_{x(at)}[T,s] = \max_{t\in[t_{o,}t_1]}\{|x(at) * o_T(t)|\}. \tag{2.22}$$

Για τον δεύτερο όρο της (2.21) έχουμε

$$v_{x(t)}[a^{-1}T,s] = \max_{t\in[at_{o,}at_1]}\{|x(t) * o_{a^{-1}T}(t)|\}. \tag{2.23}$$

Η αλλαγή των ορίων $t \in [at_{o,}at_1]$ στον υπολογισμό του μεγίστου, συμβαίνει επειδή επιθυμούμε κατά την κατάτμηση της εξόδου των ταλαντωτών $r_T(t) = x(at) * o_T(t)$, το μήκος των τμημάτων να είναι ανάλογο της περιόδου του τέμπο.



Σχήμα 2.14: Γραφική αναπαράσταση των βημάτων επεξεργασίας της ανάλυσης περιοδικότητας

Έτσι αν για το τέμπο T και για το τμήμα (segment) s γίνεται παραθύρωση στο διάστημα $[t_0, t_1]$, για το πιο αργό τέμπο $a^{-1}T$ η παραθύρωση για το τμήμα s θα πρέπει να γίνει σε μεγαλύτερο διάστημα $[at_0, at_1]$.

Επειδή $o_{a^{-1}T}(t) = o_T(a^{-1}t)$, η Εξ. 2.23 γράφεται

$$v_{x(t)}[a^{-1}T, s] = \max_{t\in[at_0, at_1]}\{|x(t) * o_T(a^{-1}t)|\}. \quad (2.24)$$

Με αλλαγή μεταβλητής $t' = a^{-1}t$ έχουμε

$$v_{x(t)}[a^{-1}T, s] = \max_{t'\in[t_0, t_1]}\{|x(at') * o_{a^{-1}T}(t')|\}. \quad (2.25)$$

Το να θέσουμε το μήκος του παραθύρου ανάλυσης για κάθε τέμπο ανάλογο της περιόδου που αντιστοιχεί στο τέμπο είναι απαραίτητη προϋπόθεση για να ισχύει η Εξ. 2.21. Με τον τρόπο αυτόν η ανάλυση περιοδικότητας αποτελεί ανάλυση του σήματος εισόδου με την ίδια ακριβώς διαδικασία σε διαφορετικές χρονικές κλίμακες, όπου κάθε κλίμακα αντιστοιχεί σε ένα μουσικό τέμπο. Στην περίπτωση που τα σήματα στις Εξ. 2.21-2.25 είναι διακριτά, η (2.21) ισχύει προσεγγιστικά, αφού υπεισέρχεται πάντα ένα σφάλμα που οφείλεται στον κβαντισμό λόγω της δειγματοληψίας. Ωστόσο το σφάλμα αυτό μικραίνει όσο η συχνότητα δειγματοληψίας αυξάνει και μπορούμε να θεωρήσουμε ότι η (2.21) «σχεδόν» ισχύει και για διακριτά σήματα.

Επιλέγοντας λοιπόν διαφορετικό παράθυρο κατάτμησης στην έξοδο κάθε ταλαντωτή, επιτυγχάνουμε την ισότητα της Εξ. 2.21. Ωστόσο από το μεταβλητό μήκος παραθύρου για κάθε τέμπο προκύπτει διαφορετικό πλήθος παραθύρων S_T και επομένως διαφορετική χρονική διακριτική ικανότητα. Για να γίνει η ταυτόχρονη επεξεργασία όλων των τέμπο για κάθε χρονική στιγμή, πρέπει να γίνει επαναδειγματοληψία των ΣΠ $v_{\mathbf{x}}[T, s]$ στην ίδια συχνότητα. Η επαναδειγματοληψία μπορεί να γίνει έμμεσα με κυβική παρεμβολή στις στήλες σε κάθε $v_{\mathbf{x}}[T, s]$ ως προς το s

$$\hat{v}_{\boldsymbol{x}}[T, s] = bicubic_s(v_{\mathbf{x}}[T, s]). \quad (2.26)$$

Το $\hat{v}_{\boldsymbol{x}}[T, s]$ μας δίνει μια εκτίμηση της συνάρτησης περιοδικότητας για κάθε χρονικό τμήμα (segment) s. Πέρα από την δυνατότητα ανεξάρτητης ρυθμικής ανάλυσης κάθε τμήματος χωριστά, αυτή η αναπαράσταση επιτρέπει την ανίχνευση αλλαγών στο ρυθμικό περιεχόμενο, ή και άλλων χαρακτηριστικών, όπως εκφραστικών αλλαγών του τέμπο. Ωστόσο, αν γίνει η παραδοχή του «σταθερού» ρυθμού, ότι δηλαδή το ρυθμικό περιεχόμενο δεν αλλάζει, μπορεί να θεωρηθεί η μέση τιμή της συνάρτησης περιοδικότητας ως προς τον χρόνο για

κάθε τέμπο. Σε αυτή την περίπτωση η επαναδειγματοληψία δεν είναι απαραίτητη.

Κατά τη διάρκεια πειραματισμών και υπολογισμών της ΣΠ παρατηρήθηκαν δύο ανεπιθύμητα φαινόμενα. Το πρώτο είναι ότι όσο μεγαλώνει η τάξη του φίλτρου παραγώγισης L (Εξ. 2.13), η ΣΠ πολώνεται προς τα αργά τέμπο και αντιστρόφως. Για να εξαλειφθεί το φαινόμενο αυτό, θα μπορούσαμε να παραγωγίσουμε κάθε συνάρτηση έμφασης με διαφορετικό L, και συγκεκριμένα με τέτοιο L ώστε $L(T) \propto \tau_T$. Αυτή η προσέγγιση όμως προϋποθέτει παραγώγιση κάθε συνάρτησης έμφασης για κάθε διαφορετικό τέμπο, το οποίο αυξάνει πολύ το υπολογιστικό κόστος. Ωστόσο αυτό μπορεί να ξεπεραστεί με την εκμετάλλευση της προσεταιριστικής ιδιότητας της συνέλιξης, γράφοντας την $r_{\mathbf{x}}[n,T]$ ως

$$r_{\mathbf{x}}[n,T] = \big((\mathbf{x} * h_{L(T)}) * o_T\big)[n] = \big(\mathbf{x} * (h_{L(T)} * o_T)\big)[n]. \tag{2.27}$$

Επομένως αντί για παραγώγιση των συναρτήσεων έμφασης, παραγωγίζουμε τις κρουστικές αποκρίσεις των ταλαντωτών. Η τάξη του κάθε φίλτρου παραγώγισης εξαρτάται από την εγγενή συχνότητα του κάθε ταλαντωτή.

Το δεύτερο ανεπιθύμητο φαινόμενο είναι η πόλωση της ΣΠ προς μικρότερα τέμπο η οποία οφείλεται στην διαφορετική νόρμα της κρουστικής απόκρισης κάθε ταλαντωτή. Από την ανισότητα Hölder για την συνέλιξη, δηλαδή $\|f * g\|_p \leq \|f\|_1 \|g\|_p, 1 \leq p \leq \infty$, και λόγω του γεγονότος ότι $\|h_L\|_1 = const, \forall L$, η (2.20) μπορεί να ξαναγραφτεί ως

$$\begin{gathered} v_{\mathbf{x}}[T,s] = \left\| r_x^{s,T}[m] \right\|_\infty = \left\| \mathbf{x} * h_{L(T)} * o_T \right\|_\infty \Rightarrow \\ v_{\mathbf{x}}[T,s] \leq \|\mathbf{x}\|_\infty \left\| h_{L(T)} * o_T \right\|_1 \leq \|\mathbf{x}\|_\infty \left\| h_{L(T)} \right\|_1 \|o_T\|_1 \leq c\|o_T\|_1 \end{gathered} \tag{2.28}$$

όπου c είναι κάποια σταθερά. Η Εξ. 2.28 υποδηλώνει ότι η συνάρτηση περιοδικότητας πρέπει να κανονικοποιηθεί με $\|o_T\|_1$, δηλ.

$$v_{\mathbf{x}}[T,s] \leftarrow v_{\mathbf{x}}[T,s] / \|o_T\|_1. \tag{2.29}$$

Το επόμενο βήμα είναι ο συνδυασμός των επιμέρους ΣΠ, οι οποίες προέρχονται από τις επιμέρους συναρτήσεις έμφασης. Ο συνδυασμός τους, όπως θα περιγραφτεί παρακάτω, μπορεί να γίνει είτε αυτές έχουν υπολογιστεί ανά τμήμα (Εξ. 2.26) είτε έχει υπολογιστεί ο μέσος όρος ως προς τον χρόνο. Μια προσέγγιση που διατηρούσε όσο το δυνατόν περισσότερη πληροφορία θα ήταν η παράθεση όλων των επιμέρους ΣΠ σε ένα διάνυσμα

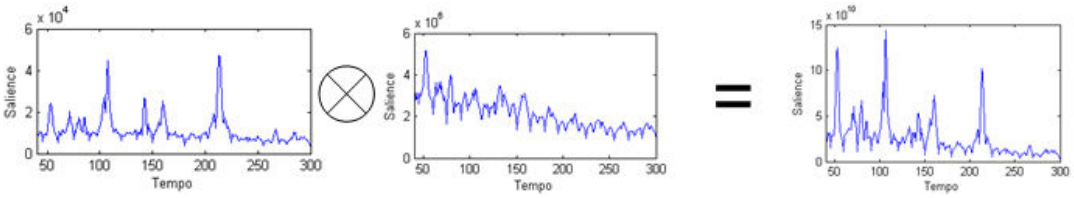
$$\mathbf{v} = \left[\mathbf{v}_e^1 \middle| \ldots \middle| \mathbf{v}_e^E \middle| \mathbf{v}_{ch}^1 \middle| \ldots \middle| \mathbf{v}_{ch}^{12}\right] \tag{2.31}$$

όπου Ε είναι το πλήθος των μπαντών κατά την εξαγωγή των χαρακτηριστικών ενέργειας.

Μία πρώτη σύμπτυξη που μπορεί να γίνει αφορά τις ΣΠ που προέρχονται από τα χαρακτηριστικά χρώματος. Λόγω του ότι τα χαρακτηριστικά χρώματος δεν περιέχουν συμπληρωματική ρυθμική πληροφορία, αυτά μπορεί να αθροιστούν σε ένα διάνυσμα περιοδικότητα ως:

$$v_{ch}[T] = \textstyle\sum_j v_{\mathbf{x}}[T],\ \mathbf{x} \in \{\mathbf{x}_{ch}^j\} \tag{2.31}$$

Πράγματι, ας θεωρήσουμε ένα κομμάτι το οποίο μεταφέρεται σε άλλη κλίμακα.



Σχήμα 2.15: Από αριστερά προς τα δεξιά: α) Η συνάρτηση περιοδικότητας από τα χαρακτηριστικά ενέργειας. β) Η συνάρτηση περιοδικότητας από τα χαρακτηριστικά χρώματος. γ) Η συνάρτηση περιοδικότητας που προκύπτει από κατά σημείο πολλαπλασιασμό. Το τέμπο του μουσικού κομματιού είναι 110 BPM.

Λόγω της μετάθεσης, οι επιμέρους ΣΠ $\mathbf{x}_{ch}^{i}$ θα είναι διαφορετικές, ωστόσο το άθροισμά τους θα πρέπει να είναι το ίδιο, καθότι το ρυθμικό περιεχόμενο δεν αλλάζει με μεταθέσεις κλίμακας. Κάτι αντίστοιχο δεν συμβαίνει και με τις ΣΠ που προέρχονται από τα χαρακτηριστικά ενέργειας. Οι ΣΠ που προέρχονται από χαμηλόσυχνες μπάντες εμφανίζουν μεγαλύτερα πλάτη στα αργά τέμπο και αντίστροφα. Επομένως μια εύρωστη αναπαράσταση του ρυθμικού περιεχομένου ενός σήματος θα χρησιμοποιούσε όλες τις ΣΠ των χαρακτηριστικών ενέργειας και το άθροισμα των ΣΠ χρώματος, δηλ.

$$\mathbf{v} = [\mathbf{v}_e^1 | \dots | \mathbf{v}_e^E | \mathbf{v}_{ch}]. \tag{2.32}$$

Μια περαιτέρω συμπαγής αναπαράσταση μπορεί να βασιστεί στην παραδοχή ότι μονάχα οι ακολουθίες των δύο χαρακτηριστικών (κρουστές / αρμονικές) περιέχουν συμπληρωματική πληροφορία. Τότε, οι δύο ακολουθίες αναλύονται και αθροίζονται ανεξάρτητα. Έτσι, όπως και την περίπτωση των χαρακτηριστικών χρώματος, αν θεωρήσουμε μια ΣΠ για τα χαρακτηριστικά ενέργειας

$$v_e[T] = \sum_i v_{\mathbf{x}}[T],\ \mathbf{x} \in \{\mathbf{x}_e^i\}. \tag{2.33}$$

Τελικά προκύπτουν δύο συναρτήσεις περιοδικότητας $v_{ch}[T]$ και $v_e[T]$ για τα χαρακτηριστικά χρώματος και ενέργειας αντίστοιχα. Τέλος, ένα συνολικό διάνυσμα περιοδικότητας μπορεί να εξαχθεί αν τα $v_{ch}[T]$ και $v_e[T]$ συνδυαστούν περεταίρω, με πολλαπλασιασμό κατά σημείο

$$v[T] = v_e[T] \cdot v_{ch}[T]. \tag{2.34}$$

Στο Σχήμα 2.15 φαίνονται οι συναρτήσεις περιοδικότητας για τα χαρακτηριστικά ενέργειας, χρώματος και ο συνδυασμός τους για ένα κομμάτι με τέμπο 110 BPM. Από τα χαρακτηριστικά ενέργειας παρατηρούμε δύο εξέχουσες κορυφές στα 110 και 220 BPM, ενώ για τα χαρακτηριστικά χρώματος παρατηρούμε την πιο εξέχουσα κορυφή στο μισό του πραγματικού τέμπο 55 BPM. Ο συνδυασμός των δύο διανυσμάτων περιοδικότητας υποδεικνύει το 110 BPMως το πιο εξέχων τέμπο. Αυτό που φαίνεται από το συγκεκριμένο παράδειγμα – και ισχύει σε πολλές από τις περιπτώσεις – είναι ότι η συνάρτηση περιοδικότητας για τα χαρακτηριστικά ενέργειας παρουσιάζει ισχυρά ακρότατα τα οποία εμφανίζονται στο σωστό τέμπο και σε πολλαπλάσια-υποπολλαπλάσια αυτού. Ωστόσο το ισχυρότερο ακρότατο δεν υποδεικνύει πάντα το σωστό τέμπο. Από την άλλη, οι συναρτήσεις περιοδικότητας των χαρακτηριστικών χρώματος έχουν λιγότερο εμφανείς κορυφές, αλλά δίνουν μια πιο πλατειά, και λιγότερο λεπτομερή εκτίμηση των τέμπο, ενώ υπάρχει μια τάση η εκτίμηση αυτή να πολώνεται προς μικρότερες τιμές. Έτσι, όταν οι δύο συναρτήσεις περιοδικότητας

συνδυάζονται, μπορούμε να πούμε ότι η συνάρτηση περιοδικότητας της ενέργειας υποδεικνύει τα ενδεχόμενα τέμπο και η συνάρτηση περιοδικότητας του χρώματος επιλέγει μία από αυτές τις κορυφές.

Κεφάλαιο 3: Εξαγωγή Χαρακτηριστικών από τη Συνάρτηση Περιοδικότητας

3.1 Εισαγωγή

Η συνάρτηση περιοδικότητας όπως περιγράφτηκε στο προηγούμενο Κεφάλαιο αποτελεί μία εύρωστη αναπαράσταση των ρυθμικών περιοδικοτήτων. Για μια σειρά από διαφορετικά χαρακτηριστικά του σήματος εισόδου, υπολογίζεται η ισχύς του σήματος για τις διάφορες περιοδικότητες. Ωστόσο μια τέτοια αναπαράσταση έχει το μειονέκτημα ότι έχει μεγάλη διάσταση. Ένα τυπικό παράδειγμα είναι η χρήση 8 μπαντών κατά την εξαγωγή των συναρτήσεων έμφασης ενέργειας, σε ένα εύρος ανάλυσης μεταξύ 25 και 300 BPM. Η συνολική ΣΠ περιοδικότητας (Εξ. 2.32) σε αυτή την περίπτωση θα είχε διάσταση (8 συναρτήσεις ενέργειας + 1 για τα χαρακτηριστικά χρώματος) (8+1)x276 = 2484. Στην περίπτωση μεταβαλλόμενου ρυθμικού περιεχομένου, όπου οι ΣΠ δεν μπορούν να αθροιστούν για όλα τα τμήματα, τότε το διάνυσμα μήκους 2484 θα αναπαριστά ένα μόνο τμήμα (segment) διάρκειας μερικών δευτερολέπτων.

Το γεγονός ότι η ΣΠ είναι πολύ υψηλής διάστασης είναι μια ανεπιθύμητη ιδιότητα, τόσο όσον αφορά θέματα αποθήκευσης της ρυθμικής πληροφορίας, όσο και σε επίπεδο χρησιμότητάς της. Στην περίπτωση που η ΣΠ χρησιμοποιηθεί ως είσοδος σε ένα σύστημα Μηχανικής Μάθησης (Machine Learning), όπως για παράδειγμα σε μία Μηχανή Υποστήριξης Διανυσμάτων (Support Vector Machine, SVM) (Ενότητα 4.2), το να διατηρηθεί η διάσταση της ΣΠ όσον το δυνατό μικρότερη είναι κρίσιμης σημασίας, λόγω της «Κατάρας της Υψηλής Διάστασης» (curse of dimensionality).

Εκτός από τα παραπάνω, μια υψηλής διάστασης ΣΠ μπορεί να περιέχει πλεονάζουσα (redundant) πληροφορία. Αν για παράδειγμα θέλουμε να βρούμε την μουσική ταχύτητα ενός κομματιού, η ΣΠ μπορεί να μην είναι κατάλληλη. Δύο ΣΠ που αντιστοιχούν σε κομμάτια με πολύ κοντινά τέμπο, μπορεί να έχουν αρκετά διαφορετικές ΣΠ, λόγω απότομων κορυφών στις ΣΠ.

Ο σκοπός αυτού του Κεφαλαίου είναι ο μετασχηματισμός της συνάρτησης περιοδικότητας σε ένα διάνυσμα μικρότερης διάστασης. Αυτό μπορεί αφενός να επιτύχει πιο «συμπαγείς» αναπαραστάσεις του ρυθμικού περιεχομένου, αφετέρου να περιορίσει τα εγγενή φαινόμενα των υψηλής διάστασης αναπαραστάσεων που προαναφέρθηκαν. Τέτοιου είδους μετασχηματισμοί συναντώνται συχνά στη βιβλιογραφία και ως εξαγωγή χαρακτηριστικών. Στη συνέχεια θα εξεταστούν δύο είδη εξαγωγής χαρακτηριστικών, Η πρώτη κατηγορία αποτελείται από «χειρονακτικά» χαρακτηριστικά, δηλαδή χαρακτηριστικά τα οποία σχεδιάζονται απευθείας από τον ερευνητή, αντανακλούν την ανθρώπινη διαίσθηση και οι παράμετροί τους υπολογίζονται με το χέρι. Η δεύτερη κατηγορία αποτελείται από μεθόδους που εξάγουν χαρακτηριστικά με αυτόματο τρόπο. Οι παράμετροι των μεθόδων αυτών υπολογίζονται συνήθως με τεχνικές εκμάθησης από τα δεδομένα.

3.2 Χειρονακτικά Χαρακτηριστικά

3.2.1 Κλιμάκωση της Συνάρτησης Περιοδικότητας

Λόγω του μικρού πλήθους των επισημειωμένων μουσικών κομματιών που είναι διαθέσιμα, υπάρχει η ανάγκη ενός αυτόματου τρόπου δημιουργίας επιπλέον τεχνητών παραδειγμάτων. Αυτό είναι εξαιρετικής σημασίας για την εφαρμογή μεθόδων Μηχανικής Μάθησης στην αυτόματη ανάλυση ρυθμού, όπου είναι απαραίτητο ένα σχετικά μεγάλο πλήθος παραδειγμάτων εκμάθησης. Μια προσέγγιση μπορεί να είναι η χρονική κλιμάκωση των μουσικών σημάτων. Δημιουργώντας τεχνητές αλλαγές στην ταχύτητα ενός μουσικού κομματιού μπορούμε να δημιουργήσουμε νέα δεδομένα. Εκμεταλλευόμενοι την ιδιότητα της Εξ. 2.21, αντί να γίνει η κλιμάκωση απευθείας στο σήμα, μπορεί να γίνει στο διάνυσμα περιοδικότητας, το οποίο έχει πολύ μικρότερο υπολογιστικό κόστος. Για κάθε διάνυσμα εκπαίδευσης $v[T]$, δημιουργούμε κλιμακώσεις αυτού σε ένα εύρος τιμών $[\underline{a}, \overline{a}]$ με βήμα δa . Η κλιμάκωση $S_a(\cdot)$ κατά τον λόγο a πραγματοποιείται με γραμμική παρεμβολή:

$$v_a[T] \xleftarrow{\text{κλιμάκωση}} S_a(v[T]) = bilinear_a(v[T]). \tag{3.1}$$

Προκειμένου να έχουν όλα τα διανύσματα περιοδικότητας που προκύπτουν από την κλιμάκωση το ίδιο μήκος, τα νέα διανύσματα περικόπτονται στο εύρος $\underline{a} \cdot T_{max}$:

$$v_a[T] \leftarrow v_a[T],\ T \in [T_{min} \dots \underline{a} \cdot T_{max}] \tag{3.2}$$

όπου T_{max} είναι το μέγιστο τέμπο ανάλυσης. Έτσι η αρχική συλλογή $\{v^k[T]\}_k, k = 1 \dots K$ έχει επεκταθεί σε ένα μεγαλύτερο υπερσύνολο

$$\left\{v_{a_i}^k[T]\right\}_{k,l},\ k = 1 \dots K,\ a_l = \underline{a} + (l-1)\delta a,\ l = 1 \dots \left\lfloor (\overline{a} - \underline{a})/\delta a \right\rfloor \tag{3.3}$$

όπου $\lfloor \cdot \rfloor$ είναι η κάτω στρογγυλοποίηση σε ακέραιο. Ανάλογα με το είδος της επισημείωσης της ρυθμικής κατηγορίας για κάποιο μουσικό κομμάτι αλλάζει αντίστοιχα και η ετικέτα. Αν έχουμε την πληροφορία ότι ένα μουσικό κομμάτι έχει τέμπο T, τότε μια κλιμακούμενη έκδοση αυτού κατά έναν παράγονται a θα έχει τέμπο aT . Παρόμοια τεχνική εφαρμόστηκε για πρώτη φορά στο [Eronen2010]. Κάθε νέο διάνυσμα επισημειώνεται με τέμπο aT. Αντίθετα με το τέμπο, η κλιμάκωση του διανύσματος περιοδικότητας δεν αλλάζει άλλα ρυθμικά χαρακτηριστικά, όπως για παράδειγμα το μουσικό μέτρο. Ένα μουσικό κομμάτι με χρονικό κλειδί 3/4 θα έχει το ίδιο χρονικό κλειδί όσο πιο γρήγορα ή αργά κι αν εκτελεστεί. Υπάρχουν όμως περιπτώσεις όπου η κλιμάκωση της ΣΠ θα πρέπει να γίνεται με προσοχή. Στη συλλογή Ballroom για παράδειγμα [Gouyon2006], τα μουσικά αποσπάσματα είναι επισημειωμένα με μία ρυθμική κατηγορία. Η επισημείωση αυτή συσχετίζεται πολύ με το τέμπο. Για παράδειγμα, υπάρχουν δύο κατηγορίες Βαλς στη συγκεκριμένη συλλογή: Τα κανονικά Βαλς, που παρουσιάζουν ένα εύρος τέμπο (~100 BPM) και τα Βιεννέζικα Βαλς, τα οποία είναι πολύ πιο γρήγορα (~180 BPM). Και οι δύο κατηγορίες χαρακτηρίζονται από το μέτρο 3/4. Η κλιμάκωση ενός κανονικού Βαλς κατά 50%, θα οδηγούσε σε ένα τεχνητό παράδειγμα, του οποίο το μέτρο θα ήταν 3/4, και το τέμπο περίπου 150 BPM. Σε αυτή την περίπτωση δεν είναι ξεκάθαρο αν το προκύπτων τεχνητό σήμα θα έπρεπε να ταξινομηθεί ως Βιεννέζικο ή κανονικό Βαλς.

3.2.2 Κωδικοποίηση της Συνάρτησης Περιοδικότητας

Όπως αναφέρθηκε, ενώ οι συναρτήσεις περιοδικότητας μας δίνουν μια καλή εκτίμηση των εξεχουσών ρυθμικών συχνοτήτων, κάποια χαρακτηριστικά όπως η ταχύτητα του μουσικού κομματιού δεν εξαρτάται τόσο από τη θέση αυτών των κορυφών όσο από τη γενική περιβάλλουσα (envelope) της συνάρτησης περιοδικότητας. Επιπλέον, οι αιχμηρές συναρτήσεις περιοδικότητας δεν είναι κατάλληλες για την τροφοδότηση μιας μηχανής εκμάθησης. Δύο μουσικά κομμάτια που έχουν κοντινή ταχύτητα ή ανήκουν στην ίδια ρυθμική κλάση μπορεί να εμφανίζουν κορυφές σε τελείως διαφορετικά σημεία, ή μπορεί να έχουν πολύ παρόμοιο σχήμα, αλλά οι κορυφές τους να μην συμπίπτουν (όπως θα συνέβαινε σε ένα διάνυσμα και σε μία κλιμακούμενη έκδοσή του). Και στις δύο περιπτώσεις, η Ευκλείδεια απόστασή τους θα ήταν σχετικά μεγάλη, κάτι που δεν είναι θεμιτό αφού ανήκουν στην ίδια ρυθμική κατηγορία. Για καλύτερη κατανόηση των παραπάνω, μπορούμε να θέσουμε σε αναλογία το διάνυσμα περιοδικότητας με τον μετασχηματισμό Fourier ενός σήματος. Ο μετασχηματισμός Fourier δεν είναι κατάλληλος για την απευθείας τροφοδότηση μηχανών εκμάθησης. Όλες οι τεχνικές που περιλαμβάνουν εκμάθηση πάνω στο φάσμα (όπως π.χ. αναγνώριση ομιλητή, αναγνώριση φωνής, κατηγοριοποίηση μουσικής), περιλαμβάνουν ένα βήμα εξομάλυνσης των κορυφών και κωδικοποίησης, όπως για παράδειγμα τα Mel Frequency Cepstral Coefficients (MFCC) χαρακτηριστικά.

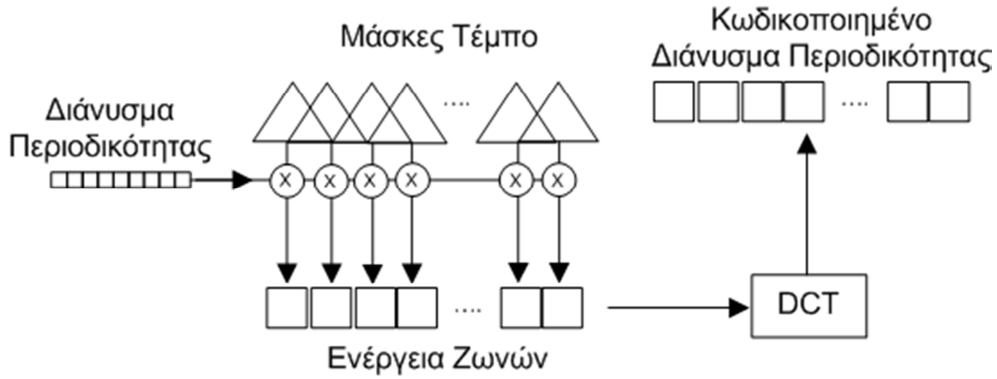
Μερικές πρόσφατες εργασίες ασχολήθηκαν με την φασματική μοντελοποίηση ρυθμικής πληροφορίας. Οι Holzapfel και Stylianou [Holzapfel2011a] εφάρμοσαν τον μετασχηματισμό κλίμακας (scale transform) στην συνάρτηση αυτοσυσχέτισης μουσικών σημάτων για να σχηματίσουν μια ρυθμική αναπαράσταση και να διερευνήσουν πτυχές της ρυθμικής ομοιότητας. Στο [Peeters2011a] ο συγγραφέας συγκρίνει διάφορους ρυθμικούς περιγραφείς ενώ στο [Peeters2010] ο DFT της συνάρτησης έμφασης υποδειγματοληπτείται σε κεντρικές συχνότητες που αντιστοιχούν σε αρμονικές των τέμπο συγκεκριμένων μετρικών δομών.

Βάσει της παραπάνω συζήτησης προτείνεται μία μέθοδος επεξεργασίας του διανύσματος περιοδικότητας με ανάλυση σε ζώνες τέμπο, κατ' αναλογία με τα χαρακτηριστικά MFCC. Το διάστημα των τέμπο στόχων $[T_{min}, T_{max}]$ χωρίζεται σε K ίσα υποδιαστήματα τέμπο, με 50% επικάλυψη μεταξύ διαδοχικών διαστημάτων. Για κάθε ένα από τα διαστήματα αυτά, υιοθετείται μια συμμετρική τριγωνική μάσκα $M[T,k]$. Για κάθε υποδιάστημα $k=1..K$ και για κάθε συνάρτηση περιοδικότητας $v_{\mathbf{x}}[T]$ για το χαρακτηριστικό $\mathbf{x}$ υπολογίζεται η ισχύς της συνάρτησης περιοδικότητας σε αυτό το διάστημα ως το εσωτερικό γινόμενο της $v_{\mathbf{x}}[T]$ διάνυσμα με την αντίστοιχη τριγωνική μάσκα $M[T,k]$

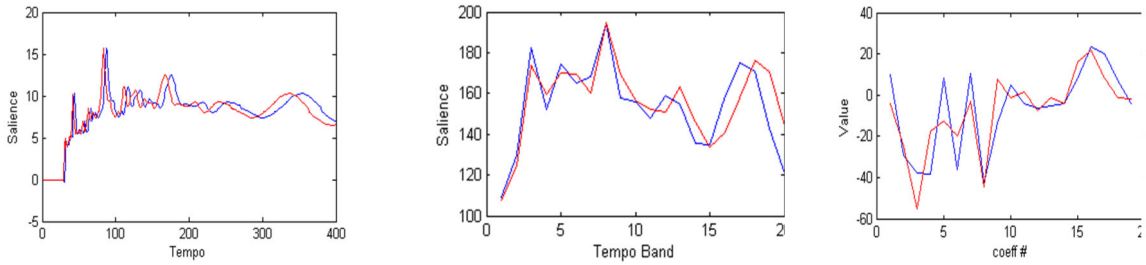
$$y_{\mathbf{x}}[k] = \sum_{T_{min}}^{T_{max}} v_{\mathbf{x}}[T] \cdot M[T,k] \;\leftrightarrow\; \mathbf{y_x} = \mathbf{v_x M}. \tag{3.4}$$

Η παραπάνω προσέγγιση έχει το μειονέκτημα ότι υπάρχει ισχυρή συσχέτιση μεταξύ των γειτονικών χαρακτηριστικών $y_{\mathbf{x}}[k]$, που οφείλεται στην επικάλυψη διαδοχικών διαστημάτων. Για να αντιμετωπίσουμε αυτό το φαινόμενο, μπορούμε να εφαρμόσουμε κάποια μέθοδο αποσυσχέτισης (decorrelation). Μία τέτοια μέθοδος αποτελεί ο Διακριτός Μετασχηματισμός Συνημίτονου (Discrete Cosine Transform, DCT) ενός διανύσματος $a_n, n = 1..N-1$, από τον οποίο προκύπτουν οι προσεγγιστικά ασυσχέτιστοι συντελεστές $\mathbf{m_x}$:

$$\mathbf{m_x} = \mathrm{m}_{\mathbf{x}}[l] = \mathrm{DCT}(\mathbf{y_x}) = \sum_{k=0}^{K-1} y_{\mathbf{x}}[k] \cos\left(\frac{\pi}{K}(l+\frac{1}{2})\right),\; l = 1 \ldots K-1. \tag{3.5}$$



Σχήμα 3.1: Διαδικασία κωδικοποίησης του διανύσματος περιοδικότητας για μία συνιστώσα χαρακτηριστικών.



α) β) γ)

Σχήμα 3.2: α) Διάνυσμα περιοδικότητας ενός κομματιού (κόκκινο) και μιας κλιμακούμενης εκδοχής του (μπλε) κατά $a=1.05$. β) Το αποτέλεσμα της επεξεργασίας από τις μάσκες (Εξ. 3.4). γ) Οι συντελεστές που προκύπτουν μετά τον DCT (Εξ. 3.5).

Ακολουθώντας παρόμοια τακτική σχετικά με την παράθεση των διαφορετικών χαρακτηριστικών με την Ενότητα 2.6 (Εξ. 2.32), συνθέτουμε τα επιμέρους $m_{\mathbf{x}}$ σε ένα διάνυσμα που το ονομάζουμε «κωδικοποίηση» της συνάρτησης περιοδικότητας

$$\mathbf{m} = [\mathbf{m}_e^1 | \dots | \mathbf{m}_e^E | \mathbf{m}_{ch}] \tag{3.6}$$

όπου $\mathbf{m}_{ch} = \text{DCT}(\mathbf{y}_{ch})$, με $\mathbf{y}_{ch} = \mathbf{v}_{\mathbf{x}}\mathbf{M}$. Στο Σχ. 3.1 φαίνεται με γραφικό τρόπο τα στάδια κωδικοποίησης του διανύσματος περιοδικότητας, ενώ στο Σχ. 3.2 φαίνεται το αποτέλεσμα της διαδικασίας για δύο συναρτήσεις περιοδικότητας για δύο κομμάτια όπου το ένα αποτελεί κλιμάκωση του άλλου.

Στο Σχ. 3.2α παρουσιάζονται δύο διανύσματα περιοδικότητας (Εξ. 2.34) που έχουν κανονικοποιηθεί ώστε να έχουν νόρμα ίση με τη μονάδα. Η ευκλείδεια απόσταση των δύο διανυσμάτων είναι ίση με 0.1579. Στη μεσαία εικόνα παρουσιάζονται αντίστοιχα τα $y[k]$ και αυτά κανονικοποιημένα. Σε αυτή την περίπτωση η ευκλείδεια απόσταση είναι μικρότερη και ίση με 0.0732. Για να καταδείξουμε σε μεγαλύτερη κλίμακα ότι η κωδικοποίηση του διανύσματος περιοδικότητας μπορεί να βελτιώσει την αναπαράσταση ρυθμικών κλάσεων, θεωρούμε το πρόβλημα της κατηγοριοποίησης σε τρεις κλάσεις ταχύτητας με το αρχικό και το κωδικοποιημένο διάνυσμα. Οι τρεις κλάσεις C_s, C_m, C_f {slow, moderate, fast} εξήχθησαν από το επισημειωμένο τέμπο στο σύνολο δεδομένων ISMIR 2004 *Songs* [Gouyon2006] (Βλ. Κεφ. 4.3) με τον εξής κανόνα κατωφλίωσης

$$c(T)=\begin{cases} slow, & T_{slow} \geq T \\ \text{mod}, & T_{slow} < T < T_{fast} \\ fast, & T \geq T_{fast} \end{cases}. \tag{3.7}$$

Η συλλογή αποτελείται από 465 μουσικά αποσπάσματα διαφόρων ειδών για τα οποία υπολογίσαμε τον δείκτη Davies-Bouldin (DB)

$$DB = \frac{1}{n}\sum_{i=1}^{n} R_i \tag{3.8}$$

όπου

$$R_i = \max_{j, j\neq i}\left\{\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right\}. \tag{3.9}$$

Με c_k συμβολίζεται το κέντρο βάρους της κλάσης k και με σ_k η μέση απόσταση των στοιχείων της κλάσεως k από το c_k. Ο δείκτης *DB* μας δείχνει την ποιότητα διαχωρισμού των τριών κλάσεων, και λαμβάνονται υπόψη τόσο οι αποστάσεις εντός της κάθε κλάσης όσο και οι αποστάσεις μεταξύ των κλάσεων. Όσο οι αποστάσεις των παραδειγμάτων σε μία κατηγορία μικραίνουν και οι αποστάσεις μεταξύ κατηγοριών μεγαλώνουν τόσο μικρότερος γίνεται ο δείκτης *DB*. Επομένως, μικρότερος δείκτης *DB* δείχνει καλύτερα διαχωρίσιμες κατηγορίες. Στον Πίνακα 3.1 παρουσιάζονται τα R_i για τις τρεις κατηγορίες (*s,m,f*) και για τα αρχικά και τα κωδικοποιημένα διανύσματα περιοδικότητας, για τις ακόλουθες μεθόδους κανονικοποίησης των διανυσμάτων εισόδου:

α) **0-1** κανονικοποίηση όλων των διαστάσεων στο διάστημα [0..1]

$$\hat{\mathbf{x}} = \mathbf{x}/\max(\mathbf{x}) \tag{3.10}$$

β) **std** κανονικοποίηση ώστε κάθε διάνυσμα να έχει μηδενική μέση τιμή και διακύμανση ίση με 1

$$\hat{\mathbf{x}} = (\mathbf{x} - \bar{\mathbf{x}})/\mathrm{Var}(\mathbf{x}) \tag{3.11}$$

γ) **Norm** κανονικοποίηση ώστε κάθε διάνυσμα να έχει L_2 νόρμα ίση με 1

$$\hat{\mathbf{x}} = \mathbf{x}/\|\mathbf{x}\| \tag{3.12}$$

Ο DCT έχει παραλειφθεί καθώς ως ορθοκανονικός μετασχηματισμός δεν αλλάζει τις αποστάσεις μεταξύ των παραδειγμάτων, ωστόσο παίζει σημαντικό ρόλο στο στάδιο της εκμάθησης όπως θα δείξουμε σε επόμενο κεφάλαιο. Από τον Πίνακα 3.1 πράγματι φαίνεται ότι ο δείκτης *DB* είναι αρκετά μικρότερος στην περίπτωση των κωδικοποιημένων διανυσμάτων περιοδικότητας.. Το ίδιο δεν συμβαίνει πάντα χρησιμοποιώντας απευθείας τα αρχικά διανύσματα περιοδικότητας.

Κανονικοποίηση	R_s *(slow)*	R_s *(moderate)*	R_f *(fast)*	R *(all)*
Αρχικά διανύσματα				
0-1	6.03	6.03	4.23	5.43
std	5.77	5.38	5.77	5.64
Norm	5.52	5.47	5.52	5.50
Κωδικοποιημένα διανύσματα				
0-1	4.79	4.31	4.79	4.63
std	4.66	3.88	4.66	4.40
Norm	4.65	3.88	4.66	4.40

Πίνακας 3.1: Ο δείκτης Davies-Bouldin για τα αρχικά διανύσματα και για τα κωδικοποιημένα διανύσματα, για κάθε κλάση χωριστά, C_s, C_m, C_f καθώς και για όλες τις κλάσεις.

3.3 Εξαγωγή Χαρακτηριστικών με Τεχνικές μη Επιβλεπόμενης Μάθησης

Σε αυτή τη παράγραφο θα παρουσιαστούν δύο τεχνικές αυτόματης εξαγωγής χαρακτηριστικών από την συνάρτηση περιοδικότητας. Η πρώτη είναι ένας γραμμικός μετασχηματισμός, με χρήση της Ανάλυσης σε Κύριες Συνιστώσες (Principal Component Analysis, PCA). Η δεύτερη είναι μια μη γραμμική απεικόνιση που προκύπτει από την εφαρμογή των Περιορισμένων Μηχανών Boltzmann (Restricted Boltzmann Machines, RBM). Οι παράμετροι και των δύο τεχνικών υπολογίζονται από μη επισημειωμένα δεδομένα, υπό την έννοια ότι δεν χρειάζεται η κατηγορία που ανήκουν τα δεδομένα ή κάποια άλλη πληροφορία σχετικά με το πρόβλημα. Για αυτό τον λόγο μπορούν να χαρακτηριστούν ως τεχνικές *μη επιβλεπόμενης μάθησης* (unsupervised learning).

3.3.1 Principal Component Analysis

Η Ανάλυση σε Κύριες Συνιστώσες (Principal Component Analysis, PCA) είναι μια στατιστική διαδικασία η οποία χρησιμοποιεί έναν ορθογώνιο γραμμικό μετασχηματισμό για να μετατρέψει ένα σύνολο παρατηρήσεων που αποτελούνται από συσχετιζόμενες (correlated) μεταβλητές σε ένα σύνολο τιμών γραμμικά μη συσχετιζόμενων (uncorrelated) μεταβλητών. Δεδομένου μιας συλλογής μουσικών δεδομένων με ΣΠ $V = \{\mathbf{v}_i\}$, $\mathbf{v}_i \in R^{N_1}$, η ανάλυση σε κύριες συνιστώσες βρίσκει έναν γραμμικό μετασχηματισμό $\mathbf{W}$ όπου $\mathbf{W} = [\mathbf{w}_1|\mathbf{w}_2|\dots|\mathbf{w}_{N_2}]], \mathbf{w}_k \in R^{N_1}$, ο οποίος μετασχηματίζει τα δεδομένα ως $\mathbf{y}_i = \mathbf{W^T v_i}$ σε έναν διανυσματικό χώρο χαμηλότερης διάστασης N_2 τέτοιο ώστε να ελαχιστοποιείται η ποσότητα $\sum_i \|\mathbf{v}_i - \mathbf{W}\mathbf{W^T}\mathbf{v}_i\|$. Τα *διατεταγμένα* ορθοκανονικά διανύσματα $\mathbf{w}_k$ ονομάζονται Κύριες Συνιστώσες (Principal Components). Οι κύριες συνιστώσες υπολογίζονται διαδοχικά, έτσι ώστε κάθε συνιστώσα να μεγιστοποιεί την διακύμανση του $V = \{\mathbf{v}_i\}$, υπό τον περιορισμό ότι είναι κάθετη σε όλες τις υπόλοιπες. Η $1^η$ συνιστώσα $\mathbf{w}_1$ υπολογίζεται ως το διάνυσμα που μεγιστοποιεί την διακύμανση ως:

$$\mathbf{w}_1 = \text{argmax}_{\|w\|=1}\{\|\mathbf{V}\mathbf{w}\|^2\}. \tag{3.13}$$

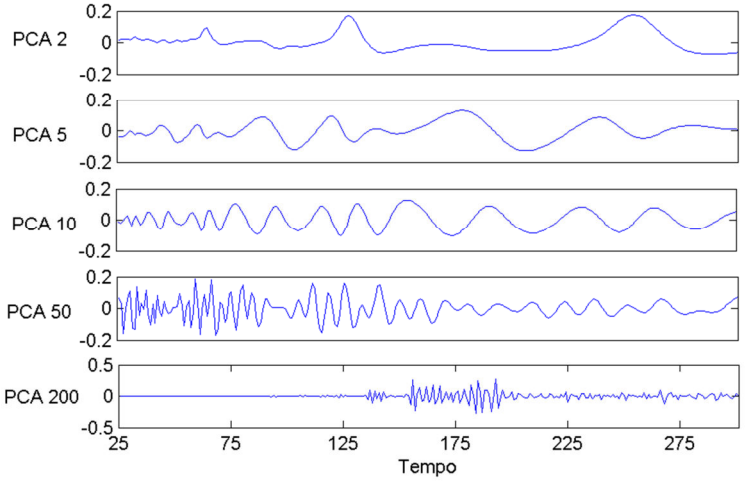
Η $2^η$ συνιστώσα υπολογίζεται με τον ίδιο τρόπο αφαιρώντας από το V την $1^η$ συνιστώσα κ.ο.κ:

$$\mathbf{w}_2 = \text{argmax}_{\|w\|=1}\{\|\mathbf{V}_2\mathbf{w}\|^2\},\ \mathbf{V}_2 = \mathbf{V} - \mathbf{V}\mathbf{w}_1\mathbf{w}_1^T. \tag{3.14}$$

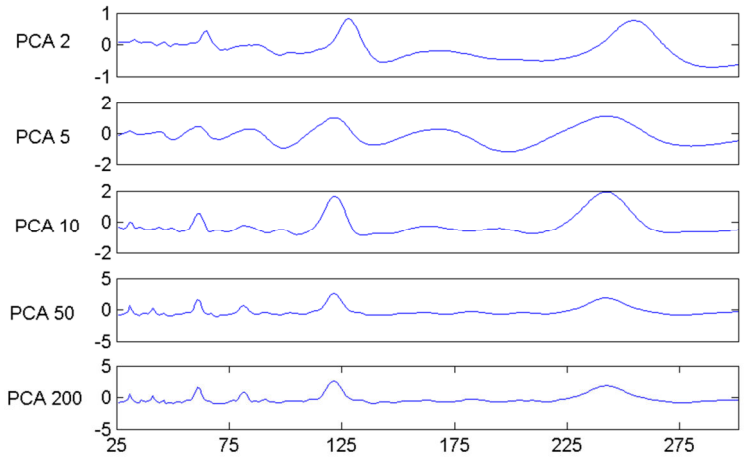
Οι συνιστώσες κατώτερης τάξης σχετίζονται με γενικά χαρακτηριστικά των δεδομένων, ενώ οι συνιστώσες ανώτερης τάξης σχετίζονται με τις λεπτομέρειες. Αν $N_1 = N_2$ τότε $\sum_i \left\| \mathbf{v}_i - \mathbf{W}\mathbf{W}^{\mathbf{T}}\mathbf{v}_i \right\| = 0$, επομένως η ποσότητα $\sum_i \left\| \mathbf{v}_i - \mathbf{W}\mathbf{W}^{\mathbf{T}}\mathbf{v}_i \right\|$ που ισούται με το σφάλμα ανακατασκευής μειώνεται καθώς αυξάνεται το N_2 και τα $\mathbf{W}\mathbf{W}^{\mathbf{T}}\mathbf{v}_i$ προσεγγίζουν τα $\mathbf{v}_i$ καθώς $N_2 \to N_1$. Αν συμβολίσουμε με λ_k την διακύμανση της συνιστώσας $\mathbf{w}_k$, τότε το σύνολο $\left\{ \mathbf{w}_k / \sqrt{\lambda_k} \right\}_k$ αποτελεί έναν άλλο μετασχηματισμό, ο οποίος ονομάζεται ως «Λευκοποιημένη Ανάλυση σε Κύριες Συνιστώσες» (whitened PCA). Η βασική διαφορά των PCA και whitened PCA είναι ότι ενώ στον PCA κάθε συνιστώσα έχει διαφορετική σημασία στον μετασχηματισμένο διανυσματικό χώρο, στην περίπτωση του whitened PCA όλες οι συνιστώσες έχουν την ίδια σημασία αφού έχουν κανονικοποιηθεί βάσει της διακύμανσής τους.

Στο Σχ. 3.3(α) παρουσιάζονται η 2^{η}, η 5^{η}, η 10^{η} , η 50^{η} και η 200^{η} συνιστώσα του PCA ο οποίος έχει εφαρμοστεί στην συνάρτηση περιοδικότητας ενός μουσικού αποσπάσματος. Για λόγους ευκολότερης παρουσίασης, στο συγκεκριμένο παράδειγμα οι συναρτήσεις περιοδικότητας που αντιστοιχούν στα διάφορα χαρακτηριστικά έμφασης έχουν συμπτυχτεί σε μία ΣΠ σύμφωνα με την Εξ. 2.34. Οι συντελεστές υπολογίστηκαν σε μία συλλογή 130.000 μουσικών αποσπασμάτων, εξάγοντας μία ΣΠ ανά απόσπασμα. Η ΣΠ υπολογίστηκε για $T_{min} = 25$ και $T_{max} = 300$ BPM με βήμα 1 BPM. Παρατηρούμε ότι οι χαμηλής τάξης συνιστώσες αντιστοιχούν σε πλατιές διακυμάνσεις της ΣΠ, ενώ οι υψηλότερης τάξης σε περισσότερες λεπτομέρειες της ΣΠ. Στο Σχ. 3.3(β) παρουσιάζεται η ανακατασκευή της ΣΠ χρησιμοποιώντας τις 2, 5, 10, 50 και 200 πρώτες συνιστώσες του PCA. Όπως ήταν αναμενόμενο, όταν χρησιμοποιείτε μεγάλο πλήθος συνιστωσών (200 συνιστώσες), αποκομίζουμε σχεδόν τέλεια ανακατασκευή. Ωστόσο, ακόμα και για σχετικά μικρό αριθμό συνιστωσών (10-50 συνιστώσες), έχουμε αρκετά καλή ανακατασκευή της αρχικής ΣΠ, αφού οι πιο εξέχουσες κορυφές είναι εμφανείς ακόμα και με τη χρήση μόλις 10 συνιστωσών. Επομένως μπορούμε να συμπεράνουμε ότι τα δεδομένα εισόδου (διάστασης 276) μπορούν να ανασυσταθούν και άρα να αναπαρασταθούν αποτελεσματικά σε έναν διανυσματικό χώρο αρκετά μικρότερης διάστασης από τον αρχικό. Στο 3.3(γ) παρουσιάζεται η ανακατασκευή της ΣΠ χρησιμοποιώντας τον whitened PCA. Παρατηρούμε ότι για μικρό πλήθος χρησιμοποιούμενων συνιστωσών, η ανακατασκευή της ΣΠ μοιάζει με την αντίστοιχη του PCA. Όσο όμως το πλήθος των συνιστωσών αυξάνει (50 συνιστώσες), η ανακατασκευασμένη ΣΠ παρουσιάζει διακυμάνσεις που δεν υπάρχουν στην αντίστοιχη ανακατασκευή με τον PCA. Το φαινόμενο αυτό είναι ακόμα πιο έντονο στην περίπτωση των 200 συνιστωσών, όπου παρατηρούμε παραμόρφωση της ΣΠ. Αυτό οφείλεται στο γεγονός ότι οι υψηλής τάξης συνιστώσες του PCA έχουν ενισχυθεί λόγω της κανονικοποίησης με τον όρο $1/\sqrt{\lambda_k}$. Επομένως, οι πολύ υψηλής τάξης συνιστώσες μπορούν να θεωρηθούν ότι αποτελούν θόρυβο, όπως φαίνεται για παράδειγμα στο Σχ. 3.3(α) (200^{η} συνιστώσα). Το φαινόμενο της παραμόρφωσης της ΣΠ λόγω των υψηλής τάξης συνιστωσών δεν είναι έκδηλο στον PCA, αφού η επίδρασή τους στην ανακατασκευή είναι μικρότερη, λόγω της μη κανονικοποίησης.

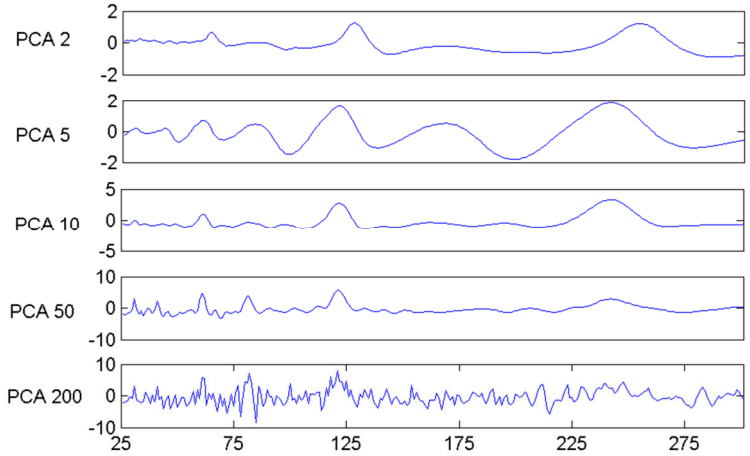
Συνοψίζοντας, μπορούμε να ισχυριστούμε ότι ο whitened PCA υπερτερεί του PCA για σχετικά μικρό πλήθος συνιστωσών, αφού λόγω της κανονικοποίησης οι συνιστώσες είναι ισοδύναμες και επομένως περιγράφουν καλύτερα την ΣΠ. Πειραματική επαλήθευση αυτού του ισχυρισμού θα παρουσιαστεί στα επόμενα κεφάλαια, καθώς θα παρουσιαστούν συγκριτικά αποτελέσματα για διάφορα προβλήματα αυτόματης ρυθμικής ανάλυσης.



(α)



(β)



(γ)

Σχήμα 3.3: (α) Η $2^{η}$, $5^{η}$, $10^{η}$, $50^{η}$ και $200^{η}$ του PCA στη συνάρτηση περιοδικότητας. (β) Η ανακατασκευή της συνάρτησης περιοδικότητας χρησιμοποιώντας τις πρώτες 2, 5, 10, 50 και 200 συνιστώσες του PCA. (γ) Η ανακατασκευή της συνάρτησης περιοδικότητας χρησιμοποιώντας τις πρώτες 2, 5, 10, 50 και 200 συνιστώσες του whitened PCA.

Ωστόσο, πέραν ενός πλήθους συνιστωσών, οι επιπλέον συνιστώσες αντιστοιχούν σε θόρυβο και η χρησιμοποίησή τους μειώνει τόσο την ποιότητα ανακατασκευής όσο και την ποιότητα της αναπαράστασης της ΣΠ. Η έρευση του βέλτιστου πλήθος των συνιστωσών για τον whitened PCA μπορεί να γίνει μόνο εμπειρικά, καθώς δεν υπάρχει κάποια μεθοδολογία. Το βέλτιστο πλήθος συνιστωσών εξαρτάται κυρίως από την διάσταση του χώρου εισόδου, από την κατανομή των δεδομένων και από τον πλεονασμό (redundancy) των χαρακτηριστικών εισόδου.

3.3.2 Restricted Boltzmann Machines

Οι Μηχανές Boltzmann είναι Γραφικά Πιθανοτικά Μοντέλα Ενέργειας (Energy Based Probabilistic Models) που αποτελούνται από έναν ορατό (visible) και ένα κρυφό (hidden) επίπεδο (layer) μεταβλητών. Έχουν την ίδια αρχιτεκτονική με τα δίκτυα Hopfield (Σχ. 3.4α). Αποτελούνται από ένα σύνολο μονάδων (units) με κατάσταση s_i, έναν πίνακα συνάψεων $\mathbf{W}$ και ένα διάνυσμα πόλωσης των μονάδων $\boldsymbol{\theta} = (\theta_i)_i$. Η «ενέργεια» του δικτύου ισούται με

$$E = -\left(\sum_{i<j} w_{ij} s_i s_j + \sum_i \theta_i s_i\right) \tag{3.15}$$

Η διαφορά μιας μηχανής Boltzmann με ένα δίκτυο Hopfield είναι ότι οι μονάδες (units) στις μηχανές Boltzmann είναι *στοχαστικές*, ο πίνακας $\mathbf{W}$ είναι συμμετρικός, ενώ $w_{ii} = 0$, δηλαδή καμία μονάδα δεν συνδέεται με τον εαυτό της. Η πιθανότητα μία μονάδα να έχει κατάσταση 1 είναι

$$p(s_i = 1) = 1/(1 + \exp(\Delta E_{\mathrm{i}})/T) \tag{3.16}$$

όπου $\Delta E_{\mathrm{i}} = \sum_j w_{ij}\, s_j + \theta_i$ και T μια παράμετρος που έχει την έννοια της θερμοκρασίας.

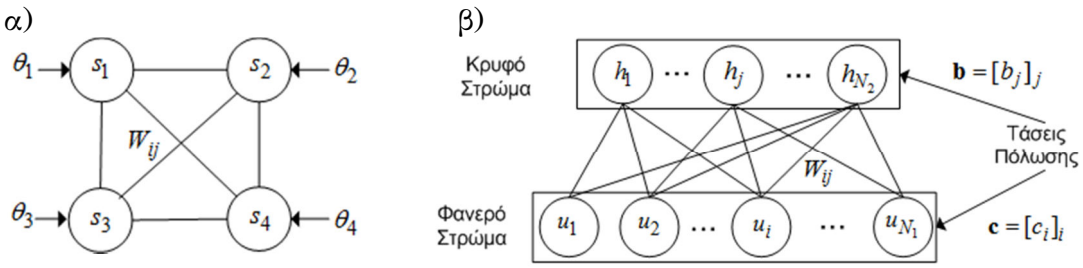
Μια ειδική κατηγορία των Μηχανών Boltzmann είναι η Περιορισμένη Μηχανή Boltzmann (Restricted Boltzmann Machine, RBM). Ένα RBM (Σχ. 3.4β) αποτελείται από δύο κατηγορίες μονάδων που σχηματίζουν δύο επίπεδα (layers). Το ορατό επίπεδο (visible layer) του οποίου τα στοιχεία συμβολίζονται με $\mathbf{v} = [v_i]$ και το κρυφό επίπεδο (hidden layer) που συμβολίζεται ως $\mathbf{h} = [h_j]$. Αντίθετα με τις Μηχανές Boltzmann, στα RBM δεν υπάρχουν συνάψεις μεταξύ των μονάδων του ίδιου στρώματος (από αυτήν την ιδιότητα προκύπτει ο όρος περιορισμένη, Restricted). Στην περίπτωση δυαδικών (binary) μονάδων, η (3.15) που δίνει την ενέργεια του δικτύου μπορεί να γραφτεί ως

$$E(\mathbf{v}, \mathbf{h}) = -\left(\sum_{i,j} w_{ij} v_i h_j + \sum_i c_i v_i + \sum_i b_j h_j\right) = -(\mathbf{v}^T \mathbf{W} \mathbf{h} + \mathbf{c}^T \mathbf{v} + \mathbf{b}^T \mathbf{h}) \tag{3.17}$$

όπου $\mathbf{W}$ είναι ο πίνακας των συνάψεων μεταξύ των μονάδων των δύο επιπέδων και $\mathbf{b}, \mathbf{c}$ είναι τα διανύσματα της τάσης πόλωσης των μονάδων του ορατού και κρυφού επιπέδου αντίστοιχα. Η πιθανότητα $p(\mathbf{v}, \mathbf{h})$ ενός ζεύγους τιμών $(\mathbf{v}, \mathbf{h})$ δίνεται από

$$p(\mathbf{v}, \mathbf{h}) = Z^{-1} e^{-E(\mathbf{v}, \mathbf{h})} \tag{3.18}$$

όπου $Z = \sum_{\mathbf{v},\mathbf{h}} p(\mathbf{v}, \mathbf{h})$ είναι ένας παράγοντας κανονικοποίησης που ονομάζεται συνάρτηση διαμελισμού (partition function). Δεδομένου ότι δεν υπάρχουν συνάψεις μεταξύ των μονάδων του ίδιου επιπέδου, οι μονάδες του κρυφού επιπέδου είναι μεταξύ τους ανεξάρτητες (independent) δεδομένου των μονάδων του ορατού επιπέδου.



Σχήμα 3.4: (α) Μία Μηχανή Boltzmann με 4 μονάδες (units). (β) Μια Περιορισμένη Μηχανή Boltzmann (Restricted Boltzmann Machine).

Αντίστοιχα, οι μονάδες του ορατού στρώματος είναι αμοιβαία ανεξάρτητες δεδομένου των μονάδων του κρυφού στρώματος. Λόγω αυτής της ανεξαρτησίας, οι αντίστοιχες δεσμευμένες πιθανότητες μπορούν να γραφούν ως γινόμενα των επιμέρους δεσμευμένων πιθανοτήτων

$$p(\mathbf{v}|\mathbf{h}) = \prod_i p(v_i|\mathbf{h}),\ p(\mathbf{h}|\mathbf{v}) = \prod_j p(h_j|\mathbf{v}). \tag{3.19}$$

Στην περίπτωση που οι μονάδες και των δύο επιπέδων είναι δυαδικές, προκύπτει ότι

$$p(v_i|\mathbf{h}) = \sigma(\mathbf{c}_i + \sum_j w_{ij}\, h_j), \quad p(h_j|\mathbf{v}) = \sigma(\mathbf{b}_j + \sum_i w_{ij}\, v_i) \tag{3.20}$$

όπου $\sigma(x) = \frac{1}{1+e^{-x}}$ είναι η λογιστική συνάρτηση.

Τα RBM είναι παραγωγικά μοντέλα (generative models), δηλαδή εκπαιδεύονται ώστε να μεγιστοποιηθεί η πιθανότητα περιθωρίου (marginal probability) $p(\mathbf{v})$ των μεταβλητών του ορατού επιπέδου. Επομένως για τον υπολογισμό των παραμέτρων $\mathbf{W}, \mathbf{c}, \mathbf{b}$ κατά την εκπαίδευση χρειάζεται η εφαρμογή της μεθόδου κλίσης (gradient ascent) στον λογάριθμο της πιθανοφάνειας (log-likelihood) της πιθανότητας περιθωρίου $p(\mathbf{v})$. Οι κανόνες μάθησης με την μέθοδο κλίσης έχουν προταθεί από τον [Hinton1983] και είναι

$$\begin{aligned} \frac{\partial \log p(\mathbf{v})}{\partial w_{ij}} &= \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \\ \frac{\partial \log p(\mathbf{v})}{\partial c_i} &= \varepsilon(\langle v_i \rangle_{data} - \langle v_i \rangle_{model}) \\ \frac{\partial \log p(\mathbf{v})}{\partial b_j} &= \varepsilon\big(\langle h_j \rangle_{data} - \langle h_j \rangle_{model}\big) \end{aligned} \tag{3.21}$$

όπου με $\langle \cdot \rangle_{data}$ συμβολίζουμε την αναμενόμενη (expectation) τιμή σε σχέση με τα δεδομένα, ενώ με $\langle \cdot \rangle_{model}$ συμβολίζεται η αναμενόμενη τιμή σε σχέση με το μοντέλο. Ο υπολογισμός των παραπάνω όρων είναι εκθετικής υπολογιστικής πολυπλοκότητας ως προς το πλήθος των μονάδων του κρυφού επιπέδου, επομένως ο υπολογισμός των παραμέτρων βάσει της (3.21) είναι πρακτικά ανέφικτος. Ένας εναλλακτικός τρόπος υπολογισμού των παραμέτρων $\mathbf{W}, \mathbf{c}, \mathbf{b}$ προτάθηκε από τον Hinton [Hinton2002] με την ελαχιστοποίηση της ποσότητας Contrastive Divergence. Οι κανόνες εκμάθησης για τα $\mathbf{W}, \mathbf{c}, \mathbf{b}$ δίνονται από τις ακόλουθες σχέσεις

$$\begin{aligned} \Delta w_{ij} &= \varepsilon\left(\langle v_i h_j\rangle_0 - \langle v_i h_j\rangle_k\right) \\ \Delta c_i &= \varepsilon(\langle v_i\rangle_0 - \langle v_i\rangle_k) \\ \Delta b_i &= \varepsilon(\langle h_i\rangle_0 - \langle h_i\rangle_k) \end{aligned} \tag{3.22}$$

όπου με $\langle\cdot\rangle_k$ συμβολίζουμε την τιμή μετά από k βήματα δειγματοληψίας Gibbs (Gibbs sampling) και ε είναι ο ρυθμός εκμάθησης (learning rate). Συγκρίνοντας τις Εξ. 3.21 και 3.22 παρατηρούμε ότι οι όροι της (3.21) που περιέχουν το $\langle\cdot\rangle_{data}$ και $\langle\cdot\rangle_{model}$ έχουν προσεγγιστεί χρησιμοποιώντας τα $\langle\cdot\rangle_0$ και $\langle\cdot\rangle_k$ αντίστοιχα. Δηλαδή η μέση τιμή της κατανομής των δεδομένων $\langle\cdot\rangle_{data}$ προσεγγίζεται με την μέση τιμή της εμπειρικής κατανομής $\langle\cdot\rangle_0$, ενώ η μέση τιμή βάσει του μοντέλου $\langle\cdot\rangle_{model}$ προσεγγίζεται με την μέση τιμή μετά από k βήματα δειγματοληψίας Gibbs.

Αφού τα RBM μαθαίνουν μία απεικόνιση από το ορατό στο κρυφό επίπεδο, μπορούμε να υποθέσουμε ότι η χρήση πολλών RBM τα οποία είναι οργανωμένα ιεραρχικά θα οδηγήσει σε μια απεικόνιση η οποία θα μπορεί να αναπαραστήσει ανώτερης πολυπλοκότητας χαρακτηριστικά των διανυσμάτων εισόδου. Ξεκινάμε από ένα RBM του οποίου το ορατό επίπεδο περιλαμβάνει τις μεταβλητές εισόδου (input layer) και το οποίο εκπαιδεύεται με βάση ένα σύνολο δεδομένων μάθησης για να υπολογιστούν οι παράμετροι $\mathbf{W}, \mathbf{c}, \mathbf{b}$. Στη συνέχεια οι μεταβλητές του κρυφού επιπέδου μπορούν να θεωρηθούν ως έξοδοι του $1^{ου}$ RBM και να χρησιμοποιηθούν ως μεταβλητές εισόδου στο $2^{ο}$ RBM, το οποίο εκπαιδεύεται ανεξάρτητα από το πρώτο RBM κ.ο.κ. Η προκύπτουσα αρχιτεκτονική ονομάζεται "Δίκτυο Βαθειάς Πίστης» (Deep Belief Network, DBN) [Hinton2006] και πρέπει να σημειωθεί ότι δεν είναι μια Μηχανή Boltzmann.

Τα RBM και DBN έχουν εφαρμοστεί με μεγάλη επιτυχία σε διάφορα προβλήματα αναγνώρισης, όπως για παράδειγμα αναγνώρισης εικόνας [Hinton2006], αναγνώριση φωνής [Mohammed2012] και "συνεργατικό φιλτράρισμα" (collaborative filtering) [Salakhutdinov2007]. Η ευρεία διάδοσή τους έχει οδηγήσει σε διάφορες παραλλαγές της βασικής αρχιτεκτονικής που παρουσιάστηκε. Ο σκοπός της παρούσας διατριβής δεν είναι η βαθύτερη ανάλυση των RBM και DBN, αλλά η εφαρμογή τους στα πλαίσια της Ανάκτησης Μουσικής Πληροφορίας και επομένως θα παρουσιαστούν οι παραλλαγές που χρησιμοποιήθηκαν στην προτεινόμενη μέθοδο.

Γκαουσσιανές Μονάδες (Gaussian Units)

Οι μονάδες του ορατού επιπέδου των RBM είναι δυαδικές, γεγονός το οποίο τις καθιστά ακατάλληλες για εφαρμογές όπου συναντάμε πραγματικές τιμές στα διανύσματα χαρακτηριστικών αναπαράστασης των δεδομένων, κάτι το οποίο είναι είναι πολύ σύνηθες στο πεδίο της επεξεργασίας ήχου. Αυτό συμβαίνει και στην περίπτωση της ρυθμικής ανάλυσης, αφού η ΣΠ παίρνει πραγματικές τιμές. Παρότι οι πραγματικές τιμές θα μπορούσαν να κανονικοποιηθούν στο διάστημα [0,1] και να υποτεθούν ως πιθανότητες των δυαδικών μονάδων, οι δυαδικές μονάδες έχουν αποδειχτεί [Hinton2010] ότι δεν είναι αποτελεσματικές στην μοντελοποίηση πραγματικών εισόδων. Αυτό μπορεί να παρακαμφθεί με την χρήση Γκαουσσιανών αντί δυαδικών μονάδων. Η ενέργεια του RBM σε αυτή την περίπτωση ισούται με

$$E(\mathbf{v}, \mathbf{h}) = \sum_i \frac{(v_i - c_i)^2}{2\sigma_i{}^2} - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{ij} - \sum_j b_j h_j \tag{3.23}$$

όπου σ_i είναι η τυπική απόκλιση (standard deviation) για την μονάδα i. Οι δεσμευμένες πιθανότητες για τις κρυφές μονάδες $p(h_j|\mathbf{v})$ σε αυτή τη περίπτωση παραμένουν ίδιες όπως στην (3.20), ενώ για τις μονάδες του ορατού επιπέδου είναι

$$p(v_i|\mathbf{h}) = N(\mathbf{c_i} + \sum_j w_{ij}\, h_j, \sigma_i) \tag{3.24}$$

όπου με $N(\mu, \sigma^2)$ συμβολίζεται η Κανονική Κατανομή (Normal Distribution) με μέση τιμή μ και τυπική απόκλιση σ. Επειδή ο υπολογισμός των σ_i είναι δύσκολος, είναι σύνηθες τα δεδομένα εκμάθησης να κανονικοποιούνται ώστε να έχουν μηδενική μέση τιμή και τυπική απόκλιση ίση με 1. Σε αυτή την περίπτωση μπορούμε να αγνοήσουμε τα σ_i στην (3.23).

Γραμμικές Μονάδες Ανόρθωσης (Linear Rectified Units)

Μια επέκταση των δυαδικών μονάδων είναι η χρησιμοποίηση διωνυμικών μονάδων, με την διαφορά ότι ενώ όλα τα στοιχεία της διωνυμικής κατανομής (binomial distribution) έχουν το ίδιο διάνυσμα βαρών **w** και την ίδια τάση πόλωσης b, στην συγκεκριμένη προσέγγιση έχουν διαφορετικό offsset στο b. Αν τα offsets των επιμέρους στοιχείων μιας διωνυμικής μονάδας είναι κατά σειρά -0.5, -1.5 ... –(N-0.5), τότε το άθροισμα των πιθανοτήτων για κάθε στοιχείο της διωνυμικής κατανομής θα είναι

$$\sum_{i=0}^{\infty} \sigma(x - i + 0.5) \approx \log(1 + e^x) \tag{3.25}$$

όπου $x = \mathbf{v}\mathbf{w}^T + b$, δηλαδή η συνολική είσοδος της μονάδας. Επομένως η συνολική έξοδος της τροποποιημένης διωνυμικής μονάδας προσεγγίζει μία λεία εκδοχή της συνάρτησης ημιανόρθωσης. Ενώ χρησιμοποιεί τον ίδιο αριθμό παραμέτρων με την δυαδική μονάδα, η προκύπτουσα μονάδα παρέχει μια πλουσιότερη συμπεριφορά και επομένως μπορεί να περιγράψει πιο αποτελεσματικά κατανομές δεδομένων. Λόγω του ότι ο υπολογισμός της (3.25) προϋποθέτει τον υπολογισμό μεγάλου πλήθος επιμέρους σιγμοειδών, ένας τρόπος να μειωθεί το υπολογιστικό κόστος είναι να προσεγγιστεί η (3.25) ως ένα Θορυβώδες Γραμμικό Στοιχείο (Noisy Rectified Linear Unit, NReLU) [Nair2010] του οποίου η έξοδος ισούται με $\max(0, x + N(0,1))$, όπου x η συνολική είσοδος της μονάδας και $N(0,1)$ είναι μια τυχαία μεταβλητή από την κανονική κατανομή με μέση τιμή 0 και απόκλιση 1.

Αραιότητα (Sparseness)

Τόσο τα δυαδικά όσο και τα NReLU στοιχεία έχουν πιθανότητα ενεργοποίησης ίση με 0.5, δηλαδή αναμένεται να είναι μη μηδενικά για το 50% των δεδομένων εκπαίδευσης κατά μέσο όρο. Αν τα στοιχεία αυτά είχανε μικρότερη πιθανότητα ενεργοποίησης, τότε η συνεισφορά του κάθε στοιχείου θα ήταν πιο ερμηνεύσιμη. Επιπλέον, έχει δειχτεί ότι οι αραιές (sparse) αναπαραστάσεις μπορούν να βελτιώσουν την επίδοση ταξινομητών που έχουν ως είσοδο αυτές τις αναπαραστάσεις, όπως για παράδειγμα στο [Nair2009] όπου παρουσιάστηκε μια μέθοδος αναγνώρισης τρισδιάστατων αντικειμένων.

Στην παρούσα διατριβή χρησιμοποιήθηκε μια απλουστευμένη εκδοχή της μεθόδου [Goh2010]. Αρχικά ορίζεται ένας «στόχος» ενεργοποίησης $\overline{p} \in [0,1]$, που αντιστοιχεί στην επιθυμητή πιθανότητα ενεργοποίησης για όλα τα στοιχεία. Οι εξισώσεις ανανέωσης των βαρών (3.22) τροποποιούνται έτσι ώστε

$$\begin{aligned} \Delta w_{ij} &= \varepsilon\big(\langle v_i h_j\rangle_{data} - \langle v_i h_j\rangle_{model}\big) - \eta\langle v_i(h_j - \overline{p})\rangle_{data} \\ \Delta b_i &= \varepsilon(\langle h_i\rangle_{data} - \langle h_i\rangle_{model}) - \eta\langle h_j - \overline{p}\rangle_{data} \\ \Delta c_i &= \varepsilon(\langle v_i\rangle_{data} - \langle v_i\rangle_{model}) \end{aligned} \tag{3.26}$$

Με $\eta > 0$ συμβολίζεται ο ρυθμός μάθησης για τον όρο της αραιότητας. Η Εξ. 3.26 υποδεικνύει ότι μονάδες του κρυφού στρώματος που τείνουν να έχουν μέση ενεργοποίηση μεγαλύτερη από το κατώφλι $\overline{p}$ «τιμωρούνται» και τα αντίστοιχα βάρη μειώνονται. Αντίστροφα, οι μονάδες του κρυφού επιπέδου που τείνουν να ενεργοποιούνται με μέση τιμή μικρότερη της $\overline{p}$ «προάγονται» αυξάνοντας τα αντίστοιχα βάρη. Παρότι η αρχική μέθοδος [Goh2010] ορίστηκε για δυαδικά στοιχεία, μπορεί να εφαρμοστεί και στα NReLU, αφού ο μηχανισμός που οδηγεί σε αραιές αναπαραστάσεις δεν αλλάζει.

Persistent Constrastive Divergence

Ένα από τα σημαντικότερα στοιχεία στην εκμάθηση ενός RBM, είναι ο υπολογισμός των όρων $\langle\cdot\rangle_{model}$ της Εξ. 3.21, που προϋποθέτει τη δειγματοληψία ενός δείγματος (sample) από το μοντέλο. Όπως προαναφέρθηκε, οι εξισώσεις εκμάθησης που ελαχιστοποιούν την Constrastive Divergence (Εξ. 3.22) αποτελούν προσέγγιση των αντίστοιχων εξισώσεων που μεγιστοποιούν την πιθανότητα $\log p(\mathrm{v})$ και οι όροι που αντιστοιχούν στα δείγματα από το μοντέλο $\langle\cdot\rangle_{model}$ προσεγγίζονται από τους όρους $\langle\cdot\rangle_k$. Με άλλα λόγια η δειγματοληψία κατά την Constrastive Divergence εκμάθηση γίνεται με μία Μαρκοβιανή Αλυσίδα (Markov Chain) k βημάτων. Λόγω του υπολογιστικού κόστους όμως, συνήθως επιλέγεται $k = 1$, μια τακτική που οδηγεί σε φτωχότερα μοντέλα συγκριτικά με την περίπτωση μεγαλύτερων k.

Ο Tieleman [Tieleman2008] πρότεινε μια απλή και αποτελεσματική μέθοδο δειγματοληψίας, που δεν αυξάνει το υπολογιστικό κόστος και οδηγεί σε καλύτερα δείγματα της κατανομής. Η μέθοδός του οφείλεται στην παραδοχή ότι για κάθε ανανέωση των παραμέτρων το μοντέλο δεν αλλάζει αισθητά. Έτσι, αντί να αρχικοποιείται μία νέα Μαρκοβιανή Αλυσίδα μετά από κάθε επανάληψη (iteration) ανανέωσης των παραμέτρων, αρχικοποιείται μία Μαρκοβιανή Αλύσίδα στην αρχή της εκμάθησης. Έτσι τα δείγματα σε κάθε επανάληψη δεν λαμβάνονται από Μαρκοβιανή Αλυσίδα με αφετηρία κάποια πρότυπα εκμάθησης, αλλά από την «Φανταστική» αλυσίδα που αρχικοποιήθηκε στο ξεκίνημα της εκμάθησης. Επειδή η αλυσίδα αυτή είναι η ίδια καθόλη τη διάρκεια της εκμάθησης προέκυψε και ο όρος «Επίμονη» (Persistent).

Κεφάλαιο 4: Αυτόματη Ρυθμική Κατηγοριοποίηση

4.1 Εισαγωγή

Στα προηγούμενα κεφάλαια δείξαμε τα στάδια υπολογισμού μιας συνάρτησης περιοδικότητας (Κεφ. 2) και την εξαγωγή συμπαγών χαρακτηριστικών από αυτή (Κεφ. 3). Στο παρόν κεφάλαιο θα δείξουμε ότι τα χαρακτηριστικά αυτά (καθώς και η ίδια η ΣΠ) αποτελούν μια εύρωστη ρυθμική αναπαράσταση και περιέχουν ρυθμική πληροφορία ικανή να χρησιμοποιηθεί για επιτυχή αυτόματη ρυθμική ανάλυση και κατηγοριοποίηση.

Για το σκοπό αυτό, χρησιμοποιούμε μια τυπική προσέγγιση ταξινόμησης ρυθμού πάνω στα χαρακτηριστικά που παρουσιάστηκαν βασισμένη στις Μηχανές Υποστήριξης Διανυσμάτων (Support Vector Machines). Η αρχιτεκτονική της μεθόδου αυτόματης ταξινόμησης παρουσιάζεται στο Σχ. 4.1. Το $1^{ο}$ στάδιο είναι η *Μη Επιβλεπόμενη Εξαγωγή Χαρακτηριστικών*. Ο όρος *μη επιβλεπόμενη* αναφέρεται στο γεγονός ότι τα δεδομένα τα οποία χρησιμοποιούνται στην εξαγωγή των χαρακτηριστικών δεν είναι απαραίτητο να είναι επισημειωμένα με κάποια ετικέτα ή πληροφορία σχετικά με τον ρυθμό, δηλαδή δεν χρησιμοποιείται τέτοιου είδους πληροφορία για την εξαγωγή των χαρακτηριστικών. Από μια μεγάλη συλλογή μουσικών αποσπασμάτων υπολογίστηκε η συνάρτηση περιοδικότητας για κάθε απόσπασμα. Το σύνολο των ΣΠ το οποίο συμβολίζουμε ως D_u (unlabeled), χρησιμοποιείται για την εξαγωγή των παραμέτρων των δύο τεχνικών εξαγωγής χαρακτηριστικών που παρουσιάστηκαν στο προηγούμενο κεφάλαιο, δηλ. οι συντελεστές των PCA και των RBM. Συμβολίζουμε με R τον μετασχηματισμό (PCA ή RBM) των δεδομένων. Στη συνέχεια, δεδομένου ενός προβλήματος κατηγοριοποίησης και ενός συνόλου μουσικών αποσπασμάτων επισημειωμένου με τις αντίστοιχες κλάσεις (κατηγορίες), εξάγονται όπως προηγουμένως οι ΣΠ που σχηματίζουν το σύνολο D_l (labeled). Στην συνέχεια το σύνολο D_l μετασχηματίζεται μέσω του R σε ένα σύνολο προβλήματος κατηγοριοποίησης $\{\mathbf{v}_i, \mathrm{c}_i\}$ όπου $\mathbf{v}_i$ είναι τα διανύσματα των χαρακτηριστικών εκπαίδευσης και c_l οι αντίχτοιχες κλάσεις. Τα $\{\mathbf{v}_i, \mathrm{c}_i\}$ χρησιμοποιούνται στη συνέχεια ως είσοδο σε έναν ταξινομητή SVM για την εκπαίδευση και κατηγοριοποίηση.

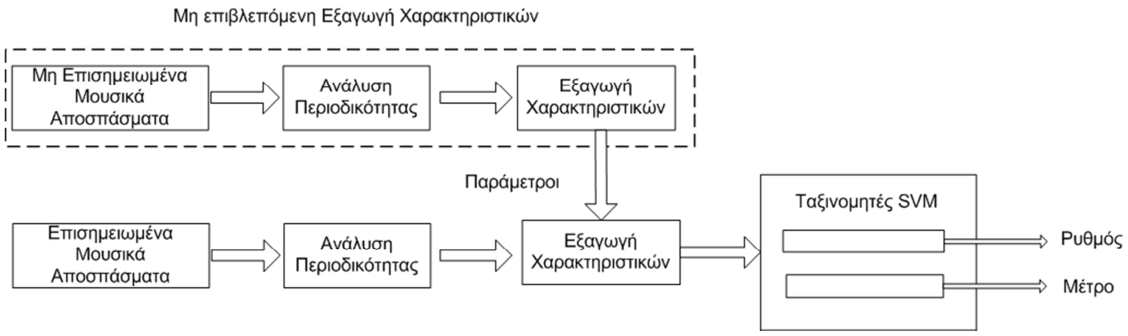
Το πρόβλημα της κατηγοριοποίησης για το SVM ορίζεται ως επιμέρους προβλήματα κατηγοριοποίησης δύο κατηγοριών (one vs. one strategy). Υπάρχουν ενδείξεις [Hsu2002] ότι αυτή η στρατηγική είναι πιο αποτελεσματική από τη συνήθη «one vs. all» όπου εκπαιδεύεται ένα SVM για κάθε κατηγορία, ειδικά στις περιπτώσεις όπου υπάρχει ανισορροπία στο μέγεθος των κατηγοριών. Στα επιμέρους προβλήματα δύο κατηγοριών τα SVM ελαχιστοποιούν την συνάρτηση κόστους

$$\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{I}\xi_i \tag{4.1}$$

σύμφωνα με τους περιορισμούς

$$y_i\left(\mathbf{w}^T\varphi(\mathbf{m}_i)\right) + b \geq 1 - \xi_i,\ \xi_i \geq 0 \tag{4.2}$$

όπου $\{(\mathbf{v}_i, \mathrm{y}_i)\}$ είναι τα δεδομένα εκμάθησης και $y_i \in \{-1,1\}$ είναι η κατηγορία στο πρόβλημα δύο κλάσεων.



Σχήμα 4.1: Διαδικασία κωδικοποίησης του διανύσματος περιοδικότητας για μία συνιστώσα χαρακτηριστικών.

Με $\varphi(\cdot)$ συμβολίζεται η Συνάρτηση Πυρήνα (Kernel Function) που απεικονίζει τα δεδομένα σε ένα χώρο υψηλότερης διάστασης και ο συντελεστής $C > 0$ «τιμωρεί» τα παραδείγματα εκμάθησης που βρίσκονται στη ζώνη του περιθωρίου των δύο κλάσεων. Στην φάση κατηγοριοποίησης, η συνάρτηση απόφασης για το άγνωστο παράδειγμα $\hat{\mathbf{v}}$ δίνεται από

$$\hat{y} = \text{sgn}(\mathbf{w}^T \varphi(\hat{\mathbf{v}}) + b) \tag{4.3}$$

Πρέπει να σημειωθεί ότι το σύνολο D_l μετασχηματίζεται με τις παραμέτρους που υπολογίστηκαν από το D_u. Στην παρούσα διατριβή, αντιμετωπίζουμε δύο προβλήματα ρυθμικής κατηγοριοποίησης. Το 1^o είναι η κατηγοριοποίηση σε είδη "χορευτικού στυλ" (dance style classification) και το 2^o είναι η εξαγωγή του χρονικού κλειδιού (time-signature) ενός μουσικού αποσπάσματος.

Για την μη επιβλεπόμενη εξαγωγή των χαρακτηριστικών χρησιμοποιήθηκε ένα υποσύνολο 130.000 αποσπασμάτων της συλλογής Million Song Dataset (MSD) [Bertin-Mahieux2011]. Το MSD αποτελείται από 1.000.000 μουσικά κομμάτια και αποτελεί μία εξαιρετική πηγή μουσικής πληροφορίας, αφού περιλαμβάνει πληθώρα μεταδεδομένων και ακουστικών χαρακτηριστικών για κάθε κομμάτι, όπως τον καλλιτέχνη, ταμπέλες χρηστών (user tags), παρόμοιους καλλιτέχνες, δημοτικότητα (popularity), κ.ά. Επιπλέον περιλαμβάνει και φασματικά χαρακτηριστικά υπολογισμένα ανά πλαίσιο (mfcc, chromas) καθώς και υψηλότερου επιπέδου χαρακτηριστικά (χρονικό κλειδί, τέμπο, θέσεις παλμού) τα οποία όμως έχουν εξαχθεί με αυτόματο τρόπο. Το MSD περιλαμβάνει κυρίως σύγχρονη δυτική μουσική, αλλά και άλλα είδη όπως κλασσική ή folk μουσική. Παρότι δεν υπάρχει συστηματική καταγραφή των ειδών μουσικής που υπάρχουν στο MSD, μια ενδεικτική κατανομή τους μπορεί να βρεθεί στα [Liang2011], [Schindler2012]. Λόγω περιορισμών πνευματικών δικαιωμάτων, μόνο τα μεταδεδομένα είναι διαθέσιμα και όχι τα αρχεία ήχου. Ωστόσο μέρος των μεταδεδομένων είναι και ένα μοναδικό αναγνωριστικό (unique identifier) από τον ιστότοπο 7digital[2], οπότε είναι δυνατή η λήψη μέσω διαδικτύου αποσπασμάτων 30" ή 60" για τα περισσότερα κομμάτια.

Οι συντελεστές PCA υπολογίστηκαν χρησιμοποιώντας την ενσωματωμένη συνάρτηση princomp της Matlab. Για την εκπαίδευση του RBM χρησιμοποιήθηκαν Γκαουσσιανές μονάδες στο ορατό επίπεδο και Γραμμικές Μονάδες Ανόρθωσης στο κρυφό. Για την δειγματοληψία από μοντέλο χρησιμοποιήσαμε τη μέθοδο Persistent Contrastive Divergence.

[2] http://developer.7digital.com/resources/api-docs/standard-response-objects

Είδος	Πλήθος	Μέτρο	Εύρος Τέμπο (BPM)
Cha Cha Cha	111	4/4	92-137
Jive	60	4/4	124-182
Quickstep	82	4/4	189-216
Rumba	98	4/4	73-45
Samba	86	4/4	138-247
Tango	86	4/4	112-135
Viennese Waltz	65	3/4	168-186
Waltz	110	3/4	78-106

Πίνακας 4.1. Στατιστικά στοιχεία των κατηγοριών της συλλογής Ballroom.

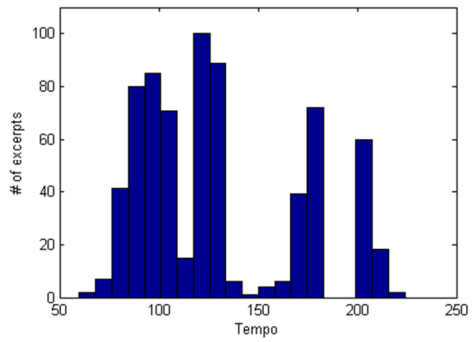
Επιπλέον υιοθετήθηκε και ο όρος αραιότητας της Εξ. 3.26. Οι ρυθμοί μάθησης ε και η της (3.26) επιλέχθηκαν ίσοι με 10^{-4} και 4×10^{-5} αντίστοιχα και ο στόχος ενεργοποίησης $\overline{p} \in [0,1]$. Το σύνολο εκμάθησης διαμερίστηκε σε μικρές ομάδες (mini-bacthes) των 100 παραδειγμάτων. Η ανανέωση των βαρών βάσει της (3.26) σε κάθε επανάληψη δεν γινόταν για όλο το σύνολο εκμάθησης, αλλά για κάθε minibatch ξεχωριστά. Η εκμάθηση του RBM τερματίστηκε όταν το κόστος ανακατασκευής του δικτύου σε ένα σύνολο επαλήθευσης (validation set) ελαχιστοποιήθηκε. Ως κόστος ανακατασκευής για ένα σύνολο $V = \{\mathbf{v}_i\}$ ενός RBM ορίζεται η ποσότητα $\sum_i \|\mathbf{v}_i - \tilde{\mathbf{v}}_i\|^2$ όπου $\tilde{\mathbf{v}}_i$ είναι οι μεταβλητές στο ορατό επίπεδο μετά από ένα βήμα δειγματοληψίας Gibbs με αφετηρία τα $\mathbf{v}_i$. Τέλος, για να μειώσουμε την επίδραση που μπορεί να είχε στα πειραματικά αποτελέσματα η στοχαστική φύση του RBM, τα πειράματα εκτελέστηκαν οκτώ φορές (εκμάθηση RBM και SVM) και ως τελικό αποτέλεσμα ελήφθη ο μέσος όρος. Η παραπάνω διαδικασία έγινε για διάφορες τιμές του μεγέθους του κρυφού επιπέδου.

Για την καλύτερη αξιολόγηση της προτεινόμενης μεθόδου και μείωση της αλλοίωσης των αποτελεσμάτων που μπορεί να προκληθεί από την τυχαιότητα με την οποία χωρίζεται ένα σύνολο δεδομένων στα υποσύνολα εκπαίδευσης, επαλήθευσης και αξιολόγησης (training / validation / testing), ακολουθήθηκε η τακτική «Τεμαχισμού σε Ν υποσύνολα δια-επαλήθευσης» (N-fold Cross Validation). Η συλλογή χωρίζεται σε N υποσύνολα ίσου μεγέθους τέτοια ώστε η κατανομή των κλάσεων σε κάθε υποσύνολο να είναι περίπου η ίδια με την κατανομή στην συνολική συλλογή. Στη συνέχεια, η ένωση των N-2 υποσυνόλων χρησιμοποιούνται ως σύνολο εκμάθησης και τα άλλα 2 ως σύνολα επαλήθευσης και αξιολόγησης. Η ίδια διαδικασία επαναλαμβάνεται N φορές, έτσι ώστε και τα N υποσύνολα να έχουν εμφανιστεί ως υποσύνολο αξιολόγησης. Τέλος, τα επιμέρους N αποτελέσματα συγχωνεύονται για τον υπολογισμό του τελικού αποτελέσματος. Στη παρούσα διατριβή χρησιμοποιήσαμε $N = 10$ για όλα τα πειράματα. Σχετικά με το SVM, χρησιμοποιήσαμε την βιβλιοθήκη LIBSVM [Chang2011]. Για πυρήνα του SVM χρησιμοποιήθηκε η Συνάρτηση Ακτινικής Βάσης (Radial Basis Function, RBF):

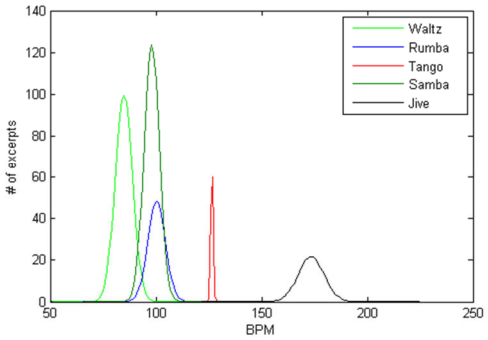
$$K(x,y) = e^{-\frac{\|x-y\|^2}{2\gamma^2}} \tag{4.4}$$

Οι βέλτιστοι παράμετροι του SVM C (Εξ. 4.1) και γ (Εξ. 4.4) υπολογίστηκαν με αναζήτηση πλέγματος (grid search). Για κάθε ένα από τα N πειράματα, οι τιμές του ζεύγους τιμών C, γ του πλέγματος που μεγιστοποιούν το ποσοστό ταξινόμησης στο σύνολο επαλήθευσης, χρησιμοποιούνται στη συνέχεια για εκμάθηση και αξιολόγηση της μεθόδου.

α)



β)



Σχήμα 4.2: (α) Η κατανομή των τέμπο στην συλλογή Ballroom. (β) Η κατανομή των τέμπο για κάθε κατηγορία ενός υποσυνόλου των κατηγοριών της συλλογής Ballroom.

Στην περίπτωση πολλών κλάσεων χρησιμοποιήθηκε η στρατηγική «ένα-έναντι-ένα» (one-vs-one) και επικράτηση της κλάσης με τον κανόνα της πλειοψηφίας (majority voting).

Στη συνέχεια αυτού του κεφαλαίου θα παρουσιαστούν αναλυτικά τα πειραματικά αποτελέσματα για την αυτόματη κατηγοριοποίηση χορευτικής μουσικής (Ενότητα 4.1) και την εξαγωγή του μέτρου (Ενότητα 4.2), ενώ στην Ενότητα 4.3 θα παρουσιαστούν επιπλέον λεπτομέρειες και θα γίνει περαιτέρω ανάλυση των αποτελεσμάτων για βαθύτερη κατανόηση των χαρακτηριστικών και της επίδρασής τους στην επίδοση.

4.2 Πειραματικά Αποτελέσματα

4.2.1 Αυτόματη Κατηγοριοποίηση Χορευτικής Μουσικής

Για την αξιολόγηση της μεθόδου στην αυτόματη κατηγοριοποίηση χορευτικής μουσικής χρησιμοποιήθηκε η συλλογή Ballroom [Gouyon2006]. Αποτελείται από 698 αποσπάσματα, το καθένα διάρκειας 30", από οκτώ κατηγορίες χορευτικών ρυθμών. Αναλυτικά στοιχεία για την κατανομή των ειδών καθώς και το μέτρο και εύρος των τέμπο για κάθε κλάση παρουσιάζονται στον Πίνακα 4.1. Πρέπει να σημειωθεί ότι υπάρχει μεγάλη συσχέτιση της κατηγορίας και του τέμπο ενός κομματιού. Στο Σχήμα 4.2α παρουσιάζεται η συνολική κατανομή των τέμπο, ενώ στο Σχήμα 4.2β η κατανομή των τέμπο ανά μουσική κλάση. Είναι φανερό ότι κάποιες κατηγορίες, όπως για παράδειγμα η Tango και η Jive είναι τελείως διαχωρίσιμες αν είναι γνωστό το τέμπο. Η παρατήρηση αυτή ενισχύεται από το γεγονός ότι στο [Gouyon2004b] παρουσιάστηκε μία μέθοδος όπου με απλή μοντελοποίηση του τέμπο κάθε κατηγορίας με μια Γκαουσσιανή Κατανομή, πέτυχε ποσοστό 80% σωστής ταξινόμησης. Από την άλλη αυτό δεν ισχύει για όλα τα ζεύγη των κατηγοριών, όπως για παράδειγμα για τις Samba, Waltz και Rumba, όπου υπάρχει επιμέρους επικάλυψη.

Στον Πίνακα 4.2 παρουσιάζονται τα αποτελέσματα ταξινόμησης για διάφορες τιμές της διάστασης εξόδου του μετασχηματισμού με την PCA, wPCA και RBM. Επιπλέον, στον Πίνακα 4.2 παρουσιάζονται και τα αποτελέσματα χρησιμοποιώντας απευθείας την ΣΠ χωρίς καμία επεξεργασία (Εξ. 2.32) σε σύγκριση με μεθόδους της διεθνούς βιβλιογραφίας.

	Διάσταση				
Χαρακτηριστικά	50	100	200	500	1000
PCA	87.5	88.1	89.1	89.1	89.4
wPCA	89.1	89.8	88.1	87.7	83.2-
RBM	85.6	87	89.5	90.9	90.7

Άλλες Μέθοδοι	Αρχική ΣΠ	[Gouyon2004c]	[Peeters2005]	[Dixon2004]
	88.9	79.6- (90.1)	81- (90.4)	84- (96+)

Πίνακας 4.2. (Πάνω) Πειραματικά αποτελέσματα της αυτόματης ταξινόμησης χορευτικής κατηγορίας για τις 3 μεθόδους εξαγωγής χαρακτηριστικών και διάφορες διαστάσεις των μετασχηματισμών. (Κάτω) Συγκριτικά πειραματικά αποτελέσματα με τη διεθνή βιβλιογραφία χωρίς εξαγωγή χαρακτηριστικών (χρήση απευθείας της συνάρτησης περιοδικότητας). Οι τιμές στην παρένθεση για τις μεθόδους αναφοράς αντιστοιχούν σε αποτελέσματα όταν το αληθές τέμπο χρησιμοποιήθηκε ως χαρακτηριστικό. Τα σύμβολα +/- δείχνουν σημαντική στατιστική διαφορά (statistical significance) συγκριτικά με την μέθοδο «Αρχική ΣΠ».

Το ποσοστό σωστής ταξινόμησης που επιτεύχθηκε χωρίς κανέναν μετασχηματισμό της ΣΠ είναι 88.9%. Όταν χρησιμοποιούμε τα PCA και RBM ως αναπαράσταση της ΣΠ, η επίδοση αυξάνεται όταν η διάσταση εξόδου του μετασχηματισμού είναι πάνω από 200. Αυτό είναι περισσότερο έκδηλο στην περίπτωση του RBM. Αντίθετα, στην περίπτωση του wPCA, η καλύτερη επίδοση που παρατηρείται είναι εφάμιλλη με αυτή των RBM και PCA, η οποία επιτυγχάνεται όμως με πολύ μικρότερη διάσταση αναπαράστασης. Ωστόσο, η επίδοση μειώνεται όσο το πλήθος των χαρακτηριστικών αυξάνει μετά τις 100 διαστάσεις. Αυτό το αποτέλεσμα επαληθεύει και τον ισχυρισμό της Ενότητας 3.3.1 ότι οι υψηλής τάξης συνιστώσες αποτελούν θόρυβο και δεν περιέχουν καμία πληροφορία.

Στον Πίνακα 4.2 παρουσιάζονται επίσης συγκριτικά αποτελέσματα της προτεινόμενης μεθόδου χωρίς την εξαγωγή των χαρακτηριστικών με μεθόδους της διεθνούς βιβλιογραφίας. Τα σύμβολα +/- υποδηλώνουν σημαντική στατιστική διαφορά (statistical significance) με την μέθοδο που χρησιμοποιεί την ΣΠ χωρίς την εξαγωγή χαρακτηριστικών. Η στατιστική διαφορά υπολογίστηκε με την δοκιμή A/B (A/B test) και τιμή $p=0.05$. Ακολουθεί μια σύντομη περιγραφή των μεθόδων αναφοράς.

Μέθοδος των Gouyon, Dixon, Pampalk και Widmer

Η μέθοδος αυτή [Gouyon2004c] χρησιμοποιεί 4 ειδών χαρακτηριστικά: (α) χαρακτηριστικά βασισμένα στο τέμπο, όπως ίδιο το τέμπο, (είτε το αληθές τέμπο, είτε το τέμπο που έχει υπολογιστεί αυτόματα από μια άλλη μέθοδο), το tatum κ.ά. (β) Χαρακτηριστικά όπως το κέντρο βάρους (centroid), η συνολική ισχύς και η υψηλότερη κορυφή, τα οποία έχουν εξαχθεί από το «ιστόγραμμα περιοδικότητας» (Periodicity Histogram), που είναι κάτι ανάλογο με την ΣΠ. (γ) Αντίστοιχα χαρακτηριστικά που έχουν εξαχθεί από τα IOIs αντί του ιστογράμματος περιοδικότητας. (δ) Τα Mel Frequency Cepstral Coefficients, (MFCC). Στη συνέχεια με μεθόδους επιλογής χαρακτηριστικών (feature selection) και έναν ταξινομητή κοντινότερου γείτονα πέτυχαν ποσοστό αναγνώρισης 79.6%.

Όταν χρησιμοποίησαν το σωστό τέμπο ως χαρακτηριστικό το ποσοστό αυτό αυξήθηκε στο 90.1%.

Μέθοδος του Peeters

Ο Peeters [Peeters2005] ακολούθησε πολύ διαφορετική προσέγγιση. Χρησιμοποίησε αυτούσια την συνάρτηση περιοδικότητας ως είσοδο σε ένα ταξινομητή μέσω παλινδρόμησης (Classification Via Regression) [Witten1999]. Η συνάρτηση περιοδικότητας υπολογίστηκε με συνδυασμό της συνάρτησης αυτοσυσχέτισης (ACF) και του DFT, η οποία στη συνέχεια κανονικοποιήθηκε με το τέμπο. Έτσι κομμάτια με το ίδιο «σχήμα» της ΣΠ αλλά διαφορετικό τέμπο (επομένως η μία θα είναι κλιμακούμενη έκδοση της άλλης, βλ. Σχ. 3.2) μετά την κανονικοποίηση θα έχουν τις ίδιες ΣΠ. Η προτεινόμενη μέθοδος πέτυχε ποσοστό αναγνώρισης 81% και 90.4% όταν το αληθές τέμπο περιλαμβάνονταν μεταξύ των χαρακτηριστικών.

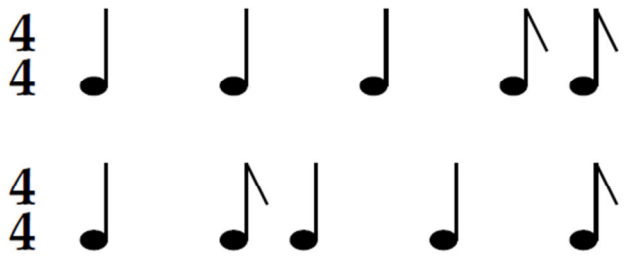
Μέθοδος των Dixon, Gouyon και Widmer

Η μέθοδος αυτή [Dixon2004] χρησιμοποιεί αποκλειστικά χρονικά χαρακτηριστικά (temporal features) για την ρυθμική κατηγοριοποίηση. Για την εξαγωγή των χαρακτηριστικών αυτών προϋποθέτει την γνώση της χρονικής στιγμής αρχής και τέλους του $1^{ου}$ μέτρου (bar) για κάθε απόσπασμα. Στη συνέχεια με τη χρήση μιας μεθόδου εξαγωγής παλμού εξάγονται οι θέσεις των υπόλοιπων μέτρων. Η συνάρτηση έμφασης κανονικοποιείται ως προς το μήκος του μέτρου και τα διανύσματα που προκύπτουν ομαδοποιούνται. Το κέντρο της πιο πολυπληθούς ομάδας θεωρείται ως το «ρυθμικό πρότυπο» του αποσπάσματος. Από αυτό το πρότυπο εξάγονται χαρακτηριστικά όπως το μέγιστο πλάτος, η διακύμανση και ο λόγος swing[3]. Σε συνδυασμό με τα χαρακτηριστικά του [Gouyon2004c] και έναν ταξινομητή κοντινότερου γείτονα η μέθοδος αυτή πέτυχε ποσοστό αναγνώρισης 85.7% και 96% όταν το αληθές τέμπο περιλαμβάνονταν μεταξύ των χαρακτηριστικών.

Οι δύο πρώτες μέθοδοι έχουν πολύ παρόμοια ποσοστά αναγνώρισης (στα όρια του στατιστικού λάθος) ακόμα και όταν δίνεται το αληθές τέμπο ως χαρακτηριστικό. Η $3^{η}$ μέθοδος υπερτερεί ελαφρά των άλλων δύο, αφού περιλαμβάνει επιπλέον χρονικά χαρακτηριστικά. Τα χρονικά αυτά χαρακτηριστικά βοηθούν στον διαχωρισμό προτύπων που μπορεί να είναι διαφορετικά, αλλά να οδηγούν στην ίδια συνάρτηση περιοδικότητας. Ένα τέτοιο παράδειγμα φαίνεται στο Σχ. 4.3 όπου δύο ρυθμικά πρότυπα ChaChaCha και Rumba δίνουν τελείως διαφορετική αίσθηση ρυθμού αλλά οδηγούν στο ίδιο ιστόγραμμα IOIs.

Από τα συγκριτικά αποτελέσματα του Πίνακα 4.2 προκύπτει ότι η προτεινόμενη μέθοδος υπερτερεί σημαντικά σε σχέση με τις τρεις μεθόδους αναφοράς, οι οποίες είναι και οι μέθοδοι με τα μεγαλύτερα ποσοστά αναγνώρισης για το σύνολο Ballroom στη διεθνή βιβλιογραφία. Ακόμα και όταν το τέμπο είναι γνωστό σε αυτές τις μεθόδους, η προτεινόμενη μέθοδος επιτυγχάνει παρόμοια αποτελέσματα (εκτός της [Dixon2004]). Μπορούμε να συμπεράνουμε ότι πέρα από την ρυθμική κατηγορία, η προτεινόμενη ΣΠ και τα εξαχθέντα από αυτή χαρακτηριστικά κωδικοποιούν και πληροφορία σχετική με το τέμπο, το οποίο θα επαληθευτεί πειραματικά στο Κεφάλαιο 5.

[3] https://en.wikipedia.org/wiki/Swing_(jazz_performance_style)



Σχήμα 4.3: Ρυθμικό πρότυπο ενός κομματιού ChaChaCha (πάνω) και ενός Rumba (κάτω). Το σχήμα προέρχεται από το [Dixon2004]

	ChaCha	Jive	Quickstep	Rumba	Samba	Tango	VWaltz	Waltz
ChaCha	**93**	0	1	4	0	1	0	1
Jive	3	**82**	0	7	0	3	0	5
Quickstep	1	0	**93**	5	1	0	0	0
Rumba	1	5	3	**82**	1	3	0	5
Samba	1	4	1	4	**88**	1	0	1
Tango	5	1	0	0	0	**94**	0	0
VWaltz	0	0	0	1	0	0	**94**	5
Waltz	0	2	0	5	0	0	1	**92**

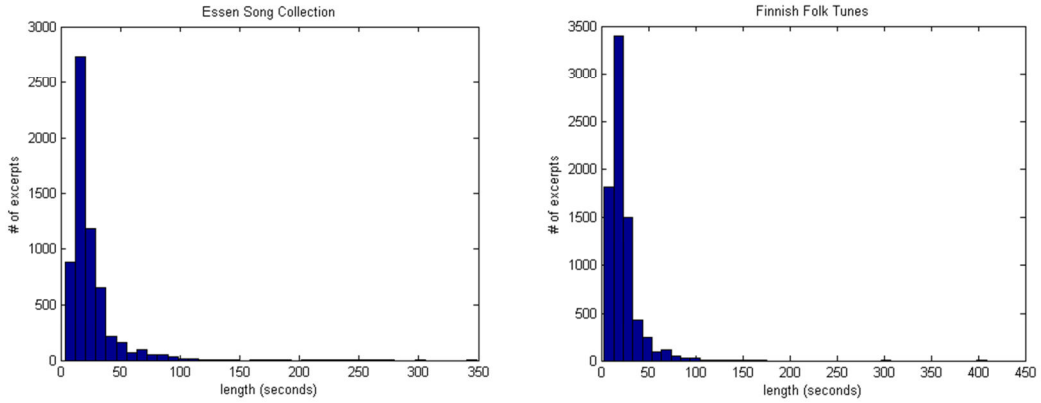
Πίνακας 4.3. Πίνακας σύγχυσης (%) μεταξύ των κατηγοριών του Ballroom χρησιμοποιώντας RBM με 200 κρυφές μονάδες. Οι στήλες αντιστοιχούν σε προβλέψεις και οι γραμμές στις πραγματικές κατηγορίες.

Στον Πίνακα 4.3 παρουσιάζεται ο πίνακας σύγχυσης (confusion matrix) για ένα από τα πειράματα που διεξήχθησαν. Συγκεκριμένα παρουσιάζονται τα αποτελέσματα χρησιμοποιώντας ως χαρακτηριστικά το RBM με 200 κρυφές μονάδες. Παρατηρούμε ότι δεν υπάρχει κάποια συστηματική σύγχυση μεταξύ των κλάσεων, δηλαδή να υπάρχει κάποιο ζεύγος κλάσεων που να είναι δυσκολότερο να διαχωριστούν μεταξύ τους. Τα χαμηλότερα αναγνώρισης επιτεύχθηκαν για τις κατηγορίες Jive και Rumba ενώ τα υψηλότερα για τις Tango, ChachaCha, Jive και Viennese Waltz.

4.2.2 Αυτόματη Εξαγωγή Μουσικού Κλειδιού

Η μέθοδος εξαγωγής μουσικού κλειδιού αξιολογήθηκε σε δύο συλλογές, οι οποίες είναι πολύ μεγαλύτερες από την Ballroom και επομένως τα αποτελέσματα σε αυτές μπορούν να οδηγήσουν σε πιο αξιόπιστα συμπεράσματα. Είναι οι συλλογές Essen Song Collection [Schaffrath1995] και Finnish Folk Tunes [Eerola2004]. Και οι δύο αποτελούνται από MIDI αρχεία, τα οποία για τους πειραματικούς σκοπούς της παρούσας διατριβής μετατράπηκαν σε αρχεία ήχου με σύνθεση Διαμόρφωσης Συχνότητας (Frequency Modulation). Η συχνότητα δειγματοληψίας του ηχητικού σήματος τέθηκε στα 22kHz, ενώ το τέμπο επιλέχτηκε σταθερό για όλα τα κομμάτια και ίσο με 120 BPM. Ως όργανο για τη σύνθεση από τα MIDI αρχεία επιλέχτηκε το φλάουτο. Η πληροφορία του χρονικού κλειδιού εξήχθη απευθείας από την διάρκεια των μέτρων στα MIDI αρχεία.

(α) (β)



Σχήμα 4.4: Η κατανομή της διάρκειας των κομματιών στην συλλογή (α) Essen Song Collection και στην συλλογή (β) Finnish Folk Tunes.

Essen Song Collection

Μέτρο	2/4	3/2	3/4	3/8	4/1	4/2	4/4	6/4	6/8	Σύνολο
Πλήθος	1333	114	1283	307	51	238	1641	121	796	5884

Finnish Folk Tunes

Μέτρο	2/4	3/2	3/4	3/8	4/4	5/2	5/4	6/4	6/8	Σύνολο
Πλήθος	3530	73	964	142	2199	38	409	79	223	7657

Πίνακας 4.4. Κατανομή των κατηγοριών των συλλογών Essen Song Collection και Finnish Folk Tunes.

Και οι δύο συλλογές αποτελούνται κυρίως από μονοφωνικές απλές μελωδίες. Το ιστόγραμμα της διάρκειας των αρχείων για κάθε συλλογή παρουσιάζεται στο Σχ. 4.4.

Τα στατιστικά στοιχεία της κατανομής των κλάσεων στα δύο σύνολα δεδομένων παρουσιάζονται στον Πίνακα 4.4. Σε αντίθεση με τη συλλογή Ballroom, η κατανομή των κλάσεων είναι πολύ ανομοιογενής. Στην συλλογή Essen είναι κυρίαρχες οι κλάσεις 2/4, 3/4 και 4/4, ενώ υπάρχουν κλάσεις που είναι εξαιρετικά σπάνιες, όπως για παράδειγμα οι κλάσεις 4/1 και 3/2. Τα σπάνια μουσικά κλειδιά μπορεί να μην είναι τα πραγματικά, αλλά να προέκυψαν ως αποτέλεσμα της μορφής του MIDI αρχείου απ'όπου προέρχονται. Αυτό συμβαίνει επειδή ένα MIDI αρχείο δεν είναι πάντοτε πιστή απόδοση της παρτιτούρας καθώς στην διαδικασία μετατροπής μιας παρτιτούρας σε ένα MIDI αρχείο υπεισέρχεται ο ανθρώπινος παράγοντας. Ακόμα πάντως και στην περίπτωση κομματιών με αληθές χρονικό κλειδί 4/1, αυτά θα μπορούσαν να ομαδοποιηθούν κάτω από την κατηγορία 4/4, αφού τα κλειδιά αυτά έχουν πολύ κοντινό ρυθμικό περιεχόμενο. Κάτι αντίστοιχο συμβαίνει και για άλλα ζεύγη κλάσεων, όπως για παράδειγμα για τις κλάσεις των κλειδιών 3/4 και 3/8. Ωστόσο, για να είναι δυνατή η απευθείας σύγκριση της προτεινόμενης μεθόδου με άλλες μεθόδους στη βιβλιογραφία, στα πειράματα διατηρήθηκαν οι κλάσεις του Πίνακα 4.4. Αντίστοιχη ανάλυση μπορεί να γίνει και για τη συλλογή Finnish Folk Tunes. Στη συλλογή αυτή συναντάμε τις ίδιες πολυπληθείς κλάσεις με την συλλογή Essen, αλλά υπάρχουν οι κλάσεις των κλειδιών 5/2, 5/4 αντί των 4/1, 4/2. Ωστόσο και σε αυτή την περίπτωση διατηρήθηκαν οι κλάσεις του Πίνακα 4.4. Τέλος, σε

συμφωνία πάντα με άλλες μεθόδους, η μέθοδος αξιολογήθηκε στα ίδια δεδομένα αφού όμως ομαδοποιήσαμε τα χρονικά κλειδιά σε δύο μεγάλες κατηγορίες, στην κατηγορία των διπλών (duple) μέτρων (2/4, 4/4, 4/1, 4/2) και στην κατηγορία των τριπλών/σύνθετων (triple/compound) μέτρων (3/2, 3/4, 3/8, 5/2, 5/4, 6/4, 6/8). Εφεξής θα χρησιμοποιήσουμε τις συντομογραφίες Essen-9 και F-Folk-9 για τις δύο συλλογές και Essen-2, F-Folk-2 για τις συλλογές όταν έχουν συμπτυχτεί οι 9 κλάσεις σε 2. Παρότι η πληροφορία του μουσικού μέτρου είναι πάρα πολύ σημαντική, το πλήθος των μεθόδων που πραγματεύονται με άμεσο τρόπο την κατηγοριοποίηση μουσικού κλειδιού είναι λίγες (Βλ. Ενότητα 1.3.2). Οι μέθοδοι αναφοράς που έχουν αξιολογηθεί στις δύο αυτές συλλογές και επιτυγχάνουν αξιοσημείωτα αποτελέσματα είναι οι ακόλουθες.

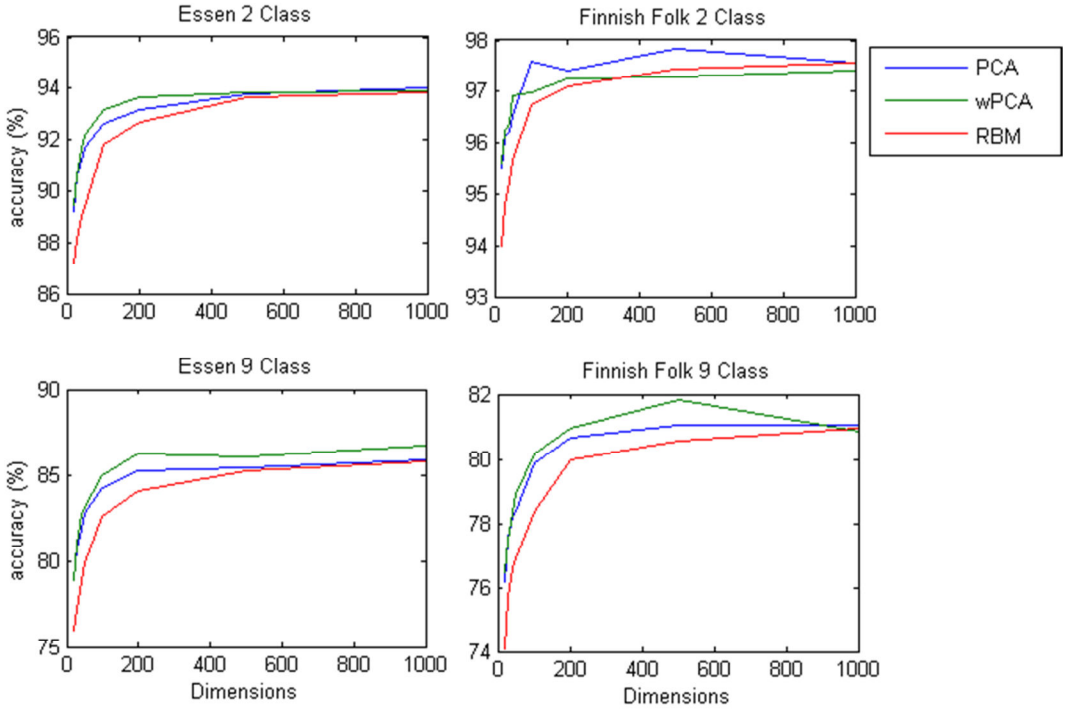
Μέθοδος των Toiviainen και Eerola

Η μέθοδος αυτή [Toiviainen2006] προϋποθέτει MIDI αρχεία ως είσοδο. Από την MIDI αναπαράσταση των κομματιών εξάγεται μια συνάρτηση έμφασης. Οι συγγραφείς αναφέρουν αποτελέσματα για 5 διαφορετικές συναρτήσεις έμφασης. Στη συνέχεια υπολογίζεται η αυτοσυσχέτιση της συνάρτηση έμφασης η οποία αποτελεί την συνάρτηση περιοδικότητας. Η προτεινόμενη μέθοδος πέτυχε ποσοστό αναγνώρισης 83.2% και 68% για τις συλλογές Essen-9 και F-Folk-9 αντίστοιχα και 95.3% και 96.4% για τις συλλογές Essen-2 και F-Folk-2. Σχετικά με τον διαχωρισμό σε σύνολα εκμάθησης/αξιολόγησης, για την ταξινόμηση ενός κομματιού μιας συλλογής χρησιμοποιήθηκαν τα υπόλοιπα κομμάτια της συλλογής ως δεδομένα εκπαίδευσης (leave-one-out cross-validation).

Μέθοδος των Eck και Casagrande

Η μέθοδος αυτή [Eck2005] επεκτείνει την έννοια στη αυτοσυσχέτισης σε έναν Πίνακα Αυτοσυσχέτισης Φάσης (Autocorrelation Phase Matrix). Για διάφορες τιμές φάσης, υπολογίζεται η αυτοσυσχέτιση έχοντας ως αποτέλεσμα έναν πίνακα του οποίου η μία διάσταση είναι η φάση [0,2π) και η άλλη η καθυστέρηση (lag). Στη συνέχεια, για κάθε lag υπολογίζεται η εντροπία ως προς τις φάσεις. Το προκύπτων διάνυσμα πολλαπλασιάζεται κατά σημείο με την συνάρτηση αυτοσυσχέτισης για να εξαχθεί η τελική συνάρτηση περιοδικότητας της μεθόδου. Στη συνέχεια ένας απλός ευριστικός κανόνας στην ΣΠ βρίσκει αν το μέτρο είναι «διπλό» ή «τριπλό». Η προτεινόμενη μέθοδος δεν αναγνωρίζει χρονικό κλειδί, αλλά μόνο την κατηγορία του μουσικού κλειδιού και πέτυχε ποσοστά επιτυχούς ταξινόμησης 90% και 93% για τις συλλογές Essen-2 και F-Folk-2 αντίστοιχα.

Στο Σχήμα 4.5 παρουσιάζεται το ποσοστό ταξινόμησης για κάθε μέθοδο/σύνολο ως προς την διάσταση των εξαγόμενων χαρακτηριστικών. Παρατηρούμε ότι για μικρές διαστάσεις τα PCA και wPCA χαρακτηριστικά υπερτερούν σημαντικά των RBM. Ωστόσο, όσο η διάσταση μεγαλώνει τα αποτελέσματα είναι παρόμοια και η διαφορά στην αναγνώριση είναι στατιστικά μη σημαντική. Μόνη εξαίρεση είναι η περίπτωση της συλλογής Essen-9. Τα wPCA χαρακτηριστικά σε αυτή την περίπτωση δίνουν καλύτερα αποτελέσματα για όλες τις διαστάσεις. Ωστόσο πρέπει να σημειωθεί ότι τα χαρακτηριστικά RBM έχουν πιο επιθυμητή συμπεριφορά ως προς την διάσταση του μετασχηματισμού. Όσο αυτή αυξάνεται, η επίδοση βελτιώνεται και δεν φαίνεται να υπάρχει ένα βέλτιστο μέγεθος του μετασχηματισμού πέρα από το οποίο τα χαρακτηριστικά να αποτελούν θόρυβο, όπως συμβαίνει με τα wPCA.



Σχήμα 4.5: Ποσοστό αναγνώρισης των τριών μεθόδων εξαγωγής χαρακτηριστικών στα 4 σύνολα δεδομένων ως προς την διάσταση των μετασχηματισμών.

	Essen Songs		Finnish Folk	
Μέθοδος	2 κλάσεις	9 κλάσεις	2 κλάσεις	9 κλάσεις
Αρχική ΣΠ	94	86	97.5	81.2
PCA (1000)	94	85.9	97.4	81.1
wPCA (500)	93.9	86.8	97.8	81.9
RBM (1000)	93.9	85.8	97.6	81
[Toiviainen2006]	95.3+	83.2-	96.4-	68-
[Eck2005]	90-	-	93-	-

Πίνακας 4.5. Συγκριτικά πειραματικά αποτελέσματα ποσοστών αναγνώρισης (%). Τα σύμβολα +/- δείχνουν σημαντική στατιστική διαφορά (statistical significance) συγκριτικά με την μέθοδο «Αρχική ΣΠ».

Στην περίπτωση των wPCA, όπως και την περίπτωση της κατηγοριοποίησης χορευτικής μουσικής, η επίδοση μειώνεται όταν αυξάνει το πλήθος των χαρακτηριστικών στην συλλογή F-Folk-9, κάτι που συμβαίνει και με τα PCA χαρακτηριστικά στην συλλογή F-Folk-2.

Ο Πίνακας 4.5 παρουσιάζει συγκριτικά αποτελέσματα με τις μεθόδους αναφοράς καθώς και με την μέθοδο «Αρχική ΣΠ», δηλαδή χωρίς την εξαγωγή χαρακτηριστικών. Για κάθε είδος χαρακτηριστικού επιλέχτηκε η διάσταση που επιτυγχάνει τα καλύτερα αποτελέσματα. Παρατηρούμε ότι δεν υπάρχει στατιστικά σημαντική διαφορά όταν υιοθετούμε κάποια τεχνική εξαγωγής χαρακτηριστικών.

	2/4	3/2	3/4	3/8	4/1	4/2	4/4	6/4	6/8
2/4	**87**	0	3	1	0	0	8	0	1
3/2	1	**67**	7	0	0	13	11	1	0
3/4	5	0	**86**	1	0	0	5	2	1
3/8	12	0	8	**52**	0	0	0	0	28
4/1	0	0	0	0	**86**	14	0	0	0
4/2	0	3	0	0	5	**87**	5	0	0
4/4	6	0	4	0	0	1	**89**	0	0
6/4	0	1	45	0	0	0	11	**43**	0
6/8	3	0	2	5	0	0	0	0	**90**

Πίνακας 4.6. Πίνακας σύγχυσης (%) μεταξύ των κατηγοριών της συλλογής Essen-9 χρησιμοποιώντας RBM με 1000 κρυφές μονάδες. Οι στήλες αντιστοιχούν σε προβλέψεις και οι γραμμές στις πραγματικές κατηγορίες. Δηλαδή τα παραδείγματα της *i* γραμμής (κλάσης) ταξινομήθηκαν στη *j* στήλη (κλάση).

Συγκριτικά με τις άλλες μεθόδους, παρατηρούμε ότι εκτός της συλλογής Essen-2 όπου η μέθοδος [Toiviainen2006] υπερτερεί ελαφρά (~1%), η προτεινόμενη μέθοδος είναι αισθητά καλύτερη από τις μεθόδους αναφοράς. Ειδικά στην περίπτωση της συλλογής F-Folk-9, η προτεινόμενη μέθοδος επιτυγχάνει πάνω από 10% υψηλότερη ακρίβεια από την μέθοδο [Toiviainen2006]. Στην συλλογή F-Folk-9 υπερτερεί ελαφρά κατά (~1%) και στην συλλογή Essen-9 κατά ~3%. Μπορούμε να συμπεράνουμε ότι παρότι η προτεινόμενη μέθοδος μπορεί να χαρακτηριστεί ως πιο «γενική», υπό την έννοια ότι δεν είναι σχεδιασμένη για την εξαγωγή του χρονικού κλειδιού, αλλά μπορεί να αντιμετωπίσει και άλλα προβλήματα ανάλυσης ρυθμού (όπως η κατηγοριοποίηση χορευτικής μουσικής), αποδίδει καλύτερα από τις δύο μεθόδους αναφοράς, οι οποίες είναι σχεδιασμένες για το συγκεκριμένο πρόβλημα.

Για την βαθύτερη κατανόηση της επίδοσης και των λαθών της μεθόδου, οι Πίνακες 4.6 και 4.7 παρουσιάζουν τον πίνακα σύγχυσης για τις συλλογές Essen-9 και F-Folk-9 αντίστοιχα. Τα αποτελέσματα πάρθηκαν χρησιμοποιώντας χαρακτηριστικά ενός RBM με 1000 κρυφές μονάδες. Τα αποτελέσματα στη συλλογή δείχνουν ότι τα περισσότερα λάθη του ταξινομητή γίνονται μεταξύ παρόμοιων κλάσεων, όπως για παράδειγμα για την κλάση 3/8, όπου το 28% των παραδειγμάτων ταξινομούνται ως 6/8, ή για την κλάση 6/4 η οποία κατά 45% ταξινομείται ως 3/4. Αυτό μπορεί εν μέρει να οφείλεται και στην ανισορροπία του πλήθους κάθε κλάσης. Η κλάση των 3/4 είναι πολύ μεγαλύτερη από την κλάση των 6/4 (βλ. Πίνακα 4.4) και επομένως «πολώνει» τον SVM ταξινομητή προς την κλάση 3/4. Είναι αξιοσημείωτο πάντως ότι υπάρχουν και αντίθετες περιπτώσεις, όπως για παράδειγμα η κλάση 4/1, η οποία ταξινομείται σωστά με ποσοστό 86%, παρότι είναι πολύ όμοια με τις μεγαλύτερες κλάσεις 4/2 και 4/4. Παρόμοια εικόνα αποκομίζουμε από τον πίνακα σύγχυσης (Πίνακα 4.7) για την συλλογή F-Folk-9. Τα περισσότερα λάθη γίνονται μεταξύ παρόμοιων κλάσεων, όπως π.χ. η κλάση 3/8, όπου το 20% των παραδειγμάτων ταξινομούνται ως 6/8 και αντίστροφα. Κάτι παρόμοιο συμβαίνει και για την κλάση 4/4, όπου το 32% των παραδειγμάτων ταξινομούνται ως 2/4. Ωστόσο υπάρχουν και συστηματικά λάθη για κλάσεις που δεν είναι παρόμοιες, όπως π.χ. οι κλάσεις 3/2, 5/2 και 6/4 που συγχέονται με την κλάση 4/4 σε ποσοστά 15 %, 18% και 14% αντίστοιχα.

	2/4	3/2	3/4	3/8	4/4	5/2	5/4	6/4	6/8
2/4	**87**	0	1	0	12	0	0	0	0
3/2	4	**73**	3	0	15	0	0	5	0
3/4	7	0	**92**	0	1	0	0	0	0
3/8	5	0	9	**65**	0	0	1	0	20
4/4	32	0	0	0	**68**	0	0	0	0
5/2	0	3	0	0	18	**79**	0	0	0
5/4	4	0	0	0	2	0	**94**	0	0
6/4	0	4	6	0	14	1	0	**75**	0
6/8	7	0	4	13	1	0	0	0	**75**

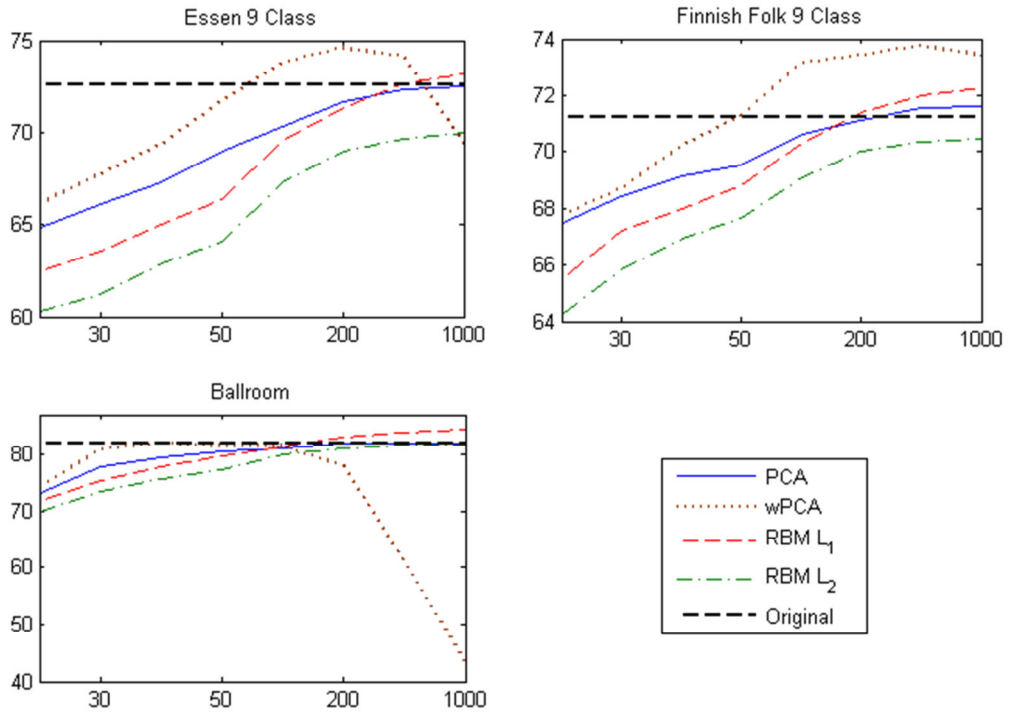
Πίνακας 4.7. Πίνακας σύγχυσης (%) μεταξύ των κατηγοριών της συλλογής F-Folk-9 χρησιμοποιώντας ως χαρακτηριστικά από ένα RBM με 1000 κρυφές μονάδες. Οι στήλες αντιστοιχούν σε προβλέψεις και οι γραμμές στις πραγματικές κατηγορίες.

4.3 Ανάλυση της Επίδοσης των Χαρακτηριστικών

Το πιο προφανές σχόλιο που μπορεί να γίνει για τα πειραματικά αποτελέσματα και των δύο προβλημάτων είναι ότι για τις περισσότερες περιπτώσεις δεν υπήρχε σημαντική διαφορά μεταξύ της επίδοσης όταν χρησιμοποιήσαμε ως αναπαράσταση της ΣΠ την έξοδο των RBM, PCA και wPCA. Σε μερικές περιπτώσεις τα PCA και wPCA είχαν ελαφριά καλύτερη επίδοση από τα RBM, όπως για παράδειγμα στην περίπτωση των 9 κλάσεων εξαγωγής μέτρου. Αντίθετα, τα χαρακτηριστικά που προέκυψαν από το RBM έδωσαν λίγο καλύτερα αποτελέσματα στην περίπτωση της ταξινόμησης σε κατηγορίες χορευτικού ρυθμού. Επιπλέον, πρέπει να παρατηρήσουμε ότι δεν υπήρχε σημαντική βελτίωση για καμιά από τις 3 μεθόδους εξαγωγής χαρακτηριστικών συγκριτικά με την χρήση της αρχικής ΣΠ ως είσοδο στον SVM ταξινομητή.

Σχετικά με τα PCA, είναι αναμενόμενο να αποδώσουν παρόμοια με την αρχική ΣΠ, ειδικά στην περίπτωση που το πλήθος των συνιστωσών είναι μεγάλο. Καθώς χρησιμοποιούμε όλο και περισσότερες συνιστώσες, η ευκλείδεια απόσταση μεταξύ των προτύπων προσεγγίζει την απόσταση των αρχικών χαρακτηριστικών της ΣΠ (βλ. Ενότητα 3.3.1), και επομένως τα αποτελέσματα που παίρνουμε με τους συντελεστές PCA είναι παρόμοια με αυτά της αρχικής ΣΠ. Αντιθέτως, στην περίπτωση του wPCA όλες οι συνιστώσες έχουν ίση σημασία και επομένως είναι αναμενόμενο να παρατηρήσουμε μείωση στην επίδοση όταν το πλήθος των συνιστωσών αυξάνει, αφού όπως προαναφέρθηκε (Εν. 3.3.1) οι υψηλής τάξης συντελεστές αντιστοιχούν σε θόρυβο. Αυτό ήταν πιο έκδηλο στην περίπτωση της ταξινόμησης σε κατηγορίες χορευτικού ρυθμού.

Σχετικά με τα RBM, στις περισσότερες περιπτώσεις παρατηρήθηκε μια μικρή αλλά στατιστικά μη σημαντική βελτίωση συγκριτικά με την αρχική ΣΠ. Αυτή η παρατήρηση δεν είναι σε συμφωνία με την γενική παραδοχή στη βιβλιογραφία ότι τα RBM μαθαίνουν σύνθετα και συμπαγή χαρακτηριστικά τα οποία βελτιώνουν την επίδοση σε μια σειρά από προβλήματα κατηγοριοποίησης. Μπορούμε να συμπεράνουμε ότι η ικανότητα των RBM να κάνουν έναν μη γραμμικό μετασχηματισμό του διανυσματικού χώρου της ΣΠ δεν είναι ωφέλιμη. Επομένως η ΣΠ που περιγράφηκε στο Κεφάλαιο 2 μπορεί να θεωρηθεί ήδη ως μια σύνθετη αναπαράσταση του ρυθμικού περιεχομένου και για την οποία το RBM δεν μπορεί να «μάθει» μια πιο πολύπλοκη δομή.



Σχήμα 4.6: Ποσοστό ανάκτησης για τα σύνολα Essen-9, F-Folk-9 και Ballroom των τριών μεθόδων εξαγωγής χαρακτηριστικών για διάφορες τιμές της διάστασής τους καθώς και της αρχικής ΣΠ.

Σε αυτή τη περίπτωση το RBM συμπεριφέρεται περισσότερο σαν μια τεχνική «μείωσης της διάστασης» (dimensionality reduction) παρά σαν μια τεχνική εξαγωγής χαρακτηριστικών. Για να το δείξουμε αυτό, προσθέσαμε δύο επιπλέον RBM πάνω από το κρυφό επίπεδο του $1^{ου}$ RBM, για να σχηματιστεί ένα Deep Belief Network. Τα πειραματικά αποτελέσματα έδειξαν μια συστηματική μείωση της επίδοσης, της τάξης του 1%, για κάθε επιπλέον RBM. Επομένως μπορούμε να συμπεράνουμε ότι προσθέτοντας RBMs δεν εξάγεται κάποια πιο σύνθετη δομή του χώρου εισόδου και το μόνο αποτέλεσμα των επιπλέον RBM είναι η αύξηση του κόστους ανακατασκευής του δικτύου η οποία οδηγεί σε χειρότερες αναπαραστάσεις.

Ο ταξινομητής SVM περιλαμβάνει και αυτός έναν μη γραμμικό μετασχηματισμό του χώρου εισόδου μέσω του RBF πυρήνα. Για να μπορέσουμε να αποκτήσουμε μια πιο διαυγή εικόνα των χαρακτηριστικών που εξάγονται με τα PCA, wPCA και RBM και να παρακάμψουμε την επίδραση του SVM στα πειραματικά αποτελέσματα, διεξήχθη το εξής πείραμα. Για κάθε απόσπασμα, υπολογίσαμε το ποσοστό των 5 κοντινότερων γειτόνων που ανήκουν στην ίδια με αυτό κλάση. Στη συνέχεια υπολογίστηκε ο μέσος όρος αυτού του ποσοστού ως προς όλα τα αποσπάσματα. Χάριν συντόμευσης, θα αναφερόμαστε σε αυτό το ποσοστό ως ποσοστό ανάκτησης (retrieval rate). Οι κοντινότεροι γείτονες υπολογίστηκαν βάσει της Ευκλείδειας L_2 νόρμας στην περίπτωση των PCA και wPCA. Σχετικά με το RBM όμως είναι πιο φυσιολογικό η απόσταση να μετρηθεί με την L_1 νόρμα, αφού η αναπαράσταση των διανυσμάτων σε αυτή τη περίπτωση είναι αραιή (sparse). Ωστόσο για λόγους πληρότητας θα αναφερθούν

αποτελέσματα ανάκτησης και για τις δύο νόρμες L_1 και L_2. Το Σχήμα 4.6 συνοψίζει τα ποσοστά ανάκτησης για διάφορες διαστάσεις των μετασχηματισμών.

Είναι προφανές ότι στην περίπτωση εξαγωγής του μέτρου, τα χαρακτηριστικά wPCA υπερτερούν των δύο άλλων μεθόδων. Ωστόσο το ποσοστό ανάκτησης μειώνεται δραματικά αν το πλήθος των συνιστωσών είναι αρκετά μεγάλο, ειδικά στη συλλογή Ballroom. Επιπλέον, δεν υπάρχει βέλτιστο πλήθος συνιστωσών για όλες τις συλλογές, το οποίο συνάδει με τα αποτελέσματα της προηγούμενης παραγράφου. Όταν χρησιμοποιούμε την L_1 στην έξοδο του RBM, τα χαρακτηριστικά PCA είναι καλύτερα για μικρές διαστάσεις, ενώ το RBM υπερτερεί για μεγαλύτερες. Ωστόσο, όπως είδαμε και προηγουμένως, μπορούμε να ισχυριστούμε ότι τα χαρακτηριστικά RBM είναι καλύτερα από τα PCA, αφού όσο μεγαλώνει η διάσταση του μετασχηματισμού, τόσο αυξάνεται και η επίδοση. Επιπλέον, χρησιμοποιώντας 200 ή περισσότερες μονάδες στο κρυφό επίπεδο, η αναπαράσταση μέσω των RBM είναι καλύτερη και από την αρχική ΣΠ. Τέλος, η σύγκριση του ποσοστού ανάκτησης για τις δύο νόρμες αναδεικνύει ότι η L_1 είναι πολύ πιο κατάλληλη να χρησιμοποιηθεί για τον υπολογισμό αποστάσεων στην περίπτωση των RBM.

Κεφάλαιο 5: Εξαγωγή Μουσικού Τέμπο

5.1 Εισαγωγή

Στο προηγούμενο κεφάλαιο είδαμε ότι η ΣΠ αποτελεί μια εύρωστη αναπαράσταση του ρυθμικού περιεχομένου. Χρησιμοποιώντας την ΣΠ καθώς και συμπαγείς αναπαραστάσεις αυτής (μέσω των μετασχηματισμών με τα PCA και τα RBM), ένας ταξινομητής SVM πέτυχε πολύ υψηλή επίδοση για δύο πολύ σημαντικά προβλήματα ρυθμικής κατηγοριοποίησης. Το παρόν κεφάλαιο πραγματεύεται την αυτόματη εξαγωγή του τέμπο από την ΣΠ. Στο Κεφάλαιο 2 είδαμε ότι η ΣΠ που προκύπτει με την συνέλιξη με τους ταλαντωτές, αφενός τονίζει όλες τις παρούσες μουσικές περιοδικότητες (και στα πολλαπλάσια και στα υποπολλαπλάσια του τέμπο), αφετέρου δίνει μεγαλύτερη έμφαση –όπου αυτό είναι δυνατό– στο αληθές τέμπο (Σχ. 2.13). Μια βασική προσέγγιση εξαγωγής του τέμπο από την ΣΠ θα ήταν να επιλέξουμε το τέμπο που αντιστοιχεί στην πιο εξέχουσα κορυφή, προσέγγιση που συναντάμε συχνά στη βιβλιογραφία [Scheirer1998] [Alonso2007]. Τέτοιου είδους μέθοδοι παρουσίασαν μια σχετική επιτυχία, ωστόσο τίποτα δεν εξασφαλίζει ότι η πιο εξέχουσα κορυφή στην ΣΠ θα αντιστοιχεί πάντοτε στο σωστό τέμπο. Επομένως, η απευθείας επιλογή της πιο εξέχουσας κορυφής δεν επαρκεί και απαιτείται μια πιο εκλεπτυσμένη επεξεργασία του διανύσματος περιοδικότητας.

Όπως αναφέρθηκε στην Ενότητα 1.3.2, οι μέθοδοι εξαγωγής τέμπο που συναντιούνται στη βιβλιογραφία μπορούν να κατηγοριοποιηθούν σε δύο κύριες οικογένειες. Σε τεχνικές που κάνουν απευθείας επεξεργασία του διανύσματος περιοδικότητας και σε τεχνικές που χρησιμοποιούν μηχανική μάθηση. Η πρώτη οικογένεια περιλαμβάνει συνήθως την επιλογή της πιο εξέχουσας κορυφής αφού πρώτα προηγηθεί κάποια επιπλέον επεξεργασία της συνάρτησης περιοδικότητας. Στο [Klapuri2006] οι συγγραφείς χρησιμοποίησαν εκ των προτέρων κατανομές του μουσικού τέμπο για να για να δώσουν βάρη στα διανύσματα περιοδικότητας. Στην εργασία [Dixon2001] οι συγγραφείς προτείνουν ομαδοποίηση (clustering) των IOI ακολουθούμενες από μια συνάρτηση βάρους. Κάθε ομάδα αντιστοιχεί σε ένα τέμπο στόχο και τα υποψήφια τέμπο κατατάσσονται βάσει του πλήθους των onsets. Η συνάρτηση βάρους αναθέτει τα onsets κάθε ομάδας και στις ομάδες με πολλαπλάσιο τέμπο με κάποιο βάρος. Άλλες μέθοδοι χρησιμοποιούν απλώς την επιλογή της πιο εξέχουσας κορυφής [Scheirer1998] [Alonso2007].

Οι μέθοδοι που χρησιμοποιούν τεχνικές μηχανικής μάθησης έχουν εμφανιστεί πρόσφατα στη βιβλιογραφία και επιδεικνύουν καλύτερη επίδοση. Ωστόσο, η έλλειψη αρκετά μεγάλων επισημειωμένων με τέμπο μουσικών συλλογών καθιστά την αξιολόγησή τους μη αξιόπιστη, αφού συνήθως οι μέθοδοι αυτές εκπαιδεύονται και αξιολογούνται σε διαφορετικά υποσύνολα της ίδιας όμως συλλογής. Οι συλλογές αναφοράς είναι συνήθως μικρές και πολύ συχνά με μεγάλη ομοιογένεια. Έτσι όταν ένα σύστημα εκπαιδευτεί σε ένα υποσύνολο μιας ομοιογενούς σχετικά συλλογής, θα επιτύχει υψηλά ποσοστά εύρεσης του τέμπο στα υπόλοιπα κομμάτια της συλλογής. Δεν έχει γίνει όμως αξιολόγηση αυτών των μεθόδων σε μεγάλης κλίμακας δεδομένα.

Στη παρούσα διατριβή αναπτύχθηκαν τρεις μέθοδοι εξαγωγής τέμπο. Η $1^{η}$ μέθοδος εξάγει το τέμπο απ' ευθείας από την ΣΠ, χωρίς καμία εξαγωγή χαρακτηριστικών και χωρίς χρήση τεχνικών Μηχανικής Μάθησης. Η $2^{η}$ μέθοδος περιλαμβάνει την εξαγωγή των χειρονακτικών χαρακτηριστικών που παρουσιάστηκαν στην Ενότητα 3.2, τα οποία χρησιμοποιήθηκαν για την

κατηγοριοποίηση της «μουσικής ταχύτητας» Η μουσική ταχύτητα στη συνέχεια συνδυάζεται με την ΣΠ για την εξαγωγή του τέμπο. Η $3^{η}$ μέθοδος αποτελεί συνδυασμό και επέκταση των άλλων δύο μεθόδων. Χρησιμοποιεί τα χαρακτηριστικά που εξάγονται με τις τεχνικές μη επιβλεπόμενης μάθησης που παρουσιάστηκαν στην Ενότητα 3.3. Στη συνέχεια η κατηγοριοποίηση μουσικής ταχύτητας της 2ης μεθόδου επεκτείνεται σε ένα σύνολο ταξινομητών οι οποίοι σχηματίζουν μια μάσκα της ΣΠ. Η μάσκα αυτή σε συνδυασμό με την ΣΠ χρησιμοποιούνται για την εξαγωγή του τέμπο. Οι τρεις επόμενες ενότητες περιγράφουν τις τρεις αυτές μεθόδους, ενώ η Ενότητα 5.5 παρουσιάζει αναλυτικά αποτελέσματα των 3 μεθόδων και σύγκρισή τους με τη διεθνή βιβλιογραφία.

5.2 Εξαγωγή Τέμπο με Χρήση Μετρικών Σχέσεων

Από την επισκόπηση των διανυσμάτων περιοδικότητας παρατηρούμε ότι οι σχετικές με τον ρυθμό περιοδικότητες είναι συνήθως (σχεδόν) ακέραια πολλαπλάσια κάποιας χαμηλής «θεμελιώδους» συχνότητας. Πράγματι, αν θεωρήσουμε τις ΣΠ που προκύπτουν από τα πρότυπα σήματα του Σχ. 2.13 (αληθές τέμπο 120 BPM) παρατηρούμε κορυφές στα υποπολλαπλάσια του 120 BPM, με πιο εξέχουσα την κορυφή στα 60 BPM. Αυτό συμβαίνει επειδή τα πρότυπα σήματα είναι διπλού μέτρου. Σε περιπτώσεις τριπλού μέτρου (π.χ. 3/4), η πιο εξέχουσα κορυφή θα ήταν στα 40 BPM.

Η προτεινόμενη μέθοδος δεν χρησιμοποιεί τις επιμέρους ΣΠ, αλλά τις συμπτύσσει σε μία τελική ΣΠ όπως περιγράφηκε στην Ενότητα 2.7 (Εξ. 2.34). Περιλαμβάνει 3 βήματα επεξεργασίας της ΣΠ. Το πρώτο περιλαμβάνει τον υπολογισμό της «θεμελιώδους περιοδικότητας». Αν $v[T]$ είναι το τελικό διάνυσμα περιοδικότητας, ορίζουμε τη «θεμελιώδη περιοδικότητα» T_0 ως

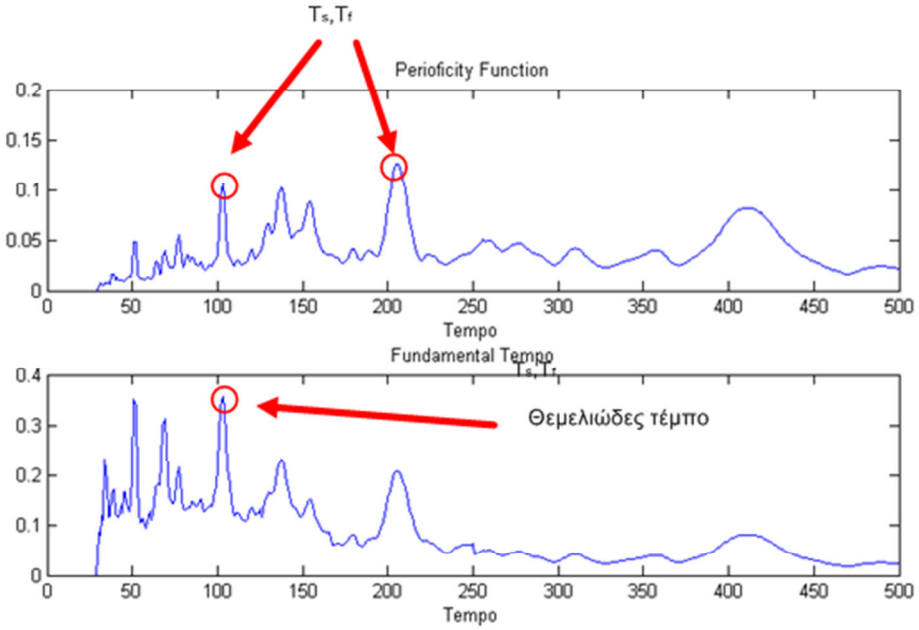
$$T_0 = \text{argmax}_T\{\sum_k v[kT]\,/\max(k)\}, k = 1..4 \tag{5.1}$$

Η Εξ. 5.1, εξ ορισμού υποδεικνύει ότι το $v[T]$ θα έχει κορυφές στα πολλαπλάσια του T_0. Βασική παραδοχή της προτεινόμενης μεθόδου είναι ότι το πραγματικό τέμπο θα είναι πολλαπλάσιο του T_0. Ωστόσο, η απευθείας επιλογή του πολλαπλασίου με το μεγαλύτερο πλάτος ως αληθές τέμπο δεν επαρκεί, αφού ο περιορισμός του διαστήματος αναζήτησης του τέμπο στα πολλαπλάσια $\{kT_0\}_k$ απλά θα παραβλέψει τις μη σχετικές με τον ρυθμό περιοδικότητες (εξ ορισμού οι πιο πολλές κορυφές θα επιτυγχάνονται στα $\{kT_0\}_k$). Ο όρος $\max(k)$ εισήχθη για κανονικοποίηση ως προς το πλήθος των όρων του αθροίσματος. Αφού το πεδίο ορισμού του $v[T]$ είναι περιορισμένο στο διάστημα $[T_{min}, T_{max}]$, το πλήθος των όρων που αθροίζονται στη ποσότητα $\sum_k v[kT]$ δεν θα είναι ο ίδιος για όλα τα T.

Το επόμενο βήμα της επεξεργασίας περιλαμβάνει τον ορισμό δύο τέμπο, ενός αργού T_{slow} και ενός γρήγορου T_{fast}. Το T_{slow} θεωρείται ως το πιο σημαντικό και διαισθητικά σχετικό, ενώ το T_{fast} αναμένουμε να είναι διπλάσιο, τριπλάσιο ή τετραπλάσιο του T_{slow}. Ορίζουμε την *αμοιβαία ισχύ* δύο τέμπο T_1 και T_2 με $T_2 > T_1$ ως:

$$J(T_1, T_2) = (v[T_1] + v[T_2]) \cdot \sum_{l=2}^{4} exp\left\{\frac{(T_2/T_1 - l)^2}{(\rho l)^2}\right\}, T_2 > T_1 \tag{5.2}$$

Η ποσότητα J είναι ανάλογη των πλατών της ΣΠ στα T_1 και T_2, και αυξάνει όσο το T_2 είναι διπλάσιο, τριπλάσιο ή τετραπλάσιο του T_1.



Σχήμα 5.1: Πάνω: Η συνάρτηση περιοδικότητας για ένα μουσικό κομμάτι με τέμπο 102 BPM. Παρατηρούμε ότι η πιο εξέχουσα κορυφή εμφανίζεται στο διπλάσιο τέμπο. Κάτω: Η συνάρτηση του θεμελιώδους τέμπο $\sum_k v[kT], k = 1..4$, που εμφανίζει μέγιστο στα 102 BPM (Εξ. 5.1). Τα δύο τέμπο που μεγιστοποιούν την ποσότητα $J(T_1, T_2)$ είναι τα 102 και 204 BPM. Ως τελικό τέμπο επιλέγεται το $T_{slow} = 102$ BPM.

Δηλαδή η J παίρνει μεγάλες τιμές για ζεύγη τέμπο τα οποία α) είναι ισχυρά στην ΣΠ και β) έχουν μεταξύ τους μια μετρική σχέση.

Το τελευταίο βήμα είναι ο υπολογισμός των T_{slow} και T_{fast} τα οποία υπολογίζονται ως τα πολλαπλάσια του θεμελιώδους τέμπο T_0 που μεγιστοποιούν την J:

$$T_{slow}, T_{fast} = \text{argmax}_{iT_0, kT_0}\{J(iT_0, kT_0)\}, iT_0, kT_0 \in [T_{min}, T_{max}], i < k, \qquad (5.3)$$

Τέλος, εξ ορισμού επιλέγουμε ως αληθές τέμπο το T_{slow}.

Η επιλογή $i, k \leq N = 4$ στις Εξ. 5.1 και 5.2 έχει δύο βασικούς σκοπούς. Στην (5.1) περιορίζει το T_0 σε εύρος τιμών $\leq T_{max}/N$, ενώ στην (5.2) μπαίνει έμμεσα ο περιορισμός ότι τα πολλαπλάσια (αρμονικές) του πραγματικού τέμπο δεν μπορούν να εκτείνονται πέρα από το T_{max} και ότι η μεγαλύτερη αρμονική μπορεί να είναι το πολύ τετραπλάσια του T_0.

Παρότι η προτεινόμενη μέθοδος αποτελεί μια ευριστική μέθοδο, επιτυγχάνει πολύ καλά αποτελέσματα όπως θα δούμε στην Ενότητα 5.5. Η διαίσθηση πίσω από τις Εξ. (5.1)-(5.3) θα παρουσιαστεί με ένα παράδειγμα. Ας θεωρήσουμε το διάνυσμα περιοδικότητας του Σχ. 5.1 (πάνω), ενός κομματιού με μουσικό τέμπο 102 BPM. Στην ΣΠ η πιο εξέχουσα κορυφή είναι στα 204 BPM και αντιστοιχεί στο διπλάσιο τέμπο από το πραγματικό. Αποτελεί ένα τυπικό παράδειγμα που μπορεί να οδηγήσει σε σύγχυση οκτάβας. Η αμέσως επόμενη σε πλάτος κορυφή εμφανίζεται στο πραγματικό τέμπο 102 BPM. Κορυφές επίσης εμφανίζονται στα 51 BPM, 137 BPM ($\sim 4/3 \cdot 102$), 77 BPM ($\sim 3/4 \cdot 102$), και 155 BPM ($\sim 3/2 \cdot 102$), που όλα αποτελούν ρυθμικά σχετικά κλάσματα του 102 BPM. Στο ίδιο σχήμα (κάτω) παρουσιάζεται η γραφική παράσταση της ποσότητας $\sum_k v[kT], k = 1..4$ ως προς το T, της οποίας το μέγιστο μας δίνει και το θεμελιώδες τέμπο. Τα

πολλαπλάσια του 102 BPM επιτυγχάνουν το μεγαλύτερο άθροισμα, επομένως $T_0 = 102$ BPM. Στην συνέχεια τα πολλαπλάσια του T_0 που μεγιστοποιούν το $J(T_1, T_2)$ (Εξ. 5.2) είναι τα $T_{slow} = 102$ BPM και $T_{fast} = 204$ BPM. Ως τελικό τέμπο θεωρείται το $T_{slow} = 102$ BPM. Η μέθοδος επιδεικνύει μια σχετικά ανθεκτικότητα ως προς το T_0. Στην περίπτωση μιας πολύ παρόμοιας ΣΠ με το ίδιο αληθές τέμπο, στην οποία θα υπολογιζόταν $T_0 = 51$ BPM (είναι πολύ κοντά οι τιμές των 51, 102 BPM), πάλι η έξοδος της μετρικής ανάλυσης θα ήταν $T_{slow} = 102$ BPM και $T_{fast} = 204$ BPM.

5.3 Εξαγωγή Τέμπο από την Μουσική Ταχύτητα με Χειρωνακτικά Χαρακτηριστικά

Όπως αναφέρθηκε στην Ενότητα 1.3, οι περισσότερες μέθοδοι εξαγωγής τέμπο υποφέρουν από τα λεγόμενα «λάθη οκτάβας» (octave errors), δηλαδή το αποτέλεσμά τους είναι συνήθως κάποιο (υπο)πολλαπλάσιο του πραγματικού τέμπο (διαφορετικό μετρικό επίπεδο). Παρότι έχουν παρουσιαστεί αρκετές μέθοδοι [Peeters2007], [Klapuri2006] οι οποίες παρουσιάζουν ακρίβεια μεγαλύτερη του 90% όταν τα λάθη οκτάβας θεωρούνται σωστά (μετρική αξιολόγησης *accuracy2*), η ακρίβεια αυτών των μεθόδων πέφτει στο 50~60% για την ακριβή τιμή του τέμπο (μετρική *accuracy1*). Η οκτάβα στην οποία βρίσκεται το εξαγόμενο τέμπο συναντάται και με τον όρο μετρικό επίπεδο (metrical level).

Όταν η έξοδος ενός αλγόριθμου αξιολογείται σωστά όταν κάνει λάθη οκτάβας διαφαίνονται δύο αντικρουόμενες παρατηρήσεις. Θα μπορούσαμε να πούμε ότι η αξιολόγηση ενός αλγόριθμου με την μετρική *accuracy2* είναι ίσως πιο κοντά στην έννοια του υποκειμενικού τρόπου αντίληψης του ρυθμού. Δύο διαφορετικοί ακροατές μπορεί να αντιληφθούν το τέμπο σε διαφορετικό μετρικό επίπεδο για το ίδιο μουσικό απόσπασμα. Ακόμα και ο ίδιος ακροατής ανάλογα με την φυσική ή ψυχική του κατάσταση ή τις συνθήκες στο εξωτερικό περιβάλλον μπορεί να επιλέξει άλλο μετρικό επίπεδο για το ίδιο μουσικό κομμάτι. Έτσι η επιλογή της σωστής οκτάβας δεν είναι πάντα κρίσιμη στην εξαγωγή του τέμπο. Από την άλλη πλευρά όμως, δεν μπορούν να θεωρηθούν όλα τα κλάσματα και πολλαπλάσια του αληθινού τέμπο σωστά. Για παράδειγμα για ένα κομμάτι σε μέτρο 4/4 και τέμπο 100 BPM, τυχόν κορυφές που μπορεί να εμφανιστούν στη συνάρτηση περιοδικότητας για τις τιμές 66 (100x2/3), 150 (100x3/2) ή 300 (100x3) BPM είναι ρυθμικά και αντιληπτικά μη σχετικές με τον ρυθμό, αφού αντιστοιχούν σε τριπλό μέτρο. Αντιθέτως, τα κλάσματα 1/2 και 2/1 (50, 200 BPM) είναι ρυθμικά σχετικά, όπως φαίνεται και στο παράδειγμα του Σχ. 5.1). Συνοψίζοντας, θα μπορούσαμε να πούμε ότι ένα μετρικό επίπεδο είναι αντιληπτικά σωστό, αν ο αντίστοιχος μουσικός παλμός (υπό την προϋπόθεση ότι έχει σωστή φάση, Ενότητα 1.2.4), δίνει στον ακροατή την αίσθηση του σωστού ρυθμού. Εδώ πρέπει να σημειωθεί ότι, ακόμα και αν δύο άνθρωποι αντιλαμβάνονται διαφορετικό μετρικό επίπεδο ως βασικό στο ίδιο κομμάτι, θα συμφωνούσαν στο αν ένα μετρικό επίπεδο είναι ρυθμικά σχετικό (ακόμα και αν δεν είναι το μετρικό επίπεδο που επέλεξαν).

Ένας τρόπος συμβιβασμού της παραπάνω αντίφασης (ότι κάποια πολλαπλάσια είναι σωστά, αλλά δεν είναι όλα σωστά) ήταν ο ορισμός την μετρικής P-score που προτάθηκε για πρώτη φορά στον διαγωνισμό αυτόματης εξαγωγής τέμπο MIREX 2005. Κάθε μουσικό απόσπασμα επισημειώθηκε με ένα τέμπο από 40 διαφορετικούς ακροατές. Τα δύο τέμπο με τις περισσότερες επισημειώσεις θεωρήθηκαν ως σωστά για κάθε κομμάτι, και για κάθε ένα από αυτά ανατέθηκε ένα σχετικό βάρος ανάλογο με το πλήθος των επισημειώσεων,

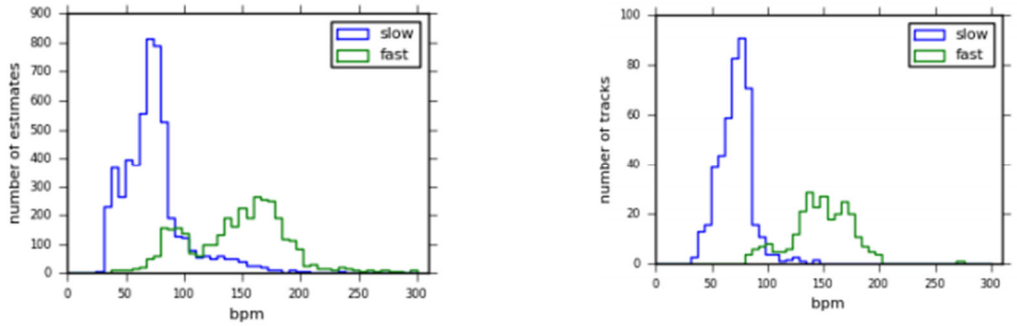
ώστε το άθροισμα των δύο βαρών να είναι ίσο με την μονάδα. Οι αλγόριθμοι που αξιολογήθηκαν έπρεπε να εξάγουν δύο τέμπο. Το P-score υπολογίζεται ως το μέσο άθροισμα των βαρών των σωστών τέμπο που βρέθηκαν με ένα περιθώριο ανοχής 8%. Με αυτή την μετρική αξιολόγησης οι αλγόριθμοι «απαλλάχθηκαν» σε μεγάλο βαθμό από το πρόβλημα εύρεσης του σωστού μετρικού επιπέδου.

Ωστόσο το πρόβλημα της επιλογής μετρικού επιπέδου και η ανθρώπινη υποκειμενικότητα σε αυτήν την επιλογή φέρνει στην επιφάνεια μία επιπλέον πτυχή του μουσικού ρυθμού: την μουσική ταχύτητα. Παρόλο που μπορεί να υπάρχει διαφορά στην προτίμηση δύο ακροατών για το ποιο είναι το βασικό μετρικό επίπεδο ενός τραγουδιού, δεν συμβαίνει το ίδιο και με την ταχύτητα. Δύο ακροατές σχεδόν πάντα θα συμφωνούν για το αν ένα κομμάτι είναι γρήγορο ή αργό, ενώ όταν διαφωνούν, αυτό θα συμβαίνει σε κομμάτια που είναι ενδιάμεσα. Επομένως, αν θεωρήσουμε μια επιπλέον τρίτη κλάση, αυτή των μουσικών κομματιών που είναι μεταξύ αργού και γρήγορου, η όποια διαφωνία θα συμβαίνει σχεδόν πάντα μεταξύ της μεσαίας κατηγορίας και των άλλων δύο.

Υπάρχουν αρκετές ερευνητικές εργασίες που μελετούν την απόκριση του ανθρώπου στις ρυθμικές περιοδικότητες είτε από ψυχοακουστική, είτε από στατιστική σκοπιά. Οι van Noorden και Molentants [Noorden1999] παρατήρησαν ότι υπάρχει κάποιο κατώφλι τέμπο, πέρα από το οποίο οι άνθρωποι δεν μπορούν να το ακολουθήσουν, με αποτέλεσμα να επιλέγουν μία οκτάβα χαμηλότερα. Κάτι ανάλογο ισχύει και για τα πολύ αργά τέμπο, όπου οι ακροατές συνήθιζαν να χτυπάνε το πόδι τους σε διπλάσιο ή τριπλάσιο τέμπο κάτω από ένα κατώφλι. Αντίστοιχη παρατήρηση έκανε ο M. Levy [Levy2011], στις αποκρίσεις των χρηστών σε ένα παιχνίδι στην μουσική ιστοσελίδα LastFm[4]. Οι χρήστες καλούνταν να πατάνε ένα πλήκτρο του υπολογιστή ακολουθώντας τον παλμό των μουσικών κομματιών που άκουγαν. Στη συνέχεια απαντούσαν στην ερώτηση αν το συγκεκριμένο κομμάτι είναι αργό, γρήγορο ή ενδιάμεσο. Συλλέγοντας δεδομένα από 6.000 περίπου χρήστες και για 4.000 μουσικά κομμάτια, παρουσιάστηκε η κατανομή των εξαγόμενων τέμπο από τους χρήστες για κάθε μία από τις τρεις κλάσεις. Στο Σχήμα 5.2α παρουσιάζεται αυτή η κατανομή για κομμάτια που επισημειώθηκαν από πέντε τουλάχιστον χρήστες, ενώ στο Σχήμα 5.2β παρουσιάζεται η κατανομή όταν για κάθε κομμάτι, λαμβάνεται υπόψη ο μέσος των επισημειώσεων. Παρατηρούμε ότι (Σχ. 5.2α) στην κατηγορία των γρήγορων κομματιών, εμφανίζεται ένα μέγιστο στα τέμπο 80~100 BPM γεγονός που –όπως ερμηνεύεται και από τον συγγραφέα– οφείλεται στην τάση των χρηστών να χτυπούν στο μισό τέμπο για πολύ γρήγορα κομμάτια. Αυτό έρχεται και σε συμφωνία με την παρατήρηση των Noorden και Molentants [Noorden1999] που προαναφέρθηκε. Όταν όμως λαμβάνεται ο μέσος όλων των επισημειώσεων το φαινόμενο αυτό εξαλείφεται, δηλαδή για τη πλειοψηφία των ακροατών το πάτημα του πλήκτρου συμφωνούσε με την αίσθηση αργού / γρήγορου.

Παρότι η μουσική ταχύτητα είναι πολύ σημαντική έννοια, μόλις πρόσφατα έχουν παρουσιαστεί εργασίες που πραγματεύονται τον αυτόματο καθορισμό της μουσικής ταχύτητας. Στο άρθρο [Eronen2010] παρουσιάστηκε και ένα υποσύστημα καθορισμού κατηγοριοποίησης της μουσικής ταχύτητας σε αργό, γρήγορο και μέσο, η οποία όμως υπολογίζεται μέσω του εξαχθέντος τέμπο. Οι Hockman και Fujinaja [Hockman2010] πρότειναν ένα σύστημα που κατηγοριοποιεί τα μουσικά κομμάτια σε αργά/γρήγορα. Οι κατηγορίες για τα κομμάτια εκμάθησης / κατηγοριοποίησης δεν εξάχθηκαν από το τέμπο, αλλά από ετικέτες (tags) χρηστών στο YouTube.

[4] www.last.fm



Σχήμα 5.2[5]: α) αριστερά: κατανομές των τέμπο στις κατηγορίες 'αργό' – γρήγορο' για μουσικά αποσπάσματα που αξιολογήθηκαν από τουλάχιστον 5 χρήστες. β) δεξιά: η ίδια κατανομή όταν λαμβάνεται ο μέσος των επισημειώσεων για κάθε μουσικό απόσπασμα.

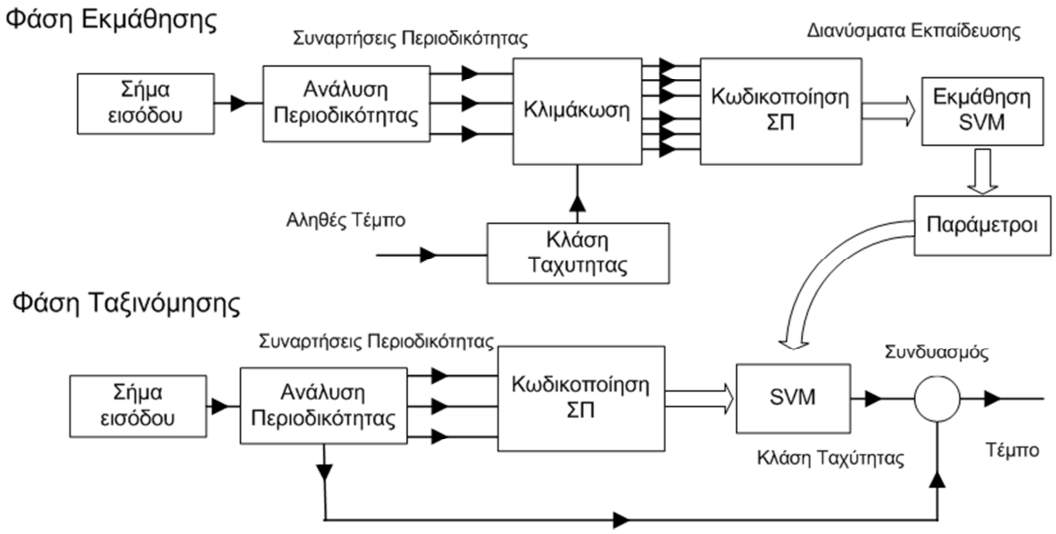
Χρησιμοποιώντας αποκλειστικά χαρακτηριστικά σε επίπεδο παραθύρου (frame) και χωρίς καμία ρυθμική ανάλυση, πέτυχαν 96% αναγνώριση με την χρήση της μεθόδου AdaBoost. Εξίσου υψηλά ποσοστά ανέφεραν και για άλλους ταξινομητές, όπως για παράδειγμα τα SVM (~95%). Στο [Smith2010] ο συγγραφέας παρουσίασε μια ευριστική μέθοδο για την ανίχνευση των λαθών οκτάβας ενός beat-tracker αναφοράς.

Το γεγονός ότι η αντίληψη της μουσικής ταχύτητας συνδέεται με την επιλογή της οκτάβας καθώς και το ότι η μέθοδος εξαγωγής τέμπο με τη χρήση μετρικών σχέσεων και παραλλαγές αυτής ([Gkiokas2010], [Gkiokas2012a]) επιτυγχάνουν εξαιρετικά υψηλά ποσοστά αναγνώρισης βάσει της μετρικής *acc2* (>90%), μας οδηγεί στην ανάπτυξη ενός συστήματος εξαγωγής της κατηγορίας της μουσικής ταχύτητας [Gkiokas2012b]. Έχοντας γνώση της κατηγορίας αυτής, περιορίζουμε την συνάρτηση περιοδικότητας στο εύρος των τέμπο της κλάσης και αναμένουμε να αυξηθεί η επίδοση του συστήματος βάσει της *acc1*.

Η προτεινόμενη μέθοδος συνίσταται σε τέσσερα βασικά μέρη/υποσυστήματα, όπως παρουσιάζονται στο Σχήμα 5.3. Ένα σύστημα ανάλυσης περιοδικότητας, ένα σύστημα εξαγωγής χαρακτηριστικών από το διάνυσμα περιοδικότητας, ένα σύστημα κλιμάκωσης ή επαναδειγματοληψίας και ένα σύστημα εκμάθησης / κατηγοριοποίησης.

Για το σύστημα ανάλυσης περιοδικότητας χρησιμοποιούμε την ΣΠ που περιγράφτηκε στο Κεφ. 2, χωρίς καμία σύμπτυξη των επιμέρους ΣΠ (Εξ. 2.30). Στην συνέχεια λαμβάνει χώρα η εξαγωγή των «χειρονακτικών» χαρακτηριστικών (Ενότητα 3.2) που περιλαμβάνει την κλιμάκωση και ανάλυση σε μπάντες της ΣΠ. Τέλος, το στάδιο εκμάθησης περιλαμβάνει τον διαχωρισμό τριών κλάσεων σχετικές με την ταχύτητα του κομματιού: αργό, μεσαίο και γρήγορο. Στο στάδιο της κατηγοριοποίησης, κάθε μουσικό απόσπασμα ανατίθεται σε μία από τις τρεις κλάσεις και στη συνέχεια το συμπτυγμένο διάνυσμα περιοδικότητας (Εξ. 2.34) περιορίζεται στα τέμπο που αντιστοιχούν στην κλάση αυτή. Το τελικό τέμπο αντιστοιχεί στην πιο εξέχουσα κορυφή στο περιορισμένο διάνυσμα περιοδικότητας.

[5] Σχήμα από το [Levy2011]



Σχήμα 5.3. Σύνοψη του συστήματος επιβλεπόμενης εξαγωγής τέμπο.

Για την εκμάθηση / κατηγοριοποίηση υιοθετήσαμε δύο μεθόδους. Την μέθοδο ενός SVM ταξινομητή και την κατηγοριοποίηση με προσέγγιση του τέμπο με συνεχή τρόπο, υιοθετώντας ένα SVM παλινδρόμησης (regression). Έστω $\{(\mathbf{m}_l, T_l), l = 1..L\}$ τα κωδικοποιημένα διανύσματα εκπαίδευσης $\mathbf{m}_l$ (Εξ. 3.6) μαζί με το επισημειωμένο τέμπο T_l. Όταν δεν είναι απευθείας διαθέσιμη η πληροφορία της κατηγορίας c_l όπως συμβαίνει στα δεδομένα [Levy2011], αυτή εξάγεται απευθείας από το αληθές τέμπο ως

$$c_l = c(T_l) = \begin{cases} 3, & T_l \leq T_{slow} \\ 2, & T_{slow} \leq T_l \leq T_{fast} \\ 3, & T_l \geq T_{fast} \end{cases}. \quad (5.4)$$

Τα κατώφλια T_{slow}, T_{fast} μπορούν είτε να εξαχθούν από τα ίδια τα δεδομένα είτε να οριστούν εξαρχής και χωρίζουν τα δεδομένα εκμάθησης στις τρεις κατηγορίες. Για την εκμάθηση και κατηγοριοποίηση των μουσικών αποσπασμάτων υιοθετήθηκαν δύο εκδόσεις των Μηχανών Διανυσμάτων Υποστήριξης (Support Vector Machines, SVM). Τα SVM κατηγοριοποίησης (Classification SVM) και τα SVM παλινδρόμησης (Regression SVM). Τα SVM κατηγοριοποίησης εκπαιδεύονται στην πρόβλεψη διακριτών κατηγοριών, ενώ τα SVM παλινδρόμησης προβλέπουν μια συνεχή τιμή.

Το πρόβλημα της κατηγοριοποίησης στην περίπτωση του Classification SVM για τα δεδομένα $\{(\mathbf{m}_l, c_l), l = 1..L\}$, ορίζεται ως επιμέρους προβλήματα κατηγοριοποίησης δύο κατηγοριών (one vs. one strategy). Υπάρχουν ενδείξεις [Hsu2002] ότι αυτή η στρατηγική είναι πιο αποτελεσματική από τη συνήθη «one vs. all» όπου εκπαιδεύεται ένα SVM για κάθε κατηγορία, ειδικά στις περιπτώσεις όπου υπάρχει ανισορροπία στο μέγεθος των κατηγοριών. Στην περίπτωση των SVM παλινδρόμησης τα δεδομένα εκμάθησης είναι επισημειωμένα απευθείας με το αληθές τέμπο και όχι με την κατηγορία ταχύτητας, δηλαδή $D = \{(\mathbf{m}_l, T_l), l = 1..L\}$. Το πρόβλημα βελτιστοποίησης στην περίπτωση αυτή περιγράφεται ως η ελαχιστοποίηση της συνάρτησης κόστους

$$\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{l=1}^{L}(\xi_l + \xi_l^*)) \quad (5.5)$$

σύμφωνα με τους περιορισμούς

$$-(\varepsilon + \xi_l^*) \leq \mathbf{w}^T \varphi(\mathbf{m}_l) + b - T_l \leq \varepsilon + \xi_l,\ \xi_l, \xi_l^* \geq 0 \tag{5.6}$$

Στην φάση πρόβλεψης στο άγνωστο διάνυσμα $\hat{\mathbf{m}}$ ανατίθεται μια εκτίμηση του τέμπο συνεχούς τιμής $\hat{T}$ ως

$$\hat{T} = \mathbf{w}^T \varphi(\mathbf{m}) + b \tag{5.7}$$

Η τιμή αυτή δεν μπορεί να θεωρηθεί ως μια ακριβής εκτίμηση του τέμπο του μουσικού κομματιού, καθώς στην αναπαράστασή του μέσω της κωδικοποιημένης ΣΠ $\hat{\mathbf{m}}$ αρκετή πληροφορία που έχει να κάνει με λεπτομέρειες της ΣΠ, όπως για παράδειγμα η ακριβής θέση των κορυφών, έχει χαθεί. Ωστόσο, η τιμή $\hat{T}$ μας δίνει μια προσέγγιση του τελικού τέμπο, που είναι ικανή να μας δώσει πληροφορία για την ταχύτητα του μουσικού κομματιού. Η κατηγορία της ταχύτητας $\hat{c}$ για το $\hat{\mathbf{m}}$ αποφασίζεται μέσω της Εξ. 5.4.

Τέλος, το τέμπο του κομματιού αποφασίζεται από τη συμπτυγμένη ΣΠ $\boldsymbol{v}[T]$ (Εξ. 2.34) σε συνδυασμό με την κατηγορία της ταχύτητας $\hat{c}$:

$$T = \text{argmax}_T\{\boldsymbol{v}[T] \cdot I_c[T]\} \tag{5.8}$$

όπου $I_c[T]$ είναι η συνάρτηση ένδειξης (Indicator function) εάν το τέμπο ανήκει στην κατηγορία c.

5.4 Εξαγωγή Τέμπο ως Πολλαπλά Προβλήματα Κατηγοριοποίησης με Αυτόματα Χαρακτηριστικά

Μία εύλογη προσέγγιση της εύρεσης του τέμπο από την ΣΠ με την χρήση τεχνικών Μηχανικής Μάθησης, θα μπορούσε να είναι η διατύπωσή του ως ένα πρόβλημα παλινδρόμησης, δηλαδή της εύρεσης μιας συνεχούς συνάρτησης $g : \mathbf{V} \rightarrow T$, όπου $\boldsymbol{V}$ είναι ο διανυσματικός χώρος των ΣΠ, και T είναι το πεδίο τιμών του τέμπο. Τέτοιου είδους προσέγγιση δεν θα μπορούσε να εφαρμοστεί στα χαρακτηριστικά της προηγούμενης μεθόδου, τα οποία σχεδιάστηκαν για την κατηγοριοποίηση της μουσικής ταχύτητας, αλλά μόνο απευθείας στην ΣΠ. Ο κύριος λόγος είναι όπως προαναφέρθηκε ότι στα χαρακτηριστικά αυτά έχει χαθεί χρήσιμη πληροφορία, όπως για παράδειγμα οι ακριβείς θέσεις των κορυφών της ΣΠ. Ωστόσο, ακόμα με την αρχική ΣΠ, η διατύπωση της εύρεσης του τέμπο ως ένα πρόβλημα παλινδρόμησης έχει μια εγγενή αδυναμία, την οποία θα αναδείξουμε με ένα απλό παράδειγμα. Έστω $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ οι ΣΠ τριών μουσικών κομματιών με πολύ κοντινό ρυθμικό περιεχόμενο, τα οποία έχουν επισημειωθεί με τέμπο ίσο με 100, 100 και 200 BPM αντίστοιχα. Στην περίπτωση που τα τρία κομμάτια είναι πολύ δύσκολο να διαχωριστούν λόγω των πολύ κοντινών ΣΠ, μια μέθοδος παλινδρόμησης πιθανόν να υπολογίσει μια τιμή ίση με 133 BPM προκειμένου να ελαχιστοποιήσει το μέσο τετραγωνικό σφάλμα. Ένα τέτοιο αποτέλεσμα θα έδινε 0% ακρίβεια στην πρόβλεψη. Αντίθετα, αν το ίδιο πρόβλημα οριστεί ως πρόβλημα κατηγοριοποίησης σε δύο κλάσεις «μικρότερο» και «μεγαλύτερο» από 150 BPM, τότε ένας ταξινομητής θα αποφανθεί και για τα τρία κομμάτια ότι έχουν τέμπο μικρότερο του 150 BPM. Ο ταξινομητής σε αυτή τη περίπτωση θα είχε 67% ακρίβεια. Το αίτιο που προκαλεί αυτό το φαινόμενο είναι η φύση του ίδιου του προβλήματος. Δύο αποσπάσματα με πολύ παρόμοιες ΣΠ, μπορεί να έχουν μεγάλη διαφορά στο τέμπο (π.χ. διπλάσιο ή μισό). Κάτι τέτοιο κάνει τις συναρτήσεις $g : \mathbf{V} \rightarrow T$ να εμφανίζουν πολύ μεγάλες μεταβολές στον χώρο T, για πολύ μικρές μεταβολές στον χώρο $\mathbf{V}$. Έτσι είναι πολύ δύσκολο να βρεθεί μία αξιόπιστη συνάρτηση g.

Αντίθετα, η κατηγοριοποίηση σε «κλάσεις τέμπο» φαίνεται να μειώνει αυτό

το φαινόμενο, το οποίο επαληθεύεται πειραματικά στην επόμενη Ενότητα. Βάσει αυτής της παρατήρησης, προτάθηκε η μέθοδος της Ενότητας 5.3, η οποία περιλαμβάνει και ένα επιπλέον χαρακτηριστικό, αυτό της χρησιμοποίησης των χειρονακτικών χαρακτηριστικών από την ΣΠ. Όπως θα δειχτεί και στην επόμενη Ενότητα, τα αποτελέσματα αυτής της μεθόδου είναι ανταγωνιστικά της διεθνούς βιβλιογραφίας. Ωστόσο έχει δύο βασικά μειονεκτήματα. Το 1^o είναι ότι όλα τα λάθη στην κατηγοριοποίηση μουσικής ταχύτητας διαδίδονται στην επιλογή του τέμπο. Το 2^o είναι ότι τα περισσότερα λάθη γίνονται για κομμάτια που το τέμπο τους είναι κοντά στα κατώφλια T_{slow} και T_{fast}. Επιπλέον, τα T_{slow} και T_{fast} επιλέγονται αυθαίρετα.

Η μέθοδος που θα παρουσιαστεί σε αυτή τη Ενότητα αποτελεί επέκταση και γενίκευση της προηγούμενης μεθόδου και ξεπερνάει τους περιορισμούς και τα προβλήματά της. Η $1^η$ επέκταση είναι η χρήση τεχνικών μηχανικής μάθησης για την εξαγωγή χαρακτηριστικών από την ΣΠ, αντί των χειρονακτικών χαρακτηριστικών. Συγκεκριμένα, χρησιμοποιήσαμε τα χαρακτηριστικά που προέκυψαν με τα RBM και τα PCA, wPCA που περιγράφηκαν στο Κεφ. 3 με τον ίδιο τρόπο και τις ίδιες παραμέτρους που χρησιμοποιήθηκαν στο πρόβλημα της ρυθμικής κατηγοριοποίησης (Κεφ. 4). Το 2^o χαρακτηριστικό αυτή της μεθόδου είναι ότι, αντί να ορίσει τρεις κλάσεις ταχύτητας, χρησιμοποιεί πολλούς ταξινομητές, με δύο μόνο κλάσεις ανά ταξινομητή.

Το εύρος των τέμπο στόχων $[T_{\min}, T_{\max}]$ χωρίζεται από μία ακολουθία k ενδιάμεσων τιμών τέμπο $\{T_k\}, T_k < T_{k+1}$. Στη συνέχεια, για κάθε T_k, ορίζεται ένας δυαδικός ταξονομητής C_k τέτοιες ώστε $C_k = 1$ αν το τέμπο ενός κομματιού είναι μεγαλύτερο από το T_k και $C_k = 0$ αν είναι μικρότερο ή ίσο με το T_k. Στη συνέχεια, για κάθε ταξινομητή C_k, δημιουργείται μία μάσκα τέμπο $M_k(T)$, τέτοια ώστε:

$$M_k(T) = C_k,\ T > T_k \text{ και } M_k(T) = 1 - C_k,\ T \le T_k. \tag{5.9}$$

Δηλ. κάθε ταξινομητής αποφασίζει εάν το τέμπο ανήκει σε ένα διάστημα ($T > T_k$ ή $T \le T_k$) και η $M_k(T)$ είναι ίση 1 σε αυτό το διάστημα απόφασης. Οι επιμέρους μάσκες στη συνέχεια αθροίζονται για τον σχηματισμό μας ολικής μάσκας

$$M(T) = \sum_k M_k(T). \tag{5.10}$$

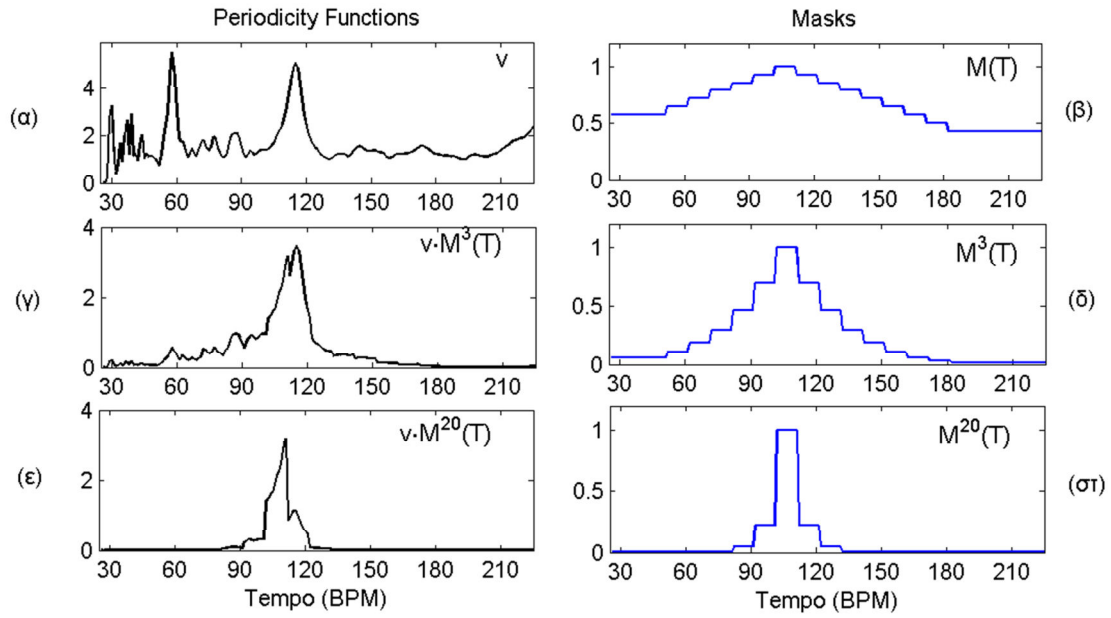
Η $M(T)$ στη συνέχεια κανονικοποιείται με τη μέγιστη τιμή της

$$M(T) \leftarrow M(T)/\max(M(T)). \tag{5.11}$$

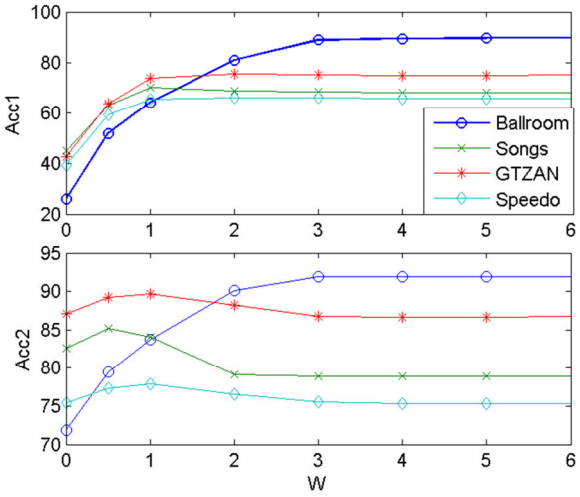
Η $M(T)$ μπορεί να θεωρηθεί ως μία συνάρτηση βάρους της ΣΠ, η οποία παρέχει μία γενικότερη εκτίμηση του τέμπο. Επομένως, εφαρμόζεται στην συμπτυγμένη ΣΠ $\boldsymbol{v}[T]$ (Εξ. 2.34) ως

$$\boldsymbol{v}^*[T] = M^W[T] \cdot \boldsymbol{v}[T]. \tag{5.12}$$

Τέλος, η πιο εξέχουσα κορυφή της $\boldsymbol{v}^*[T]$ μπορεί να επιλεχτεί ως το σωστό τέμπο. Ο εκθέτης $W \ge 0$ ρυθμίζει την επίδραση που θα έχει η $M(T)$ στην ΣΠ. Αν $W = 0$ τότε έχουμε $v^*[T] = v[T]$ και όσο το W αυξάνεται, η $M^W[T]$ γίνεται πιο «αιχμηρή», και στο όριο γίνεται μια δυαδική μάσκα. Στο Σχ. 5.4 παρουσιάζεται η διαδικασία υπολογισμού της μάσκας $M(T)$ και η εφαρμογή της στην ΣΠ. Το αληθές τέμπο του κομματιού είναι 115 BPM. Το Σχ. 5.4α δείχνει την συμπτυγμένη ΣΠ. Πέρα από τα 115 BPM, η ΣΠ παρουσιάζει μία κορυφή στο μισό τέμπο (57 BPM), η οποία είναι ισχυρότερη από την κορυφή στα 115 BPM.



Σχήμα 5.4. Σύνοψη του συστήματος επιβλεπόμενης εξαγωγής τέμπο.



Σχήμα 5.5. *acc1* και *acc2* σε τέσσερις συλλογές ως προς τον εκθέτη W (Εξ. 512).

Στο Σχ. 5.4β παρουσιάζεται η τελική μάσκα $M(T)$, υπολογισμένη από τις επιμέρους μάσκες στα τέμπο $T_k = 50 + (k-1)10, T_k \leq 180$ BPM. Η κορυφή της μάσκας είναι στο εύρος των τέμπο [100, 110] BPM, το οποίο δεν συμπίπτει με το αληθές τέμπο. Στα Σχ. 5.4δ φαίνεται η μάσκα υψωμένη στη δύναμη $W = 3$ και στο Σχ. 5.4γ η αντίστοιχη ΣΠ. Παρατηρούμε ότι μετά την εφαρμογή της μάσκας $M^3(T)$ η πιο εξέχουσα κορυφή της ΣΠ παρουσιάζεται στο σωστό τέμπο. Επομένως σε αυτή τη περίπτωση μια προσέγγιση επιλογής της πιο εξέχουσας κορυφής επαρκεί για την εύρεση του σωστού τέμπο. Ωστόσο, αν το W είναι πολύ μεγάλο ($W = 20$, Σχ. 5.4στ) το λάθος της μάσκας επηρεάζει το τελικό τέμπο, αφού στην προκύπτουσα ΣΠ (Σχ. 5.4ε) η πιο εξέχουσα κορυφή εμφανίζεται στα 110 BPM. Ο εκθέτης W είναι πάρα πολύ κρίσιμος για την ακρίβεια της μεθόδου. Η βέλτιστη τιμή του διαφέρει από κομμάτι σε κομμάτι και δεν μπορεί να υπολογιστεί με αυτόματο ή αναλυτικό τρόπο. Στο Σχ. 5.5 παρουσιάζεται η

ακρίβεια βάσει των μετρικών *acc1* και *acc2* σε τέσσερα διαφορετικά σύνολα δεδομένων, το οποίο επαληθεύει την μη ύπαρξη βέλτιστης τιμής για το W. Επομένως, η τιμή του W εκτιμάται με αναζήτηση πλέγματος. Η τιμή του W που μεγιστοποιεί την ακρίβεια (*acc1*) στο σύνολο επαλήθευσης χρησιμοποιείται και στη φάση αξιολόγησης. Όπως και στην περίπτωση της μουσικής κατηγοριοποίησης, στις δύο μεθόδους εξαγωγής τέμπο που υιοθετούν σχήμα κατηγοριοποίησης, ακολουθήσαμε την τακτική «Τεμαχισμού σε Ν υποσύνολα» για την εκμάθηση, επαλήθευση και κατηγοριοποίηση. Αναλυτική περιγραφή των συλλογών και των πειραματικών αποτελεσμάτων θα παρουσιαστούν στη συνέχεια.

5.5 Πειραματικά Αποτελέσματα

5.5.1 Εισαγωγή

Οι προτεινόμενες μέθοδοι εξαγωγής τέμπο αξιολογήθηκαν σε συνολικά εφτά συλλογές δεδομένων. Επίσης δύο μέθοδοι υποβλήθηκαν στον διεθνή διαγωνισμό MIREX τα έτη 2010 και 2011. Μια μέθοδος που υπολογίζει το τέμπο ως τη πιο εξέχουσα κορυφή της ΣΠ (έτος 2010), και μέθοδος που υιοθετεί το μοντέλο μετρικών σχέσων (έτος 2012). Τα αποτελέσματα καταδεικνύουν ότι όλες οι προτεινόμενες μέθοδοι εξαγωγής τέμπο (μοντέλο μετρικών σχέσεων, SVMs) είναι άκρως ανταγωνιστικές με τις μεθόδους αιχμής και σε αρκετές περιπτώσεις επιτυγχάνουν υψηλότερες επιδόσεις. Κατά την αξιολόγηση του τέμπο υπάρχει ένα διάστημα ανοχής σε σχέση με το σωστό τέμπο. Αυτό το διάστημα ορίζεται ως το ποσοστό 4% σε σχέση με το αληθές τέμπο. Οι περισσότερες μέθοδοι πάσχουν από το να εντοπίσουν το τέμπο στη σωστή οκτάβα, δηλαδή εξάγουν τέμπο που συχνά είναι πολλαπλάσιο ή υποπολλαπλάσιο του πραγματικού. Ωστόσο, επειδή η έννοια του «σωστού» τέμπο είναι στενά συνδεδεμένη και με την έννοια του «αντιληπτικού» τέμπο (Κεφ. 1), τα λάθη οκτάβας δεν μπορούν εν γένει να ληφθούν ως λάθη. Αν για ένα κομμάτι με πραγματικό τέμπο 90 BPM ο αλγόριθμος Α εξάγει 180 BPM και ο αλγόριθμος Β 150 BPM, προφανώς η έξοδος του Α είναι περισσότερο ρυθμικά συνεπής με το επισημειωμένο τέμπο ενώ δεν ισχύει το ίδιο για τον αλγόριθμο Β.

Προκειμένου να αξιολογείται μία μέθοδος και στην ικανότητά του να βρίσκει το σωστό μετρικό επίπεδο αλλά και να αξιολογείται θετικά στα λάθη οκτάβας, είναι κοινώς αποδεκτές δύο μετρικές αξιολόγησης:

Accuracy1

Το ποσοστό επιτυχίας της ικανότητα μιας μεθόδου να βρει το πραγματικό τέμπο σε ένα διάστημα ανοχής 4%.

Accuracy2

Το ποσοστό επιτυχίας της ικανότητας μιας μεθόδου να βρει το πραγματικό τέμπο, το διπλάσιο, τριπλάσιο, μισό ή το 1/3 αυτού σε ένα διάστημα ανοχής 4%.

Συνήθως όλες οι επιστημονικές εργασίες αναφέρουν αποτελέσματα και για τις δύο μετρικές. Η accuracy2 μπορεί να θεωρηθεί περισσότερο ενδεικτική για την αποτελεσματικότητα της ανάλυσης περιοδικότητας μιας μεθόδου. Ένας αλγόριθμος που επιτυγχάνει βάσει της accuracy2, τότε σχεδόν σίγουρα η ΣΠ θα έχει κορυφή στο σωστό τέμπο, αλλά η υψηλότερη κορυφή της θα είναι σε άλλη

οκτάβα. Αντίθετα, η accuracy1 αντικατοπτρίζει την ικανότητα μιας μεθόδου να βρίσκει το σωστό μετρικό επίπεδο.

Ωστόσο, ενώ θα μπορούσε κάποιος να ισχυριστεί ότι η μετρική accuracy2 είναι κοντινότερη στην έννοια του αντιληπτικού τέμπο, αυτό δεν είναι απόλυτα σωστό, καθώς δεν είναι όλα τα κλάσματα και πολλαπλάσια ρυθμικά συνεπή με το πραγματικό τέμπο. Το ποια πολλαπλάσια είναι σχετικά και ποια όχι εξαρτάται σε μεγάλο βαθμό από τη μετρική δομή ενός κομματιού.

MIREX P-Score:

Μια προσπάθεια να υπερκεραστεί αυτή η ασάφεια για το ποια πολλαπλάσια/κλάσματα του πραγματικού τέμπο είναι σωστά έγινε με την εισαγωγή της μετρικής αξιολόγησης P-score στον διαγωνισμό MIREX που γίνεται στα πλαίσια του διεθνούς συνεδρίου International Society for Music Information Retrieval (ISMIR). Κάθε μουσικό απόσπασμα επισημειώθηκε από 40 ειδικούς οι οποίοι χτυπούσαν πάνω στον μουσικό παλμό ένα πλήκτρο. Από κάθε ειδικό εξάχθηκε ένα τέμπο. Τα δύο συχνότερα τέμπο T_1, T_2 θεωρήθηκαν ως σωστά και για καθένα από αυτά εξάχθηκε ένα σχετικό βάρος ανάλογο με το ποσοστό των επισημειωτών που συμφώνησαν σε αυτό. Τα δύο βάρη w_1, w_2 κανονικοποιήθηκαν ώστε να έχουν άθροισμα στη μονάδα. Κάθε μέθοδος έπρεπε να εξάγει δύο τέμπο. Το P-score για κάθε κομμάτι είναι το άθροισμα των w_1, w_2 για τα οποία η μέθοδος υπολόγισε σωστά τα T_1, T_2 σε ένα διάστημα ανοχής 8%. Η μετρική P-score «απαλλάσσει» κατά μία έννοια τις μεθόδους να επιλέξουν το σωστό μετρικό επίπεδο επομένως διαισθητικά είναι πιο κοντά στη μετρική accuracy2.

Η προτεινόμενη μέθοδος εξαγωγής τέμπο με χρήση του μοντέλου μετρικών σχέσεων αξιολογήθηκε στις εξής 7 συλλογές:

ISMIR 2004 *Songs* Dataset

Αποτελείται από 465 αποσπάσματα των 20s από ένα πλήθος μουσικών ειδών όπως Rock, Classical, Electronic, Latin, Samba, Jazz, Afro-beat, Flamenco, Balkan και Greek. Δημιουργήθηκε στα πλαίσια του πρώτου διαγωνισμού αλγορίθμων που έγινε στα πλαίσια του ISMIR 2004 που αποτέλεσε το έναυσμα για την καθιέρωση του διαγωνισμού MIREX από το 2005 και μετά. Η συλλογή είναι ελεύθερη για κατέβασμα από το διαδίκτυο[6].

ISMIR 2004 *Ballroom* Dataset

Αποτελείται από 698 αποσπάσματα των 30s χορευτικών (κυρίως Latin) κομματιών από τα εξής 8 είδη: Cha Cha Cha, Jive, Quickstep, Rumba, Samba, Tango, Viennese Waltz και Slow Waltz. Η συλλογή είναι ελεύθερη για κατέβασμα από το διαδίκτυο[7].

MIREX 2005 McKinney Dataset (*MCK*)

Αποτελείται από 160 αποσπάσματα των 30s. Κάθε απόσπασμα έχει επισημειωθεί με δύο τέμπο, που προκύπτουν ως τα δύο πιο εξέχοντα τέμπο από επισημειώσεις 40 ειδικών [McKinney2004]. Από τα 160 αποσπάσματα μόνο τα

[6] http://www.iua.upf.edu/mtg/ismir2004/contest/tempoContest/data3.tar.gz
[7] http://www.iua.upf.edu/mtg/ismir2004/contest/tempoContest/data1.tar.gz

20 είναι διαθέσιμα για κατέβασμα από το διαδίκτυο[8]. Τα υπόλοιπα 140 παραμένουν κρυφά για τους σκοπούς του διαγωνισμού MIREX.

GTZAN Genre Dataset

Αποτελείται από 1000 αποσπάσματα των 30s για δέκα διαφορετικά είδη μουσικής [Tzanetakis2002]. Αποτελεί μια πολύ διαδεδομένη συλλογή για αξιολόγηση μεθόδων αυτόματης κατηγοριοποίησης σε είδος (genre classification). Πολύ πρόσφατα έγινε περαιτέρω επεξεργασία με επισημείωσεις του τέμπο [Tzanetakis2013, Percival2014] και αφαίρεση των διπλότυπων (duplicates) [Sturm2014]. Είναι και αυτή διαθέσιμη στο διαδίκτυο[9].

LastFm *Speedo* Dataset

Δημιουργήθηκε από δεδομένα του ιστότοπου LastFm [Levy2011]. Η αρχική έκδοση της συλλογής αποτελούταν από 4200 περίπου κομμάτια, τα οποία ήταν επισημειωμένα από χρήστες του ιστότοπου με ένα τέμπο και μία κλάση ταχύτητας (αργό, γρήγορο, ενδιάμεσο). Κάθε κομμάτι επισημειώθηκε από πολλούς χρήστες. Ο Peeters [Peeters2012] επεξεργάστηκε ώστε να αφαιρεθούν αντικρουόμενες ή μη αξιόπιστες επισημειώσεις. Η τελική μορφή της συλλογής προέκυψε στο [Percival2014] όπου έγιναν διαθέσιμα και τα αρχεία ήχου. Η τελική έκδοση της συλλογής αποτελείται από 1400 επισημειωμένα αποσπάσματα με μήκος 30s ή 60s[10].

HAINSWORTH Dataset

Αποτελείται από 222 μουσικά αποσπάσματα που χρησιμοποιήθηκαν από τον συγγραφέα του [Haisworth2004].

AFRICAN Dataset

Αποτελείται από 70 αποσπάσματα αφρικάνικων ρυθμών που έχουν συλλεχτεί από τον Olmo Cornelis (University of Ghent), ο οποίος αξιολόγησε 15 διαφορετικές μεθόδους εξαγωγής τέμπο [Cornellis2013].

Οι τρεις διαφορετικές μέθοδοι εξαγωγή τέμπο αναπτύχθηκαν και αξιολογήθηκαν σε διαφορετικά στάδια έρευνας στα πλαίσια της εκπόνησης αυτής της διατριβής. Επομένως τα μεταξύ τους αποτελέσματα δεν είναι απολύτως συγκρίσιμα. Για παράδειγμα δεν αξιολογήθηκαν όλες οι μέθοδοι σε όλες τις συλλογές, ή οι επισημειώσεις κάποιων συλλογών αλλάξανε (διορθώθηκαν) κατά τη διάρκεια εκπόνησης της διατριβής. Συνεπώς, τα αποτελέσματα των τριών μεθόδων θα παρουσιαστούν χωριστά στις 3 επόμενες υποενότητες. Τέλος, στην Ενότητα 5.5.5 θα παρουσιαστούν τασυγκριτικά αποτελέσματα με μεθόδους της διεθνούς βιβλιογραφίας.

5.5.2 Αξιολόγηση της μεθόδου με χρήση μετρικών σχέσεων – επίδραση των χαρακτηριστικών – διαχωρισμού πηγών

Η πρώτη μέθοδος εξαγωγής τέμπο με την χρήση μετρικών σχέσεων είναι η μόνη που αξιολογήθηκε και στις 7 συλλογές. Ο κύριος λόγος αυτού είναι το γεγονός ότι

[8] http://www.music-ir.org/mirex/wiki/2005:Audio_Tempo_Extraction
[9] http://opihi.cs.uvic.ca/sound/genres.tar.gz
[10] http://opihi.cs.uvic.ca/tempo/

δεν χρειάζεται εκμάθηση, κάτι που καθιστά αρκετά εύκολο την εκτέλεση πειραμάτων από άλλους ερευνητές. Για τις 5 από τις 7 συλλογές τα πειράματα εκτελέστηκαν από άλλους ερευνητές που ήθελαν να συγκρίνουν τις δικές τους μεθόδους τους με αυτήν της παρούσας διατριβής. Κάτι τέτοιο δεν είναι εξίσου εύκολο και πρακτικό για τις άλλες δύο μεθόδους, καθώς χρειάζεται εκμάθηση, αναζήτηση πλέγματος για τις παραμέτρους, κατάτμηση των συλλογών σε N διαμερίσεις κ.ο.κ. Μια τέτοια διαδικασία είναι χρονοβόρα τόσο σε επίπεδο υλοποίησης όσο και χρόνου εκτέλεσης.

Στην συνέχεια θα παρουσιαστούν αναλυτικά τα αποτελέσματα για τις δύο συλλογές Ballroom και Songs για τις οποίες τα πειράματα εκτελέστηκαν τον συγγραφέα της παρούσας διατριβής. Τα αποτελέσματα για τις άλλες συλλογές θα παρουσιαστούν στην ενότητα των συγκριτικών αποτελεσμάτων (Εν. 5.5.5). Τα βασικά συστατικά αυτής της μεθόδου είναι ο διαχωρισμός κρουστών/αρμονικών πηγών, η εξαγωγή δύο ειδών χαρακτηριστικών (ενέργειας και χρώματος) και ο συνδυασμός τους.

Ο Πίνακας 5.1 παρουσιάζει αναλυτικά την αξιολόγηση της προτεινόμενης μεθόδου με τις μετρικές *acc1* και *acc2* στις συλλογές Ballroom και Songs για όλες τις δυνατές παραμετροποιήσεις. Η $1^{η}$ ομάδα του Πίνακα 5.1 παρουσιάζει τα αποτελέσματα με τη χρήση μόνο του μετρικού μοντέλου. Επομένως το βήμα διαχωρισμού πηγών έχει παρακαμφθεί και τα χαρακτηριστικά χρώματος/ενέργειας εξάγονται απευθείας από το συνολικό φασματογράφημα. Η $2^{η}$ ομάδα παρουσιάζει τα αποτελέσματα με τη χρήση μόνο του διαχωρισμού πηγών, χωρίς τη χρήση του μετρικού μοντέλου. Στη μέθοδο αυτή το τέμπο υπολογίζεται ως το πιο εξέχων στην ΣΠ. Η $3^{η}$ ομάδα παρουσιάζει τα αποτελέσματα της πλήρους μεθόδους. Και για τις τρεις προσεγγίσεις, η ΣΠ υπολογίστηκε: (α) μόνο από τα χαρακτηριστικά χρώματος, (β) μόνο από τα χαρακτηριστικά ενέργειας και (γ) συνδυάζοντας τα δύο είδη χαρακτηριστικών (α) και (β). Επομένως η τελευταία γραμμή του Πίνακα 5.1 αντιστοιχεί στα αποτελέσματα της πλήρους μεθόδου. Το ελάχιστο/μέγιστο τέμπο ορίστηκαν σε $T_{min} = 30$ BPM και $T_{max} = 500$ BPM (Εξ. 5.3). Η μεγάλη τιμή για το μέγιστο τέμπο T_{max} είναι απαραίτητη για το μοντέλο μετρικών σχέσεων και στον υπολογισμό των T_0, T_{slow}, T_{fast}.

Όταν υιοθετείται μονάχα το μετρικό μοντέλο (ΜΜ), παρατηρούμε ότι ο συνδυασμός των δύο ειδών χαρακτηριστικών βελτιώνει την επίδοση στην περίπτωση της συλλογής Ballroom, ενώ μειώνεται αισθητά στην συλλογή Songs, όπου με τη χρήση μόνο των χαρακτηριστικών ενέργειας επιτυγχάνεται *acc1* σχεδόν 68%. Η χρήση όμως μόνο των χαρακτηριστικών ενέργειας είναι ανεπαρκής για τη συλλογή Ballroom, αφού επιτυγχάνεται ποσοστό αναγνώρισης μόνο ~39%. Ωστόσο, λαμβάνοντας υπόψη και τις δύο συλλογές, ο συνδυασμός των δύο χαρακτηριστικών επιτυγχάνει καλύτερη συνολική επίδοση. Ωστόσο, πρέπει να σημειωθεί ότι πολύ μεγάλη διαφορά στην επίδοση παρατηρείται κυρίως για την μετρική *acc1*. Αν συγκρίνουμε π.χ. την *acc1* για τη συλλογή Ballroom για τα χαρακτηριστικά χρώματος/ενέργειας, παρατηρούμε μια διαφορά της τάξης του 20% για την *acc1* και μόλις στο 4.5% περίπου για την *acc2*. Το αντίστροφο συμβαίνει για τη συλλογή Songs, με ακόμα μεγαλύτερες διαφορές για τα δύο είδη χαρακτηριστικών.

Μπορούμε να συμπεράνουμε ότι η τόσο μεγάλη διαφορά στα δύο χαρακτηριστικά για την *acc1* δεν οφείλεται τόσο στο ότι η μία ΣΠ αναπαριστά καλύτερα το ρυθμικό περιεχόμενο από την άλλη ΣΠ, αλλά στο γεγονός ότι η ΣΠ που προκύπτει από τα χαρακτηριστικά αυτά μπορεί να «πολώνεται» περισσότερο ή λιγότερο στην σωστή οκτάβα για κάποια συγκεκριμένη συλλογή.

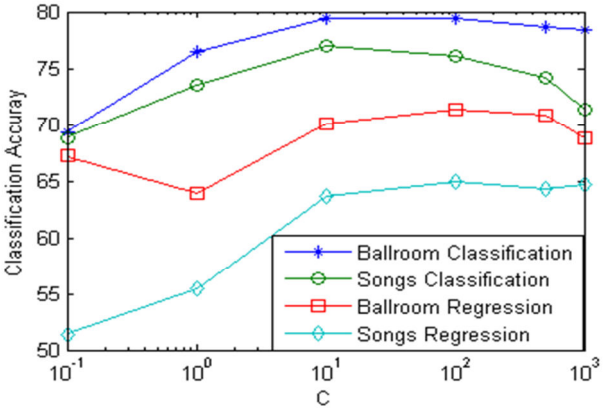
		Ballroom		Songs	
	Χαρακτηριστικά	**Acc1**	**Acc2**	**Acc1**	**Acc2**
Μετρικό Μοντέλο (ΜΜ)	Ενέργειας	38.68	84.53	67.96	88.82
	Χρώματος	56.30	88.83	44.52	79.78
	Συνδυασμός	58.17	92.41	60.00	89.03
Διαχωρισμός Πηγών (ΔΧ)	Ενέργειας	52.29	88.40	24.52	88.60
	Χρώματος	46.13	88.40	15.05	70.54
	Συνδυασμός	53.30	90.97	21.08	88.17
Μετρικό Μοντέλο + Διαχωρισμός Πηγών	Ενέργειας	48.42	90.40	57.85	88.39
	Χρώματος	54.58	82.95	40.86	73.12
	Συνδυασμός	59.89	93.27	58.49	89.89

Πίνακας 5.1: Επίδραση των χαρακτηριστικών (ενέργειας/χρώματος), του διαχωρισμού πηγών και του μοντέλου μετρικών σχέσεων στην επίδοση της εξαγωγής τέμπο.

Επιπλέον παρατηρούμε ότι η διαφορά στη σωστή οκτάβα που προκύπτει από τα δύο χαρακτηριστικά «εξισορροπείται» όταν τα συνδυάζουμε. Για παράδειγμα στη συλλογή Ballroom συνδυάζοντας τα δύο χαρακτηριστικά παίρνουμε ακόμα καλύτερα αποτελέσματα και για τις δύο μετρικές. Αν π.χ. σε ένα κομμάτι με τέμπο 120 BPM, το ένα είδος χαρακτηριστικού έδινε τέμπο 60 BPM και το άλλο 240 BPM, χρησιμοποιώντας την ΣΠ που προκύπτει από συνδυασμό τους μπορεί να εξάγουμε το πραγματικό τέμπο. Κάτι αντίστοιχο παρατηρούμε και στη συλλογή Songs, αλλά η επίδοση μειώνεται σε σχέση με τα χαρακτηριστικά ενέργειας.

Όταν δεν υιοθετείται το μετρικό μοντέλο (Πίνακας 5.1, $2^{η}$ ομάδα) αλλά μόνο ο διαχωρισμός πηγών, η προτεινόμενη μέθοδος εμφανίζει πολλά λάθη «οκτάβας». Ειδικά στη συλλογή songs, η *acc1* είναι της τάξης του 20% ενώ η *acc2* είναι ~80-90%. Μπορούμε να συμπεράνουμε ότι η χρήση μετρικών σχέσεων είναι εξαιρετικής σημασίας για τη μείωση λαθών οκτάβας, και ιδιαίτερα σε συλλογές που περιέχουν «δύσκολα» ρυθμικά μουσικά κομμάτια. Στην συλλογή Ballroom αυτό το φαινόμενο δεν είναι τόσο έκδηλο. Στο ~50% των περιπτώσεων η πιο εξέχουσα κορυφή στην ΣΠ αντιστοιχεί στο αληθές τέμπο.

Στην $3^{η}$ ομάδα του Πίνακα 5.1 παρατηρούμε ότι οι διαφορές της επίδοσης των δύο χαρακτηριστικών και του συνδυασμού τους έχει μειωθεί σε σχέση με τις άλλες δύο εκδοχές της μεθόδου. Μόνη εξαίρεση είναι η περίπτωση των χαρακτηριστικών χρώματος για τη συλλογή Songs, όπου και για τις δύο μετρικές η ακρίβεια είναι πολύ μικρή. Βάσει αυτής της παρατήρησης μπορούμε να ισχυριστούμε ότι η ταυτόχρονη υιοθέτηση του διαχωρισμού πηγών και των μετρικών σχέσεων, καθιστά λιγότερο ευαίσθητη την μέθοδο ως προς τα χαρακτηριστικά έμφασης. Επιπλέον, ο συνδυασμός των δύο ειδών χαρακτηριστικών βελτιώνει την επίδοση σε σχέση με τα μεμονωμένα χαρακτηριστικά και για τις δύο συλλογές. Επομένως, μπορούμε να πούμε ότι η τελική έκδοση της μεθόδου υπερτερεί όλων των επιμέρους συνδυασμών. Παρότι υπάρχουν εκδοχές της μεθόδου που αποδίδουν καλύτερα σε συγκεκριμένες συλλογές (π.χ. χαρακτηριστικά ενέργειας χωρίς διαχωρισμό πηγών για τη συλλογή Songs), η τελική έκδοση της προτεινόμενης μεθόδου επιτυγχάνει την



Σχήμα 5.6. Ποσοστό αναγνώρισης κατηγορίας ταχύτητας σε συνάρτηση με τις τιμές της παραμέτρου C (Εξ. 4.1, 5.5) των SVM κατηγοριοποίησης και SVM παλινδρόμησης για τις συλλογές Ballroom και Songs.

καλύτερη συνολική επίδοση. Είναι επίσης αξιοσημείωτο ότι η χρήση του μετρικού μοντέλου, όχι μόνο μειώνει τα λάθη οκτάβας, αλλά αυξάνει και την *acc2*, δηλαδή βελτιώνει την ικανότητα της μεθόδου να βρίσκει σχετικές με τον ρυθμό περιοδικότητες που δεν είναι τόσο έκδηλες στην αρχική ΣΠ.

5.5.3 Αξιολόγηση της μεθόδου εξαγωγής της μουσικής ταχύτητας και μουσικού τέμπο με χειρονακτικά χαρακτηριστικά και SVM ταξινομητή

Η αξιολόγηση μιας οποιαδήποτε μεθόδου τεχνικής απαιτεί επισημειωμένα δεδομένα εκμάθησης, διαφορετικά από τα δεδομένα αξιολόγησης, τα οποία συνήθως είναι δύσκολο να βρεθούν. Μια συνήθης τακτική όπως προαναφέρθηκε είναι ο «διαμερισμός σε N-υποσύνολα» όπου το σύνολο δεδομένων χωρίζεται με τυχαίο τρόπο σε N ισοπληθή υποσύνολα ανά δύο διαφορετικά μεταξύ τους. Ένα από τα N υποσύνολα χρησιμοποιείται σαν σύνολο αξιολόγησης, ένα ως σύνολο επαλήθευσης και τα υπόλοιπα N-2 σαν σύνολο εκμάθησης. Η συνολική επίδοση υπολογίζεται ως ο μέσος όρος στα N χωριστά πειράματα. Μια τέτοια μεθοδολογία δεν είναι πάντα αξιόπιστη, ειδικά όταν το σύνολο δεδομένων παρουσιάζει μεγάλη ομοιογένεια. Σε αυτή την περίπτωση είναι πολύ πιθανόν να βρεθούν παρόμοια δεδομένα στα σύνολα εκμάθησης και αξιολόγησης. Προφανώς, όσο το N μεγαλώνει τόσο θα αυξάνεται η επίδοση κάποιου αλγορίθμου. Για την συγκεκριμένη μέθοδο εξαγωγής μουσικής ταχύτητας και τέμπο ακολουθήσαμε την μέθοδο τεμαχισμού με $N = 3$, χωρίς όμως τη χρήση συνόλου επαλήθευσης. Έτσι χρησιμοποιήθηκαν κάθε φορά τα 2/3 των δειγμάτων ως σύνολο εκπαίδευσης και 1/3 ως σύνολο αξιολόγησης. Σχετικά με τις παραμέτρους του SVM ταξινομητή θέσαμε $\gamma = 2.5\text{x}10^{-3}$(Εξ. 4.4) και αναφέρουμε αποτελέσματα για διάφορες τιμές της παραμέτρου C (Εξ. 4.1).

Η συγκεκριμένη μέθοδος αξιολογήθηκε στις συλλογές Ballroom και Songs. Στο Σχ. 5.6 παρουσιάζονται τα ποσοστά κατηγοριοποίησης με τη μεθοδολογία αξιολόγησης N-cross fold validation με $N = 3$ στις τρεις κλάσεις (αργό, μέτριο, γρήγορο) για τις συλλογές Ballroom και songs με τις μεθόδους SVM κατηγοριοποίησης και SVM παλινδρόμησης, για διάφορες τιμές της παραμέτρου C (Εξ. 4.1, 5.5) των SVM. Η κλιμάκωση των διανυσμάτων περιοδικότητας (Εξ. 3.3) έγινε στο διάστημα [0.8, 1.2] με βήμα $\delta\alpha = 0.02$. Οι κλάσεις ταχύτητας προέκυψαν θέτοντας στην Εξ. 5.3 τα κατώφλια $T_{slow} = 80$ BPM και $T_{fast} = 130$ BPM.

	Ballroom			Songs		
	Αργό	Μεσαίο	Γρήγορο	Αργό	Μεσαίο	Γρήγορο
Αργό	**23**	33	44	**79**	16	5
Μεσαίο	2	**86**	12	17	**82**	1
Γρήγορο	1	32	**67**	46	28	**26**

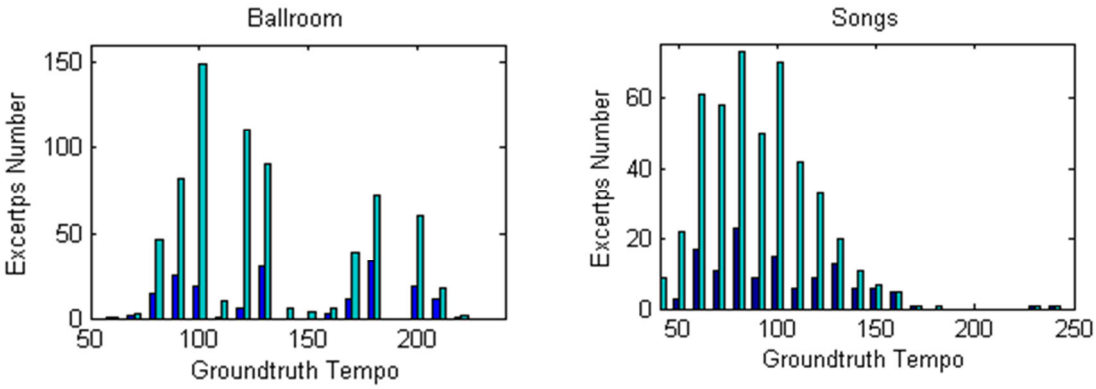
Πίνακας 5.2. Πίνακες σύγχυσης της κατηγοριοποίησης (%) σε μουσική ταχύτητα. Οι στήλες αντιστοιχούν στις προβλέψεις του SVM και οι γραμμές στις πραγματικές κατηγορίες.

Το πλήθος των μπαντών των τέμπο τέθηκε ίσο με $K = 20$ (Εξ. 3.4) και τα διανύσματα εισόδου κανονικοποιήθηκαν στο διάστημα [-1,+1]. Σχετικά με τις παραμέτρους των SVM παλινδρόμησης, δοκιμάστηκαν διάφορες τιμές της παραμέτρου ε (Εξ. 5.6). Τα πειραματικά αποτελέσματα έδειξαν ότι διακυμάνσεις του ε δεν επηρέασαν την συνολική επίδοση.

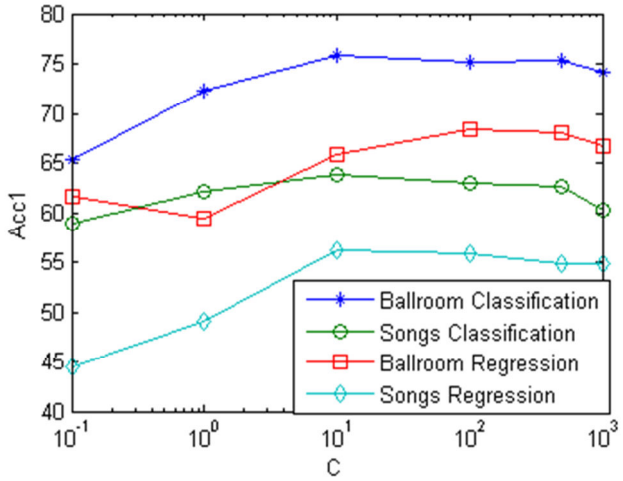
Συνοψίζοντας, από το Σχ. 5.6 φαίνεται ότι η ακρίβεια κατηγοριοποίησης δεν αλλάζει αισθητά για ένα μεγάλο εύρος τιμών της παραμέτρου C (C~$1..500$). Επιπλέον παρατηρούμε ότι και για τις δύο συλλογές τα SVM κατηγοριοποίησης υπερτερούν αισθητά των SVM παλινδρόμησης. Αυτό μπορεί να εξηγηθεί από το γεγονός ότι τα SVM παλινδρόμησης χρειάζονται πολύ περισσότερα δεδομένα εκμάθησης προκειμένου να εξάγουν αξιόπιστα στατιστικά στοιχεία.

Για να αποκτήσουμε μία βαθύτερη εικόνα των πειραματικών αποτελεσμάτων στον Πίνακα 5.2 παρουσιάζεται ο πίνακας σύγχυσης μεταξύ των κατηγοριών για $C = 100$, ενώ στο Σχήμα 5.7 παρουσιάζεται η κατανομή των λαθών στις διάφορες τιμές του τέμπο συγκρινόμενη με την συνολική κατανομή των τέμπο στις δύο συλλογές. Από τον Πίνακα 5.2 φαίνεται ότι πολλά από τα αργά κομμάτια έχουν κατηγοριοποιηθεί ως γρήγορα στην συλλογή Ballroom, ενώ το αντίθετο συμβαίνει για την συλλογή Songs. Αυτό μπορεί να εξηγηθεί από το γεγονός ότι η συλλογή Ballroom περιέχει πολύ λίγα αργά κομμάτια, ενώ αντίστοιχα η συλλογή Songs πολύ λίγα γρήγορα κομμάτια. Επομένως τα SVM αποτυγχάνουν να βρουν αξιόπιστα όρια για τις κλάσεις αυτές. Παρατηρώντας το Σχ. 5.7 φαίνεται ότι κοντά στα όρια $T_{slow} = 80$ BPM και $T_{fast} = 130$ που χωρίζουν τις δύο κλάσεις, παρατηρούνται περισσότερα λάθη συγκριτικά με το πλήθος των παραδειγμάτων που βρίσκονται στις περιοχές αυτές. Μία βέλτιστη επιλογή αυτών των ορίων είναι εξαρτώμενη από τα ίδια τα δεδομένα. Για παράδειγμα στην συλλογή Ballroom η επιλογή $T_{slow} = 110$ και $T_{fast} = 150$ θα έδινε πολύ καλύτερα αποτελέσματα, αφού οι 3 κατηγορίες θα ήταν ευκολότερα διαχωρίσιμες. Ωστόσο, παρότι τα κατώφλια $T_{slow} = 80$ και $T_{fast} = 130$ επιλέχθηκαν βάσει της μουσικής διαίσθησης και όχι βάσει των δεδομένων - οδηγώντας σε άνισες το πλήθος κλάσεις που μειώνει την επίδοση της μεθόδου- επιτυγχάνονται ικανοποιητικά αποτελέσματα, που όπως θα δούμε είναι ανταγωνιστικά με μεθόδους της διεθνούς βιβλιογραφίας.

Στο Σχ. 5.8 παρουσιάζονται τα ποσοστά αναγνώρισης τέμπο (*acc1*) που προκύπτει από τον περιορισμό του διανύσματος περιοδικότητας στο εύρος που αντιστοιχεί στην κλάση που εξάγεται από τα SVM κατηγοριοποίησης, για διάφορες τιμές της παραμέτρου C.



Σχήμα 5.7. Κατανομή των λαθών κατηγοριοποίησης (σκούρο μπλε) σε σύγκριση με την κατανομή των επισημειωμένων τέμπο (ανοιχτό μπλε).



Σχήμα 5.8. Ποσοστό εξαγωγής τέμπο (*acc1*) σε συνάρτηση με τις τιμές της παραμέτρου *C* (Εξ. 4.1 - 5.5) των SVM κατηγοριοποίησης και SVM παλινδρόμησης για τις συλλογές Ballroom και Songs.

Συγκρίνοντας τα Σχ. 5.6 και 5.8 παρατηρούμε ότι τα ποσοστά αναγνώρισης κλάσης και τέμπο είναι παρόμοια για την συλλογή Ballroom. Το ποσοστό κατηγοριοποίησης στη σωστή κλάση ταχύτητας είναι ~79% ενώ η μετρική *acc1* είναι ~ 75%.

Αντίθετα, δεν ισχύει το ίδιο και για την συλλογή Songs, όπου τα ποσοστά εύρεσης του τέμπο είναι πάνω από 10% χαμηλότερα από τα ποσοστά εύρεσης της κλάσης της μουσικής ταχύτητας. Επομένως προκύπτει το συμπέρασμα ότι για την συλλογή Songs είναι πιο δύσκολη η εξαγωγή του τέμπο, ακόμα και με γνώση της μουσικής ταχύτητας. Για να επαληθεύσουμε το συμπέρασμα αυτό, υπολογίσαμε τα ποσοστά αναγνώρισης τέμπο της προτεινόμενης μεθόδου έχοντας γνώση της σωστής μουσικής ταχύτητας. Για τη συλλογή Ballroom το ποσοστό εξαγωγής *acc1* ήταν 88%, ενώ για τη συλλογή Songs μόλις 76%. Το αποτέλεσμα αυτό αναδεικνύει τους περιορισμούς της προτεινόμενης μεθόδου. Ακόμα και με γνώση της σωστής κατηγορίας, η πιο εξέχουσα κορυφή στο αντίστοιχο εύρος της ΣΠ δεν αντιστοιχεί στο σωστό τέμπο. Αυτό μπορεί να οφείλεται στο γεγονός ότι το εύρος της κάθε κλάσης είναι αρκετά μεγάλο, ώστε να έχουμε δύο ρυθμικά σχετιζόμενες εξέχουσες κορυφές μέσα στην ίδια κλάση. Μια λύση θα ήταν η χρήση περισσότερων και άρα πιο στενών κλάσεων. Ωστόσο, ακόμα και στην περίπτωση «κατάλληλου» εύρους, η ΣΠ μπορεί να εμφανίσει μία μη σχετική με

το ρυθμό κορυφή, η οποία βρίσκεται στο εύρος της κλάσης της ταχύτητας του κομματιού και είναι πιο εξέχουσα από την κορυφή που αντιστοιχεί στο σωστό τέμπο.

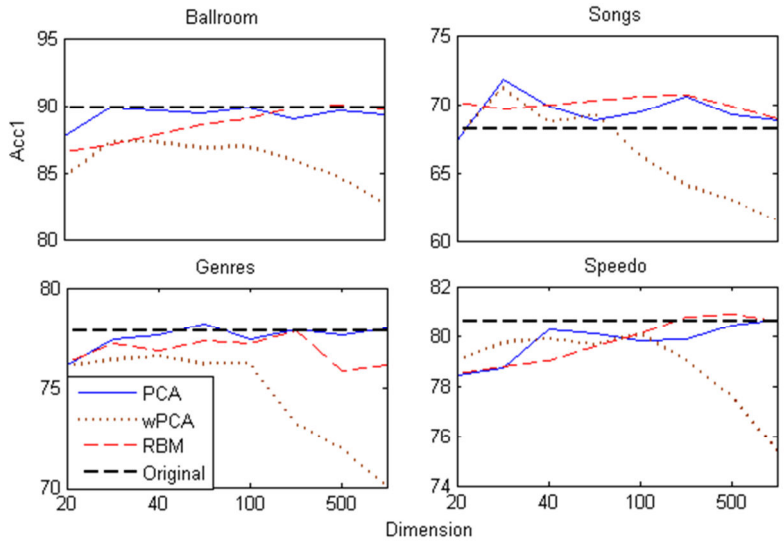
Η μέθοδος αυτή παρότι επιτυγχάνει καλά αποτελέσματα, αποτελεί προνύμφη της επόμενης μεθόδου, η οποία γενικεύει τόσο την διαδικασία εξαγωγής χαρακτηριστικών, όσο και τον φορμαλισμό της κατηγοριοποίησης. Αντί για τρεις αυστηρές ζώνες τέμπο, η κατηγοριοποίηση ταχύτητας γίνεται πολλές φορές σε αυθαίρετες ζώνες, οι οποίες όταν συνδυαστούν δίνουν μια πιο αξιόπιστη μάσκα της ΣΠ που βοηθάει σημαντικά την επιλογή του σωστού μετρικού επιπέδου.

5.5.4 Αξιολόγηση της μεθόδου εξαγωγής τέμπο με πολλαπλούς ταξινομητές σε αυτόματα χαρακτηριστικά.

Η προτεινόμενη μέθοδος αξιολογήθηκε στις συλλογές Ballroom, Songs, Genres και Speedo και χρησιμοποιήσαμε τις μετρικές *acc1* και *acc2* για την αξιολόγηση της επίδοσης. Το πρόβλημα της κατηγοριοποίησης προσεγγίστηκε για κάθε συλλογή χωριστά με τον διαμερισμό σε N υποσύνολα, με $N = 10$. Ο εκθέτης W (Εξ. 5.12) και οι παράμετροι C, γ (Εξ. 4.1, 4.4) του SVM υπολογίστηκαν με αναζήτηση πλέγματος στο σύνολο επαλήθευσης. Το βέλτιστο W είναι ανεξάρτητο από τα βέλτιστα C, γ οπότε πρώτα έγινε η εύρεση των βέλτιστων C, γ και στη συνέχεια η εύρεση του W. Τα βέλτιστα C, γ υπολογίστηκαν για κάθε ταξινομητή χωριστά, ενώ το βέλτιστο W για κάθε ένα από τα N υποσύνολα. Τα κατώφλια των τέμπο T_k (Εξ. 5.9, 5.10) επιλέχθηκαν αυθαίρετα μεταξύ 50 και 180 BPM, με βήμα 10 BPM. Όπως θα δειχθεί παρακάτω, τα πειραματικά αποτελέσματα αναδεικνύουν ότι η επιλογή αυτή δεν είναι κρίσιμη και η επίδοση της μεθόδου δεν είναι ευαίσθητη σε αυτή την επιλογή. Για τις μεθόδους εξαγωγής χαρακτηριστικών χρησιμοποιήθηκαν οι ίδιοι παράμετροι με αυτές που υπολογίστηκαν στην ταξινόμηση μέτρου και χορευτικής μουσικής.

Σε αντίθεση με τα προβλήματα κατηγοριοποίησης του Κεφ. 4, δεν βρέθηκε βέλτιστη διάσταση για κάθε μετασχηματισμό της ΣΠ για όλες τις συλλογές. Το Σχ. 5.9 παρουσιάζει τη μετρική *acc1* για κάθε συλλογή/χαρακτηριστικό ως προς τη διάσταση του μετασχηματισμού. Παρατηρούμε ότι για την περίπτωση του wPCA, η επίδοση είναι συστηματικά χειρότερη από τις άλλες δύο μεθόδους, κάτι που γίνεται ακόμα πιο έκδηλο όσο η διάσταση του wPCA μεγαλώνει. Χρησιμοποιώντας πάνω από 100 συντελεστές του wPCA η *acc1* μειώνεται κατά 5-10 %. Η μόνη εξαίρεση αφορά τη συλλογή Songs όπου τα wPCA εμφανίζουν υψηλή επίδοση, για μικρό όμως πλήθος συντελεστών. Μπορούμε επομένως με ασφάλεια να συμπεράνουμε ότι τα χαρακτηριστικά wPCA είναι υποδεέστερα από τα PCA και RBM χαρακτηριστικά. Σχετικά με τον PCA, όπως αναμενόταν η επίδοση είναι παρόμοια με την χρήση της αρχικής ΣΠ, ειδικά όταν ο αριθμός των συνιστωσών μεγαλώνει. Στην περίπτωση των συλλογών Ballroom και Genres η επίδοση είναι ελάχιστα χαμηλότερη και προσεγγίζει ασυμπτωτικά αυτή της αρχικής ΣΠ. Κάτι αντίστοιχο συμβαίνει και στην συλλογή Speedo, με τη διαφορά ότι για λίγους συντελεστές η διαφορά με την *acc1* είναι λίγο μεγαλύτερη (~2%). Για μεγάλο πλήθος συντελεστών όμως η *acc1* είναι ίδια. Αντίθετα στη συλλογή Songs η επίδοση είναι καλύτερη από την αρχική ΣΠ.

Τα χαρακτηριστικά που εξήχθησαν από το RBM είναι καλύτερα από τα υπόλοιπα στην περίπτωση των συλλογών Songs και Speedo, εφάμιλλα με τα PCA και την αρχική ΣΠ στη συλλογή Ballroom και λίγο χειρότερα στη συλλογή Genres. Συγκεκριμένα, στην περίπτωση των Songs, ακόμα και με πολύ λίγες κρυφές μονάδες, τα χαρακτηριστικά RBM υπερτερούν όλων των άλλων μεθόδων.

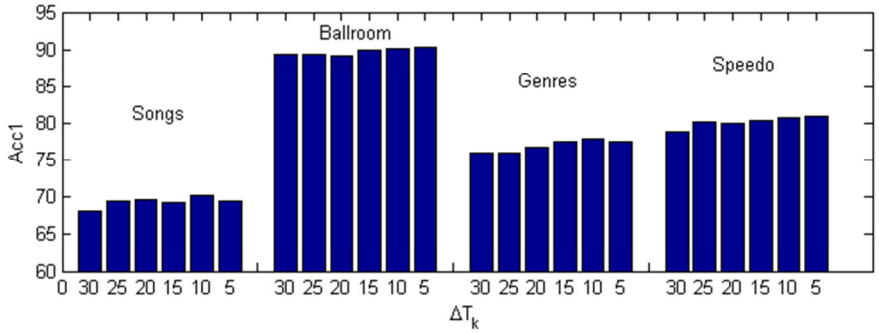


Σχήμα 5.9. Ποσοστό εξαγωγής τέμπο (*acc1*) των τριών μεθόδων PCA, wPCA και RBM εξαγωγής χαρακτηριστικών ως προς την διάσταση των μετασχηματισμών στα 4 σύνολα δεδομένων.

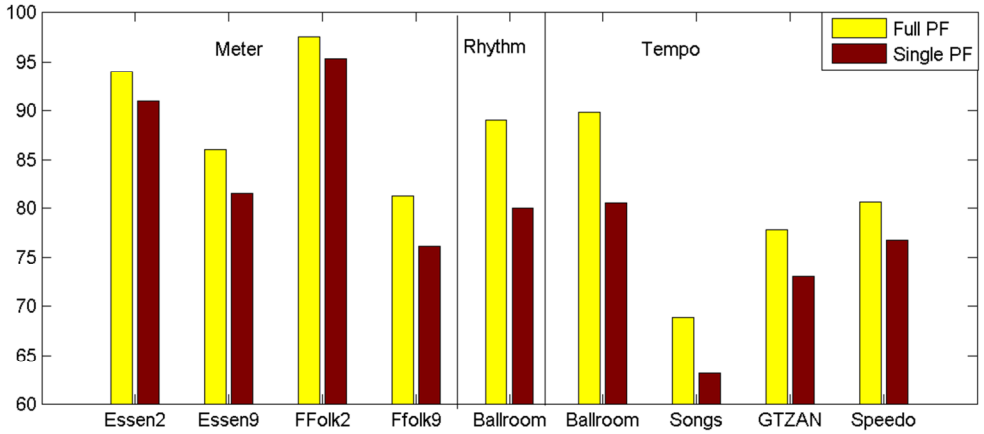
Μέθοδος	Ballroom	Songs	Speedo	GTZAN
Αρχική ΣΠ	89.8/96.4	68.8/80.9	80.6/95	77.8/91.2
PCA (40)	89.7//96.3	69.9/81.5	80.3/95	77.6/90.9
wPCA (30)	87.3/95.1	71.2/83.7	79.7/94.6	76.4/90.7
RBM (200)	89.8/96.1	70.7/83.6	80.8/95.3	77.8/91.8

Πίνακας 5.2. Ποσοστό εξαγωγής τέμπο (*acc1/acc2*) για τις τρεις μεθόδους εξαγωγής χαρακτηριστικών και για την αρχική ΣΠ. Οι τιμές στην παρένθεση αντιστοιχούν στη διάσταση των χαρακτηριστικών του κάθε μετασχηματισμού.

Στη περίπτωση του Speedo αποδίδουν καλύτερα από τα PCA όταν ο αριθμός των χαρακτηριστικών είναι πάνω από 50 και καλύτερα από την αρχική ΣΠ όταν είναι πάνω από 100. Κάτι ανάλογο συμβαίνει και στη συλλογή Ballroom. Έχουμε μεγαλύτερο *acc1* σε σχέση με τα PCA, αλλά ποτέ η *acc1* δεν είναι μεγαλύτερη από αυτή με την αρχική ΣΠ· είναι όμως σχεδόν ίση όταν η διάσταση της απεικόνισης είναι πάνω από 200. Στην περίπτωση της συλλογής Genres, τα RBM χαρακτηριστικά εμφάνισαν χειρότερη επίδοση από τα PCA και την αρχική ΣΠ. Πάντως μπορούμε να συμπεράνουμε ότι όπως και στην περίπτωση της κατηγοριοποίησης, τα RBM χαρακτηριστικά έχουν πιο επιθυμητή συμπεριφορά ως προς το μέγεθος του μετασχηματισμού. Ο Πίνακας 5.2 συνοψίζει τα αποτελέσματα των τριών μεθόδων και της αρχικής ΣΠ. Λόγω του ότι δεν υπάρχει βέλτιστη διάσταση για κάθε μέθοδο, παρουσιάζονται τα αποτελέσματα για την διάσταση που επιτυγχάνει την καλύτερη μέση *acc1* ως προς τις συλλογές. Μπορούμε να συμπεράνουμε ότι (αν και με μεγαλύτερο πλήθος συντελεστών) η RBM μέθοδος έχει λίγο καλύτερη συνολική επίδοση σε σχέση με τα υπόλοιπα χαρακτηριστικά. Ωστόσο πρέπει να σημειωθεί ότι η διαφορά αυτή είναι στατιστικά μη σημαντική βάσει της A/B δοκιμής.



Σχήμα 5.10. Ποσοστό εξαγωγής τέμπο (*acc1*) βάσει του βήματος κατωφλίωσης ΔT_k (Εξ. 5.9-5.10) για τις 4 συλλογές.



Σχήμα 5.11. Ποσοστά επίδοσης για τα προβλήματα ταξινόμησης μέτρου, ταξινόμησης σε χορευτικό είδος και εξαγωγής του τέμπο (*acc1*) για τις δύο εκδοχές (Εξ. 2.32 και Εξ. 2.34) της ΣΠ σε διάφορες συλλογές.

Όπως προαναφέρθηκε, τα κατώφλια T_k του τέμπο που χρησιμοποιήθηκαν για τον υπολογισμό της μάσκας $M(T)$ (Εξ. 5.9-5.10) επιλέχτηκαν αυθαίρετα. Ωστόσο τα πειραματικά αποτελέσματα δείχνουν ότι η επιλογή των T_k δεν είναι κρίσιμη. Θέτοντας $T_0 = 50$ και $T_K = 180$ δοκιμάστηκαν διάφορες τιμες του βήματος ΔT_k κατά τον υπολογισμό των κατωφλιών $T_k = T_0 + (k-1)\Delta T_k$. Στο Σχ. 5.10 παρουσιάζεται η *acc1* σε σχέση με το ΔT_k για όλες τις συλλογές. Ως μέθοδο εξαγωγής χαρακτηριστικών χρησιμοποιήσαμε τα RBM με 200 κρυφές μονάδες. Μπορούμε να συμπεράνουμε ότι η *acc1* αυξάνεται όταν το βήμα ΔT_k μικραίνει. Αυτό καταδεικνύει ότι η μέθοδος εξαγωγής με τη συστοιχία των ταξινομητών τέμπο καλώς ορισμένο, όσο περισσότεροι ταξινομητές χρησιμοποιηθούν τόσο αυξάνεται η *acc1*. Ωστόσο, πρέπει να παρατηρήσουμε ότι ακόμα και για μεγάλο ΔT_k η επίδοση είναι ακόμα σχετικά υψηλή.

Σε αυτό το σημείο θα παρουσιαστούν τα πειραματικά αποτελέσματα για δύο παραλλαγές της ΣΠ. Λόγω του ότι η μέθοδος εξαγωγής τέμπο που παρουσιάζεται χρησιμοποιεί την ίδια ΣΠ και τα ίδια χαρακτηριστικά της ΣΠ με τις μεθόδους κατηγοριοποίησης του Κεφ. 4, θα αναφερθούν αποτελέσματα και για αυτές τις μεθόδους. Η 1η παραλλαγή της ΣΠ που αξιολογήθηκε είναι η χρήση διαφόρων τιμών του μέγιστου τέμπο ανάλυσης $T_{\max}$, από 300 έως 500 BPM (Εξ. 2.20). Τα πειραματικά αποτελέσματα έδειξαν ότι δεν υπάρχει σημαντική

μεταβολή στην επίδοση αυξάνοντας το $T_{\max}$ πέρα των 300 BPM. Επομένως η τιμή $T_{\max} = 300$ BPM είναι επαρκής, κάτι που είναι σε συμφωνία με τα ευρήματα στη βιβλιογραφία.

Η $2^{η}$ παραλλαγή είναι η σύμπτυξη των επιμέρους ΣΠ (Εξ. 2.34) αντί της χρησιμοποίησης της πλήρους μορφής (Εξ. 2.32). Στο Σχ. 5.11 παρουσιάζεται η σύγκριση για τις δύο εκδοχές της ΣΠ της επίδοσης (ποσοστά ακρίβειας για τα προβλήματα ταξινόμησης και *acc1* για το πρόβλημα εξαγωγής τέμπο) για όλες τις συλλογές. Είναι έκδηλο ότι η παράθεση των επιμέρους ΣΠ για τις διάφορες συναρτήσεις έμφασης αντί για τη σύμπτυξή τους αυξάνει δραματικά (που μπορεί να φτάσει σε διαφορά της τάξης του 10%) την επίδοση όλων των μεθόδων.

5.5.5 Σύγκριση των μεθόδων με την διεθνή βιβλιογραφία

Σε αντίθεση με την εύρεση του μουσικού κλειδιού και του χορευτικού ρυθμού, το πρόβλημα της εξαγωγής τέμπο τυγχάνει πολύ μεγάλου ενδιαφέροντος στην ερευνητική κοινότητα και υπάρχει μεγάλο εύρος μεθοδολογιών στη διεθνή βιβλιογραφία. Στα πλαίσια της παρούσας διατριβής επιλέχτηκαν οι πιο επιτυχημένοι (βάσει αποτελεσμάτων) μέθοδοι που έχουν αξιολογηθεί σε κάποια(ες) από τις συλλογές δεδομένων που παρουσιάστηκαν, προκειμένου να γίνει η σύγκριση των προτεινόμενων μεθόδων εξαγωγής τέμπο. Οι «μέθοδοι αναφοράς» παρουσιάζονται παρακάτω και χωρίζονται σε δύο μεγάλες κατηγορίες, ανάλογα με το αν περιλαμβάνουν ή όχι κάποιο υποσύστημα Μηχανικής Μάθησης.

Μέθοδοι χωρίς Μηχανική Μάθηση

Μέθοδος VAMP

Το VAMP[11] είναι ένα Plug-in για επεξεργασία μουσικών σημάτων. Περιλαμβάνει μια πληθώρα βιβλιοθηκών αυτόματης ανάλυσης μουσικής, όπως εξαγωγή κλειδιού, μουσικής ομοιότητας καθώς και μεταγραφής. Υλοποιεί τον αλγόριθμο εύρεσης παλμού που περιγράφεται στο [Davies2007], από τον οποίο στη συνέχεια προκύπτει η τιμή του τέμπο. Η μέθοδος αυτή είναι βασισμένη στην υιοθέτηση της αυτοσυσχέτισης ως ΣΠ, ενώ η έξοδος ενός ταλαντωτή που αντιστοιχεί στο τέμπο χρησιμοποιείται για να εξαχθούν οι θέσεις του παλμού. Στη συνέχεια ορίζεται μία «κατάσταση γενικού πλαισίου» (context-dependent state) η οποία χειρίζεται θέματα συνέχειας και ασυνέχειας του παλμού (π.χ. πήδημα φάσης, χειρισμός λαθών οκτάβας κ.ο.κ.).

Μέθοδος των Tzanetakis και Percival [Tzanetakis2013]

Αποτελεί μια ευριστική μέθοδο που επίσης βασίζεται στην αυτοσυσχέτιση για την εξαγωγή της ΣΠ, από την οποία εξάγονται κάποια υποψήφια τέμπο. Στη συνέχεια για κάθε τέμπο υπολογίζεται η ετεροσυσχέριση (cross-correlation) με έναν παλμό αντίστοιχης συχνότητας. Από τις δύο αναπαραστάσεις υπολογίζονται τρία υποψήφια τέμπο, στα οποία στη συνέχεια εφαρμόζεται μια ευριστική σχέση για την επιλογή του σωστού τέμπο. Η μέθοδος αυτή έχει κάποια κοινά χαρακτηριστικά με την μέθοδο των μετρικών σχέσεων. Επιπλέον σε αυτή την ερευνητική εργασία διεξήχθη ένα εκτεταμένο πείραμα διαφόρων μεθόδων σε διάφορες συλλογές δεδομένων.

[11] http://vamp-plugins.org/

Μέθοδος των Eck και Casagrande [Eck2005]

Είναι η ίδια μέθοδος που χρησιμοποιήθηκε και στα συγκριτικά αποτελέσματα εξαγωγής μέτρου (Ενότητα 4.2.2). Ο απλός ευριστικός κανόνας στην ΣΠ που βρίσκει αν το μέτρο είναι «διπλό» ή «τριπλό» χρησιμοποιείται μετέπειτα για την επιλογή του σωστού τέμπο από την ΣΠ. Η μέθοδος αυτή αξιολογήθηκε στις συλλογές Ballroom και Songs.

Μέθοδος των Klapuri, Eronen και Astola [Klapuri2006]

Αποτελεί μια από τις πιο αξιοσημείωτες μεθόδους εξαγωγής τέμπο. Ήταν η καλύτερη σε επίδοση μέθοδος όταν προτάθηκε και παραμένει στο State-of-the-Art ακόμα και σήμερα. Δεν είναι μια μέθοδος αποκλειστικά για την εξαγωγή του τέμπο, αλλά για την πλήρη εξαγωγή του μουσικού μέτρου, το οποίο μοντελοποιείται σε τρία μετρικά επίπεδα: του μέτρου, του βασικού παλμού και του tatum. Αρχικά γίνεται επεξεργασία των συναρτήσεων έμφασης από μια συστοιχία ταλαντωτών. Η προκύπτουσα ΣΠ χρησιμοποιείται ως είσοδος σε ένα Μπεϋζιανό Δίκτυο που μοντελοποιεί τις σχέσεις των τριών μετρικών επιπέδων που αναφέρθηκαν. Στην συνέχεια εξάγονται οι παλμοί για κάθε μετρικό επίπεδο. Παρότι η μέθοδος αρχικά αξιολογήθηκε σε μία προσωπική συλλογή δεδομένων των συγγραφέων η οποία δεν είναι διαθέσιμη, μετέπειτα αξιολογήθηκε από άλλους ερευνητές σε πληθώρα συλλογών οπότε είναι άμεσα συγκρίσιμη με τις άλλες μεθόδους της βιβλιογραφίας.

Μέθοδοι με Μηχανική Μάθηση

Μέθοδος των Percival και Tzanetakis [Percival2014]

Αποτελεί επέκτασης της μεθόδου που περιγράφτηκε προηγουμένως [Tzanetakis2013]. Η κύρια διαφορά με την προηγούμενη μέθοδο είναι ότι απλοποιήθηκε η εξαγωγή της ΣΠ. Επιπλέον η ευριστική σχέση στο τελικό στάδιο επιλογής του τέμπο αντικαταστάθηκε από έναν ταξινομητή που αποφασίζει αν θα διπλασιαστεί ή όχι το υποψήφιο τέμπο.

Μέθοδος των Seyerlehner, Widmer και Schnitzer [Seyerlehner2007]

Αποτελεί την πρώτη μέθοδο που όρισε το πρόβλημα της εύρεσης του τέμπο ως ένα πρόβλημα κατηγοριοποίησης. Ένας ταξινομητής των k-κοντινότερων γειτόνων εφαρμόστηκε στην ΣΠ. Το μέσο τέμπο των k-κοντινότερων γειτόνων θεωρείται ως το αληθές τέμπο. Χρησιμοποιήθηκαν δύο μέθοδοι για την εξαγωγή της ΣΠ. Η πρώτη είναι η αυτοσυσχέτιση και η δεύτερη τα «πρότυπα διακύμανσης» (fluctuation patterns). Η μέθοδος αξιολογήθηκε στις συλλογές Ballroom και Songs.

Μέθοδος των Schreiber και Muller [Schreiber2014]

Η μέθοδος αυτή αποτελείται από δύο υποσυστήματα. Το ένα είναι η εξαγωγή μιας ΣΠ, χρησιμοποιώντας μια απλή συνάρτηση έμφασης βασισμένη στον λογάριθμο της ενέργειας. Η ΣΠ υπολογίζεται με ένα DFT στις συναρτήσεις έμφασης. Στην συνέχεια το χαρακτηριστικό Mean Spectral Novelty (SNM) χρησιμοποιείται για να φτιάξουν μία μέθοδο εκτίμησης της μουσικής οκτάβας. Υιοθέτησαν ένα γραμμικό μοντέλο παλινδρόμησης από την SNM στο αληθές τέμπο. Στη συνέχεια η εκτίμηση οκτάβας συνδυάζεται με την ΣΠ για την εξαγωγή του τελικού τέμπο. Η μέθοδος αξιολογήθηκε σε πολλές συλλογές και επέδειξε αξιόλογα αποτελέσματα. Ωστόσο πρέπει να σημειωθεί ότι το μοντέλο παλινδρόμησης εκπαιδεύτηκε στις συλλογές στις οποίες η μέθοδος αξιολογήθηκε.

Μέθοδος του Peeters [Peeters2012]

Η μέθοδος αυτή υιοθετεί παλινδρόμηση με Μοντέλα Γκαουσιανών Μίξεων (Gaussian Mixture Model Regression) σε μια ομάδα χαρακτηριστικών που αποτελούνται από την μεταβολή της ενέργειας (energy variation), την μεταβολή του αρμονικού περιεχομένου (harmonic variation), την μεταβολή της φασματικής ισορροπίας (spectral balance variation) και την επανάληψη γεγονότων βραχέως χρόνου (short-term event repetition). Ο στόχος της παλινδρόμησης είναι ο υπολογισμός είτε της κλάσης ταχύτητας, είτε του τέμπο. Η μέθοδος αξιολογήθηκε στις συλλογές Ballroom και Speedo. Ωστόσο οι συγγραφείς υιοθέτησαν 8% ανοχή (αντί 4%) για την μετρική *acc1*.

Μέθοδος EchoNest

Η μέθοδος EchoNest αποτελεί εμπορική μέθοδο που έχει αναπτυχθεί στα πλαίσια της αντίστοιχης πλατφόρμας[12], η οποία επεξεργάζεται αυτόματα μεγάλο όγκο μουσικών δεδομένων και παρέχει μεταδεδομένα αλλά και API συναφών τεχνολογιών. Η μέθοδος αυτή αξιολογήθηκε στο [Percival2014]. Πληροφορίες για την μέθοδο δεν είναι διαθέσιμες, ωστόσο υπάρχουν ενδείξεις ότι χρησιμοποιεί κάποιο υποσύστημα Μηχανικής Μάθησης.

Ο Πίνακας 5.3 παρουσιάζει τα αποτελέσματα των προτεινόμενων μεθόδων σε σύγκριση με τις μεθόδους αναφοράς που περιγράφηκαν, για τις 4 συλλογές για τις οποίες διεξήχθησαν αναλυτικά πειράματα στο πλαίσιο της παρούσας διατριβής. Τα αποτελέσματα σε συλλογές στις οποίες ο συγγραφέας δεν είχε πρόσβαση και εκτελέστηκαν από τρίτους θα παρουσιαστούν χωριστά στη συνέχεια. Χάριν συντομίας οι τρεις προτεινόμενες μέθοδοι θα αναφέρονται εφεξής στο κείμενο ως ΜΜΣ (Μέθοδος Μετρικών Σχέσεων), ΜΚΤ (Μέθοδος Κατηγοριοποίησης Ταχύτητας) και ΜΠΤ (Μέθοδος Πολλαπλών Ταξινομητών). Στη ΜΠΤ, όπου χρειάζεται, θα αναφέρεται και το είδος των χαρακτηριστικών (π.χ. ΣΠΤ-PCA, ΣΠΤ-ΣΠ κ.ο.κ). Τα σύμβολα +/- υποδηλώνουν σημαντική στατιστική διαφορά βάσει της δοκιμής A/B με p=0.05. Για την ΜΠΤ χρησιμοποιήθηκαν οι διαστάσεις των μετασχηματισμών που πέτυχαν την καλύτερη μέση ακρίβεια για όλες τις συλλογές. Πρέπει να σημειωθεί ότι στην εργασία [Percival2014] οι συγγραφείς διαπίστωσαν λάθος επισημειώσεις για τις συλλογές Ballroom, Speedo και Genres. Το ποσοστό αυτών των λαθών ήταν περίπου στο 5% των κομματιών κάθε συλλογής. Όπως αναφέρουν οι συγγραφείς, τα λάθη αυτά δεν προέκυψαν από λάθος επιλογή οκτάβας, αλλά οφείλονταν κυρίως σε επισημειώσεις τέμπο που δεν ήταν σχετικά με τη ρυθμική δομή. Στην ίδια εργασία παρείχαν και τις διορθωμένες επισημειώσεις. Επειδή κάποιες μέθοδοι στη βιβλιογραφία κατά τη δημοσίευσή τους ανέφεραν αποτελέσματα πριν την διόρθωση των συλλογών, οι μέθοδοι αυτοί είναι επισημειωμένες με * στον Πίνακα 5.3. Επιπλέον, για να παρουσιαστούν αξιόπιστα συγκριτικά αποτελέσματα με αυτές τις μεθόδους, η ΣΠΤ-ΣΠ αξιολογήθηκε και με τις παλιές επισημειώσεις.

Η 1^{η} γραμμή του Πίνακα 5.3 αναφέρεται στα αποτελέσματα της ΜΜΣ, ενώ η 2^{η} γραμμή στην μέθοδο ΜΚΤ. Οι επόμενες 4 γραμμές αντιστοιχούν στη ΜΠΤ και είναι ίδιες με αυτές του Πίνακα 5.2 (Ενότητα 5.4). Στη συνέχεια παρουσιάζονται τα αποτελέσματα των δύο ομάδων των μεθόδων αναφοράς.

[12] http://the.echonest.com/

	Ballroom	Songs	Speedo	Genres
Προτεινόμενες Μέθοδοι				
Μετρικές Σχέσεις-ΜΜΣ	63.2- / 98	58.5- / 91+	72.7- / 98+	71.7- / 93.9+
SVM+Κωδ. ΣΠ - ΜΚΤ	75.9-	63.9	-	-
Αρχική ΣΠ	89.8 / 96.4	68.8 / 80.9	80.6 / 95	77.8 / 91.2
Αρχική ΣΠ*	88.2 / 95.1	-	74.5 / 90.7	76.9 / 90.4
PCA (40)	89.7 / 96.3	69.9 / 81.5	80.3 / 95	77.6 / 90.9
wPCA (30)	87.3 / 95.1	71.2 / 83.7	79.7 / 94.6	76.4 / 90.7
RBM (200)	89.8 / 96.1	70.7 / 83.6	80.8 /95.3	77.8 / 91.8
Μέθοδοι Μηχανικής Μάθησης				
Percival (2014)	65.6- / 95	61.1-	73.3-	78.3
Peeters (2012)*	87	-	72.9	-
Seyerlehner (2007)*	78.5-	40.9-	-	-
Schreiber (2014)	66.3- / 96.4	73.1 / 91.8+	76.1- / 96	77.0 / 92.6
EchoNest	89.8 / 96.3	63.2 / 86+	72.1- / 94.9	72.5- / 91.6
Μέθοδοι χωρίς χρήση Μηχανικής Μάθησης				
VAMP [Davies2007]	66.9- / 90.8-	43- / 79.8	63.9- / 93-	58.8- / 87.7-
Tzanetakis (2013)*	61.3 / 89.3-	58.5- / 83.4	65.2- / 87.9-	73.6 / 90.1
Eck (2005) *	63- / 91-	60- / 79	-	-
Klapuri (2006)	64.9- / 92.8-	58.5- / 89.5+	68.9- / 96.9+	70.5- / 92.5

Πίνακας 5.3. Συγκριτικά αποτελέσματα των προτεινόμενων μεθόδων και μεθόδων αναφοράς. Τα σύμβολα +/- δείχνουν σημαντική στατιστική διαφορά (statistical significance) συγκριτικά με την μέθοδο «Αρχική ΣΠ».

Ωστόσο και οι άλλες δύο μέθοδοι επιτυγχάνουν πολύ καλά αποτελέσματα συγκριτικά με τις μεθόδους αναφοράς. Σχετικά με τη ΜΜΣ, αυτή επιτυγχάνει αποτελέσματα αντίστοιχα των άλλων μεθόδων που δεν κάνουν χρήση Μηχανικής Μάθησης βάσει της *acc1*, ενώ υπερτερεί σχεδόν όλων των μεθόδων βάσει της *acc2*. Η ΜΚΤ αξιολογήθηκε μόνο στις συλλογές Ballroom και Songs με τις παλιές (λάθος) επισημειώσεις.

Η επίδοσή της είναι καλύτερη από κάποιες μεθόδους αναφοράς και υπερτερεί της μεθόδου ΜΜΣ, καθώς φαίνεται ότι η κατηγοριοποίηση σε κλάσεις μουσικής ταχύτητας μειώνει τα λάθη οκτάβας όπως αναμενόταν. Η ΜΚΤ όμως είναι σίγουρα υποδεέστερη της ΜΠΤ. Η ΜΠΤ, που αποτελεί και την πιο ολοκληρωμένη προσέγγιση εξαγωγής τέμπο της παρούσας διατριβής, υπερτερεί αισθητά όλων των μεθόδων αναφοράς για όλες σχεδόν τις συλλογές. Μόνη εξαίρεση είναι η μέθοδος [Schreiber2014], η οποία υπερτερεί στη συλλογή Songs. Ωστόσο πρέπει να τονιστεί ότι αυτή η μέθοδος είναι έμμεσα εκπαιδευμένη στα δεδομένα αξιολόγησης λόγω της μεθόδου γραμμικής παλινδρόμησης που αναφέρθηκε νωρίτερα.

Συλλογή	#αρχείων	ΜΜΣ	Tzanetakis	zplane	Klapuri	echonest	ibt	VAMP
Αποτελέσματα Accuracy1 (%)								
Speedo	1410	72.7	**73.3**	70.1	68.9	72.1	63	63.9
Ballroom	698	63.2	65.6	66.9	64.9	**89.8**	64.3	66.9
Genres	999	71.7	**78.3**	68.9	70.5	72.5	61	58.8
Hainsworth	222	64.4	69.8	69.8	71.6	72.1	**72.5**	68
Songs	465	57	61.1	56.3	58.1	**63.2**	46.7	43
SMC	217	**35**	27.6	18.4	18	18.9	17.5	12.4
Μ.Ο / συλλογή		60.7	62.6	58.4	58.6	64.8	54.2	52.2
Μ.Ο / κομμάτι		66.5	69.1	64.8	64.7	71.4	58.9	58.2
Αποτελέσματα Accuracy2 (%)								
Speedo	1410	**98**	97.1	93.8	96.9	94.9	93.2	93
Ballroom	698	**98**	95	94.8	92.8	96.3	90.3	90.8
Genres	999	93.9	**94.7**	89.1	92.5	91.6	87	87.7
Hainsworth	222	84.7	**86.9**	82.4	84.2	84.2	82	77.5
Songs	465	**91**	86.7	82.6	89.5	86	76.6	79.8
SMC	217	**51.6**	45.6	31.8	41.9	34.1	36.9	30.9
Μ.Ο / συλλογή		86.2	84.3	79.1	83	81.2	77.6	76.6
Μ.Ο / κομμάτι		92.9	91.6	87.5	90.6	89.4	85.5	85.6

Πίνακας 5.4: Συγκριτικά αποτελέσματα μεθόδων εξαγωγής τέμπο όπως αναφέρονται στο [Percival2014]

Η μέθοδος ΜΜΣ αξιολογήθηκε επίσης σε επιπλέον συλλογές από τους Percival και Tzanetakis [Percival2014], όπου διεξήχθη ένα εκτενές συγκριτικό πείραμα με 8 αλγορίθμους σε 6 συλλογές. Τα συγκριτικά αποτελέσματα παρουσιάζονται στον Πίνακα 5.4. Τα αποτελέσματα είναι ιδιαίτερης αξίας, καθώς δεν αναφέρονται από τους ερευνητές που ανάπτυξαν τις μεθόδους, αλλά από τον ίδιο τον συγγραφέα του άρθρου, ο οποίος έτρεξε όλες τις μεθόδους. Εμφανίζονται δύο μέθοδοι αναφοράς που δεν έχουν περιγραφεί: η μέθοδος zplane[13] είναι μια εμπορική μέθοδος που χρησιμοποιείται σε διάφορα προϊόντα, και η μέθοδος ibt είναι μια δημοφιλής μέθοδος ανοιχτού κώδικα εξαγωγής παλμού που παρουσιάστηκε στο [Oliveira2012]. Επιπλέον εμφανίζονται δύο επιπλέον συλλογές, οι Hainsworth και SMC. Η συλλογή Hainsworth περιέχει αποσπάσματα με μεταβαλλόμενο τέμπο, ενώ η συλλογή SMC είναι σχεδιασμένη να είναι μια «δύσκολη» συλλογή. Περιέχει αποσπάσματα στα οποία ένα πλήθος μεθόδων εξαγωγής παλμού έβγαζαν διαφορετικά αποτελέσματα, δηλαδή δεν συμφωνούσαν ποιος είναι ο βασικός παλμός.

Παρατηρούμε ότι η προτεινόμενη μέθοδος υπερτερεί αισθητά από όλες τις υπόλοιπες μεθόδους βάσει της μετρικής *accuracy2* σε όλα τις συλλογές εκτός από τη συλλογή Hainsworth, όπου είναι η δεύτερη καλύτερη. Βάσει της *accuracy1* είναι δεύτερη στον μέσο όρο των κομματιών και 4η στον μέσο όρο των συλλογών.

[13] http://www.beat-tracking.com

Μέθοδος	P-score	Και τα δύο τέμπο σωστά	Τουλάχιστον ένα τέμπο σωστό
Bock2015	0.898	66,43	99,29
Bock2014	0.876	62,86	99,29
Elowsson2013	0.857	69,29	94,29
Gkiokas2011	0.829	62.14%	94.29%
Wu2013	0.826	55,00	95,71
Gkiokas2010	0.8099	50.00%	96.43%
Klapuri2006	0.806	61.43%	94.29%
Wack2010	0.7875	50.00%	91.43%
Wu2012	0.7783	55.71%	90.00%
Davies2006	0.776	45.71%	92.86%

Πίνακας 5.5: Συγκριτικά αποτελέσματα των καλύτερων μεθόδων στον διαγωνισμό MIREX για όλες τις χρονιές.

Ωστόσο η διαφορά στην επίδοση είναι μικρή. Ειδικά στην «δύσκολη» συλλογή SMC, η ΜΜΣ αποδίδει πολύ καλύτερα από όλες τις άλλες μεθόδους. Επιπλέον, οι συγγραφείς του άρθρου αναφέρουν το αν η διαφορά των αποτελεσμάτων είναι στατιστικής σημασίας βάσει της δοκιμής McNemar με p=0.01. Βάσει της μετρικής *accuracy1* βρέθηκε ότι οι πέντε καλύτερες μέθοδοι δεν έχουν στατιστική διαφορά, ενώ βάσει της μετρικής *accuracy2* βρέθηκε ότι η μέθοδός μας υπερτερούσε στατιστικά από όλες τις υπόλοιπες.

Τέλος, ο Πίνακας 5.5 παρουσιάζει τους καλύτερους σε επίδοση αλγορίθμους εξαγωγής τέμπο στον διεθνή διαγωνισμό MIREX. Λόγω του ότι τα δεδομένα αξιολόγησης παραμένουν τα ίδια όλα τα χρόνια (και κρυφά ακόμα σε όλους) παρουσιάζονται συγκριτικά αποτελέσματα για όλες τις χρονιές (2005 έως σήμερα). Ο αλγόριθμος που υποβλήθηκε το 2010 δεν έκανε χρήση του διαχωρισμού πηγών ούτε του μετρικού μοντέλου. Παρόλα αυτά επέδειξε επιδόσεις καλύτερες από όλες τις υπόλοιπες μεθόδους που είχαν υποβληθεί στον διαγωνισμό εκείνη τη στιγμή. Η χρήση του μετρικού μοντέλου και του διαχωρισμού πηγών (Έτος 2011) επέδειξε ακόμα καλύτερα αποτελέσματα και υπερτερεί σημαντικά από πολλές από τις υπόλοιπες μεθόδους.

Κεφάλαιο 6 : Εξαγωγή Παλμού

6.1 Εισαγωγή

Η διαδικασία εξαγωγής παλμού (beat tracking task) περιλαμβάνει την εύρεση των χρονικών στιγμών του παλμού στο βασικό μετρικό επίπεδο (Κεφ. 1) του κομματιού, δηλαδή τις χρονικές στιγμές που θα χτύπαγε ένας ιδανικός μετρονόμος. Όπως και στην περίπτωση του τέμπο, υπάρχει ο υποκειμενικός ανθρώπινος παράγοντας που εξαρτάται από ψυχοκοινωνικές συνθήκες, τη μουσική εκπαίδευση, τη μουσική προτίμηση κ.ά. (Κεφ. 1). Εναλλακτικά, η εξαγωγή παλμού μπορεί να οριστεί ως η πρόβλεψη των χρονικών στιγμών στις οποίες ένας ακροατής θα χτυπούσε το χέρι ή το πόδι του με τον ρυθμό. Πέρα από το ενδιαφέρον για την μελέτη και μοντελοποίηση της ανθρώπινης απόκρισης στον ρυθμό, η εύρεση παλμού μπορεί να βρει πολλές εφαρμογές όπως στην αυτόματη μεταγραφή, σε συστήματα μουσικής αλληλεπίδρασης [Robertson2007], εφέ ήχου βάσει περιεχομένου [Stark2007] [Hockman2008], καθώς και τη χρονική κατάτμηση μουσικών κομματιών για άλλες εφαρμογές όπως για την αναγνώριση συγχορδίας [Bello2005b], τη δομική κατάτμηση [Levy2008] και την μουσική ομοιότητα [Ellis2008].

Ενώ όπως είδαμε στην περίπτωση της εξαγωγής του μουσικού τέμπο και της περιοδικής ανάλυσης γενικότερα, υπάρχει μια συστηματική μεθοδολογία και οι πιο πολλές μέθοδοι έχουν κάποια κοινά χαρακτηριστικά, στην περίπτωση της αναζήτησης μουσικού παλμού δεν συναντάται κάτι τέτοιο. Παρουσιάζεται μια μεγάλη ποικιλία μεθόδων οι οποίες είναι δύσκολο να ομαδοποιηθούν σε οικογένειες προσεγγίσεων. Δύο δυνατές ομαδοποιήσεις μπορούν να γίνουν βάσει της αιτιατότητας μιας μεθόδου και του τρόπου με τον οποίο υπολογίζεται ο παλμός σε σχέση με το μουσικό τέμπο.

Αιτιατότητα των μεθόδων:

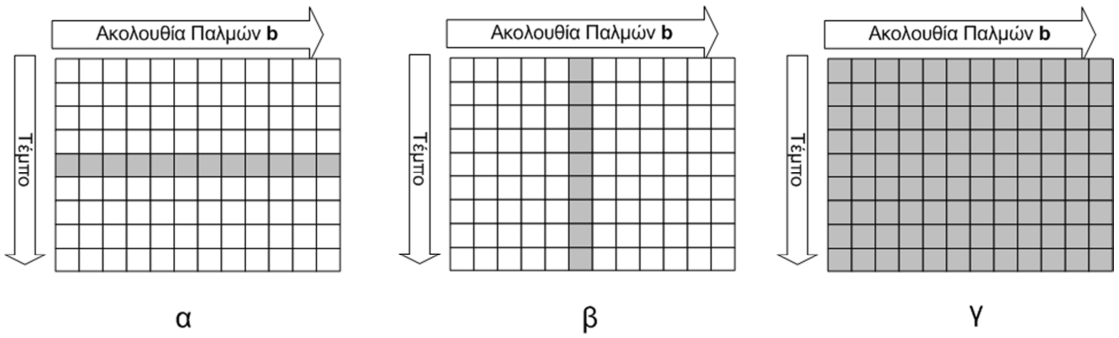
Αιτιατές μέθοδοι: Λέγονται και «εντός γραμμής» (on-line). Η εύρεση μιας θέσης παλμού γίνεται με βάση τις παρελθοντικές μόνο τιμές του σήματος. Όταν η επεξεργασία ενός αποσπάσματος χρονικής διάρκειας t διαρκεί λιγότερο από t, τότε οι αλγόριθμοι αυτοί λέγονται και «πραγματικού χρόνου» (real time).

Μη αιτιατές μέθοδοι: Χρειάζονται και μελλοντικά δείγματα στο χρόνο για την εύρεση της θέσης του παλμού. Προφανώς οι μη αιτιατές μέθοδοι επιτυγχάνουν καλύτερα αποτελέσματα.

Σχέση Παλμού - Τέμπο

Κάθε σύστημα εύρεσης παλμού προφανώς θα πρέπει να υπολογίζει μια ΣΠ και να εξάγει με έμμεσο ή άμεσο τρόπο μια τιμή του τέμπο. Ο τρόπος με τον οποίο συνδυάζονται (ή όχι) η εύρεση του παλμού και του τέμπο μάς δίνει μια άλλη κατηγοριοποίηση των μεθόδων:

- Ο υπολογισμός του τέμπο προηγείται της εξαγωγής του μουσικού παλμού. Επομένως η εξαγωγή του παλμού προϋποθέτει γνώση κάποιου μουσικού τέμπο.



Σχήμα 6.1. Εύρεση παλμού και τέμπο. Ο οριζόντιος άξονας αντιστοιχεί στις ακολουθίες παλμών $b_i \subseteq R^n$ ενώ ο κάθετος σε μία πραγματική τιμή του τέμπο. Κάθε στοιχείο του πλέγματος αντιστοιχεί σε μία ακολουθία παλμών για μία τιμή του τέμπο. α) Η αναζήτηση ακολουθίας παλμών γίνεται για μία συγκεκριμένη τιμή του τέμπο. β) Η εξαγωγή του παλμού γίνεται χωρίς τη γνώση του τέμπο. Το τέμπο εξάγεται από τον παλμό. γ) Η αναζήτηση παλμού – τέμπο γίνεται ταυτόχρονα.

- Η εξαγωγή του παλμού γίνεται χωρίς την γνώση κάποιου αρχικού τέμπο. Το τέμπο συμπεραίνεται εκ των υστέρων από τη συχνότητα του εξαχθέντος παλμού.
- Ο υπολογισμός του τέμπο και του παλμού γίνεται ταυτόχρονα.

Δεδομένου ενός σταθερού τέμπο σε ένα κομμάτι, θα μπορούσαμε να πούμε ότι το πρόβλημα της εξαγωγής παλμού ισοδυναμεί με την έρευση της χρονικής στιγμής ενός μόνο παλμού. Γνωρίζοντας αυτή τη θέση, οι θέσεις των υπολοίπων παλμών προκύπτουν έμμεσα προσθέτοντας ή αφαιρώντας διαδοχικά τη περίοδο που αντιστοιχεί στο τέμπο. Γι' αυτό το λόγο, στο πλαίσιο της ρυθμικής ανάλυσης μπορεί να συναντήσουμε την ορολογία ότι η εύρεση του τέμπο αντιστοιχεί στην συχνότητα ενώ του παλμού στην εύρεση της φάσης (θεωρώντας ότι μπορούμε να αναπαραστήσουμε τον ρυθμό με ένα στοιχείο που ταλαντώνεται σε μία συχνοτητα και μία φάση), όπως στο παράδειγμα του «πηδήματος φάσης» (Ενότητα 1.2.5).

Αν θέλαμε να αναπαραστήσουμε γραφικά τον τρόπο με τον οποίο αυτές οι τρεις ομάδες πραγματοποιούν την εύρεση παλμού, θα είχαμε ένα πλέγμα συχνότητας (ή τέμπο) και φάσης (ή ακολουθιών παλμών), όπου η μία διάσταση θα αναπαριστούσε την συχνότητα (τέμπο) ενώ η δεύτερη την ακολουθία των παλμών, όπως φαίνεται στο Σχ. 6.1. Κάθε σημείο αυτού του πλέγματος θα αντιστοιχούσε και σε μία λύση στο πρόβλημα της εύρεσης παλμού και τέμπο. Η πρώτη ομάδα μεθόδων θα περιόριζε την εύρεση της λύσης σε μία γραμμή κατά μήκος των τέμπο (Σχ. 6.1α), αφού πρώτα θα εύρισκε το τέμπο και στη συνέχεια δεδομένου του τέμπο θα αναζητούσε τη βέλτιστη ακολουθία παλμών. Αντίθετα, η δεύτερη ομάδα θα εύρισκε πρώτα τις θέσεις των παλμών και στη συνέχεια θα αναζητούσε το τέμπο σε μία γραμμή κατά μήκος των θέσεων του παλμού (Σχ. 6.1β). Αντιθέτως, η τρίτη μέθοδος θα αναζητούσε ταυτόχρονα το τέμπο και τις θέσεις των παλμών σε όλο τον χώρο (Σχ. 6.3γ).

Ένα από τα κοινά χαρακτηριστικά που μοιράζονται οι περισσότερες μέθοδοι είναι μία συνάρτηση έμφασης, παρόμοια με αυτήν που περιγράψαμε στην ανάλυση περιοδικότητας (Κεφ. 2). Σε αυτή την περίπτωση η συνάρτηση έμφασης μπορεί να θεωρηθεί είτε σαν μια πρότερη πιθανότητα κάθε χρονικής στιγμής να

είναι θέση beat (συνεχείς αναπαραστάσεις), είτε σαν τις υποψήφιες θέσεις των beats (διακριτές αναπαραστάσεις της συνάρτηση έμφασης, π.χ. onsets). Χωρίς καμία εκ των προτέρων γνώση για το τέμπο, όσο πιο ισχυρή είναι η συνάρτηση έμφασης μια χρονική στιγμή, τόσο πιο πιθανό είναι αυτή η χρονική στιγμή να είναι θέση του παλμού.

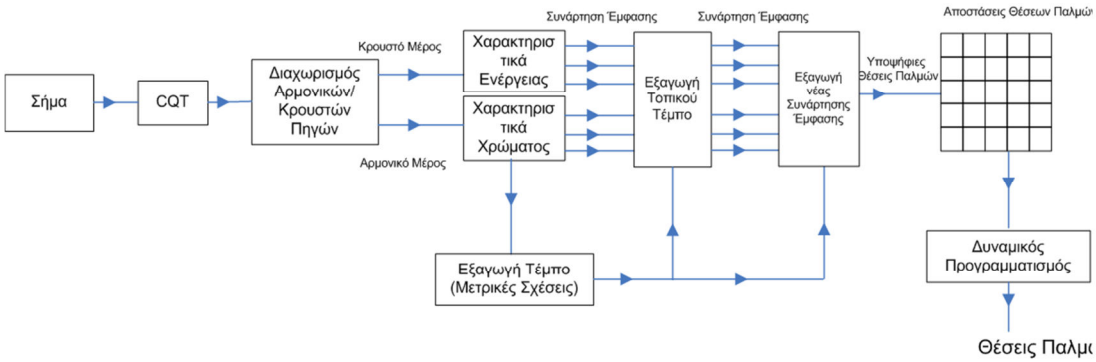
Τυπικές Μέθοδοι Εύρεσης Παλμού από τη βιβλιογραφία

Ο S. Dixon [Dixon2001] πρότεινε μια μέθοδο βασισμένη σε "πράκτορες" (agents) για να αναλύσει υποψήφιες ακολουθίες παλμών, η οποία δεχόταν ως είσοδο είτε συμβολικές αναπαραστάσεις (midi), είτε απευθείας την κυματομορφή. Στη δεύτερη περίπτωση η εξαγωγή διακριτών onsets προηγούταν οποιαδήποτε ρυθμικής ανάλυσης. Οι Davies και Plumbley [Davies2007] υιοθέτησαν την Φασματική Μιγαδική Διαφορά (Spectral Complex Difference, Εξ. 2.1) ως συνάρτηση έμφασης. Πρότειναν ένα μοντέλο δύο καταστάσεων για να περιγράψουν τις ασυνέχειες της ακολουθίας παλμών που οφείλονταν σε αλλαγές μετρικών επιπέδων. Στα [Eyben2010, Bock2011] οι συγγραφείς επέκτειναν ένα σύστημα ανίχνευσης onsets βασισμένο σε πολυκατευθυντικά αναδρομικά νευρωνικά δίκτυα βραχείας-μακράς μνήμης (Bidirectional Long Short-Term Memory Recurrent Neural Networks) σε ένα σύστημα παρακολούθησης παλμών. Ως είσοδος στο νευρωνικό δίκτυο χρησιμοποιήθηκαν οι παράγωγοι της ενέργειας φίλτρων στην κλίμακα mel μετά από ημιανόρθωση. Το προτεινόμενο σύστημα πέτυχε από τις καλύτερες επιδόσεις στον διαγωνισμό beat-tracking MIREX τις χρονιές 2010-2012. Στο [Ellis2007] ο συγγραφέας διατύπωσε το πρόβλημα της εύρεσης παλμού ως πρόβλημα βελτιστοποίησης και υιοθέτησε τεχνικές δυναμικού προγραμματισμού για την επίλυσή του. Παρόμοια με τα προηγούμενα, χρησιμοποίησε τις παραγώγους των ενεργειών των εξόδων φίλτρων στη κλίμακα mel ακολουθούμενη από ημιανόρθωση. Οι Peeters και Papadopoulos [Papadopoulos2011] χρησιμοποίησαν τεχνική ταιριάσματος σε «υποδείγματα ακολουθιών παλμών» (beat templates) σε στοχαστικό περιβάλλον για να εξάγουν τον παλμό. Στο [Laroche2003] ο συγγραφέας χρησιμοποίησε την φασματική ροή (Spectral Flux) ως συνάρτηση έμφασης. Στο [Holzapfel2012] οι συγγραφείς χρησιμοποίησαν τέσσερα υπάρχοντα συστήματα και βάσει της «αμοιβαίας συμφωνίας» (mutual agreement) προέκυπτε η εξαγόμενη ακολουθία παλμών. Στο [Zapata2013] οι συγγραφείς πρότειναν την εφαρμογή τεχνικών «ακύρωσης φωνής» (voice suppression) σαν στάδιο προεπεξεργασίας για την βελτίωση της επίδοσης υπαρχόντων συστημάτων.

6.2 Αρχιτεκτονική της Προτεινόμενης Μεθόδου

Λόγω της υψηλής επίδοσης όλων των προτεινόμενων μεθόδων εξαγωγής τέμπο που παρουσιάστηκαν στο προηγούμενο κεφάλαιο, επιλέχτηκε η μέθοδος εύρεσης του παλμού να ακολουθεί την εξαγωγή του τέμπο ($1^η$ περίπτωση Σχ. 6.1).

Αρχικά εξάγεται η τιμή του τέμπο ενός μουσικού κομματιού, και στη συνέχεια η τιμή αυτή χρησιμοποιείται ως πρότερη πληροφορία για την εξαγωγή των παλμών. Στο Σχήμα 6.2 παρουσιάζεται η συνολική αρχιτεκτονική της προτεινόμενης μεθόδου. Ένα μεγάλος μέρος της παρουσιάζει ομοιότητες με την εξαγωγή του τέμπο. Στο σήμα εισόδου εφαρμόζεται ο CQT και στη συνέχεια η μέθοδος διαχωρισμού κρουστών/αρμονικών πηγών. Στη συνέχεια εξάγονται τα χαρακτηριστικά έμφασης όπως περιγράφτηκε στην Ενότητα 2.4.



Σχήμα 6.2: Συνολικό διάγραμμα των σταδίων επεξεργασίας της μεθόδου αυτόματης εξαγωγής παλμών. Το στάδιο της εξαγωγής τέμπο αντιστοιχεί στη μέθοδο ΜΜΣ. Τα 4 πρώτα στάδια είναι ταυτόσημα με αυτά της μεθόδου εξαγωγής της συνάρτησης περιοδικότητας.

Από τη συνάρτηση έμφασης εξάγεται ένα γενικό τέμπο για όλη τη διάρκεια του κομματιού με κάποια από τις προαναφερόμενες μεθόδους (Κεφ. 5). Η διαδικασία της εύρεσης παλμού ξεκινάει με την εύρεση ενός «τοπικού» τέμπο, όπου βάσει της τιμής του τέμπο και της συνάρτησης έμφασης υπολογίζει ασθενείς μεταβολές του τέμπο ως προς τον χρόνο. Στη συνέχεια εξάγεται μια νέα συνάρτηση έμφασης η οποία χρησιμοποιείται στην εύρεση παλμού. Αυτή η συνάρτηση έμφασης είναι διαφορετική από αυτήν που χρησιμοποιείται για την εύρεση του τέμπο και ενσωματώνει πληροφορία που σχετίζεται με το τέμπο και τις μεταβολές του. Από τη νέα συνάρτηση έμφασης υπολογίζονται οι υποψήφιες θέσεις του παλμού. Αν οι υποψήφιες θέσεις είναι k σε πλήθος, τότε υπολογίζονται $k \times k$ αποστάσεις μεταξύ τους. Όσο πιο «κοντά» είναι δύο υποψήφιες θέσεις, τόσο πιο πιθανό είναι να είναι η μία θέση παλμού δεδομένου ότι είναι και η άλλη θέση παλμού. Η απόσταση μεταξύ δύο υποψηφίων παλμών λαμβάνει υπόψη τη ρυθμική συνοχή μεταξύ τους, καθώς και την συνάρτηση έμφασης στις θέσεις αυτές. Η εξαγωγή της ακολουθίας παλμών ολοκληρώνεται με δυναμικό προγραμματισμό, όπου υπολογίζεται το μονοπάτι για το οποίο η απόσταση των διαδοχικών beats να ελαχιστοποιείται.

Η μέθοδος που παρουσιάζεται στο Σχ. 6.2 μπορεί να χρησιμοποιήσει οποιαδήποτε από τις μεθόδους εύρεσης τέμπο που παρουσιάστηκαν στο Κεφ. 5. Ωστόσο επιλέχτηκε η μέθοδος με τη χρήση μετρικών σχέσεων. Αυτό έγινε για δύο λόγους. Ο πρώτος είναι ότι η προτεινόμενη μέθοδος αναπτύχθηκε χρονικά την ίδια περίοδο που αναπτύχθηκε και η ΜΜΣ και μοιράζονται αρκετά κοινά χαρακτηριστικά. Ο δεύτερος είναι ότι η ΜΜΣ δεν περιλαμβάνει κάποιο βήμα μάθησης. Δεδομένου ότι η προτεινόμενη μέθοδος εξαγωγή παλμού είναι «μη-επιβλεπόμενη», δηλαδή δεν χρειάζεται επισημειωμένα παραδείγματα για εκπαίδευση και δεν περιλαμβάνει κάποιο υποσύστημα Μηχανικής Μάθησης, είναι πιο «φυσικό» να εμπεριέχει μια αντίστοιχη μέθοδο για την εύρεση του τέμπο ως υποσύστημά της. Η χρήση της ΜΜΣ αντί των άλλων ενισχύεται περισσότερο από το γεγονός ότι η ΜΜΣ υπολείπεται των άλλων δύο μεθόδων όσον αφορά την *acc1* αλλά όχι την *acc2*. Η *acc2* στην ΜΜΣ είναι ελαφρώς μεγαλύτερη, αφού στις δύο άλλες μεθόδους τα ενδεχόμενα λάθους επιλογής κατηγορίας ταχύτητας (στην περίπτωση της ΜΚΤ) ή της δημιουργίας μάσκας που δεν είναι ρυθμικά σχετική (λόγω λαθών των επιμέρους ταξινομητών) στη περίπτωση της ΜΠΤ, μπορούν να αναδείξουν ρυθμικά μη σχετικά τέμπο ως τα πιο εξέχοντα, μειώνοντας έτσι την *acc2*. Τα λάθη στην *acc2* είναι πολύ πιο κρίσιμα για την εύρεση του παλμού απ'

ότι τα λάθη οκτάβας. Αν για παράδειγμα το πραγματικό τέμπο είναι 100 BPM για ένα κομμάτι, ένας παλμός που θα εξαχθεί στα 90 BPM θα είναι τελείως άσχετος με το ρυθμικό περιεχόμενο του κομματιού, αντίθετα με έναν παλμό στα 50 ή 200 BPM.

Στην Ενότητα 6.3 θα περιγραφεί η μέθοδος εξαγωγής του τοπικού τέμπο. Στην συνέχεια στην Ενότητα 6.4 θα παρουσιαστεί η 2^{η} συνάρτηση έμφασης (Σχ. 6.2) και η διαίσθηση πίσω από την επιλογή η συνάρτηση έμφασης να ενσωματώνει και πληροφορία για το πραγματικό τέμπο. Η Ενότητα 6.5 περιγράφει την επιλογή των υποψήφιων θέσεων παλμών, τον ορισμό της απόστασης μεταξύ τους και την εύρεση του βέλτιστου μονοπατιού. Αναλυτικά πειραματικά αποτελέσματα και σύγκριση με τη διεθνή βιβλιογραφία παρουσιάζονται στην Ενότητα 6.6.

6.3 Εξαγωγή Τοπικού Τέμπο και Μεταβολών

Η διαδικασία εξαγωγής του τοπικού τέμπο γύρω από το τέμπο T_0 που υπολογίστηκε με την ΜΜΣ γίνεται με παρόμοιο τρόπο, εφαρμόζοντας μια σειρά τροποποιήσεων των παραμέτρων του συστήματος ώστε να είναι κατάλληλες για την αντιμετώπιση των τοπικών μεταβολών. Ουσιαστικά χρησιμοποιείται η *ίδια διαδικασία ανάλυσης περιοδικότητας* με λιγότερους ταλαντωτές που είναι πιο ευαίσθητοι στην αλλαγή του τέμπο. Με αυτό τον τρόπο επιτυγχάνεται αφενός η ενοποίηση των δύο μεθόδων, αφετέρου επιτρέπει την τροποποίηση των παραμέτρων αυτών κατά περίπτωση. Κατά συνέπεια η προτεινόμενη αρχιτεκτονική προσφέρει ένα πλαίσιο εργασίας για την εύρεση απότομων μεταβολών τέμπο ή και την αλλαγή ρυθμού. Εφεξής θα χρησιμοποιήσουμε τα ίδια σύμβολα που χρησιμοποιήθηκαν στις Ενότητες 2.5-2.6 για τον υπολογισμό της ΣΠ.

Στην περίπτωση εξαγωγής του τοπικού τέμπο, το εύρος των τέμπο στόχων περιορίζεται στο διάστημα $[(1-\beta)T_0, (1+\beta)T_0]$ για κάποιο μικρό $\beta > 0$ και για όλες τις δυνατές διακριτές περιόδους $\{\tau_k\}_k$ που αντιστοιχούν σε αυτό το εύρος των τέμπο. Επομένως μπορούμε να ξαναγράψουμε την Εξ. 2.19 ως

$$r_{\mathbf{x}}[m,T] = (\mathbf{x} * \mathbf{o}_T)[m],\ T \in [(1-\beta)T_0, (1+\beta)T_0. \tag{6.1}$$

Επιπλέον, το μήκος του παραθύρου κατά την κατάτμηση (το οποίο όπως έχει αναφερθεί στην Ενότητα 2.6 είναι μεταβλητό και ανάλογο της περιόδου που αντιστοιχεί στο τέμπο) επιλέγεται αρκετά μικρότερο, ίσο με $Q = 2$ κύκλους του τέμπο αναφοράς και με ολίσθηση $Q_{hop} = 1$ κύκλο. Αντίστοιχα, το $\boldsymbol{Q_0}$ των ταλαντωτών που ορίζει το μήκος της (μη μηδενικής) κρουστικής τους απόκρισης (Εξ. 2.18) τίθεται και αυτό ίσο με 2.

Αναμένουμε ότι η χρήση μικρού $\boldsymbol{Q_0}$ θα έχει δύο επιδράσεις: α) θα μειώσει την αξιοπιστία του υπολογισμού του τέμπο, αφού όσο πιο μεγάλο είναι το πλήθος των ταλαντώσεων του ταλαντωτή, τόσο πιο «αξιόπιστη» είναι η ΣΠ και β) θα αυξήσει την χρονική διακριτική ικανότητα. Ο περιορισμός του εύρους των τέμπο στόχων εξομαλύνει την επίδραση (α) αφού ο χώρος αναζήτησης του τέμπο (γύρω από μία κεντρική τιμή που είναι κοντά στο πραγματικό τέμπο) μικραίνει, διατηρώντας ταυτόχρονα την επιθυμητή επίδραση (β). Μια άλλη σημαντική τροποποίηση της μεθόδου εξαγωγής της ΣΠ για να ακολουθεί τις μεταβολές του τέμπο είναι η τροποποίηση του υπολογισμού της απόκρισης των ταλαντωτών, ώστε να έχουμε μηδενική καθυστέρηση φάσης. Στην περίπτωση εξαγωγής του ολικού τέμπο, η καθυστέρηση φάσης των ταλαντωτών δεν επηρεάζει το τελικό αποτέλεσμα, αφού στο τέλος γίνεται άθροιση κατά μήκος ολόκληρου του

κομματιού. Στην περίπτωση όμως του υπολογισμού του τοπικού τέμπο, αυτή η καθυστέρηση φάσης μπορεί να δημιουργήσει σφάλματα.

Για να επιτευχθεί μηδενική καθυστέρηση φάσης χρησιμοποιείται η τεχνική Forward-Backward Filtering. Η ίδια η ονομασία καταδεικνύει την φιλοσοφία της τεχνικής. Στο πρώτο βήμα (forward) τα σήματα εισόδου **x** (τα χαρακτηριστικά έμφασης στην περίπτωσή μας) συνελίσσονται με τους ταλαντωτές $o_T[m]$

$$r_{\mathbf{x}}[m,T] = (\mathbf{x} * \mathbf{o}_T)[m] \tag{6.2}$$

Στη συνέχεια το σήμα εξόδου $r_{\mathbf{x}}[n,T]$ φιλτράρεται πάλι από τον ταλαντωτή αλλά από την αντίθετη κατεύθυνση (backward). Συγκεκριμένα το ανάστροφο του $\tilde{r}_{\mathbf{x}}[m,T] = r_{\mathbf{x}}[-m,T]$ συνελίσσεται με τον ταλαντωτή $o_T]$

$$\check{r}_{\mathbf{x}}[m,T] = (\tilde{r}_{\mathbf{x}}[\cdot,T] * o_T[\cdot])[m] \tag{6.3}$$

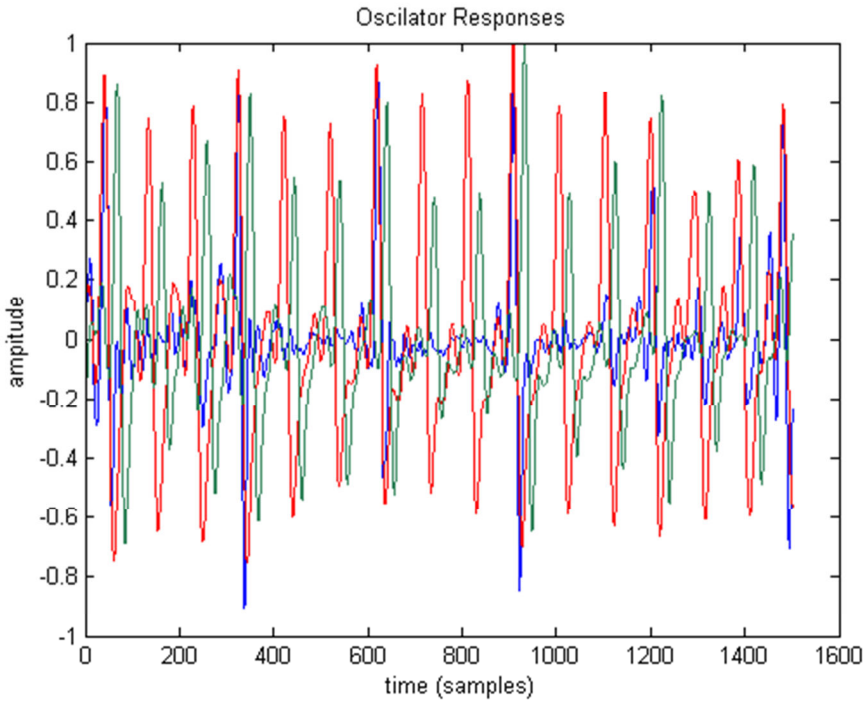
Το τελικό σήμα εξόδου λαμβάνεται με αναστροφή πάλι του $\check{r}_{\mathbf{x}}[m,T]$:

$$\hat{r}_{\mathbf{x}}[m,T] = \check{r}_{\mathbf{x}}[-m,T] \tag{6.4}$$

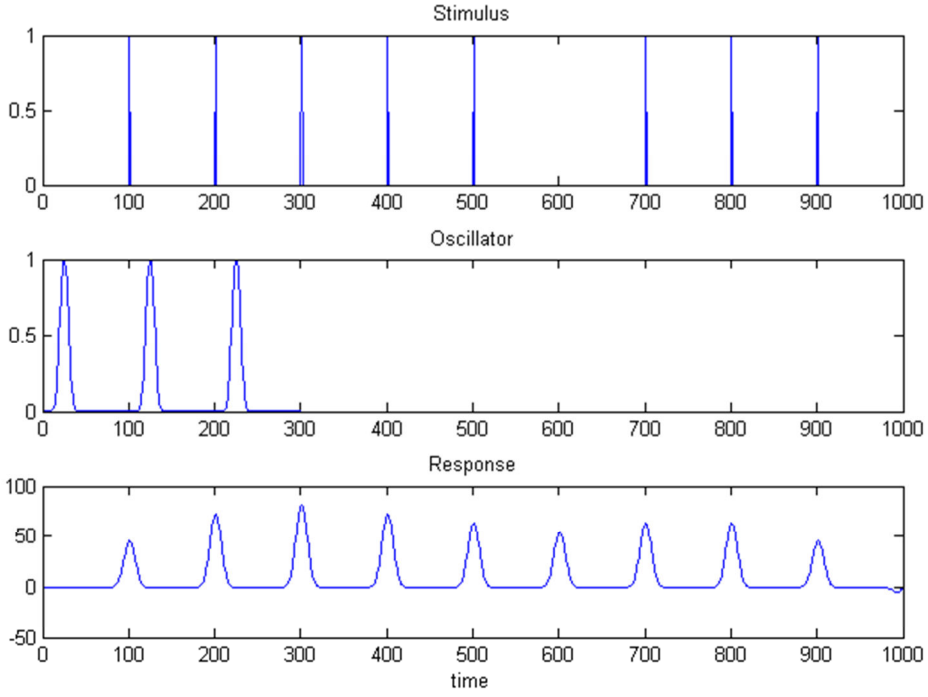
Το προκύπτων σήμα έχει μηδενική απόκριση φάσης και απόκριση πλάτους ίση με το τετράγωνο του αρχικού φίλτρου $o_T[m]$.

Στο Σχήμα 6.2 παρουσιάζεται η απόκριση του αιτιατού φίλτρου και του φίλτρου μηδενικής φάσης με είσοδο την παράγωγο της ενέργειας της τέταρτης μπάντας για κάποιο μουσικό κομμάτι. Η συχνότητα ταλάντωσης των δύο φίλτρων αντιστοιχεί σε 124 BPM (το πραγματικό τέμπο του κομματιού), η παράμετρος Q_0 είναι ίση με 4 και η συχνότητα δειγματοληψίας ίση με 200 samples/sec. Παρατηρούμε ότι οι κορυφές στην περίπτωση του αιτιατού φίλτρου έχουν καθυστέρηση ~20-25 samples που αντιστοιχεί σε χρονικό διάστημα >100ms. Οι διάφορες μετρικές αξιολόγησης μεθόδων αυτόματης εξαγωγής παλμού επιτρέπουν διάστημα ανοχής στις εξαγόμενες τιμές το πολύ 70ms (MIREX), επομένως η εξάλειψη του φαινομένου της διαφοράς φάσης είναι εξαιρετικά κρίσιμο. Αντίθετα, στην περίπτωση του φιλτραρίσματος μηδενικής φάσης παρατηρούμε ότι το σήμα εισόδου και η απόκριση εμφανίζουν κορυφές στις ίδιες ακριβώς χρονικές στιγμές. Επιπλέον το Σχήμα 6.2 καταδεικνύει και κάποια άλλη σημαντική ιδιότητα της διαδικασίας της συνέλιξης με ταλαντωτές.

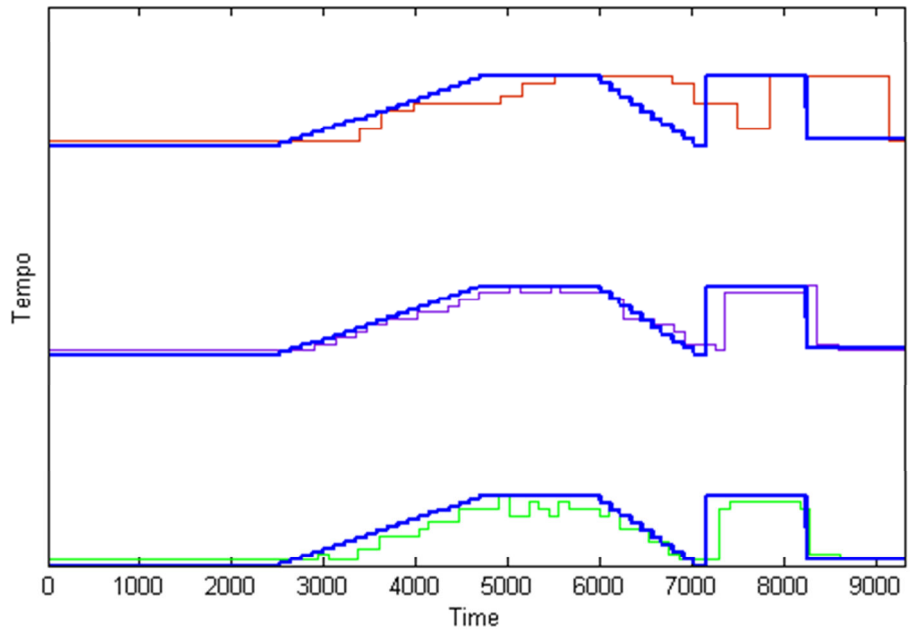
Λόγω του ότι η κρουστική απόκριση των ταλαντωτών είναι μεγαλύτερη από την περίοδο του τέμπο (κατά λόγο Q_0) οι ταλαντωτές δεν εμφανίζουν κορυφές μόνο στις χρονικές στιγμές που και το σήμα εισόδου εμφανίζει κορυφές, αλλά και σε χρονικές στιγμές που είναι ρυθμικά συνεπείς με το εξαγόμενο τέμπο αλλά μπορεί σε αυτές το σήμα εισόδου να μη εμφανίζει καθόλου (ή έστω ασθενείς) κορυφές. Με τον τρόπο αυτό οι ταλαντωτές «υπονοούν» υποψήφιες θέσεις του παλμού που μπορεί να μην είναι παρατηρήσιμες απευθείας από το σήμα εισόδου. Ας θεωρήσουμε το εξής απλό παράδειγμα που φαίνεται στο Σχ. 6.3. Έχουμε μία ακολουθία παλμών με σταθερή περίοδο, από την οποία λείπει ένας παλμός ($6^{ος}$ παλμός) και έναν ταλαντωτή με ίδια εγγενή περίοδο. Στην απόκριση όμως του ταλαντωτή η έλλειψη του $6^{ου}$ παλμού του σήματος εισόδου δεν είναι καθόλου εμφανής. Ωστόσο, αν έλειπαν 4 παλμοί, τότε η απουσία του $4^{ου}$ παλμού θα φαινόταν στην έξοδο του ταλαντωτή αφού η κρουστική του απόκριση έχει μήκος 3 παλμών. Περαιτέρω ανάλυση αυτής της ιδιότητας θα γίνει στην επόμενη παράγραφο.



Σχήμα 6.2 Μπλε: Σήμα εισόδου (συνάρτηση έμφασης). Πράσινο: Απόκριση αιτιατού φίλτρου. Κόκκινο: Απόκριση φίλτρου μηδενικής φάσης.



Σχήμα 6.3 Πάνω: Ισόχρονος παλμός που λείπει μία κρουστική (frame 600). Μέση: Ταλαντωτής ίδιας συχνότητας με τον παλμό. Κάτω: Η απόκριση του ταλαντωτή. Ο παλμός που λείπει δεν είναι σχεδόν καθόλου εμφανής στην απόκριση του ταλαντωτή.



Σχήμα 6.4. Απόκριση στις μεταβολές του τέμπο. Μπλε: Πραγματική μεταβολή τέμπο. Κόκκινο: Απόκριση αιτιατού συστήματος ($Q = 8$). Μωβ: Απόκριση μη αιτιατού συστήματος ($Q = 4$). Πράσινο: Απόκριση μη αιτιατού συστήματος ($Q = 2$)

Η εξαγωγή του τοπικού τέμπο γίνεται για τις χρονικές στιγμές $t_k = k \cdot \tau_0,\ k = \mathbf{1} \dots K$, όπου τ_0 είναι η περίοδος του «κεντρικού» τέμπο T_0. Το τοπικό τέμπο σε κάθε χρονική στιγμή επιλέγεται ως το τέμπο του ταλαντωτή με την μέγιστη απόκριση στο διάστημα $t_k - Q \cdot \frac{\tau_0}{2},\ t_k + Q \cdot \tau_0/\mathbf{2}$, δηλ.

$$T_L[t_k] = \mathbf{argmax}_T\{\sum_{\mathbf{x}} \hat{r}_{\mathbf{x}}[m,T]\ ,\ m \in [\underline{M}, \overline{M}]\},\ \underline{M} = t_k - Q \cdot \frac{\tau_0}{\mathbf{2}},\ \overline{M} = t_k + Q \cdot \tau_0/\mathbf{2} \quad (6.5)$$

όπου Q είναι το (σχετικό) μήκος του παραθύρου ανάλυσης. Η καμπύλη $T_L[t_k]$ μας δείχνει τις μεταβολές του τέμπο σε συνάρτηση με τον χρόνο και θα χρησιμοποιηθεί σαν πληροφορία τόσο στην εξαγωγή της συνάρτησης έμφασης όσο και στον ορισμό των «αποστάσεων» μεταξύ των υποψηφίων θέσεων beats. Εφεξής θα συμβολίζουμε με $T[m]$ το τοπικό τέμπο τη χρονική στιγμή m όπως αυτό προκύπτει από την Εξ. 6.5.

Για να δείξουμε την ικανότητα της προτεινόμενης μεθόδου να αποκρίνεται επαρκώς στις τοπικές μεταβολές του τέμπο, δημιουργήσαμε ένα πρότυπο σήμα που αποτελείται από χτυπήματα «μπότας» (τύμπανο ντραμς) με μεταβλητό τέμπο. Στο Σχήμα 6.4 παρουσιάζεται το πραγματικό τέμπο σε σχέση με τον χρόνο, μαζί με

- Την απόκριση του αιτιατού ταλαντωτή ($Q_0 = 8$)
- Την απόκριση δύο μη-αιτιατών ταλαντωτών με τη τεχνική Forward-Backward και με τιμές $Q_0 = 2$ και $Q_0 = 4$

Παρατηρούμε ότι στην περίπτωση του αιτιατού ταλαντωτή υπάρχει μία σημαντική καθυστέρηση της απόκρισης. Στην περίπτωση του μη-αιτιατού ταλαντωτή με $Q_0 = 4$ παρατηρείται πάλι μια διαφορά φάσης, αλλά αρκετά

μικρότερη. Η διαφορά φάσης πάντως δεν οφείλεται τόσο στην μη αιτιατότητα αλλά κυρίως στην σχετικά υψηλή τιμή του Q_0. Όταν το Q_0 μικρύνει ($Q_0 = 2$) παρατηρούμε ότι η απόκριση του συστήματος είναι πιο κοντά στην πραγματική αλλά με μεγαλύτερο σφάλμα στις εκτιμώμενες τιμές του τέμπο (απόκλιση κατά τον κάθετο άξονα). Ωστόσο η χρήση της τεχνικής Forward-Backward Filtering είναι απαραίτητη στη συνάρτηση έ,φασης.

6.4 Συνάρτηση ΄Εμφασης

Συνήθως τα χαρακτηριστικά έμφασης συναντούνται στη βιβλιογραφία ως η μεταβολή κάποιου φασματικού χαρακτηριστικού ενέργειας. Στη παρούσα μέθοδο, η υιοθέτηση των χαρακτηριστικών έμφασης που προκύπτουν από τον διαχωρισμό πηγών και χρησιμοποιούνται στη συνέχεια στην ανάλυση περιοδικότητας είναι μία υποψήφια λύση. Ωστόσο υπάρχουν περιθώρια βελτίωσης. Η χρησιμοποίηση της πληροφορίας του σωστού τέμπο για περαιτέρω επεξεργασία των χαρακτηριστικών πιθανόν να έδινε μια πιο αποτελεσματική αναπαράσταση. Μια τέτοια επεξεργασία θα πρέπει να «φιλτράρει» εμφατικές περιοχές στο σήμα που είναι μη σχετιζόμενες με τον σωστό τέμπο, ενώ θα πρέπει να ενισχύει τις σχετικές. Επιπλέον θα πρέπει να διαχειρίζεται διάφορα γεγονότα όπως για παράδειγμα παύσεις και μη σχετικές με τον ρυθμό κορυφές στα χαρακτηριστικά έμφασης. Στο προηγούμενο παράδειγμα (Σχ. 6.3) είδαμε πώς ένας παλμός που «λείπει» στη συνάρτηση έμφασης, μπορεί να αναπληρωθεί στην έξοδο του ταλαντωτή. Στο Σχ. 6.5 παρουσιάζεται η διέγερση του Σχ. 6.3 αλλά περισσότερο αλλοιωμένη. Συγκεκριμένα, εκτός από τον παλμό που λείπει, έχουν προστεθεί δύο παλμοί που δεν είναι σχετικοί με τον ρυθμό. Παρά την αλλοίωση της διέγερσης, η έξοδος του ταλαντωτή παραμένει συνεπής με τον αρχικό (μη αλλοιωμένο) παλμό. Επομένως για ένα σύστημα εξαγωγής παλμού που γνωρίζει εκ των προτέρων το σωστό τέμπο, θα ήταν πιο εύκολο να εξαχθούν οι στιγμές του παλμού από την έξοδο του ταλαντωτή παρά από την διέγερση.

Επομένως, βάσει των παραπάνω, ως συνάρτηση έμφασης $a_{\mathbf{x}}[\cdot]$ για το χαρακτηριστικό $\mathbf{x} \in \{\mathbf{x}_{ch}^{i} \cup \mathbf{x}_{e}^{k}\}$ επιλέγεται η έξοδος του ταλαντωτή μηδενικής φάσης που αντιστοιχεί στο τοπικό τέμπο. Για κάθε από τα χρονικά διαστήματα της Εξ. 6.5 επιλέγεται ο αντίστοιχος ταλαντωτής

$$a_{\mathbf{x}}[m] = \hat{r}_{\mathbf{x}}[m, T[m]] \tag{6.6}$$

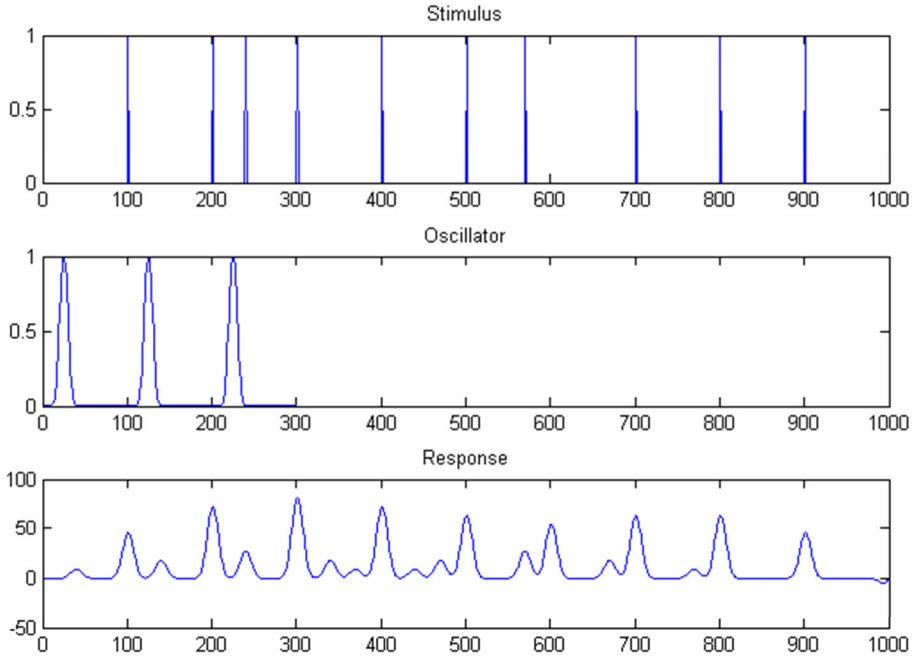
όπου $T[n]$ όπως προαναφέρθηκε είναι το τοπικό τέμπο τη χρονική στιγμή n. Ως καθολική συνάρτηση έμφασης υπολογίζεται το άθροισμα των συναρτήσεων έμφασης για να τα επιμέρους χαρακτηριστικά.

$$\hat{a}[m] = \sum_{\mathbf{x}} a_{\mathbf{x}}[m]\,,\ \mathbf{x} \in \{\mathbf{x}_{ch}^{i} \cup \mathbf{x}_{e}^{k}\} \tag{6.7}$$

Η μηδενική απόκριση φάσης μέσω της τεχνικής Forward-Backward Filtering είναι απαραίτητη, αφού τυχόν διαφορές φάσης των $a_{\mathbf{x}}[m]$ και $x[m]$ θα οδηγούσε σε θέσεις παλμών που αποκλίνουν αντίστοιχα από τις πραγματικές.

6.5 Εξαγωγή Ακολουθίας Παλμών

Το τελικό στάδιο της εξαγωγής παλμού περιλαμβάνει τρία βήματα: α) Τον ορισμό των υποψηφίων θέσεων παλμού, β) τον ορισμό μιας απόστασης μεταξύ τους και γ) την εύρεση ενός βέλτιστου μονοπατιού σε αυτά.



Σχήμα 6.5 Πάνω: Ισόχρονος παλμός που λείπει μία κρουστική (frame 600) και έχουν προστεθεί δύο «μη ρυθμικοί» παλμοί (frames 240 και 575). Μέση: Ταλαντωτής ίδιας συχνότητας με τον παλμό. Κάτω: Η απόκριση του ταλαντωτή. Οι αλλοιώσεις της διέγερσης δεν είναι τόσο εμφανείς στην απόκριση του ταλαντωτή.

Οι υποψήφιες θέσεις του παλμού πρέπει να καλύπτουν δύο ανταγωνιστικές προϋποθέσεις. Να περιέχουν όλα τα πραγματικά beats (ή όσο το δυνατόν περισσότερα) και να είναι όσο το δυνατόν λιγότερες ώστε να μην επιβαρύνουν υπολογιστικά την προτεινόμενη μέθοδο. Η επιλογή όλων των διακριτών χρονικών στιγμών που υπαγορεύει η συχνότητα δειγματοληψίας f_r θα ικανοποιούσε την πρώτη προϋπόθεση, όχι όμως τη δεύτερη. Οι θέσεις του παλμού συναντώνται συνήθως είτε σε κάποιο μουσικό γεγονός, όπως η έναρξη μιας νότας, είτε σε θέσεις που δεν υπάρχουν μουσικά γεγονότα, αλλά υπονοείται η θέση του παλμού από γειτονικά μουσικά γεγονότα. Για παράδειγμα, μια δεσπόζουσα παύση θα μπορούσε να είναι υποψήφια θέση παλμού. Ενώ οι υποψήφιες θέσεις της πρώτης κατηγορίας συνήθως παρουσιάζουν κάποια κορυφή στα αρχικά χαρακτηριστικά έμφασης, αυτό δεν συμβαίνει στη δεύτερη περίπτωση. Όπως είδαμε, αυτό συμβαίνει στην περίπτωση των εξόδων των ταλαντωτών, οι οποίοι κάτω ακόμα και από την απουσία μουσικών γεγονότων ή σε συνθήκες παύσης συνεχίζουν για κάποιο χρονικό διάστημα (που εξαρτάται από το Q_0) να ταλαντώνονται στην ίδια φάση και συχνότητα.

Τα παραπάνω καταδεικνύουν τα μέγιστα ακρότατα των εξόδων των ταλαντωτών ως κατάλληλες υποψήφιες θέσεις παλμού, οι οποίες επιλέγονται ως το σύνολο των κορυφών των εξόδων του ταλαντωτή που αντιστοιχεί στο εξαχθέν τοπικό τέμπο και υπολογιζόμενες για όλες τις διαστάσεις των χαρακτηριστικών. Αν συμβολίσουμε με $\{b_k\}_k$ τις υποψήφιες θέσεις παλμών, τότε μπορούμε να γράψουμε:

$$\{b_k\} = \text{arglocalmax}_m\{a_{\mathbf{x}}[m]\},\ \mathbf{x} \in \{\mathbf{x}_{ch}^i \cup \mathbf{x}_e^k\}. \quad (6.8)$$

Αν και περιμένουμε με αυτή τη διαδικασία να μειωθούν οι υποψήφιες θέσεις παλμών και να περιοριστεί αρκετά ο χώρος αναζήτησης, στην πραγματικότητα τα $\{b_k\}$ καλύπτουν ένα μεγάλο χρονικό εύρος, γεγονός που καταδεικνύει και τον πλούτο της πληροφορίας στις διάφορες διαστάσεις της συνάρτησης έμφασης. Σε ένα απλό πείραμα, για δέκα μουσικά κομμάτια διαφόρων ειδών μουσικής, υπολογίστηκε ο λόγος $\#\{b_k\}/N$ ο οποίος προέκυψε ίσος με 0.3115, όπου N είναι η χρονική διάρκεια (σε samples) των συναρτήσεων έμφασης.

Η επιλογή της συνάρτησης απόστασης αποτελεί ίσως την πιο καθοριστική επιλογή της σχεδίασης του συστήματος αυτόματης εξαγωγής παλμού. Έστω b_i, b_j δύο υποψήφιες θέσεις παλμών. Η απόστασή τους θα πρέπει να ελαττώνεται όσο η χρονική απόσταση αυτών των δύο θέσεων είναι συνεπής με το τέμπο του μουσικού κομματιού. Αντίθετα, όσο η χρονική τους απόσταση από την περίοδο του μουσικού τέμπο αποκλίνει, θα πρέπει αυτή η απόσταση να αυξάνει. Επιλέγουμε μια τέτοια απόσταση να περιέχει έναν παράγοντα ανάλογο με την ποσότητα

$$d_T(b_i, b_j) = 1 - exp\left\{-\frac{1}{\sigma^2}\ln^2\left(\left(b_i - b_j\right)/\tau_T\right)\right\} \quad (6.9)$$

η οποία μηδενίζεται για $b_j = b_i + \tau_T$. Με τ_T συμβολίζεται η περίοδος που αντιστοιχεί στο τέμπο T. Ο λογάριθμος καταδεικνύει ότι ίσες ποσοστιαίες αποκλίσεις του διαστήματος που ορίζουν τα b_i, b_j από την περίοδο τ_T θα συνεπάγονται ίσες αποστάσεις των b_i, b_j $d_T(b_i, b_j)$. Το σ ελέγχει το πόσο η απόκλιση αυτή θα επηρεάζει τον όρο $d_T(b_i, b_j)$. Μεγάλες τιμές του σ επιτρέπουν μεγαλύτερες αποκλίσεις της απόστασης των b_i, b_j από το τ_T ενώ μικρές απαιτούν αυστηρότερη συμμόρφωσή τους στο τέμπο.

Επιπλέον αν για κάποιο b_i και για δύο υποψήφιες θέσεις b_j, b_k η ποσότητα της Εξ. 6.9 είναι η ίδια, το $b \in \{b_j, b_k\}$ που επιτυγχάνει μεγαλύτερο πλάτος στη συνάρτηση έμφασης θα πρέπει να είναι πιο πιθανό να είναι θέση παλμού. Επομένως, η Εξ. 6.9 εμπλουτίζεται και με ένα επιπλέον όρο, ανάλογο της ποσότητας $\hat{a}[b]$:

$$d(b_i, b_j) = \gamma \cdot d_T(b_i, b_j) + (1 - \gamma) \cdot \hat{a}[b_j] \quad (6.10)$$

Η παράμετρος $\gamma \in (0,1)$ ελέγχει την σχετική συνεισφορά των δύο όρων. Όσο η παράμετρος γ πλησιάζει τη μονάδα η απόσταση $d(b_i, b_j)$ δίνει έμφαση στη ρυθμική συνέπεια, ενώ πλησιάζοντας το μηδέν δίνεται έμφαση στα δεσπόζοντα μουσικά γεγονότα. Για τον περιορισμό του χώρου αναζήτησης της ακολουθίας $\{b_k\}_k$ όλες οι αποστάσεις μεταξύ υποψηφίων θέσεων που απέχουν χρονικά περισσότερο από $2\tau_T$ τίθενται ίσες με ∞. Στην περίπτωση που για κάποιο b_i όλα τα επόμενα b_j απέχουν παραπάνω από $2\tau_T$ με αποτέλεσμα όλες οι αποστάσεις να είναι ίσες με ∞ το κατώφλι μέχρι $2\tau_T$ μεγαλώνει μέχρι να εξασφαλιστεί ότι υπάρχει μία τουλάχιστον δυνατή μετάβαση από το b_i.

Έστω $\{b_l\}, l \in L \subseteq \{1 \dots N\}$ μία ακολουθία παλμών. Η βέλτιστη ακολουθία παλμών $\{b_i^*\}, i \in I \subseteq L$ θα πρέπει να ελαχιστοποιεί την αντικειμενική συνάρτηση

$$O(\{b_i^*, i \in I\}) = \sum_{i \in I} d(b_{i-1}^*, b_i^*) \quad (6.11)$$

Η εύρεση της βέλτιστης $\{b_i^*\}$ επιτυγχάνεται με δυναμικό προγραμματισμό. Συμβολίζουμε με $C^*(b_i)$ το ελάχιστο κόστος για να φτάσουμε στην θέση b_i ως εξής

$$C^*(b_i) = \min_{b_k}\{d(b_k, b_i) + C^*(b_k)\}, k < i \quad (6.12)$$

και

$$path(b_i) = \underset{b_k}{\operatorname{argmin}}\{d(b_k, b_i) + C^*(b_k)\}, k < i \quad (6.13)$$

όπου με $path(b_i)$ συμβολίζεται η προηγούμενη θέση παλμού προκειμένου να φτάσουμε στο b_i με βέλτιστο τρόπο. Τα κόστη $C^*(b_i)$, $i = 1 \ldots I$ υπολογίζονται αναδρομικά από την Εξ. 6.12. Για την εξαγωγή της βέλτιστης ακολουθίας επιλέγουμε ένα υποσύνολο $\{b'_m\}$ στο τέλος του κομματιού, $2\tau_T$ πριν το τελευταίο b_i. Με τη χρήση του αλγορίθμου Viterbi η τελευταία θέση της βέλτιστης ακολουθίας παλμών υπολογίζεται ως

$$b_k^* = \underset{b'_m}{\operatorname{argmin}}\{C^*(b'_m)\} \quad (6.14)$$

Η υπόλοιπη βέλτιστη ακολουθία παλμών προκύπτει αναδρομικά μετακινούμενοι προς τα πίσω

$$b_{k-1}^* = path(b_k^*), k = K \ldots 2 \quad (6.15)$$

6.6 Αξιολόγηση της Μεθόδου Εξαγωγής Ακολουθίας Παλμών

Η έλλειψη διαθέσιμων δεδομένων με επισημειώσεις των θέσεων παλμών κάνει την αξιολόγηση της μεθόδου εξαγωγής παλμού δύσκολη. Η προτεινόμενη μέθοδος αξιολογήθηκε στον διεθνή διαγωνισμό MIREX τα έτη 2011 και 2012. Σε αντίθεση με την εξαγωγή τέμπο, υπάρχει πολύ μεγαλύτερη ποικιλία μετρικών αξιολόγησης, οι οποίες λαμβάνουν υπόψη διαφορετικά κριτήρια. Στον διαγωνισμό MIREX αναφέρονται αποτελέσματα για όλες τις μετρικές αυτές, οι βασικότερες των οποίων είναι οι ακόλουθες.

F-Measure

Η συνήθης μετρική F-Measure που δίνεται ως ο γεωμετρικός μέσος της ακρίβειας (Pr) και της ανάκλησης (Ee):

$$F = \frac{2 \cdot Pr \cdot Re}{Pr + Re} \quad (6.16)$$

Το διάστημα ανοχής για να θεωρηθεί ένας εξαγόμενος παλμός σωστός είναι 70 ms. Παρότι η μετρική αυτή είναι διαδεδομένη, έχει το μειονέκτημα ότι δεν λαμβάνει την ρυθμική ή όχι συνέπεια του παλμού.

P-Score [McKinney2007]

Το P-Score ορίζεται ως:

$$P = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{NP}\sum_{m=-W_s}^{W_s}\sum_{n=1}^{N} y[n]a_s[n-m] \quad (6.17)$$

όπου $y[n]$ είναι ο εξαγόμενος παλμός, S είναι το πλήθος των επισημειώσεων για κάθε κομμάτι, s ο δείκτης της κάθε επισημείωσης και $a_s[n] = 1$ αν τη χρονική στιγμή n είναι θέση παλμού για την επισημείωση s, αλλιώς $a_s[n] = 0$. N είναι το πλήθος των παλμών της επισημείωσης s, W_s είναι το παράθυρο ανοχής για το οποίο οι εξαγόμενοι παλμοί θεωρούνται σωστοί και $NP = \max(\sum y[n], \sum a_s[n])$.

Cemgil [Cemgil2000]

Οι μέθοδοι αξιολογούνται με βάση την εξής συνάρτηση αξιολόγησης:

$$\rho(\mathbf{a},\mathbf{y}) = \frac{\sum_i \max_j W(a[i]-y[j])}{(I+J)/2} \times 100 \tag{6.18}$$

όπου $a[i], i = 1 \dots I$ είναι η αληθής ακολουθία παλμών, $y[j]$ είναι η εξαγόμενη ακολουθία παλμών και $W(d) = e^{-\frac{d^2}{2\sigma_e^2}}$ με $\sigma_e^2 = 0.04$.

Μετρικές βασισμένες στην συνέχεια

Έστω a_i η ακολουθία των επισημειωμένων παλμών για ένα κομμάτι και y_j η ακολουθία των εξαγόμενων παλμών και $\Delta_i = a_i - a_{i-1}$, $\Delta_j = y_j - y_{j-1}$ οι αντίστοιχες ακολουθίες διαστημάτων που ορίζονται από αυτά. Η έννοια της «συνέχειας» προκύπτει από την δημιουργία ενός διαστήματος ανοχής $\theta = 0.175$ γύρω από κάθε επισημειωμένο παλμό $a_i : [a_i - \theta\Delta_i, a_i + \theta\Delta_i]$. Ο κοντινότερος παλμός y_j στην επισημείωση a_i θεωρείται σωστός αν βρίσκεται μέσα στο διάστημα ανοχής του a_i, και επιπλέον ο προηγούμενος παλμός y_{j-1} βρίσκεται στο διάστημα ανοχής της προηγούμενης επισημείωσης a_{i-1}. Με ένα επιπλέον περιορισμό σχετικά με την συνέχεια των Δ_i και Δ_j προκύπτουν οι ακόλουθοι τρεις τελικοί περιορισμοί «συνέχειας»:

(α) $a_i - \theta\Delta_i < y_j < a_i + \theta\Delta_i$

(β) $a_{i-1} - \theta\Delta_{i-1} < y_{j-1} < a_{i-1} + \theta\Delta_{i-1}$

(γ) $(1-\theta)\Delta_i < \Delta_j < (1+\theta)\Delta_i$

Συγκρίνοντας κάθε y_j με κάθε a_i βάσει των περιορισμών (α)-(γ) βρίσκουμε τα τμήματα $\Psi_m, m = 1 \dots M$ για τα οποία έχουν βρεθεί συνεχόμενοι παλμοί που ικανοποιούν τις (α)-(γ). Η πρώτη μετρική CML_c (Correct Metrical Level, continuity required), είναι το σχετικό μέγεθος του μεγαλύτερου σε διάρκεια Ψ_m:

$$\mathrm{CML}_c = \frac{\max(\#(\Psi_m))}{I} \times 100\% \tag{6.19}$$

όπου I είναι το πλήθος των a_i και $\#(\Psi_m)$ το μέγεθος του Ψ_m. Η μετρική CML_c περιέχει πληροφορία μόνο για το μεγαλύτερο τμήμα σωστών παλμών. Για παράδειγμα, αν για ένα κομμάτι όλοι οι εξαγόμενοι παλμοί είναι σωστοί εκτός από έναν που βρίσκεται στη μέση κομματιού, τότε η CML_c θα είναι 50%.

Για να ληφθούν υπόψη και οι υπόλοιπες (μικρότερες) ακολουθίες σωστών εξαγόμενων παλμών, προτείνεται η λιγότερη αυστηρή μετρική CML_t (Correct Metrical Level, continuity not required):

$$\mathrm{CML}_t = \frac{\sum_{m=1}^{M} \#(\Psi_m)}{I} \times 100\% \tag{6.20}$$

Προκειμένου να συμπεριληφθούν στις παραπάνω μετρικές και τα λάθη οκτάβας, προτείνονται δύο επιπλέον μετρικές, όπου η ακολουθία a_i επεκτείνεται ώστε να επιτρέπει παλμούς στο μισό ή διπλάσιο τέμπο. Με αυτόν τον τρόπο ορίζονται οι AML_c (Allowed Metrical Level, continuity required) και AML_t (Allowed Metrical Level, continuity not required). Ωστόσο η μέθοδος αυτή μπορεί να τεθεί υπό αμφισβήτηση καθώς δεν είναι το μισό και το διπλάσιο τέμπο πάντα ρυθμικά σωστά.

Μέθοδος	F-Meas	Cemgil	P-score	CMLc	CMLt	AMLc	AMLt	D (bits)	Dg (bits)
KB1	40.74	30.51	50.03	12.81	19.21	26.61	45.07	1.00	0.147
GKC2	**36.60**	**27.90**	**51.67**	**17.73**	**26.76**	**24.43**	**40.02**	**1.02**	**0.125**
GP3	35.16	26.25	46.16	11.81	16.79	24.96	39.50	0.97	0.109
GP2	36.35	27.24	47.61	13.37	20.14	23.63	38.59	0.99	0.118
ZDG2	37.06	28.49	47.74	12.71	17.06	23.60	37.58	0.99	0.134
ZDG1	35.52	27.50	46.98	11.83	16.29	23.65	36.97	0.98	0.124
FK1	39.71	30.50	49.97	14.24	22.34	22.88	36.66	0.99	0.190
KFRO1	32.81	25.51	47.51	16.09	22.68	24.44	35.04	1.01	0.132
FW4	34.11	26.32	46.16	13.47	20.35	19.07	32.68	0.85	0.078
GP4	36.89	27.89	48.79	10.94	18.43	17.43	31.36	0.93	0.128
SB6	35.03	27.44	46.06	13.09	18.04	18.92	31.24	0.83	0.082

Πίνακας 5.1. Τα αποτελέσματα του διαγωνισμού εξαγωγής παλμού MIREX 2012 για τη συλλογή SCM. Τα αποτελέσματα είναι ταξινομημένα κατά την μετρική αξιολόγησης AML-t. Αναλυτικότερα αποτελέσματα υπάρχουν στην ιστοσελίδα του MIREX.[14]

Μέθοδος	F-Meas	Cemgil	P-score	CMLc	CMLt	AMLc	AMLt	D (bits)	Dg (bits)
ZDG2	53.39	40.62	58.24	25.01	33.38	51.76	66.66	1.81	0.313
GKC2	**50.10**	**37.83**	**55.16**	**25.81**	**32.94**	**51.04**	**64.23**	**1.69**	**0.273**
ZDG1	51.61	38.82	57.38	23.72	32.34	49.45	65.10	1.80	0.264
ODGR1	50.50	38.21	55.50	21.56	29.99	49.38	64.15	1.66	0.257
GP3	50.32	37.27	56.56	23.96	33.69	49.27	66.45	1.78	0.252
GP2	50.09	37.00	56.18	23.26	32.30	48.58	64.89	1.78	0.241
KFRO1	51.13	38.97	56.03	25.01	32.02	47.09	58.84	1.66	0.292
ODGR2	50.38	38.17	55.44	22.36	30.39	47.04	62.70	1.61	0.267
ODGR3	49.75	37.72	55.03	21.83	29.74	44.23	59.74	1.54	0.258
FW4	52.13	39.50	57.68	23.68	34.52	42.44	59.14	1.64	0.259
FK1	56.73	42.70	61.16	22.25	35.08	41.48	63.27	1.66	0.313

Πίνακας 5.2. Τα αποτελέσματα του διαγωνισμού εξαγωγής παλμού MIREX 2012 για τη συλλογή MCK. Τα αποτελέσματα είναι ταξινομημένα κατά την μετρική αξιολόγησης AML-t.

[14] http://nema.lis.illinois.edu/nema_out/mirex2012/results/abt/smc/summary.html

	F_Meas	Cemgil Acc	Goto Acc	PScore	CMLc	CMLt	AMLc	AMLt	Inf Gain	AML Cemg
Aubio	49,35	40,09	18,85	54,07	26,43	35,12	37,73	50,57	1,58	49,10
Beatit	52,68	42,77	7,07	47,90	6,98	8,69	43,64	60,95	1,62	61,28
Beatroot	61,73	50,52	32,40	60,92	29,05	35,70	53,51	70,84	1,98	63,73
BeatUJAEN	33,87	25,88	8,47	41,46	10,45	17,17	26,84	41,63	1,18	38,56
Bock	66,64	56,91	21,50	65,78	31,46	43,48	42,20	58,74	1,98	63,41
Davies	62,84	52,69	46,32	65,21	46,82	50,79	69,28	75,88	2,26	65,49
Degara	65,27	55,24	45,80	66,39	46,04	50,17	69,89	77,72	2,27	67,93
Ellis	55,13	42,08	8,84	50,64	10,66	14,02	38,54	60,03	1,76	58,83
Essentia	51,66	38,69	14,95	53,79	13,82	21,60	34,38	57,32	1,43	51,31
Gkiokas	59,62	48,89	36,67	62,05	41,47	47,10	62,73	72,75	2,10	59,76
Hainsworth	51,14	43,16	31,81	54,31	34,28	37,24	54,08	59,62	1,85	56,04
IBT (causal)	55,18	44,80	18,04	56,59	25,27	30,82	47,05	58,07	1,67	55,99
IBT (No causal)	60,47	49,18	31,59	60,33	32,54	36,88	63,97	73,76	1,92	63,16
Klapuri	65,54	55,11	45,80	66,63	47,75	52,71	69,79	77,70	2,32	66,90
Lee	48,79	38,30	0,07	46,51	1,61	7,06	5,87	26,38	1,09	51,11
Scheirer	56,16	45,81	12,22	58,50	21,19	34,52	30,38	48,97	1,69	53,68
Stark	59,49	49,99	41,02	62,75	41,68	47,32	61,64	70,99	2,03	62,24

Πίνακας 5.3: Συγκριτικά αποτελέσματα 17 μεθόδων όπως προέκυψαν από πειράματα που έκανε ο J. Zapata.

Ο Πίνακας 5.1 παρουσιάζει τα αποτελέσματα των έντεκα καλύτερων μεθόδων (από ένα σύνολο 20) του διαγωνισμού MIREX 2012 για το σύνολο δεδομένων SMC που αποτελείται από 289 αποσπάσματα. Όπως αναφέρθηκε και στην Ενότητα 5.5.5, το SMC περιέχει μουσικά κομμάτια που έχουν σύνθετο ρυθμικό περιεχόμενο και που είναι πολύ δύσκολο να εξαχθεί ο παλμός με αυτόματο τρόπο. Στον Πίνακα 5.2 παρουσιάζονται τα αντίστοιχα αποτελέσματα για το σύνολο McKinney που είναι το ίδιο με αυτό της αξιολόγησης των μεθόδων εξαγωγής τέμπο. Καθότι οι αλγόριθμοι που αξιολογούνται είναι πολλοί και δεν αντιστοιχούν πάντα σε δημοσιευμένα άρθρα (γίνεται υποβολή μόνο extended abstract στο διαγωνισμό MIREX), δεν θα γίνει αναλυτική περιγραφή των μεθόδων όπως έγινε στα προβλήματα εξαγωγής μέτρου, χορευτικού ρυθμού και τέμπο. Ο αναγνώστης μπορεί να βρει πληροφορίες και τυχόν λεπτομέρειες για τις μεθόδους στην ιστοσελίδα του διαγωνισμού[1].

Παρατηρούμε ότι και στις δύο συλλογές η προτεινόμενη μέθοδος (GKC2) είναι από τις καλύτερες σε επίδοση για όλες τις μετρικές. Η χρήση πολλών μετρικών κάνει δύσκολη την κατάταξη των μεθόδων, ωστόσο η μέθοδός μας βρίσκεται στις πέντε πρώτες για όλες σχεδόν τις μετρικές. Λόγω του ότι η προτεινόμενη μέθοδος επιτυγχάνει αρκετά ακριβή εκτίμηση του τέμπο καθώς και υψηλά ποσοστά αναγνώρισης του σωστού μετρικού επιπέδου, επιτυγχάνονται πολύ υψηλά ποσοστά βάσει των μετρικών CML_c και CML_t. Η προτεινόμενη μέθοδος είναι η πρώτη σε επίδοση μέθοδο βάσει της CML_c και στις δύο συλλογές, ενώ βάσει της CML_t είναι πρώτη στην συλλογή SCM και 5^{η} στην McKinney.

Τέλος στον Πίνακα 5.3 παρουσιάζεται ένα πείραμα σύγκρισης που διεξήγαγε ο J. Zapata [Zapata2013] από το Universitat Pompeu Fabra, σε ένα ειδικά επιλεγμένο σύνολο μουσικών κομματιών το οποίο θεωρείται «δύσκολο». Σε συμφωνία με τα αποτελέσματα του διαγωνισμού MIREX η προτεινόμενη μέθοδος είναι στις καλύτερες 5-6 για όλες τις μετρικές, ωστόσο υστερεί σημαντικά σε σχέση με τις τρεις πρώτες μεθόδους για κάθε μετρική.

Κεφάλαιο 7: Προσεγγίζοντας έναν Ρυθμικό Μετασχηματισμό

7.1 Εισαγωγή

Σε αυτό το Κεφάλαιο θα παρουσιαστεί μία ρυθμική αναπαράσταση η οποία να μπορεί να χρησιμοποιηθεί σε προβλήματα ρυθμικής ανάλυσης (όπως έγινε με την ΣΠ) και να είναι προσεγγιστικά αντιστρέψιμη, δηλαδή να μπορούμε να ανακατασκευάσουμε ένα μουσικό σήμα από αυτήν. Για να επιτυχθεί αυτό υιοθετούμε ένα τυπικό σχήμα εξαγωγής μιας ΣΠ, προσπαθώντας κάθε βήμα της επεξεργασίας να είναι (προσεγγιστικά) αντιστρέψιμο, δηλαδή να μπορούμε να ανακατασκευάσουμε την είσοδο από την έξοδο. Στη συνέχεια η ΣΠ αποτελεί την είσοδο σε μια συστοιχία RBM των οποίων η έξοδος χρησιμοποιείται από έναν ταξινομητή για να την αντιμετώπιση των προβλημάτων εξαγωγής μέτρου και χορευτικού ρυθμού, όπως έγινε και στο Κεφ. 4.

Οι αντίστροφοι μετασχηματισμοί κατέχουν ένα πολύ κεντρικό ρόλο στην επεξεργασία μουσικών σημάτων. Από την καθημερινή χρήση μουσικής, όπως για παράδειγμα ρυθμίζοντας το equalizer στο ηχοσύστημα στο σπίτις μας, έως και την επαγγελματική παραγωγή μουσικής (ηχογράφηση, μίξη κ.λπ.), ο μετασχηματισμός του ακουστικού σήματος αποτελεί ένα πολύ σημαντικό στοιχείο αυτών των εφαρμογών. Οι πιο πολλές από αυτές βασίζονται στον μετασχηματισμό Fourier βραχέως χρόνου (Short-Time Fourier Transform, STFT), ο οποίος παρέχει έναν πολύ απλό και διαισθητικό τρόπο να επεξεργαστούμε και να τροποποιήσουμε ηχητικά σήματα. Ωστόσο ο κύριος περιορισμός του STFT είναι η γραμμική κατανομή των συχνοτήτων, που είναι μειονέκτημα για την ανάλυση μουσικών σημάτων. Ο CQT (βλ. Παρ. 2.2) προτάθηκε το 1991 από τον Brown [Brown1991] και παρότι υπήρξε εναλλακτικός μετασχηματισμός του STFT που υπερτερούσε, δεν έτυχε της εκτίμησης που αναμενόταν. Αυτό οφείλεται σε ένα βαθμό στο υπολογιστικό του κόστος, αλλά και το γεγονός ότι μέχρι και πρόσφατα δεν είχε βρεθεί ο αντίστροφός του. Μόλις πρόσφατα προτάθηκαν μέθοδοι για σχεδόν τέλεια ([Schorkhuber2010]) και τέλεια ([Velasco2011]) ανακατασκευή.

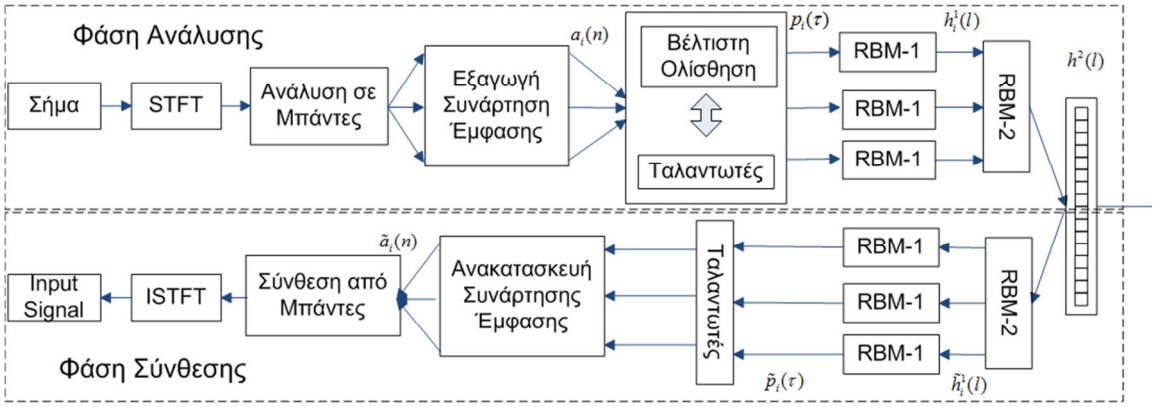
Οι περισσότερες εφαρμογές επεξεργασίας και μετασχηματισμού ακουστικών σημάτων περιλαμβάνουν αλλαγές μόνο στο συχνοτικό περιεχόμενο, δηλαδή επεξεργάζονται κάθε πλαίσιο (frame) του μετασχηματισμού χωριστά. Σχηματικά μπορούμε να πούμε ότι η επεξεργασία σε μια αναπαράσταση χρόνου-συχνότητας γίνεται μόνο ως προς τον άξονα του χρόνου. Αντιθέτως, μετασχηματισμοί στον άξονα του χρόνου δεν έχουν μελετηθεί ιδιαίτερα. Τέτοιου είδους μετασχηματισμοί θα είχαν επίδραση στη χρονική οργάνωση ενός σήματος, που στην περίπτωση της μουσικής θα ήταν πολύ συνδεδεμένοι με τον ρυθμό. Σε αυτό το κεφάλαιο προτείνεται μία ρυθμική αναπαράσταση ενός μουσικού σήματος η οποία μπορεί να χρησιμοποιηθεί σε προβλήματα ανάλυσης ρυθμού και που είναι αντιστρέψιμη, υπό την έννοια ότι ένα ρυθμικά σχετικό ακουστικό σήμα μπορεί να ανακατασκευστεί από αυτή. Πέρα από το αλγοριθμικό ενδιαφέρον για τη δημιουργία ενός τέτοιου μετασχηματισμού, θα βοηθούσε και στην καλύτερη κατανόηση ρυθμικά σχετιζόμενων χαρακτηριστικών των μουσικών σημάτων.

Όπως έχει προαναφερθεί, το προτεινόμενο σύστημα ανάλυσης, όπως και οι περισσότερες μέθοδοι στη βιβλιογραφία, αποτελείται από 2 συστατικά. Την εξαγωγή μιας συνάρτησης έμφασης από την αρχική αναπαράσταση χρόνου-συχνότητας και την ανάλυση περιοδικότητας. Στην παρούσα διατριβή προτείνεται και ένα επιπλέον βήμα επεξεργασίας της ΣΠ. Για να δημιουργηθεί ένας

αντιστρέψιμος ρυθμικός μετασχηματισμός, θα πρέπει το καθένα από τα βήματα επεξεργασίας να είναι αντιστρέψιμο. Παρότι η εξαγωγή της συνάρτησης έμφασης είναι μία *μη αντιστρέψιμη* διαδικασία, στα πλαίσια της ρυθμικής ανάλυσης και υπό κάποιες προϋποθέσεις μπορεί να θεωρηθεί ως (προσεγγιστικά) *αντιστρέψιμη*. Με τον όρο *προσεγγιστικά αντιστρέψιμη* και πάντα στο πλαίο της *ρυθμικής ανάλυσης* εννοούμε ότι είναι δυνατό να ανακατασκευάσουμε κάποιο ηχητικό σήμα το οποίο να διατηρεί τα βασικά *ρυθμικά* χαρακτηριστικά του αρχικού σήματος. Ένα τέτοιο παράδειγμα παρουσιάστηκε στο [Scheirer1998], όπου διεξήχθη το εξής πείραμα. Οι περιβάλλουσες των ενεργειών φασματικών μπαντών ενός μουσικού σήματος, χρησιμοποιήθηκαν για να διαμορφώσουν ένα σήμα θορύβου. Στη συνέχεια ακροατές αξιολόγησαν το προκύπτων σήμα ως προς το ρυθμικό του περιεχόμενο. Τα αποτελέσματα έδειξαν ότι το ρυθμικό περιεχόμενο του ανακατασκευασμένου σήματος έμοιαζε πολύ με το αρχικό. Το ελάχιστο πλήθος μπαντών που βρέθηκε ότι χρειάζεται για αυτό ήταν μόλις πέντε. Επομένως από τη περιβάλλουσα των ενεργειών φασματικών μπαντών (η οποία θα μπορούσε να χρησιμοποιηθεί ως συνάρτηση έμφασης σε κάποιο σύστημα ρυθμικής ανάλυσης) ανακατασκευάστηκε ένα ακουστικό σήμα με παρόμοιο ρυθμικό περιεχόμενο.

Δυστυχώς κάτι ανάλογο δεν συμβαίνει και για την ανάλυση περιοδικότητας. Δηλαδή δεν είναι δυνατό να ανακατασκευάσουμε την συνάρτηση έμφασης από την ΣΠ. Παρότι οι περισσότερες μέθοδοι ανάλυσης περιοδικότητας διατηρούν (σε κάποιο βαθμό) την φασματική πληροφορία (το πλάτος των περιοδικοτήτων) του αρχικού σήματος, απορρίπτουν ή αγνοούν πληροφορία που έχει να κάνει με την φάση ή το χρονισμό των περιοδικοτήτων. Αυτό καθιστά αδύνατη την ανακατασκευή οποιουδήποτε ρυθμικά σχετικού σήματος από την ΣΠ και επομένως αποτελεί εμπόδιο στην δημιουργία ενός αντιστρέψιμου ρυθμικού μετασχηματισμού. Μία μέθοδος εξαγωγής παλμού θα μπορούσε (αφαιρετικά) να θεωρηθεί μία αντίστροφη διαδικασία από τον χώρο της περιοδικότητας στο χώρο του χρόνου, αφού από την ΣΠ προκύπτει μια ακολουθία παλμών. Ωστόσο δεν μπορεί να θεωρηθεί ως ένας αντίστροφος μετασχηματισμός. Μια εξαίρεση προς αυτή τη κατεύθυνση αποτελεί η μέθοδος εξαγωγής παλμού [Holzapfel2011b]. Ο Μη-Στάσιμος Μετασχηματισμός Gabor (Non-Stationary Gabor Transform, NSGT) εφαρμόστηκε στη συνάρτηση περιοδικότητας. Στη συνέχεια από την προκύπτουσα μιγαδική συχνοτική αναπαράσταση (κατ'αναλογία ΣΠ), έγινε επαναδειγματοληψία και επιλογή των πιο εξέχουσων κορυφών. Στη συνέχεια εφαρμόστηκε ο αντίστροφος NSGT στην επεξεργασμένη ΣΠ για να υπολογιστεί μια εκτίμηση των θέσεων του παλμού. Παρότι αυτή η μέθοδος είναι στοχευμένη στην εύρεση του παλμού, έχει τα στοιχεία μιας αντιστρέψιμης ρυθμικής αναπαράστασης αφού ακολουθεί το σχήμα *συνάρτηση έμφασης → συνάρτηση περιοδικότητας → επεξεργασία → αντίστροφη ΣΠ → παλμός*.

Σε αυτό το κεφάλαιο θα παρουσιαστεί μια (προσεγγιστικά) αντιστρέψιμη ρυθμική αναπαράσταση, που περιέχει δύο σημαντικά χαρακτηριστικά. Το πρώτο είναι μία μέθοδος ανάλυσης περιοδικότητας που επιτρέπει μερική ανακατασκευή των συναρτήσεων έμφασης από την ΣΠ. Το δεύτερο σημαντικό στοιχείο είναι η επεξεργασία της ΣΠ με μια συστοιχία RBMs, με τρόπο παρόμοιο με αυτόν που παρουσιάστηκε στο Κεφ. 3. Η συστοιχία των RBM επιτρέπει αφενός (α) την καλύτερη αναπαράσταση της ΣΠ και (β) την ανακατασκευή της ΣΠ από την κρυφή κατάσταση των RBM. Επιπλέον, όπως θα παρουσιαστεί και στο τέλος αυτού του κεφαλαίου, τα RBM επιτρέπουν την *δειγματοληψία* τυχαίων παραδειγμάτων.



Σχήμα 7.1: Η αρχιτεκτονική της προτεινόμενης μεθόδου αντιστρέψιμης ρυθμικής ανάλυσης.

Στην επόμενη Ενότητα θα παρουσιαστούν οι λεπτομέρειες της προτεινόμενης μεθόδου. Στην Ενότητα 7.3 θα αναλυθούν τα πειραματικά αποτελέσματα τόσο στο επίπεδο της αντιστρεψιμότητας (δηλ. ποιότητας ανακατασκευής) της μεθόδου όσο και την ικανότητά της να αναπαριστά επαρκώς και με ακρίβεια το ρυθμικό περιεχόμενο (δηλ. ικανότητα ταξονόμησης). Τέλος, θα παρουσιαστεί μια μέθοδος δειγματοληψίας τυχαίων (συνθετικών) παραδειγμάτων μιας ρυθμικής κατηγορίας.

7.2 Προτεινόμενη Μέθοδος

7.2.1 Προεπισκόπηση Μεθόδου

Στο Σχ. 7.1 παρουσιάζεται η σύνοψη της προτεινόμενης μεθόδου, η οποία αποτελείται από δύο κύριες φάσεις, την φάση ανάλυσης και την φάση σύνθεσης. Στην φάση ανάλυσης το αρχικό σήμα αναλύεται σε I συχνοτικές ζώνες και οι αντίστοιχες συναρτήσεις έμφασης $a_i[m]$ εξάγονται για κάθε ζώνη i. Στη συνέχεια εφαρμόζεται η ανάλυση περιοδικότητας για κάθε $a_i[m]$ προκειμένου να εξαχθούν οι αντίστοιχες ΣΠ $v_i[T]$. Η ανάλυση περιοδικότητας σχεδιάστηκε έτσι ώστε να διατηρεί τόσο τη φασματική όσο και την πληροφορία φάσης της ΣΠ, έτσι ώστε να είναι δυνατό να ανακατασκευαστούν τα $a_i[m]$ από τα $v_i[T]$. Στη συνέχεια τα $v_i[T]$ από όλα τα δεδομένα εκμάθησης χρησιμοποιούνται για να εκπαιδευτεί ένα RBM, το οποίο συμβολίζουμε με RBM-1. Το RBM-1 μαθαίνει την κατανομή όλων των επιμέρους ΣΠ, ανεξάρτητα από ποια διάσταση i της συνάρτησης έμφασης προέρχονται. Στη συνέχεια, οι έξοδοι του RBM-1 που συμβολίζονται με $h_i^1[l]$ ενώνονται σε ένα διάνυσμα h^1 ως:

$$\mathbf{h}^1 = [\mathbf{h}_1^1| \dots |\mathbf{h}_i^1| \dots |\mathbf{h}_I^1]. \tag{7.1}$$

Τα διανύσματα $\mathbf{h}^1$ χρησιμοποιούνται για να εκπαιδεύσουν το RBM-2. Το κίνητρο για αυτή τη διάταξη των RBM είναι ότι το RBM-1 θα μάθει την κατανομή όλων των επιμέρους ΣΠ ανεξάρτητα από ποιά ζώνη i προέρχονται, ενώ το RBM-2 θα μάθει την κατανομή των συνδυασμών των επιμέρους ΣΠ στις I μπάντες των συναρτήσεων έμφασης στο σύνολο εκμάθησης. Στη συνέχεια η έξοδος του RBM-2 που συμβολίζεται με $h^2[l]$ χρησιμοποιείται ως είσοδος σε κάποιο ταξινομητή όπως για παράδειγμα σε ένα SVM, που εκπαιδεύεται σε συγκεκριμένα προβλήματα κατηγοριοποίησης. Στη φάση ανακατασκευής, το συνολικό δίκτυο είναι ικανό να ανακατασκευάσει τις συναρτήσεις έμφασης. Ξεκινώντας από την

έξοδο του RBM-2 $h^2[l]$, υπολογίζονται διαδοχικά τα $\tilde{h}_i^1[l]$, $\tilde{v}_i[\tau]$ και $\tilde{a}_i[m]$ όπως φαίνεται στο κάτω μέρος του Σχ. 7.1. Τέλος, από τα ανακατασκευασμένα $\tilde{a}_i[m]$, είναι δυνατό να εξαχθεί μια κυματομορφή που διατηρεί τα πιο σημαντικά ρυθμικά χαρακτηριστικά του αρχικού σήματος.

7.2.2 Εξαγωγή και Ανακατασκευή Συνάρτησης Έμφασης

Αρχικά το σήμα εισόδου επαναδειγματοληπτείται στα 22.05 kHz και υπολογίζεται ο STFT με ένα ολισθαίνων παράθυρο με μήκος 1024 δείγματα και επικάλυψη 512 δείγματα μεταξύ συνεχόμενων παραθύρων. Από το φασματογράφημα X, εξάγεται η ενέργεια σε I μπάντες, τις οποίες συμβολίζουμε με $e_i[m]$, $i = 1..I$. Τα $e_i[m]$ υπολογίζονται με ισοκατανεμημένα τριγωνικά φίλτρα στην κλίμακα mel. Μπορούμε να γράψουμε:

$$\mathbf{E} = \mathbf{X} \cdot \mathbf{M} \tag{7.2}$$

όπου $\mathbf{E} = [\mathbf{e}_1|\mathbf{e}_2|..\mathbf{e}_I]$ και $\mathbf{M}$ είναι ο πίνακας των φίλτρων. Στη συνέχεια θεωρούμε τον λογάριθμο των $e_i[m]$ και παραγωγίζουμε ως προς τον χρόνο για να εξαχθούν οι συναρτήσεις έμφασης $a_i[m]$, $i = 1..I$. Η διαδικασία αυτή είναι πολύ παρόμοια με την μέθοδο εξαγωγής των συναρτήσεων έμφασης που παρουσιάστηκε στο Κεφ. 2. Η χρήση λογαρίθμων ακολουθούμενων από παραγώγιση αναδεικνύουν τις μεταβολές των $e_i[m]$ σε σχέση με το επίπεδο έντασής τους [Klapuri1999]. Στην συνέχεια κάθε $a_i[m]$ τεμαχίζεται από ένα ολισθαίνων τετραγωνικό παράθυρο μήκους N και ολίσθησης $N/2$. Επιλέγοντας N =512 προκύπτουν παράθυρα μήκους 12sec περίπου. Τέλος, τα τμήματα $a_i^s[m]$ (segments) που προκύπτουν κανονικοποιούνται με τη μέση τιμή και τη τυπική απόκλισή τους. Για να ανακατασκευάσουμε ένα ηχητικό σήμα από τα $\tilde{a}_i^s[m]$ η αντίστροφη διαδικασία εφαρμόζεται στα $\tilde{a}_i^s[n]$, η οποία μπορεί να συνοψιστεί σε

$$\tilde{e}_i^s[m] = exp\left\{\sum_{n=1}^{m}(a_i^s[n] \cdot \sigma_{a_i^s} + \mu_{a_i^s})\right\} \tag{7.3}$$

όπου τα $\mu_{a_i^s}$, $\sigma_{a_i^s}$ συμβολίζουν τη μέση τιμή και την τυπική απόκλιση του $\tilde{\mathbf{a}}_i^s$. Στη συνέχεια το ανακατασκευασμένο φασματογράφημα $\tilde{\mathbf{X}}$ προκύπτει από το $\tilde{\mathbf{E}} = [\tilde{\mathbf{e}}_1|\tilde{\mathbf{e}}_2|..\tilde{\mathbf{e}}_I]$ ως

$$\tilde{\mathbf{X}} = \tilde{\mathbf{E}}\mathbf{M}^{\mathbf{T}} \circ \left[e^{j\angle\mathbf{X}}\right] \tag{7.4}$$

όπου με $\circ$ συμβολίζεται το γινόμενο κατά σημείο (γινόμενο Hadamard) και με $\angle\mathbf{X}$ συμβολίζονται οι φάσεις του αρχικού φασματογραφήματος $\mathbf{X}$. Πρέπει να τονιστεί ότι η Εξ. 7.4 δίνει προσεγγιστική ανακατασκευή του $\mathbf{X}$, αφού $I < M/2$, όπου M είναι το μέγεθος του FFT. Επιπλέον, αφού $I \ll M/2$, το μεγαλύτερο μέρος του «αρμονικού» περιεχομένου έχει χαθεί.

7.2.3 Ανάλυση Περιοδικότητας

Ο σκοπός αυτής της παραγράφου είναι η περιγραφή της ανάλυσης περιοδικότητας, η οποία πρέπει να έχει μια πολύ σημαντική ιδιότητα: να είναι (προσεγγιστικά) αντιστρέψιμη. Για να επιτευχθεί αυτό, θα ξεκινήσουμε από την μέθοδο ανάλυσης περιοδικότητας που περιγράφτηκε στην Ενότητα 2.5, και θα περιγράψουμε τις απαραίτητες τροποποιήσεις για να γίνει αυτή αντιστρέψιμη.

Όπως είδαμε στην Ενότητα 2.5, μία τυπική προσέγγιση ανάλυσης περιοδικότητας η οποία αποτελεί και τη μέθοδο της παρούσας διατριβής είναι η

συνέλιξη κάθε συνάρτησης έμφασης $a_i[m]$ με μία τράπεξα ταλαντωτών $\mathbf{o}_T$, όπου T είναι το τέμπο του κάθε ταλαντωτή. Η μέγιστη τιμή της συνέλιξης σε κάποιο συγκεκριμένο χρονικό παράθυρο δίνει μια εκτίμηση της ΣΠ για το τέμπο Τ για αυτό το παράθυρο, δηλαδή $v_a(T) = \max(\mathbf{o}_T * \mathbf{a})$, όπου $v_a(T)$ είναι η ΣΠ για το $a[nm]$. Αν τα $\mathbf{o}_T$ και $\mathbf{a}$ είχαν το ίδιο μήκος, ένας εναλλακτικός υπολογισμός μιας ΣΠ θα ήταν

$$v_a(T) = \max_k\{\mathbf{a}^T \mathbf{o}_T^k\} \tag{7.5}$$

όπου το $\mathbf{o}_T^k$ συμβολίζει κυκλική ολίσθηση του $\mathbf{o}_T$ κατά k δείγματα (samples). Η $v_a(\tau)$ αντιστοιχεί στο «καλύτερο ταίριασμα» μεταξύ των $\mathbf{o}_T$ και $\mathbf{a}$ και επομένως μπορεί να θεωρηθεί ως ΣΠ.

Το ζητούμενο είναι να εξαχθεί μια αντιστρέψιμη ΣΠ, δηλαδή να μπορεί να ανακατασκευαστεί το $\mathbf{a}$ *εξ'ολοκλήρου* από το $v_a(\tau)$. Αντίστοιχα με το Κεφ. 2 (Εξ. 2.18) χρησιμοποιούνται οι ταλαντωτές που προτάθηκαν στο [Large1994], με τη διαφορά ότι οι ταλαντωτές παραγωγίζονται

$$o_T[m] = d_L[m] * \left[1 + \tanh\left(\gamma(\cos(2\pi f_T m) - 1)\right)\right] \tag{7.6}$$

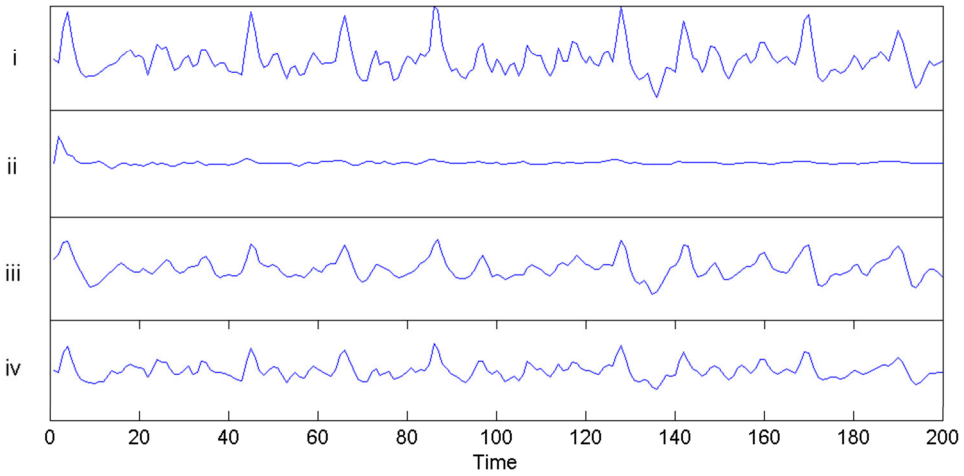
όπου με d_L συμβολίζεται (Εξ. 2.13) ένα μη-αιτιατό φίλτρο παραγώγισης τάξης L, f_T είναι η συχνότητα που αντιστοιχεί στο τέμπο T, και γ το κέρδος εξόδου.

Έστω $\mathbf{O^k} = [\mathbf{o}_{T_1}^{k_1}|\mathbf{o}_T^{k_2}|..|\mathbf{o}_T^{k_M}]$ όπου $\mathbf{k} = [k_1, k_2, .., k_M]$ είναι ένα διάνυσμα που περιέχει τις διαφορετικές ολισθήσεις των ταλαντωτών και $\mathbf{O}^k = \left[\mathbf{o}_{T_1}^k|\mathbf{o}_{T_2}^k|..|\mathbf{o}_{T_M}^k\right]$ όπου k ένας ακέραιος, δηλαδή έχουμε την ίδια ολίσθηση για όλους τους ταλαντωτές. Πως θα μπορούσαμε να ανακατασκευάζαμε το $\mathbf{a}$ από το $\mathbf{v}_a$; Μια απλοϊκή προσέγγιση θα ήταν

$$\tilde{\mathbf{a}} = \mathbf{O}^0 \mathbf{v}_a^T \tag{7.7}$$

δηλαδή να θεωρήσουμε μηδενική ολίσθηση για όλους τους ταλαντωτές. Ωστόσο μια τέτοια προσέγγιση θα οδηγούσε σε πολύ χαμηλής ποιότητας ανακατασκευή, αφού χρειάζονται οι ολισθήσεις των ταλαντωτών k_T που μεγιστοποιούν το $\mathbf{a^T o}_T^k$ ($k_\tau = \text{argmax}_k\left(\mathbf{a}^\mathrm{T}\mathbf{o}_\tau^k\right)$). Για να αναδειχθεί αυτό, στο Σχ. 7.2 (i) και (ii) παρουσιάζονται η αρχική $\mathbf{a}$ και ανακατασκευασμένη συνάρτηση έμφασης $\tilde{\mathbf{a}}$ και είναι φανερό ότι το ρυθμικό περιεχόμενο του $\mathbf{a}$ χάθηκε κατά την ανακατασκευή. Αντίθετα, στο Σχ. 7.2 (iii) φαίνεται η ανακατασκευή του $\mathbf{a}$ όταν στην Εξ. 7.7 χρησιμοποιήθηκαν οι ολισθήσεις των ταλαντωτών k_τ , δηλαδή $\tilde{\mathbf{a}} = \mathbf{O}^{\boldsymbol{k_T}}\mathbf{v}_a^T$.

Αν πάλι υπολογίσουμε την ΣΠ ως $\bar{v}_a(T) = \mathbf{a^T o}_\tau$ αντί της Εξ. 7.5, δηλαδή θέσουμε k=0 για όλους τους ταλαντωτές και θεωρήσουμε απευθείας το εσωτερικό γινόμενο, το ανακατασκευασμένο σήμα που προκύπτει από την αντίστροφη διαδικασία $\bar{\mathbf{a}} = \mathbf{O}\bar{\mathbf{v}}_a^\mathrm{T}$ είναι παρόμοιο με το αρχικό και η πληροφορίας φάσης και χρονισμού έχει διατηρηθεί. Η ανακατασκευασμένη συνάρτηση έμφασης σε αυτή τη περίπτωση παρουσιάζεται στο Σχ. 7.2 (iv). Το ίδιο αποτέλεσμα θα ίσχυε για οποιοδήποτε σταθερό k. Δηλαδή, για οποιαδήποτε σταθερή ολίσθηση k, η τράπεζα ταλαντωτών $\mathbf{O}^k$ παρέχει μια συνεπή ανακατασκευή των $\mathbf{a}$. Επομένως μπορούμε να συμπεράνουμε ότι η τράπεζα ταλαντωτών $\mathbf{O}^k$ είναι μία κατά προσέγγιση βάση των χαρακτηριστικών έμφασης. Πρέπει να σημειωθεί ότι ενώ το $\bar{v}_a(T)$ χρησιμοποιείται για την ανακατασκευή των $\mathbf{a}$, η πραγματική ΣΠ, δηλαδή η ποσότητα που αναπαριστά την ισχύ των ρυθμικών περιοδικοτήτων είναι η απόλυτη τιμή της $|\bar{v}_a(T)|$.



Σχήμα 7.2: Ανακατασκευή των συναρτήσεων έμφασης από την συνάρτηση περιοδικότητας.

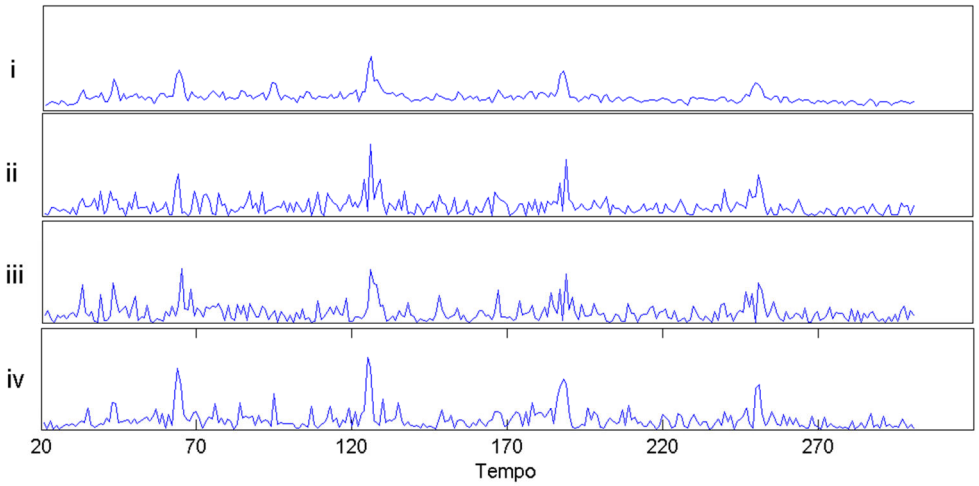
Επομένως, συμπεραίνουμε ότι θεωρώντας μόνο το εσωτερικό γινόμενο και αγνοώντας τις ολισθήσεις των ταλαντωτών κατά τον υπολογισμό της ΣΠ, η ανακατασκευή των συναρτήσεων έμφασης είναι ικανοποιητική. Ωστόσο η συγκεκριμένη ΣΠ $\bar{v}_a(T)$ φαίνεται να είναι μια φτωχή ΣΠ αν υπολογιστεί για αυθαίρετη ολίσθηση k. Στο Σχ. 7.3 (i) παρουσιάζεται τη ΣΠ όπως προκύπτει από την Εξ. 7.5, ενώ στα Σχ. 7.3(ii) και Σχ. 7.3(iii) παρουσιάζεται η $|\bar{v}_a(T)|$ για τυχαίες τιμές της ολίσθησης. Είναι φανερό όρι παρότι οι $v_a(\tau)$ και $|\bar{v}_a(T)|$ παρουσιάζουν κορυφές στις ίδιες τιμές του τέμπο, το πλάτος αυτών των κορυφών είναι αρκετά διαφορετικό. Επιπλέον συγκρίνοντας τα Σχ. 7.3(ii) και (iii) προκύπτει ότι η $|\bar{v}_a(T)|$ είναι ευαίσθητη σε ολισθήσεις των ταλαντωτών ή ισοδύναμα σε ολισθήσεις των συναρτήσεων έμφασης, κάτι που δεν είναι επιθυμητό. Συνοψίζοντας, μπορούμε να ισχυριστούμε ότι η εξαγωγή μιας αξιόπιστης ΣΠ και η ανακατασκευή των συναρτήσεων έμφασης είναι δύο ανταγωνιστικά φαινόμενα. Όταν υπολογίζεται μια αξιόπιστη ΣΠ, η ανακατασκευή είναι αδύνατη, εκτός αν γνωρίζουμε τις ολισθήσεις όλων των ταλαντωτών. Η γνώση όμως όλων των ολισθήσεων δενείναι πρακτική προσέγγιση σε ένα σύστημα αυτόματης ανάλυσης. Από την άλλη, αν πετύχουμε επαρκή ανακατασκευή των συναρτήσεων έμφασης, η ΣΠ έχει διαφορετικά χαρακτηριστικά από την επιθυμητή και επιπλέον είναι ευαίσθητη σε ολισθήσεις της συνάρτησης έμφασης.

Για να ξεπεράσουμε αυτόν τον περιορισμό ώστε να έχουμε μια ικανοποιητική ΣΠ και ταυτόχρονη ανακατασκευή των $\mathbf{a}$, προτείνουμε την ακόλουθη μέθοδο. Το πρώτο βήμα είναι ο υπολογισμός μιας «ιδανικής» ΣΠ, η οποία υπολογίζεται όπως στην Εξ. 7.5 και στη συνέχεια αθροίζεται ως προς τις ζώνες i

$$\hat{v}_a(T) = \sum_i \left| v_{a_i}(T) \right| = \sum_i \left| \max_k \{ \mathbf{a}^T \mathbf{o}_T^k \} \right|. \tag{7.8}$$

Στην συνέχεια, υπολογίζουμε την «πραγματική» ΣΠ για διάφορες τιμές της ολίσθησης k ως

$$v_a^k(T) = \sum_i \mathbf{a}_i^T \mathbf{o}_T^k, \ \mathbf{v}_a^k = \sum_i \mathbf{a}_i^T \mathbf{O}_T^k. \tag{7.9}$$



Σχήμα 7.3: Η συνάρτηση περιοδικότητας λαμβάνοντας τις βέλτιστες ολισθήσεις (i), λαμβάνοντας τυχαίες τιμές ίδιες για όλους τους ταλαντωτές (ii) και (iii), και η ΣΠ της προτεινόμενης μεθόδου (iv).

Πρέπει να σημειωθεί ότι η ολίσθηση k είναι η ίδια για *όλους* τους ταλαντωτές και για *όλες* τις διαστάσεις της συνάρτησης έμφασης i. Τέλος, επιλέγεται η ολίσθηση k_0 που επιτυγχάνει την μεγαλύτερη ομοιότητα της «πραγματικής» $\mathbf{v}_a^k$ και της «ιδανικής» ΣΠ $\hat{\mathbf{v}}_a$. Ως μετρική ομοιότητας υιοθετήθηκε η ομοιότητα συνημιτόνου:

$$k_0 = \text{argmax}_k \left(\frac{|\mathbf{v}_a^k|^T \hat{\mathbf{v}}_a}{\|\mathbf{v}_a^k\| \|\hat{\mathbf{v}}_a\|} \right). \tag{7.10}$$

Η ΣΠ $v_a^{k_0}(T)$ που αντιστοιχεί στην ολίσθηση k_0 μπορεί να ερμηνευτεί ως η καλύτερη προσέγγιση της «ιδανικής» ΣΠ $\hat{v}_a(T)$ από την «πραγματική». Οι επιμέρους ΣΠ για κάθε $\mathbf{a}_i$ προκύπτουν θεωρώντας το εσωτερικό γινόμενο θέτοντας $k = k_0$ για όλους τους ταλαντωτές:

$$\mathbf{v}_{a_i} = \mathbf{a}_i^T \mathbf{O}_T^{k_0}. \tag{7.11}$$

Τέλος, οι συναρτήσεις έμφασης μπορούν να ανακατασκευαστούν από τα $\mathbf{v}_{a_i}$ως:

$$\hat{\mathbf{a}}_i^T = \mathbf{v}_{a_i} \left(\mathbf{O}_T^{k_0} \right)^T. \tag{7.12}$$

Η επιλογή της ίδιας ολίσθησης k στον υπολογισμό της Εξ. 7.9 διασφαλίζει μια καλή ισορροπία μεταξύ της ανακατασκευής των $\mathbf{a}_i$ και της προσέγγισης της «ιδανικής» ΣΠ. Αφού η ολίσθηση k_0 θα είναι ίδια για όλους τους ταλαντωτές, τα $\tilde{\mathbf{a}}_i$ θα είναι κοντά στα $\mathbf{a}_i$, όπως παρουσιάζεται στο Σχ. 7.2 (iv). Ακόμα και στην περίπτωση που μόνο η ΣΠ $\mathbf{v}_{a_i}$ είναι γνωστή, χωρίς καν τη γνώση της τιμής k_0, οι ανακατασκευασμένες συναρτήσεις έμφασης θα διαφέρουν μονάχα κατά μία σταθερή τιμή ολίσθησης ίση με k_0. Επιπλέον, χρησιμοποιώντας μια μοναδική τιμή ολίσθησης κατά τον υπολογισμό της ΣΠ για τις διαφορετικές μπάντες i, δεν χρειάζεται να γνωρίζουμε άλλη τιμή της ολίσθησης για κάθε i προκειμένου να περιγράψουμε τις ΣΠ για όλες τις μπάντες. Ταυτόχρονα, η ΣΠ σε αυτή την περίπτωση Σχ. 7.3 (iv) είναι όσο το δυνατόν πιο όμοια στην «ιδανική» ΣΠ. Με αυτή τη μέθοδο τόσο η συχνοτική (επαρκής ΣΠ) όσο και η χρονική (επαρκής ανακατασκευή) ικανότητα της διατηρούνται. Η ανάλυση περιοδικότητας είναι αναλλοίωτη ως προς ολισθήσεις των συναρτήσεων έμφασης και το

ανακατασκευασμένο σήμα είναι μονάχα ολισθημένο κατά μία σταθερή τιμή k_0 η οποία είναι ήδη γνωστή. Ποσοτική ανάλυση των παραπάνω θα παρουσιαστεί στην Ενότητα 7.3.

7.2.4 Εκμάθηση Χαρακτηριστικών με RBM

Όπως είδαμε στην Ενότητα 3.3.2 εκπαιδεύτηκε ένα RBM λαμβάνοντας τιμές εισόδου από την ΣΠ και τα εξαχθέντα χαρακτηριστικά χρησιμοποιήθηκαν για την αποτελεσματική αντιμετώπιση 3 προβλημάτων ρυθμικής ανάλυσης. Αντίστοιχη τακτική θα ακολουθήσουμε και στην ανάπτυξη του αντιστρέψιμου ρυθμικού μετασχηματισμού. Μια σημαντική διαφορά είναι ότι αντί να θεωρήσουμε την απόλυτη τιμή του βέλτιστου ταιριάσματος όπως κάναμε στην Εξ. 7.8, θεωρούμε τις τιμές που προκύπτουν από την Εξ. 7.11. Επομένως η ΣΠ θα περιέχει και αρνητικές τιμές. Τότε, οι *πολύ μικρές* (αρνητικές) τιμές της ΣΠ θα αντιστοιχούν σε *εξέχουσες* περιοδικότητες. Αυτή η επιλογή είναι απαραίτητη για να είναι δυνατή η ανακατασκευή. Αναμένουμε τότε ότι οι αρνητικές εξέχουσες τιμές της ΣΠ θα κωδικοποιηθούν από τα RBM ως «ισχυρές» περιοδικότητες.

Η επιλογή των RBM αντί για άλλων αντίστοιχων τεχνικών όπως για παράδειγμα του Αυτόματου-Κωδικοποιητή (Auto-Encoder) οφείλεται αφενός στο ότι τα RBM μαθαίνουν καλύτερα χαρακτηριστικά, αφετέρου είναι παραγωγικά (generative) μοντέλα, μια ιδιότητα που χρησιμοποιήθηκε για την δημιουργία τυχαίων τεχνητών παραδειγμάτων όπως θα περιγραφεί αργότερα.

Αν το κόστος ανακατασκευής του RBM δικτύου είναι αρκετά μικρό, τότε και το συνολικό κόστος ανακατασκευής της συνολικής μεθόδου θα είναι χαμηλό. Για την εκπαίδευση του δικτύου, οι ΣΠ εξήχθησαν για όλα τα δεδομένα εκμάθησης, τις οποίες συμβολίζουμε με $v_i^l[T]$, όπου τα l, i είναι οι δείκτες των δεδομένων και των μπαντών των συναρτήσεων έμφασης αντίστοιχα. Στη συνέχεια οι $v_i^l[T]$ συμπτύχθηκαν σε έναν σύνολο εκμάθησης D_1, ανεξάρτητα από το i. Το D_1 χρησιμοποίηθηκε για την εκπαίδευση του RBM-1. Στη συνέχεια, για κάθε παράδειγμα εκμάθησης $p_i^l[T]$, οι αντίστοιχες έξοδοι του RBM-1 $\mathbf{h}_{i,l}^1$ συμπτύχθηκαν σε ένα διάνυσμα όπως περιγράφεται και από την Εξ. 7.1

$$\mathbf{h}_m^1 = [\mathbf{h}_{1,m}^1| \dots |\mathbf{h}_{i,m}^1| \dots |\mathbf{h}_{I,m}^1]. \tag{7.13}$$

Συμβολίζουμε $D_2 = \{\mathbf{h}_l^1\}_l$ το σύνολο εκμάθησης για το RBM-2. Αν N_1 είναι η διάσταση του κρυφού επιπέδου του RBM-1 , τότε η διάσταση των χαρακτηριστικών του D_2 θα είναι $I \cdot N_1$. Επομένως, ενώ το RBM-1 μαθαίνει την κατανομή των επιμέρους ΣΠ ανεξάρτητα από τη μπάντα i, το RBM-2 μαθαίνει μια συνολική κατανομή των συνδυασμών των επιμέρους ΣΠ σε ένα μουσικό κομμάτι. Συμβολίζουμε την έξοδο του RBM-2 ως $\mathbf{h}_m^2$.

7.2.5 Ρυθμική Κατηγοριοποίηση

Για να αναδείξουμε την δυνατότητα των εξαχθέντων χαρακτηριστικών, υιοθετήσαμε έναν ταξινομητή SVM παρόμοια με την μέθοδο κατηγοριοποίησης που παρουσιάστηκε στο Κεφ. 4. Δεδομένου ενός συνόλου εκμάθησης με συναρτήσεις έμφασης $\{\mathbf{a}_i^l\}_{l=1..L}$ και κατηγοριών $\{c^l\}_{l=1..L}$, υπολογίζονται οι αντίστοιχες έξοδοι του συνολικού δικτύου $\{\mathbf{h}_l^2\}_{l=1..L}$ και τα ζεύγη $\{(\mathbf{h}_m^2, c^m)\}_{m=1..M}$ τροφοδοτούν τον SVM ταξινομητή. Για την εκμάθηση, επαλήθευση και κατηγοριοποίηση ακολουθήθηκε η τακτική τεμαχισμού σε 10 υποσύνολα. Το SVM εκπαιδεύτηκε σε επίπεδο τμημάτων (segments). Κατά την κατηγοριοποίηση ενός κομματιού, η απόφαση της κλάσης έγινε βάσει πλειοψηφίας (majority voting).

7.3 Πειραματικά Αποτελέσματα

Η προτεινόμενη μέθοδος αξιολογήθηκε συνολικά στα δύο προβλήματα κατηγοριοποίησης που παρουσιάστηκαν και στο Κεφ. 4, την εξαγωγή χρονικού κλειδιού και την κατηγοριοποίηση σε χορευτικό ρυθμό. Η αξιολόγηση έγινε στις ίδιες συλλογές δεδομένων και πιο συγκεκριμένα στις συλλογές Essen Folk Song και Finish Folk Collections για την εξαγωγή του μέτρου και την συλλογή Ballroom για την ρυθμική κατηγοριοποίηση. Καθότι οι τρεις αυτές συλλογές δεν είναι τόσο αντιπροσωπευτικές (οι δύο περιέχουν MIDI αρχεία), η ικανότητα ανακατασκευής της προτεινόμενης μεθόδου αξιολογήθηκε και σε δύο επιπλέον συλλογές δεδομένων που αποτελούνται από πραγματικά αρχεία ήχου, τις συλλογές Genres και Speedo (βλ. Ενότητα 5.5.1).

Το δίκτυο των RBM εκπαιδεύτηκε στο ίδιο υποσύνολο του MSD αποτελούμενο από 130.000 αποσπάσματα, στο οποίο εκπαιδεύτηκαν τα RBM και στα προηγούμενα πειράματα (βλ. Ενότητα 4.1). Λόγω του ότι όλα τα μουσικά αποσπάσματα κατατμήζονται σε τμήματα των ~12" (βλ. Ενότητα 7.2.2) προκύπτουν περίπου 4.5×10^6 πρότυπα εκπαίδευσης για το RBM-1 και περίπου 9×10^5 πρότυπα εκπαίδευσης για το RBM-2. Τα πρότυπα εκπαίδευσης κανονικοποιήθηκαν ώστε να έχουν μηδενική μέση τιμή και μοναδιαία τυπική απόκλιση σε κάθε διάσταση. Στο RBM-1 υιοθετήθηκαν Γκαουσιανά στοιχεία στο φανερό επίπεδο και στοιχεία ημιανόρθωσης στο κρυφό επίπεδο (NReLUs). Στο RBM-2 και τα δύο επίπεδα αποτελούνται από NReLUs. Το πλήθος των συναρτήσεων έμφασης ανά κομμάτι επιλέχτηκε $I = 5$ και η συνάρτηση περιοδικότητας υπολογίστηκε για $T_{\min} = 20$ και $T_{\max} = 300$ BPM με βήμα $\delta\tau = 1$. Το πλήθος των κρυφών στοιχείων επιλέχθηκαν $N_1 = 300$ και $N_2 = 1500$ για τα RBM-1 και RBM-2 αντίστοιχα, οδηγώντας σε μια αρχιτεκτονική 281x300 για το RBM-1 και (1500x500) για το RBM-2. Και τα δύο RBM εκπαιδεύτηκαν με την μέθοδο Contrastive Divergence με ένα βήμα δειγματοληψίας Gibbs.

Στην επόμενη Ενότητα θα παρουσιαστούν τα αποτελέσματα αξιολόγησης της «αντιστρεψιμότητας» της μεθόδου, δηλαδή το πόσο ικανοποιητικά ανακατασκευάζονται οι συναρτήσεις έμφασης από την έξοδο του δικτύου. Στην Ενότητα 7.3.2 θα παρουσιαστούν τα αποτελέσματα κατηγοριοποίησης, ενώ στο τέλος (Ενότητα 7.3.3) θα παρουσιαστεί μια μέθοδος κατασκευής τυχαίων παραδειγμάτων από ρυθμικές κατηγορίες.

7.3.1 Αξιολόγηση της Αντιστρεψιμότητας

Στη φάση σύνθεσης του Σχ. 7.1 για την ανακατασκευή των $\tilde{a}_i[m]$ από το $\mathbf{h}^2$, τα λάθη ανακατασκευής κάθε βήματος συσσωρεύονται στο ολικό λάθος ανακασκευής. Σε αυτή τη ενότητα θα παρουσιαστεί μία αναλυτική περιγραφή και βαθύτερη ανάλυση των λαθών ανακατασκευής της προτεινόμενης μεθόδου ανάλυσης περιοδικότητας. Επιπλέον θα αναφερθούν αποτελέσματα στην προσέγγιση της «ιδανικής» ΣΠ (Εξ. 7.11) και του λάθος ανακατασκευής του δικτύου RBM.

Έστω $\hat{a}_i[m]$ η ανακατασκευή των συναρτήσεων έμφασης $a_i[m]$ μόνο από την ΣΠ, δηλαδή τα $\hat{a}_i[m]$ υπολογίζονται από τις Εξ. 7.11-7.12 και έστω ότι με $\tilde{a}_i[m]$ συμβολίζεται η ανακατασκευή από όλο το δίκτυο (από την έξοδο $\mathbf{h^2}$ του RBM-2). Τότε η διαφορά μεταξύ των $(\hat{a}_i[m], a_i[m])$ θα αντιστοιχούν στο κόστος ανακατασκευής που οφείλεται αποκλειστικά στην ανάλυση περιοδικότητας (τα RBM αγνοούνται). Η διαφορά μεταξύ των $(\tilde{a}_i[m], \hat{a}_i[m])$ θα αντιστοιχεί στο λάθος ανακατασκευής που εισάγεται έμμεσα από το δίκτυο RBM, αφού τα RBM συνεισφέρουν στο συνολικό λάθος μέσω του λάθους ανακατασκευής της ΣΠ.

Τέλος, η διαφορά μεταξύ των $(\tilde{a}_i[m], a_i[m])$ αντιστοιχεί στο κόστος ανακατασκευής του συνολικού δικτύου.

Σχετικά με την προσέγγιση της «ιδανικής» ΣΠ, η διαφορά της $\hat{v}_a(T)$ (Εξ. 7.8) και της $v_a(T) = \sum_i |v_{a_i}(T)|$ (Εξ. 7.11) αντιστοιχεί στο λάθος προσέγγισης της ΣΠ, που οφείλεται στην ανάλυση περιοδικότητας. Αν συμβολίσουμε με $\tilde{v}_{a_i}(T)$ (Σχ. 7.1) την ΣΠ που προέρχεται από την ανακατασκευή των RBM, τότε η διαφορά μεταξύ των $\tilde{v}_{a_i}(T)$ και $v_{a_i}(T)$ αντιστοιχεί στο λάθος ανακατασκευής της ΣΠ από το δίκτυο των RBM. Τέλος, η διαφορά μεταξύ των $\hat{v}_a(T)$ και $\tilde{v}_a(T) = \sum_i \tilde{v}_{a_i}(T)$ αντιστοιχεί στο λάθος της προσέγγισης της «ιδανικής» ΣΠ από όλο το δίκτυο (ανάλυση περιοδικότητας + RBM).

Για να ποσοτικοποιηθεί η επίδοση της μεθόδου ως προς την ανακατασκευή των συναρτήσεων έμφασης και την προσέγγιση της «ιδανικής» ΣΠ, χρησιμοποιήθηκε ως μετρική ομοιότητας η ομοιότητα συνημιτόνου $R_{\mathbf{x}\tilde{\mathbf{x}}} = \mathbf{x}^T\tilde{\mathbf{x}}/\|\mathbf{x}\|\|\tilde{\mathbf{x}}\|$. Η επιλογή της ομοιότητας συνημιτόνου αντί άλλων συμβατικών μετρικών όπως για παράδειγμα η Ευκλείδια απόσταση $\|\mathbf{x} - \tilde{\mathbf{x}}\|_2$ βασίζεται στο γεγονός ότι η ομοιότητα συνιμητόνου είναι πιο σχετική στο παρών πλαίσιο. Για παράδειγμα, η ομοιότητα συνημιτόνου της $a_i[m]$ με μια ενισχυμένη έκδοσής της $A \cdot a_i[m]$ θα είναι μονάδα, το οποίο δεν ισχύει για την L^2 νόρμα. Το ίδιο ισχύει και την σύγκριση των ΣΠ.

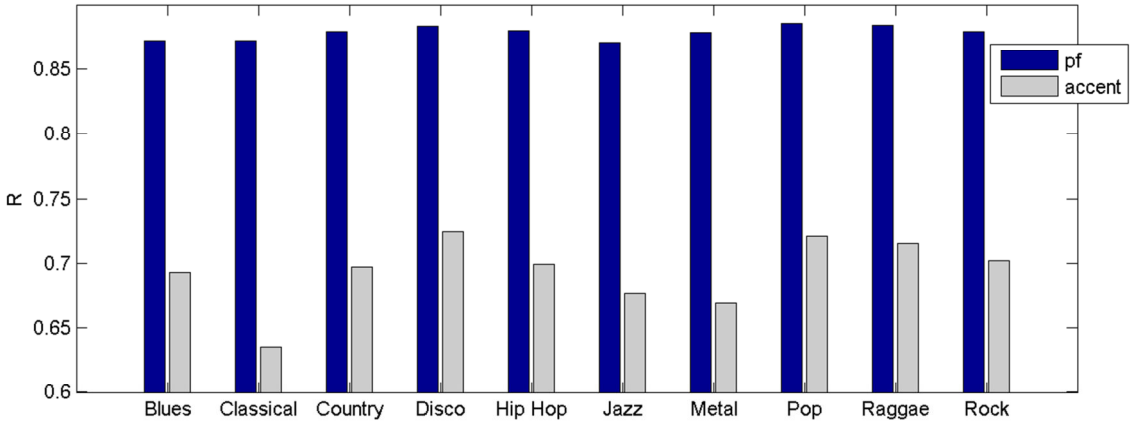
Ο Πίνακας 7.1 συνοψίζει τα αποτελέσματα (α) ανακατασκευής των συναρτήσεων έμφασης και (β) προσέγγισης της ΣΠ για την ανάλυση περιοδικότητας, για το RBM δίκτυο και για ολόκληρη την αρχιτεκτονική, για τις πέντε συλλογές που προαναφέρθηκαν. Οι τιμές αντιστοιχούν στον μέσο όρο για κάθε συλλογή. Σχετικά με την προσέγγιση της «ιδανικής» ΣΠ ($R_{\hat{v}v}$), είναι αξιοσημείωτο ότι με την χρήση μιας μόνο τιμής για την βέλτιστη ολίσθηση των ταλαντωτών (Εξ. 7.8 και 7.11) επιτυγχάνεται προσέγγιση της ιδανικής ΣΠ με ομοιότητα μεγαλύτερη από 0.9 για όλες τις συλλογές. Η μέση ομοιότητα της ανακατασκευασμένης από τα RBM ΣΠ $\tilde{v}_{a_i}(T)$ είναι περίπου 0.92 στο σύνολο των συλλογών, ενώ η ομοιότητα μεταξύ των $\tilde{v}_a(T) = \sum_i \tilde{v}_{a_i}(T)$ και $\hat{v}_a(T)$ (η προσέγγιση της «ιδανικής» ΣΠ από όλο το δίκτυο) είναι 0.87. Με άλλα λόγια, η προτεινόμενη αρχιτεκτονική μαθαίνει μια κρυμμένη (latent) αναπαράσταση $\mathbf{h}^2$ που είναι ικανή να προσεγγίζει την «ιδανική» ΣΠ με ομοιότητα 0.87. Επίσης είναι αξιοσημείωτο ότι η ομοιότητα προσέγγισης είναι περίπου ίδια για όλες τις συλλογές, ακόμα και για αυτές που αποτελούνται από αρχεία MIDI.

Για την ανακατασκευή των συναρτήσεων έμφασης, η πρώτη στήλη του Πίνακα 7.1 αναδεικνύει ένα μεγαλύτερο λάθος λόγω της ανάλυσης περιοδικότητας (Εξ. 7.12), όπου επιτυγχάνεται ομοιότητα της τάξης του 70-80%. Αντίθετα το λάθος που εισάγεται από το δίκτυο των RBM είναι αρκετά μικρό, αφού επιτυγχάνεται μια μέση ομοιότητα των $\hat{a}$ και $\tilde{a}$ της τάξης του 95%. Η συνολική ωστόσο ποιότητα ανακατασκευής του συνολικού δικτύου $R_{\tilde{a}a}$ είναι σχετικά χαμηλή με τις υπόλοιπες, αφού είναι περίπου 70% για τις συλλογές Ballroom, Genres και Speedo και 60% περίπου για τις συλλογές (MIDI) Essen και Finnish-Folk.

Για να αποκτήσουμε καλύτερη εικόνα των λαθών ανακατασκευής, το Σχ. 7.4 παρουσιάζει τα $R_{\tilde{a}a}$ και $R_{\tilde{p}\hat{p}}$ για καθένα από τα δέκα είδη της συλλογής Genres. pop. Από την άλλη είναι αξιοσημείωτο ότι η μετρική προσέγγισης $R_{\tilde{v}\hat{v}}$ της «ιδανικής» ΣΠ είναι περίπου η ίδια για όλα τα μουσικά είδη.

	Ανάλυση Περιοδικότητας		Δίκτυο RBM		Συνολικό Δίκτυο	
Συλλογή	$R_{\hat{a}a}$	$R_{\hat{\nu}v}$	$R_{\hat{a}\tilde{a}}$	$R_{\tilde{\nu}v}$	$R_{\tilde{a}a}$	$R_{\hat{\nu}\tilde{\nu}}$
Essen	74.0	89.7	94.7	91.2	63.6	87.0
F-Folk	72.0	90.5	92.5	89.9	59.4	87.8
Genres	78.7	90.4	94.9	92.6	69.3	87.9
Speedo	79.5	90.4	95.1	92.8	70.4	87.9
Ballroom	78.6	89.8	95.2	92.9	69.9	87.6

Πίνακας 7.1: Αναλυτικά αποτελέσματα ανακατασκευής των συναρτήσεων έμφασης και προσέγγισης της «ιδανικής» ΣΠ από την ανάλυση περιοδικότητας, το δίκτυο των RBM και της συνολικής αρχιτεκτονικής. Οι τιμές δίνονται σε % επί της ομοιότητας συνημιτόνου.



Σχήμα 7.4: Τα $R_{\tilde{a}a}$ (accent) και $R_{\tilde{p}\hat{p}}$ (pf) για καθένα από τα δέκα είδη της συλλογής Genres.

Όπως θα περίμενε κανείς, το λάθος ανακατασκευής των συναρτήσεων έμφασης για κάποια είδη μουσικής όπως η κλασσική και η jazz είναι μεγαλύτερη, ενώ είναι μικρότερο για κάποια είδη με πιο σταθερό και εμφανή ρυθμό, όπως η disco και η

7.3.2 Πειραματικά Αποτελέσματα Κατηγοριοποίησης

Και για τα δύο προβλήματα κατηγοριοποίησης χρησιμοποιήθηκε η ίδια ακριβώς πειραματική διάταξη του Κεφ. 4. Σχετικά με την εξαγωγή μέτρου χρησιμοποιήσαμε εννιά κλάσεις (2/4, 3/2, 3/4, 3/8, 4/1, 4/2, 4/4, 6/4, 6/8 για τη συλλογή Essen Song Collection και τις κλάσεις (2/4, 3/2, 3/4, 3/8, 4/4, 5/2, 5/4, 6/4, 6/8) για την Finnish Folk Tunes, ενώ αναφέρουμε αποτελέσματα και για τα προβλήματα των δύο κατηγοριών (διπλό/τριπλό). Τα αποτελέσματα παρουσιάζονται τον Πίνακα 7.2. Η προτεινόμενη μέθοδος πέτυχε ποσοστό κατηγοριοποίησης 80.7% και 75% στις Essen και Finish Folk συλλογές σε επίπεδο κομματιού.

	Essen Songs		Finnish Folk	
Μέθοδος	2 κλάσεις	9 κλάσεις	2 κλάσεις	9 κλάσεις
Αντιστρέψιμη ΣΠ	89.2	80.7	93.8	75
PCA (1000)	94+	85.9+	97.4+	81.1+
wPCA (500)	93.9+	86.8+	97.8+	81.9+
RBM (1000)	93.9+	85.8+	97.6+	81+
[Toiviainen2006]	95.3+	83.2+	96.4+-	68-
[Eck2005]	90-	-	93	-

Πίνακας 7.2. Συγκριτικά πειραματικά αποτελέσματα με τη διεθνή βιβλιογραφία και με την μέθοδο που παρουσιάστηκε στο Κεφ.4. Τα σύμβολα +/- δείχνουν σημαντική στατιστική διαφορά (statistical significance) συγκριτικά με την μέθοδο «Αντιστρέψιμη ΣΠ».

	2/4	3/2	3/4	3/8	4/1	4/2	4/4	6/4	6/8
2/4	**83**	0	3	1	0	0	12	0	1
3/2	0	**42**	9	0	0	31	18	0	0
3/4	6	0	**83**	0	0	0	10	1	0
3/8	36	0	12	**19**	0	0	2	0	31
4/1	0	0	0	0	**86**	14	0	0	0
4/2	1	1	0	0	2	**92**	4	0	0
4/4	5	0	3	0	0	1	**91**	0	0
6/4	0	0	61	0	0	0	13	**26**	0
6/8	5	0	6	2	0	0	1	0	**86**

Πίνακας 7.3. Πίνακας σύγχυσης (%) μεταξύ των κατηγοριών της συλλογής Essen-9. Οι στήλες αντιστοιχούν σε προβλέψεις και οι γραμμές στις πραγματικές κατηγορίες.

Είναι εμφανές ότι το κόστος της αντιστρεψιμότητας της ΣΠ στην επίδοση κατηγοριοποίησης είναι μεγάλο. Με την ίδια μέθοδο κατηγοριοποίησης και στα ίδια ακριβώς δεδομένα η ακρίβεια ταξινόμησης μειώθηκε κατά 5% περίπου. Ωστόσο πρέπει να σημειωθεί ότι η μέθοδος παραμένει ανταγωνιστική με τις μεθόδους αναφοράς. Για καλύτερη επισκόπηση των αποτελεσμάτων ο Πίνακας 7.3 παρουσιάζει τον πίνακα σύγχυσης για τη συλλογή Essen. Παρατηρούμε αντίστοιχα αποτελέσματα με τον Πίνακα 4.6, με τη διαφορά ότι τα ποσοστά είναι μειωμένα. Τα αποτελέσματα στη συλλογή δείχνουν ότι τα περισσότερα λάθη του ταξινομητή γίνονται μεταξύ παρόμοιων κλάσεων, όπως για παράδειγμα στη κατηγορία 3/8, όπου το 31% των παραδειγμάτων ταξινομούνται ως 6/8 και το 12% ως 3/4, ή η κλάση 6/4 όπου κατά 61% ταξινομείται ως 3/4. Ωστόσο παρατηρούμε και συστηματική σύγχυση μεταξύ μη παρόμοιων μέτρων, όπως π.χ. στην κλάση 3/8 όπου το 36% των παραδειγμάτων ταξινομείται ως 2/4.

Σχετικά με την κατηγοριοποίηση χορευτικού ρυθμού χρησιμοποιήθηκε η συλλογή Ballroom. Τα συγκριτικά αποτελέσματα της μεθόδου παρουσιάζονται στον Πίνακα 7.4. Όπως και στην περίπτωση εξαγωγής μέτρου, η αντιστρέψιμη ΣΠ παρουσιάζει μειωμένη διακριτικά ικανότητα σε σχέση με την ΣΠ των Κεφ. 2 και Κεφ. 3, επιτυγχάνοντας ακρίβεια ίση με 82% έναντι της τάξης του 90% της αρχικής ΣΠ. Ωστόσο και σε αυτή τη περίπτωση, η προτεινόμενη μέθοδος αποδίδει εφάμιλλα με τις μεθόδους αναφοράς.

Μέθοδος	Ακρίβεια
Αντιστρέψιμη ΣΠ	81.95%
PCA (1000)	89.4+
wPCA (100)	89.8+
RBM (500)	90.9+
[Gouyon2004α]	79.6 (90.1+)
[Peeters2005]	81 (90.4+)
[Dixon2004]	84 (96+)

Πίνακας 7.4. Συγκριτικά πειραματικά αποτελέσματα της κατηγοριοποίησης στη συλλογή Ballroom με τη διεθνή βιβλιογραφία και με την προηγούμενο μέθοδο που παρουσιάστηκε στο Κεφ.4. Τα σύμβολα +/- δείχνουν σημαντική στατιστική διαφορά (statistical significance) συγκριτικά με την μέθοδο «Αντιστρέψιμη ΣΠ».

	ChaCha	Jive	Quickstep	Rumba	Samba	Tango	VWaltz	Waltz
ChaCha	**88**	0	3	4	1	5	0	0
Jive	3	**60**	0	3	3	2	5	23
Quickstep	0	0	**98**	2	0	0	0	0
Rumba	1	4	1	**78**	0	4	5	7
Samba	1	1	5	8	**79**	1	1	4
Tango	6	0	0	0	0	**90**	0	5
VWaltz	0	2	0	2	0	0	**59**	39
Waltz	0	3	0	2	0	1	5	**90**

Πίνακας 7.5. Πίνακας σύγχυσης (%) μεταξύ των κατηγοριών της συλλογής Ballroom. Οι στήλες αντιστοιχούν σε προβλέψεις και οι γραμμές στις πραγματικές κατηγορίες.

Τέλος, στον Πίνακα 7.5 παρουσιάζεται ο πίνακας σύγχυσης για τη συλλογή Ballroom. Η υψηλότερη ακρίβεια επετεύχθη για την κατηγορία Quickstep, με ακρίβεια 97%, ενώ τα Tango, ChaChaCha και Waltz αναγνωρίστηκαν με επιτυχία στο ~90% των περιπτώσεων. Αντίθετα, τα ποσοστά αναγνώρισης ήταν πολύ χαμηλά για τις κατηγορίες Jive και Viennese Waltz, όπου μεγάλο ποσοστό τους κατηγοριοποιήθηκε ως Waltz.

Πρέπει να σημειωθεί ότι η προτεινόμενη μέθοδος δεν κάνει κάποια παραδοχή για αυτά τα δύο προβλήματα, αφού αποτελεί μια γενική μέθοδος ρυθμικής κατηγοριοποίησης. Παρόλα αυτά –ακόμα και στην αντιστρέψιμη εκδοχή της ΣΠ– επιτυγχάνει αποτελέσματα τα οποία είναι συγκρίσιμα με αυτά των μεθόδων αναφοράς.

7.3.3 Δειγματοληψία από Ρυθμικές Κλάσεις

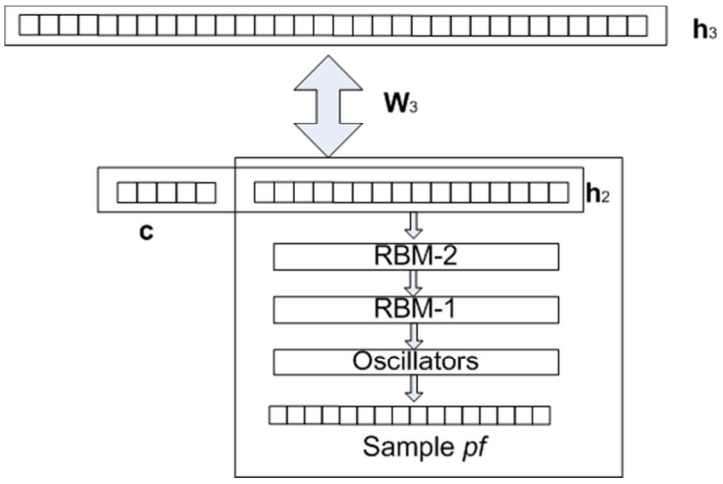
Για να κατανοήσουμε καλύτερα τα χαρακτηριστικά που εξάγονται από το δίκτυο των RBM, αλλά και για να αποκτήσουμε μια επισκόπηση της ικανότητας ανακατασκευής της μεθόδου, υλοποιήθηκε μια μέθοδος *δειγματοληψίας* ηχητικών παραδειγμάτων από τις κατηγορίες της συλλογής Ballroom. Με τον όρο δειγματοληψία δεν εννοούμε την τυχαία επιλογή ενός παραδείγματος από κάθε κλάση, αλλά την δημιουργία *τεχνητών* παραδειγμάτων με δειγματοληψία από την κατανομή του RBM δικτύου. Για να γίνει αυτό, υιοθετήθηκε τακτική παρόμοια με αυτή του Hinton [Hinton2006] για την δειγματοληψία ψηφίων από τις 10

κατηγορίες της συλλογής MNIST. Η αρχιτεκτονική της προσέγγισης παρουσιάζεται στο Σχ. 7.5. Ένα διάνυσμα $\mathbf{c}$ με δυαδικές τιμές που αναπαριστά την κατηγορία (η θέση του 1 είναι η κατηγορία), παρατέθηκε στην έξοδο του δικτύου $\mathbf{h}^2$ ώστε να σχηματίζει το φανερό επίπεδο $\mathbf{v} = [\mathbf{h}^2|\mathbf{c}]$ ενός επιπλέον RBM (RBM-3) με N=3000 κρυφά στοιχεία τύπου Bernoulli (δυαδικά). Τα στοιχεία του διανύσματος $\mathbf{c}$ θεωρήθηκαν ως soft-max στοιχεία. Το RBM-3 εκπαιδεύτηκε με τη μέθοδο Persistent Contrastive Divergence. Επομένως το RBM-3 μαθαίνει την από κοινού αμοιβαία κατανομή του διανύσματος των κατηγοριών $\mathbf{c}$ και της εξόδου του δικτύου ρυθμικής ανάλυσης $\mathbf{h}^2$. Μετά την εκπαίδευση, προκειμένου να παράγουμε *δείγματα* από μία κατηγορία, ακολουθούμε την εξής διαδικασία: Το $\mathbf{h}^2$ αρχικοποείται σε μια τυχαία τιμή $\mathbf{h_0^2}$, ενώ το διάνυσμα $\mathbf{c}$ αρχικοποιείται στην αντίστοιχη κατηγορία από την οποία θέλουμε να πάρουμε δείγμα. Στην συνέχεια εφαρμόζουμε μια αλυσίδα Gibbs με αρχική κατάσταση $\mathbf{v_0} = [\mathbf{h_0^2}|\mathbf{c}]$, κρατώντας όμως το $\mathbf{c}$ σταθερό σε όλη την αλυσίδα. Η διαδικασία αυτή ονομάζεται και «σύσφιξη» (clamping) του $\mathbf{c}$. Στη συνέχεια, από μια κατάσταση $\mathbf{v}_k = [\mathbf{h}_k^2|\mathbf{c}]$ της αλυσίδας μπορούμε να ανακατασκευάσουμε το ηχητικό σήμα του δείγματος $\mathbf{h}_k^2$ εφαρμόζοντας τα βήματα ανακατασκευής του Σχ. 7.1.

Εφαρμόσαμε τη παραπάνω μέθοδο για να δημιουργήσουμε δείγματα από τις οκτώ κατηγορίες της συλλογής Ballroom. Προκειμένου να έχουμε μια μέτρηση της «ποιότητας» των δειγμάτων, ακολουθήσαμε στην εξής διαδικασία. Για κάθε κατηγορία, μετά από 2000 αρχικές επαναλήψεις της δειγματοληψίας Gibbs, κρατούσαμε ένα δείγμα κάθε 250 επαναλήψεις. Προκειμένου τα δείγματα να είναι αντιπροσωπευτικά της κατηγορίας δειγματοληψίας, αγνοήθηκαν αυτά των οποίων η πιθανότητα της αντίστοιχης κλάσης του $\mathbf{c}_k$ πριν την σύσφιξη ήταν μικρότερη από 0.7. Η παραπάνω διαδικασία εκτελέστηκε μέχρι να παραχθούν 50 δείγματα από κάθε κλάση.

Ένας τρόπος να αξιολογηθεί η μέθοδος δειγματοληψίας θα ήταν η ακουστική παρατήρηση και αξιολόγηση των ανακατασκευασμένων ακουστικών δειγμάτων. Ωστόσο μια τέτοια προσέγγιση είναι αφενός υποκειμενική, αφετέρου δεν ποσοτικοποιείται. Για να αξιολογηθεί η ποιότητα των παραχθέντων δειγμάτων, ακολουθήθηκε η εξής διαδικασία. Για κάθε δείγμα, υπολογίστηκαν τα 10 πιο κοντινά παραδείγματα του συνόλου εκμάθησης. Η ομοιότητα υπολογίστηκε ως το συνημίτονο της ΣΠ του παραδείγματος εκπαίδευσης και της ΣΠ που ανακατασκευάστηκε από το δείγμα $\mathbf{h}_k^2$. Στη συνέχεια υπολογίστηκε ο πίνακας σύγχυσης (Πιν. 7.6) για τους κοντινότερους γείτονες των δειγμάτων. Για κάθε κατηγορία (γραμμή του Πίν. 7.6), υπολογίστηκε η κατανομή των κατηγοριών των 10 κοντινότερων γειτόνων (στήλες) των δειγμάτων αυτής της κατηγορίας. Η τελευταία γραμμή αντιστοιχεί στη μέση τιμή του διανύσματος $\mathbf{c}$ ως προς όλα τα βήματα δειγματοληψίας Gibbs.

Παρατηρούμε ότι για μερικές κατηγορίες τα δείγματα είναι πολύ κοντινά στα παραδείγματα εκπαίδευσης της ίδιας κατηγορίας, όπως για παράδειγμα στις κλάσεις ChaChaCha και Samba. Ενδιαφέρον είναι επίσης και το γεγονός ότι συγχέονται πολύ οι κλάσεις Walz και Viennese Waltz. Αυτό είναι κατά το ήμισυ σε συμφωνία με τον Πίνακα 7.5, όπου πολλά παραδείγματα Viennese Waltz ταξινομούνται ως Waltz, δεν ισχύει όμως και το αντίστροφο. Η Rumba συγχέεται με τη Samba και όμοια με τον Πίνακα 7.5, πολλά δείγματα από τις κατηγορίες Jive και Viennese Waltz είναι παρόμοια με παραδείγματα που ανήκουν στη κλάση Waltz.



Σχήμα 7.5: Σύνοψη της μεθόδου δειγματοληψίας τυχαίων παραδειγμάτων.

	ChaChaCha	Jive	Quickstep	Rumba	Samba	Tango	VWaltz	Waltz
ChaChaCha	**100**	0	0	0	0	0	0	0
Jive	1	**40**	0	3	0	0	12	44
Quickstep	2	2	**40**	9	45	1	0	1
Rumba	1	3	5	**50**	38	0	1	3
Samba	0	0	1	13	**86**	0	0	0
Tango	33	6	0	1	0	**56**	2	2
VWaltz	1	7	0	3	4	1	**56**	28
Waltz	1	32	1	3	1	2	14	**44**
c	0.983	0.682	0.772	0.559	0.911	0.787	0.584	0.644

Πίνακας 7.6. Πίνακας σύγχυσης (%) μεταξύ των τυχαίων δειγμάτων και των κατηγοριών της συλλογής Ballroom. Οι στήλες αντιστοιχούν στα τυχαία δείγματα κάθε κλάσης και οι γραμμές στις κατηγορίες των κοντινότερων γειτόνων.

Επίσης παρατηρούμε ότι πολλά δείγματα από τις κλάσεις Quickstep και Tango συγχέονται με τις κλάσεις Samba και ChaCha, κάτι που δεν είναι εμφανές στον αντίστοιχο Πίνακα 7.5. Ωστόσο πρέπει να σημειωθεί ότι τα αποτελέσματα του Πίνακα 7.6 αντιστοιχούν σε ένα τυχαίο πείραμα. Η επισκόπηση πολλών εκτελέσεων αυτού του πειράματος έδειξε ότι υπάρχει μεγάλη ευαισθησία των αποτελεσμάτων σε σχέση με το πλήθος των εποχών εκπαίδευσης του RBM και του πλήθους βημάτων κατά την Gibbs δειγματοληψία. Αυτό όμως είναι αναμενόμενο λόγω της στοχαστικής φύσης της ίδιας της μεθόδου. Συνοψίζοντας πάντως μπορούμε να ισχυριστούμε ότι σε γενικές γραμμές τα αποτελέσματα του Πίνακα 7.6 είναι σε συμφωνία με αυτά του Πίνακα 7.5 και ότι τα δείγμα των κατηγοριών δεν είναι αυθαίρετα δείγματα αλλά είναι κοντινότερα στην κατηγορία στην οποία ανήκουν. Ακουστικά παραδείγματα ανακατασκευασμένων μουσικών σημάτων μπορούν να βρεθούν στην ιστοσελίδα http://mir.ilsp.gr.

Κεφάλαιο 8: Μουσική Ομοιότητα Βάσει Περιεχομένου

8.1 Εισαγωγή

Το παρόν κεφάλαιο πραγματεύεται τον ορισμό και υπολογισμό της ομοιότητας μουσικών κομματιών βάσει περιεχομένου. Η προτεινόμενη μέθοδος ενσωματώνει ένα μεγάλο μέρος των τεχνικών και μεθόδων που περιγράφτηκαν προηγουμένως και αποτελείται από τρία κύρια μέρη. Το πρώτο περιλαμβάνει την εξαγωγή χαρακτηριστικών που είναι βασισμένα στον ρυθμό, στο ηχόχρωμα και στις αρμονικές αλλαγές. Το δεύτερο μέρος περιλαμβάνει την εκπαίδευση ομάδας από Deep Belief Network (DBN), καθένα από τα οποία εκπαιδεύεται σε ένα είδος χαρακτηριστικών. Τέλος, οι αποστάσεις μεταξύ των τριών συνιστωσών (ρυθμός, χροιά, αρμονία) συνδυάζονται για να υπολογιστεί η απόσταση (ή ομοιότητα) μεταξύ κομματιών.

Η αύξηση της αγοράς μουσικής μέσω διαδικτύου, ο μεγάλος πλέον όγκος των προσωπικών μουσικών συλλογών σε ηλεκτρονική μορφή που έχει ο καθένας σπίτι του, καθώς και ο αυξανόμενος αριθμός νέων και ανεξάρτητων καλλιτεχνών που διανέμουν τη μουσική τους μέσω του διαδικτύου, καθιστά αναγκαία τη χρήση τεχνολογιών για την αυτόματη οργάνωση μουσικών συλλογών. Τέτοιες τεχνολογίες θα πρέπει να ενσωματώνουν διάφορα υποσυστήματα αυτόματης μουσικής ανάλυσης όπως π.χ. κατηγοριοποίηση γένους (genre classification), αυτόματη επισημείωση (automated annotation) όπως για παράδειγμα πρόβλεψη ετικέτας (tag prediction) ή διάθεσης (mood), «σύσταση» μουσικής (music recommendation) ή και δημιουργία αυτόματων playlists.

Η «σύσταση μουσικής» ορίζεται ως το πρόβλημα ανάκτησης ενός συνόλου μουσικών κομματιών που είναι σχετικά με κάποια συγκεκριμένα κριτίρια, όπως για παράδειγμα το γένος, το στυλ ή ακόμα και βάσει παραδειγμάτων, όπως π.χ. μια λίστα με αγαπημένα κομμάτια ή αγαπημένους καλλιτέχνες. Παρότι οι διαδεδομένες εμπορικές εφαρμογές όπως το Spotify [15] και το LastFm [16] χρησιμοποιούν κυρίως τεχνικές «συνεργατικού φιλτραρίσματος» (collaborative filtering) πάνω στα μεταδεδομένα των συλλογών και στα δεδομένα χρηστών, είναι εξαιρετικά σημαντική η ανάπτυξη μεθόδων σύστασης μουσικής βάσει περιεχομένου. Ένα αξιοσημείωτο παράδειγμα τέτοιου συστήματος [Gasser2009] είναι το FM4 soundpark[17] του Αυστριακού κρατικού ραδιοφώνου. Η μηχανή σύστασης εφαρμόστηκε σε μια πλατφόρμα διανομής μουσικού υλικού, όπου – κυρίως ανεξάρτητοι – καλλιτέχνες ανέβαζαν τη μουσική τους. Μια μηχανή σύστασης βάσει περιεχομένου αποτελεί τη μόνη λύση σε ένα τέτοιο σενάριο, αφού δεν υπάρχουν επαρκή δεδομένα για τη δημιουργία συνεργατικών φίλτρων ή άλλα μεταδεδομένα όπως ομοιότητα καλλιτεχνών που θα μπορούσαν να χρησιμοποιηθούν σε μια μηχανή σύστασης. Η μηχανή αυτή βασίστηκε στα απλά χαρακτηριστικά MFCC, τα οποία συνδέονται με την χροιά. Για κάθε μουσικό κομμάτι υπολογίστηκε μια Γκαουσιανή στον διανυσματικό χώρο των MFCC. Η μουσική ομοιότητα δύο κομματιών υπολογίστηκε ως η απόκλιση Kullback–Leibler (Kullback–Leibler Divergence). Παρότι απλή, αποδείχτηκε μία επιτυχής μέθοδος στη χρήση της καθώς ο ημερήσιος αριθμός των μοναδικών κομματιών που κατέβαινε (unique track downloads) διπλασιάστηκε μετά την ενσωμάτωση της μεθόδου στην πλατφόρμα.

[15] www.spotify.com
[16] www.last.fm
[17] http://fm4.orf.at/soundpark

Ο πυρήνας ενός συστήματος αυτόματης σύστασης είναι ο υπολογισμός της ομοιότητας ή διαφοράς μεταξύ δύο μουσικών κομματιών. Όπως είπαμε, μια απλή αλλά αποτελεσματική προσέγγιση όπως η [Gasser2009] είναι η χρησιμοποίηση των MFCC υπό την παραδοχή ότι περιγράφουν την χροιά. Αντίστοιχα, οι Logan και Salomon [Logan2001] παρουσίασαν ένα σύστημα μουσικής ομοιότητας με ομαδοποίηση χαρακτηριστικών του Cepstrum, ενώ οι Aucouturier και Pachet [Aucouturier2002], [Pachet2004] υιοθέτησαν Μίξεις Γκαουσιανών Μοντέλων (Gaussian Mixture Models, GMM) στα MFCC. Ωστόσο όλες αυτές οι προσπάθειες δεν μπορούσαν να αξιολογηθούν με αξιόπιστο τρόπο και να συγκριθούν μεταξύ τους γιατί δεν υπήρχαν διαθέσιμες συλλογές δεδομένων ούτε διαδεδομένες μέθοδοι αξιολόγησης των μεθόδων. Το πρόβλημα μουσικής ομοιότητας (Audio Music Similarity) του διαγωνισμού MIREX[18] που πρωτοεμφανίστηκε το 2006 ήταν μια λύση σε αυτό το πρόβλημα. Έκτοτε, μια μεγάλη ποικιλία μεθόδων εμφανίστηκε στη βιβλιογραφία. Οι Barrington et al. [Barrington2007] υπέβαλαν την μέθοδό τους στο MIREX το 2007, ο οποίος απεικόνιζε κάθε μουσικό κομμάτι σε έναν σημασιολογικό χώρο (semantic space), όπως π.χ. το είδος, τη διάθεση, τα συναισθήματα κ.ο.κ. και η KL-divergence χρησιμοποιήθηκε ως μετρική ομοιότητας. Στο [Lukashevich2010] διερευνήθηκαν παραλλαγές της KL-divergence μεταξύ Γκαουσιανών Μίξεων που είχαν προσαρμοστεί σε φασματικά και cepstral χαρακτηριστικά. Στο [Bosteels2008] οι συγγραφείς πρότειναν τη χρήση ασαφών (fuzzy) μετρικών ομοιότητας σε φασματικά και ρυθμικά χαρακτηριστικά. Οι Mandel et al. [Mandel2006] υιοθέτησαν την «ενεργή εκμάθηση» των SVM (SVM active learning) στη στατιστική των MFCC. Στο [Bogdanov2011] προτάθηκε μία υβριδική μετρική μουσικής ομοιότητας βασισμένη σε αποστάσεις υπολογισμένες σε χαμηλού (low level) και υψηλού επιπέδου (high level) χαρακτηριστικά οι οποίες συνδυάστηκαν.

Στο [Aucouturier2008] οι συγγραφείς έρχονται αντιμέτωποι με το «φαινόμενο των κόμβων» (hubness phenomenon), δηλαδή το φαινόμενο της ύπαρξης κομματιών «κόμβων» που τείνουν να είναι κοντά σε πολλά άλλα κομμάτια και «ορφανών» κομματιών, που απέχουν μεγάλη απόσταση από τα περισσότερα κομμάτια, με συνέπεια να μη συγκαταλέγονται μεταξύ των προτεινόμενων κομματιών. Το φαινόμενο των κόμβων δεν έχει μουσικολογικά αίτια. Τα κομμάτια κόμβοι δεν είναι κομμάτια που μοιάζουν περισσότερο με πολλά κομμάτια και αντίστροφα, τα ορφανά δεν είναι κομμάτια που δεν μοιάζουν με κανένα. Το φαινόμενο των κόμβων οφείλεται στην ίδια την αναπαράσταση των κομματιών στον χώρο των χαρακτηριστικών. Επομένως για την ίδια συλλογή, δύο διαφορετικές αναπαραστάσεις των δεδομένων θα οδηγήσουν σε διαφορετικούς κόμβους και ορφανά. Επιπλέον, το φαινόμενο των κόμβων έχει αποδειχτεί ότι γίνεται πιο έντονο όσο αυξάνεται η διάσταση των χώρου των χαρακτηριστικών [Radovanovic2010], αποτελώντας ένα ακόμη παράδειγμα της κατάρας της υψηλής διάστασης. Στο [Flexer2010] παρατηρήθηκε ότι ο συνδυασμός χαρακτηριστικών μειώνει το φαινόμενο των κόμβων. Οι ίδιοι συγγραφείς [Schnitzer2011] πρότειναν μία μέθοδο κλιμάκωσης των αποστάσεων που μειώνει αυτό το φαινόμενο. Εκτενής ανάλυση του φαινομένου των κόμβων για τους αλγορίθμους που υποβλήθηκαν στο MIREX το 2011 και της επίδρασης στην επίδοσή τους μπορεί να βρεθεί στο [Flexer2012]. Στο [Charbuillet2011] χρησιμοποιήθηκε ένα Καθολικό Μοντέλο Υπόβαθρου (Universal Background Model, UBM) αποτελούμενο από GMM το οποίο εκπαιδεύτηκε στα MFCC και σε φασματικά χαρακτηριστικά.

[18] http://www.music-ir.org/mirex/wiki/2006:Audio_Music_Similarity_and_Retrieval_Results

Επιπλέον υιοθετήθηκε ένα στάδιο κανονικοποίησης των αποστάσεων για να μειωθεί το φαινόμενο των κόμβων.

Στο [Pohle2009] προτάθηκε μια μέθοδος που συνδύαζε χροιά και ρυθμό. Είναι μία από τις πιο επιτυχημένες μεθόδους από το 2009 στον διαγωνισμό MIREX. Με αντίστοιχη προσέγγιση, οι Seyerlehner et al. [Seyerlehner2010a] από το 2010 έως και σήμερα έχουν καταθέσει πληθώρα παραλλαγών της μεθόδου που περιγράφεται στο [Seyerlehner2010b] και έχουν επιτύχει την καλύτερη επίδοση τις περισσότερες φορές από τότε. Στο [Schluter2011] προτάθηκε η χρήση της Περιορισμένης Μηχανής Boltzmann Μέσης Συνδιακύμανσης (mean-covariance RBM). Τα MFCC προεπεξεργάστηκαν με wPCA και τα χαρακτηριστικά που προέκυψαν χρησιμοποιήθηκαν ως είσοδος στο RBM. Σε μια παρόμοια προσέγγιση [Schluter2013] αξιολογήθηκαν διάφορες μέθοδοι δυαδικής αναπαράστασης (συμπεριλαμβανομένων και των RBM) και αναζήτησης μουσικών κομματιών υπό το πρίσμα της γρήγορης αναζήτησης σε μεγάλες κλίμακας συλλογές.

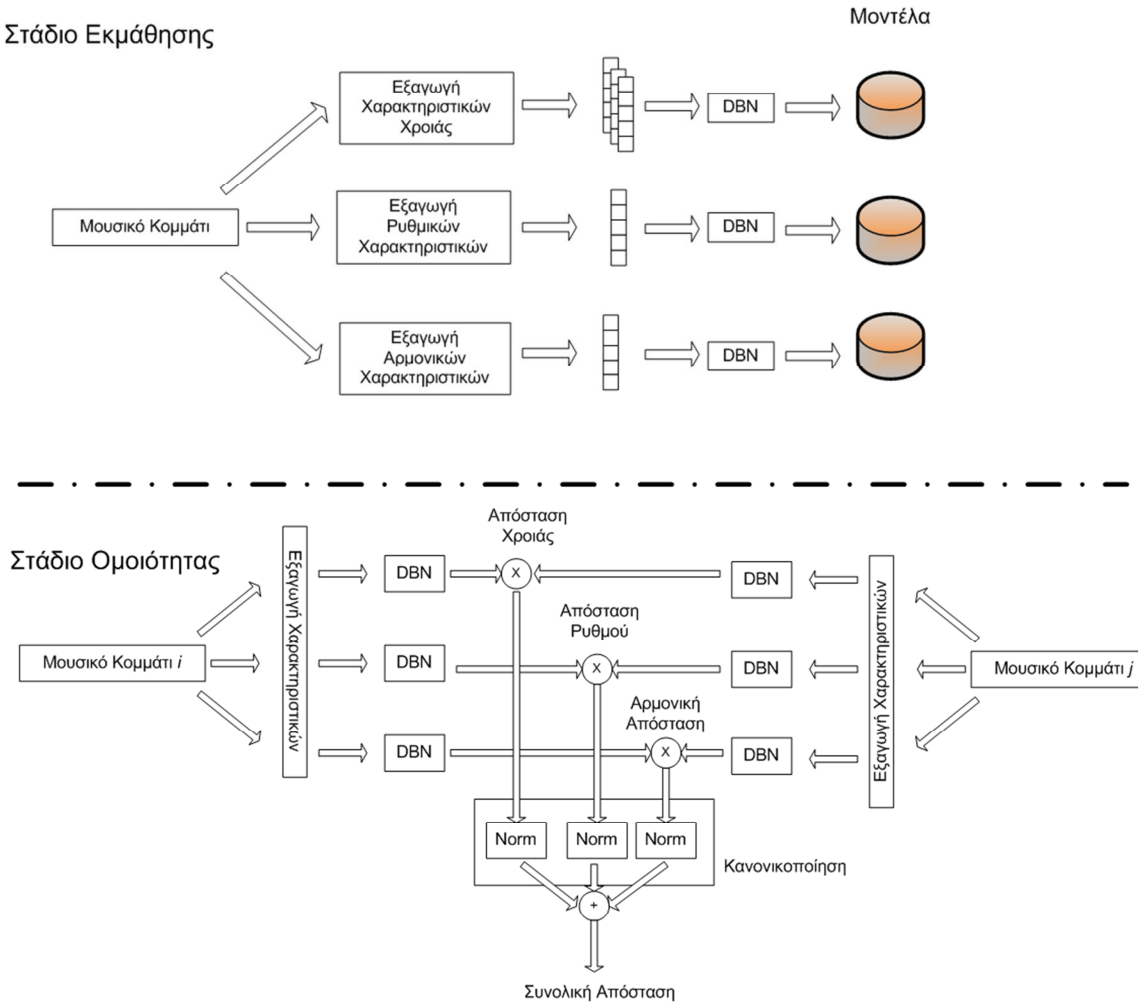
Στην επόμενη Ενότητα θα παρουσιαστεί η προτεινόμενη μέθοδος υπολογισμού μουσικής ομοιότητας. Συγκεκριμένα στην Ενότητα 8.2.1 θα γίνει η σύνοψη της προτεινόμενης μεθόδου και στην Ενότητα 8.2.2 θα περιγραφεί η εξαγωγή των χαρακτηριστικών. Στην Ενότητα 8.2.3 θα οριστεί η απόσταση μεταξύ μουσικών κομματιών και θα περιγραφή η κανονικοποίηση των αποστάσεων για την μείωση του φαινομένου των κόμβων. Στην Ενότητα 8.3 θα γίνει αξιολόγηση της μεθόδου, σύγκρισή της με άλλες μεθόδους και ανάλυση των αποτελεσμάτων. Τέλος, στην Ενότητα 8.4 θα γίνει μια σύντομη περιγραφή της online πλατφόρμας στην οποία έχει εφαρμοστεί η μέθοδος ομοιότητας.

8.2 Εύρεση Μουσικής Ομοιότητας

8.2.1 Προεπισκόπηση

Η προτεινόμενη μέθοδος υπολογισμού μουσικής ομοιότητας είναι βασισμένη στην εφαρμογή των DBN σε ένα σύνολο χαρακτηριστικών. Τα DBN έχουν εφαρμοστεί επιτυχώς σε μια πληθώρα προβλημάτων στο πεδίο της ανάκτησης μουσικής πληροφορίας. Στο [Hamel2010] εφαρμόστηκε ένα DBN απευθείας στον STFT. Με την χρήση ενός ταξινομητή SVM, τα παραχθέντα χαρακτηριστικά πέτυχαν 5% μεγαλύτερα ακρίβεια σε σύγκριση με τα MFCC χαρακτηριστικά. Αντίστοιχα, οι Nam et al. [Nam2011] εκπαίδευσαν ένα DBN στο φασματογράφημα και έναν ταξινομητή SVM για το πρόβλημα της μεταγραφής μουσικής νότας. Στο πεδίο της μουσικής ομοιότητας, η μόνη μέθοδος που έχει προταθεί και κάνει χρήση των DBN είναι η [Schluter2011]. Επιπλέον είδαμε ότι (Κεφ. 3, 4, 5) ότι τα χαρακτηριστικά που προέκυψαν από το RBM που εφαρμόστηκε στην ΣΠ χρησιμοποιήθηκαν επιτυχώς σε τρία διαφορετικά προβλήματα ρυθμικής ανάλυσης.

Η σύνοψη της προτεινόμενης μεθόδου παρουσιάζεται στο Σχ. 8.1. Αρχικά, από το ηχητικό σήμα εξάγονται τριών ειδών χαρακτηριστικά. Τα ρυθμικά χαρακτηριστικά που είναι βασισμένα στην ρυθμική ανάλυση της παρούσας διατριβής. Τα αρμονικά χαρακτηριστικά που σχετίζονται κυρίως με το μουσικό κλειδί και τις αλλαγές των συγχορδιών. Τα φασματικά χαρακτηριστικά, κατ' αναλογία με τα MFCC, συνδέονται κυρίως με την χροιά ενός κομματιού. Στην συνέχεια εκπαιδεύεται ένα DBN για κάθε μία από τις τρεις ομάδες χαρακτηριστικών σε μία μεγάλη συλλογή. Για τον υπολογισμό της απόστασης μεταξύ δύο κομματιών, αρχικά υπολογίζεται η απόσταση των επιμέρους DBN.



Σχ. 8.1. Αρχιτεκτονική της προτεινόμενης μεθόδου υπολογισμού της μουσικής ομοιότητας.

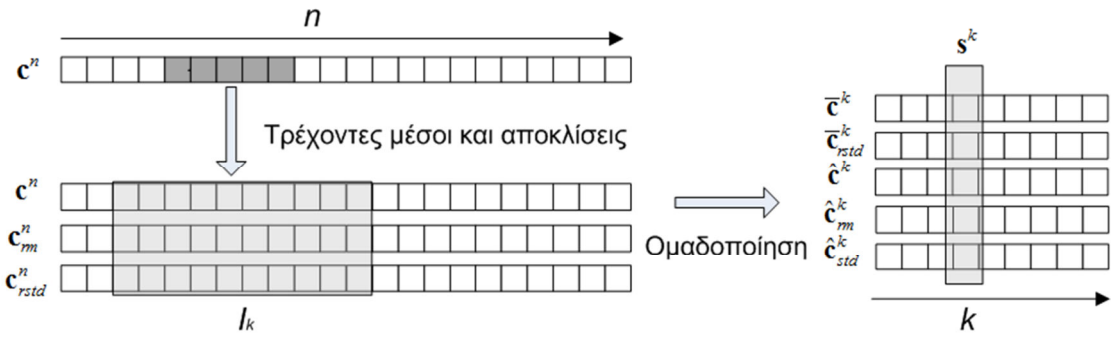
Στη συνέχεια οι τρεις αποστάσεις κανονικοποιούνται και συνδυάζονται για τον υπολογισμό της τελικής απόστασης.

8.2.2 Εξαγωγή Χαρακτηριστικών

Σε αυτή την Ενότητα θα παρουσιαστούν αναλυτικά τα τρία είδη χαρακτηριστικών που σχετίζονται με την χροιά, τον ρυθμό και το αρμονικό περιεχόμενοι που στη συνέχεια χρησιμοποιούνται για την εκπαίδευση των DBN. Τα ρυθμικά και αρμονικά χαρακτηριστικά υπολογίζονται στο επίπεδο του κομματιού (1 διάνυσμα ανά κομμάτι-απόσπασμα), ενώ τα χαρακτηριστικά χροιάς σε επίπεδο τμήματος (segment).

Χαρακτηριστικά σχετικά με τη χροιά

Για τη μοντελοποίηση της χροιάς της μουσικής, εξήχθησαν μια σειρά από φασματικά και Cepstral χαρακτηριστικά. Αρχικά το σήμα επαναδειγματοληπτείται στη συχνότητα 22.05kHz και υπολογίζεται ο SFTF με μήκος παραθύρου 46.4ms (1024 δείγματα) με κατά το ήμισυ επικάλυψη (512 δείγματα) μεταξύ διαδοχικών παραθύρων. Στη συνέχεια, για κάθε παράθυρο ανάλυσης υπολογίζονται:



Σχ. 8.2. Γραφική αναπαράσταση της διαδικασίας εξαγωγής των χαρακτηριστικών χροιάς.

- 20 συντελεστές MFCC που έχουν προκύψει από 50 φίλτρα και συμβολίζονται με **m**.
- Ο φασματικός μέσος (centroid), η φασματική διακύμανση (variance), η φασματική συμπάγεια (compactness), η φασματική ροή (flux) και η συχνότητα μέχρι την οποία κατανέμεται το 95% της φασματικής ενέργειας (Roll Off at 95%). Συμβολίζουμε αυτά τα χαρακτηριστικά με (s_c, s_v, s_{comp}, s_{flux}, s_{roll})
- Η 2η, 3η και 4η φασματική ροπή τις οποίες συμβολίζουμε με (μ_2, μ_3, μ_4)

Τα χαρακτηριστικά αυτά ενώνονται σε ένα διάνυσμα 28 διαστάσεων που συμβολίζουμε με $c = [\mathbf{m}, s_c, s_v, s_{comp}, s_{flux}, s_{roll}, \mu_2, \mu_3, \mu_4]$. Αν $\mathbf{c}_n$ είναι το διάνυσμα για το παράθυρο n, στην συνέχεια υπολογίζονται οι τρέχοντες μέσοι (running means) $\mathbf{c}_n^{rm}$ και οι τρέχουσες τυπικές αποκλίσεις (running standard deviation) $\mathbf{c}_n^{rstd}$ των $\mathbf{c}_n$ για τμήματα 10 διαδοχικών παράθυρων ανάλυσης (~235ms). Στη συνέχεια, οι προκύπτουσες ακολουθίες χαρακτηριστικών ομαδοποιούνται σε μεγάλα τμήματα των 100 παραθύρων ανάλυσης (~2.35s) με 50 παράθυρα επικάλυψη (~1.18s). Για κάθε τμήμα $[\mathbf{c}_i, \mathbf{c}_i^{rm}, \mathbf{c}_i^{rstd}], i \in I_k$, όπου το I_k συμβολίζει το σύνολο των παραθύρων i που ανήκουν στο τμήμα k, τα παρακάτω στατιστικά στοιχεία εξάγονται:

- $\bar{\mathbf{c}}_k = \mathbf{mean}(\mathbf{c}_i) = \mathbf{mean}(\mathbf{c}_i^{rm}),\ i \in I_k$
- $\bar{\mathbf{c}}_k^{rstd} = \mathbf{mean}(\mathbf{c}_i^{rstd}),\ i \in I_k$
- $\hat{\mathbf{c}}_k = \mathbf{std}(\mathbf{c}_i),\ i \in I_k$
- $\hat{\mathbf{c}}_k^{rm} = \mathbf{std}(\mathbf{c}_i^{rm})\ i \in I_k$
- $\hat{\mathbf{c}}_k^{rstd} = \mathbf{std}(\mathbf{c}_i^{rstd})\ i \in I_k$

Τέλος, τα παραπάνω στατιστικά ενώνονται σε ένα διάνυσμα $\mathbf{s}_k = [\bar{\mathbf{c}}_k, \bar{\mathbf{c}}_k^{rstd}, \hat{\mathbf{c}}_k, \hat{\mathbf{c}}_k^{rm}, \hat{\mathbf{c}}_k^{rstd}]$ συνολικής διάστασης 140. Επομένως, κάθε μουσικό κομμάτι παριστάνεται με ένα σύνολο $\mathbf{S} = \{\mathbf{s}_k\}_{k=1\ldots K}$ K διανυσμάτων, όπου K είναι ο συνολικός των τμημάτων του μουσικού κομματιού. Η διαδικασία εξαγωγής των χαρακτηριστικών χροιάς παρουσιάζεται στο Σχ. 8.2.

Χαρακτηριστικά σχετικά με τον ρυθμό

Το γεγονός ότι η ΣΠ που περιγράφτηκε στο Κεφ. 2 χρησιμοποιήθηκε με επιτυχία σε διάφορα προβλήματα ρυθμικής ανάλυσης, μας οδηγεί στο συμπέρασμα ότι μουσικά κομμάτια με «κοντινές» ΣΠ θα έχουν και παρόμοια ρυθμικό περιεχόμενο. Επομένως χρησιμοποιούμε ως ρυθμική αναπαράσταση την ΣΠ όπως ορίζεται από την Εξ. 2.32. Συμβολίζουμε εφεξής με **v** την ΣΠ.

Χαρακτηριστικά σχετικά με το αρμονικό περιεχόμενο

Για την αρμονική ομοιότητα προτείνεται ένα σύνολο χαρακτηριστικών που προέρχεται από το διάνυσμα χρώματος και το οποίο περιγράφει την τονική εξέλιξη ενός κομματιού. Παρότι τα χαρακτηριστικά χρώματος περιέχουν πληροφορία σχετική με τις συγχορδίες ή το μουσικό κλειδί, δεν είναι είναι κατάλληλα για να διαχωρίσουν σχετικά μεταξύ τους κλειδιά. Για παράδειγμα δύο μουσικά κομμάτια με κλειδιά C Major και A minor αντίστοιχα θα έχουν παρόμοια ιστογράμματα των χαρακτηριστικών χρώματος. Αυτό δεν αποτελεί πάντα πρόβλημα για εφαρμογές όπως η αναγνώριση συγχορδίας, αλλά δεν είναι επιθυμητό για την μοντελοποίηση ιδιοτήτων της μουσικής που σχετίζονται με την τονικότητα. Ένας ακροατής θα αντιληθφεί πολύ διαφορετικά ένα κομμάτι που είναι γραμμένο σε Μείζονα κλίμακα σε σχέση με ένα κομμάτι που είναι γραμμένο στην σχετική της Ελάσσονα κλίμακα. Με την παραδοχή ότι το είδος του μουσικού κλειδιού (Μείζων ή Ελάσσων) συνδέεται με αντιληπτικά χαρακτηριστικά όπως το στυλ ή τη διάθεση (mood) ενός κομματιού, η απευθείας υιοθέτηση των χαρακτηριστικών χρώματος σε ένα σύστημα μουσικής ομοιότητας δεν θα είναι αποτελεσματική. Επιπλέον, μουσικά κομμάτια με παρόμοιες (σχετικές) ακολουθίες συγχορδιών συχνά δίνουν μια αίσθηση ομοιότητας στον ακροατή.

Μία ιδιότητα που μπορεί να ξεχωρίσει τα μείζονα από τα ελάσσονα τονικά κλειδιά είναι η μετάβαση των συγχορδιών. Δύο σχετικά κλειδιά έχουν τους ίδιους τόνους, αλλά διαφορετικές μεταβάσεις από τόνο σε τόνο. Στην παρούσα διατριβή προτείνεται η «σχετική διαφορά χρώματος» (relative chroma difference, RCD) που συμβολίζεται με **rcd**[·]. Η **rcd**[·] είναι ένα διάνυσμα δώδεκα διαστάσεων όπου κάθε στοιχείο του αντιστοιχεί σε μία από τις δώδεκα τιμές χρώματος. Η **rcd**[·]υπολογίζεται για διαφορετικές χρονικές στιγμές. Για μια χρονική στιγμή m, η **rcd**$[i,m]$ υπολογίζεται για δύο διαδοχικά διανύσματα χρώματος $x_{ch}[j,m-1]$, $x_{ch}[j,m]$, $j=1\ldots12$, όπου j είναι ο δείκτης του χρώματος. Το i στοιχείο του αναπαριστά την ισχύ της σχετικής μεταβάσης από το προηγούμενο διάνυσμα χρώματος $x_{ch}[j,m-1]$ στο τρέχον διάνυσμα χρώματος $x_{ch}[j,m]$ κατά i τόνους. Η $rcd[i,m]$ υπολογίζεται ως:

$$rcd[i,m]=\frac{\sum_{\mathbf{k=1}}^{\mathbf{12}} x_{ch}[mod(i+k,12),m]\cdot x_{ch}[k,m-1]}{\|x_{ch}[\cdot,m]\|\cdot\|x_{ch}[\cdot,m-1]\|} \tag{8.1}$$

Δηλαδή η $rcd[i,m]$ είναι το συνημίτονο του $x_{ch}[\cdot,m-1]$ με την τονική μετάθεση του $x_{ch}[\cdot,m]$ κατά i τόνους. Η RCD έχει την επιθυμητή ιδιότητα ότι είναι αναλλοίωτη σε μεταθέτεις τόνου. Μετατεθιμένες τονικά εκδόσεις του ίδιου κομματιού θα έχουν την ίδια RDC και το i στοιχείο της μία χρονική στιγμή θα συμβολίζει την ισχύ της κίνησης του διανύσματος χρώματος κατά i τόνους εκείνη τη χρονική στιγμή.

Ωστόσο πριν την πρακτική εφαρμογή της RCD χρειάζεται κάποια προεπεξεργασία. Η Εξ. 8.1 έχει νόημα μόνο σε ισχυρές μεταβολές του χαρακτηριστικών χρώματος ή σε χρονικές στιγμές που παρατηρείται αλλαγή συγχορδίας. Επομένως πριν τον υπολογισμό της RCD από την Εξ. 8.1 προηγείται

ένα στάδιο εύρεσης των χρονικών στιγμών αλλαγής συγχορδίας. Για τον εντοπισμό αλλαγής συγχορδίας θεωρούμε την διαφορά ως προς τον χρόνο των χαρακτηριστικών χρώματος:

$$\Delta x_{ch}[m] = \sum_i \| x_{ch}[i,m] - x_{ch}[i,m-1] \| \tag{8.2}$$

Οι κορυφές του Δx_{ch} ορίζει ομάδες χρώματος όπου η μέση τιμή $\bar{x}_{ch}[i,l]$ υπολογίζεται ως:

$$\bar{x}_{ch}[i,l] = \sum_{m=p_{l-1}}^{p_l - 1} x_{ch}[i,m] \,/ (p_l - p_{l-1} + 1) \tag{8.3}$$

όπου με p_l συμβολίζεται η θέση της l κατά σειρά κορυφής. Στη συνέχεια η $\mathbf{rcd}[\cdot]$ υπολογίζεται όπως στην 8.1 αντικαθιστώντας τα $x_{ch}[i,m]$ με $\bar{x}_{ch}[i,l]$. Ο τελικός αρμονικός περιγραφέας $\mathbf{h}$ είναι ένα διάνυσμα 36 διαστάσεων που σχηματίζεται με παράθεση της μέσης τιμής, της τυπικής απόκλισης και της μέγιστης τιμής της RCD ως προς τον χρόνο

$$\mathbf{h} = [\text{mean}_l(\mathbf{rcd}[i,l]) \ldots \mid \text{std}_l(\mathbf{rcd}[i,l]) \mid \max_l(\mathbf{rcd}[i,l])] \tag{8.4}$$

8.2.3 Υπολογισμός Αποστάσεων Μουσικών Κομματιών

Έστω οι 3 περιγραφείς $y_j = \left(\{\mathbf{s}_k^j\}, \mathbf{v}^j, \mathbf{h}^j\right)$ του j κομματιού. Όπως περιγράφτηκε στην Ενότητα 8.2.1, αυτοί οι περιγραφείς μετασχηματίζονται από τα τρία αντίστοιχα DBN. Στη συνέχεια ο υπολογισμός των αποστάσεων γίνεται στις εξόδους των DBN. Οι έξοδοι των DBN συμβολίζονται με $\{\tilde{\mathbf{s}}_k^j\}, \tilde{\mathbf{v}}^j, \tilde{\mathbf{h}}^j$ αντίστοιχα. Στη συνέχεια, από τα χαρακτηριστικά χροιάς $\tilde{\mathbf{s}}_k^j$ θεωρούμε τη μέση τιμή $\bar{\mathbf{s}}^j = mean(\tilde{\mathbf{s}}_k^j)$. Επομένως, το μουσικό κομμάτι j αναπαρίσταται τελικά ως $\tilde{y}_j = \left(\bar{\mathbf{s}}^j, \tilde{\mathbf{v}}^j, \tilde{\mathbf{h}}^j\right)$. Για να υπολογιστεί η απόσταση ενός ερωτήματος (query) $\tilde{q} = \left(\bar{\mathbf{s}}^q, \tilde{\mathbf{v}}^q, \tilde{\mathbf{h}}^q\right)$ από ένα απόσπασμα $\tilde{y}_j$, οι επιμέρους αποστάσεις για κάθε από τις τρεις συνιστώσες υπολογίζονται με την χρήση της $\|\cdot\|_1$:

$$d^f\left(\tilde{q}, \tilde{y}_j\right) = \left\| f^q - f^j \right\|_1, \; f \in \{\bar{\mathbf{s}}, \tilde{\mathbf{v}}, \tilde{\mathbf{h}}\} \tag{8.5}$$

Η επιλογή της L_1 νόρμας βασίζεται στο γεγονός ότι οι έξοδοι των DBN $\tilde{y}_j = \left(\bar{\mathbf{s}}^j, \tilde{\mathbf{v}}^j, \tilde{\mathbf{h}}^j\right)$ είναι αραιές αναπαραστάσεις λόγω των δυαδικών στοιχείων των κρυφών επιπέδων των RBM που απαρτίζουν τα DBN και επομένως η L_1 αναπαριστά πιο αποτελεσματικά τις αποστάσεις σε αυτόν τον χώρο. Κάτι αντίστοιχο διαπιστώθηκε πειραματικά και στην ανάλυση της Ενότητας 4.3.

Στη συνέχεια, για κάθε μία από τις συνιστώσες $f \in \{\bar{\mathbf{s}}, \tilde{\mathbf{v}}, \tilde{\mathbf{h}}\}$ υπολογίζεται η μέση τιμή μ_f^q και η τυπική απόκλιση σ_f^q των αποστάσεων των K κοντινότερων γειτόνων του ερωτήματος $\tilde{q}$. Οι τιμές αυτές χρησιμοποιούνται για να κανονικοποιήσουν τις αποστάσεις του $\tilde{q}$ (Εξ. 8.5) από το τυχαίο παράδειγμα της συλλογής $\tilde{y}_j$:

$$\hat{d}^f\left(\tilde{q}, \tilde{y}_j\right) = \frac{d^f(\tilde{q}, \tilde{y}_j) - \mu_f^q}{\sigma_f^q}, f \in \{\bar{\mathbf{s}}, \tilde{\mathbf{v}}, \tilde{\mathbf{h}}\} \tag{8.6}$$

Ας σημειωθεί ότι οι κανονικοποιημένες αποστάσεις δεν είναι πια συμμετρικές. Τέλος, η συνολική απόσταση μεταξύ των $\tilde{q}$και $\tilde{y}_j$είναι το σταθμισμένο άθροισμα των επιμέρους αποστάσεων:

$$\hat{d}\left(\tilde{q}, \tilde{y}_j\right) = w_{\mathbf{s}} \hat{d}^{\mathbf{s}}\left(\tilde{q}, \tilde{y}_j\right) + w_{\mathbf{v}} \hat{d}^{\mathbf{v}}\left(\tilde{q}, \tilde{y}_j\right) + w_{\mathbf{h}} \hat{d}^{\mathbf{h}}\left(\tilde{q}, \tilde{y}_j\right) \tag{8.7}$$

Η κανονικοποίηση των αποστάσων γίνεται για δύο λόγους. Ο πρώτος είναι για να κανονικοποιηθεί το εύρος τιμών των αποστάσεων για τις τρεις συνιστώσες $f \in \{\bar{\mathbf{s}}, \tilde{\mathbf{v}}, \tilde{\mathbf{h}}\}$ έτσι ώστε καμία να μην κυριαρχεί των υπολοίπων. Αφού μετά την κανονικοποίηση οι αποστάσεις είναι σταθμισμένες, η προτίμηση σε κάποια από τις διαστάσεις μπορεί να τεθεί χειρονακτικά μέσω των βαρών $w_{\mathbf{s}}$, $w_{\mathbf{v}}$, $w_{\mathbf{h}}$ της Εξ. 8.7. Ο δεύτερος λόγος ήταν η μείωση του φαινομένου κόμβων. Πράγματι, η κανονικοποίηση της Εξ. 8.6 είναι παρόμοια με την κανονικοποίηση που προτάθηκε στο [Schnitzer2012]. Η κανονικοποίηση των αποστάσεων λαμβάνοντας υπόψη μόνο τους κοντινότερους γείτονες για τον υπολογισμό των μ_f^q και σ_f^q και όχι ολόκληρης της συλλογής, μειώνει ανεπιθύμητα φαινόμενα που μπορεί να οφείλονται σε μη σχετικά παραδείγματα $\tilde{y}_m$ που έχουν πολύ μεγάλη απόσταση με το ερώτημα $\tilde{q}$. Αυτός ο ισχυρισμός μπορεί να εξηγηθεί ποιοτικά με το ακόλουθο παράδειγμα. Ας θεωρήσουμε ότι στην συλλογή προστίθεται ένα νέο κομμάτι $\tilde{y}_m$, που έχει πολύ μεγάλη απόσταση από το ερώτημα $\tilde{q}$. Τότε οι αποστάσεις του $\tilde{q}$ από όλα τα υπόλοιπα παραδείγματα δεν θα αλλάξει, το οποίο είναι επιθυμητό αφού το $\tilde{y}_m$ είναι πολύ μακρινό από το $\tilde{q}$ και δεν θα έπρεπε να το επηρεάζει. Στην περίπτωση όμως που λαμβάναμε υπόψη όλα τα δεδομένα της συλλογής κατά τον υπολογισμό των όρων κανονικοποίησης μ_f^q και σ_f^q, τότε το $\tilde{y}_m$ θα επηρέαζε τις αποστάσεις του $\tilde{q}$ με τους γείτονές του. Αντίθετα, αν το $\tilde{y}_m$ ήταν «κοντινό» του $\tilde{q}$ τότε θα επηρέαζε τις αποστάσεις με τους γείτονές του, κάτι που είναι επιθυμητό. Απλούστερα, ένα κομμάτι $\tilde{y}_m$ θα πρέπει να επηρεάζει την απόσταση του $\tilde{q}$ με τους γείτονές του μόνο αν ανήκει και αυτό στη γειτονιά του $\tilde{q}$, προϋπόθεση που ικανοποιείται λόγω της τοπικής κανονικοποίησης.

8.3 Πειραματικά Αποτελέσματα

Όπως και στην περίπτωση των RBM στα προηγούμενα πειράματα, το DBN εκπαιδεύτηκε σε ένα υποσύνολο του Million Song Datset αποτελούμενο από 130.000 μουσικά αποσπάσματα. Τα αρχικά χαρακτηριστικά $y_j = \left(\{\mathbf{s}_k^j\}, \mathbf{v}^j, \mathbf{h}^j\right)$ κανονικοποιήθηκαν ώστε σε κάθε διάσταση να επιτυγχάνεται μηδενική μέση τιμή και μοναδιαία τυπική απόκλιση σε όλη τη συλλογή. Για το πρώτο επίπεδο του DBN υιοθετήθηκαν Γκαουσιανά στοιχεία ενώ για τα κρυφά επίπεδα στοιχεία Bernoulli. Το πλήθος των κρυφών επιπέδων επιλέχτηκε ίσο με 3 για όλα τα DBN. Τα DBN των ρυθμικών χαρακτηριστικών και των χαρακτηριστικών χροιάς είχαν 500 στοιχεία σε κάθε κρυφό επίπεδο, ενώ για το DBN των αρμονικών χαρακτηριστικών επιλέχτηκαν 300 στοιχεία. Ο στόχος «αραιότητας» (Εξ. 3.27) τέθηκε ίσος με $\overline{p} = 0.1$. Ο αλγόριθμος εκμάθησης των RBM ήταν ο Constrastive Divergence με ένα βήμα δειγματοληψίας Gibbs (k=1, Εξ. 3.23). Ο ρυθμός μάθησης ε τέθηκε ίσος με 0.01 για το RBM του 1ου επιπέδου και 0.1 για τα υπόλοιπα δύο. Το εύρος των τέμπο της ανάλυσης περιοδικότητας (Εξ. 2.20) ήταν $T_{\min} = 30$ BPM και $T_{\max} = 500$ BPM με βήμα $\Delta T = 1$ BPM. Ο αριθμός των φίλτρων ενέργειας E (Εξ. 2.30) επιλέχτηκε όπως και στα υπόλοιπα πειράματα ίσος με E=8. Επομένως η διάσταση της ΣΠ είναι (8+1)x471=4239. Το πλήθος των κοντινότερων γειτόνων για την κανονικοποίηση των αποστάσεων επιλέχτηκε K=50.

Πειραματικά αποτελέσματα στον διαγωνισμό MIREX

Για να μετρηθεί η επίδραση του χαρατηριστικού RCD δύο εκδόσεις της προτεινόμενης μεθόδου υποβλήθηκαν στο διεθνή διαγωνισμό MIREX το έτος 2013. Η πρώτη έκδοση είναι η πλήρης μέθοδος όπως περιγράφτηκε στο παρόν κεφάλαιο και θα συμβολίζεται εφεξής ως GKC1.

Μέθοδος	DM1	DM2	**GKC1**	**GKC2**	PS1	RA1	SS2	SSPK
Cat Score	0.926	0.992	0.896	0.978	1.142	0.276	1.184	1.164
Fine Score	46.26	48.07	43.80	48.04	53.81	18.86	55.21	54.10

Πίνακας. 8.1. Συγκριτικά αποτελέσματα του διαγωνισμού MIREX για το έτος 2013. Οι προτεινόμενες μέθοδοι συμβολίζονται με GKC1 και GKC2 .

Η δεύτερη έκδοση της μεθόδου αγνοεί το RCD και λαμβάνει υπόψει μόνο τις συνιστώσες ρυθμού και χροιάς. Τα αποτελέσματα όλων των μεθόδων συνοψίζονται στον Πίνακα 8.1. Οι αλγόριθμοι αξιολογήθηκαν ως εξής: για ένα σύνολο από 50 ερωτήματα (queries), κάθε αλγόριθμος επέστρεψε τα 5 πιο όμοια κομμάτια (μετά από φιλτράρισμα καλλιτέχνη) στο ερώτημα. Στη συνέχεια, κάθε αποτέλεσμα αξιολογήθηκε από ακροατές σε δύο κλίμακες. Η μία κλίμακα (Cat Score) αποτελείται από τρεις βαθμονομημένες κατηγορίες (0: μη όμοιο, 1: ενδιάμεσο, 2: όμοιο) και η δεύτερη (Fine Score) είναι κλίμακα ομοιότητας στο διάστημα [0..100]. Τα βάρη της άθροισης των επιμέρους αποστάσεων της Εξ. 8.7 $w_{\mathbf{s}}$, $w_{\mathbf{v}}$, $w_{\mathbf{h}}$ επιλέχθηκαν ίσα μεταξύ τους.

Όπως φαίνεται από τον Πίνακα 8.1, τα πειραματικά αποτελέσματα υποδεικνύουν ότι η υιοθέτηση των RCD χαρακτηριστικών μείωσαν σημαντικά την επίδοση της μεθόδου. Μια πιθανή εξήγηση αυτής της μείωσης είναι ότι τα χαρακτηριστικά δεν έχουν αξιολογηθεί σε πραγματικά δεδομένα αλλά σε απλά τεχνητά παραδείγματα, επομένως μπορεί τα ίδια τα χαρακτηριστικά RCD να μην είναι αποτελεσματικά. Ωστόσο δεν υπάρχει ένδειξη μέχρι τώρα στη βιβλιογραφία ότι χαρακτηριστικά που περιέχουν αρμονική πληροφορία όπως το RCD μπορούν τα βελτιώσουν τα αποτελέσματα μιας μεθόδου μουσικής ομοιότητας. Πρέπει τα χαρακτηριστικά RCD να αξιολογηθούν και σε άλλα προβλήματα στο πλαίσιο της αρμονική ανάλυσης όπως για παράδειγμα στην αναγνώριση μουσικού κλειδιού, έτσι ώστε να ελεγχθεί η εγκυρότητά τους. Επιπλέον πριν την χρησιμοποίηση των RCD στη μουσική ομοιότητα μπορούν να χρησιμοποιηθούν πιο εδραιωμένες τεχνικές αρμονικής ανάλυσης ώστε να μελετηθεί η θετική επίδραση ή μη ενός συστήματος αρμονικής ανάλυσης σε μια μέθοδο μουσικής ομοιότητας. Κάτι τέτοιο ωστόσο υπόκειται στους περιορισμούς που αναφέρθηκαν προηγουμένως σχετικά με το προαπαιτούμενο τα χαρακτηριστικά να μένουν αναλλοίωτα σε μεταθέσεις τόνου.

Μια άλλη πιθανή εξήγηση της αρνητικής επίδρασης αποτελεί το ισοσταθμισμένο ζύγισμα των επιμέρους αποστάσεων στην Εξ. 8.7. Παρά το γεγονός ότι οι επιμέρους αποστάσεις έχουν κανονικοποιηθεί ώστε να έχουν ίδιο εύρος τιμών και επομένως την ίδια επίδραση στην συνολική απόσταση, μπορεί να μην είναι σωστό να θεωρηθούν αυτές οι αποστάσεις με ισοδύναμα βάρη. Οι ακροατές τείνουν να έχουν προτιμήσεις όσον αφορά κάποιες μουσικές διαστάσεις που λαμβάνουν υπόψη για την αντίληψη της μουσικής ομοιότητας. Είναι πιο πιθανό κάποιος ακροατής να αντιληφθεί ως όμοια δύο τραγούδια που ανήκουν στο ίδιο γένος (μικρότερη απόσταση στη διάσταση της χροιάς) ακόμα και αν έχουν τελείως διαφορετικό ρυθμικό και αρμονικό περιεχόμενο, παρά δύο κομμάτια που έχουν παρόμοια αρμονική δομή αλλά διαφορετικά χροιά και ρυθμό. Μια εύλογη ιεραρχία των χαρακτηριστικών που αντιλαμβάνεται ο ακροατής για να κρίνει την ομοιότητα μεταξύ δύο μουσικών κομματιών είναι πρώτα η χροιά, μετά ο ρυθμός και τέλος τη αρμονία. Επομένως τα βάρη $w_{\mathbf{s}}$, $w_{\mathbf{v}}$, $w_{\mathbf{h}}$ μπορούν να οριστούν με αντίστοιχο τρόπο.

Αν και οι προτεινόμενες μέθοδοι αποδίδουν υποδεέστερα από τις καλύτερες μεθόδους (PS1, SS2, SSPK), τα πειραματικά αποτελέσματα είναι ενθαρρυντικά.

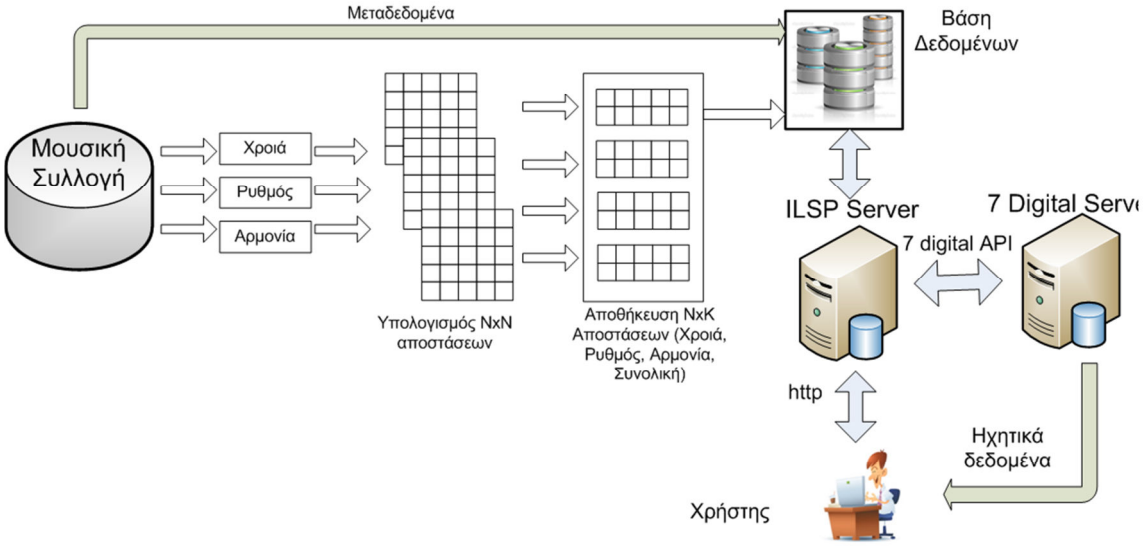
	DM1	DM2	**GKC1**	**GKC2**	PS1	RA1	SS2	SSPK
Ορφανά (%)	8.16	6.94	10.9	7.46	12	27.5	3.79	5.47
Μέγιστο Κέντρο	42	24	39	27	81	1916	35	33
Γένος (%)	74	76.3	56.1	61.5	77.5	18.1	77.4	77.5

Πίνακας 8.2. Συγκριτικά στατιστικά στοιχεία των αποτελεσμάτων του διαγωνισμού MIREX για το έτος 2013. Οι προτεινόμενες μέθοδοι συμβολίζονται με GKC1 και GKC2 .

Επιπλέον, υπάρχουν πολλά περιθώρια βελτίωσης, αφού οι παράμετροι της μεθόδου υπολογίστηκαν με πρόχειρα ακουστικά πειράματα, λόγω της έλλειψης επισημειωμένων συλλογών σχετικά με τη μουσική ομοιότητα. Επομένως πρέπει να διευρευνηθούν τρόποι υπολογισμού των (υπο)βέλτιστων παραμέτρων και χωρίς διαθέσιμα δεδομένα. Επιπλέον, μπορούν να γίνουν τροποποιήσεις της ίδιας της μεθόδου, όπως για παράδειγμα υιοθετώντας Γραμμικές Μονάδες Ανόρθωσης στα κρυφά επίπεδα ή την ενσωμάτωση ενός επιπλέον βήματος εκπαίδευσης των DBN, όπως για παράδειγμα την εφαρμογή back propagation ώστε τα εξαγόμενα χαρακτηριστικά να διακρίνουν πιο εύκολα διάφορες κλάσεις. Μια τέτοια προσέγγιση θα έχει κοινά στοιχεία με τις μεθόδους [Barrington2007] και [Charbuillet2011] που - προβάλουν τα κομμάτια σε χαρακτηριστικά υψηλότερου επιπέδου.

Πέρα από τα αποτελέσματα αξιολόγησης οι διοργανωτές του MIREX παρείχαν και κάποια στατιστικά στοιχεία που προσφέρουν μια βαθύτερη κατανόηση των ιδιοτήτων των μεθόδων. Τα σταστιστικά αυτά στοιχεία παρουσιάζονται στον Πίνακα 8.2. Η $1^{η}$ γραμμή του Πίνακα 8.2 περιέχει το ποσοστό των κομματιών που δεν βρέθηκαν ποτέ στη λίστα των 5 πιο κοντινών αποτελεσμάτων. Αποτελεί αντιπροσωπευτική ποσότητα του ποσοστού των ορφανών. Η $2^{η}$ γραμμή παρουσιάζει το μέγιστο πλήθος των φορών που κάποιο τραγούδι ήταν στα 5 κοντινότερα αποτελέσματα. Για παράδειγμα, για τον αλγόριθμο DM1 υπήρξε κομμάτι που ήταν 42 φορές στα 5 πρώτα αποτελέσματα για κάθε ερωτήματος. Η τιμή αυτή αποτελεί τρόπο μέτρησης του φαινομένου των κέντρων. Η $3^{η}$ γραμμή παρουσιάζει το ποσοστό των 5 αποτελεσμάτων που ανήκαν στο ίδιο γένος με το ερώτημα. Αν συνδυάσουμε τον Πίνακα 8.2 με τα αποτελέσματα του Πίνακα 8.1, μπορούμε να κατηγοριοποίησουμε τις μεθόδους σε τέσσερις ομάδες (αγνοώντας τη μέθοδο RA1). Οι SS2 και SSPK επιτυγχάνουν υψηλά ποσοστά ομοιότητας, ανάκτησης όσον αφορά το γένος και μικρό hubness. Η PS1 έχει και αυτή υψηλή επίδοση αλλά υψηλότερο hubness. Οι DM1 και DM2 έχουν μέτρια επίδοση στην ομοιότητα, μέτριο hubness και υψηλή επίδοση στην ανάκτηση κατηγορίας. Η προτεινόμενη μέθοδος GKC2 επιτυγχάνει μέτριο hubness, μέτρια βαθμολόγηση ομοιότητας και μέτρια ποσοστό ανάκτησης γένους.

Συγκρίνοντας τις μεθόδους SS2, SSPK και PS1 λαμβάνοντας υπόψη τα παραπάνω, μπορούμε να ισχυριστούμε ότι η SS2 είναι καλύτερη. Αυτό δεν οφείλεται τόσο στο γεγονός ότι επιτυγχάνει λίγο μεγαλύτερη βαθμολογία ομοιότητας, αλλά στο γεγονός ότι πέτυχε πολύ μεγαλύτερη κάλυψη της συλλογής στα αποτελέσματα. Μόλις το 3.9% των κομματιών της συλλογής δεν εμφανίστηκαν στο αποτέλεσμα κανενός ερωτήματος. Οι μέθοδοι DM1 και DM2 συμπεριφέρονται περισσότερο ως ταξινομητές είδους, παρά ως μηχανές υπολογισμού ομοιότητας. Παρότι επιτυγχάνουν υψηλό ποσοστό ανάκτησης είδους στα αποτελέσματα (~76% και εφάμιλλο με τις SS2, SSPK και PS1), τα αποτελέσματα αυτά έχουν σχετικά μικρό βαθμό ομοιότητας με τα ερωτήματα.



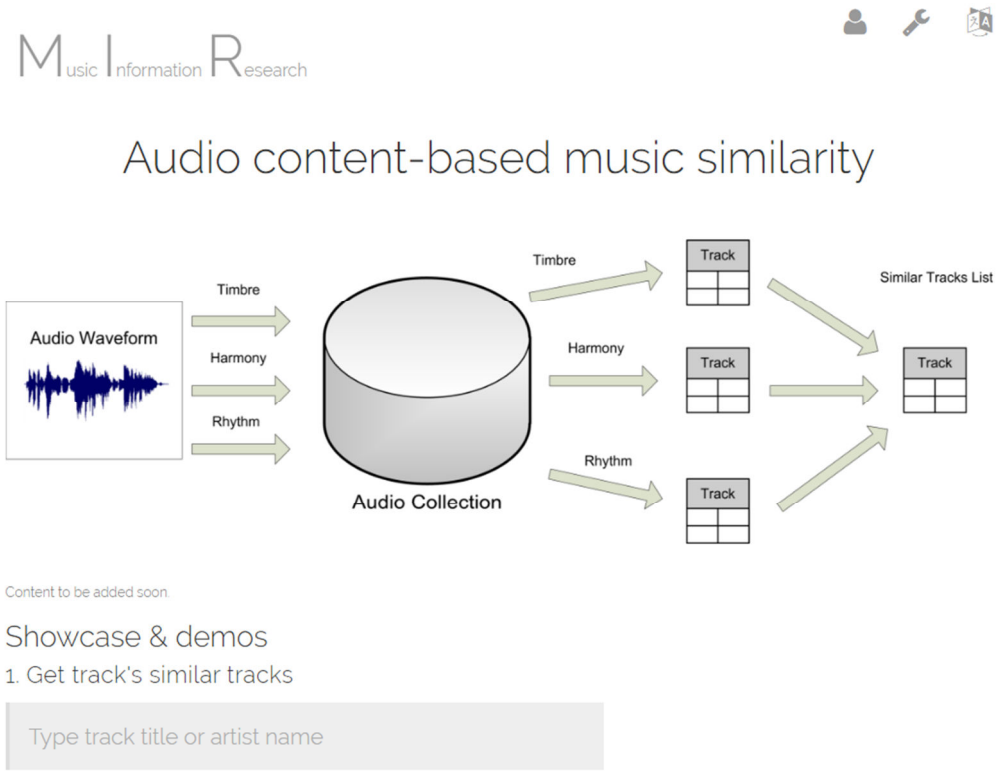
Σχ. 8.3. Αρχιτεκτονική της διαδικτυακής πλατφόρμας ανάκτησης μουσικής.

Αντίθερα, οι προτεινόμενες μέθοδοι GKC1 και GKC2 έχουν την τελείως αντίθετη συμπεριφορά. Ενώ επιτυγχάνουν τον ίδιο περίπου βαθμό ομοιότητας με τις DM1 και DM2, τα αποτελέσματα ανήκουν στο ίδιο είδος με το ερώτημα σε πολύ μικρότερο ποσοστό. Δηλαδή η προτεινόμενη μέθοδος προτείνει κομμάτια που ενώ είναι ίδιου βαθμού ομοιότητας με τις DM1 και DM2, ανήκουν σε άλλο είδος. Αυτό είναι ένα φαινόμενο που οφείλεται κυρίως στον μετασχηματισμό των χαρακτηριστικών από τα DBN, αφού (α) μοντελοποιούν πιο αφαιρετικές δομές της εισόδου και (β) ενσωματώνουν γνώση από τα δεδομένα εκπαίδευσης. Αυτές είναι πολύ ενδιαφέρουσες ιδιότητες, που υποδηλώνουν επίσης ότι υπάρχει περιθώριο βελτίωσης της μεθόδου. Επιπλέον, το γεγονός ότι τα αποτελέσματα για τα ερωτήματα κανανέμονται σε ένα μεγαλύτερο εύρος μουσικών ειδών διατηρώντας ένα υψηλό ποσοστό ομοιότητας, καθιστούν την προτεινόμενη μέθοδο ένα κατάλληλο πλαίσιο για μελλοντική έρευνα στην μουσική ομοιότητα μεταξύ διαφορετικών ειδών μουσικής.

8.4 Η Διαδικτυακή Πλατφόρμα Ανάκτησης Μουσικής

Η παραπάνω μέθοδος ανάκτησης μουσικής πληροφορίας υιοθετήθηκε για την δημιουργία μιας πλατφόρμας εύρεσης παρόμοιας μουσικής. Το μεγαλύτερο μέρος της πλατφόρμας υλοποιήθηκε από τον Δρ. Ιωάννη Καρύδη στο πλαίσιο του έργου LangTERRA[19]. Η αρχιτεκτονική της πλατφόρμας παρουσιάζεται στο Σχ. 8.3. Από μια συλλογή υποσύνολο του MSD αποτελούμενη από 130.000 κομμάτια, εξήχθησαν τα χαρακτηριστικά χροιάς, ρυθμού και αρμονίας, προκειμένου να υπολογιστούν οι $N \times N$ αποστάσεις ($N = 130.000$) μεταξύ των κομματιών. Οι αποστάσεις κανονικοποιήθηκαν προκειμένου να υπολογιστεί και η συνολική απόσταση μεταξύ των κομματιών. Στη συνέχεια, από τις 4 ομάδες (χροιά, ρυθμός, αρμονία, συνολική) των $N \times N$ αποστάσεων αποθηκεύτηκε ένα υποσύνολο. Για κάθε κομμάτι και για κάθε τύπο ομοιότητας αποθηκεύτηκαν οι αποστάσεις των K πρώτων γειτόνων. Το K επιλέχθηκε ίσο με $K = 200$.

[19] http://www.langterra.eu/



Σχ. 8.4. Η αρχική σελίδα της πλατφόρμας ανάκτησης μουσικής.

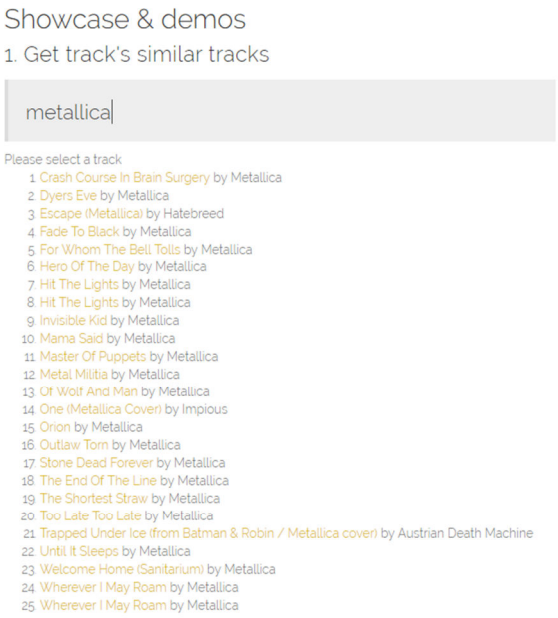
Οι $4 \times N \times K$ αποστάσεις αποθηκεύτηκαν τελικά σε μία Βάση Δεδομένων (ΒΔ) SQL μαζί με τα μεταδεδομένα της συλλογής, όπως τίτλος τραγουδιού, καλλιτέχνης, έτος έκδοσης κ.ο.κ. Στην συνέχεια υλοποιήθηκε η ιστοσελίδα http://mir.ilsp.gr η οποία υλοποιεί ένα διαδραστικό περιβάλλον επικοινωνίας με τη ΒΔ. Την υποδομή για την εγκατάσταση και λειτουργίας της πλατφόρμας (web server και ΒΔ) παρείχε το Ερευνητικό Κέντρο «Αθηνά».

Ο χρήστης έχει τη δυνατότητα να αναζητήσει τραγούδια και καλλιτέχνη στη ΒΔ, και να βρει τα πιο όμοια κομμάτια του ερωτήματος που θέτει. Επιπλέον μπορεί να επιλέξει ποιά απόσταση από τις 4 θα χρησιμοποιηθεί για τα αποτελέσματα. Αφού θέσει το ερώτημα, ο χρήστης μπορεί να ακούσει τα μουσικά αποτελέσματα.

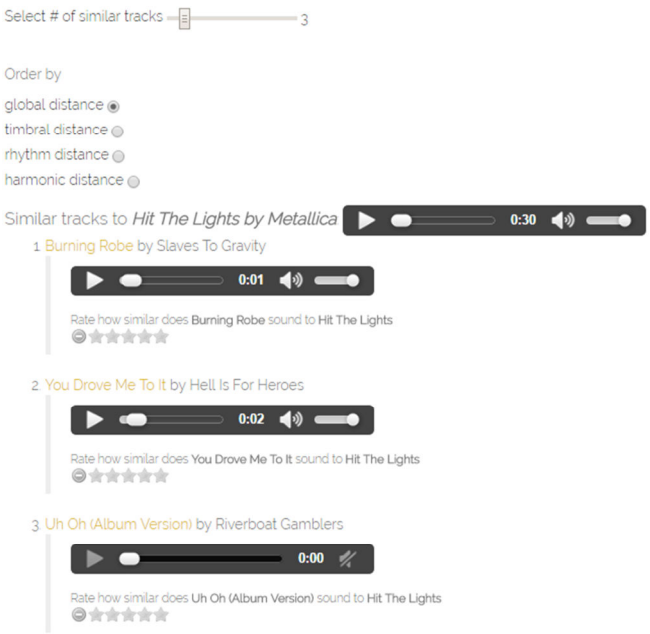
Λόγω περιορισμών στα πνευματικά δικαιώματα, τα ηχητικά αρχεία των αποτελεσμάτων δεν ήταν δυνατόν να μεταδοθούν από τον διακομιστή του Ε.Κ. «Αθηνά». Αντ'αυτού χρησιμοποιήσαμε την διαδικτυακή πύλη 7digital[20] , η οποία παρέχει τη δυνατότητα μέσω της Διεπαφής Προγραμματισμού Εφαρμογών (Application Programming Interface, API) να ακούσει κάποιος αποσπάσματα των 30'' ή 60''. Στιγμιότυπο της ιστοσελίδας παρουσιάζεται στο Σχ. 8.4.

Στην 1η σελίδα ζητείται από τον χρήστη να πληκτρολογήσει το όνομα ενός μουσικού ή ενός καλλιτέχνη που θέλει να θέσει ως ερώτημα στην ΒΔ. Στην συνέχεια ο χρήστης μπορεί να επιλέξει κάποιο από τα τραγούδια που εμφανίζονται βάσει του ονόματος που έθεσε (Σχ. 8.5). Αφού επιλέξει κάποιο από τα κομμάτια, εμφανίζεται μια νεά σελίδα με το ερώτημα και τα πιο όμοια αποτελέσματα. Ο χρήστης μπορεί να αλλάξει το πλήθος των αποτελεσμάτων καθώς και το είδος της απόστασης (Σχ. 8.6).

[20] https://www.7digital.com/



Σχ. 8.5. Η σελίδα αναζήτησης καλλιτέχνη ή μουσικού κομματιού της πλατφόρμας ανάκτησης μουσικής.



Σχ. 8.6. Η σελίδα αναζήτησης καλλιτέχνη ή μουσικού κομματιού της πλατφόρμας ανάκτησης μουσικής.

Πέρα από το ενδιαφέρον της παρουσίασης των αποτελεσμάτων, η πλατφόρμα μπορεί να αποκτήσει μια πάρα πολύ σημαντική χρήση. Όπως φαίνεται από το Σχ. 8.6, υπάρχει η δυνατότητα το χρήστης να αξιολογήσει τα αποτελέσματα της ομοιότητας. Αυτή η δυνατότητα είναι σε στάδιο υλοποίησης και δεν είναι ακόμα ενεργοποιημένη. Μετά το πέρας της ολοκλήρωσής της, οι αξιολογήσεις των χρηστών θα αποθηκεύονται στη βάση ενώ τα αποτελέσματα θα

αναδιαμορφώνονται μετά από την εισαγωγή μιας νέας αξιολόγησης στη ΒΔ. Μια τέτοια λειτουργία της πλατφόρμας θα είναι εξαιρετικής σημασίας καθώς θα δημιουργηθεί για πρώτη φορά μια συλλογή με επισημειώσεις μουσικής ομοιότητας.

Κεφάλαιο 9: Συμπεράσματα και Σχολιασμός

9.1 Συνεισφορά της διατριβής

Σκοπός της παρούσας διδακτορικής εργασίας είναι η αυτόματη ανάκτηση μουσικής πληροφορίας με έμφαση τον ρυθμό. Συγκεκριμένα μελετώνται τεχνικές για την συμπαγή και αποτελεσματική περιγραφή του ρυθμικού περιεχομένου του μουσικού σήματος, ώστε να χρησιμοποιηθεί στη συνέχεια σε προβλήματα ανάλυσης ρυθμού. Ξεκινώντας από τον ορισμό μιας συνάρτησης περιοδικότητας, η οποία αποτελεί το ρυθμικό ανάλογο του φάσματος, ξεδιπλώνονται μια σειρά από μέθοδοι αυτόματης ανάλυσης και επεξεργασίας ρυθμικού περιεχομένου που περιλαμβάνουν τεχνικές κατηγοριοποίησης (εξαγωγή μέτρου, χορευτικού στυλ), τεχνικές ανάλυσης (εξαγωγή τέμπο, παλμού) και τεχνικές ανάκτησης (μουσική ομοιότητα βάσει περιεχομένου). Οι προτεινόμενες μέθοδοι αξιολογήθηκαν εκτενώς και τα συγκριτικά αποτελέσματα με άλλες μεθόδους τις κατατάσσουν στις μεθόδους αιχμής. Οι ερευνητικές συνεισφορές μπορούν να συνοψιστούν στα ακόλουθα:

Εξαγωγή εύρωστης Συνάρτησης Περιοδικότητας

Προτάθηκε μια συνάρτηση περιοδικότητας (ΣΠ), η οποία όπως καταδείχτηκε και από τα πειραματικά αποτελέσματα, αποτελεί μια πολύ εύρωστη αναπαράσταση του ρυθμικού περιεχομένου. Σημαντικές συνεισφορές στον υπολογισμό της ΣΠ αποτελούν η χρήση του μετασχηματισμού σταθερού Q, καθώς και ο διαχωρισμός αρμονικών και κρουστών πηγών, όπου προτάθηκε μια απλή, αποτελεσματική και καινοτόμος μέθοδος διαχωρισμού κρουστών / αρμονικών πηγών με την υιοθέτηση τεχνικών επεξεργασίας εικόνας. Από τις δύο συνιστώσες του σήματος εξήχθησαν δύο διαφορετικοί τύποι χαρακτηριστικών έμφασης. Επιπλέον, διερευνήθηκαν με ποιοτικό τρόπο διάφορες τεχνικές εξαγωγής περιοδικότητας, και αναδείχτηκαν τα πλεονεκτήματα της χρήσης μιας συστοιχίας ταλαντωτών για την εξαγωγή της ΣΠ.

Εξαγωγή Χαρακτηριστικών από την Συνάρτηση Περιοδικότητας

Προκειμένου να εξαχθούν συμπαγή χαρακτηριστικά από την ΣΠ διερευνήθηκαν διάφορες τεχνικές, οι οποίες χωρίζονται σε δύο κατηγορίες, των χειρονακτικών χαρακτηριστικών και των χαρακτηριστικών που εξάγονται με τεχνικές μηχανικής μάθησης. Σχετικά με τα χειρονακτικά χαρακτηριστικά, προτάθηκε μια καινοτόμος μέθοδος ανάλυσης της ΣΠ σε «ζώνες τέμπο», τα οποία σχεδιάστηκαν για να αναπαραστήσουν αποτελεσματικά την έννοια της μουσικής ταχύτητας. Για την αυτόματη εξαγωγή χαρακτηριστικών χρησιμοποιήθηκαν δύο δημοφιλείς τεχνικές. Ένας γραμμικός μετασχηματισμός βασισμένος στην Ανάλυση σε Κύριες Συνιστώσες (Principal Component Analysis, PCA) και μια μη γραμμική απεικόνιση τεχνική βασισμένη στις Περιορισμένες Μηχανές Boltzmann (Restricted Boltzmann Machines, RBM). Τα χαρακτηριστικά που εξήχθησαν αξιολογήθηκαν εκτενώς για όλα τα επιμέρους προβλήματα ρυθμικής ανάλυσης που αντιμετωπίζονται στην παρούσα διατριβή, ενώ έγινε και βαθύτερη ανάλυση των χαρακτηριστικών που εξάγονται. Η συστηματική χρήση και αξιολόγηση της εξαγωγής των χαρακτηριστικών από την ΣΠ, καθώς και η χρησιμοποίηση των ίδιων χαρακτηριστικών τους για πληθώρα προβλημάτων αποτελεί μία μεγάλη συνεισφορά και καινοτομία της παρούσας διατριβής.

Κατηγοριοποίηση ρυθμού

Δύο προβλήματα που στην βιβλιογραφία αντιμετωπίζονταν με πολύπλοκες μεθόδους που ήταν σχεδιασμένες για αυτά τα προβλήματα, αντιμετωπίστηκαν με μια απλή προσέγγιση κατηγοριοποίησης. Τα προβλήματα της εύρεσης του μέτρου (meter estimation) και της κατάταξης σε ρυθμική κατηγορία (dance style classification), αντιμετωπίστηκαν ως προβλήματα κατηγοριοποίησης με τη χρήση SVM ταξινομητών στα χαρακτηριστικά που εξήχθησαν από την ΣΠ. Τα αποτελέσματα κατατάσσουν την προτεινόμενη μέθοδο στις μεθόδους αιχμής. Πρέπει να σημειωθεί, ότι και για τα δύο προβλήματα χρησιμοποιήθηκαν τα ίδια χαρακτηριστικά, τα οποία εξήχθησαν ανεξάρτητα από το πρόβλημα. Επομένως τα χαρακτηριστικά αυτά μπορούν να θεωρηθούν ως «πολλαπλής χρήσης» ή «γενικά», αφού μας δίνουν την δυνατότητα να αντιμετωπίσουμε πληθώρα σύνθετων προβλημάτων με με απλή μεθοδολογία.

Ορισμός εξαγωγής τέμπο ως πολλά υποπροβλήματα κατηγοριοποίησης

Μια διαδεδομένη προσέγγιση στη βιβλιογραφία για τον περιορισμό των λαθών οκτάβας κατά τον υπολογισμό του τέμπο, είναι η υιοθέτηση μιας μάσκας πρότερων κατανομών στην ΣΠ και η εφαρμογή ευριστικών μεθόδων για την επιλογή του τέμπο από την ΣΠ. Στην παρούσα διατριβή προτάθηκε μια καινοτόμος επαναπροσέγγιση του προσδιορισμού του τέμπο. Μια συστοιχία δυαδικών ταξινομητών ταχύτητας με διάφορες τιμές κατωφλίων συνδυάζονται για την εξαγωγή μιας μάσκας του τέμπο. Αυτή η μάσκα συνδυάζεται με την ΣΠ για την εξαγωγή του τελικού τέμπο. Η προτεινόμενη μέθοδος πέτυχε από τα υψηλότερα ποσοστά ακρίβειας που αναφέρονται στη διεθνή βιβλιογραφία σε πληθώρα συλλογών διαφορετικών ειδών μουσικής.

Εξαγωγή παλμού

Προτάθηκε μια μέθοδος εξαγωγής παλμού, η οποία ενσωματώνει την ίδια ανάλυση περιοδικότητας και την επεκτείνει με δύο επιπρόσθετα βήματα για τον υπολογισμό των τοπικών μεταβολών του τέμπο και για τον επαναπροσδιορισμό των συναρτήσεων έμφασης ώστε να περιγράφουν πιο αποτελεσματικά το ρυθμικό περιεχόμενο. Η προτεινόμενη μέθοδος αξιολογήθηκε στον διεθνή διαγωνισμό MIREX και κατατάχθηκε στις καλύτερες μεθόδους.

Αντιστρέψιμη Συνάρτηση Περιοδικότητας

Μια σημαντική συνεισφορά της παρούσας διατριβής αφορά στην επέκταση της συνάρτησης περιοδικότητας έτσι ώστε να είναι αντιστρέψιμη. Τα επιμέρους στοιχεία ανάλυσης περιοδικότητας απλοποιήθηκαν και τροποποιήθηκαν με τέτοιο τρόπο ώστε να είναι δυνατή η εξαγωγή ενός ακουστικού σήματος από την συνάρτηση περιοδικότητας. Αυτή αποτελεί μια πολύ σημαντική καινοτομία καθώς δεν έχει υπάρξει παρόμοια μέθοδος μέχρι τώρα. Δίνει τη δυνατότητα ανάπτυξης σε μελλοντικές εργασίες πολύ εξεζητημένων τεχνικών ανάλυσης και αναπαράστασης μουσικών σημάτων όπως για παράδειγμα την αλλαγή του ρυθμικού περιεχομένου ενός κομματιού, την «ηχητικοποίηση» ρυθμικών χαρακτηριστικών ή την δημιουργία τεχνητών ηχητικών παραδειγμάτων.

Ενσωμάτωση ρυθμικών χαρακτηριστικών σε μηχανή αναζήτησης βάσει περιεχομένου

Τα αποτελέσματα της έρευνας που έγινε στο επίπεδο της ρυθμικής ανάλυσης, συνδυάστηκε επιτυχώς με άλλες τεχνικές μουσικής ανάκτησης, που είναι εστιασμένες στην επεξεργασία και ανάλυση της χροιάς και της μουσικής αρμονίας, προκειμένου να υλοποιηθεί ένα σύστημα υπολογισμού μουσικής ομοιότητας και ανάκτησης βάσει περιεχομένου. Τα αρχικά χαρακτηριστικά ρυθμού, χροιάς και αρμονίας αποτέλεσαν την είσοδο σε Δίκτυα Βαθέως Πίστεως (Deep Belief Networks, DBN) για την εξαγωγή υψηλότερου επιπέδου χαρακτηριστικών. Η προτεινόμενη μέθοδος, λαμβάνει υπόψη τα εξαγόμενα χαρακτηριστικά ρυθμού, χροιάς και αρμονίας και αξιολογήθηκε στον διεθνή διαγωνισμό MIREX με ενθαρρυντικά αποτελέσματα. Επιπλέον χρησιμοποιήθηκε για να υπολογίσει την ομοιότητα μεταξύ των κομματιών μιας συλλογής αποτελούμενη από 130.000 μουσικά κομμάτια και τα αποτελέσματα ενσωματώθηκαν σε μια διαδικτυακή πλατφόρμα μουσικής ανάκτησης.

9.2 Κατευθύνσεις Μελλοντικής Έρευνας

Παρότι η διδακτορική διατριβή αποτελεί μια ολοκληρωμένη και συμπαγής προσέγγιση, θα προταθούν επεκτάσεις των παρόντων μεθόδων και μελλοντικοί ερευνητικοί άξονες. Συγκεκριμένα:

- Όλες οι μέθοδοι, εκτός από την αναζήτηση παλμού, κάνουν την υπόθεση του «σχεδόν σταθερού» ρυθμικού περιεχομένου. Αυτό αντανακλάται συνήθως στην άθροιση των ΣΠ ως προς τον χρόνο. Ωστόσο, η προτεινόμενη μέθοδος ανάλυσης περιοδικότητας μπορεί εύκολα να επεκταθεί σε ένα «τεμπογράφημα», δηλαδή να προκύπτει η ΣΠ ως συνάρτηση του χρόνου. Μια τέτοια αναπαράσταση θα μπορεί στην συνέχεια να χρησιμοποιηθεί για διάφορες εφαρμογές ανίχνευσης αλλαγών στο ρυθμικό περιεχόμενο όπως για παράδειγμα μεταβολές του τέμπο, αλλαγές του χρονικό κλειδιού, κατάτμηση του κομματιού στα μέρη του (εισαγωγή, ρεφρέν, κουπλέ).
- Τα εξαγόμενα χαρακτηριστικά από την ΣΠ μέσω των RBM και PCA και τα αποτελέσματα που προκύπτουν αναδεικνύουν ότι η πολύ καλή επίδοση όλων των μεθόδων οφείλεται πρωτίστως στο γεγονός ότι η προτεινόμενη ΣΠ αποτελεί μια εύρωστη αναπαράσταση του ρυθμικού περιεχομένου και δευτερευόντως στην αποτελεσματικότητα των RBM και PCA. Πρόχειρα πειραματικά αποτελέσματα έδειξαν ότι ακόμα και «τυχαίοι γραμμικοί μετασχηματισμοί» (random projections) αποδίδουν καλά αποτελέσματα. Επομένως είναι σημαντικό σε μελλοντική έρευνα να διερευνηθούν και άλλες τεχνικές μετασχηματισμού της ΣΠ και εξαγωγής χαρακτηριστικών.
- Παρότι η διατριβή εστίασε σε μεγάλο βαθμό στην χρήση αυτόματων μεθόδων επεξεργασίας στο πλαίσιο της ρυθμικής ανάλυσης, πολλές από τις παραμέτρους των προτεινόμενων υποσυστημάτων είναι ορισμένες «χειρονακτικά». Εξέχων ενδιαφέρον θα είχε σε μελλοντική έρευνα η χρήση αυτόματων τεχνικών μάθησης όλων των παραμέτρων. Για παράδειγμα, η κρουστική απόκριση των ταλαντωτών και η τάξη των φίλτρων παραγώγισης κατά τον υπολογισμό της ΣΠ θα μπορούσαν να

υπολογιστούν από δεδομένα μέσω κάποιας προσέγγισης εκμάθησης, που θα βελτιστοποιεί κάποια συνάρτηση κόστους. Μια τέτοια προσέγγιση θα μπορούσε να επεκταθεί περαιτέρω στην υιοθέτηση μεθόδων «Βαθείας Εκμάθησης» (Deep Learning) Τεχνητών Νευρωνικών Δικτύων (ΤΝΔ). Αφού θα οριστεί μια γενική αρχιτεκτονική ενός ΤΝΔ που θα περιγράφει ένα σύστημα ρυθμικής ανάλυσης, οι παράμετροι του δικτύου θα εκτιμώνται από δεδομένα.

- Παρόμοια προσέγγιση μπορεί να γίνει και στην εύρεση του παλμού. Οι παράμετροι της συνάρτησης απόστασης μεταξύ των παλμών μπορούν να υπολογιστούν με εκμάθηση (π.χ. Baum-Welch).
- Η μέθοδος υπολογισμού μιας αντιστρέψιμης ΣΠ έχει πολλά περιθώρια βελτίωσης τόσο σε σχέση με την επίδοση ανάλυσης (~5% χαμηλότερη επίδοση από την μη αντιστρέψιμη ΣΠ), όσο με την ανακατασκευή του αρχικού σήματος. Αυτό μπορεί να επιτευχθεί με την υιοθέτηση πιο αποτελεσματικών αναπαραστάσεων σε σχέση με τους ταλαντωτές, χρήση περισσότερων συναρτήσεων έμφασης και επαναπροσδιορισμού του δικτύου RBM. Επιπλέον η λογική υπολογισμού της «αντιστρέψιμης ΣΠ» μπορεί να επεκταθεί και πέραν του ρυθμού, π.χ. στην έννοια των «αντιστρέψιμων χαρακτηριστικών χροιάς».
- Η μέθοδος υπολογισμού ομοιότητας έχει πολλές παραμετροποιήσεις που πρέπει να ελεγχθούν πειραματικά. Για παράδειγμα πρέπει να γίνει εξέταση του κατά πόσο η χρήση ή όχι των DBN βελτιώνει τα αποτελέσματα και αν ναι, ποιά είναι η βέλτιστη διάσταση και βάθος του δικτύου. Επιπλέον μπορεί να διερευνηθεί τρόπος υπολογισμού των βέλτιστων τιμών των βαρών στις αποστάσεις των τριών ειδών χαρακτηριστικών. Επίσης πρέπει να διερευνηθεί σε βάθος κατά πόσο χαρακτηριστικά χρώματος όπως τα RCD μπορούν να είναι αποτελεσματικά σε ένα σύστημα υπολογισμού ομοιότητας βάσει περιεχομένου. Τέλος, θα είναι εξαιρετικής ερευνητικής αξίας η επέκταση της ιστοσελίδας μουσικής ομοιότητας έτσι ώστε να μπορεί να αποθηκεύει τις αξιολογήσεις των αποτελεσμάτων από τους χρήστες και να δημιουργηθεί μία συλλογή επισημειωμένης μουσικής ομοιότητας.

Παραπομπές

[Alonso2007] Alonso M., Richard G., David B., "Accurate Tempo Estimation Based on Harmonic + Noise Decomposition", *EURASIP Journal on Applied Signal Processing*, Volume 2007, Issue 1, January 2007

[Antonopoulos2007] Antonopoulos I., Pikrakis A., Theodoridis S., Cornelis O., Moelants D., and Leman M. "Music Retrieval by Rhythmic Similarity Applied on Greek and African Traditional Music," in *Proceedings of the 8th International Conference on Music Information Retrieval*, Vienna, Austria, 2007, (pp. 297-300).

[Aucouturier2002] Aucouturier, J. J., & Pachet, F., "Music Similarity Measures: What's the Use?" in *Proceedings of the 3th International Conference on Music Information Retrieval*, Paris, France, 2002.

[Aucouturier2008] Aucouturier, J. J., and Pachet F. "A Scale-Free Distribution of False Positives for a Large Class of Audio Similarity Measures." *Pattern Recognition* 41(1), 2008, pp. 272-284.

[Barrington2007] Barrington, L., Chan, A., Turnbull, D., & Lanckriet, G., "Audio Information Retrieval Using Semantic Similarity." in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 2, Honolulu, HI, USA, 2007 (pp. II-725).

[Bello2005a] Bello J. P., Daudet L., Abdallah S., Duxbury C., Davies M., and Sandler M.B., "A Tutorial on Onset Detection in Music Signals." *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pt. 2, Sep. 2005, pp. 1035-1047.

[Bello2005b] Bello J. P. and Pickens J., "A Robust Mid-Level Representation for Harmonic Content in Music Signals," in *Proceedings of 6th International Conference on Music Information Retrieval*, London, United Kingdom, 2005, (pp. 304-311).

[Bertin-Mahieux2011] Bertin-Mahieux T., Ellis D. P. W., Whitman B., and Lamere P.. "The Million Song Dataset," in *Proceedings of 12th International Conference on Music Information Retrieval*, Miami, USA, 2011 (pp. 591-596).

[Bock2011] Bock S. and Schedl M., "Enhanced Beat Tracking with Context-Aware Neural Networks," in *Proceedings of 14th International Conference on Digital Audio Effects*, Paris, France, 2011.

[Bogdanov2011] Bogdanov D., Serra J., Wack N., Herrera P., and Serra X. "Unifying Low-Level and High-Level Music Similarity Measures." *IEEE Transactions on Multimedia*, 13(4), 2011, pp. 687-701.

[Bosteels2008] Bosteels K., and Etienne E.K. "Fuzzy Audio Similarity Measures Based on Spectrum Histograms and Fluctuation Patterns." *Computational Intelligence in Multimedia Processing: Recent Advances*. Springer Berlin Heidelberg, 2008, pp. 213-231.

[Brown1991] Brown J. C. "Calculation of a Constant Q Spectral Transform." *The Journal of the Acoustical Society of America*, 89(1), 1991, pp. 425-434.

[Cemgil2000] Cemgil A.T., Taylan A., Kappen B., Desain P., and Honing H., "On Tempo Tracking: Tempogram Representation and Kalman Filtering." *Journal of New Music Research*, 24(9), 2000, pp. 259-273.

[Chang2011] Chang C.C., and Lin C.J., "LIBSVM: a Library for Support Vector Machines." *ACM Transactions on Intelligent Systems and Technology*,

2(3), 2011, pp. 27.

[Charbuillet2011] Charbuillet C., Tardieu D., and Peeters G. "Gmm Supervector for Content Based Music Similarity," in *Proceedings of 14th International Conference on Digital Audio Effects*, Paris, France, 2011.

[Cooper1963] Cooper G., and Meyer L. B., *The Rhythmic Structure of Music*. Vol. 118. University of Chicago Press, 1963.

[Cornelis2013] Cornelis O., Six J., Holzapfel A., and Leman M., "Evaluation and Recommendation of Pulse and Tempo Annotation in Ethnic Music." *Journal of New Music Research*, 42(2), 2013, pp.131-149.

[Davies2007] Davies M., Plumbley M., "Context-Dependent Beat Tracking of Musical Audio." *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 3, March 2007.

[Davies2009] Davies M.E.P, Degara N., and Plumbley M., "Evaluation Methods for Musical Audio Beat Tracking Algorithms," Queen Mary University of London, Centre for Digital Music, Tech. Rep. C4DM-TR-09-06 (2009).

[Desain1992] Desain P., "A (de) Composable Theory of Rhythm Perception." *Music Perception : An Interdisciplinary Journal*, 1992, pp. 439-454.

[Dixon2001] Dixon S., "Automatic Extraction of Tempo and Beat from Expressive Performances." *Journal of New Music Research*, 30(1), 2001, pp. 39-58.

[Dixon2004] Dixon S., Gouyon S., and Widmer G., "Towards Characterisation of Music via Rhythmic Patterns," in *Proceedings of 5th International Conference on Music Information Retrieval*, Barcelona, Spain, 2004.

[Drake2000a] Drake C., Penel A., and Bigand E.. "Why Musicians Tap Slower than non Musicians." *Rhythm Perception and Production*, 2000, pp. 245-248.

[Drake2000b] Drake C., Jones M.R., and Baruch C., "The Development of Rhythmic Attending in Auditory Sequences: Attunement, Referent Period, Focal Attending," *Cognition* 77, no. 3, 2000, pp. 251-288.

[Duong2011] Duong N.Q.K, Tachibana H., Vincent E., Ono N., Gribonval R., and Sagayama S., "Multichannel Harmonic and Percussive Component Separation by Joint Modeling of Spatial and Spectral Continuity," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 2011.

[Duxbury2002] Duxbury C., Sandler M., and Davies M., "A hybrid approach to musical note onset detection," in *Proceedings of 5th International Conference on Digital Audio Effects*, Hamburg, Germany, 2002 (pp. 33-38).

[Eck2005] Eck D., and Casagrande N., "Finding Meter in Music Using an Autocorrelation Phase Matrix and Shannon Entropy," in *Proceedings of 6th International Conference on Music Information Retrieval*, London, UK, 2005 (pp. 504-509).

[Eerola2004] Eerola T. and Toiviainen P., "Digital Archive of Finnish Folk Tunes," Computer Database, University of Jyvaskyla, 2004.

[Ellis2007] Ellis D. "Beat Tracking by Dynamic Programming." *Journal of New Music Research*, 26(1), 2007, pp. 51-60.

[Ellis2008] Ellis D., Cotton C., and Mandel M., "Cross-Correlation of Beat-Synchronous Representations for Music Similarity," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal*

Processing, Las Vegas, USA, April 2008. (pp. 57-60).

[Elowsson2013] Elowsson A., and Friberg A., "Modelling Perception of Speed in Music Audio," in *Proceedings of the Sound and Music Computing Conference*, Stockholm, Sweden, 2013 (pp. 735-741).

[Eronen2001] Eronen A. "Automatic Musical Instrument Recognition." Memoire de DEA, Tempere University of Technology, April 2001, 178.

[Eronen2010] Eronen A., and Klapuri A, "Music Tempo Estimation with k-NN Regression." *IEEE Transactions on Audio, Speech, and Language Processing*, 18(1), 2010, pp. 50-57.

[Eyben2010] Eyben, F., Bock, S., Schuller, B., and Graves, A., "Universal Onset Detection with Bidirectional Long Short-Term Memory Neural Networks," in *Proceedings of 11th International Conference on Music Information Retrieval*, Utrecht, Netherlands, 2010.

[FitzGerald2009] FitzGerald D., Coyle E., and Cranitch M., "Using Tensor Factorisation Models to Separate Drums from Polyphonic Music," in *Proceedings of 12th International Conference on Digital Audio Effects*, Como, Italy, 2009.

[FitzGerald2010] Fitzgerald D., "Harmonic/Percussive Separation Using Median Filtering," in *Proceedings of 13th International Conference on Digital Audio Effects*, Graz, Austria, 2010.

[Flexer2010] Flexer A., Schnitzer D., Gasser M. and Pohle T., "Combining Features Reduces Hubness in Audio Similarity," in *Proceedings of 11th International Conference on Music Information Retrieval*, Utrect, Netherlands, 2010.

[Flexer2012] Flexer A., Schnitzer D., and Schluter J. "A MIREX meta-analysis of Hubness in Audio Music Similarity," in *Proceedings of 13th International Conference on Music Information Retrieval*, Porto, Portugal, 2012.

[Frierer2004] Frieler K., "Beat and Meter Extraction Using Gaussified Onsets," in *Proceedings of 5th International Conference on Music Information Retrieval*, Barcelona, Spain, 2004.

[Gasser2009] Gasser M, and Flexer A,. "Fm4 soundpark: Audio-Based Music Recommendation in Everyday Use," in *Proceedings of the 6th Sound and Music Computing Conference*, Porto, Portugal. 2009.

[Gillet2008] Gillet O. and Richard G., "Transcription and Separation of Drum Signals from Polyphonic Music." *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 3, 2008, pp. 529-540.

[Gkiokas2010] Gkiokas A., Katsouros V., and Carayiannis G., "Tempo Induction Using Filterbank Analysis and Tonal Features," in *Proceedings of the 11th International Conference on Music Information Retrieval*, Utrecht, Netherlands, 2010.

[Gkiokas2012a] Gkiokas A., Katsouros V., Carayannis G., and Stafylakis T., "Music Tempo Estimation and Beat Tracking by Applying Source Separation and Metrical Relations," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, 2012 (pp. 421-424).

[Gkiokas2012b] Gkiokas A., Katsouros V., and Carayannis G., "Reducing Tempo Octave Errors by Periodicity Vector Coding and SVM Learning," in *Proceedings of 13th International Conference on Music Information Retrieval*, Porto, Portugal, 2012.

[Goh2010] Goh H., Thome N., and Cord M., "Biasing Restricted Boltzmann

Machines to Manipulate Latent Selectivity and Sparsity," in *Processings of the NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, Vancouver, Canada, 2010

[Goto1998] Goto M., and Muraoka T., "Music Understanding at the Beat Level: Real-Time Beat Tracking for Audio Signals." *Computational Auditory Scene Analysis*, 1998, pp. 157-176.

[Gouyon2004a] Dixon S., Gouyon F., and Widmer G., "Towards Characterisation of Music via Rhythmic Patterns," in *Proceedings of 5th International Conference on Music Information Retrieval*, Barcelona, Spain, 2004.

[Gouyon2004b] Gouyon F. and Dixon S., "Dance Music Classification: A Tempo-Based Approach," in *Proceedings of 5th International Conference on Music Information Retrieval*, Barcelona, Spain, 2004.

[Gouyon2004c] Gouyon F., Dixon S., Pampalk E., and Widmer G., "Evaluating Rhythmic Descriptors for Musical Genre Classification," in *Proceedings of the AES 25th International Conference*, London, UK (pp. 196-204).

[Gouyon2005] Gouyon F. and Dixon S., "A Review of Automatic Rhythm Description Systems," *Computer Music Journal*, 29(1), pp. 34-54, 2005.

[Gouyon2006] Gouyon F., Klapuri A., Dixon S., Alonso M., Tzanetakis G., Uhle C., and Cano P., "An Experimental Comparison of Audio Tempo Induction Algorithms", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 5, September 2006, pp. 1832-1844.

[Hainsworth2004] Hainsworth S.W., Techniques for the Automated Analysis of Musical Audio, Ph.D. thesis, University of Cambridge, UK, September 2004.

[Hamel2010] Hamel P. and Eck D. "Learning Features from Music Audio with Deep Belief Networks," in *Proceedings of the 11th International Conference on Music Information Retrieval*, Utrecht, Netherlands, 2010.

[Helen2005] Helen M. and Virtanen T., "Separation of Drums from Polyphonic Music Using Non-Negative Matrix Factorisation and Support Vector Machine," in *Proceedings of the European Signal Processing Conference*, Anatalya, Turkey, 2005 (pp. 1-4).

[Hinton1983] Hinton G, Sejnowski T., "Optimal Perceptual Inference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, New York, 1983, (pp. 448-453).

[Hinton2002] Hinton G., "Training Products of Experts by Minimizing Contrastive Divergence." *Neural Computation* 14(8), 2002, pp. 1771-1800.

[Hinton2006] Hinton G, Osindero S., and The Y., "A Fast Learning Algorithm for Deep Belief Nets." *Neural Computation*, 18, 2006, pp. 1527-1554.

[Hinton2010] Hinton G., "A Practical Guide to Training Restricted Boltzmann Machines." *Momentum* 9(1), 2010.

[Hockman2008] Hockman J., Bello J., Davies M,, and Plumbley M., "Automated Rhythmic Transformation of Musical Audio," in *Proceedings of 11th International Conference on Digital Audio Effects*, Espoo, Finland, 2008 (pp. 177-180).

[Hockman2010] Hockman J. and Fujinaga I., "Fast vs Slow: Learning Tempo

Octaves from User Data," in *Proceedings of the 11th International Conference on Music Information Retrieval*, Utrecht, Netherlands, 2010.

[Holzapfel2011a] Holzapfel A. and Stylianou Y., "Scale Transform in Rhythmic Similarity of Music." *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 19, No. 1, 2011, pp. 176-185.

[Holzapfel2011b] Holzapfel A., Velasco G.A., Holighaus N., Dorfler M., and Flexer A., "Advantages of Nonstationary Gabor Transforms in Beat Tracking," in *Proceedings of the 1st International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies*, Scottsdale, AZ, USA, 2011 (pp. 45-50).

[Holzapfel2012] Holzapfel A., Davies M., Zapata J., Oliveira J., and Gouyon F., "Selective Sampling for Beat Tracking Evaluation." *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 9, 2012, pp. 2539-2548.

[Holzapfel2014] Holzapfel A., Krebs F., and Srinivasamurthy A., "Tracking the "Odd": Meter Inference in a Culturally Diverse Music Corpus," in *Proceedings of the 15th International Conference on Music Information Retrieval*, Taipei, Taiwan, 2014.

[Hsu2002] Hsu C.W. and Lin C.J., "A Comparison of Methods for Multiclass Support Vector Machines." *IEEE Transactions on Neural Networks*, Vol. 13(2), March 2002, pp.415-425.

[Klapuri1999] Klapuri A., "Sound Onset Detection by Applying Psychoacoustic Knowledge," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Phoenix, USA, 1999 (pp. 3089-3092).

[Klapuri2006] Klapuri A., Eronen A. and Astola J., "Analysis of the Meter of Acoustic Musical Signals." *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 14, No. 1, January 2006, pp. 342-355.

[Krebs2013] Krebs F., Bock S., and Widmer G., "Rhythmic Pattern Modeling for Beat and Downbeat Tracking in Musical Audio," in *Proceedings of the 14th International Conference on Music Information Retrieval*, Curitiba, Brazil, 2013 (pp. 227-232).

[Lapidaki2000] Lapidaki E., "Stability of Tempo Perception in Music Listening." *Music Education Research* Vol. 2, No. 1, 2000, pp. 25-44.

[Large1994] Large E. and Kolen J., "Resonance and the Perception of Musical Meter", *Connection Science* 6(1), 1004, pp. 177-208.

[Laroche2003] Laroche, J., "Efficient Tempo and Beat Tracking in Audio Recordings." *Journal of the Audio Engineering Society*, 51(4), 2003, pp. 226-233.

[Lerdahl1985] FLerdahl F. and Jackendoff R., *A Generative Theory of Tonal Music*. MIT press, 1985.

[Levitin1996] Levitin D. J., and Cook P. R. "Memory for Musical Tempo: Additional Evidence that Auditory Memory is Absolute." *Perception & Psychophysics*, 58, 1996, pp. 927-935.

[Levy2008] Levy M. and Sandler M.,, "Structural Segmentation of Musical Audio by Constrained Clustering." *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 156, No. 2, 2008, pp. 318-326.

[Levy2011] Levy M. "Improving Perceptual Tempo Estimation with Crowd-Sourced Annotations," in *Proceedings of the 12th International*

Conference on Music Information Retrieval, Miami, USA, 2011.

[Liang 2011] Liang D., , Gu H., , H. and O'Connor H., "Music Genre Classification with the Million Song Dataset," *Machine Learning Department, CMU*, 2011.

[Logan2011] Logan B. and Salomon A., "Music Similarity Function Based on Signal Analysis," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, Tokyo, Japan, 2001.

[Lukashevich2010] Lukashevich H., Dittmar C., and Bastuck C., "Applying Statistical Models and Parametric Distance Measures for Music Similarity Search." *Advances in Data Analysis, Data Handling and Business Intelligence*, Springer Berlin Heidelberg, 2010, pp. 409-418.

[Madison2010] Madison G. and Paulin J., "Ratings of Speed in Real Susic as a Function of Both Original and Manipulated Beat Tempo." *The Journal of the Acoustical Society of America*, 128(5), 2010, pp. 3032-3040.

[Mandel2006] Mandel M. I., Poliner G.E., and Ellis, D.P., "Support Vector Machine Active Learning for Music Retrieval." *Multimedia Systems*, 12(1), 2006, pp. 3-13.

[Maragos1987] Maragos P., and Schafer R.W., "Morphological Filters - Part II: Their Relations to Median, Order-Statistic, and Stack Filters." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 35, No. 8, Aug. 1987, pp.1170-1184.

[Marsi1996] Masri P., *Computer Modeling of Sound for Transformation and Synthesis of Musical Signals.* Ph.D. dissertation, Univ. of Bristol, Bristol, U.K., 1996.

[McKinney2004] McKinney M.F. and Moelants D., "Deviations from the Resonance Theory of Tempo Induction," in *Proceesings of the Conference on Interdisciplinary Musicology*, Graz, Austria, 2004.

[McKinney2007] McKinney M.F, Moelants D, Davies ME, Klapuri A. "Evaluation of Audio Beat Tracking and Music Tempo Extraction Algorithms." *Journal of New Music Research*, 36(1), 2007, pp. 1-6.

[Mohammed2012] Mohamed A.R., Dahl G.E., Hinton G., "Acoustic Modeling Using Deep Belief Networks." *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1), 2012, pp. 14-22.

[Nair2009] Nair V., Hinton G.E., "3D Object Recognition with Deep Belief Nets." in *Advances in Neural Information Processing Systems*, 2009 pp. 1339-1347.

[Nair2010] Nair, V. and Hinton, G.E., "Rectified Linear Units Improve Restricted Boltzmann Machines," in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010 (pp. 807-814).

[Nam2011] Nam J., Ngiam J, Lee H. and Slaney M. "A Classification-Based Polyphonic Piano Transcription Approach Using Learned Feature Representations," in *Proceedings of the 12th International Conference on Music Information Retrieval*, Miami, USA, 2011.

[Ni2012] Ni Y., McVicar M., Santos-Rodriguez R., De Bie T., "Using Hyper-genre Training to Explore Genre Information for Automatic Chord Estimation," in *Proceedings of the 11th International Conference on Music Information Retrieval*, Porto, Portugal, 2012 (pp. 109-114).

[Noorden1999] van Noorden, L., and Moelants D., "Resonance in the Perception of Musical Pulse." *Journal of New Music Research* 28(1), 1999, pp.

43-66.

[Oliveira2012] Oliveira J.L., Davies M., Gouyon F. and Reis P., "Beat Tracking for Multiple Applications: A Multi-Agent System Architecture with State Recovery." *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 10, December 2012, pp. 2696-2706.

[Ono2008a] Ono N., Miyamoto K., Le Roux J., Kameoka H., and Sagayama S., "Separation of a Monaural Audio Signal into Harmonic/Percussive Components by Complementary Diffusion on Spectrogram," in *Proceedings of the EUSIPCO 2008 European Signal Processing Conference*, Lausanne, Switcherland, Aug. 2008.

[Ono2008b] Ono N., Miyamoto K., Kameoka H., and Sagayama S., "A Real-Time Equalizer of Harmonic and Percussive Components in Music Signals," in *Proceedings of the 9th International Conference on Music Information Retrieval*, Philadelphia , USA, 2008 (pp. 139-144).

[Pachet2004] Pachet, F, and Aucouturier J. J., "Improving Timbre Similarity: How high is the sky?" *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004, pp. 1-13.

[Papadopoulos2011] Papadopoulos H. and Peeters G.,"Joint Estimation of Chords and Downbeats From an Audio Signal." *IEEE Transactions on Audio, Speech & Language Processing*, Vol. 19, No. 1, 2001, pp. 138-152.

[Parncutt1994] R. Parncutt. "A Perceptual Model of Pulse Salience and Metrical Accent in Musical Rhythms." *Music perception*, 11(4), 1994, pp. 409-464.

[Peeters2005] Peeters G., "Rhythm Classification Using Spectral Rhythm Patterns", in *Proceedings of the 6th International Conference on Music Information Retrieval*, London, UK, 2005

[Peeters2007] Peeters G., "Template-Based Estimation of Time Varying Tempo." *EURASIP Journal on Applied Signal Processing*, Volume 2007 Issue 1, 2007, pp. 158.

[Peeters2010] Peeters G., "Template-Based Estimation of Tempo: Using Unsupervised or Supervised Learning to Create Better Spectral Templates," in *Proceedings of 13th International Conference on Digital Audio Effects*, Graz, Austria, 2010.

[Peeters2011a] Peeters G., "Spectral and Temporal Periodicity Representations of Rhythm for the Automatic Classification of Music Audio Signal." *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 19, No. 5, 2011, pp. 1242-1252.

[Peeters2011b] Peeters G. and Papadopoulos H., "Simultaneous Beat and Downbeat Tracking Using a Probabilistic Framework: Theory and Large-Scale Evaluation." *IEEE Transactions on Audio, Speech, and Language Processing*, Vol 19, No. 6, Aug. 2011, pp. 1754-1769.

[Peeters2012] Peeters G. and Flocon-Cholet J., "Perceptual Tempo Estimation Using GMM-Regression," in *In Proceedings of the 2nd International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies*, Nara, Japan, 2012.

[Percival2014] Percival G. and Tzanetakis G., "Streamlined Tempo Estimation Based on Autocorrelation and Cross-Correlation with Pulses." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. Vol. 22, No. 12, Dec. 2014, pp. 1765-1776.

[Pikrakis2004] Pikrakis A., Antonopoulos I., and Theodoridis S., "Music Meter and Tempo Tracking from Raw Polyphonic Audio," in *Proceedings of the 5th International Conference on Music Information Retrieval*,

Barcelona, Spain, 2004.

[Pohle2009] Pohle T., Schnitzer D., Schedl M., Knees P., and Widmer G., "On Rhythm and General Music Similarity," in *Proceedings of the 10th International Conference on Music Information Retrieval*, Kobe, Japan, 2009.

[Radovanovic2010] Radovanovic M, Nanopoulos A, Ivanovic M. "Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data." *The Journal of Machine Learning Research*, Vol. 11, Sept. 2010, pp. 2487-2531.

[Ranjani2013] Ranjani H.G., and Sreenivas T.V., "Hierarchical Classification of Carnatic Music Forms," in *Proceedings of the 14th International Conference on Music Information Retrieval*, Curitiba, Brazil, 2013 (pp. 251-256).

[Robertson2007] Robertson A. and Plumbley M.D., "B-Keeper: A Beat-Tracker for Live Performance," in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, New York, USA, June, 2007, (pp. 234-237).

[Robine2005] Robine M., Hanna P., and Lagrange M., "Meter Class Profiles for Music Similarity and Retrieval," in *Proceedings of the 6th International Conference on Music Information Retrieval*, London, UK, 2005 (pp. 639-644).

[Rosenthal1992] Rosenthal D., *Machine Rhythm--Computer Emulation of Human Rhythm Perception*. PhD diss., Massachusetts Institute of Technology, 1992.

[Salakhutdinov2007] Salakhutdinov R., Mnih A. and Hinton G., "Restricted Boltzmann Machines for Collaborative Filtering," in *Proceedings of the 24th International Conference on Machine Learning*, Corvalis, OR, USA, 2007.

[Schaffrath1995] Schaffrath H. and Huron D., "The Essen Folksong Collection in the Humdrum Kern Format," Menlo Park, CA: Center for Computer Assisted Research in the Humanities, 1995.

[Scheirer1998] Scheirer E., "Tempo and Beat Analysis of Acoustic Musical Signals." *The Journal of the Acoustical Society of America*, Vol. 103, No. 1, January 1998, pp. 588-601.

[Schindler2012] Schindler A., Mayer R., and Rauber A., "Facilitating Comprehensive Benchmarking Experiments on the Million Song Dataset," in *Proceedings of the 13th International Conference on Music Information Retrieval*, Porto, Portugal, 2012 (pp. 469-474).

[Schluter2011] Schluter, J. and Osendorfer C. "Music Similarity Estimation with the Mean-Covariance Restricted Boltzmann Machine," in *Machine Learning and Applications and Workshops (ICMLA)*, 2011 10th International Conference on. Vol. 2. IEEE, 2011 (pp. 118-123).

[Schluter2013] Schluter J. "Learning Binary Codes for Efficient Large-Scale Music Similarity Search." in *Proceedings of the 14th International Conference on Music Information Retrieval*, Curitiba, Brazil, 2013.

[Schnitzer2011] Schnitzer D., Flexer A., Schedl M. and Widmer G., "Using Mutual Proximity to Improve Content-Based Audio Similarity," in *Proceedings of the 12th International Conference on Music Information Retrieval*, Miami, FL, USA, 2011.

[Schnitzer2012] Schnitzer, D., Flexer, A., Schedl, M., & Widmer, G. "Local and Global Scaling Reduce Hubs in Space." *The Journal of Machine*

Learning Research, 13(1), 2012, pp. 2871-2902.

[Schorkhuber2010] Schorkhuber, C., and Klapuri, A., "Constant-Q Transform Toolbox for Music Processing," in *7th Sound and Music Computing Conference*, Barcelona, Spain ,2010

[Schreiber2014] Schreiber H. and Muller M., "Exploiting Global Features for Tempo Octave Correction," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, 2014 (pp. 639-643).

[Serra1982] Serra J., *Image Analysis and Mathematical Morphology v. 1*, Academic Press, 1982.

[Seyerlehner2007] Seyerlehner K., Widmer G., and Schnitzer D., "From Rhythm Patterns to Perceived Tempo", in *Proceedings of the 8th International Conference on Music Information Retrieval*, Vienna, Austria, 2007

[Seyerlehner2010a] Seyerlehner, K. Schedl, M. Pohle, T. Knees, P. "Using Block-Level Features for Genre Classification, Tag Classification and Music Similarity Estimation," in *online Proceedings of the 6th Annual Music Information Retrieval Evaluation eXchange* (MIREX-2010), Utrecht, Netherlands, 2010.

[Seyerlehner2010b] Seyerlehner, K., Widmer, G., & Pohle, T., "Fusing Block-Level Features for Music Similarity Estimation," in *Proceedings of 13th International Conference on Digital Audio Effects*, Graz, Austria, 2010.

[Smith2010] Smith L.M., Beat-Critic: Beat-Tracking OctaveError Identification by Metrical Profile Analysis," in *Proceedings of the 11th International Conference on Music Information Retrieval*, Utrecht, Netherlands, 2010.

[Stark2007] Stark, A.M., Plumbley M. D., and Davies M., "Audio Effects for Real-Time Performance Using Beat Tracking," in *Proceedings of the 122nd AES Convention*, Vienna, Austria, May 2007.

[Sturm2014] Sturm, B. "The State of the Art ten Years after a State of the Art: Future Research in Music Information Retrieval." *Journal of New Music Research*, 43(2), 2014, pp. 147-172.

[Thoshkahna2011] Thoshkahna B. and Ramakrishnan K.R., "A Postprocessing Technique for Improved Harmonic/Percussion Separation for Polyphonic Music," in *Proceedings of the 12th International Conference on Music Information Retrieval*,, Miami, USA, October 2011.

[Tieleman2008] Tieleman T., "Training Restricted Boltzmann Machines Using Approximations to the Likelihood Gradient," in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008 (pp. 1064-1071).

[Toiviainen2006] Toiviainen P. and Eerola T., "Autocorrelation in Meter Induction: the Role of Accent Structure," *The Journal of the Acoustical Society of America*, 119(2), 2006, pp. 1164-1170.

[Tzanetakis2002] Tzanetakis G. and Cook P., "Musical Genre Classification of Audio Signals." *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 5, 2002, pp. 293-302.

[Tzanetakis2013] Tzanetakis G. and Percival G., "An Effective, Simple Tempo Estimation Method Based on Self-Similarity and Regularity," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, 2013.

[Velasco2011] Velasco, G. A., Holighaus, N., Dorfler, M., and Grill, T.,

"Constructing an Invertible Constant-Q Transform with Non-Stationary Gabor Frames," in *Proceedings of 14th International Conference on Digital Audio Effects*, Paris, France, 2011.

[Vincent2006] Vincent E., Fevotte C., and R. Gribonval R., "Performance Measurement in Blind Audio Source Separation," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 14, No. 4, 2006, pp. 1462-1469.

[Vincent2012] Vincent E., Araki S., Theis F.J., Nolte G., Bofill P., Sawada H., Ozerov A., Gowreesunker B.V., Lutter D., and Duong N.Q.K., "The Signal Separation Evaluation Campaign (2007-2010): Achievements and remaining challenges." *Signal Processing*, 92(8), 2012, pp. 1928-1936.

[Witten1999] Witten I. and Frank E., Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation. Morgan Kaufmann, San Fransisco, CA, 1999.

[Yoshii2007] Yoshii K., Goto M., and Okuno H., "Drum Sound Recognition for Polyphonic Audio Signals by Adaptation and Matching of Spectrogram Templates with Harmonic Structure Suppression." *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, 2007, pp. 333-345.

[Zapata2013] Zapata, J. R., & Gomez, E. "Using Voice Suppression Algorithms to Improve Beat Tracking in the Presence of Highly Predominant Vocals," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, 2013.

Παράρτημα Α': Δημοσιεύσεις Διατριβής

1. **Gkiokas A.**, Katsouros V., and Carayannis G., "Towards Multi-Purpose Spectral Rhythm Features: An Application to Dance Style, Meter and Tempo Estimation", *IEEE/ACM Transactions on Audio Speech and Language Processing*, Volume 24, Issue 11, Nov. 2016, pp. 1885-1896.
2. **Gkiokas A.**, Lattner S, Katsouros V., Flexer A., and Carayannis, G. "Towards an Invertible Rhythm Representation," in *Proceedings of the 18th International Conference on Digital Audio Effects (DAFX15).* Trondheim, Norway, December 2015.
3. **Gkiokas A**., Katsouros, V., and Carayannis, G. "Deploying Deep Belief Nets for content based audio music similarity," in *Proceedings of the 5th IEEE International Conference Information, Intelligence, Systems and Applications (IISA)*, Chania, Greece, July 2014.
4. Vincent, E., **Gkiokas A**., Schnitzer, D., and Flexer, A., "An Investigation of Likelihood Normalization for Robust ASR," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Singapore, September 2014.
5. **Gkiokas A**., Katsouros V. and Carayannis G., "Recuding Tempo Octave Errors by Periodicity Vector Coding and SVM Learning," in *Proceedings of the 13th International Conference on Music Information Retrieval*, Portugal, October 2012
6. **Gkiokas, A.**, Papavassiliou, V., Katsouros, V. and Carayannis, G. "Deploying Nonlinear Image Filters to Spectrogram for Harmonic/Percussive Separation," *in Proceedings of the 15th International Conference on Digital Audio Effects (DSFX12).* York, UK., September 2012.
7. **Gkiokas, A.,** Katsouros, V., Carayannis, G. and Stafylakis, T. "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.
8. **Gkiokas A.**, Katsouros V. and Carayiannis G., "Tempo Induction Using Filterbank Analysis and Tonal Features," in *Proceedings of the 11th International Conference on Music Information Retrieval*, Utrecht, Netherlands, August 2010.